

Mobile Phone-based Rheumatic Heart Disease Detection



David Brian Springer

St. Edmund Hall

Supervised by

Professor Lionel Tarassenko

Professor Gari D. Clifford

Trinity Term, 2015

This thesis is submitted to the Department of Engineering Science, University of Oxford, in fulfillment of the requirements for the degree of Doctor of Philosophy

David Brian Springer
St. Edmund Hall

Doctor of Philosophy
Trinity Term, 2015

Mobile Phone-based Rheumatic Heart Disease Detection

Abstract

Rheumatic heart disease (RHD), the permanent damage of the heart valves caused by an untreated “strep throat” infection, is the leading cause of cardiovascular mortality and morbidity in children and young adults worldwide. Simple penicillin treatment after the early diagnosis of RHD can stop recurring bouts of the condition, which lead to the most severe valvulopathy, and ultimately, heart failure. However, RHD is an under-diagnosed condition in the developing world, as such a diagnosis requires, at a minimum, a trained clinician to perform auscultation to detect pathological heart sounds. Trained medical personnel are scarce in the countries where RHD is most prevalent. A low-cost, mobile phone-based automatic diagnostic tool offers a potential solution, allowing a non-medically trained individual to screen for RHD in those countries.

An essential feature of such a device is feedback on the *signal quality* of heart sound recordings. The first major contribution of this thesis is the investigation of features and algorithms for the automatic signal quality assessment of heart sound recordings. These algorithms are able to differentiate between good- and poor-quality recordings in over 80% of cases when using both a low-cost mobile phone-based stethoscope and an electronic stethoscope.

Once the quality of recordings is ensured, the positions of the first and second heart sounds need to be located in a process called *segmentation*. This thesis extends the state-of-the-art hidden semi-Markov models by: investigating additional features; extending the Viterbi algorithm; incorporating logistic regression into the model to form a hybrid generative-discriminative model; and investigating a discriminative duration-dependent probabilistic model – a conditional random field. These extensions are found to outperform the state-of-the-art method.

Lastly, the period between the first and second heart sounds can be analysed for the presence of a pathological murmur. This thesis presents *automated systolic murmur classification* algorithms based on wavelet and mel-frequency cepstral coefficient-based features along with denoising via cycle averaging. These algorithms outperform three methods from the literature when detecting valvulopathy, while also outperforming a cardiologist and commercial software when detecting RHD in mobile phone-based heart sound recordings.

Acknowledgments

I would have neither started nor finished this thesis were it not for the support and guidance of many people over the last couple of years.

Firstly, my sincerest gratitude to my supervisors, Prof. Lionel Tarassenko and Prof. Gari Clifford.

Gari, two examples of your guidance stand out for me during my DPhil; the first was when I was discussing approaches to identifying heart sounds in noisy signals with the lab. You overheard, stuck your head around the corner, and said, “try hidden Markov models”. That ended up being my favourite chapter in my thesis. The second was when I emailed you about the generator noise in the background of my heart sound recordings, which I thought was completely fatal to my research. Your positivity and excitement about the signal processing methods we could apply was such a relief in one of the darker moments of my DPhil. Thank you.

Lionel, thank you for making me feel like I was your only student during every meeting we had, and for taking such an interest in everything I did outside of my studies. In our first meeting you spent as much time asking me about my extracurricular activities as my interest in biomedical signal processing, which I greatly appreciated. Your scrupulous proofreading never ceased to amaze me.

I’d like to thank my examiners, Prof. Stephen Payne and Prof. Leif Sörnmo, for the fulfilling conversation we had about my thesis. A friend said to me that I should try to enjoy my viva, as it is the one time that two people who are as interested in my research as me would sit down and have a three hour discussion about it with me. That was exactly the feeling I left with - thank you.

I would not have found this project if it were not for the guidance of Dr. Tom Brennan. Thank you, Tom, for initiating this project and allowing me to continue it. Thank you, too, for getting me involved in the StAR study, during which I learnt so much.

Thank you to my lab mates, in both Gari and Lionel's research groups, for the support, guidance and questioning. Without the discussions we had in the lab, or the questions you asked after my presentations, this thesis would be in a far worse state. To Lisa, I'm glad that there was someone working on a similar topic who I could turn to for advice (and for annotating so much data!). To Arvind, thank you for shouting "Hello Dave!" at me each time I walked into the lab.

Kirsty, Carmen, Liezel and the rest of the StAR team - working alongside you on StAR was one of the highlights of my DPhil. Thank you for providing such a fulfilling distraction during the times when I was bored of heart sounds. Kirsty, thank you for hosting me so often during my trips to Cape Town - I so enjoyed staying with you. It was during these times that I learnt how to question every fact and figure.

One of the most selfless contributions to this thesis was the work of Dr. Liesl Zühlke. Liesl, thank you so much for collecting the data I needed to complete this research, alongside all the data annotations and guidance you provided. It is no exaggeration to say that without your help, this thesis would not have been completed. Many thanks too to Prof. Bongani Mayosi, who oversaw the data collection in South Africa and who allowed me to use the data (after one of the most stressful meetings of my DPhil).

I would not have made it to Oxford were it not for my uncle and aunt, Paddy and Sue. Thank you for allowing me the privilege of attending St. Andrew's, which gave me the chance of getting a scholarship to Oxford, and which set me on a path to writing these acknowledgments today.

To my referees in my Rhodes applications, and in particular Mr. Terry, Mr. Clark and Dr. Wilde, thank you for believing in me. Thank you to the Rhodes selection committee for granting me the privilege of doing a DPhil at Oxford. I did not realise the magnitude of the gift of a Rhodes Scholarship five years ago, and I am only now beginning to grasp what a privilege the scholarship was. I hope to live up to it. To Rhodes house, thank you for your guidance on, and more often distraction from, this thesis. Thank you to the Rhodes class of

2011, of which I was equally proud and intimidated to be a part, for the many discussions, coffees, drinks, cycles and parties. I wish I had spent more time with you.

Thank you to Teddy Hall for their support over the last five years. Thank you to the Principal, Prof. Keith Gull, for his support, and to Prof. Rob Whittaker for being a superb Dean to work under. Thank you to the Dean Teams over the years for making the job more fun. Thank you to the MCR committee of 2011-2012 for making my job so easy. And to the Hockey team, of which I was so proud to be a part.

I feel so grateful for starting at Teddy Hall in 2010 alongside such an amazing group of people. To the Teddy Hall family, thank you for distracting me so effectively for the first three years of my Oxford career, and then supporting me in the two years when I needed it the most. You are my response to the question, “what was the best part about Oxford?”. A special thanks to Dan and Lucy, for letting me stay with you for the final stretch of thesis-writing mayhem.

To Clive, thank you for sharing the Oxford experience with me. With us chatting over cups of tea in your Upper bedroom at St. Andrew’s in 2004, to doing the same thing in your room in Trinity College in 2014, I look forward to see where we will be in 2024.

Thank you to Louise, for being more excited about the completion of this thesis than me, for being ever supportive during those “six to twelve months”, and without whom I would probably still be writing this thesis.

And finally, to my family, thank you for all the love and support, always. Mom, thank you for always making me feel that anything was achievable and that nothing was out of my reach. I can’t decide whether Dad would not believe he was in the acknowledgements of a Oxford engineering DPhil, or if he’d say, “of course Dave got a DPhil - he takes after his father”. I wish he could be at my graduation.

Relevant publications:

- [a] D. B. Springer, T. Brennan, L. J. Zühlke, H. Y. Abdelrahman, N. Ntusi, G. D. Clifford, B. M. Mayosi, and Lionel Tarassenko, "Signal quality classification of mobile phone-recorded phonocardiogram signals." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, Florence, Italy, pp. 1335-1339. IEEE.
- [b] D. B. Springer, T. Brennan, J. Hitzeroth, B. M. Mayosi, L. Tarassenko, and G. D. Clifford, "Robust heart rate estimation from noisy phonocardiograms." In *Computing in Cardiology (CinC)*, 2014, Cambridge, Massachusetts, USA, pp. 613-616. IEEE.
- [c] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Support vector machine hidden semi-Markov model-based heart sound segmentation." In *Computing in Cardiology (CinC)*, 2014, Cambridge, Massachusetts, USA, pp. 625-628. IEEE.
- [d] M. A. F. Pimentel, M. D. Santos, D. B. Springer, and G. D. Clifford. "Hidden semi-Markov model-based heartbeat detection using multimodal data and signal quality indices." In *Computing in Cardiology (CinC)*, 2014, Cambridge, Massachusetts, USA, pp. 553-556. IEEE.
- [e] M. A. F. Pimentel, M. D. Santos, D. B. Springer, and G. D. Clifford. "Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices." *Physiological measurement*, vol 36, no. 8, pp. 1717-1727, July 2015.
- [f] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic Regression HSMM-based Heart Sound Segmentation", *IEEE Transactions on Biomedical Engineering*, vol 63, no. 4, pp. 822-832, March 2016.

Contents

1	Introduction	1
1.1	Mobile health	2
1.1.1	The potential for mobile health in developing countries	2
1.1.2	The evidence for mobile health	4
1.1.3	Mobile health for heart sound analysis	5
1.2	Rheumatic heart disease	6
1.3	The cardiac cycle, heart sounds and murmurs	9
1.3.1	The heart structure	10
1.3.2	The cardiac cycle	12
1.3.3	Heart sounds and murmurs	13
1.3.4	Murmurs	18
1.4	The motivation for mobile phone-based rheumatic heart disease detection	24
2	Datasets and fieldwork	29
2.1	Mobile stethoscope design	29
2.2	Mobile stethoscope pilot dataset	31
2.3	Rheumatic heart disease dataset	33
2.4	Massachusetts Institute of Technology automated auscultation system (MAAS) dataset	39
2.5	Conclusion	41
3	Phonocardiogram signal quality classification	43
3.1	Background	44
3.1.1	Noise reduction in PCG signals	45
3.1.2	Identification of clean PCG segments	47
3.1.3	Signal quality of cardiac signals	49
3.1.4	PCG signal quality classification	51
3.2	Methods	52
3.2.1	Dataset	52
3.2.2	Annotations and annotator agreement	53
3.2.3	Combining annotations	54

3.2.4	Signal quality indices	55
3.2.5	Classification methods for signal quality classification	74
3.2.6	Evaluation	74
3.3	Results	79
3.3.1	PCG annotations and annotator agreement	79
3.3.2	Evaluation of SQIs	80
3.3.3	Classification of signal quality	81
3.4	Discussion	83
3.5	Conclusion	86
4	Probabilistic model-based phonocardiogram segmentation	89
4.1	Background	91
4.2	Methods	96
4.2.1	Database	96
4.2.2	Data exclusion based on ECG signal quality	97
4.2.3	Heart sound duration distributions	104
4.2.4	ECG-derived heart sound labelling	106
4.2.5	Hidden Markov models	106
4.2.6	Logistic regression HSMM	115
4.2.7	Conditional random fields	116
4.2.8	Feature extraction	123
4.2.9	Model training and evaluation	128
4.3	Results	130
4.4	Discussion	134
4.5	Conclusion	139
5	Systolic murmur classification in phonocardiograms	141
5.1	Background	143
5.1.1	Early classification methods	143
5.1.2	Artificial neural network-based classification	149
5.1.3	Support vector machine-based classification	153
5.1.4	Classification using hidden Markov models	155
5.1.5	Clustering-based classification	156
5.1.6	Other methods for heart sound classification	158
5.2	Methods	159
5.2.1	Data	159
5.2.2	Exclusion of noisy data	162
5.2.3	Heart cycle segmentation	165
5.2.4	Classification procedures	167

5.2.5	Feature extraction	167
5.2.6	Cycle averaging	173
5.2.7	Noisy heart cycle rejection	175
5.2.8	Selection of cycles based on k -means clustering	175
5.2.9	Feature downsampling	177
5.2.10	Data labelling	178
5.2.11	Classification methods and evaluation	179
5.2.12	Feature normalisation and class imbalance	183
5.2.13	Baseline methods for comparison	184
5.3	Results	185
5.3.1	Performance of cardiologist and Cardioscan software	185
5.3.2	Results of baseline methods	185
5.3.3	Results of proposed automatic classification algorithms	186
5.4	Discussion	189
5.5	Importance of signal quality assessment	194
5.6	Impact of external noise	195
5.7	Conclusion	196
6	Discussion	199
6.1	Summary of work	199
6.2	Limitations and future work	202
6.2.1	Signal quality classification	203
6.2.2	Heart sound segmentation	204
6.2.3	Pathology classification	206
6.2.4	Implementation considerations	207
A	Cohen's kappa	209
A.1	Extending Cohen's κ to more than two annotators	211
B	Logistic regression	213
C	Support vector machines	217
D	Wavelet decomposition	223

List of Figures

1.1	The number of physicians around the world, in density per thousand.	3
1.2	The projected number of mobile phone connections in Africa	4
1.3	Example of a mitral heart valve damaged from rheumatic heart disease	7
1.4	Prevalence of rheumatic heart disease in children between 5 and 14 years of age	8
1.5	The structure of the heart	10
1.6	The four heart valves	11
1.7	The relationship between different physiological events in the cardiac cycle . . .	13
1.8	The four auscultation areas of the heart	15
1.9	The audible range of heart sounds and murmurs	15
1.10	Relationship between cardiac sounds	18
1.11	Ejection and regurgitant systolic murmurs	20
1.12	Diastolic and continuous murmurs	22
1.13	Image of innocent murmurs and where they are best heard	23
1.14	Flow diagram illustrating the major design steps of an automated mobile phone-based auscultation device.	26
2.1	iPhone mobile phone-based stethoscope	31
2.2	Frequency response plot for the mobile phone sound cards in an iPhone 3G and a Nokia 3500 mobile phone	33
2.3	A mobile phone-based heart sound recording taking place at the Cardiac Clinic at Groote Schuur Hospital in Cape Town	33
2.4	Frequency response plot for the Samsung Galaxy SII I9100 mobile phone	34
2.5	The mobile clinic used for rheumatic heart disease screening by the University of Cape Town	35
2.6	The numbers of patients in each echocardiography-diagnosed RHD sub-category.	39
2.7	The number of RHD patients with cardiologist-diagnosed murmurs	40
3.1	Distributions for the annotations for the iPhone and Littmann recordings. . . .	53
3.2	Examples of good- and poor-quality PCG recordings.	56
3.3	An example of the envelope of a PCG recording	57
3.4	Examples of the autocorrelation of the envelope of good- and poor-quality PCG recordings	58

3.5	PCG and autocorrelation signal of an arrhythmic patient	59
3.6	Truncated autocorrelations of good- and bad-quality PCGs	60
3.7	Example of the calculation of <i>SampEn</i>	62
3.8	Example of a fitted cosine, used to calculate the <i>ccSQI</i> feature.	66
3.9	A scatterplot matrix of the iPhone SQIs	72
3.10	A scatterplot matrix of the Littmann SQIs	73
3.11	The effect of the λ regularisation parameter on a LASSO fit for the iPhone recordings	77
3.12	Single SQI test set classification accuracy	81
3.13	Average values of β coefficients for LASSO LR models	82
3.14	The proportional number of times a feature was selected for the SVM-based classification model	82
3.15	Example of decision boundaries for SVM-based signal quality classification of iPhone recordings	84
4.1	Example of an ECG-labelled PCG	90
4.2	Block diagram of the steps employed in the evaluation of the segmentation algorithms	97
4.3	Example of the accurate R-peak and end-T-wave positions derived in a noisy ECG signal	102
4.4	Heart sound duration distributions	105
4.5	Graphical representations of a hidden Markov model	108
4.6	The relationship between naïve Bayes, logistic regression, HMMs and linear-chain CRFs.	118
4.7	Determination of wavelet family and decomposition level.	127
4.8	Comparison of the shapes of typical S_1 and S_2 sounds with the selected wavelets. 131	
4.9	Distributions of F_1 scores for the 100 iterations of algorithms 1a and 3a from Table 4.4.	132
4.10	Example of a correctly segmented, noisy, PCG using the LR-HSMM method . .	136
4.11	Example of the weakness of the standard Viterbi algorithm	138
4.12	Example of an incorrectly segmented arrhythmic PCG using the LR-HSMM method.	139
5.1	Example of a pronounced systolic murmur due to mitral valve regurgitation. . .	142
5.2	Example of holosystolic murmur indicative of RHD recorded using the mobile phone-based stethoscope.	161
5.3	Example of an innocent systolic murmur recorded using mobile phone-based stethoscope.	162
5.4	Example of a heart cycle in the presence of the electricity generator.	163
5.5	The analysis procedures evaluated in the detection of pathology in PCG recordings 167	
5.6	The shape and frequency response of the Daubechies 10-tap wavelet (D10 or db5).	168

5.7	Wavelet features derived from a single cycle of a patient with severe systolic murmur.	169
5.8	MFCC filterbank with 13 linear filters followed by 27 logarithmically-spaced filters.	171
5.9	Illustration of the steps for MFCC computation.	172
5.10	MFCC computation in the absence of a murmur.	173
5.11	MFCC features derived from a single cycle of a patient with severe systolic murmur.	173
5.12	Cycle exclusion based on RMS error from the median systole	176
5.13	Example of k -means clustering-based selection of heart cycles	177
5.14	Feature downsampling to three systolic windows	178
5.15	Evaluation procedure for the SVM classification algorithms	180
5.16	Evaluation procedure for the LASSO LR classification algorithms	182
5.17	Example of an SVM decision boundary with features derived from the MAAS dataset.	188
5.18	Example of RHD_ES SVM decision boundary for MFCC features	190
5.19	Example of SVM decision boundary for MFCC features derived from RHD_MP dataset	190
6.1	PCG and autocorrelation of a good-quality PCG signal of an arrhythmic patient showing a lack of peaks	203
B.1	The response of the logistic function	214
C.1	An example of a large margin separating hyperplane between two sets of linearly separable data in 2D	218
C.2	Comparison of linear hyperplanes with and without slack variables	220
C.3	Non-linear decision boundaries generated using an SVM with a Gaussian kernel.	222
D.1	A two-level wavelet filter bank	225

List of Tables

2.1	Recording devices used for University of Cape Town mobile phone dataset . . .	34
2.2	Breakdown of numbers of patients with RHD, those with pathological blood flow, and those with murmurs.	38
3.1	PCG labelling scheme	53
3.2	The number of recordings in the smaller class after separating the good- and poor-quality recordings based on thresholding.	55
3.3	Summary of SQIs investigated for automatic PCG signal quality estimation . .	74
3.4	Weighted Cohen’s kappa for iPhone recordings	80
3.5	Weighted Cohen’s kappa for Littmann recordings	80
3.6	Conger’s Kappa, using quadratic weights for iPhone and Littmann devices . . .	81
3.7	The results of the PCG signal quality classification	83
4.1	Summary of previous segmentation work	91
4.2	The selection and exclusion rates of the different R-peak and end-T-wave detectors	103
4.3	Example of a transition matrix for a four-state heart sound HMM.	108
4.4	Gross results of segmentation algorithms on train and test sets using various input features	133
4.5	Average results of algorithms <i>1a</i> , <i>4b</i> and <i>6b</i> from Table 4.4 when averaged across all recordings rather than per patient.	135
5.1	Summary of previous classification work	144
5.2	Segmentation performance of the CRF-based algorithm on the mobile phone-based RHD recordings	166
5.3	Performance of the cardiologist (when using manual auscultation) and the CardioScan software when detecting RHD patients with functional valve damage .	185
5.4	Classification results from baseline comparison feature extraction methods . . .	186
5.5	MAAS dataset average results for proposed automatic classification algorithms	187
5.6	RHD_ES dataset average results for proposed automatic classification algorithms	189
5.7	RHD_MP dataset average results for proposed automatic classification algorithms	191
5.8	LASSO LR classification results of best-performing algorithms on each dataset	192
5.9	Classification results when including poor-quality recordings in the RHD_ES and RHD_MP datasets	194

5.10	Classification results when including generator-contaminated recordings in the RHD_ES and RHD_MP datasets	195
A.1	Benchmarks for the levels of agreement when using Cohen's Kappa	210

List of common abbreviations

CRF - Conditional random field

ECG - Electrocardiogram

Db5 - Daubechies 10-tap wavelet function

FHS - Fundamental heart sounds

HMM - Hidden Markov model

HSMM - Hidden semi-Markov model

LR - Logistic regression

MFCC - Mel-frequency cepstral coefficient

MR - Mitral regurgitation

MVP - Mitral valve prolapse

PCG - Phonocardiogram

RHD - Rheumatic heart disease

SVM - Support vector machine

S_1 - The first heart sound

S_2 - The second heart sound

S_3 - The third heart sound

S_4 - The fourth heart sound

Chapter 1

Introduction

“...with each movement of the heart, when there is the delivery of a quantity of blood from the veins to the arteries, a pulse takes place, and can be heard within the chest.” - William Harvey, 1628

*“Nor we, poor deafs, nor any other doctor in Venice can hear (heart sounds).
But happy may be he who hears them in London” - Æmylius Parisanus, 1632*

It is not commonly known that “strep throat” can lead to heart failure. Yet, this bacterial infection, if left untreated, can cause severe heart damage, known as rheumatic heart disease (RHD). RHD often leads to heart failure later in life. If children with RHD are diagnosed at an early stage, they can be prescribed simple penicillin which halts the progression of the disease. However, such a diagnosis requires, at a minimum, a trained clinician. Trained medical personnel are scarce in the developing areas of the world where RHD is most prevalent. Hence, RHD remains a major burden of disease in developing countries.

This thesis investigates the design of an automatic mobile phone-based diagnostic tool to screen for RHD, with the aim of identifying those with a high probability of having RHD, while not requiring a trained clinician. Such a device falls under the umbrella of “mobile health”, which is outlined in Section 1.1. A further explanation of RHD and the impact it has on the developing world is given in Section 1.2. The diagnostic tool designed in this thesis aims to identify individuals with RHD by analysing their recorded heart sounds. Therefore, an outline of the cardiac cycle, the heart sounds, and the pathological sounds produced by a diseased heart is given in Section 1.3. The motivation for the remainder of the thesis is given in Section 1.4.

1.1 Mobile health

Mobile health can be broadly defined as the support of medical and public health practice by mobile phones, patient monitoring devices, personal digital assistants and other wireless devices [1, 2]. Interest in mobile health has grown rapidly in recent years, with the World Health Organization (WHO) stating that mobile health has the “potential to transform the face of health service delivery across the globe” [3]. Mobile technology is capable of improving access to healthcare, enhance diagnostics and lower healthcare costs [1, 4, 5].

Of particular interest is the application of mobile health in the developing world, where it is thought to have the potential for the greatest impact [6]. This stems from three main phenomena: the burden of disease in the developing world; the lack of healthcare workers in the developing world; and the prevalence of mobile phones in these regions [2, 7].

1.1.1 The potential for mobile health in developing countries

The majority of the world’s population receive healthcare in low-resource areas [7], as people in developing countries have been found to be more unhealthy than those in developed countries [8]. This can be attributed to a number of factors, including: inadequate electricity and clean water; a shortage of healthcare workers; poor infrastructure; lack of funding; and poor healthcare training [7, 9, 10]. Smith [11], and Pritchett and Summers [8] have established the link between poor health and low household income and wealth. As a result, developing countries have a much higher prevalence of disease.

The majority of deaths in low- and middle-income countries (LMICs) are a result of infectious and parasitic diseases [12]. However, in addition to communicable diseases such as HIV, malaria and tuberculosis (TB), developing countries also have to deal with non-communicable, chronic conditions such as diabetes, cancer and cardiovascular disease which are traditionally associated with the First World [13]. This has been termed a “double burden” of disease in developing countries [12]. As Banatvala and Donaldson [13] comment, this combination would challenge even the most advanced and well-resourced healthcare systems. As a result of this high burden

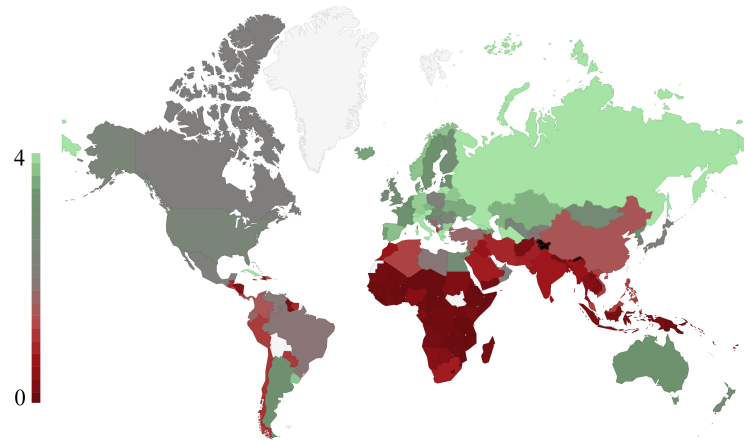


Figure 1.1 – The number of physicians around the world, in density per thousand. Africa and South-East Asia can be seen to have the largest shortfall of physicians [20].

of disease, the introduction of mobile health has the potential to make a significant impact in developing countries [6, 14].

The burden of disease faced by developing nations is compounded by a massive shortage of healthcare workers [15, 16, 17]. This is coupled with inadequate training, lack of technical staff and poor infrastructure that is unable to provide sufficient consumables and spare parts for medical equipment [6, 9, 10]. For example, sub-Saharan Africa carries over 24% of the world's burden of disease, yet is home to only 3% of the global health workforce [15]. It has been established that 57 countries, predominantly in Africa, have a critical shortage of healthcare workers, with a global deficit of 2.4 million doctors, nurses and midwives [18]. Some African countries would need to scale their workforce up by 140% in order to meet needs [16].

South-East Asia and Africa have the fewest healthcare workers and the highest burdens of disease in the world [17], while the 15 countries with the highest population per doctor are all from Africa, all with fewer than one doctor per 10,000 people. Liberia has one doctor to 70,000 people [19]. The most recent audit of the number of physicians around the world echoed these findings, as shown in Figure 1.1.

In areas with a severe lack of healthcare workers, the use of mobile health has been identified as a possible way to provide access to healthcare [5] while decreasing costs and maintaining or improving quality of care [21]. Alongside the lack of healthcare personnel, people living in developing countries are now more likely to have access to a mobile phone than to clean drinking

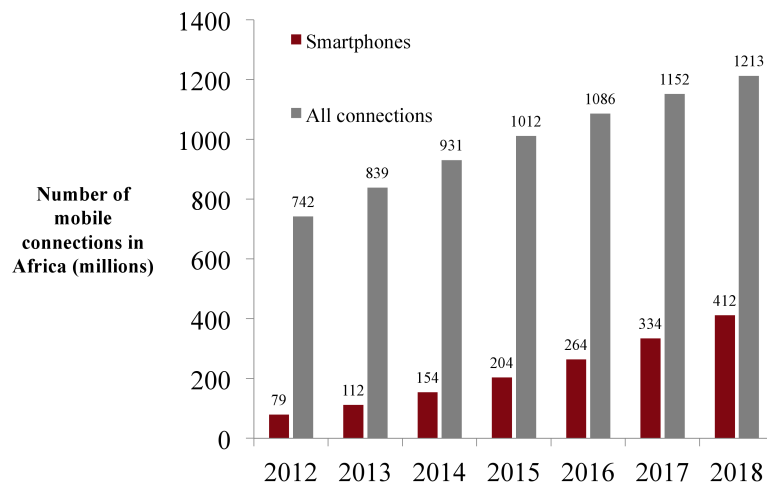


Figure 1.2 – The projected number of mobile phone connections in Africa, including smartphones. The number of mobile connections is steadily increasing, while the number of smartphones can be seen to be increasing exponentially. It is projected that one third of all phones in Africa will be smartphones by 2018 [27].

water and electricity [22]. Mobile phone usage in developing countries has risen sharply; it is estimated that there are now over 75 mobile subscriptions per 100 people in the developing world [23].

The growth of mobile phone connections in Africa is shown in Figure 1.2. While there has been, and is projected to be, a steady increase in numbers of all mobile phone connections, of particular interest is the projected growth of smartphones. These devices, with enhanced processing capabilities and internet access, have greater potential for enabling health communication with under-resourced areas, making data collection and real-time diagnostics possible on a large scale [24]. The use of mobile health could reduce isolation and provide healthcare services and support to otherwise sequestered areas [25, 26].

1.1.2 The evidence for mobile health

The field of mobile health has generated much hype over recent years. However, to date, there is limited evidence underpinning the hype and confirming the potential [28]. This is thought to be because of this field developing as a patchwork of incompatible applications and pilot studies [1, 29]. Tomlinson et al. [5] have been particularly scathing of mobile health, stating that it currently lacks a foundation of even basic evidence. They highlight, as Estrin and Sim [29] also do, the need for collaboration between mobile health application developers, and

as a result, the development of robust, interoperable platforms for large-scale mobile health evaluation. A large-scale review by Chib et al. [30] did find a growing body of evidence for mobile health in developing countries, but concluded that more robust studies were still needed.

One area where evidence for mobile health does exist is in the field of text message reminders. There is evidence that such reminders improve attendance at clinics in the developed world [31, 32]. Similarly, two recent large systematic reviews by Free et al. [26, 33] found evidence that mobile health technology does improve healthcare delivery and brings about positive behaviour change. However, these researchers also emphasise that high-quality, adequately powered clinical trials, particularly in the developing world, are necessary. It is only now that such studies are beginning to emerge. For instance, text message reminders have been shown to improve medication adherence in antiretroviral treatment in Kenya [34, 35], while the largest trial-to-date of text message-based support for the self-management of hypertension was recently completed in South Africa [36]. Promisingly, Labrique et al. [37] argue that mobile health is moving into a new era of substantive, rigorous studies and evidence, while Nilsen et al. [4] attribute the lack of large-scale evaluation of mobile health technologies to the high pace of their development.

Within the field of mobile health applications, there is debate as to whether applications should take advantage of the most basic functionality of a mobile phone, such as text messages and phone calls, in order to provide the best coverage, or whether they should take advantage of the superior processing power and sensing capabilities of the less widely available smartphone [1, 4]. This thesis argues for the latter, by exploiting a smartphone’s intrinsic ability to record sound, and the possibility to develop applications to analyse these sounds. Of particular focus in this thesis is the analysis of heart sounds for the detection of sounds indicative of pathology.

1.1.3 Mobile health for heart sound analysis

The recording of heart sounds and their subsequent analysis, both using a mobile phone, have been explored before [38, 39]. These researchers explored the design of a low-cost stethoscope and worked on heart sound detection in the signals recorded by a mobile phone.

There are a number of heart sound recordings “apps” and accessories commercially available. The earliest of these was the iStethoscope Pro and associated iAuscultate applications [40, 41],

which allow the visualisation of recorded heart sounds (without a stethoscope attachment) and of their frequency content, but provides no further analysis. Similarly, the Thinklabs stethoscope app [42] allows the user to visualise and listen to heart sounds recorded directly on the phone using a custom-built stethoscope, but the app does not perform any analysis of the recorded heart sounds.

Others have developed hardware to record heart sounds on a mobile phone. StratoScientific have advertised a 3D printed stethoscope attachment for an iPhone device, named StethIO [43]. In their brief online summary, they advertise the ability to record heart sounds, visualise these sounds, and provide remote diagnostic capabilities. However, this device has been available for “pre-order” for over a year, while no information about the “diagnostic procedure” is given. More recently, CliniCloud [44] and Eko Devices [45] have developed bespoke devices to record heart sounds on a mobile phone. Eko’s device, the Eko Core, is an attachment for a regular stethoscope that allows the transfer of recorded heart sounds to a mobile phone via BluetoothTM, while CliniCloud’s device is an attachment designed specifically for connection to a mobile phone. While both of these devices allow the sharing of the recorded sounds (with the possibility of remote diagnosis by a clinician), neither of them offer any diagnostic capabilities.

1.2 Rheumatic heart disease

Rheumatic heart disease (RHD) is the result of valvular heart damage caused by the immune system’s response to untreated *Group A Streptococcal* (GAS) infections that often manifest as pharyngitis or “strep throat” [46]. GAS infections lead to acute rheumatic fever, most commonly in children and adolescents [47]. During acute rheumatic fever, the body releases antibodies to fight the infection. However, due to the similarity in structure between the bacterium and cells in the heart valves, the antibodies cause damage to the valves within the heart, leading to inflammation and fibrosis which can narrow the heart valve (stenosis) [48, 49]. It is this permanent damage to the heart valves caused by rheumatic fever that is termed RHD. Up to 60% of individuals develop RHD after a single acute rheumatic fever attack [50].

While an initial acute rheumatic fever attack can cause valve damage [51], it is also the case that RHD most often results from recurrent acute rheumatic fever infections [46]. Indeed,

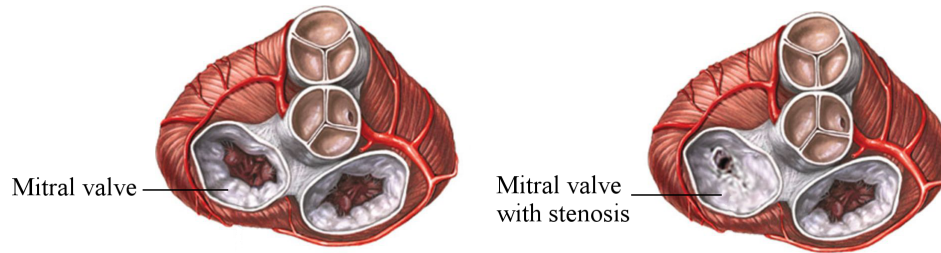


Figure 1.3 – Example of a damaged mitral heart valve as a result of RHD. The cross-section of the top of the heart is shown, displaying the four heart valves. The example on the left shows the undamaged mitral valve; the example on the right shows the thickening of the valve leaflets caused by RHD, and narrowing of the valve area (stenosis). Adapted from [48].

recent evidence suggests that acute rheumatic fever is not a prerequisite for RHD [49, 52]. An example of the damage caused by RHD to the mitral valve, most commonly affected by RHD [53, 54], can be seen in Figure 1.3.

RHD is most common in children and young adults from families with low socioeconomic status [52, 55, 56, 57, 58]. Risk factors for rheumatic fever are poverty, overcrowding and malnutrition [46]. For this reason, this disease remains a major burden in developing countries (particularly sub-Saharan Africa [50, 59, 60]) and is the leading cause of cardiovascular mortality and morbidity in children and young adults worldwide [46, 61]. The disease causes the highest number of disability-adjusted life-years of all listed cardiovascular diseases among 10 to 14-year-olds and the second highest number among children aged between the ages of five and nine years [62].

The prevalence of RHD in South African children is estimated to be 20.2 cases per 1000 [60], while it is estimated that between 15.6 and 19.6 million people are living with RHD worldwide, leading to an estimated 1.4 million deaths annually [63]. Between 44 and 80 per cent of children with untreated RHD will progress to severe cardiac failure. Without medication, most affected individuals die before the age of 30 [49]. The prevalence of RHD in children between five and fourteen years of age is shown in Figure 1.4. It can be seen that those areas most affected by RHD are also those with the fewest doctors, as shown in Figure 1.1.

Clinical diagnosis of RHD is based on auscultation or echocardiography [46, 64]. During auscultation, RHD diagnosis is based upon the detection of an audible pathological valvular heart murmur by a trained clinician using a stethoscope [46, 54]. Mitral valve damage is the most common valvular involvement in early stages of RHD, while mixed aortic and mitral

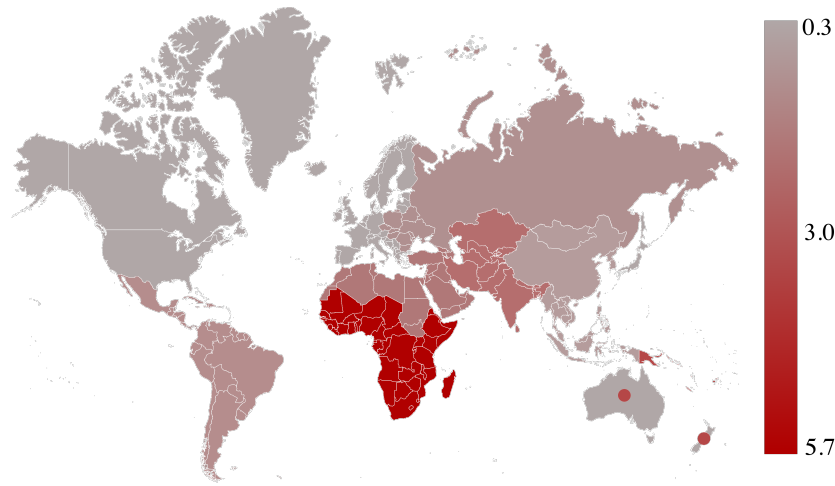


Figure 1.4 – Prevalence of RHD in children between five and fourteen years of age, in cases per thousand. The bar on the right-hand-side gives an indication of the prevalence, with Africa having the largest burden (5.7 cases/1000), while indigenous populations of Australia and New Zealand (indicated with circles) and the Pacific islands have the second largest (3.5). Adapted from [50].

valve damage is common in adolescents [63]. Damage to these valves causes systolic murmurs in most cases (see Section 1.3.4.5).

Echocardiography^a has been shown to be far more sensitive than auscultation when diagnosing RHD, and it has been documented that human auscultation has low sensitivity and may underestimate the prevalence of RHD [46, 49, 58, 59, 65]. However, the use of echocardiography is not feasible in low-resource areas [66] as it is expensive (an estimated \$15,000 for a single portable machine [67]) and requires a highly skilled operator [49, 57]. In addition, the use of echocardiography alone may be over sensitive, resulting in a high number of false positive diagnoses, as the progression of *latent* or *subclinical* RHD, or RHD without clinical symptoms, has not yet been adequately documented [49, 58, 68]. There have been recent efforts to develop lower-cost, hand-held echocardiography devices [69, 64]. However, extensive training is still required on these devices. In addition, while hand-held devices may prove to be more sensitive than auscultation, they are not able to meet the gold-standard diagnostic criteria for RHD due to their technical limitations [64]. In addition, even the so-called “low-cost” devices remain prohibitively expensive, with a cost of \$9000 for a single device [70].

In under-developed areas, diagnosis of RHD is usually made at a late stage, when valve lesions have become severe and symptomatic and significant intervention (such as surgery) is required (which is the least cost-effective treatment and consumes the majority of the funds devoted to

^aEchocardiography, or simply “echo”, is the imaging of the heart using 2D or 3D Doppler ultrasound.

RHD treatment) [59, 71]. However, for patients at an advanced stage of the disease, access to cardiac surgery is limited: even in developed countries only 11% of patients have surgery [72]. These patients are normally below 20 years of age [72].

Since severe RHD most often results from recurrent acute rheumatic fever infections [46, 73], primary treatment of rheumatic fever is a beneficial intervention but rare in developing countries, due to the lack of facilities and healthcare workers. Secondary prophylaxis using penicillin after initial diagnosis of RHD can prevent recurrent acute rheumatic fever infections and has been shown to be cost effective [57, 73]; penicillin can reduce the severity of RHD, thus decreasing the risk of valvular malfunction [46, 49, 74]. The WHO recommends school-based screening for RHD in high-prevalence areas to identify previously undiagnosed patients [49]. Those with a mild murmur are meant to be given penicillin for ten years, while those with more severe murmurs should receive lifelong prophylaxis [59, 75].

In summary, RHD is a major health burden, most prevalent in developing areas of the world. In these areas, the diagnosis of RHD, by either auscultation or echocardiography, is difficult because of the need for trained individuals and/or expensive equipment. However, once individuals have been diagnosed with early-stage RHD, they can receive low-cost penicillin treatment which would significantly reduce their risk of developing severe valvular damage. Therefore, a low-cost screening device, requiring little to no training, would be ideal to identify young adults and children with a high probability of having RHD and refer them on for treatment. A mobile phone-based automatic auscultation device has the potential to identify those individuals with a high probability of having RHD without requiring either expert training or expensive equipment, provided that the apparatus is designed appropriately.

1.3 The cardiac cycle, heart sounds and murmurs

During the cardiac cycle, the heart emits sounds that are audible using a stethoscope. These sounds are due to the vibrations of structures or the turbulent flow of blood within the heart. This section outlines the relationship between the mechanical activity of the heart and the cardiac sounds. The structure of the heart is first outlined. The cardiac cycle, prompted by electrical activity that subsequently gives rise to the heart's mechanical activity, is then described. The sounds produced by the vibration of the heart as a result of this mechanical

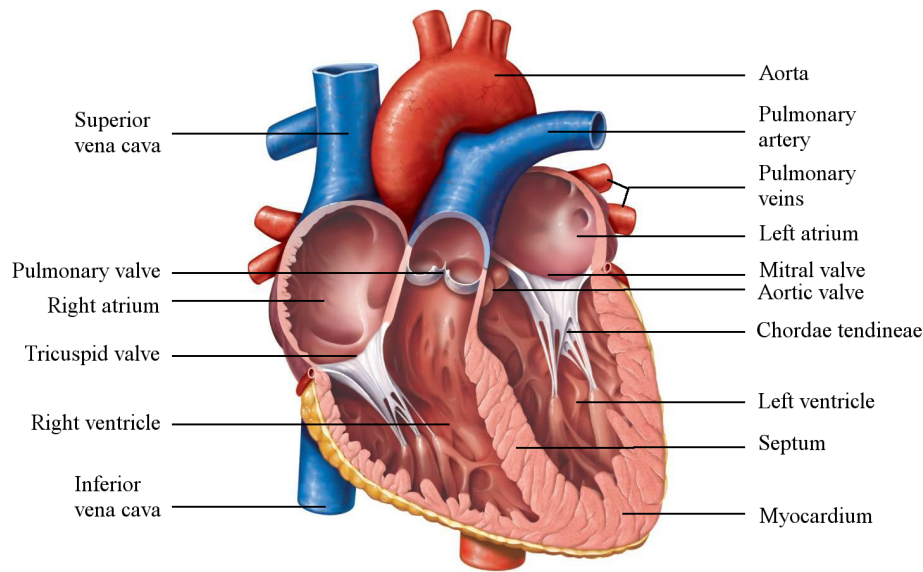


Figure 1.5 – The structure of the heart, showing the major veins and arteries, the valves, atria and ventricles [77].

activity are described, followed by an introduction to murmurs – the sounds produced by the turbulent flow of blood within the heart.

1.3.1 The heart structure

The heart consists of two *mirrored* sides, which support different circulatory systems but which beat in a synchronised manner. Each side of the heart, left and right, has two chambers: one *atrium*, where blood is collected when it enters the heart; and the *ventricle*, where blood is ejected from the heart into circulation around the body [76]. The direction of blood flow is controlled by four heart valves. The flaps that make up the valves are called leaflets or cusps. These cusps open to allow the flow of blood and close to prevent the flow of blood in the incorrect direction. The atria, ventricles and four heart valves are shown in Figure 1.5.

The *mitral valve*^b halts the flow of blood from the left ventricle to the left atrium during ventricular contraction. In order to stop this valve everting into the atrium during periods of high pressure in the ventricle, *chordae tendineae* connect the valve leaflets to the ventricle wall. The muscles to which these tendinous chords connect contract at the same time as the ventricle, holding the valve leaflets in place. The *tricuspid valve* serves the same purpose between the

^bThe mitral valve is so named due to its resemblance to a bishop's mitre

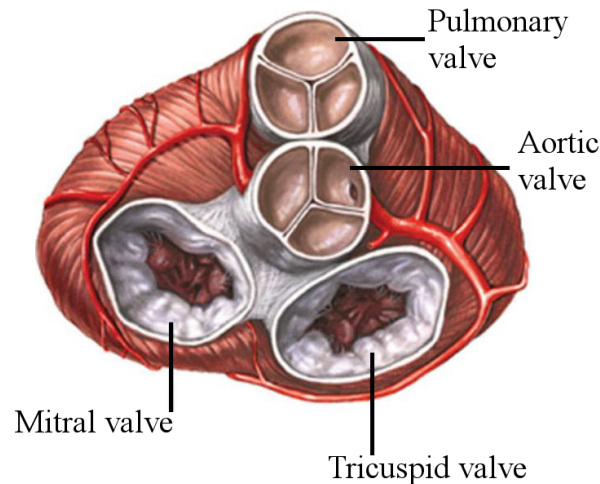


Figure 1.6 – The four heart valves, viewed from the top of the heart. Adapted from [48].

right atrium and right ventricle, and also has chordae tendineae. Together, the mitral and tricuspid valves are called the *atrioventricular* valves. These valves are shown in Figure 1.6.

Between the ventricles and the major arteries are the *aortic* and *pulmonary* valves, on the left and right side of the heart respectively. These valves both consist of three cusps but, due to their shape, do not require chordae tendineae to prevent prolapse. Together, these are called the *semilunar* valves.

The *myocardium*, or the wall of the heart, is made up of muscle which produces the force responsible for moving blood around the body during contraction. The myocardium also contains the conduction system that allows electrical impulses to travel rapidly through the heart. These electrical impulses are responsible for initiating the cardiac cycle, and cause the contraction of the heart muscle [76].

During one heart beat, as is discussed in Section 1.3.2 below, de-oxygenated blood flows from the body into the right atrium via the vena cava, where it is collected before being pushed into the right ventricle. When the right ventricle contracts, this de-oxygenated blood is forced into the lungs via the pulmonary artery. Oxygenated blood from the lungs returns to the heart via the pulmonary veins. It is collected in the left atrium, and forced into the left ventricle. Upon the contraction of the left ventricle, this oxygenated blood is pumped around the body via the aorta. These major vessels are shown in Figure 1.5.

1.3.2 The cardiac cycle

The cardiac cycle refers to a single complete beat of the heart. It is initiated by the *depolarisation* of the pacemaker cells which make up the *sinoatrial* (SA) *node*, located in the upper part of the right atrium. In the absence of pathology, this mass of cells depolarises and repolarises at regular intervals, controlled by the autonomic nervous system. The resulting electrical impulse propagates through the heart's conduction system in order to produce an ordered sequence of events.

The heart's electrical activity can be measured on the surface of the body, with surface electrodes used to record the *electrocardiogram* (ECG). A typical ECG trace is shown in Figure 1.7 alongside the pressure waveforms and the heart sounds, which are discussed in Section 1.3.3 below.

The cardiac cycle commences with *atrial systole*, or the contraction of the two atria, initiated by the depolarisation of the SA node. This causes an electrical impulse to propagate through the heart muscle, pushing blood into the ventricles, ready for ejection into the pulmonary and aortic arteries. Atrial depolarisation gives rise to the P-wave in the ECG.

The electrical impulse is then collected at the *atrioventricular* (AV) *node*, which delays the further propagation of the electrical impulse. This delay allows further filling of the ventricles with blood before the contraction of the ventricles, named *ventricular systole*. The electrical impulse reaches the ventricles via the *bundle of His*, which spreads into rapidly conducting fibres, called *Purkinje fibres*. This rapid conduction, corresponding to the QRS peak in the ECG, allows a unified contraction of the ventricles. Ventricular systole has three phases:

- Isovolumetric contraction, during which pressure rises within the ventricles, but before the pressure has risen enough to overcome the pressure threshold to open the pulmonary and aortic valves.
- Ventricular ejection, during which the pulmonary and aortic valves open and blood is ejected to the lungs (via the pulmonary artery) and the body (via the aorta).
- Isovolumetric relaxation, during which the ejection of blood ceases, the aortic and pulmonary valves close, and pressure within the ventricles falls.

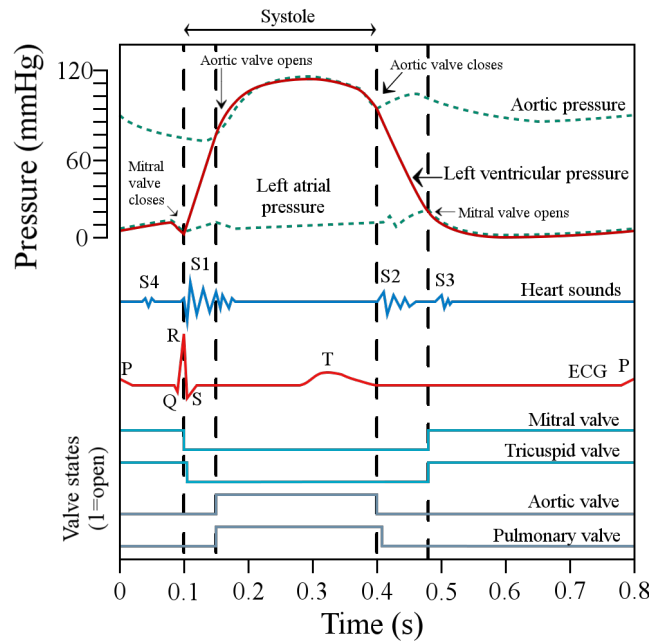


Figure 1.7 – The relationship between different physiological events in the cardiac cycle. The rapid increase in left ventricular pressure at the start of systole, coinciding with the R-peak of the ECG, causes the left ventricular pressure to exceed the left atrial pressure. This causes the mitral valve (and tricuspid valve on the right side of the heart) to close, giving rise to the first heart sound, S_1 . Ejection of blood into the aorta and pulmonary veins occurs when left ventricular pressure exceeds aortic pressure. The closure of the aortic and pulmonary valves gives rise to S_2 , which occurs when the left ventricular pressure drops below the aortic pressure, coinciding approximately with the end of the T-wave in the ECG. The changes of the four valve states, demarcated with dotted lines, are shown at the bottom of the figure. Adapted from [78].

Ventricular systole, which is often abbreviated to only *systole* (not to be confused with atrial systole), is shown at the top of Figure 1.7. The remainder of the cardiac cycle is termed *diastole*. During the second phase of ventricular systole (ventricular ejection), the atria relax and are filled with venous blood from the body and lungs. This, then, allows passive filling of the ventricles during isovolumetric relaxation. The T-wave in the ECG represents ventricular repolarisation.

1.3.3 Heart sounds and murmurs

As described above, the electrical activity of the heart during the cardiac cycle causes mechanical activity in the form of atrial and ventricular contractions. This in turn forces blood between the chambers of the heart and around the body, as a result of the opening and closure of the heart valves. This mechanical activity, and the sudden start or stop of the flow of blood within the heart, gives rise to vibrations of the entire cardiac structure [79]. These vibrations

are audible on the chest wall, and listening for specific sounds can give an indication of the health of the heart.

The audible vibrations of the heart can be broadly split into two categories: heart sounds and heart murmurs. Heart sounds are generated by either the commencement or the halting of the flow of blood within the heart. Murmurs are caused by the vibration of the valves or the turbulent flow of blood within the heart. A brief introduction to auscultation is given first, followed by a discussion of the various heart sounds and murmurs.

1.3.3.1 Auscultation and phonocardiography

Auscultation, derived from the Latin verb *auscultare* meaning “to listen” is the act of listening to the internal sounds of the body, most often with a stethoscope, in order to detect pathology. A phonocardiogram (PCG) is the graphical representation of a heart sound recording.

When performing auscultation of the heart the four locations most often used to listen to the heart are:

- Aortic area - centred at the second right intercostal space
- Pulmonic area - in the second intercostal space along the left sternal border
- Tricuspid area - in the fourth intercostal space along the left sternal edge
- Mitral area - at the cardiac apex, in the fifth intercostal space on the midclavicular line.

These auscultation areas are shown in Figure 1.8. They are named according to the valve which can be best heard at these positions.

The detection of pathological sounds, such as murmurs, through auscultation is difficult. This is because the human auditory system is not well suited to the detection of the low-amplitude and low-frequency oscillations of murmurs [80]. The sounds that the heart makes are near the threshold of human hearing [79], as shown in Figure 1.9. Clinical auscultation skills have been found to be poor, even with high levels of training, with medical students and physicians in practice achieving only between 20 and 30% accuracy when diagnosing pathological heart sounds [81, 82, 83]. Furthermore, only 20-30% of patients referred due to suspicion of heart murmurs actually have pathology [84]. This can be explained by the low

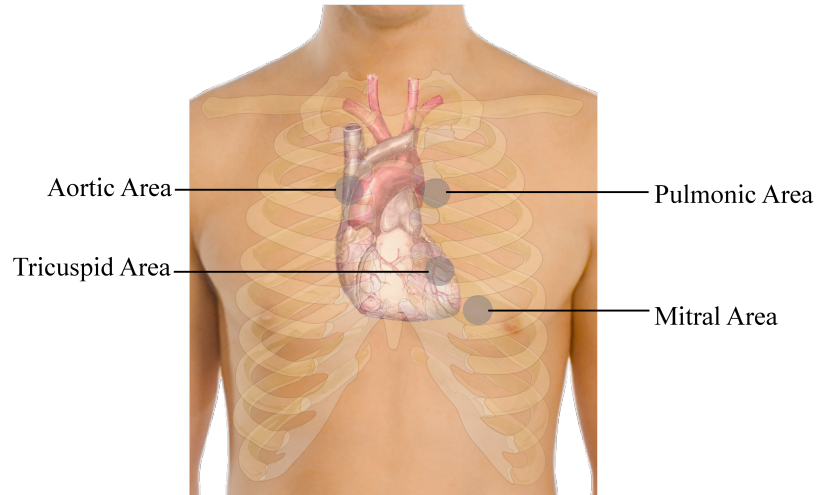


Figure 1.8 – The four auscultation sites of the heart, named according to the valve which can be best heart using a stethoscope at these positions. Adapted from [54].

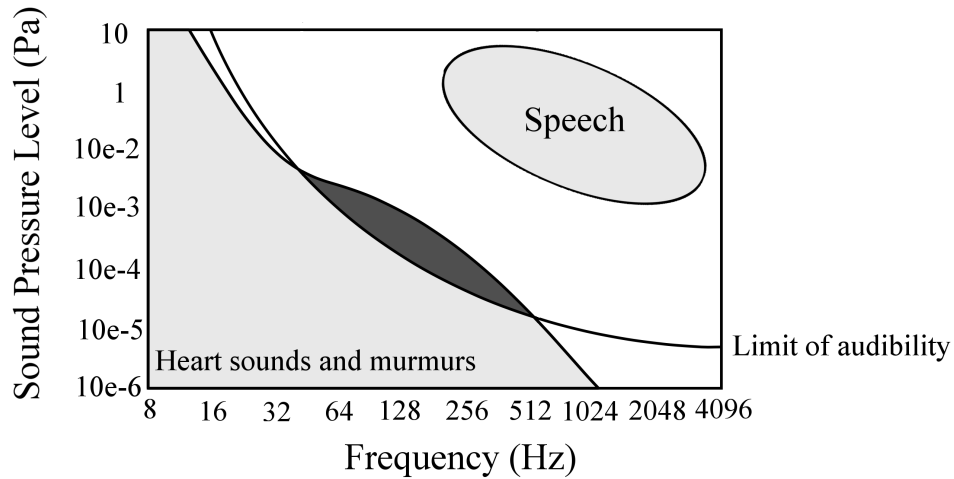


Figure 1.9 – Illustration of the audible range of heart sounds and murmurs. Adapted from [79].

amplitude and low frequency of pathological murmurs, which is close to the threshold of human hearing (see Figure 1.9). The advent of more advanced electronic stethoscopes and computer-aided auscultation (CAA) has the proven potential to improve auscultation accuracy [85, 86], especially in areas with major shortages of physicians, such as developing regions of Africa [87].

1.3.3.2 The fundamental heart sounds

The fundamental heart sounds (FHSs), or primary heart sounds, refer to the first and second heart sounds (S_1 and S_2). These are often the most pronounced and recognisable heart sounds, giving rise to the characteristic “*lub-dub...lub-dub...lub-dub*” sound cycle of the heart.

The first heart sound (S_1) occurs at the beginning of isovolumetric ventricular contraction, when the mitral and tricuspid (atrioventricular) valves close due to the rapid increase in pressure within the ventricles [88]. However, it is thought that the sound itself is not due to the collision of the valve cusps when the valve closes, but rather to the vibrations of cardiac structures and blood within the heart [79, 80]. The start of S_1 is synchronous with the QRS peak of the ECG [89], as shown in Figure 1.7.

Shortly after the start of isovolumetric contraction, when ventricular pressure rises above aortic and pulmonary artery pressure, the pulmonic and aortic valves open without any sound [79, 90]. There may be a slight splitting of mitral and tricuspid elements of S_1 , of up to 0.03 s [79, 80]. Splitting of a longer duration is almost always abnormal, and can be caused by electrical issues such as bundle branch block, ectopic beats and tachycardia, or mechanical delays from a damaged mitral or tricuspid valve [79, 80].

The second heart sound (S_2) occurs at the beginning of diastole with the closure of the aortic and pulmonic valves [90]. It is approximately synchronous with the end of the T-wave in the ECG [89], as shown in Figure 1.7. As with the first heart sound, it is not the actual closing of the valves that causes the sound. The valve closures occur milliseconds before the onset of the second heart sound, which demonstrates that the origin of the sounds, like S_1 , is from the vibrations of the entire cardiac structure and blood within the heart rather than motion of the valves [79, 91].

The S_2 sound normally comprises the louder aortic valve closure component, and the quieter pulmonary valve closure component, often manifesting as a “split” second heart sound [54]. A narrowly split S_2 (0.02-0.08 s during inspiration) is physiologically normal, and is most often found during or at the end of inspiration in children and young adults [79, 91]. This phenomenon is due to the increase in volume of the right atrium from increased blood flow from the vena cava and diminished flow from the lungs to the left atrium during inspiration [79, 90]. Persistent splitting of S_2 may be due to delayed electrical activation of the right ventricle and mitral regurgitation^c, while the absence of respiratory variations in the degree of S_2 splitting may, too, be pathological [54, 79].

^cRegurgitation, in a cardiac sense, refers to the backward flow of blood through a closed, leaking valve.

For normal subjects over the age of six months, the dominant frequency content in the FHSs is at 50 Hz approximately. S_1 has a slightly narrower range of frequencies (24-104 Hz) than S_2 sounds (24 - 144 Hz) [91, 92]. The main frequency content of heart sounds is below 150 Hz [92]. Estimates of the normal duration of the fundamental heart sounds vary [93, 94, 95, 96, 97, 98, 99], but more recent measurements have estimated the duration of S_1 to be approximately 0.12 s while the duration of S_2 is of the order of 0.08 s [98, 99].

1.3.3.3 Other heart sounds

While the FHSs are the most recognisable sounds of the heart cycle, the mechanical activity of the heart may cause other audible sounds.

The rapid filling of the ventricles with blood during diastole may cause vibrations loud enough to be audible. These vibrations are called the third heart sound (S_3), which is often heard in children and young adults [54, 79]. The third heart sound occurs approximately 0.15 s after S_2 , but may often be confused with S_2 , especially in cases of mitral regurgitation, where S_2 may be obscured by a murmur [79].

A fourth heart sound (S_4), heard slightly before S_2 , is a sign of forceful atrial contraction in an attempt to overcome ventricular abnormality, such as an enlarged size (hypertrophy) or a stiff ventricular wall [54, 79]. It is always pathological, yet not common in infants and children [54].

Early systolic ejection sounds, sometimes called systolic *ejection clicks* (EC), are caused by the halting of the upward movement of the aortic and pulmonary valves with the onset of ejection into the aorta and pulmonary artery. This is seen in cases of aortic and pulmonary valve disease and with the dilation of the aorta and pulmonary artery, as in the case of pulmonary hypertension [79]. These sounds can often be loud and at a similar frequency to the S_1 sound.

Mid-systolic clicks (MC) are rarely heard, but may occur with systolic murmurs caused by RHD or mitral valve prolapse. The exact timing of these clicks is dependent on the position of the patient during auscultation, and can shift into late systole [54, 79].

Diastolic sounds or *opening snaps* (OS) may also be heard due to the halting of the atrioventricular valves during high flow, because of stenosis or other abnormalities of the valves. These occur just after S_2 , but before, and at a higher frequency than, S_3 [79].

An illustration of the relationship between the various heart sounds is shown in Figure 1.10.

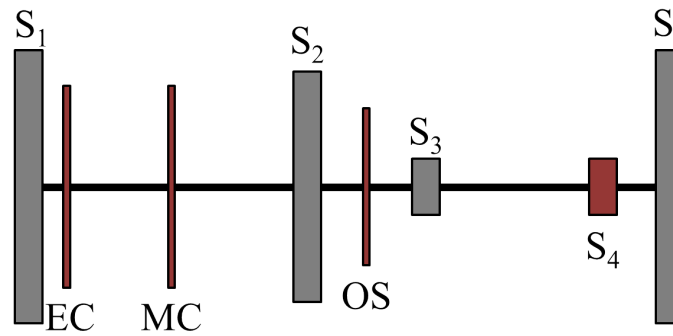


Figure 1.10 – Relationship between cardiac sounds, showing their timing, relative duration and relative amplitude. The first, second, third and fourth heart sounds (S_1 - S_4) are shown, as well as early and mid-systolic clicks (EC, MC) and the diastolic opening snap (OS). Pathological sounds are shown in red. Adapted from [54].

1.3.4 Murmurs

A heart murmur is caused by turbulent, high-velocity flow of blood, often across an abnormal heart valve. It is described as a “whooshing” sound that is audible using a stethoscope [79, 88]. Murmurs have a high frequency content (in comparison to the S_1 and S_2 sound) extending up to approximately 600 Hz [100]. The intensity of a cardiac murmur depends on the size of the orifice through which the blood flows, the pressure gradient across the orifice and the volume of blood flowing across the site [90].

Murmurs are often described as being caused by either *regurgitation* or *stenosis*. Stenosis, when referring to cardiac structures, refers to the narrowing of a heart valve due to the pathological fibrosis or scarring of the valve tissue. Regurgitation is also called valve insufficiency or incompetence.

Heart murmurs are defined by the region of the cardiac cycle in which they appear, being either systolic, diastolic or continuous (spanning both systole and diastole). Therefore, a murmur can only be classified once both the S_1 and S_2 sounds have been correctly identified [79]. While murmurs are often an indication of pathology, they can also be “innocent” and physiologically normal. A description of the various murmurs and their aetiology follows.

1.3.4.1 Systolic murmurs

Systolic murmurs can be classified as *ejection systolic murmurs*, due to the ejection of blood into the aorta and pulmonary artery, or *regurgitant murmurs*, caused by the flow of blood back

through closed valves [54].

Ejection systolic murmurs occur during the phase of maximum velocity of blood flow into the aorta and pulmonary artery during ventricular contraction. These murmurs are most often mid-systolic, as there is a delay in the time between S_1 and the opening of the aortic and pulmonary valves. Their intensity wanes as the pressure in the ventricles decreases towards the end of systole. The exact position of the peak intensity of the murmur varies with the pressure in the aorta and pulmonary artery (controlling when the aortic and pulmonary valves open), and depends on the presence of stenosis, in which case the peak intensity may be late-systolic. Ejection systolic murmurs may be caused by damage to the aortic or pulmonary valves (as in the case of stenosis), dilation of the aorta or pulmonary artery or are the most common form of innocent murmur (see Section 1.3.4.4) [54].

Systolic regurgitant murmurs are most often pan- or holosystolic, since they are caused by the flow of blood from the ventricle at higher pressure throughout systole, to the atria at lower pressure through a leaking valve. The pressure difference between the atria and ventricles rises rapidly after S_1 , causing these murmurs to start with, or very soon after, the first heart sound, and end at the start of S_2 , as the pressure in the ventricles is still high until the closing of the semilunar valves. Systolic regurgitant murmurs are caused by mitral or tricuspid valve regurgitation (MR and TR), mitral valve *prolapse* (MVP)^d, and *ventricular septal defect* (VSD)^e [54, 101].

More specifically, MVP, which is investigated more in Chapter 5, refers to the displacement of the leaflets of the mitral valve during systole. As outlined in Section 1.3, the mitral valve is responsible for halting the flow of blood from the left ventricle into the left atrium during the contraction of the heart. However, due to valve abnormalities and/or pathological leaflet thickening, these valves can be pushed back into the atrium during systole. This often leads to a high-pitched systolic regurgitant murmur, observed in late systole, with an occasional mid-systolic click [102].

Leatham [79] described a method to discriminate between ejection and regurgitant murmurs: if the murmur is of short duration, of crescendo shape, peaking in mid-systole and ending

^dProlapse, in the cardiac sense, refers to the pathological displacement of heart valves into the adjacent atrium.

^eA ventricular septal defect refers to a hole in the heart wall dividing the left and right ventricles.

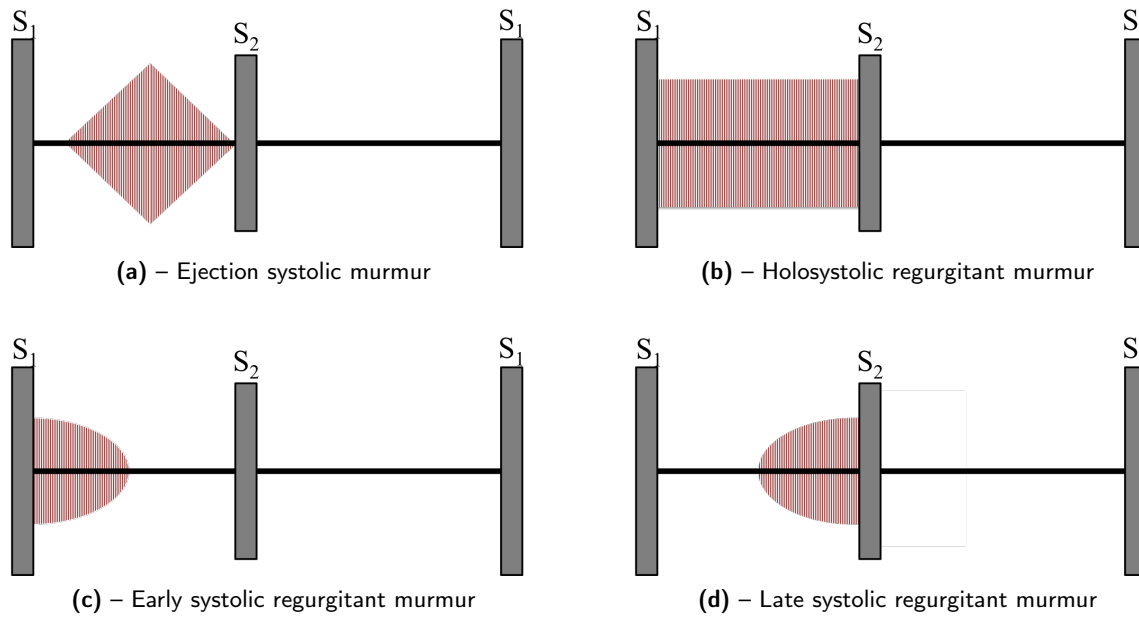


Figure 1.11 – Ejection and regurgitant systolic murmurs in the cardiac cycle, shown in red, with the FHSs shown in grey. The ejection murmur can be seen to start after the S_1 sound, with a peak intensity at mid-systole and fading before S_2 . The regurgitant murmurs can be seen to be either holosystolic or having a peak intensity in early- or late-systole. Adapted from [54].

before both components of the S_2 sound, it is an ejection murmur. If the murmur reaches both components of the S_2 sound (holosystolic), it is likely to be a regurgitant murmur. However, more recently, regurgitant-type murmurs have been further divided into early-, holosystolic and late-systolic [54]. Early-systolic murmurs most commonly occur due to VSD, but can be present as a result of MVP, MR and TR [54]. MVP and cases of acute MR result in late-systolic murmurs [54]. An illustration of various systolic murmurs is shown in Figure 1.11.

1.3.4.2 Diastolic murmurs

Diastolic murmurs are not as prevalent as systolic murmurs as blood flows through the heart valves during ventricular diastole at a low pressure. However, regurgitation of the semilunar valves or stenosis of an atrioventricular valve can cause a diastolic murmur. While some systolic and continuous murmurs are innocent, diastolic murmurs are always pathological [79]. These murmurs can be grouped into semilunar regurgitant, ventricular filling and atrial systolic murmurs. An illustration of diastolic murmurs is shown in Figure 1.12.

Semilunar regurgitant murmurs, or early diastolic murmurs, start immediately after the closure of the leaking aortic or pulmonary valve, as a result of the higher pressure in the aorta and

pulmonary artery as compared to the ventricles [54]. The intensity of these murmurs may rise as the pressure in the ventricles continues to fall. These are the most common pathological murmurs found in the general population, followed closely by mitral regurgitation [79]. An early diastolic murmur is shown in Figure 1.12a.

Ventricular filling murmurs, or mid-diastolic murmurs (illustrated in Figure 1.12b), are caused by stenosis of the atrioventricular valves or by increased atrioventricular blood flow [54]. These murmurs occur in mid- to late-diastole as a result of the delay in the ventricular pressure falling below that of the atria (see Figure 1.7). With stenosis, the intensity of the murmur does not give a direct indication of the severity of the valve damage. Rather, the duration of the murmur is elongated because of the prolonged diastolic times, caused by narrowing of the heart valves [79]. In the case of mitral stenosis, a mid-diastolic opening snap may occur 0.03 to 0.14s after S_2 . In addition, a mid- or late-diastolic murmur may be present [103].

Atrial systolic murmurs, or late-diastolic murmurs (illustrated in Figure 1.12b), occur during the contraction of the atria, and hence occur slightly before the first heart sound. They are an indicator of atrioventricular valve stenosis or complete heart block - a condition for which the electrical impulse of the SA node does not reach the ventricles [54]. These murmurs are often confused with a splitting of the first heart sound, because of their short duration and proximity to S_1 [79].

1.3.4.3 Continuous murmurs

Continuous murmurs start during systole and extend up to and beyond S_2 . These are most common as an innocent venous hum (see Section 1.3.4.4), over the neck of a child [79]. However, a continuous murmur can be an indication of patent ductus arteriosus (PDA), caused by an abnormal connection between the aorta and pulmonary artery which normally closes during infancy. This pathological murmur is best heard at the pulmonary auscultation position. A continuous murmur is illustrated in Figure 1.12c.

1.3.4.4 Innocent murmurs

Innocent murmurs, also called functional, benign, innocuous or physiological, are present during auscultation in the absence of structural or physiological disease [90]. They are common in

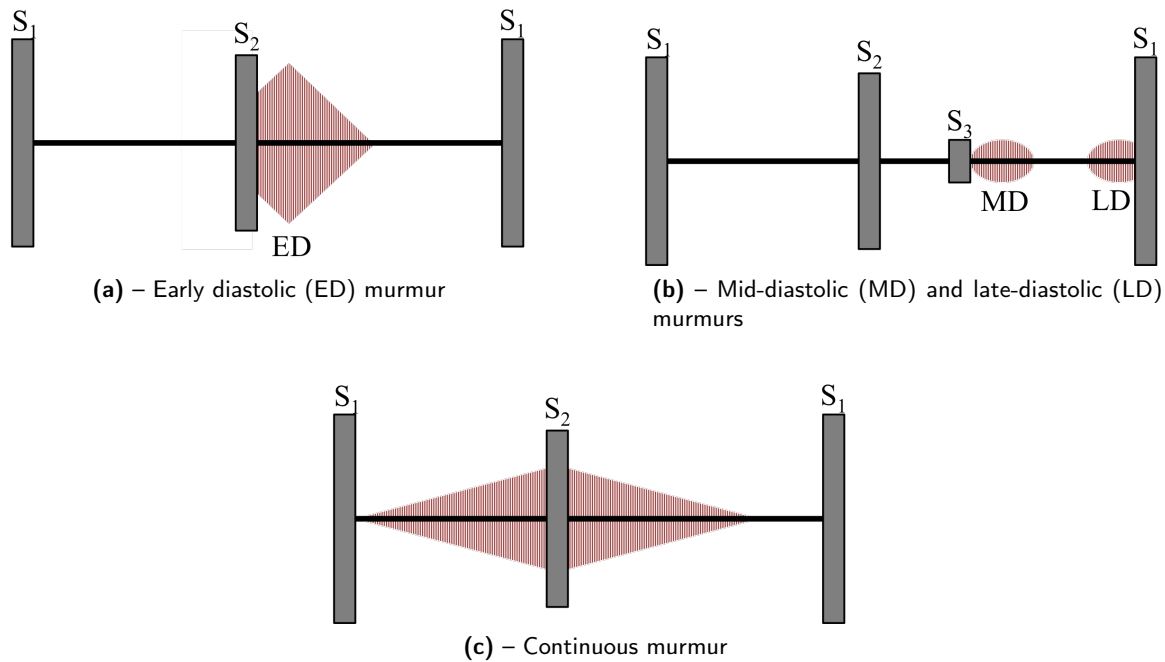


Figure 1.12 – Diastolic and continuous murmurs in the cardiac cycle, shown in red, with the heart sounds shown in grey. Adapted from [54].

children, with up to 80% of children having an innocent murmur of some sort [54]. They are generally quiet, and vary greatly with the position of auscultation and the patient's orientation [104]. There are six variations of innocent murmur, occurring during systole or continuously [54]. Innocent murmurs never occur during diastole. An illustration of innocent murmurs and where they are best heard during auscultation is given in Figure 1.13.

The most common innocent murmur is “Still’s murmur” [105], or the vibratory systolic murmur [90, 106]. This is typically audible between the ages of two and six, and is often heard up until adolescence. It is confined to early- to mid-systole, is low-intensity and has a musical quality. It is thought to be caused by the ejection of blood into the aorta, and may be easily confused with murmurs caused by mitral regurgitation and VSD. However, in contrast to the pathological murmurs of mitral regurgitation and VSD, Still’s murmur is confined to early- to mid-systole and is less “harsh”, with a frequency below 200 Hz [107, 108].

The pulmonary flow murmur, or pulmonary ejection murmur, is heard in children, adolescents and young adults. It is a type of ejection systolic murmur, caused by the exaggerated normal vibrations of the pulmonary artery during ejection. It is confined to the aortic auscultation area and not heard at the apex (see Section 1.3.3.1).

The peripheral pulmonary arterial stenosis murmur, also called pulmonary flow murmur of

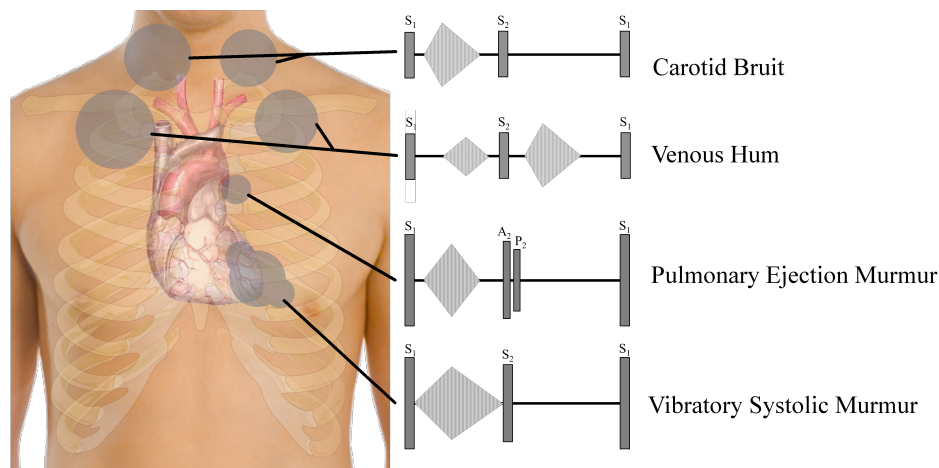


Figure 1.13 – Illustration of innocent murmurs, showing the relative intensities of the various sounds and the auscultation areas where they are best heard. The S_1 and S_2 sounds are shown in darker grey, while murmurs are shown in a lighter grey shade. The aortic and pulmonary components (A_2 & P_2) of S_2 are shown for the pulmonary ejection murmur. Adapted from [54].

newborns, is heard frequently in infants and newborns. It is an ejection systolic murmur of low intensity, starting in early or mid-systole.

The supraclavicular systolic murmur, or carotid bruit, is heard in children and young adults, in the first two-thirds of systole. It is caused by turbulence in the carotid arteries, and hence, is best heard over the carotid arteries in the neck.

The venous hum, a continuous murmur, is heard in children, with varying intensity. It is caused by disturbances in the flow of blood in the veins. The diastolic component is louder than the systolic, and is most audible below the clavicles. This murmur is very similar to that caused by pathological PDA. However, the venous hum of PDA is loudest during systole.

The final innocent murmur, mammary arterial soufflé, is a continuous murmur that is most common in pregnancy and in lactating women. It is normally confined to systole, but may extend into diastole.

1.3.4.5 Murmurs specific to rheumatic heart disease

Rheumatic heart disease (RHD), discussed in Section 1.2, is the damage to the heart valves caused by the immune system's response to untreated *Group A Streptococcal* (GAS) infections. The mitral valve is the most commonly damaged heart valve in RHD, followed by the aortic valve [109]. Children under ten-years old most often have pure mitral regurgitation, while

mixed mitral and aortic valve disease (regurgitation and stenosis) is most common in adults, from 20 years of age onwards [63]. Murmurs are present in 50% of patients who have had acute rheumatic fever [109].

As highlighted in Section 1.3.4.1, the murmur caused by rheumatic mitral valve regurgitation is holosystolic, meaning that it starts with S_1 and extends to the aortic component of S_2 . It may reach a crescendo in late systole [101]. However, especially in acute cases of RHD with minimal regurgitation, the murmur is likely to be early- or late-systolic rather than holosystolic [79, 109]. These murmurs are best heard at the apex of the heart [109] (see Section 1.3.3.1), and therefore may be easily confused with the innocent vibratory systolic murmur. Children with mitral regurgitation are often asymptomatic, apart from the heart murmur [54].

In summary, pure mitral regurgitation, and hence a systolic regurgitant murmur (Figures 1.11b - 1.11d), is the most common early valvular manifestation in RHD. This can be accompanied by aortic valve damage after the second decade of life, leading to an ejection-type murmur of aortic stenosis during systole (Figure 1.11a). Because of the involvement of mitral valve damage in cases of RHD, murmurs are most likely to be audible when auscultating at the apex and tricuspid auscultation areas.

1.4 The motivation for mobile phone-based rheumatic heart disease detection

Deaths caused by RHD are eminently preventable: simple penicillin treatment can save someone with RHD from heart failure. However, RHD is still responsible for 1.4 million deaths each year, with most of these cases being children and young adults [110].

The success and cost-effectiveness of penicillin-based secondary prophylaxis to halt the progression of RHD have been proven [67, 74]. However, such treatment first requires at-risk individuals to be diagnosed with the disease before they develop serious valve damage. Sadly, due to the lack of healthcare workers and resources in the areas where RHD is most prevalent, the children with the highest risk of having the disease have the lowest chance of being diagnosed [74].

Currently, people with RHD are diagnosed at a late stage, when they have advanced disease [63]. If treatment is available at this stage, it is often surgery, which is the least cost-effective treatment of RHD [59, 71] and a luxury only available to those in high-income countries [63]. The need for screening devices was highlighted by Carapetis [110], who identified the urgent need to “explore ways to diagnose cases earlier in their natural history... so that secondary prophylaxis can be delivered at a stage when progression to severe RHD might potentially be prevented”. In order to reach those in developing areas which have the highest prevalence of RHD, such a screening device needs to be low-cost [10] and usable with minimal training.

A mobile phone provides the basis for such a screening device. The growing number of mobile phones in Africa, including smart phones, was identified in Section 1.1. Apart from their ubiquity, mobile phones are an attractive platform as they inherently contain many of the components needed for an automatic screening device. They contain a battery, which is relatively easy to keep charged and negates the need for a constant power supply - often lacking in low-resource settings [7]. They are equipped with various sensors, such as microphones, and smart devices contain the computing power necessary for the development of an automatic screening device. Their intrinsic data communication capabilities would allow the transfer of screening data to a central medical record repository, enabling the implementation of an organised screening programme across various sites.

The detection of pathological murmurs via auscultation, which is both low-cost and non-invasive, is used to identify patients with RHD. The use of computer-aided auscultation has been identified as a promising solution to identify patients with heart pathology [85, 86, 87], specifically for RHD in low-resource settings [111]. The equivalent sound-processing capabilities of mobile phones offer an attractive solution for RHD detection.

An automatic mobile phone-based auscultation system is, therefore, a possible screening solution for RHD. An application running on a smartphone may not require extensive training of users. Screenings for RHD using such an application could be undertaken by school teachers. The potential for school-based screening programmes have been highlighted by Engel et al. [52], who identified “the need for national health systems to consider the development of school health services to address preventable chronic conditions such as RHD”. Equipped with such a screening device, teachers could automatically screen their classes for RHD. Children

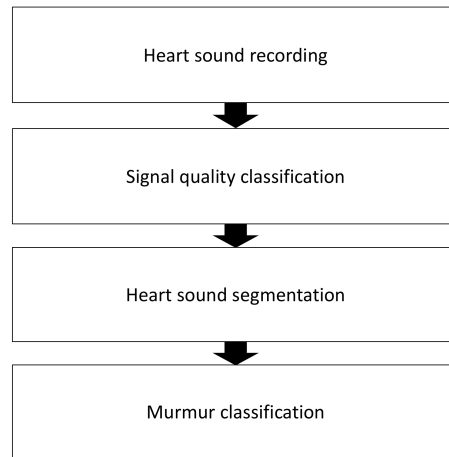


Figure 1.14 – Flow diagram illustrating the major design steps of an automated mobile phone-based auscultation device.

at high-risk of having the condition could then be identified and referred for gold-standard echocardiography screening.

This thesis outlines the development of such a diagnostic tool, from the initial hardware design, to the automatic classification of RHD from recordings made with such a device. The major design steps are shown in the flow-chart of Figure 1.14.

Necessary for such a device to record adequate heart sounds is a stethoscope-like attachment for the mobile phone. The design of such an attachment that does not require any additional specific equipment, but that can rather be made with locally-available resources, has the obvious benefits of circumventing possibly inadequate supply chains and extra expense. Such a design is discussed in Section 2.1, along with the various datasets used in this thesis.

When analysing heart sound recordings made using a mobile phone, it is important to assess the *signal quality* of the recordings (discussed in Chapter 3). This is necessary, as many of the recordings will be of low-quality because of the low-cost nature device and the minimal medical training of those expected to use the application. To ensure high-quality recordings which are necessary for the accurate detection of pathology, immediate feedback on the acceptability of the recordings needs to be given to the user.

A major requirement in the analysis of heart sound recordings is the identification of the fundamental heart sounds, S_1 and S_2 , or *heart sound segmentation*. Murmurs, pathological or otherwise, can only be classified once the positions of these sounds have been identified [79].

The development of heart sound segmentation methods are discussed in Chapter 4 of this thesis.

The final step in the automatic analysis of heart sound signals is the *classification of murmurs*. The classification of pathological murmurs in heart sound recordings, made with both a low-cost mobile phone-based stethoscope and electronic stethoscopes, is investigated in Chapter 5.

The thesis concludes with an summary of the presented work, highlighting areas requiring future work. The barriers that would need to be overcome to implement such a screening tool for RHD are also discussed.

Chapter 2

Datasets and fieldwork

This chapter will outline the heart sound datasets that were investigated in this thesis. A total of three different datasets were analysed, comprising PCGs recorded using four different devices. Two of these datasets contained recordings made with a mobile phone equipped with a low-cost stethoscope attachment. The design of this device is initially discussed (Section 2.1), followed by a description of the two datasets that employed such a device (Sections 2.2 and 2.3). Finally, a dataset of electronic stethoscope recordings supplied by the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology, recorded from patients with mitral valve prolapse, is outlined in Section 2.4.

2.1 Mobile stethoscope design

The need for a low-cost, mobile phone-based stethoscope for the automatic detection of heart pathologies has been motivated in Section 1.4. The chosen design of such a stethoscope in this work was based on the work of Kuan [39]. Kuan postulated that a simple stethoscope attachment, constructed with locally-sourced parts, attached to a mobile phone, would circumvent the need to source pre-manufactured stethoscopes, which are expensive and rely on an already established supply chain. In Kuan’s thesis, she investigated the suitability of a number of mobile phones, microphone attachments and low-cost stethoscope designs for the recording of physiological signals.

A key factor for physiological audio recording is the frequency response of the mobile phone's sound card. Because the main frequencies in human speech are above 250 Hz (see Figure 1.9), lower-cost mobile phones have focussed their audio processing capabilities above the frequencies present in heart sounds (24-144 Hz - see Section 1.3.3.2). However, due to the need for improved audio playback for the high-quality music, mobile phone manufacturers have increasingly focussed on producing undistorted low-frequency audio. This allows the duplex sound cards of these mobile devices to record minimally-distorted physiological audio down to 20 Hz.

The key findings of Kuan's investigation into low-cost stethoscope attachment design were:

- The stethoscope attachment should be constructed from a solid material, such as stainless steel or titanium. Such a material is less prone to sound attenuation as compared to a soft material, such as plastic. In addition, a solid material has the added benefit of amplifying higher frequencies which may otherwise be attenuated [79].
- The stethoscope should have a funnel shape. Such a shape amplifies sound waves that enter the larger diameter mouth of the funnel by the time they reach the neck of the funnel. The typical size for the mouth of a modern stethoscope is 40-50 mm in diameter.
- When performing recordings, there must be an airtight seal between the stethoscope mouth and the body. Without an airtight seal, sound waves are easily dissipated before reaching the recording microphone.
- The electret microphone of a mobile phone hands-free kit, placed into the back of such a low-cost stethoscope attachment, provides adequate quality recordings of physiological signals.

Based on these findings, a stethoscope attachment was constructed using a stainless steel egg-cup, as shown in Figure 2.1a. A 15 mm diameter hole was drilled into the neck of such a cup, into which the microphone of a mobile phone's hands-free kit was placed. This microphone was held in place using a putty adhesive, which kept the microphone securely in position, while providing an air-tight seal around the microphone. A rubber ring was glued to the 48 mm mouth of the egg cup to ensure patient comfort during recordings. A diaphragm was not used, because of the difficulty in securing a diaphragm to the mouth of such a cup. The final design of the stethoscope attachment is shown in Figure 2.1b.

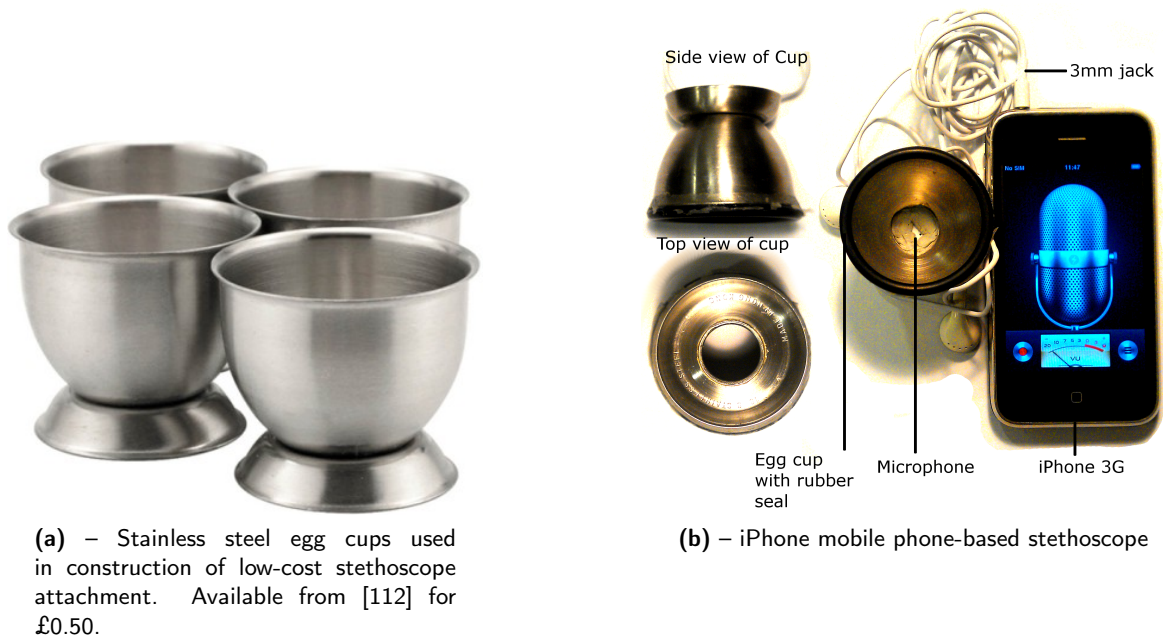


Figure 2.1 – iPhone mobile phone-based stethoscope, constructed out of stainless steel egg-cups (2.1a) and the microphone from a mobile phone hands-free kit plugged into an iPhone 3G (2.1b).

2.2 Mobile stethoscope pilot dataset

The first dataset collected as part of this thesis formed the initial pilot study for the mobile phone-based stethoscope device. This data was collected in collaboration with clinicians from the University of Cape Town. This data was initially used to assess whether such a device could record high-quality heart sound recordings. This, as well as the automatic classification of signal quality, is discussed in Chapter 3.

This study was approved by the Human Research Ethics Committee based at the Health Science Faculty of University of Cape Town (HREC REF: 568/2010). It was conducted in the Cardiac Clinic at Groote Schuur Hospital in Cape Town, South Africa.

All adult patients attending the clinic were eligible to participate in the study. Each potential participant was approached and informed of the study in a private consultation room. Written consent to participate in the study was obtained. Over a period of one month, 151 participants (97 female) were recruited for the study, dictated by the throughput of the hospital. Age (51.4 ± 16.7 years) and BMI (26.61 ± 3.26 kg/m²) measurements of the participants were recorded and the data were de-identified, with each participant being assigned a unique numeric identifier. Of the 151 patients recruited, 100 had valve replacements (with a mixture of natural and

prosthetic valves), 25 had congestive heart failure, 12 had pacemakers and 11 had congenital disorders.

Three different devices were used to record data from each patient. Firstly, a Nonin (Plymouth, Minnesota, USA) Onyx II 9560 finger pulse oximeter was used to capture photoplethysmography (PPG) data from each patient. This was used as the reference signal to compute the heart rate, with which later analysis of the phonocardiogram (PCG) recordings will be compared. The sampling frequency of this device was 75 Hz. The PPG data from the pulse oximeter was streamed in real-time via BluetoothTM to a study laptop (Apple (Cupertino, California, USA) 15" MacBook Pro).

Concurrently with the pulse oximeter readings, PCG recordings were made using two different devices. The first device was a 3M (Saint Paul, Minnesota, USA) Littmann 3200 electronic stethoscope. The device used BluetoothTM to stream data to proprietary software (Zargis CardioScan), which exported the PCG recordings to a standard uncompressed “Wave” file with a sampling frequency of 4 kHz and an amplitude resolution of 16 bits (manufacturer specifications).

The second PCG recording device used was an Apple iPhone 3G mobile phone, which recorded PCG data using the phone’s native audio sampled at 44.1 kHz and stored it as a “Wave” file with 16-bit resolution. This device was fitted with the low-cost stethoscope attachment outlined in Section 2.1, as shown in Figure 2.1b. The frequency response of the iPhone 3G can be seen in Figure 2.2, showing an almost flat response even down to low frequencies, leading to minimal distortion. This is important, as heart sound frequencies reach as low as 24 Hz [92], while a lower-cost mobile phone with a poor-quality sound card may distort the heart sounds. All the cardiac auscultation recordings collected using the iPhone device were copied to the study laptop using BluetoothTM.

All recordings were performed by a non-medically trained research assistant in order to replicate the role of an untrained healthcare worker. These recordings were made with the patient in the supine position, at the mitral auscultation position, after the patient had been resting for five minutes. A recording in progress is shown in Figure 2.3. All recordings were 60 s in length and all data recorded were stored using a matching numeric identifier between the different recordings in a password-protected data partition on the laptop’s hard drive. This was backed

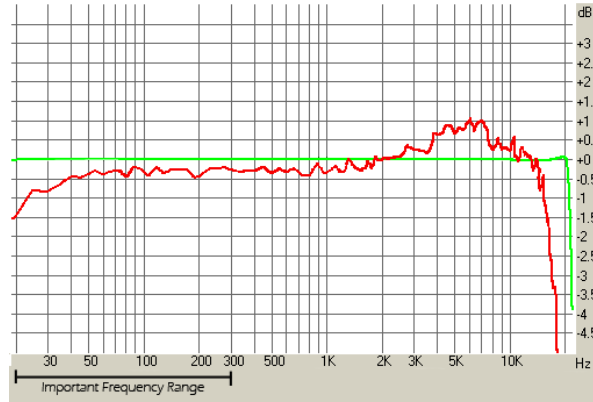


Figure 2.2 – Frequency response plot for the mobile phone sound cards in an iPhone 3G (green) and a Nokia 3500 (red). The iPhone device shows an almost flat response down to low frequencies, while the lower-cost Nokia device can be seen to add distortion to audio recordings. The important frequency range for heart sound analysis is shown. Adapted from [113] and [114].



Figure 2.3 – A mobile phone-based heart sound recording taking place at the Cardiac Clinic at Groote Schuur Hospital in Cape Town.

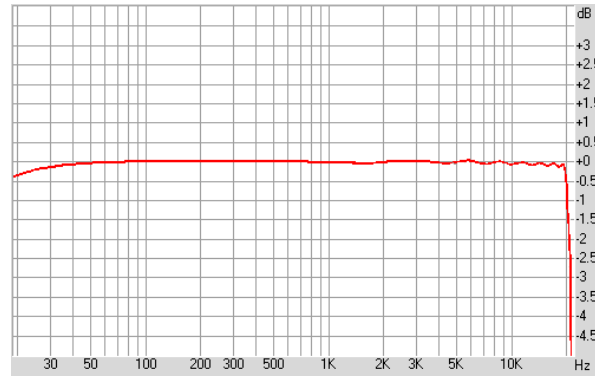
up daily onto a password-protected external hard drive. The data were then stored on a secure server within the Institute of Biomedical Engineering at the University of Oxford. A summary of the devices used is given in Table 2.1.

2.3 Rheumatic heart disease dataset

The dataset collated for the purpose of RHD detection was collected in collaboration with doctors from the University of Cape Town. This study was approved by the Human Research Ethics Committee based at the Health Science Faculty of University of Cape Town (HREC REF 028/2006), while Oxford’s collaboration with this study was approved by the Oxford Tropical Research Ethics Committee (OXTREC minimal-risk reference number 42-12).

Table 2.1 – Recording devices used for University of Cape Town mobile phone dataset

Recording type	Device Type	Sampling Frequency (kHz)
PPG	Nonin Onyx II 9560 pulse oximeter	0.075
PCG	Littmann 3200 electronic stethoscope	4.0
	iPhone 3G mobile phone	44.1

**Figure 2.4** – Frequency response plot for the Samsung Galaxy SII I9100 mobile phone's sound card. The Galaxy SII device shows an almost flat response down to low frequencies. Adapted from [116].

In this trial, researchers from the University of Cape Town aimed to assess the efficacy of different diagnostic tools (manual stethoscope, electronic stethoscope, hand-held echocardiography and standard echocardiography) for detecting RHD [115] by recalling young adults who had been identified with RHD and normal controls from a previous study, which had identified a total of 2720 children and young adults [52]. In addition to these diagnostic tests, the paediatric cardiologist (Dr. Liesl Zühlke) made recordings using the mobile phone-based stethoscope, the design of which is outlined in Section 2.1.

The mobile phone used to perform PCG recordings in this trial was a Samsung (Suwon, South Korea) Galaxy SII I9100. At the time of the trial, this was Samsung's premier smart phone, boasting 1GB of RAM, 32GB of internal memory and a flat frequency response down to the low frequencies required for heart sound recordings, as shown in Figure 2.4 [116]. Attached to this mobile device was a low-cost stethoscope attachment identical to that shown in Figures 2.1b, but equipped with the Samsung device's native microphone.

A custom application was developed in order to capture heart sound recordings on the mobile stethoscope. This application recorded uncompressed audio at 11.025 kHz with 16-bit quantisation. It was ensured that neither automatic gain control nor speech processing filters were employed during audio recording. This application guided the user through the five 30-second recordings that were performed per patient. These included: a recording from each of the



Figure 2.5 – The mobile clinic used for rheumatic heart disease screening by the University of Cape Town, pictured at the Vanguard Clinic outside of Cape Town where many recordings took place.

four classic auscultation positions on the chest (see Figure 1.8); and a final recording from the auscultation position where a murmur, if present, was heard (and the mitral area otherwise).

The electronic stethoscope employed in this study was a 3M (Saint Paul, Minnesota, USA) Littmann 3200.

While the majority of participants were screened in a hospital clinic at Groote Schuur hospital in Cape Town, many of the recordings took place in a *mobile clinic*, pictured in Figure 2.5. The mobile clinic was equipped with an examination bed (positioned behind the driver’s seat) and a waiting area (in the rear of the van), and was used to perform screening for RHD at schools and local clinics, rather than the participants having to travel to the hospital in the city to be screened. Apart from the noise sources that can be identified in Figure 2.5 (such as people talking and cars or trucks driving past), the clinic needed to generate its own electricity for the medical equipment using a portable generator. This generator was mounted below the driver’s seat - almost directly below the examination bed where heart sound recordings were performed. While a relatively quiet generator was used, this noise still contaminated the heart sound recordings performed using the mobile stethoscope.

The procedure for the screening of each trial participant was as follows:

1. A paediatric cardiologist made mobile phone-based stethoscope recordings from the four classic auscultation positions while the participant lay in the supine position. These recordings were stored using the unique patient identifier that was assigned during the initial trial which identified the subjects. This same identifier was used to match these recordings with the subsequent recordings and diagnoses.
2. The cardiologist, blinded to the subject’s true diagnosis, then performed a screening of the participant using *manual* auscultation, deciding whether the patient had no murmur,

- a benign murmur or a pathological murmur *without* the assistance of echocardiography or electronic stethoscope recordings.
3. The cardiologist then performed a recording on the mobile phone-based stethoscope at the auscultation position she determined to be the best to hear the murmur. If no murmur was present, the mitral area was chosen. The position of auscultation on the chest was recorded using the custom-coded application.
 4. The cardiologist then performed a screening using the Littmann electronic stethoscope, recording 20-second PCGs from the four auscultation positions. These recordings were subsequently transferred via BluetoothTM to a laptop computer. The presence of a murmur was detected using Zargis Cardioscan, which is commercial software that automatically analyses PCG recordings for pathology.
 5. The cardiologist then performed a first echocardiographic screening using a hand-held echocardiography device, Vscan, produced by GE Healthcare (Cleveland, Ohio, USA) [117].
 6. Finally, the cardiologist performed a full echocardiographic screening using a Philips (Best, Netherlands) CX50 CompactXtreme ultrasound system [118]. This was used as the gold-standard reference for detecting valvular pathology in the participants. Patients were diagnosed using the World Heart Federation (WHF) criteria for diagnosing RHD [64]. Using these criteria, patients were stratified into three classes: definite RHD, borderline RHD and normal echocardiographic findings. Then, within each of these classes, the diagnoses were further stratified. These divisions are illustrated in Box 1 on the next page. The morphological features that are mentioned refer to valve leaflet thickening, restricted leaflet motion and valvular prolapse. Such features indicate heart valves that may be structurally damaged, but this might *not* lead to valvular malfunction. Importantly, grade “A” for borderline cases of RHD contains patients with some pathological morphological features, but without pathological valvular regurgitation or stenosis. These patients have no abnormal blood flow, and hence no murmur. In addition, “normal” patients may also have some degree of abnormal valvular morphology and physiological regurgitation.

Box 1. 2012 WHF criteria for echocardiographic diagnosis of RHD

Definite RHD (Either A, B, C or D):

- A) Pathological MR and at least two morphological features of RHD of the MV
- B) MS mean gradient ≥ 4 mmHg
- C) Pathological AR and at least two morphological features of RHD of the AV
- D) Borderline disease of both the AV and MV (or pathological AR and at least two morphological features of RHD of the MV for patients > 20 years)

Borderline RHD (either A, B or C):

- A) At least two morphological features of RHD of the MV without pathological MR or MS
- B) Pathological MR
- C) Pathological AR

Normal echocardiographic findings (all of A, B, C and D):

- A) MR that does not meet all four Doppler echocardiographic criteria (physiological MR)
- B) AR that does not meet all four Doppler echocardiographic criteria (physiological AR)
- C) An isolated morphological feature of RHD of the MV (for example, valvular thickening) without any associated pathological stenosis or regurgitation
- D) An isolated morphological feature of RHD of the AV without any associated pathological stenosis or regurgitation

Abbreviations: MR, mitral regurgitation; MV, mitral valve; MS, mitral stenosis; AR, aortic regurgitation; AV, aortic valve.

After a day's examinations, the mobile phone-based recordings were uploaded directly from the mobile phone, along with their anonymised patient identifiers, to a secure FTP server based at the University of Oxford. These recordings were then transferred to a hidden folder on the mobile device as a data backup. The echocardiographic findings for each participant

Table 2.2 – Breakdown of numbers of patients with RHD, those with pathological blood flow, and those with murmurs.

	Total	Pathological blood flow	Pathological Murmur	Innocent Murmur	Uncertain Murmur	Missing murmur data
Definite RHD	15	14	4	0	1	1
Borderline RHD	25	16	1	2	0	4
No RHD	69	2	1	6	0	5

were retrieved at a later date from the University of Cape Town. These mobile phone-based recordings formed the RHD_MP dataset. The electronic stethoscope recordings were stored in a password-controlled netbook and downloaded to a portable hard-drive for archiving and analysis. These electronic stethoscope recordings formed the RHD_ES dataset. The automatic classification of murmurs in the RHD_MP and RHD_ES datasets is investigated in Chapter 5.

In total, 109 participants were seen during this trial. The mean age of the participants was 17.7 ± 4.6 years of age, with mean weight of 63.7 ± 49.8 kg and height of 160.4 ± 10.6 cm. Female participants made up 65% of the dataset^a.

The breakdown of the numbers of participants diagnosed with RHD (using echocardiography) and with murmurs (by clinical auscultation) is shown in Table 2.2. The breakdown of the patients in each diagnostic sub-category (from Box 1) are shown in Figure 2.6, while the number of patients with diagnosed murmurs in each diagnostic category are shown in Figure 2.7.

An important consideration from Figure 2.6 is the number of patients with definite or borderline RHD, but without associated valvular malfunction (sub-categories D for definite cases and A for borderline cases). These patients, while having morphologically pathological valves, would have no abnormal flow of blood in the heart which may cause a murmur. Those patients with pathological heart damage, leading to the abnormal flow of blood in the heart, are defined as having *functional valve damage*.

As shown in Table 2.2 and Figure 2.7, patients without RHD may still have a murmur (most commonly an innocent murmur) while one patient had a severe systolic murmur due to a ventricular septal defect while another had patent ductus arteriosus. In addition, those with

^aBased on the 97 participants with complete age, height and weight data, and 100 participants with gender information.

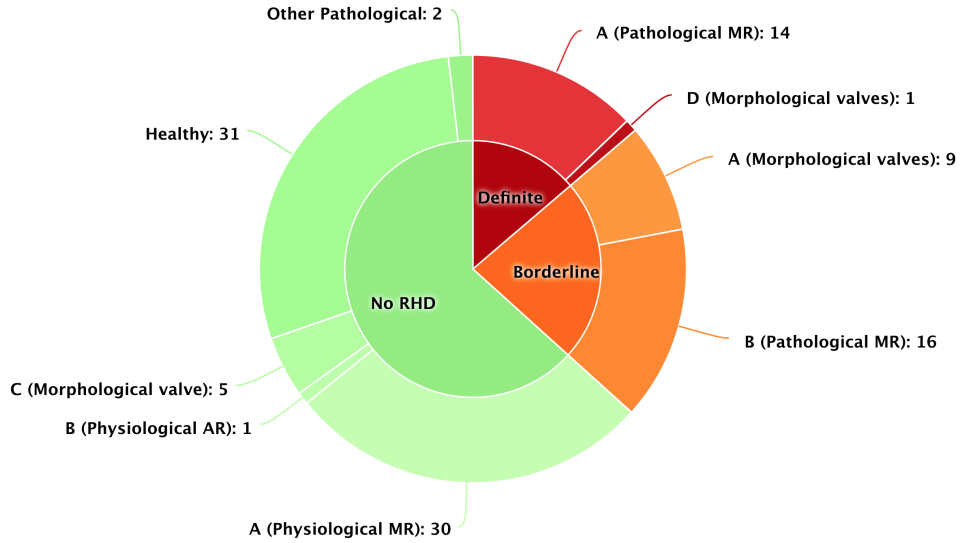


Figure 2.6 – The numbers of patients in each echocardiography-diagnosed RHD sub-category, taken from Box 1. The inner circle shows the definite, borderline or no RHD diagnosis, while the outer circle shows the number of patients in each diagnostic sub-category for that diagnosis.

diagnosed RHD may often *not* have an audible murmur, or have an innocent murmur not caused by RHD. In one of the cases, the cardiologist was unsure as to whether the murmur was innocent or pathological.

Some data were missing from the dataset. Firstly, eight subjects did not have clinical auscultation data (cardiologist-diagnosed murmurs). Secondly, three subjects did not have mobile phone recordings. Lastly, seven subjects did not have electronic stethoscope recordings.

2.4 Massachusetts Institute of Technology automated auscultation system (MAAS) dataset

The data collected as part of the Massachusetts Institute of Technology (MIT) automated auscultation system (MAAS) [119, 120] consist of 405 synchronous 30-40s PCG and ECG recordings from 123 de-identified adult patients. These patients were either known to have mitral valve prolapse (MVP), or were referred to the Massachusetts General Hospital for cardiac screening due to suspicion of having MVP. These recordings were either performed during in-home visits or in the hospital.

Recordings were made using a Welch Allyn Meditron (Skaneateles Falls, New York, USA) electronic stethoscope, with a frequency response of 20 Hz to 20 kHz, which was connected

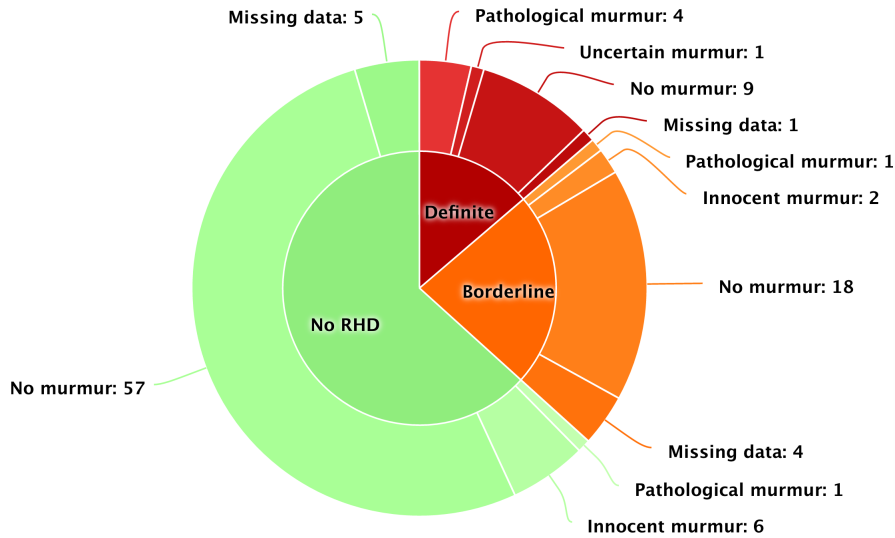


Figure 2.7 – The number of RHD patients with murmurs, as diagnosed by a cardiologist. The inner circle shows the echocardiography-based RHD diagnoses, while the outer circle shows the number of patients with each type of murmur.

directly to a PC that sampled audio at 44.1 kHz with 16-bit quantisation on one channel of a stereo recording. The synchronous ECG signals were recorded on the second channel of the stereo recording [119]. These recordings were performed in an uncontrolled environment, resulting in many of these PCG recordings being corrupted by various sources of noise, such as talking, dogs barking and children playing. Other noise sources included stethoscope motion, breathing and intestinal sounds.

Of the 123 patients, 38 were normal controls, 37 had murmurs relating to MVP, 36 had innocent or benign murmurs, five had aortic disease and seven had other miscellaneous conditions.

Recordings were made from nine different recording positions and patient orientations. From the tricuspid or parasternal auscultation area, 115 recordings were taken with the patient in the supine position, 88 with the patient sitting forward and 42 with the patient squatting. From the mitral auscultation area, 115 recordings were made with the patient in the supine position, 30 with the patient sitting forward and seven with the patient squatting. Four recordings were made at the second right intercostal space (aortic area) and three from the second left intercostal space (pulmonic area) with the patient in the supine position. One recording was taken from the carotid artery (an illustration of the auscultation sites is shown in Figure 1.8). All diagnoses were verified by a cardiologist through echocardiographic examination.

This dataset is difficult to analyse for various reasons. The first is the uncontrolled environment of the recordings, leading to many recordings being corrupted by the various noise sources

mentioned above. Secondly, the patients with no murmur or innocent murmurs in this dataset had been referred by primary care physicians for echocardiographic screening on the suspicion that they had MVP. Therefore, the majority of the non-MVP patients were false positives, hence a false positive detection rate on these recordings of less than 100% would already be an improvement on the clinical diagnosis by the primary care physicians.

2.5 Conclusion

This chapter has described the data collected as part of this thesis. This includes an outline of the design of the mobile phone-based stethoscope that was employed to collect heart sound recordings in two of the described data sets for the distinct purpose of analysis in this thesis. The first set of recordings (from 151 patients) was performed to validate the use of such a device and to perform signal quality analysis, while the second set of recordings (from 109 patients) was performed to directly compare the RHD diagnosis potential of a low-cost automatic diagnostic tool against that of a clinician performing auscultation and echocardiography.

A third dataset, collected by clinicians at the Massachusetts General Hospital, consisted of 405 recordings of synchronous ECG and PCG from patients with suspected MVP. This formed the dataset against which segmentation algorithms were trained and tested.

In total, this represents one of the largest collective sources of real-world heart sound recordings (see the literature reviews in Chapters 4 and 5), which allowed the rigorous analysis of the algorithms presented in this thesis.

Chapter 3

Phonocardiogram signal quality classification

“To err is human. To err repeatedly is research” - @AcademicsSay

An essential task in the initial analysis of heart sound recordings is to ensure that they are of high quality. The quality of the PCG recordings has an obvious impact on the success of the diagnosis of pathologies, whether the diagnosis is made by a clinician or by an automatic classification algorithm. Noise sources that may corrupt the PCG are likely to lead to the misclassification of PCGs, by not detecting pathological murmurs (false negatives) or by generating a false alarm (false positives) for a healthy patient. This is because common noise sources have an overlap with the salient heart sounds and murmurs in both the time and frequency domain. Common noise sources include motion artefacts, physiological sounds caused by breathing and digestion, talking and medical equipment noise (such as in a hospital or clinic setting). The degradation of heart sounds due to such extraneous noise has been documented for many years [93].

Signal quality classification is especially necessary when analysing PCGs recorded by non-medically trained healthcare workers, who may have neither the necessary experience nor adequate training to know how to perform high-quality recordings and to select the best position to place the stethoscope. The need for automated signal quality estimation is likely to be

greater when a low-cost mobile phone-based stethoscope is used, as the chance of lower-quality recordings is greater.

This chapter addresses the first step in the analysis of heart sound recordings, as shown in Figure 1.14. The feasibility of using a mobile phone-based stethoscope to record high-quality PCGs is first investigated. Then, mobile phone-based PCGs are annotated for signal quality, and the level of agreement between these annotations is assessed. Finally, this chapter investigates the automatic classification of the signal quality of these recordings and of recordings performed by a non-medically trained individual when using a top-of-the-range commercial electronic stethoscope. The aim of signal quality estimation algorithms is to provide real-time, automated feedback to healthcare workers in the field, thereby ensuring the recording of high-quality signals. This, in turn, would allow the accurate analysis of such recordings for pathology.

The principal contributions of this chapter to the field of heart sound analysis are:

- A feasibility study of using a low-cost mobile phone-based stethoscope to record high-quality heart sounds
- An assessment of the degree of agreement between five annotators when rating each heart sound recording's quality
- The investigation of nine signal quality indices, based on a review of the literature concerning the quantification of the signal quality of cardiac signals for the automatic classification of the signal quality of heart sound recordings
- The development of novel algorithms for the automatic classification of the signal quality of heart sound recordings, which are robustly evaluated on held-out test datasets. While other researchers have developed algorithms to perform a similar task, the evaluation of these algorithms was handicapped by poor methodology

3.1 Background

Auscultation recordings of the heart are inherently susceptible to noise from a number of sources: endogenous noise from breathing and the digestive system, recording artefacts arising

because of changes in contact between the stethoscope and the chest, as well as ambient noise. Noise from these sources often overlaps with heart sounds in the time and frequency domain [78, 121]. This noise affects the analysis of heart sounds by changing the morphology of sections of the signal that may contain information about heart disease. For these reasons, the identification of noisy PCGs, so as not to attempt to extract heart sounds from these noisy recordings, is a key step in PCG analysis. Braun-Menendez documented the impact of noise sources on the PCG in 1938 [93], stating that noise from extraneous sources was the greatest problem encountered when recording heart sounds. He stated that “I hope that by a careful analysis of all the factors involved the electrical methods will in the future solve all these problems” [93].

The elimination of noise from PCGs is still an open problem. A simple approach to reduce the impact of noise in PCG signals is to apply filtering techniques. Such methods are reviewed first. Then, techniques for identifying potentially noisy segments in the PCG signal are investigated. However, these methods do not explicitly classify the signal quality of the signal. Publications dealing with the explicit classification of the signal quality of cardiovascular signals are few, with most of the literature dealing with ECG signal quality classification. A brief review of the assessment of the signal quality of cardiovascular signals is presented, followed by a review of the methods applied specifically to PCG signal quality determination.

3.1.1 Noise reduction in PCG signals

A straightforward approach to removing noise in PCG recordings, based on the assumption of Gaussian distributed noise (the frequency of which does not overlap with the sounds of interest) is to apply frequency filtering techniques.

In order to attenuate high-frequency noise, researchers have low-pass filtered the PCG at 2.5 kHz [122], 2 kHz [123, 124], 750 Hz [125] and 650 Hz [126]. Higher cut-off frequencies ensure that frequency content due to murmurs, normally up to 600 Hz, is preserved.

In order to attenuate low-frequency noise from motion artefacts and physiological noise sources, researchers have employed band-pass filtering, with cut-off frequencies of 75 and 1500 Hz [108], 20 and 800 Hz [127], 20 and 250 Hz [128], 20 and 300 Hz [129] and 25 and 400 Hz [99].

Moukadem et al. [130] high-pass filtered the signal at 30 Hz, after downsampling the signal to 2000 Hz while Kumar et al. [131] high-pass filtered the PCG at 40 Hz.

Researchers have adapted methods from speech processing to denoise signals. Paul et al. [132] applied spectral subtraction to heart sound signals, employing a modified minimum-mean squared error estimator to estimate the noise spectrum in a signal. Synchronous ECG gating was employed to identify the location of the heart sounds, after which the noise spectrum was subtracted from the signal spectrum. However, only qualitative results were reported.

The above methods may prove ineffectual if the noise sources are non-Gaussian, non-stationary or coloured [133]. In order to overcome this, Tang et al. [133] employed non-linear time scaling to align the different cycles of the PCG to ensure cyclostationarity of the heart cycles, and then performed noise reduction using averaging. The same researchers applied time-frequency decomposition and fuzzy logic to separate heart sound frequency components from noise [134]. While these researchers report high levels of signal-to-noise (SNR) ratio improvement (from 0 dB to 13 dB) as well as correlation coefficients of 0.94 with the original signals after denoising, their methods are only presented for a few test cases (10 recordings of nine pathologies in the first study and five recordings of three pathologies in the second study). They also require precise cycle segmentation provided by synchronous ECG gating.

A number of researchers have applied wavelet-based denoising to PCG signals before further analysis. Hall et al. [135] and Messer et al. [136] report that the wavelet coefficients of heart sounds are much larger than those due to noise. Using this premise, many researchers have sought to eliminate noise by thresholding out the wavelet coefficients not associated with the fundamental heart sounds [137, 138, 139, 140, 141, 142, 143]. Gradolewski and Redlarski assessed various wavelet-based denoising parameters for removing pink noise from mobile phone-based stethoscope recordings [144]. These researchers determined that the fifth order Coiflet wavelet at the tenth decomposition level led to the greatest SNR and “fit” coefficient value when removing pink noise from clean PCGs, improving the SNR from 5 to 13 dB. However, their results were restricted to a small sample size (ten recordings of roughly 5 seconds) without a separate test set on which to evaluate the proposed denoising method.

Wavelet-based denoising alone does not remove noise sources with spectral content which overlaps the heart sounds of interest. To overcome this, Messer et al. [136] averaged the

heart cycles in a PCG recording using ECG gating in order to reduce the effect of noise spikes. This was applied after deriving and applying the optimal wavelet and decomposition level for denoising heart sounds, based on maximising the SNR, on a database of unknown size. This resulted in marked increases in SNR (from 0 dB to 28 dB). Almasi et al. [145] adapted a dynamic model for generating ECG signals to generate PCG signals, and then used this model, combined with an extended Kalman filter, to remove noise in PCGs. These researchers found that their method resulted in marginally higher levels of SNR after denoising 30 recordings with added Gaussian noise (improvement from 0 to 11 dB) when compared to the wavelet denoising method of Messer et al. [136] (improvement from 0 to 10.5 dB without the cycle averaging), though their method, again, requires synchronous ECG recording.

3.1.2 Identification of clean PCG segments

The above filtering techniques work well when eliminating noise outside the frequency distribution of the heart sounds of interest. However, such filtering becomes an issue when noise sources with overlapping spectral characteristics, such as motion artefacts, breathing sounds, coughing and talking, corrupt the PCG signal. For this reason some researchers have attempted to detect noise-free cycles of the PCG for further analysis, rather than suppressing the source of the noise.

Barschdorff et al. [146] identified noisy systolic regions by comparing the variance of the PCG in each half of systole to a threshold determined from the variance of all systolic regions over the entire recording.

Beritelli and Spadaccini [147], in their work on using heart sounds as a biometric identifier, compared the cepstral^a distance between S_1 and S_2 heart sounds within four adjacent heart cycles. They defined the reciprocal of the summation of the cepstral distance between all the sounds as a quality metric. Therefore, sequences of sounds that were most similar in their cepstral content would result in a higher signal quality metric. However, these researchers did not perform a rigorous assessment of their method.

^aThe cepstrum is defined as the squared magnitude of the inverse Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal [148]. More is discussed on cepstral features in Section 5.2.5.2.

Li et al. [149, 150] found noise-free segments of the PCG based on the degree of periodicity, by estimating the cycle frequency spectral density from the autocorrelation signal. The researchers postulated that higher-quality PCG segments exhibit higher periodicity in the form of repetitive peaks in the autocorrelation signal, and illustrated this graphically in their detection of high-quality PCG segments in three simulations and one real-life example. However, no rigorous evaluation of their method was performed.

Syed et al. [120] first discarded ECG-gated heart sound cycles that were corrupted by noise sources which were larger in amplitude than S_1 or S_2 . Then, the remaining cycles were clustered using a k -means algorithm ($k=2$) based on the position of the peaks in their frequency bands. After aligning the cycles using dynamic time warping, the cycles from each cluster were averaged. A prototypical beat was extracted by selecting the cluster which had the greater energy content in the frequency bands other than those corresponding to the S_1 and S_2 frequency content, in order to emphasise pathological murmurs. This prototypical beat extraction process was found to improve the accuracy of their automatic murmur detection algorithm in their dataset of 39 patients (from 66% to 90%). However, their evaluation lacked a separate test data set.

While the aim of the above methods is to identify segments or cycles of the PCG that are clean, they do not explicitly classify the signal quality of a given segment of a signal. They implicitly assume that within a PCG signal, there is a “clean” segment against which to compare the entire PCG. In a PCG recording significantly corrupted by noise, in a best-case scenario, these methods may find the heart cycles that are the least corrupted by noise. However, these cycles may still be too corrupted by noise for accurate analysis or pathology classification.

The remainder of this literature review focuses on the explicit classification of the degree of signal quality of biomedical signals, rather than the identification of segments that are least corrupted by noise. A brief review of the assessment of signal quality for cardiac signals in general is presented first, followed by a detailed review of the two studies directly related to the estimation of the signal quality of PCG signals.

3.1.3 Signal quality of cardiac signals

Moody and Mark [151], in their early work on the detection of noisy beats in the ECG, used the first five components from Principal Component Analysis (PCA) to model the QRS complex. The residual after projecting a QRS complex onto these components gave an indication of the level of noise. Zong et al. [152] classified the signal quality of arterial blood pressure (ABP) signals by identifying each beat, and extracting a number of features (including systolic, diastolic and mean blood pressure, pulse slope information, and time intervals between pulses and the ECG-ABP delay time). Fuzzy feature representation and reasoning was then used to produce the signal quality index. This signal quality assessment led to a sensitivity of 99.8% and a positive predictivity of 99.3% for detecting true blood pressure alarms and a sensitivity of 98.2% and a positive predictivity of 99.4% for detecting false alarms.

Li et al. [153] analysed the quality of a multi-lead ECG using a number of features: the number of matched beats between leads; the kurtosis of each ECG signal; and the ratio of the sum of the spectral power in the ECG between 5Hz and 14 Hz to the sum of the power between 5 Hz and 50 Hz. They then used a Kalman filter to update the heart rate estimate, using their signal quality score to change the measurement noise covariance of the filter. Therefore, the signal quality was used as a confidence measure of the heart rate estimates. Two of the same researchers, Li and Clifford [154], assessed the signal quality of pulsatile signals, specifically photoplethysmography (PPG) signals, using the correlation of an individually generated beat template to beats stretched using dynamic time warping as a feature with a neural network classifier. These researchers achieved 95% binary classification accuracy on unseen test data. Monasterio et al. [155] classified true and false neonatal apnoea desaturations using information about oxygen saturation, heart rate, respiratory rate and a signal quality index (SQI) derived from physiological signals (ECG, impedance pneumogram (IP) and pulse photoplethysmogram (PPG)) with a support vector machine. The features that led to the highest classification accuracy included the minimum heart rate within a 30 s interval before a desaturation, spectral purity of the IP, PPG, PPG- and ECG-derived respiratory rate signals and the slopes of the heart rate and oxygen saturation signals. These researchers achieved 90% classification accuracy of the relevance (true/false) of saturation events.

Redmond et al. [156] assessed the quality of the ECG by locating sections in the signal which

were corrupted by high-frequency noise, low power (due to the lack of a signal) and amplifier saturation. The remaining sections were then analysed for signal quality by classifying all the beats into one of two classes (in order to differentiate between normal and ectopic beats) and comparing each beat with the averaged beat in each class using seven different metrics. The signal quality of each beat was classified using these metrics as features with a Parzen Window classifier [157]. These researchers reported an accuracy of 78.7% when locating poor quality segments of the ECG.

The detection of poor-quality segments in the ECG was the topic of the 2011 Physionet Challenge [158]. The highest official score (93.2%) was achieved by Xia et al. [159] who used SQIs such as the identification of missing leads, thresholding of the autocorrelation and cross-correlation functions and flat line detection. These researchers improved on their approach the next year by using a “matrix of regularity”: a 12×12 matrix with each column corresponding to a lead of the ECG with a row of 12 quality heuristics [160]. The elements were then summed and thresholded to provide a classification of the quality of the 12-lead ECG, achieving 95.1% binary classification accuracy without a train-test split in the data. The aforementioned approach of Li et al. [153] achieved the second-highest score in this challenge (92.6%).

Clifford et al. [161] built on this algorithm of Li et al. [153], and introduced more features, including the skewness of the ECG, the percentage of the signal that appeared to be flat line and the power in the baseline of the signal. These features were used to train either a SVM or neural network classifier on an individual lead, or combined for each of the 12 leads. These researchers reported 96.5% accuracy on the individual leads using four of the signal quality indices and 95.9% on all 12 leads when identifying poor-quality ECG traces from an unseen test dataset.

The signal quality classification of arrhythmic ECG data was performed by Behar et al. [162]. These researchers made use of seven features, including the kurtosis, skewness, QRS power, baseline power, a PCA-based feature and two features based on the agreement between two QRS detectors. While these researchers achieved up to 99% classification accuracy of normal sinus rhythm and up to 95% for arrhythmias, they acknowledged that performance varied greatly depending on the type of rhythm. Therefore, they suggested a signal quality classification model trained specifically for each type of rhythm.

3.1.4 PCG signal quality classification

Two previous publications have directly addressed the classification of the signal quality of PCG recordings. Kumar et al. [121] developed a method to identify noisy segments in the PCG by finding a segment that was noise free and then comparing other segments to it. However, these researchers optimised their method on the entire dataset. They started by computing the autocorrelation signal from the envelope of the PCG. They then estimated the heart cycle duration by analysing the singular values from the singular value decomposition (SVD) of a matrix of different sized windows of the autocorrelation (see Equation 3.12).

This value was then used to find peaks within the autocorrelation function, by finding the first greatest peak and then subsequent peaks not greater than the heart sound cycle length from the previous peak. A radial distance between samples of identified cycles in the autocorrelation greater than a value of 0.8 (set by trial and error) represented heart cycles with similar shapes. The time-frequency domain was then used to split the signal into 15 frequency bands between 0-600 Hz and the autocorrelations of the energy of those bands was then computed. Singular value decomposition was performed on five bands at a time, with a monotonicity criterion on the singular values being used for detecting a “clean” signal. Peak detection was performed on the time-frequency band autocorrelation functions. If 80% of the peaks were within a 5% error window in all 15 frequency bands and the signal satisfied the monotonicity criterion, the PCG segment was labelled as uncontaminated. PCG segments that met both the radial distance and time-frequency criteria were then compared to the entire PCG signal, and the correlation coefficient between the two spectral contents was calculated. Those segments with a correlation coefficient greater than 0.98 were considered to be noise free. These researchers achieved sensitivity and specificity of 95.88% and 97.56% respectively when detecting hand-labelled noise sources in 4781 uncontaminated heart cycles and 1659 heart cycles contaminated with vocal, sensor, physiological and ambient noise.

Naseri and Homaeinezhad [163] developed a heuristic-based signal quality classifier for PCGs. They manually labelled 400 two-minute PCGs (with unknown numbers of normal or pathological patients) as acceptable or unacceptable. They then extracted three features from each PCG after performing normalisation, wavelet-based denoising and baseline wander removal. The first feature was the logarithmic energy in the PCG signal, computed from Gaussian-

windowed segments (of unknown length). The second was the standard deviation of the level of noise above 700 Hz, computed during the process of wavelet denoising. The third feature was the width of each S_1 and S_2 sound, segmented using peak detection in a signal envelope combining frequency and amplitude information (computed by trial and error). The researchers then set upper and lower limits on the acceptable range of these three features, with PCGs having features outside these bounds being marked as poor quality. They reported a specificity of 96.0% and sensitivity of 93.0%. However, their features and limits of acceptable PCGs were optimised on their entire dataset. Their analysis is not repeatable as they left out many important details and values, including the width of many of the analysis windows, the values of the thresholds and details of the smoothing parameters. In addition, the presence of a single noise spike or a split heart sound (lasting longer than 120 ms) in a PCG signal would result in the entire recording being labelled as being of bad quality.

This chapter builds on the current literature by developing novel signal quality classification algorithms that are rigorously evaluated on a dataset of both electronic stethoscope and mobile phone-based heart sound recordings. These recordings were manually annotated by five cardiologists and researchers and the level of inter- and intra-rater agreement was assessed. Rather than applying denoising techniques or identifying the segment of the PCG that is least corrupted by noise (which would fail in very noisy recordings), these algorithms employ a binary classification scheme to assess the clinical acceptability of PCGs. Novel features, based on a review of the literature, were computed and evaluated while the effectiveness of the classification algorithm was rigorously evaluated on unseen test data.

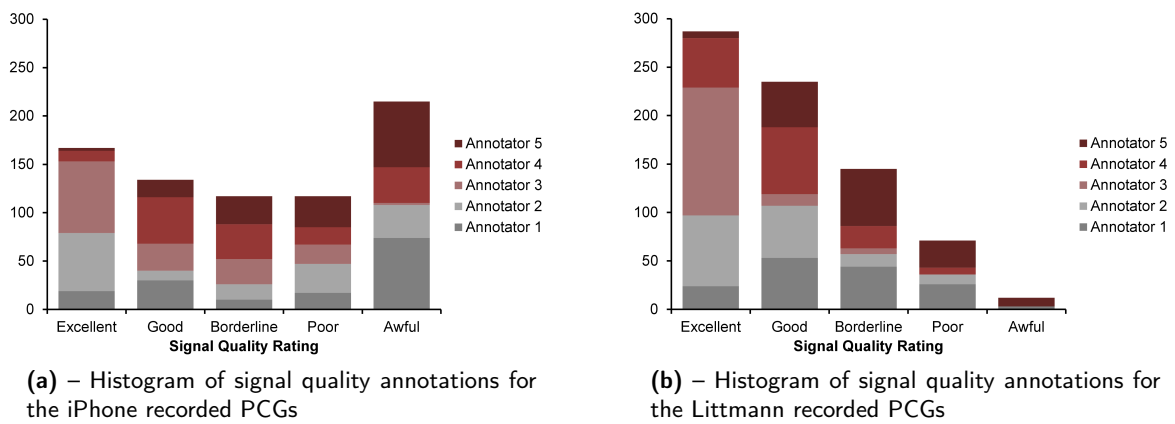
3.2 Methods

3.2.1 Dataset

In order to develop automatic signal quality classification methods, the data recorded in the mobile stethoscope pilot dataset (see Section 2.2) were employed in this chapter. This dataset contained 151 PCG recordings from two devices: a Littmann electronic stethoscope and an iPhone mobile phone-based stethoscope.

Table 3.1 – PCG labelling scheme used by the PCG signal quality annotators.

Rating	Quality Description
1	Excellent quality. An unequivocal diagnosis can be made, with little to no noise in the recording.
2	Good quality - like auscultation with noise but still easily heard. Interpretable but some noise present which means expert judgement is needed.
3	Borderline quality - very faint and poorly heard heart sounds and fairly difficult to interpret.
4	Poor quality - mostly noise with some heart sounds.
5	Awful quality - no heart sounds, only noise.

**Figure 3.1** – Distributions for the annotations for the iPhone and Littmann recordings. These show the numbers of recordings assigned by each annotator to each of the five signal quality levels.

3.2.2 Annotations and annotator agreement

Gold-standard annotations for the signal quality of the recordings were needed in order to develop automatic classification algorithms. These gold-standard annotations were derived from the manual annotation of the signal quality of each recording by three cardiologists and two researchers.

Each annotator performed these annotations in a quiet office, listening to the recordings using an Apple Macbook Pro with a pair of BoseTM SoundTrueTM on-ear headphones. Each recording from both the Littmann and iPhone device was assigned a quality rating of one to five, based on the labelling scheme shown in Table 3.1. To assess how well annotators agreed with themselves (intra-rater agreement), 47 of the 151 recordings were repeated without the knowledge of the annotators. The distributions for the annotations from the five annotators for both devices can be seen in Figure 3.1.

Both intra-rater and inter-rater agreement were analysed. These measurements indicate the difficulty of the labelling task: a high level of agreement between annotators and with themselves, indicates that the labelling task was straightforward, while low levels of agreement points to difficulties in the task. Intra-rater agreement also gives an indication of the consistency of an individual annotator.

The percentage agreement between annotators (the percentage of recordings they assigned to the same class) can give a simple estimate of the level of consensus between annotators. However, such a measurement neither takes into account chance agreement between annotators, nor does it account for ordinal structure in the classes. For these reasons, Cohen’s and Conger’s kappa was employed to measure intra- and inter-rater agreement. Cohen’s and Conger’s kappa is discussed in detail in Appendix A. The results of the agreement between annotators are shown in Tables 3.4, 3.5 and 3.6, and discussed in Section 3.4.

3.2.3 Combining annotations

It was necessary to combine the five annotations into a single, “gold-standard”, signal quality annotation for each iPhone and Littmann recording. While it is known that each annotator would have their own bias and variance, it is expected that when combining the ratings from a large number of annotators, the bias and variance would be compensated for.

The iPhone gold-standard annotation for each record was found by calculating the median of the five separate annotations. The median was selected, as it is more robust to outliers than finding the mean. In order to create a binary classification problem, a threshold between good- and bad-quality records had to be decided. Records with a median value of 1 or 2 were deemed good-quality, based on the annotation scale in Table 3.1. Those with a median annotation of greater than 2 were deemed to be poor-quality. This resulted in 58 good- and 93 poor-quality iPhone PCGs. As expected for a low-cost, mobile phone-based stethoscope, the good-quality recordings form the smaller class.

When finding the median of the Littmann annotations, a threshold of 2 resulted in only 38 poor-quality PCGs. In the classification approach followed in this Chapter (leave-two-out cross-validation, see Section 3.2.6.2), this would result in only 37 recordings per class upon which to train the classifier. This small dataset might lead to inadequate training of the classifier, and

Table 3.2 – The number of recordings in each class, after separating the good- and poor-quality recordings based on the selected thresholding rule (black). The numbers of recordings in each class using the alternative thresholding method are shown in grey.

Device	Merging of Annotations	Good-quality recordings	Poor-quality recordings
iPhone	Median	58	93
	Mean	54	97
Littmann	Mean	92	59
	Median	113	38

hence poor classification performance. Therefore, the *mean* annotation was rather selected as the “gold-standard” for the Littmann recordings. With a threshold of 2, this led to 92 good-quality PCGs, leaving 59 records with a poor-quality annotation of greater than 2. The larger number of poor-quality records was more likely to lead to adequate training of the classifier. The method for merging the annotations and the number of records in each class are summarised in Table 3.2.

An example of both a good- and poor-quality PCG recordings made using the iPhone device is shown in Figure 3.2. From this figure, it can be seen that good-quality PCGs exhibit large amplitude (as compared to the baseline noise level), quasiperiodic heart sounds. In contrast, the poor-quality PCG shows only background noise, with some spikes that may be heart sounds or noise spikes.

3.2.4 Signal quality indices

For the automatic classification of the signal quality of each PCG recording, features, or signal quality indices (SQIs), that indicate different attributes of the signal need to be extracted. Then, a classifier can differentiate between good-quality and poor-quality PCGs by analysing these SQIs. Nine SQIs, selected from a review of the literature, were calculated for each recording.

3.2.4.1 Preprocessing

Before extraction of the SQIs, each PCG recording was downsampled to 1000 Hz using a polyphase anti-aliasing filter. The majority of the frequency content within the fundamental heart sounds is below 150 Hz [92] and hence the Nyquist-Shannon sampling criterion [164] was satisfied.

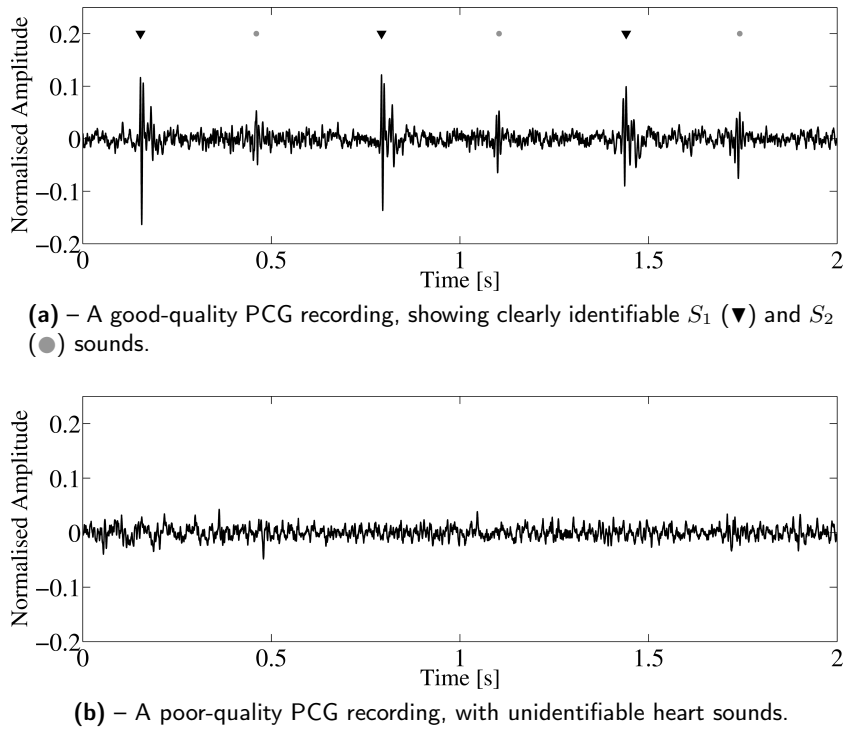


Figure 3.2 – Examples of good- and poor-quality PCG recordings.

3.2.4.2 The autocorrelation function

Several of the extracted SQIs were based on the autocorrelation of the envelope of the PCG recording. The autocorrelation function was employed as it emphasises repeating patterns in noisy signals. The autocorrelation has been used by other researchers for the assessment of signal quality and heart rate in the PCG, based on the assumption that the PCG is a quasiperiodic signal [121, 149, 150].

The autocorrelation of the *envelope* of the PCG was used, as this ensures positive-valued peaks in the autocorrelation, while also removing high-frequency noise in the original signal. The envelope of the signal was estimated using the Hilbert transform as previously proposed [121, 136, 165]. The Hilbert transform has an amplitude response of unity and a 90-degree phase lag to the original signal. This allows the extraction of the analytic signal which negates the negative frequency components of a signal [166]. The Hilbert transform of a continuous signal at time t , $x(t)$, can be defined by:

$$H[x(t)] = \pi \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau = x(t) * \frac{1}{\pi t} \quad (3.1)$$

where $*$ indicates the convolution operator and τ is a delay step of variable size. The signal envelope is calculated by estimating the magnitude, E_t , of the complex analytic signal $A[x(t)]$, which can be expressed as:

$$A[x(t)] = x(t) + jH[x(t)] = E_t e^{j\varphi_t} \quad (3.2)$$

where φ_t is the phase of the analytic signal and:

$$E_t = \sqrt{x(t)^2 + H[x(t)]^2} \quad (3.3)$$

The Hilbert envelope of the signal shown in Figure 3.2a can be seen in Figure 3.3.

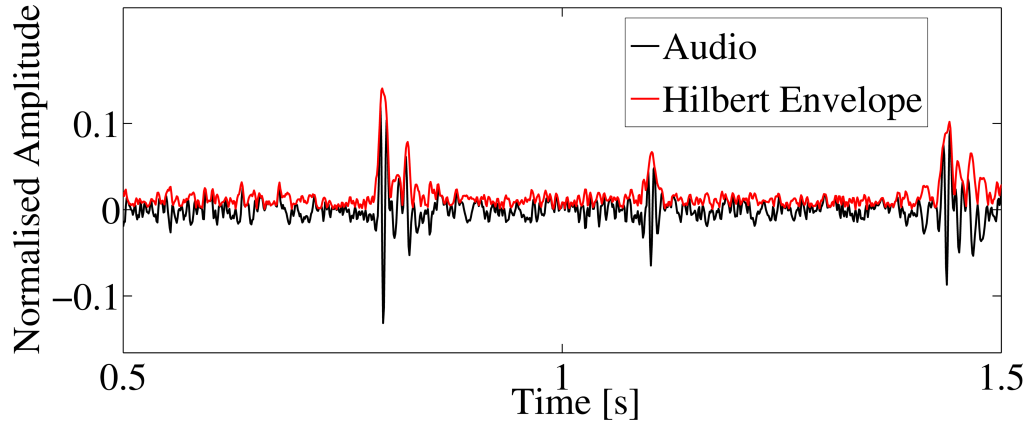


Figure 3.3 – An example of the envelope of a PCG recording (red line), overlaid on the downsampled PCG recording from Figure 3.2a.

The autocorrelation of the signal envelope, r_k , is the cross-correlation between a signal and a copy of itself, once the copy has been shifted by a *lag* of k . This is defined as

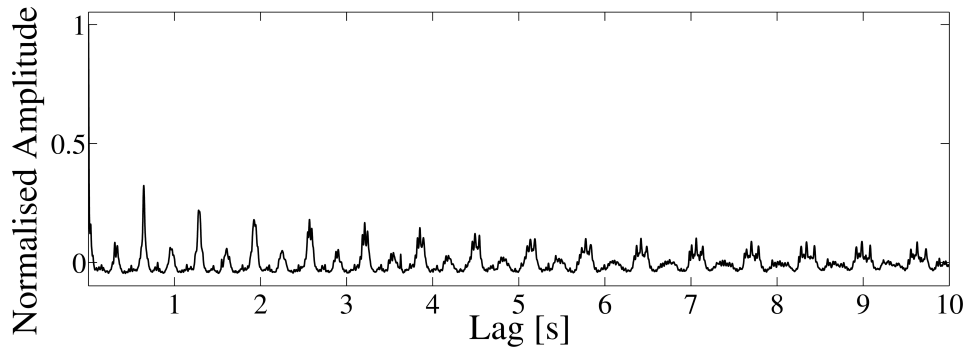
$$r_k = \frac{p_k}{p_0} \quad (3.4)$$

with

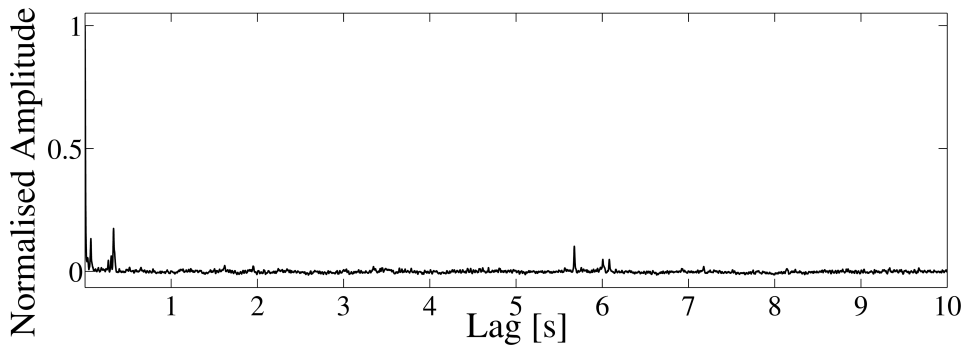
$$p_k = \frac{1}{N} \sum_{t=0}^{N-k} (z_t - \bar{z})(z_{t+k} - \bar{z}) \quad k = 0, 1, 2, 3 \dots K \quad (3.5)$$

where $\{z_t\}_{t=0}^N$ is a discrete sequence of sample length N (the signal envelope)^b, $\bar{z}$ is the mean of the periodic sequence and K is the maximum lag (set to the length of the entire signal envelope in this case) [168] and p_0 is the value of p_k at zero lag. Therefore, the value of r_0 is

^bStrictly speaking, p_k defines the *sample autocorrelation*, as we only have N samples of the signal [167].



(a) – The autocorrelation of the envelope of a good-quality PCG recording, showing prominent peaks caused by the overlap of clear S_1 and S_2 heart sounds.



(b) – The autocorrelation of the envelope of a poor-quality PCG recording, without the distinct peaks seen in a good-quality recording.

Figure 3.4 – Examples of the autocorrelation of the envelope of the good- and poor-quality PCG recordings shown in Figure 3.2. The autocorrelation function is normalised by the maximum correlation value, occurring at a lag of zero. The first ten seconds of the autocorrelation functions are shown, while these signals extend to the length of the recording.

always equal to 1 and is the global maximum value of the autocorrelation signal, as this is the cross-correlation of the signal with an exact copy of itself. Then, one expects higher values of cross-correlation at lag values for which repeated patterns in the signals occur. This is useful when analysing a quasiperiodic signal such as the PCG, as heart sounds occur repetitively. The autocorrelation is also resistant to the influence of white noise. For instance, the autocorrelation of pure white noise results in an impulse function, with a peak of 1 at zero lag [169].

The autocorrelations of the envelopes of the signals shown in Figure 3.2 are shown in Figure 3.4. As can be seen in Figure 3.4a, distinct peaks occur at lags for which the S_1 heart sounds overlap because of the higher correlation between these sounds. Smaller peaks are also seen as a result of the overlap of the S_1 and S_2 sounds. A lack of peaks is seen in the autocorrelation for a poor-quality PCG, as shown in Figure 3.4b.

While autocorrelation highlights the peaks of a semi-periodic PCG, heart sounds become

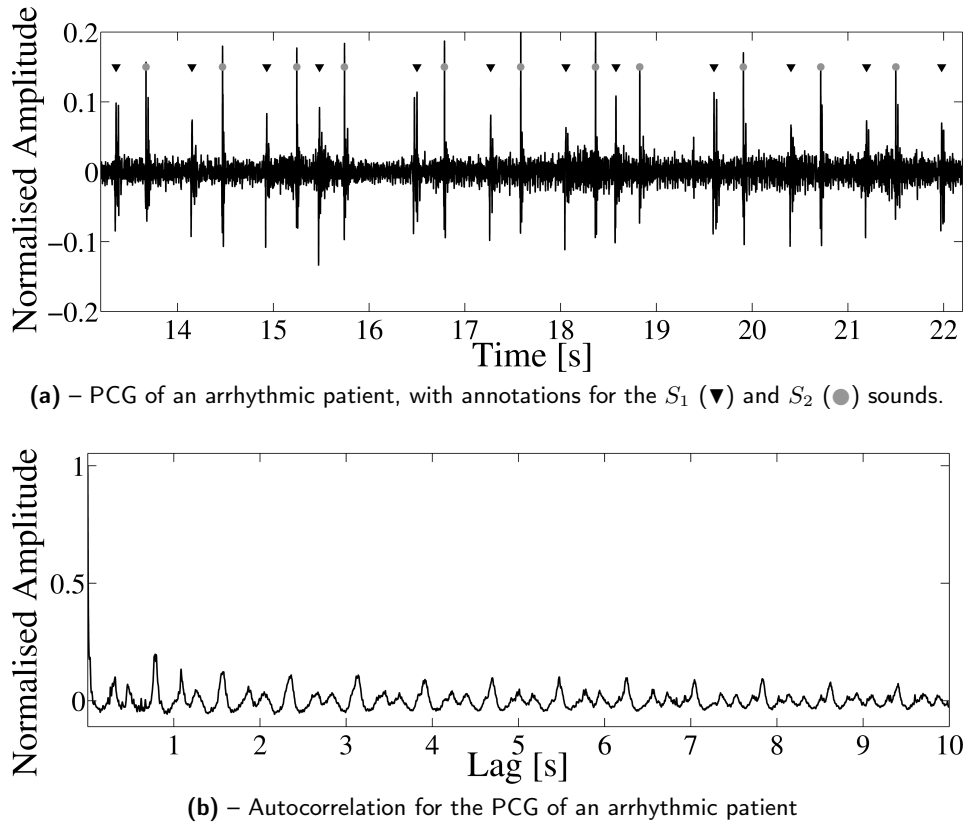


Figure 3.5 – PCG and autocorrelation for an arrhythmic patient, showing large variation in the duration of diastole in the PCG. Even with these large variations, clear peaks are still visible in the autocorrelation because of the shorter segments of periodic heart beats between arrhythmic beats.

irregular with cardiac pathologies such as arrhythmia. In such cases, even in a good-quality recording, because of the lack of periodicity, the autocorrelation may lack clear peaks. The heart sounds of a patient with an arrhythmia can be seen in Figure 3.5a. Large variations in the duration of the heart beat can be seen in this recording, and so the peaks in the autocorrelation would not be expected to be strong for such a signal. However, as can be seen in Figure 3.5b, which shows the autocorrelation of such an arrhythmic PCG, the segments of periodic heart beats between the arrhythmic beats still lead to identifiable (though less pronounced as compared to Figure 3.4a) peaks in the autocorrelation signal. Hence, autocorrelation can still be used to identify the presence of heart beats in arrhythmic patients.

The autocorrelation signal was band-pass filtered using a first-order, forward-backward Butterworth filter with 0.1 and 15 Hz cut-off frequencies. This was introduced to limit the influence of high-frequency noise, resulting in smoother peaks in the autocorrelation signal while removing baseline wander that occurs in recordings with large noise spikes.

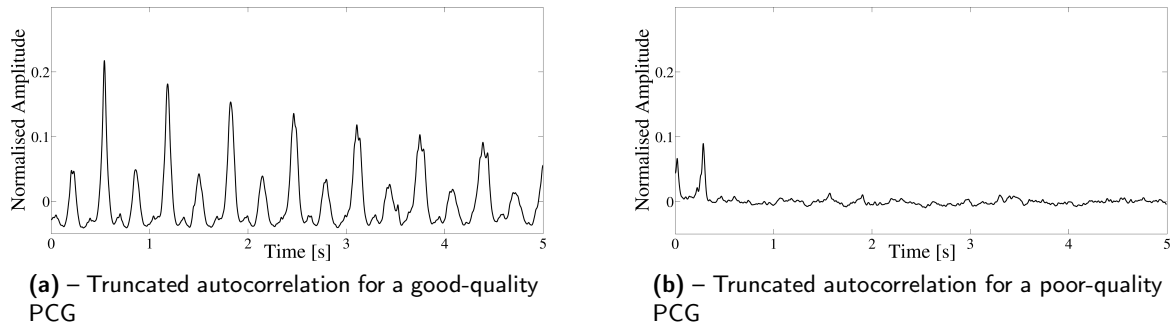


Figure 3.6 – Truncated autocorrelations of good- and bad-quality PCGs

The peaks, if present in the autocorrelation function, can be seen to decrease as the lag parameter increases, because of the natural variation in the beat-to-beat interval in PCG recordings. For this reason, and to reduce the amount of data to be analysed when extracting the SQIs, the first five seconds of the autocorrelation function only were analysed. In most cases, except in the very unusual circumstances of a heart rate below 35 beats per minute, this would allow the analysis of at least three overlapping heart cycles in the PCG and provide enough information to assess the regularity of the signal.

The peak of 1.0 in the autocorrelation at zero lag is much larger than the rest of the signal. This peak was excluded in order to reduce its influence on the computed SQI values. This was performed by locating the first zero-crossing of the first derivative of the autocorrelation signal (greater than 0.2s in order to exclude the entire first peak, with expected width of over 0.12s due to the duration of the S_1 sound – see Section 1.3.3.2), indicating a turning point from a negative to a positive gradient. Inflection points, indicating a gradient of zero but not a change from a negative to a positive gradient, were ignored. The first zero crossing of the derivative indicates the position of the minimum of the first peak in the original function. The autocorrelation was then truncated to the values between this first turning point and the data point five seconds later. The low-pass filtered, truncated autocorrelations of Figure 3.4 are shown in Figure 3.6.

3.2.4.3 $seSQI$ – Sample entropy

The sample entropy is a measure of the *complexity* of a signal, being low in regular signals while being high in irregular signals [170, 171]. It was developed as an extension to *approximate*

entropy, as bias in this measurement leads to inconsistent values between datasets, and it depends on the length of the signal being analysed [170].

In simple terms, the sample entropy measures the extent to which two *epochs* in a signal that are *similar* for m time points, will be similar for $m + 1$ points. Similarity, in this case, is defined as having the maximum absolute difference between the matched pairs of the two *epochs* below a prespecified tolerance, r .

For a discrete time signal, x_i of length N , with $i = 1 \dots N - m + 1$, let:

$$\mathbf{u}_i = \{x_i, x_{i+\tau}, x_{i+2\tau} \dots x_{i+(m-1)\tau}\} \quad (3.6)$$

where m is the length of each epoch under consideration and τ is a constant time lag, set to one sample in this work. A distance vector is then found by:

$$d[\mathbf{u}_i, \mathbf{u}_j] = \max[|\mathbf{u}_i[k] - \mathbf{u}_j[k]| : 0 \leq k \leq m - 1] \quad (3.7)$$

where $j = 1 \dots N - m + 1$, and $\mathbf{u}_i[k]$ is the k^{th} index of vector $\mathbf{u}_i$. Equation 3.7 can be interpreted as the maximum absolute value found when subtracting the individual elements from two of the u vectors. The probability of then finding a vector within a threshold value r from the template epoch $\mathbf{u}_i$ is:

$$\mathbf{C}_i^m[r] = \{\text{number of } j, j \neq i, j \leq N - m + 1, \text{ such that } d[\mathbf{u}_i, \mathbf{u}_j] \leq r\} / (N - m + 1) \quad (3.8)$$

This can be interpreted as the number of distance vectors (Equation 3.7) that are below a threshold r , divided by the total number of distance vectors, giving the probability of finding a distance vector below the threshold. Importantly, this equation excludes self-matches (the case when $i = j$). The probability of finding a distance vector below the threshold, averaged over all the template epochs, is then:

$$\Phi^m[r] = (N - m + 1)^{-1} \sum_{i=1}^{N-m+1} \mathbf{C}_i^m[r] \quad (3.9)$$

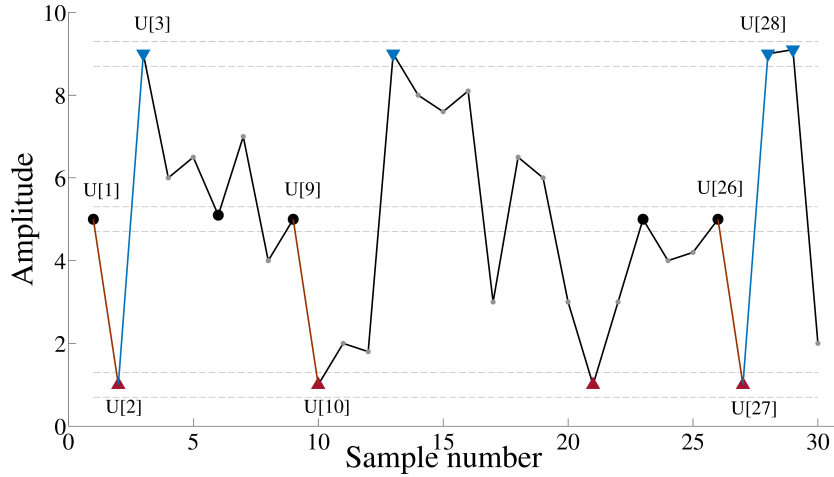


Figure 3.7 – The calculation of $SampEn$ in a simulated time series, $U[1]...U[N]$, for $m = 2$. The grey horizontal bands illustrate the tolerance ($\pm r$) around the first three template points. Points within the tolerance of $U[1]$ ($U[2]$)($U[3]$) are marked as ● (▲) (▼). Points that match the $U[1] - U[2]$ sequence ($U[9] - U[10]$ and $U[26] - U[27]$) are shown with a red line, while those that match $U[2] - U[3]$ are shown in blue. Therefore, there are two other ●-▲ sequences that match the template $U[1] - U[2]$ sequence, while only one ●-▲-▼ sequence ($U[26] - U[28]$) that matches the three-sample template sequence of $U[1] - U[3]$. These values are used to calculate the $C_i^m[r]$ variable of Equation 3.8. These calculations are then repeated for all the two- and three-sample template sequences ($U[2] - U[3]$ and $U[2] - U[4]$, etc). The number of sequences that match each of the two- and three-sample template sequences are again summed and added to the previous values. Then, the natural logarithm of the ratio between the total number of two-sample matches and the total number of three-sample template matches gives $SampEn$. Adapted from [172].

which is the sum of all the probabilities of the different template vectors being below r . The sample entropy is then given by [171]:

$$SampEn(m, r, N) = -\ln [\Phi^m[r]/\Phi^{m+1}[r]] \quad (3.10)$$

with x_i in Equation 3.6, in this case, being the truncated autocorrelation signal computed in Section 3.2.4.2. $SampEn$ is therefore the logarithmic probability that vectors, which are close in m points, are also close in $m + 1$ points. If the probability is high that m points as well as $m + 1$ points will be similar, then $SampEn$ will be low, indicating a signal that is regular. However, a signal that has different probabilities for m and $m + 1$ points will have a high $SampEn$, indicating a more complex signal. It is expected that the autocorrelation of good-quality PCG signals would have a high value for $SampEn$ because of its large peaks, as shown in Figure 3.4a, while $SampEn$ would have a low value for poor-quality PCG recordings because of the lack of variation in their autocorrelation. An example of how $SampEn$ is calculated is shown in Figure 3.7.

As can be seen in Equation 3.10, the calculation of $SampEn$ depends on three parameters: N ,

m and r . In theory, $SampEn$ is not dependent on the length of the signal, N . However, as found by Kuusela et al. [171], when $N < 200$, the value of $SampEn$ has unacceptably large confidence bounds. In our case, $N = 5000$, and therefore this does not impact the calculation. The value for m was set to 2, as chosen by Kuusela et al. [171] in their evaluation of nonlinear methods used for the analysis of heart rate and blood pressure measurements. The tolerance, r , is often set to be 20% of the standard deviation of the signal [171]. However, when comparing the complexity between normalised signals, a standard level of comparison is needed.

If the duration of a heart sound can be presumed to be 0.12 s (see the expected durations of heart sounds in Section 1.3.3.2), then in the autocorrelation function of a good-quality PCG, a peak resulting from the overlap of two such heart sounds would be 0.24 s in width. A conservative estimate of the magnitude of such a peak, after considering Figure 3.6, is 0.1. If the peak can be presumed to be a triangular shape with a sampling frequency of 1000 Hz, then the gradient on one side of this peak would be $\frac{dy}{dx} = 0.1/((0.24 \times 1000)/2) = 0.0008\dot{3}$. Therefore, in an autocorrelation signal that contains peaks, a difference between adjacent samples exceeding this value would indicate a large gradient and be indicative of a peak caused by overlapping heart sounds. Therefore, a rounded value of $r = 0.0008$ was chosen, leaving the sample entropy SQI defined as^c:

$$seSQI = SampEn(2, 0.0008, N) \quad (3.11)$$

3.2.4.4 $svdSQI$ – The singular value decomposition of windows of the autocorrelation signal

This SQI was adapted from Kumar et al. [121], who used singular value decomposition (SVD) to find windows of the autocorrelation function that were most similar. Let y_τ be semi-periodic with period T' , so $y_{\tau+T'} \approx y_\tau$. In this case, y_τ was the truncated autocorrelation of the PCG

^cA retrospective analysis of the robustness of the classification when varying m and r was performed when using single-feature $seSQI$ classification with regularised logistic regression (see Section 3.2.5). It was found that increasing m led to a decrease in classification accuracy for both the iPhone and Littmann devices, while the results varied little (less than 2%) when varying r between 0.0008 and 0.002. Values of r outside of this range led to decreased performance.

recording. Let $\mathbf{Y}_{T'}$ be a matrix constructed by stacking y_τ samples every T' ms in the form:

$$\mathbf{Y}_{T'} = \begin{pmatrix} y_0 & \cdot & \cdot & \cdot & y_{T'} \\ y_{T'+T_s} & \cdot & \cdot & \cdot & y_{2T'} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ y_{(m-1)T'+T_s} & \cdot & \cdot & \cdot & y_{mT'} \end{pmatrix} \quad (3.12)$$

where m is the number of segments of length T' in y_τ and T_s is the time between successive samples. Then,

$$\rho = (\sigma_2/\sigma_1)^2 \quad (3.13)$$

where σ_n is the n^{th} singular value of the SVD of $\mathbf{Y}_{T'}$, with $\sigma_1 \geq \sigma_2 \geq \sigma_3 \dots \geq \sigma_m$. If the rows of $\mathbf{Y}_{T'}$ are similar, then the matrix will have a dominant initial singular value. Hence, σ_1 will be greater than σ_2 . Otherwise they will be of similar magnitude. T' was varied between 0.43 s and 2 s (corresponding to heart rates between 140 and 30 bpm^d), with a ρ value found for each T' . The lowest value of ρ was chosen as the SQI. Better quality PCG recordings are expected to have more regular peaks in the autocorrelation signal (Figure 3.4a), leading to lower values of this SQI.

3.2.4.5 *haSQI* – The Hjorth activity of the autocorrelation signal

Hjorth activity, normally used for electroencephalography (EEG) analysis, gives the power in a given signal [76]. The Hjorth activity SQI can be defined by:

$$haSQI = \frac{2\pi}{N} \sum_{k=1}^N (r_k)^2 \quad (3.14)$$

where r_k is the k^{th} value of the truncated autocorrelation signal. Good-quality signals are expected to have higher values for *haSQI*, due to the more pronounced peaks seen in the autocorrelation signal of a good-quality signal compared to the lack of peaks seen in a poor-quality signal, as shown in Figure 3.4.

^dThis range was selected, as resting heartrates outside of these bounds were not expected based on the 1st and 99th percentile value of heart rates in individuals above 4 years of age [173].

3.2.4.6 *ccSQI* – The correlation coefficient between the autocorrelation signal and a fitted cosine

As can be seen in Figure 3.4a, the main peaks in the autocorrelation of the envelope of a good-quality PCG signal have a similar shape to a rectified sinusoid. Therefore, a half-wave rectified cosine was fitted to the autocorrelation signal, and the correlation coefficient between the two signals was calculated as an SQI. The first task was to find the best-fitting cosine to the autocorrelation signal^e:

Let the number of samples in a single cosine cycle be defined by:

$$n = f_s / f_{cos}, \quad f_{cos} = \{0.50, 0.51 \dots 2.33\} \text{ Hz} \quad (3.15)$$

where f_s is the sampling frequency of the autocorrelation and f_{cos} is the frequency of the cosine, varying between 0.5 and 2.33 Hz (30 - 140 bpm). Then, n represents the number of samples between peaks in the cosine signal and an integer multiple of n , $j \times n$ gives the position of the j^{th} peak of the cosine signal. The cumulative total of the amplitude of the autocorrelation signal, in windows centred on the peaks of the cosine function, is then found using:

$$\varepsilon = \sum_{j=1}^5 \sum_{k=j.n-120}^{j.n+120} (r_k) \quad (3.16)$$

where r_k is the k^{th} sample of the autocorrelation signal. Equation 3.16 finds the cumulative total in windows centred on the peaks of the cosine function. These windows are set to 120 samples either side of each peak (corresponding to the expected duration of a single heart sound for a 1 kHz sampling frequency), over the first five peaks in the cosine function. The frequency of the best fitting rectified sinusoid is then found by finding the f_{cos} that results in the largest ε , denoted by f_{cos}^* .

The rectified sinusoid is then defined as:

$$rcos(2\pi f_{cost}) = \begin{cases} \cos(2\pi f_{cost}) & \text{if } \cos(2\pi f_{cost}) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.17)$$

^eIn this case, the autocorrelation signal was not truncated, as a rectified cosine has similar shape to the non-truncated autocorrelation with a peak at zero lag.

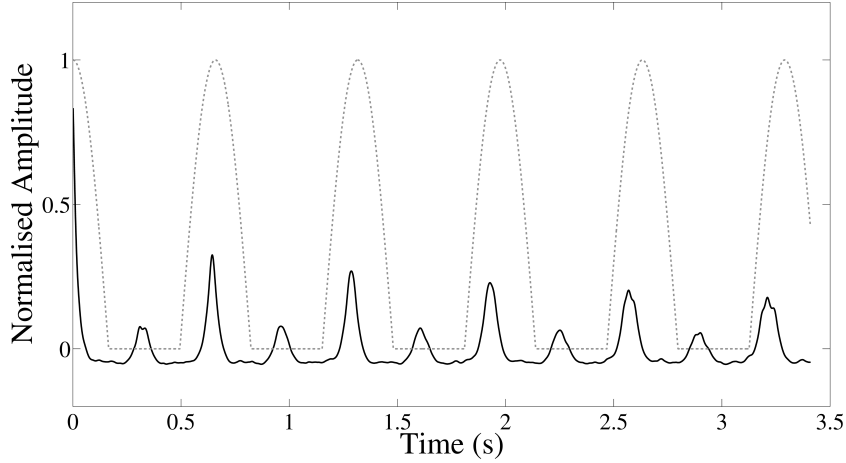


Figure 3.8 – Example of a fitted cosine, used to calculate the *ccSQI* feature. The solid black line shows the autocorrelation signal, while the dotted grey line shows the fitted rectified cosine. The correlation coefficient in this case was 0.46.

The rectified sinusoid that maximises ε is then denoted by $rcos(2\pi f_{cos}^* t)$. Hence $rcos(2\pi f_{cos}^* t)$ represents the rectified cosine that best fits the autocorrelation signal. The *SQI* is then found by calculating the correlation between this best-fitting cosine, $rcos(2\pi f_{cos}^* t)$, and the autocorrelation signal, r :

$$ccSQI = \frac{E\left\{\left(r - \mu_r\right)\left(rcos(2\pi f_{cos}^* t) - \mu_{rcos(2\pi f_{cos}^* t)}\right)\right\}}{\sigma_r \sigma_{rcos(2\pi f_{cos}^* t)}} \quad (3.18)$$

where $E\{x\}$ is the expected value of signal x , μ_x is the mean of signal x and σ_x is the standard deviation of x . The correlation coefficient between the cosine function and the autocorrelation is expected to be higher for better-quality signals because of the more pronounced peaks in the autocorrelation signal. An example of a fitted cosine is illustrated in Figure 3.8.

3.2.4.7 *sdrSQI* – The spectral distribution ratio of the PCG

As identified by Arnott, Pfeiffer and Tavel [92], the dominant frequencies for the first and second heart sounds lie between 24 and 144 Hz, irrespective of sex, body surface areas and blood pressure. Comparing the frequency content within this spectral band to a spectral band for which we do not expect any frequency content may provide a measure of noise content within the signal, and hence a measure of signal quality. The ratio of the total power found between 24 and 144 Hz was compared to the power found from 600 - 1000 Hz^f, thereby excluding

^fIn the case of the calculation of the *sdrSQI* feature, the PCG signals were downsampled to 2000 Hz, rather than 1000 Hz as in the case of the other *SQIs*. Therefore, frequencies up to 1000 Hz were still preserved.

the frequency range where murmurs may occur (up to 600 Hz [100]) that may obscure the fundamental heart sounds.

The power spectral density (PSD) estimate of the downsampled PCG signal was estimated using Welch's method: the signal was divided into two-second segments with 50% overlap, each windowed with a Hamming window function. A two-second segment allows for the inclusion of at least one heart cycle, even for very low heart rates of 30 bpm. The periodogram of each section was then found using the discrete Fourier transform (DFT), with 4096 DFT points, resulting in a frequency resolution of 0.48 Hz/bin. The power in decibels was calculated and the periodograms were then averaged. *sdrSQI* was defined to be the ratio of the sum of the power, P , of the PCG between frequencies, of 24 Hz and 144 Hz to the power between 600 Hz and 1000 Hz as follows:

$$sdrSQI = \frac{\sum_{f_1=24}^{144} P_{f_1}}{\sum_{f_2=600}^{1000} P_{f_2}} \quad (3.19)$$

where P_{f_i} is the estimated power in the PCG at frequency f_i . A better-quality PCG recording with less noise corruption is expected to have a higher *sdrSQI*. A similar method has been employed by other researchers when performing ECG signal quality classification [153, 161].

3.2.4.8 *vSQI* – The variance of the autocorrelation signal

The variance of a signal is a measure of how widely dispersed the signal is from the mean, defined by:

$$vSQI = \frac{1}{N} \sum_{k=1}^N (r_k - \mu_r)^2 \quad (3.20)$$

where r_k is sample k of the truncated autocorrelation signal, and μ_r is the mean value of this autocorrelation signal. A good-quality PCG signal is expected to have a higher variance compared to a poor-quality PCG due to the more pronounced peaks in the autocorrelation signal, as shown in Figure 3.4.

3.2.4.9 *mSQI* – The maximum peak in the autocorrelation signal

The autocorrelation of a good-quality PCG is likely to have large peaks in the autocorrelation function, as can be seen in Figure 3.4a. Therefore, the size of the first peak, other than the peak at zero lag, may give an indication of the quality of the PCG.

The amplitude of the maximum peak in the autocorrelation signal, between a lag of 500 and 2330 samples (corresponding to heart rates between 30 and 140 bpm at a sampling frequency of 1000 Hz), was found, as defined by:

$$mSQI = \max\{r_k\}, k = \{500, 501, \dots, 2330\} \quad (3.21)$$

with higher values of $mSQI$ expected to be present in higher-quality PCG recordings.

3.2.4.10 $kSQI$ – Kurtosis

From the central limit theorem, it is known that random uncorrelated processes, such as thermal noise, tend to have Gaussian distributions. However, correlated processes tend to have non-Gaussian distributions. A measure of the “Gaussianity” of a distribution is the *kurtosis*. The kurtosis is the fourth standardised moment of a distribution and is defined by:

$$kSQI = \frac{1}{N} \sum_{i=1}^N \left[\frac{x_i - \hat{\mu}_x}{\hat{\sigma}_x} \right]^4 \quad (3.22)$$

where N is the number of samples in the signal x_i , and $\hat{\mu}_x$ and $\hat{\sigma}_x$ are the empirical estimate of the mean and standard deviation of x_i respectively. The kurtosis is normally used to assess the Gaussianity of a distribution, with both the peakedness and heavier tails of the distribution influencing its value. However, kurtosis can also give an indication of how prone the distribution is to outliers, with large kurtosis values indicating infrequent, extreme deviations from the mean [174]. The kurtosis of a Gaussian distribution is equal to 3.0. Therefore, noisy PCG signals, with few deviations away from the mean due to noise-obscured heart sounds, are expected to have kurtosis values close to 3.0, while good-quality PCGs, with large fluctuations away from the mean due to loud heart sounds, are expected to have larger kurtosis values. A similar method has been employed by other researchers when performing ECG signal quality classification [153, 161].

3.2.4.11 $bSQI$ – Level of agreement between two independent beat detectors

In good-quality PCG signals, it is expected that independent automatic beat detectors will agree on the position of the detected beats. In noisy signals, it is expected that the level of

agreement between these same independent detectors will decrease, because their individual performance will be less consistent in noisier signals. The same premise has been used to assess the quality of single lead ECG signals [153, 161, 162].

The two beat detectors employed for this metric were a threshold-based method developed by Liang et al. [175] and a hidden Markov model-based approach developed by Schmidt et al. [99]. The beat detection algorithm developed by Liang et al. was selected as it was one of the first segmentation algorithms developed and remains the most cited heart beat segmentation algorithm in the literature. The method developed by Schmidt et al., as described in Chapter 4, was selected as it is the state-of-the-art segmentation method.

The first beat detection algorithm, developed by Liang et al. [175], employs a simple amplitude threshold-based beat detection strategy, ensuring that inter-beat durations fall within (adaptive) thresholds. The procedure for peak detection using this method is:

1. Extract the Shannon envelope of the PCG signal. This envelope is computed by initially normalising the PCG signal: dividing the signal by its absolute maximum value and therefore limiting the amplitude to between $[-1, 1]$. Then, the Shannon energy is computed using

$$E_s = \frac{-1}{N} \sum_{i=1}^N x_i^2 \log x_i^2 \quad (3.23)$$

where x_i is the normalised PCG at time i^g , and N is the window size in samples. N is set to 20 (corresponding to a window size of 0.02 s), with a 10-sample overlap between windows [175]. The Shannon envelope is then extracted by normalising the resulting time series by ensuring zero mean and unit standard deviation.

2. Locate the peaks in the Shannon envelope by locating the zero-crossings of the first derivative of the signal, in the positive to negative direction. Exclude those peaks below a threshold set to be the 50th percentile value of the signal [175].
3. Compute the mean and standard deviation of the time interval between peaks. Set a low-limit and high-limit for the interval between beats, based on the mean and standard deviation, setting these to the mean $\pm$ two standard deviations^h.

^gZero-valued samples of x were set to the next smallest value in x , due to the logarithm of zero being undefined.

^hThe details of the derivation of the low- and high-limits in the original publication are not clear, only stating that these are based on the mean and standard deviation of the time intervals between detected peaks. If the

4. If the interval between adjacent peaks is below the low-limit, there must be an extra, erroneous, peak detected. Therefore, select the peak with greater envelope amplitude and discard the other.
5. If the interval between adjacent peaks exceeds the high-limit, a peak has been missed. Therefore, reduce the threshold by one percentage point to find the lost peaks.
6. Repeat steps 4 to 5 until no peaks lie outside of the low- and high-limits, or until the threshold reaches the 20th percentile.
7. Locate the longest interval between adjacent peaks, marking this as a diastolic region. Compute the mean diastolic and systolic intervals in both a forward and backward direction from this interval.
8. Perform a second pass forwards and backwards from the longest identified interval, marking intervals between peaks consecutively as either systole or diastole, provided the intervals are within a tolerance of the mean systolic and diastolic durations (20% and 40% of systolic and diastolic mean durations, respectively [175]). If the interval between peaks violates these tolerances, then exclude the peak with lower amplitude as being an artefact.

The second beat detection algorithm, developed by Schmidt et al. [99], employs a hidden semi-Markov model (HSMM) that takes into account the duration that the PCG signal is expected to remain in each of the heart sounds, systole and diastole. More detail about heart sound segmentation and this method (and its performance) can be found in Chapter 4.

As the original data which Schmidt et al. used to train their model was not available, the HSMM method in this work was trained on the entire MAAS dataset (Section 2.4). The heart rate parameter needed for this method was computed from the position of the maximum peak in the autocorrelation function, calculated for *mSQI* (Section 3.2.4.9). Schmidt et al.'s algorithm segments the PCG signal into four discrete states, S_1 , systole, S_2 and diastole. The centres of the S_1 and S_2 states were labelled as the positions of the heart sounds.

The level of agreement between the two beat detection algorithms was assessed by using a beat-to-beat comparison. For each beat detected using the HSMM method, the nearest peak mean minus two standard deviations was below 0.2 seconds (corresponding to a heart rate higher than 150 bpm), the low-limit was set to this value.

in time to the detected beat using Liang’s method was located. If this closest beat was within a tolerance from the HSMM beat, it was deemed to indicate a match between the two methods. In ECG beat detection, the standard tolerance used is 150 ms [176]. However, as this is longer than the expected duration of the S_1 and S_2 heart sounds, this tolerance was decreased to 100 ms.

The matches between the two detectors were labelled as true positives (TP). The HSMM-detected beats without a match were labelled as false negatives (FN) and the Liang-detected beats without a match were labelled as false positives (FP). This is the standard method for assessing the performance of ECG beat-detection algorithms [176]. $bSQI$ was then computed as the F_1 score of the detections between the two algorithms. The F_1 score, defined as

$$bSQI = F_1 = \frac{2 \times P_+ \times Se}{P_+ + Se} \quad (3.24)$$

where P_+ is the positive predictivity ($P_+ = TP/(FP + TP)$) and Se is the sensitivity ($Se = TP/(TP + FN)$), and the F_1 score is the harmonic mean between positive predictivity and sensitivity (or precision and recall). The F_1 score was used as it gives a single harmonic mean of Se and P_+ .

3.2.4.12 Post-processing

The above SQIs were normalised across recordings by subtracting their mean and dividing by their standard deviation. The mean and standard deviation values were computed from a training set of data, and the same mean and standard deviation were used to normalise a hold-out test set of data. The details of the train-test split are discussed in Section 3.2.6.

The various SQIs investigated in this work are summarised in Table 3.3. The relationship between all the different pairs of features for both the iPhone and Littmann recording devices are shown in Figures 3.9 and 3.10.

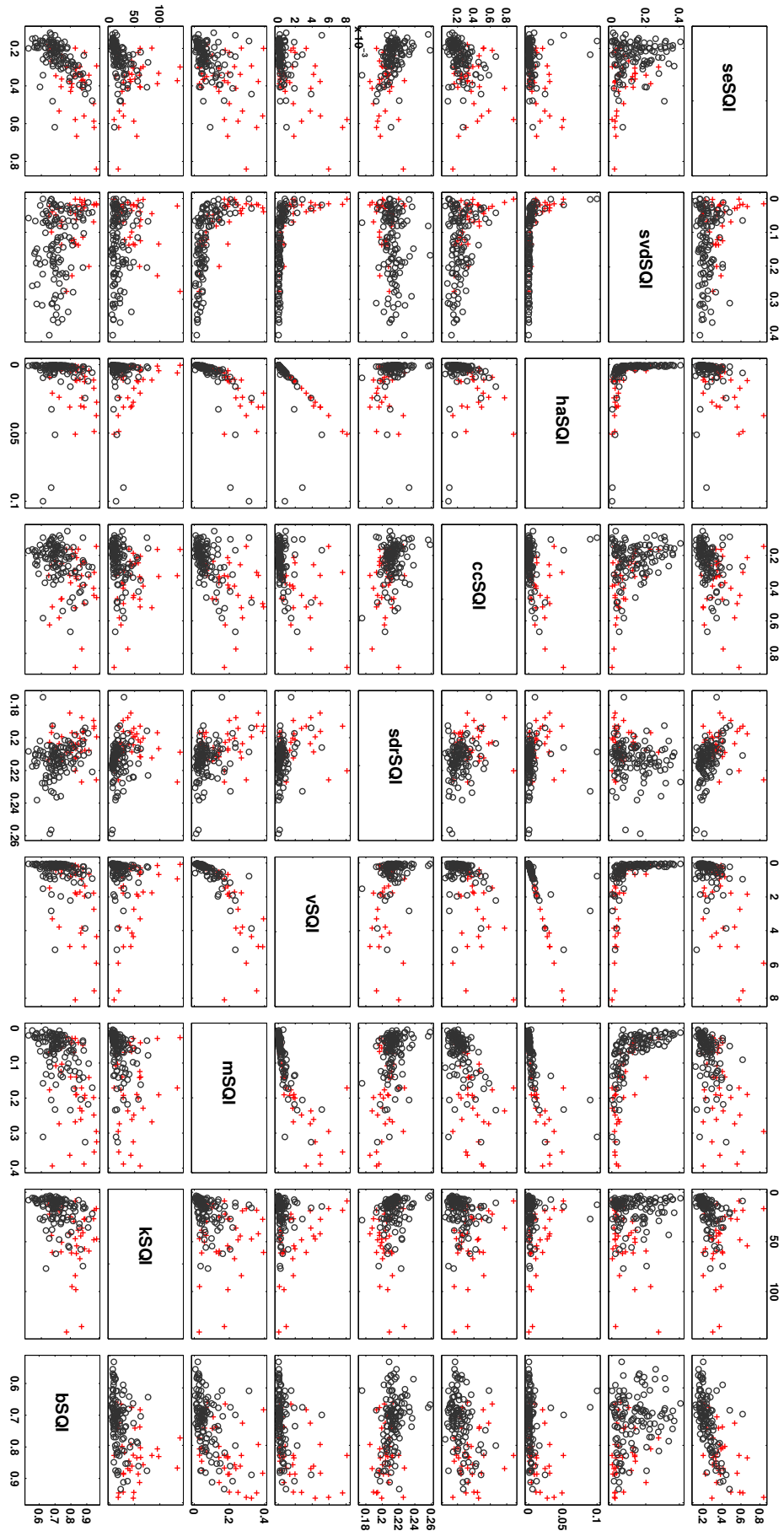


Figure 3.9 – A scatterplot matrix of the iPhone SQIs. Each plot shows a pair of SQIs. The SQI values from the poor-quality recordings are shown as grey circles ($\bullet$), while the SQIs from the good-quality recordings are shown as red crosses ($+$). The overlap between the two different classes for all the different features can be noted.

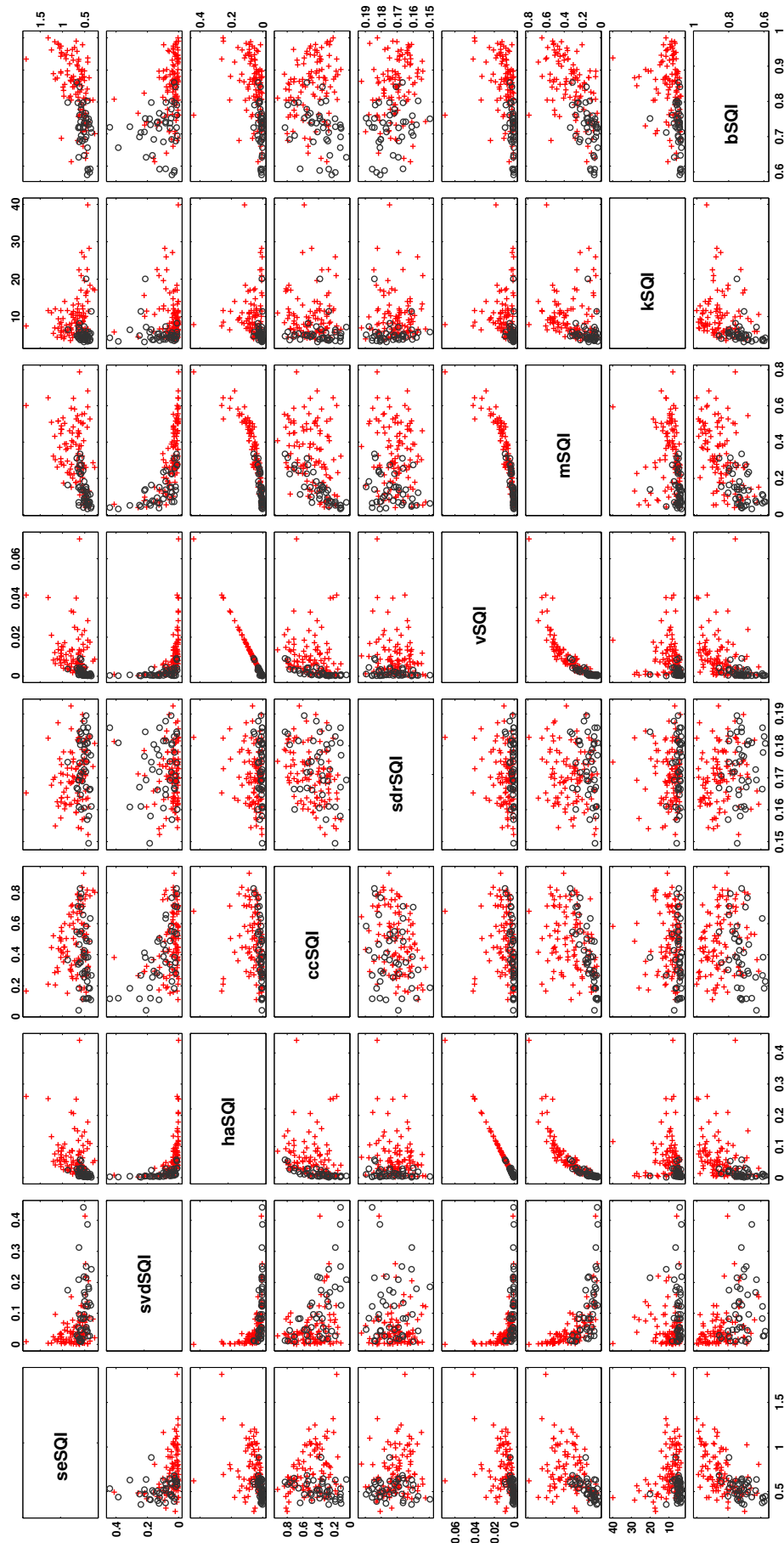


Figure 3.10 – A scatterplot matrix of the Littmann SQIs. Each plot shows a pair of SQIs. The SQI values from the poor-quality recordings are shown as grey circles ($\bullet$), while the SQIs from the good-quality recordings are shown as red crosses ($+$).

Table 3.3 – Summary of SQIs investigated for automatic PCG signal quality estimation

SQI Name	Description
<i>seSQI</i>	Sample entropy of the autocorrelation signal
<i>svdSQI</i>	Ratio of second to first singular values of the SVD of windows of the autocorrelation
<i>haSQI</i>	Hjorth activity of the autocorrelation signal
<i>ccSQI</i>	Correlation between the autocorrelation signal and a fitted sinusoid
<i>sdrSQI</i>	Spectral distribution ratio between the expected frequencies within the fundamental heart sounds and noise
<i>vSQI</i>	Variance of the autocorrelation signal
<i>mSQI</i>	Maximum peak in the autocorrelation signal between lag values of 0.5 and 2.33 seconds
<i>kSQI</i>	Kurtosis of the PCG signal
<i>bSQI</i>	Level of agreement between two beat detectors

3.2.5 Classification methods for signal quality classification

The two methods employed for signal quality classification were regularised logistic regression (LR) and a support vector machine (SVM)ⁱ with a Gaussian or radial basis function kernel. LR was selected, as this classifier is well-known and provides a linear decision boundary. It also provides a low-complexity algorithm against which to compare other classification methods. The SVM classifier with Gaussian kernel was chosen over other machine learning algorithms (such as neural networks) as the training procedure always produces a global optimum because training the classifier is a convex optimisation problem. Therefore, for the same input data, an SVM always produces the same result. In addition, the Gaussian kernel was selected as this produces a non-linear decision boundary in the original input space, thereby providing a comparison between linear and non-linear decision boundaries of the two classifiers. These methods are outlined in more detail in Appendices B and C.

3.2.6 Evaluation

3.2.6.1 Data balancing

Robust training of the classification models required a *balanced* dataset (equal numbers of both classes of data) [178]. Random subsampling of the majority class was performed to achieve this [179], as described below.

ⁱThe LIBSVM package was employed for SVM implementation [177].

For each device, the number of recordings in the smaller class was found. For instance, for the iPhone device, there were fewer good-quality than poor-quality recordings while the opposite was true for the Littmann device (see Table 3.2). Then, this number of recordings was selected randomly from the two classes, resulting in a balanced set of good- and poor-quality recordings.

This, however, results in a large proportion of the recordings from one class not being included in the analysis, with different combinations of the selected data possibly having a profound impact on the success of the classification algorithms. For this reason, the above procedure of randomly selecting a balanced dataset from the total dataset was repeated ten times, resulting in ten different balanced datasets for evaluation of the algorithms. The total results, after being averaged across all folds of the leave-two-out (LTO) cross-validation (explained in the next section), were also averaged across these ten dataset iterations.

3.2.6.2 Leave-two-out cross-validation

As there were only a few recordings in each balanced dataset, LTO cross-validation was performed to provide a robust estimate of the performance of the classifiers. Cross-validation is a process whereby a set of recordings, often called a *fold*, is held out from the total dataset to be a separate evaluation set. The classification algorithm is trained on the remaining data, and its performance is evaluated on the held-out fold. The held-out fold is then re-incorporated into the training set, with a separate fold held out for evaluation. This process is repeated until all folds have been in the held-out set once, and the average result over all folds is computed. LTO cross-validation refers to each fold consisting of a single recording from each of the two classes. When using a balanced dataset, LTO cross-validation ensures that the dataset remains balanced even when performing cross-validation. For each fold, the training and held-out data were normalised by subtracting the mean and dividing by the standard deviation of the data in the training set.

3.2.6.3 Training and evaluation of the LASSO LR model

The optimal λ parameter for the LR model (which controls the degree of regularisation in the model – see Appendix B) was selected by again employing LTO cross-validation on each

training set of recordings. This resulted in a held-out *validation* set against which to optimise the classifier's parameters.

For each training set, a range of 100 λ values was tested. The maximum λ value was selected by estimating the value at which all β coefficients^j would be shrunk to zero, with the remaining 99 values of λ being logarithmically spaced between this maximum value and 0.01% of this maximum value. For each λ value, the average *deviance* across the cross-validation folds was computed. Then, the maximum λ value that resulted in a deviance within one standard error of the minimum deviance was selected. This was a conservative approach to select the least complex model, as it is acknowledged that the deviance values are estimated with an error due to cross-validation^k. The model was re-trained on the entire training dataset with the selected λ and evaluated on the held-out test samples. The classification results were determined by averaging the results over all the test folds. A balanced dataset was randomly selected from the full dataset again, and the entire process repeated. The final results were determined by averaging the test results over all ten dataset iterations.

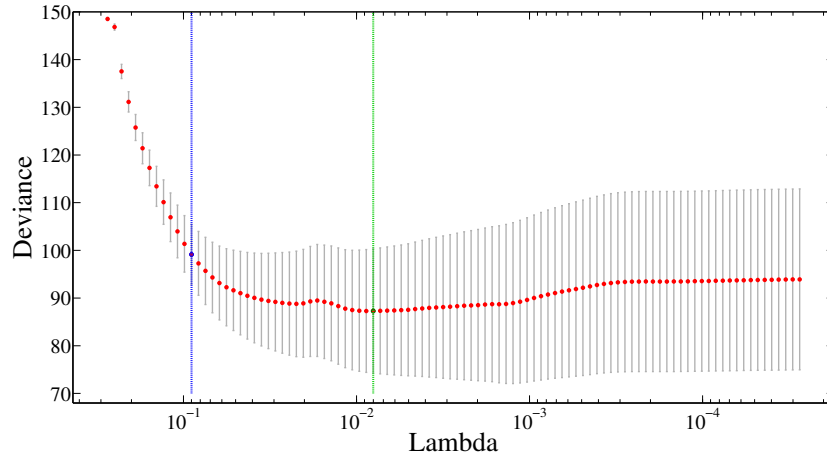
The effect of λ on the deviance and β coefficients is illustrated in Figure 3.11. This figure shows the point of minimum deviance, as well as the selected λ value within one standard error of the minimum deviance.

The LASSO LR model was used to evaluate the classification performance of each *individual* SQI alone. This was performed by following the above procedure, but optimising the model when only using one SQI at a time. In order to identify the most influential SQIs when optimising the model using the *full feature space*, the β coefficients for each SQI were analysed. The β coefficients for each SQI were averaged across each test fold of the LTO cross-validation, and averaged again over the 10 data set iterations, with the coefficients with larger absolute amplitude having a larger influence on the classifier^l.

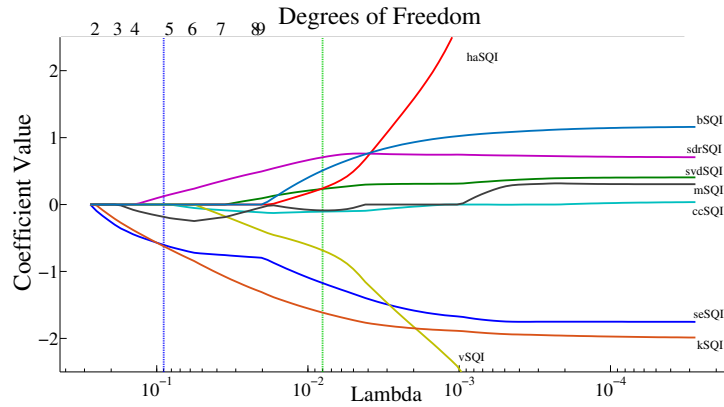
^jThe β coefficients refer to the weighting of each predictor variable, or feature, in the model. More details are given in Appendix B.

^kSee Section 7.10 of [180].

^lA simple comparison of this sort would not be permissible if the SQIs were not normalised, as their magnitudes would influence the magnitude of the β coefficients. However, as the SQIs are normalised to have zero mean and unit variance, a direct comparison of their β coefficients in the LASSO LR model can be made.



(a) – The effect of the λ regularisation parameter on the deviance of a LASSO fit for the iPhone recordings. The deviance is shown with red dots with error bars (in grey). The λ value of minimum deviance is shown with a green vertical line (0.008), while the selected λ parameter (0.0899), is the largest value of λ within one standard error of the minimum deviance, is shown with a blue vertical line.



(b) – The effect of the λ regularisation parameter on the β coefficients for the predictors of a LASSO fit for the iPhone recordings. The λ values resulting in the minimum deviance and the selected λ are shown, as above, with green and blue dotted lines. The degrees of freedom, or the number of predictors with non-zero coefficients, are shown at the top of the plot.

Figure 3.11 – The effect of the λ regularisation parameter on a LASSO fit for the iPhone recordings, showing the deviance and β coefficient values against λ .

The poor-quality recordings were labelled as output target $y_i = 1$ while good-quality recordings were labelled target output $y_i = 0$. As balanced datasets were used (see Section 3.2.6.1), a decision threshold of 0.5 was selected to discriminate between the good- and poor-quality classes. Any output of greater than or equal to 0.5 was classified as poor quality.

3.2.6.4 Training and evaluation of the SVM

As performed with the LASSO LR model, LTO cross-validation was employed for each training set of recordings in order to optimise the model parameters using a held-out validation set. There were four parameters that needed to be optimised for the SVM classifier. These were the number of features to use, the combination of features and the C and γ parameters of the RBF-SVM (which control the penalty on errors in the classification and the width of the Gaussian kernel, both of which contribute to the flexibility of the classifier – see Appendix C). In order to avoid over-fitting of the SVM on the training data, the number of features (or SQIs) was limited to between two and four, based upon the observation of the selected features from the LASSO model. Every combination of features was then evaluated. For example, when selecting three features from a set of nine, there are 84 unique combinations of features and 126 when selecting four features. As suggested by Hsu et al. [181], a grid-search over exponentially growing sequences of C and γ was performed, with C and γ equal to 2^n , $n \in \{-2, -1, 0, 1, 2\}$. These ranges were determined through initial experimentation.

The parameters (number of features, the combination of features, the C and the γ values) that resulted in the highest average classification accuracy over all the validation held-out pairs were then used to train the SVM model on the entire training dataset and evaluated on the held-out test pair. If multiple combinations of parameters led to equal classification accuracy, the combination with the smallest variance was selected, as this would indicate a more stable classifier. This was repeated for all test folds of the LTO cross-validation, with the average test results found by averaging over all test folds. Then, as performed for the LR model, another balanced dataset was randomly selected from the full dataset and the entire process repeated. The final results were determined by averaging the test results over all ten dataset iterations. The performance of *each SQI alone* was also evaluated with the SVM classifier, following the above procedure, but limiting the number of features to one.

In order to assess the influence of each SQI when performing SVM-based classification on the *full feature set*, the process followed above for the LASSO LR model (which involved finding the average β coefficients in the LR model) cannot be performed. This is because the weight vector in the SVM, w , cannot be explicitly computed when employing the kernel trick transformation [182]. Therefore, as the SVM-based model selected the features that led to the

highest average classification accuracy over the held-out validation folds, the number of times a specific feature was included in the model was counted. The ratio of the number of times a feature was selected to the total number of folds would give an indication of the effectiveness of this feature in the SVM-based classification. These ratios were then averaged across all LTO cross-validation folds, and then averaged again over the 10 dataset iterations.

3.3 Results

3.3.1 PCG annotations and annotator agreement

From the distributions of annotations in Figure 3.1 it can be observed that, as expected, the Littmann recordings have far more annotations below 2 than above, indicating a greater proportion of good-quality recordings. The opposite is true for the iPhone recordings: there are more poor-quality annotations than good-quality. 30.4% of the Littmann annotations were borderline or worse. For the iPhone recordings, 40.1% of the annotations were for either “excellent” or “good quality”.

The inter- and intra-rater agreement between individual annotators for the iPhone and Littmann device are given in Tables 3.4 and 3.5. It can be observed from the diagonal elements of Table 3.4 that for the iPhone device intra-rater agreement is at least *substantial* (according to Table A.1), with a minimum value of 0.68 for all the annotators. The highest level of intra-rater agreement for the Littmann recordings is 0.72, but rater 2 has an intra-rater agreement of only 0.18.

The off-diagonal elements of Tables 3.4 and 3.5 show the inter-rater agreement between pairs of annotators. For the iPhone device, *moderate* agreement (0.41-0.6) can be observed for most pairs of raters. *Fair* agreement (0.21-0.4) is observed between most pairs of annotators for the Littmann device. Annotator 1 and 5 exhibited moderate agreement for the Littmann recordings.

Using Conger’s κ to assess the agreement between numerous annotators (Table 3.6), the level of agreement for the iPhone recordings was moderate (0.52), while there was only fair agreement for the Littmann device (0.24).

Table 3.4 – Weighted Cohen's kappa, with quadratic weights, comparing the annotations from each pair of annotators for iPhone recordings. Diagonal elements indicate intra-rater agreement

Annotator	1	2	3	4	5
1	0.68				
2	0.58	0.87			
3	0.31	0.50	0.92		
4	0.63	0.63	0.56	0.90	
5	0.71	0.50	0.29	0.70	0.96

Table 3.5 – Weighted Cohen's kappa, with quadratic weights, comparing the annotations from each pair of annotators for Littmann recordings. Diagonal elements indicate intra-rater agreement

Annotator	1	2	3	4	5
1	0.66				
2	0.19	0.18			
3	0.10	0.26	0.68		
4	0.41	0.21	0.25	0.72	
5	0.57	0.12	0.10	0.43	0.64

3.3.2 Evaluation of SQIs

The average results of the single SQI classification on each of the held-out test datasets for each recording device and classification algorithm can be seen in Figure 3.12. The best-performing single SQI for the iPhone device, regardless of classification algorithm, was *seSQI* ($\approx 78\%$), while *mSQI* was the second best-performing SQI ($\approx 75\%$). For the Littmann recordings, *haSQI* and *vSQI* exhibited the best performance ($\approx 81\%$), while *bSQI*, *kSQI* and *mSQI* all had similar performance for the LR and SVM classification algorithms ($\approx 78.5\%$).

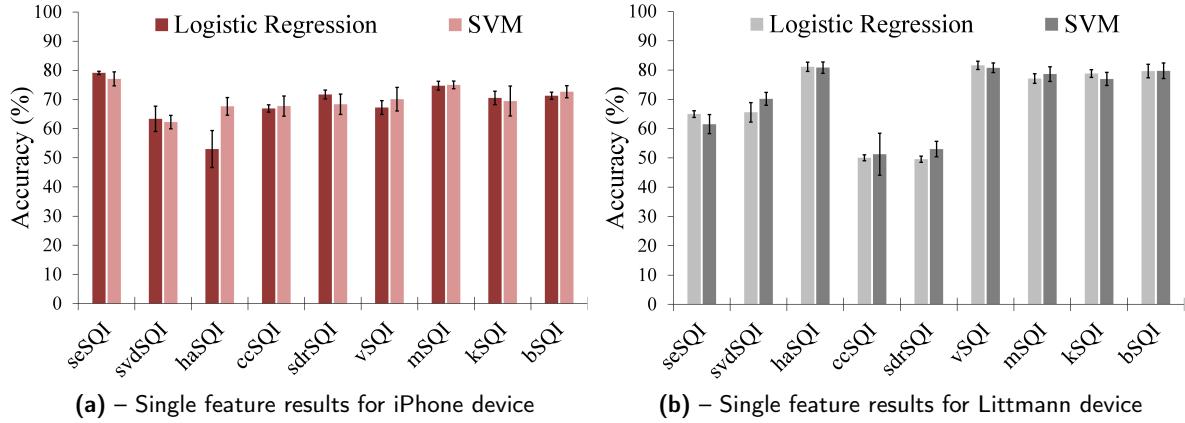
The average values of the β coefficients in the LASSO LR model, for each recording device, are shown in Figure 3.13. These values show the average weight of each SQI in the linear summation within the LR algorithm, and hence give an indication of the influence of each SQI in the model.

For classification of the Littmann recordings, *mSQI* (-0.69), *kSQI* (-0.55) and *bSQI* (-0.47) can be observed to have the coefficients with greatest absolute values. The other SQIs, including *haSQI* and *vSQI* which had the best performance when using a single SQI, had little influence on the classifier.

When classifying the iPhone recordings using the LASSO LR model, *seSQI* (-0.51), *sdrSQI* (0.29), *mSQI* (-0.51) and *kSQI* (-0.54) had the largest absolute coefficient values, while *ccSQI* (-0.06) has a small coefficient value.

Table 3.6 – Conger’s Kappa, using quadratic weights for iPhone and Littmann devices

	iPhone	Littmann
Conger’s κ	0.52	0.24

**Figure 3.12** – Single SQI test set classification accuracy, across recording devices and for each classification algorithm.

For the SVM-based classification, the proportional number of times a feature was selected, as compared to the total number of folds in the LTO cross-validation, is shown in Figure 3.14. This number gives an indication of how influential each SQI was in the signal quality classification.

There is large variation in the feature selection for the SVM, as shown by the large (as compared to the mean) error bars in Figure 3.14. For the iPhone device, *seSQI* is most frequently selected (0.94 ± 0.12), similarly to the LASSO LR coefficient values. In contrast to the LASSO LR, *ccSQI* and *haSQI* were frequently selected in the best-performing model on the validation set, while *sdrSQI* was not.

kSQI can be observed to be selected in almost every fold for the Littmann device with mean annotations (0.99 ± 0.02). *bsSQI* is second place in the list of most-often selected features for the Littmann device, with a proportion of 0.79 ± 0.25 . The selection of these two SQIs agrees with the selection of features in the LASSO LR model for the Littmann recordings. However, in contrast to the LASSO model, *hSQI* is third most selected (0.7 ± 0.2), rather than *mSQI*.

3.3.3 Classification of signal quality

The results of the average classification performance for each recording device and classification algorithm can be observed in Table 3.7. As outlined in Section 3.2.6, these results are the

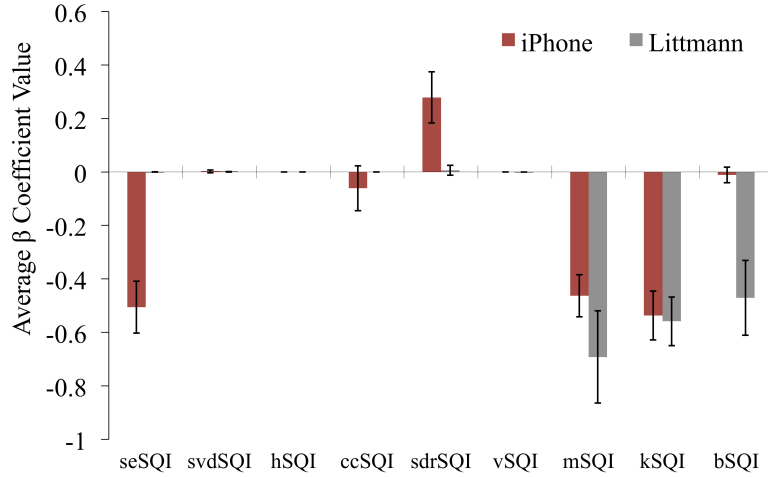


Figure 3.13 – Average values of β coefficients for LASSO LR models for each recording device.

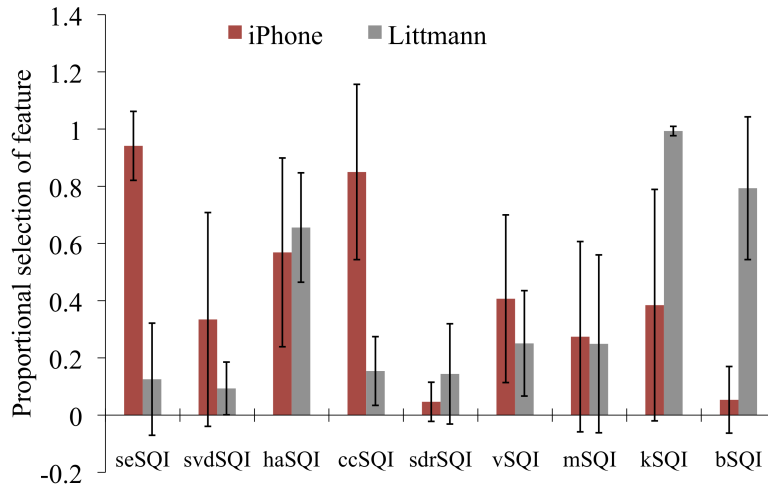


Figure 3.14 – The proportional number of times a feature was selected for the SVM-based classification model, with a value of 1 indicating that the feature was selected for every fold of the LTO cross-validation.

average results over each fold of the LTO cross-validation, and then averaged again over 10 dataset iterations. For the iPhone classification, the LASSO LR model had the highest test set accuracy ($82.2 \pm 1.9\%$) and specificity ($86.7 \pm 2.6\%$), while the SVM-based classification had higher sensitivity ($83.6 \pm 5.2\%$), indicating it is better at identifying poor-quality recordings as compared to the LR model. The LASSO LR model outperformed the SVM across all metrics when classifying the Littmann recordings, achieving an average accuracy of $86.5 \pm 2.2\%$ on the held-out test sets.

The average λ values for the iPhone and Littmann LASSO LR models were 0.076 ± 0.006 and 0.084 ± 0.01 . The average number of features selected, the average C and average γ values for the SVM model were 3.9 ± 0.1 , 1.0 ± 0.3 and 1.2 ± 0.3 for the iPhone classification, and

Table 3.7 – The results of the PCG signal quality classification, showing the two recording devices and the method of classification. The average accuracy (Ac), sensitivity (Se) and specificity (Sp) are shown (%), averaged for every fold of the LTO cross-validation and again across the ten dataset iterations. The average training set results are shown in grey, while the best-performing method for each device is shown in **bold**.

Device	Classification method	Ac	Se	Sp
iPhone	LASSO LR	84.4±1.6	80.0±1.7	88.7±1.8
		82.2±1.9	77.8±1.9	86.7±2.6
	SVM	89.3±1.3	91.1±1.6	87.5±2.4
		80.2±4.3	83.6±5.2	76.7±4.4
Littmann	LASSO LR	87.4±2.3	83.0±3.4	91.8±1.4
		86.5±2.2	81.9±3.2	91.2±1.8
	SVM	89.2±1.7	83.3±3.0	95.0±0.8
		82.8±4.6	77.5±4.9	88.1±5.0

3.5 ± 0.2 , 0.7 ± 0.4 and 0.4 ± 0.2 for the Littmann recordings. An example of the classification boundaries when using the SVM classifier with the iPhone recordings is shown in Figure 3.15.

3.4 Discussion

An investigation into the annotations assigned to the iPhone recordings (see Figure 3.1) has highlighted that 40.1% of these annotations were for excellent or good quality. This illustrates that cardiologists believe that such a mobile phone-based stethoscope is able to record PCGs of diagnostic quality in over a third of cases. A similar number of annotations (30.4%) were borderline or worse for the Littmann device. This indicates that even when using a high-end electronic stethoscope, the quality of PCGs recorded by a non-medically trained healthcare worker can often be of non-diagnosable quality. Therefore, when using either device, it is necessary to start with signal quality assessment to ensure that the PCG recording is of high enough quality before performing any further analysis.

The levels of agreement between annotators in Table 3.5 highlight that there was more disagreement for the Littmann device than the iPhone, shown in Table 3.4. This indicates that these recordings may have been more difficult to annotate. This is thought to be because the iPhone device, if incorrectly placed, records pure noise. Therefore, it was easy to distinguish between good- and poor-quality recordings for the iPhone device. The Littmann device, however, as it is more sensitive, may have recorded short segments of heart sounds, even in poor-quality

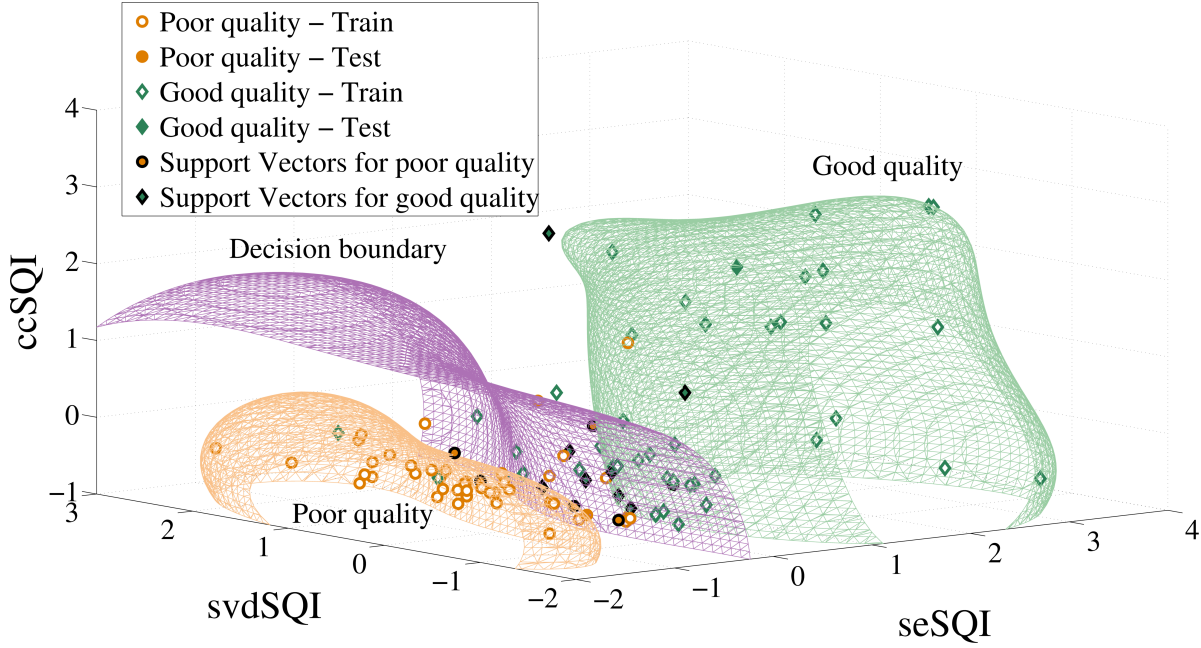


Figure 3.15 – Example of decision boundaries for SVM-based signal quality classification of iPhone recordings, when performing leave-two-out cross-validation. The C and γ parameters in this RBF-SVM were 0.5 and 0.25 respectively. The decision boundary is shown in purple, while the areas in the feature space corresponding to a decision value of 1 for the poor-quality (amber) and good-quality (green) signals are shown. The points chosen as support vectors to generate these decision boundaries are shown in black outline.

recordings. This would lead to some annotators marking it as poor-quality, while others may mark it as good-quality because of the presence of some high-quality beats.

Figures 3.9 and 3.10 illustrate the difficulty of the classification problem, due to the overlap between the classes for each of the nine features (chosen on the basis of the literature review).

When using the LASSO LR model, $mSQI$ and $kSQI$ are the SQIs with highest β coefficients for both the iPhone and Littmann device (see Figure 3.13), confirming the hypothesis that higher-quality PCGs have a large initial peak in the autocorrelation function and are less Gaussian distributed, regardless of recording device. $bSQI$ was influential in the LR-based classification of the Littmann recordings, but not for the iPhone recordings. This was thought to be because Liang’s peak detector [175] is ineffectual in noisier iPhone recordings, and the HMM beat detector was trained on electronic stethoscope data. $sdrSQI$ had large β coefficients for the iPhone recordings, but not for the Littmann-based signal quality classification. This is probably explained by the in-built frequency filters in the Littmann device (which include ambient noise reduction).

Three SQIs had surprising β coefficient values. Firstly, $seSQI$ had a large β coefficient for the iPhone recordings, but an almost zero value for the Littmann recordings. This could be because

the parameters used for this model ($m = 2$, $r = 0.008$) are inappropriate for the Littmann recordings. Secondly, $vsQI$ and $haQI$, upon investigation of Figures 3.9 and 3.10, appear as if they may be useful for differentiating the two classes, because of the separation between these classes when using these features. However, both of these features had negligible β coefficient values for both the iPhone and Littmann devices. This may be because these features have little discriminating ability when used in conjunction with other features such as mQI and kQI .

The large variances in Figure 3.14, indicating the proportional number of times a particular feature was selected for the SVM, indicates inconsistency in feature selection when using the SVM model. For the iPhone classification, similarly to the LASSO LR model, $seQI$ is selected often. However, in contrast to the LASSO LR model, $ccQI$ is also selected often, while $sdrQI$, mQI and kQI are not.

For the Littmann recordings, $haQI$, kQI and bQI were the most often selected features when using the SVM classification method. The selection of kQI and bQI agrees with the higher β coefficients of the LR model, as shown in Figure 3.13, while the frequent selection of $haQI$ and infrequent selection of mQI does not. This may be due to over-fitting of the SVM model on the training data.

The disagreement in feature ordering between the LR and SVM models, along with the large variance in the selection of SQIs for the SVM, may also indicate over-fitting of the input data by the SVM model. This over-fitting is confirmed by Table 3.7, where large discrepancies between the training and test results when using the SVM classifier can be observed. The LASSO LR models, which choose a less complex model, have similar training and test results, indicating that the model did not over-fit the data. Therefore, as expected, the LR model can be seen to outperform the SVM classifier, achieving $82.2 \pm 1.9\%$ vs. $80.2 \pm 4.3\%$ for the iPhone recordings and $86.5 \pm 2.2\%$ vs. $82.8 \pm 4.6\%$ for the Littmann recordings. With a larger dataset, which should reduce the likelihood of over-training for the SVM, the latter might perform at least as well as the LASSO LR model.

In an attempt to reduce over-fitting, the SVM-based classification algorithm was re-run only selecting the three-best performing features as inputs. These were $seQI$, $haQI$ and $ccQI$ for the iPhone and $haQI$, kQI and bQI for the Littmann classification. When only using

these three features, the average iPhone classification accuracy increased from 80.2 ± 4.3 to $84.1 \pm 1.5\%$. For the Littmann recordings, the test accuracy went up to $86.7 \pm 2.8\%$ from $82.8 \pm 4.6\%$. This illustrates clearly that over-fitting is a cause for sub-optimal SVM-based classification.

Interestingly, the Littmann classification was more accurate than the iPhone recordings when using the LASSO LR model, even though the level of agreement between the annotators was higher for the iPhone recordings. This is counter-intuitive: if the level of agreement between annotators is used as an indicator of the ease with which they could classify the recordings, then these human annotators found it easier to classify the iPhone recordings, while the automated algorithms had better classification performance on the Littmann recordings. Reasons for annotator disagreement for the Littmann recordings are discussed above, and these are based on annotator subjectivity. This subjectivity (after finding “gold-standard” annotations from the mean of five annotator’s ratings) is negated when using automated algorithms, which could explain the high accuracy even with high levels of disagreement between annotators.

3.5 Conclusion

This chapter has outlined the development of automatic signal quality classification algorithms for PCG recordings from two different devices. The main contributions and findings of this chapter were:

- Based on an investigation into the annotations assigned to the iPhone mobile phone-based PCG recordings, we can conclude that such a device is capable of recording diagnosable-quality recordings.
- Recordings from a Littmann electronic stethoscope are of poor-quality in a third of cases when performed by non-medically trained healthcare worker. Therefore, such a device also requires automated signal quality assessment.
- Nine signal quality indices were investigated after a review of the literature, many of them based on the autocorrelation function.
- The $mSQI$ and $kSQI$ features were the most influential in the LASSO LR model, regardless of recording device. $seSQI$ was favoured by the iPhone classification algorithms,

while *bSQI* had a large β coefficient when classifying Littmann recordings using the LASSO LR.

- The SVM classification did not perform as well as the LASSO LR model, due to over-fitting even when using LTO cross-validation. When using the SVM classifier, *seSQI* was the most-selected feature for the iPhone classification, whereas *kSQI* and *bSQI* were most often selected for the Littmann classification. When limiting the number of features for the SVM-based classification, average accuracy improved for both the iPhone and Littmann devices, indicating that over-fitting had been a reason for poorer SVM performance in comparison to the LASSO LR model.
- These classification algorithms are able to classify the signal quality of mobile phone-based PCGs with a sensitivity of $77.8 \pm 1.9\%$ and a specificity of $86.7 \pm 2.6\%$, while classifying the signal quality of electronic stethoscope PCGs with a sensitivity and specificity of $81.9 \pm 3.2\%$ and $91.2 \pm 1.8\%$. Two papers in the literature ([121, 163]) report sensitivity and specificity values of approximately 96% when performing binary classification of the signal quality of PCGs. However, their methods were optimised without a held-out test dataset on which to be assessed.

The implementation of an automatic signal quality classifier for heart sounds is essential when performing recordings using a low-cost mobile phone device. A fully automated system is even more important when such recordings are made by a non-medically trained healthcare worker. A clinician or cardiologist would be able to perform a heart sound recording and listen back to the recording to evaluate its quality. However, a non-medically trained healthcare worker, with minimal experience of listening to heart sounds, would not be able to judge the signal quality of such recordings. The algorithms developed in this chapter are able to give feedback to the user of a mobile phone-based stethoscope, indicating the signal quality of a recording, and whether it is fit for further processing. Otherwise, the recording would have to be redone.

This chapter addresses the first step in the analysis of heart sound recordings, as shown in Figure 1.14. After a good-quality signal has been acquired, further processing (in the form of segmentation and murmur classification) can be performed, which are described in the chapters to follow.

Chapter 4

Probabilistic model-based phonocardiogram segmentation

“I don’t do graphical models. The cliques, DAGs, message passing, expectation-maximisation, and belief propagation remind me of high school.” - @ML_Hipster

The previous chapter has described methods for the automatic classification of the quality of PCGs, so that those that are of low quality are re-recorded before further processing. The next step in the automated analysis of these recordings is the detection of the exact positions of the S_1 and S_2 sounds, or *heart sound segmentation*, as depicted in the flow diagram in Figure 1.14.

This chapter addresses the segmentation of the S_1 and S_2 sounds, or fundamental heart sounds (FHSs), within noisy, real-world recordings from healthy and pathological patients using probabilistic models. This is performed without the use of a reference signal, such as an ECG. The accurate localisation of the FHSs is a prerequisite for the identification of the systolic or diastolic regions of a PCG, allowing the subsequent recognition of pathological murmurs in these regions [79]. Importantly, the process of segmentation aims to ignore any pathological sounds at this stage, in order to find the exact positions of S_1 and S_2 . The detection of pathological sounds is then performed subsequently to segmentation.

While the segmentation of heart sounds is relatively simple in noise-free recordings, it becomes a difficult task when the recordings are corrupted by in-band noise. Common noise sources include endogenous or ambient speech, motion artefacts and physiological sounds such as

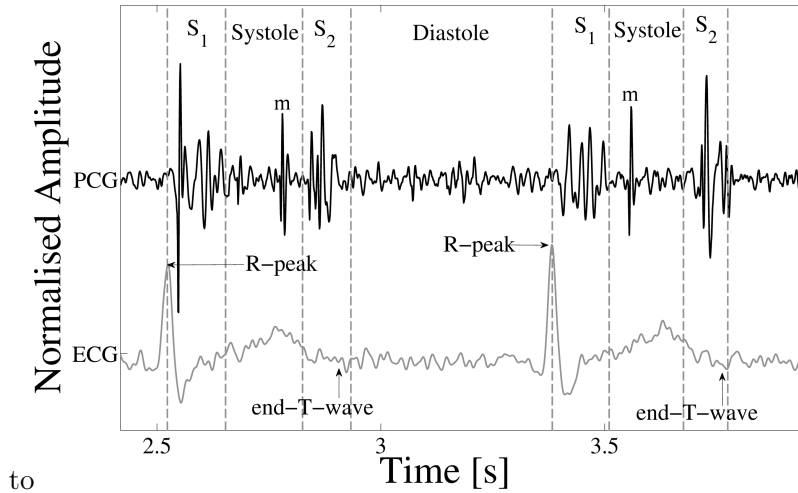


Figure 4.1 – Example of an ECG-labelled PCG, with the ECG, PCG and four states of the heart cycle (S_1 , systole, S_2 and diastole) shown. A mid-systolic click can be seen in each of the heart cycles, labelled as “m”. The R-peak and end-T-wave are labelled as references for defining the approximate positions of S_1 and S_2 respectively.

intestinal and breathing sounds. Other physiological sounds of interest, such as murmurs, clicks, splitting of the FHSs and additional S_3 and S_4 sounds, can also complicate the identification of the FHSs. For example, Figure 4.1 shows a PCG with a loud mid-systolic click. This figure shows the timing relationship between the PCG and ECG; S_1 commences with the R-peak of the ECG, while S_2 occurs at approximately the end of the T-wave. More information on the components of the ECG can be found in Section 1.3.3.2.

This chapter initially presents an extensive review of the heart sound segmentation literature, and the state-of-the-art method is identified. Secondly, four R-peak detectors and four end-T-wave detectors are used to find these points in ECGs recorded synchronously with PCGs. The R-peaks and end-T-waves provide *gold-standard* reference positions for the S_1 and S_2 sounds. Therefore, noisy ECGs with unreliable R-peak and end-T-wave annotations are identified at this point. The gold-standard reference heart sound positions are used as markers to train a modified hidden Markov model (HMM) and a conditional random field (CRF), after feature extraction. Lastly, the effectiveness of the modified HMM and CRF for identifying the S_1 and S_2 sounds are compared to that of the state-of-the-art algorithm.

The principal contributions of this chapter to the field of heart sound segmentation are:

- An exploration of features for heart sound segmentation, including a robust selection of the wavelet family and decomposition level when using the discrete wavelet transform.

- An enhanced hidden semi-Markov model, which includes duration dependencies and logistic regression-based emission probabilities.
- The use of a conditional random field for heart sound segmentation, also extended to include duration dependencies.
- The implementation of an extension to the Viterbi algorithm, for use with duration dependent probabilistic models such as hidden semi-Markov models (HSMs) and duration dependent CRFs.
- A rigorous evaluation of the proposed segmentation approaches using one of the largest published datasets, and a comparison to the state-of-the-art segmentation algorithm.

4.1 Background

Table 4.1 summarises the relevant background literature. The table presents the size of the datasets used, numerical results from the studies, and important notes, such as whether any independent test set was used to evaluate the algorithm under consideration.

Many methods of heart sound segmentation use an amplitude threshold after various transformations of the PCG signal, and many researchers have adopted this approach, using a variety of features. The normalised, average Shannon energy^a was proposed for PCG envelope extraction by Liang et al. [175], which it was claimed emphasised medium-intensity sounds, while attenuating low-intensity signals. While this method achieved 93% accuracy in a dataset of 37 recordings acquired from children with murmurs, it has been criticised for poor performance in noisy recordings and in the presence of murmurs [183]. A second iteration of this method involved wavelet decomposition before estimation of the Shannon envelope, which improved the segmentation results from 84% to 93% accuracy on a noisy dataset [137].

Similarly, Kumar et al. [131] computed the Shannon envelope after wavelet decomposition, then employed a threshold-based technique that was optimised on their entire dataset to achieve 98% sensitivity and specificity on this (in-sample) dataset.

Table 4.1 – Summary of previous segmentation work. *Ac* denotes accuracy, *Se* sensitivity, *Sp* specificity and *P₊* positive predictivity

^aThe Shannon energy of a signal is defined by $E_s = -x^2 \log x^2$ [175].

Authors	Dataset	Reported Metrics and Results	Notes
[175]	37 recordings (515 cycles) from children with murmurs (14 being pathological)	93.0% Ac	Unsupervised. Optimised on entire dataset
[137]	77 (1165 cycles) recordings from children with both pathological and physiological murmurs	94.6% Ac	Unsupervised. Optimised on entire dataset
[184]	One simulated and four teaching-quality PCGs	No quantitative results	
[131]	55 recordings (7530 cycles), 51 with valve replacements	97.95% Se , 98.2% Sp	Unsupervised. Optimised on entire dataset
[185]	71 recordings (357 cycles), 9 different pathologies	97.47% Ac	No split between train, test sets
[186]	166 teaching-quality heart cycles from normal and pathological patients	84.0% Ac	Unsupervised. No stated segmentation tolerance
[187]	500 teaching-quality heart cycles (unknown numbers of healthy/pathological)	100% (normal) and 88.2% (abnormal) Ac	No split between train, test sets. No stated segmentation tolerance
[188]	42 adult patients with bioprosthetic valves (21 normal), 13 seconds per patient	90.5% Ac	Unsupervised. No stated segmentation tolerance
[125]	41 recordings (340 cycles). Mix of normal (32%), systolic (36%) and diastolic murmurs (32%)	90.29% Ac	Unsupervised
[38]	27 recordings of 30 s (997 cycles) from healthy subjects	92.1% Se , 88.4% P_+	Unsupervised
[124]	30 clean recordings (20 s) from healthy subjects	96.2% Ac	No split between train, test sets
[189]	120 recordings from children, 80 with congenital heart disease (totalling 1200 s, 823 cycles in test set)	93.6% Ac on test set	50% train-test split
[190]	9 recordings (less than 5s). 55% pathological	99.0% Ac on whole cycle detection	No split between train, test sets
[191]	9426.8 s of recordings, normal (22.2%) and various pathologies (ASD, PDA, VSD and RHD)	S_1 : 98.53% Ac , S_2 : 98.31% Ac , Cycles: 97.37% Ac	Unsupervised. No stated segmentation tolerance.
[192]	80 recordings from an unknown number of patients of 6-12 s (40 healthy, 40 pathological recordings)	96% and 97% Se , 95% and 95% P_+ (healthy and pathological)	Unsupervised algorithm. No stated segmentation tolerance
[193]	26 teaching-quality recordings (565 cycles), 3 healthy subjects and 23 with various pathologies	94.9% and 95.9% Ac (S_1 and S_2)	No split between train, test sets and no stated segmentation tolerance
[194]	50 two-minute healthy and pathological recordings	99.0% Se and 98.6% P_+	No split between train, test sets. Results reported on 20% of dataset
[195]	64 teaching-quality recordings of less than 10 s (701 cycles). Various pathologies	93.06% Ac , 99.43% Se , 93.56% P_+	No split between train, test sets. Results reported on portion of dataset. No stated segmentation tolerance.
[196]	52 recordings (14 controls, 38 with murmurs), 43 in test set (2602 cycles)	94.56 ± 6.58 $G - measure$	Ac denoted for correctly segmented cycles. $G - measure$ is geometric mean of Se and P_+
[123]	80 patients, 8 pathological. Recordings of 20 s from four auscultation sites (10045 S_1 , 9818 S_2 sounds)	S_1 : 94.6% Se and 97.7% P_+ S_2 : 95.2% Se and 96.1% P_+	No split between train, test sets
[129]	46 clean recordings from 8 patients (2286 s). No pathologies mentioned	97.6% Ac	Ac computed from average of 8 fold cross-validation

[128]	17 patients, 44 recordings (30-60 s). No pathologies mentioned	S_1 : 98.6% Se and 96.9% P_+ S_2 : 98.3% Se and 96.5% P_+	Results computed from average of 4 fold cross-validation
[99]	113 recordings of 8 s, 8% with coronary artery disease	98.8% Se , 98.6% P_+ on test set	73 test, 40 training recordings

Nigam and Priemer [184] found their method of employing complexity-based features for PCG segmentation subjectively outperformed the Shannon energy approach when using synthetic data, finding that the Shannon energy was not able to detect murmurs due to their low amplitude and overlap with S_1 . However, their method was found to be ineffective on real-world data [186]. Ari et al. [185] found that the use of a simple squared-energy envelope outperformed the Shannon envelope in normal and pathological cases when employing a threshold-based detection method, achieving 97.5% accuracy on their dataset of 71 recordings.

Vepa et al. [186] built on previous work by including wavelet-derived complexity-based features, and incorporated the beat detection scheme employed by Ari et al. [185], yet had poor results (84% accuracy) on teaching-quality recordings. Choi and Jiang [187] compared a method they devised for envelope detection to the two popular methods mentioned above, the Shannon Energy and the Hilbert transform. Their method involved modelling the heart sounds as the input to a spring-damper system, with the *cardiac sound characteristic waveform* (CSCW) being the output from such a system [197]. They found their method to be superior to the Shannon envelope and Hilbert transform methods, concluding that their method was less susceptible to noise and provided more accurate segmentation results (100% and 88.2% accuracy on normal and pathological patients, as compared to 78.2% and 89.4% for the Shannon envelope and 51.4% and 47.3% for the Hilbert transform).

Instead of using a threshold-based technique, Rajan et al. [188] developed an unsupervised segmentation method based on Morlet wavelet decomposition and the comparison of the amplitude of the eigenvalues after singular value decomposition, on a dataset of 42 adult patients with bioprosthetic valves. Chen et al. [38] analysed a portion of the dataset used in this chapter, and adapted ECG analysis methods based on the use of a threshold and k -means clustering to identify the heart sounds, achieving 92.1% sensitivity and 88.4% positive predictivity. Similarly, k -means clustering of the systolic and diastolic durations was employed

by Gupta et al. [125] to identify the S_1 and S_2 sounds, after identifying the peaks in the homomorphic envelope with similar results (90.3% accuracy).

Oskiper and Watrous [124] used a neural network to segment the fundamental heart sounds in a database of 30 recordings from healthy subjects, using the continuous Morlet wavelet decomposition to achieve 96.2% accuracy. However, these researchers incorporated amalgamated training and test data set when evaluating their method. Sepehri et al. [189] also used a neural network classifier. However, these researchers argued against using wavelet decomposition as it generated large, inefficient feature sets [198]. They based their features on autoregression-derived frequency bands and the normalised cross-correlation between other peaks in the signal. These researchers reported segmentation accuracy of 93.6% on a separate test set (half of a dataset of 120 recordings).

Yan et al. [190] derived a characteristic moment waveform, based on the variance of the PCG over different time-scales, and showed promising results (99% accuracy) on a small dataset (of nine recordings), and without any train-test split. Sun et al. [191] used the same envelope as Yan et al. [190], but then from this derived a modified Hilbert transform-based envelope which resulted in 97% accuracy. Moukadem et al. [192] found the Shannon envelope after applying time-frequency decomposition to locate the FHSs while Tang et al. [193] employed dynamic clustering after atom decomposition, but neither stated a localisation tolerance. Naseri and Homaeinezhad [194] employed a combined frequency-energy feature, while Varghees and Ramachandran [195] employed the phase from the analytical signal after finding the Shannon entropy, but both optimised their approach on their entire dataset and reported results on a small portion of this. The results from Papdaniil et al. [196] outperformed many of these methods using empirical mode decomposition and kurtosis features to select non-Gaussian intrinsic mode functions (IMFs), and detected the start and end positions of heart sounds within the selected IMFs. This method achieved 95% G – *measure*, or the geometric mean between sensitivity and positive predictivity.

The introduction of generative probabilistic models for heart sound segmentation led to improved accuracy. Gamero and Watrous [123] applied two separate hidden Markov models (HMMs), trained using mel-frequency cepstral coefficients, to a large dataset of mostly healthy patients with notable success ($>95\%$ sensitivity and positive predictivity). Ricke et al. [129]

used embedded HMMs within each state, along with mel-frequency cepstral coefficients, Shannon energy and regression coefficients on a small dataset. They found that the cepstral coefficients gave poor results, but achieved 97.6% segmentation accuracy on 46 clean recordings using the Shannon energy features alone. Gill et al. [128] were the first researchers to incorporate timing durations within HMMs for heart sound segmentation. These researchers included the time between peaks in the homomorphic envelope as input features in their unsupervised HMM. Their method yielded high segmentation accuracy (98% sensitivity and 96.5% positive predictivity) on a small dataset of 44 recordings from 17 healthy subjects.

Schmidt et al. [99] were the first researchers to explicitly model the expected duration of heart sounds within the HMM using a hidden semi-Markov model. These researchers hand-labelled the positions of the S_1 and S_2 sounds in their dataset of 113 noisy, real-world recordings, then used the average duration of these sounds and autocorrelation analysis of systolic and diastolic durations to derive Gaussian distributions for the expected duration of each of these four states (S_1 , systole, S_2 and diastole). The features analysed were the homomorphic envelope and three frequency band features (25-50, 50-100 and 100 - 150 Hz), that, along with the hand-labelled positions of the states, were used to derive Gaussian distribution-based emission probabilities for the HMM. The duration distributions were then incorporated into the forward and backward paths of the Viterbi algorithm^b. The homomorphic envelope alone was found to yield the best results on the separate test set, with 98.8% sensitivity and 98.6% positive predictivity. The rigorous assessment of this method on a noisy dataset established this method as the current state-of-the-art segmentation method. However, limitations of this work include a small dataset of only 113, eight-second recordings, the investigation of a limited number of input features and the use of the standard Viterbi algorithm, that can introduce segmentation errors in short recordings when employing an HSMM.

The work in this chapter builds on the work of these researchers by using a substantially larger dataset, the assessment of additional input features, an extension of the Viterbi algorithm, more discriminative emission probability derivation using logistic regression (LR) within the HSMM as well as the assessment of a duration-dependent conditional random field for heart sound segmentation.

^bSee Section 4.2.5.1 for a description of the Viterbi algorithm.

4.2 Methods

The flow diagram in Figure 4.2 illustrates the seven steps employed in the evaluation of the heart sound segmentation methods.

The first step involves the detection of the *gold-standard* positions of the heart sounds from the ECG, which are used to train and evaluate the heart sound segmentation algorithms (while not being used to detect the heart sounds themselves). These are the R-peak and end-T-wave positions, which are also used to determine which ECGs are of high-quality (Section 4.2.2).

The second step involves the labelling of the positions of the heart sounds based on the R-peak and end-T-wave positions (Section 4.2.4). The third step involves the splitting of the dataset into training and test sets, which is outlined in Section 4.2.9.1. The fourth and fifth steps, involving the extraction of the features used for segmentation, are discussed in Section 4.2.8.

The sixth step involves the training of the segmentation methods. The methods evaluated in this chapter are outlined in Sections 4.2.5, 4.2.6 and 4.2.7, while the training procedure for these methods is described in Section 4.2.9.2.

Lastly, the evaluation of the methods is performed on the separate train and test sets. The evaluation procedure is described in Section 4.2.9.3.

4.2.1 Database

The dataset employed for the evaluation of the various heart sound segmentation methods was the MAAS dataset, described in Section 2.4.

The study database consisted of 405 synchronous 30-40 s PCG and ECG recordings from 123 de-identified adult patients (roughly three recordings per patient) attending the Massachusetts General Hospital for cardiac screening as well as in-home recordings of people suffering from mitral valve prolapse (MVP). Many of these PCG recordings were corrupted by various sources of noise, including stethoscope motion, breathing, intestinal sounds and talking. Comparing the size of this database to others in Table 4.1, it can be seen that this dataset represents one of the largest published in terms of patient numbers and recording length.

Of the 123 patients, 38 were normal controls, 37 had murmurs relating to MVP, 36 had benign murmurs, five had aortic disease and seven had other miscellaneous conditions (tricuspid

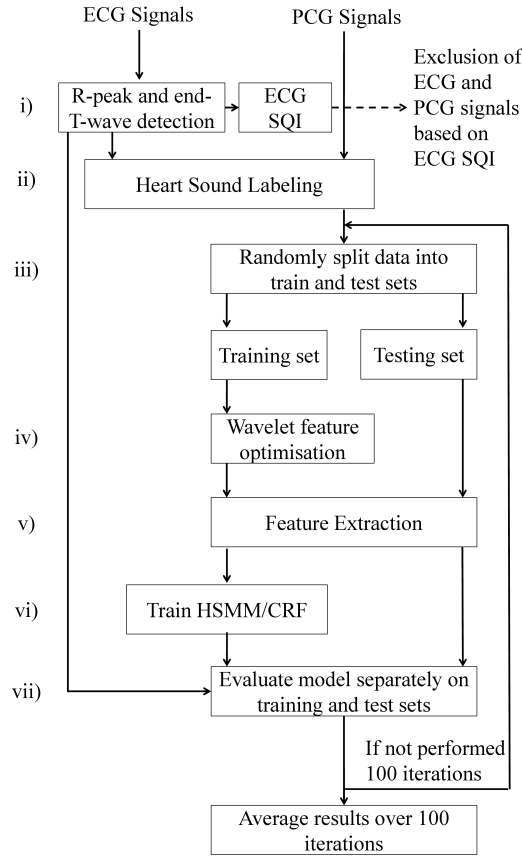


Figure 4.2 – Block diagram of the steps employed in the evaluation of the segmentation algorithms. Step *i* involves the detection of the R-peaks and end-T-waves in the ECG, and the derivation of the ECG signal quality index (SQI). These points are used to label the positions of the heart sounds (*ii*). Feature extraction (*iv* and *v*) is performed after the train and test split (*iii*). Training (*iv*) and evaluation (*vii*) of the segmentation methods are then performed.

regurgitation, endocarditis, asymmetric septal hypertrophy). Recordings were made from the parasternal, apical, aortic and pulmonic auscultation positions, with patients sitting forward, squatting or in the supine position. All diagnoses were verified by a single clinical expert through echocardiographic examination.

4.2.2 Data exclusion based on ECG signal quality

This section refers to step *i* in Figure 4.2. The gold-standard reference positions for the FHSs in the PCG were derived from the synchronous ECG recordings. The R-peak and the end-T-wave in the ECG correspond to the positions of the S1 and S2 sounds in the PCG respectively [89] (see Figure 4.1). The R-peaks and the end-T-waves form the reference positions for PCG-labelling in Section 4.2.4 and evaluation of the segmentation algorithms in Section 4.3 (as shown in steps *ii* and *vii* in Figure 4.2). Therefore, it was essential to ensure the correct detection of the

R-peak and the end-T-wave in the ECG. The correct positions of these markers in the ECG were found by comparing the agreement between four R-peak and four end-T-wave detectors, as described below.

Intuitively, the presence of noise and artefacts in the ECG will lead to disagreement between semi-independent detectors for both the R-peak and end-T-wave. Therefore, in order to ensure high-quality periods of ECG signal and accurate localisation of the R-peak and end-T-wave, the agreement between the R-peak and end-T-wave detectors was assessed to derive an ECG signal quality index (SQI), as proposed by [153, 161, 162].

Four R-peak and end-T-wave detectors were employed in this chapter. The choice of four was dictated by the requirement of three detectors to be within a specified tolerance of each other, while the fourth, worst performing detector on a particular record, was discarded. Therefore, this choice allowed one of the detectors to perform poorly on a particular record, while the requirement of agreement within a tolerance of the remaining three remaining detectors ensures robust R-peak and end-T-wave detection. A key factor in the success of robust R-peak and end-T-wave detection in such a detection scheme is that different ECG algorithms are sensitive to different types of noise [199]. Therefore, the agreement between different detectors in noisy signals allows for more robust R-peak and end-T-wave detection than a single detector. Further, the agreement between three detectors, rather than two, allows for even more robust detection.

The four R-peak detectors employed in this study were:

1. *Wavelet QRS* – a wavelet-based ECG delineator [200], which employs a modified version of the wavelet-based R-peak detection scheme originally proposed in [201]. This method decomposes the ECG signal using the discrete wavelet transform with a quadratic spline wavelet over four scales. Then, “maximum modulus lines” which exceed thresholds are located, starting in the higher scales, to identify R-peaks. Rules are used to discard isolated and redundant R-peaks.
2. *Zhang QRS* – an algorithm based on parabolic fitting, where a 2nd degree polynomial is fitted to ECG segments within a sliding window. The product of the two polynomial coefficients then produces an indicator function, whose maxima correspond to R-peaks which are found using thresholding [202].

3. “jQRS” – a method (used in two previous studies [203, 204]) based on the method of Pan and Tompkins [205]. This method employs a band-pass filter, differentiation, squaring and moving window integration to accentuate R-peaks, then thresholding (with a refractory period of 250 ms) is used to identify R-peaks.
4. “gQRS” – (available on Physionet [206]), which is not yet published.

The four end-T-wave detectors employed in this study were:

1. *Wavelet end-T-wave* – the same wavelet-based delineator mentioned previously [200], which after detecting the R-peaks, defines a search window relative to each R-peak and the computed beat duration. Within this window, the positions of the significant gradients (above a threshold) of the wavelet transform are found, with the zero crossing between these positions identifying the T-wave peak. The end-T-wave position is located by finding the point where the absolute wavelet transform falls below a threshold, or a local minimum in the absolute wavelet transform after the T-wave, whichever is closer to the T-wave peak position.
2. “ecgpuwave” [207] – a method (available on Physionet [206]) commonly referred to as the low-pass differentiation (LPD) method. The ECG signal is first differentiated and low-pass filtered. The position of the T-wave is located in this signal by finding the maximum and minimum value in a window between R-peaks (the size and position of which are dictated by the heart rate). If these maximum and minimum values are greater than 2% of the maximum slope of the QRS complex, then a T-wave is located, the peak of which is found at the zero crossing between the maximum and minimum points. To the right of the zero-crossing position, the peak in the differentiated signal, pk , is identified, indicating the position of maximum slope in the second half of the T-wave. The end-T-wave position is located at the point where the differentiated signal crosses a threshold based on the value at pk .
3. *Zhang end-T-wave* – a method based on maximising the area in a sliding window between successive R-peaks. The position in the ECG that maximises the area of the sliding window is deemed to identify the location of the end-T-wave [208].

4. *Trapezium end-T-wave* – a method based on maximising the area of a trapezium fixed at points within the ECG. Between each pair of R-peaks, the position of maximum absolute gradient after the T-wave peak is found, as well as a point on the T-P isoelectric segment past the end of the T-wave. Then, a third point, between the previous two, which maximises the area of the trapezium generated by the three points, indicates the position of the end-T-wave [209].

The wavelet-based ECG delineator was selected as it is one of the most cited methods in the literature [200]. In the original wavelet-based delineator publication, the authors compare their end-T-wave localisation method directly to “ecgpuwave” [207], which achieves similar end-T-wave detection accuracy. For this reason, “ecgpuwave” was also included. The R-peak and end-T-wave detector developed by Zhang et al. [202, 208] were selected, as these methods were found to outperform the wavelet-based delineator and ecgpuwave [208]. Finally, jQRS [203, 204] was selected as this provides a relatively simple detector, based on the original Pan and Tompkins algorithm [205] (the most cited R-peak detector in the literature). The simple end-T-wave detector of Vazquez-Seisdedos et al. [209] was included, as it was found to outperform the “ecgpuwave” method by the original authors. Lastly, gQRS was included as it performed well on preliminary tests, while being available as open-source code.

The performance of R-peak detectors is usually assessed by beat-to-beat comparisons between the detected beats and the reference beats. The standard adult tolerance window for candidate R-peaks is 150 milliseconds (ms) [176]. However, this is longer than the expected duration of a heart sound [99] and was therefore deemed to be too wide a tolerance in this case. In this study, in order to acquire a more robust estimate of the correct location of peaks, if R-peak detectors agreed to within 100 ms, they were said to have agreed on the position of the R-peak. The same tolerance was followed for the end-T-wave positions.

The process for R-peak detection was:

- Band pass filter the ECG signal between 0.7 and 50 Hz using zero-phase, forward-backward second-order Butterworth filters to ensure no baseline wander and the exclusion of high-frequency noise.
- Measure the agreement between all pairs of R-peak detectors, where agreement is measured as per the “bxb” algorithm, available from Physionet [206], which evaluates the

agreement between two sets of annotations within a specified tolerance. The agreement was measured using the F_1 score (see Section 4.2.9.3). Then, select the three R-peak detectors that had the highest product of their F_1 scores, thereby excluding the R-peak detector that had the lowest overall agreement with the other detectors over the entire record. This allows the exclusion of a detector that may perform poorly on a particular record, while keeping three independent R-peak detectors.

- Over four-second windows, the F_1 score of agreement between the chosen three detectors was taken to be the ECG SQI. The choice of a four-second window was based on the need for at least two heart cycles, even in the case of low heart rates. Two heart cycles are needed, as many of the end-T-wave detectors employed subsequently rely on the detection of at least two R-peaks. Limiting the analysis to two cycles results in short-duration analysis windows, allowing high resolution when discarding poor-quality ECG segments, and the inclusion of more ECG data that is of high quality as compared to larger analysis windows. The evaluation of ECG signal quality over windows by comparing their R-peak agreement is the same as methods employed by other authors, but with a smaller window size [153, 162]. After SQI computation, the window was shifted by one second, leading to second-by-second SQI values centred on the successive windows.
- In windows of 100% F_1 score, or complete agreement between the three peak detectors within the 100 ms tolerance, the adjudicated position of the R-peaks was defined to be the R-peak position from the three detectors which corresponded to the maximum absolute value in the ECG. Windows of less than 100% F_1 score were not used for further analysis, as they were deemed to have untrustworthy annotations due to the three R-peak detectors not having complete agreement. Although this discards some potentially good-quality ECG, it is a robust approach.

An example of R-peak detection is shown in Figure 4.3. In this case, the gQRS was excluded as being the R-peak detector with the lowest level of agreement with the other three R-peak detectors. The adjudicated R-peak is shown, which corresponds to the *Wavelet QRS* position, as it was the detected R-peak with the largest absolute amplitude.

The process for end-T-wave detection was:

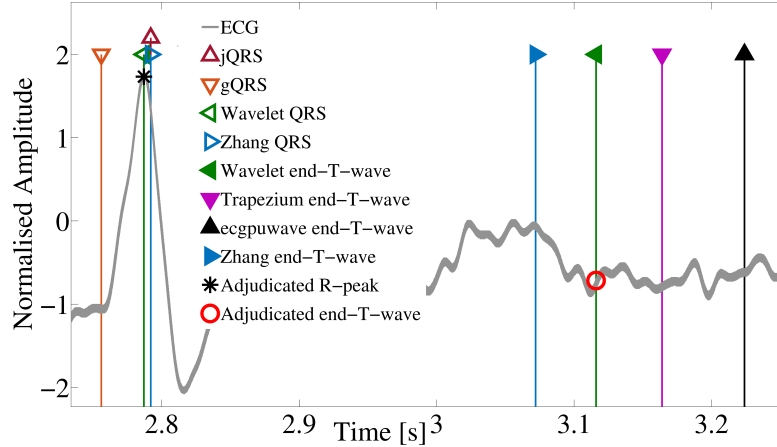


Figure 4.3 – Example of the accurate R-peak and end-T-wave positions derived in a noisy ECG signal, based on the agreement between R-peak and end-T-wave annotations. The positions of the four R-peak detections and four end-T-wave detections are shown. The algorithm adjudicated R-peak and end-T-wave, based on the four detections, is shown.

- In windows of 100% F_1 score based on the R-peak detections, find the end-T-wave positions from the four detectors.
- Exclude the annotation furthest from the median of these four annotations, allowing the exclusion of an outlier annotation. If only three end-T-wave detections were present, due to a missed detection, no exclusion was performed. If fewer than three annotations were present, a missed end-T-wave was marked and the ECG SQI was set to zero.
- If the remaining three annotations were all within the 100 ms tolerance of each other, the adjudicated end-T-wave position was marked as the median positions of the three annotations. Otherwise, the ECG SQI was set to zero to indicate an untrustworthy portion of the signal.

An example of end-T-wave detection is shown in Figure 4.3. The *ecgpuwave* annotation can be seen to be an outlier compared to the other three end-T-wave detectors, and therefore was the excluded annotation. Then, the adjudicated end-T-wave position can be seen to coincide with the *Wavelet end-T-wave* annotation.

ECG signals (and the corresponding PCG signals) of at least two heart cycles with 100% SQI were retained for further analysis, while periods of ECG with SQI values less than 100%, based on the lack of agreement between R-peak and end-T-wave annotations, were excluded as their annotations were possibly erroneous. Two cycles of continuous high-quality ECG were selected as this would give the hidden semi-Markov models and CRFs employed in this study sufficient

Table 4.2 – The selection and exclusion rates of the different R-peak and end-T-wave detectors, shown as percentages of all the detected R-peaks and end-T-waves.

(a) – Selection and exclusion rates for R-peak detectors		
R-peak	Selected (%)	Excluded (%)
jQRS	27.9	2.2
gQRS	3.8	5.4
WaveletQRS	36.9	55.4
ZhangQRS	31.3	36.9
(b) – Selection and exclusion rates for end-T-wave detectors		
end-T-wave	Selected (%)	Excluded (%)
Wavelet	32.1	11.8
Trapezium	6.6	35.2
ecgpuwave	17.2	44.6
Zhang	44.1	8.4

data to assign state labels accurately. Such segments of data would consist of at least one complete systolic and diastolic period, the durations of which help to differentiate between the S_1 and S_2 sounds, as systole is expected to be shorter than diastole (see Section 4.2.3).

The selection and exclusion rates of the different R-peak and end-T-wave detectors can be seen in Table 4.2. This table illustrates that all four detectors for both the R-peak and end-T-wave contribute to the selection of these points. While gQRS and the Trapezium end-T-wave detectors are selected the least, both of these detectors can be seen to have low exclusion percentages. In addition, the WaveletQRS and ecgpuwave detectors are excluded the most. However, they are also the selected detectors for a large proportion of the R-peaks and end-T-waves.

The remaining data consisted of 10171.85 seconds of PCG, including 12181 R-peaks and 11627 end-T-waves^c (corresponding to the same number of S1 and S2 sounds), from 112 patients. It is important to note that the ECG signals (and the derived R-peak and end-T-wave locations) were only used for heart sound labelling and evaluation of the segmentation models. They were not used in the heart sound segmentation process itself. Furthermore, ECG signal quality has little to no impact on PCG signal quality, except in the case of motion artefact, which is limited

^cThe number of R-peaks and end-T-waves are not equal due to a number of heart cycles beginning and ending with R-peaks.

in such a dataset of recordings performed on adults. Therefore, selection of high-quality ECG does not bias this study towards the selection of high-quality PCG recordings.

4.2.3 Heart sound duration distributions

A key component of the models employed to segment the PCG is an estimate of the probability density function of the expected duration for each state. In the case of heart sound analysis when using a four-state HSMM or CRF, a state is defined to be each of the four major components of a heart cycle. These are: 1) the S_1 sound, 2) the systolic period between S_1 and the S_2 sound, 3) the S_2 sound and 4) the diastolic period between S_2 and S_1 (see Figure 4.1). The duration of each of these components was modelled as described by Schmidt et al. [99], who expressed the duration of each state as a Gaussian distribution derived from their own annotated dataset.

The durations of the S_1 and S_2 sounds, D_{S_1} and D_{S_2} , were defined as:

$$D_{S_1} = \mathcal{N}(\bar{S}_1, \sigma_{S_1}^2), \bar{S}_1 = 122, \sigma_{S_1} = 22 \quad (4.1)$$

$$D_{S_2} = \mathcal{N}(\bar{S}_2, \sigma_{S_2}^2), \bar{S}_2 = 92, \sigma_{S_2} = 22 \quad (4.2)$$

where $\bar{S}_i$ and σ_{S_i} are the mean and standard deviation of the duration of sound S_i in ms. The mean duration of a heart cycle, $\overline{D_{cycle}}$ was estimated from the reference heart rate. The reference heart rate for each PCG was calculated by finding the median time between the R-peak locations computed in Section 4.2.2^d. Then, the mean total systolic duration including the S_1 sound, $\overline{D_{sysTotal}}$, was estimated by finding the largest peak in the autocorrelation of the homomorphic envelope (see Section 4.2.8.1) between 200 ms from the start and half of the heart cycle duration [99].

Using the above values, the durations of the systolic, D_{sys} , and diastolic, D_{dias} , states can be estimated. The mean duration of systole, without the S_1 sound, can be estimated using:

$$\overline{D_{sys}} = \overline{D_{sysTotal}} - \bar{S}_1 \quad (4.3)$$

^dAs this was an experiment in segmentation, the ECG was used to compute an accurate heart rate estimate. Using the PCG alone may lead to an inaccurate estimate, and hence, poor segmentation results.

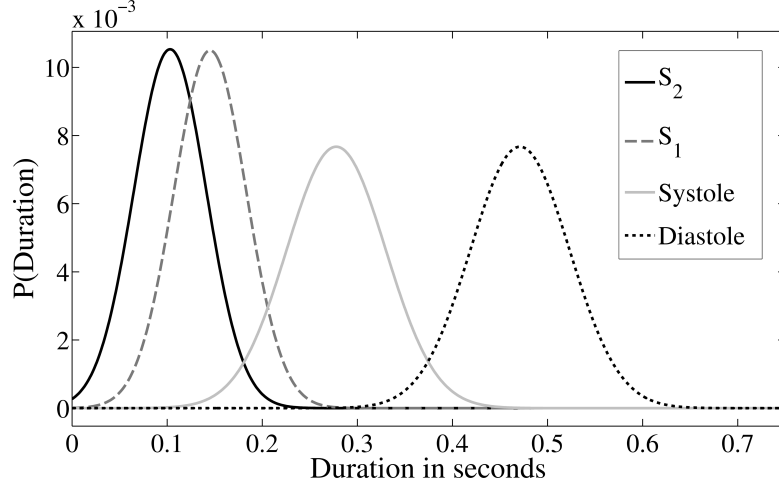


Figure 4.4 – Heart sound duration distributions, showing the expected duration of S_1 , S_2 , systole and diastole. The duration of a heart cycle is computed from the heart rate, which is 60 beats per minute in this figure. As expected, S_2 has the shortest duration followed by S_1 (both of which remain constant regardless of heart rate). The duration of systole and diastole depend on the heart rate, while systole is generally shorter than diastole, except at high heart rates.

while σ_{sys} was estimated to be 25 ms [99]. The mean duration of diastole, $\overline{D_{dias}}$, can be computed using

$$\overline{D_{dias}} = (\overline{D_{cycle}} - \overline{D_{sysTotal}}) - \bar{S}_2 \quad (4.4)$$

while Schmidt et al. estimated σ_{dias} to be $\sigma_{dias} = 0.07\overline{D_{dias}} + 6$. Therefore, the expected duration of systole and diastole can be written as:

$$D_{sys} = \mathcal{N}(\overline{D_{sys}}, \sigma_{sys}^2) \quad (4.5)$$

$$D_{dias} = \mathcal{N}(\overline{D_{dias}}, \sigma_{dias}^2) \quad (4.6)$$

An example of the duration of the different states is shown in Figure 4.4. The expected durations of S_1 and S_2 remain constant, regardless of heart rate, while the expected durations of systole and diastole are heart rate dependent, with diastole varying more than systole. The minimum duration of the S_1 and S_2 sounds were set to 0.02 s (one sample at the 50 Hz sampling frequency), while the maximum duration of any state was set to the duration of the entire heart cycle, derived from the reference heart rate. As will be shown in Algorithm 4.2, this was to ensure tractability of the HSMM.

4.2.4 ECG-derived heart sound labelling

The S_1 and S_2 sounds within the PCG recordings were labelled using the reference R-peak and end-T-wave positions, computed from the synchronous ECG signals (Section 4.2.2 and step *ii* in Figure 4.2).

Since the start of S_1 coincides with the R-peak of the ECG [89], the period of $\overline{S_1}$ duration from the position of each R-peak in the PCG was labelled as S_1 .

The S_2 sound occurs at *approximately* the end-T-wave in the ECG [89]. Therefore, labelling based on only the end-T-wave position would be erroneous. Therefore, the centre of the S_2 sound was found by searching for the maximum peak in the Hilbert envelope (see Section 4.2.8) of the PCG signal within a specified search window in the vicinity of the end-T-wave. This search window was set to be the position of the end-T-wave, plus and minus the longest expected duration of S_2 ($\overline{S_2} \pm \sigma_{S_2}$, found from [99]). The position of the maximum value of the Hilbert envelope within this search window was marked as the centre of the S_2 sound. An interval equal to the length of the expected S_2 sound, centred on this maximum position, was labelled as the S_2 sound. The period between the S_1 and S_2 sounds was labelled as systole, while those periods between S_2 and S_1 were labelled as diastole. An example of a labelled PCG recording is shown in Figure 4.1.

4.2.5 Hidden Markov models

HMMs are *generative* probabilistic models that describe the statistical relationship between an observable sequence, $\mathbf{O}$, and a sequence of discrete *hidden states*, $\mathbf{Q}$. These hidden states are governed by an underlying Markov model. HMMs operate by making inferences about the likelihood of being in one of the discrete hidden states, moving between these states, and observing what features are generated by each state. Therefore, they are a doubly stochastic process, where one stochastic process generates the sequence of hidden states, while a second stochastic process generates the features observed in each state [210].

In this work, the HMM was first order, with the hidden sequence consisting of the four states (S_1 , systole, S_2 , diastole), while the observed sequence was the features computed from the PCG signal.

An HMM is governed by two independence assumptions that allow tractable inference. These assumptions are:

- The *Markov property*, which states that each state at each time step is only dependent on its immediate predecessor. Therefore, the state q_t is independent of all its predecessors $(q_1, q_2 \dots q_{t-2})$ given q_{t-1} .
- The observation probabilities, $\mathbf{O}_t$, depend only on the current state, q_t , and cannot be influenced by observations outside of the current time-step.

This is valid for heart sounds, as each successive state can only be reached from one particular previous state.

Let the hidden states be defined as $\xi = [\xi_1, \xi_2, \dots, \xi_N]$ where N is the total number of states. In this work, $N = 4$ and ξ_1 refers to S_1 , ξ_2 refers to systole, ξ_3 refers to S_2 and ξ_4 refers to diastole. Let the duration of an entire sequence be T and the state at time t be q_t with the entire sequence of states being $\mathbf{Q}$. Let the observation sequence be $\mathbf{O} = \{\mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_T\}$, where $\mathbf{O}_t$ is a vector of feature values at time t . Then, an HMM can be defined as:

$$\lambda = (A, B, \pi) \quad (4.7)$$

where A is the transmission matrix, B is the emission or observation distribution and π is the initial state distribution [210].

The *initial state distribution*, $\pi = \{\pi_i\}$, $1 \leq i \leq N$, defines the probability of being in state i at the first time point: $p(q_1 = \xi_i)$. The discrete nature of the states within an HMM dictates that this distribution is always a probability mass function.

$A = \{a_{ij}\}$, $1 \leq i, j \leq N$, is the *transition matrix*, and defines the probability of moving from state i at time t , to state j at time $t + 1$: $p(q_{t+1} = \xi_j | q_t = \xi_i)$. A four-state heart sound HMM is a case of a “non-ergodic” HMM, where each state is only accessible from one specific other state. For example, S_2 has to precede diastole; diastole cannot be preceded by systole or S_1 ^e. In this case, with four-state HMM, A is a 4×4 stochastic matrix of probabilities, with each row summing to 1. An example of a transition matrix for the four-state heart sound HMM, taken from [99], is shown in Table 4.3.

^eIn contrast, an *ergodic* HMM is one where each state is accessible from every other state.

Table 4.3 – Example of a transition matrix for a four-state heart sound HMM, taken from [99]. The current state, i , is shown on the left of the table, while the subsequent state, j , is shown at the top.

$i \downarrow j \rightarrow$	S_1	Systole	S_2	Diastole
S_1	0.84	0.16	0	0
Systole	0	0.91	0.09	0
S_2	0	0	0.77	0.23
Diastole	0.04	0	0	0.96

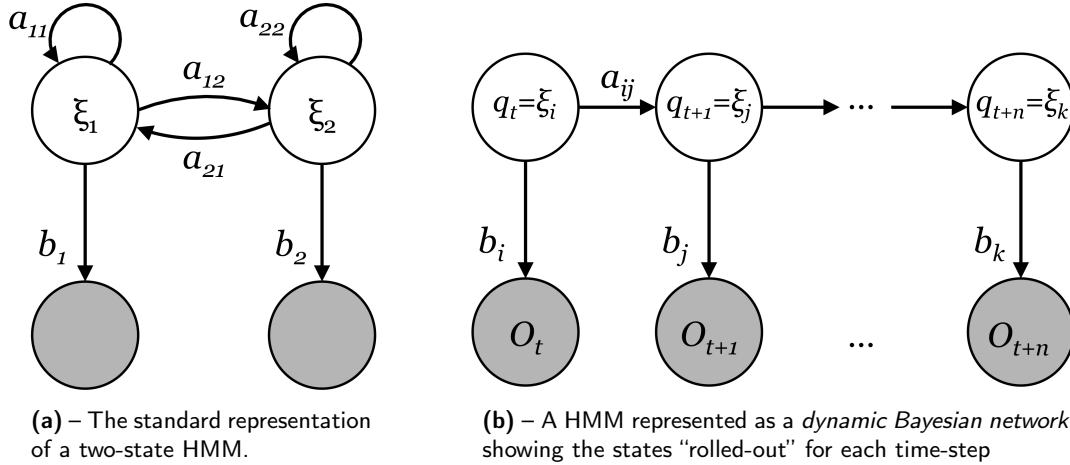


Figure 4.5 – Graphical representations of a hidden Markov model. The grey circles represent observed variables, while the white circles represent the underlying states. Each arrow represents a probability.

$B = \{b_j(\mathbf{O}_t)\}$, $1 \leq j \leq N, 1 \leq t \leq T$, the *observation* or *emission probability distributions*, define the probability that state j generates the observation vector $\mathbf{O}_t$ at time t : $p(\mathbf{O}_t | q_t = \xi_j)$. If these observations are discrete, then the observation distributions will be a probability mass function. In the case of continuous observations, probability density functions are employed, most often in the form of Gaussian or Gamma distributions [99, 211].

Therefore, A and π govern the underlying Markov model which describes the sequence of states. It is B , the observation probability distributions, that differentiates this *hidden* Markov model from a standard Markov model, as it is only through these probability distributions that we can infer what the underlying hidden states may be. Two graphical representations of such a model (with only two states) are shown in Figure 4.5. The standard representation of an HMM, as shown in Figure 4.5a, illustrates the two separate hidden states while showing the observations *generated* by each of those states. The second representation of the HMM, as shown in Figure 4.5b, shows the HMM over each time-step rather than each state of the HMM. This is the representation most often employed for the representation of dynamic Bayesian networks [212].

In the case of heart sound segmentation, a utility of the HMM is the computation of the optimal state sequence, given a model, λ , and an observation sequence, $\mathbf{O}$, where optimality is defined as the most likely sequence of states. Cycling through every combination of $\mathbf{Q}$ or order to find the optimal sequence is not feasible, even for short time sequences, because the space of possible label sequences is exponentially large with the size of the input. For example, if we were required to enumerate every possible state sequence for an ergodic four-state HMM with an input of 100 samples, we would have $4^{100} = 1.6 \times 10^{60}$ possible state sequences. Therefore, a dynamic programming method called the Viterbi algorithm is employed to compute the most likely state sequence [210].

4.2.5.1 The Viterbi algorithm

We define $\delta_t[i]$ as the likelihood of the *most probable* state sequence that accounts for the first t observations and ends in state i at time t :

$$\delta_t[i] = \max_{q_1, q_2, \dots, q_{t-1}} \{p(q_1, q_2, \dots, q_t = i, \mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_t | \lambda)\} \quad (4.8)$$

Then, the calculation of the δ variable at the next time-step can be shown to be:

$$\begin{aligned} \delta_{t+1}[j] &= \max_{q_1, q_2, \dots, q_t} \{p(q_1, q_2, \dots, q_{t+1} = j, \mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_{t+1} | \lambda)\} \\ &= \max_{1 \leq i \leq N} \left\{ \max_{q_1, q_2, \dots, q_{t-1}} \{p(q_1, q_2, \dots, q_t = i, \mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_t | \lambda)\} p(q_{t+1} = j | q_t = i) \right\} \\ &\quad \times p(\mathbf{O}_{t+1} | q_{t+1} = j) \end{aligned}$$

Therefore, the calculation of $\delta_{t+1}[j]$ can be seen to include a maximisation over the δ variable from the previous time-step, $\delta_t[i]$, multiplied by a transition probability. This expression identifies the *most probable* previous state. This probability is then multiplied by the emission probability. Therefore, by incorporating the information from the previous time-step, $\delta_{t+1}[j]$ can be calculated by induction using:

$$\delta_{t+1}[j] = \max_{1 \leq i \leq N} \{\delta_t[i] a_{ij}\} \times b_j(\mathbf{O}_{t+1}) \quad (4.9)$$

The algorithm is initialised at $t = 1$ using the initial state distribution, $\{\pi_i\}$:

$$\delta_1[i] = \pi_i b_i(\mathbf{O}_1) \quad (4.10)$$

The argument (or state) which maximises Equation 4.9 is stored in the matrix $\psi_t[j]$:

$$\psi_{t+1}[j] = \arg \max_{1 \leq i \leq N} \{\delta_t[i] a_{ij}\} \quad (4.11)$$

This can be seen as a “back-pointer”; it keeps a record of the most likely previous state, if in state j at time $t + 1$. This is used to identify the most probable sequence of states. Then, the most-likely state at time T , is identified:

$$q_T^* = \arg \max_{1 \leq i \leq N} \{\delta_T[i]\} \quad (4.12)$$

and by using the back-pointing ψ variables, the most probable state sequence can be found:

$$q_t^* = \psi_{t+1}[q_{t+1}^*], t = T - 1, T - 2, \dots, 1 \quad (4.13)$$

Therefore, the Viterbi algorithm can be seen to consist of two *passes*: the *forward pass* consists of the calculation of the δ and ψ variables from time $t = 1$ to T . The *backward pass* refers to the backtracking procedure used to extract the most probable sequence of states. The pseudo code for the standard Viterbi algorithm is shown in Algorithm 4.1.

4.2.5.2 The duration-dependent Viterbi algorithm and hidden semi-Markov models

A major limitation of the standard HMM is that it does not explicitly incorporate any information about the expected duration of each state. Without incorporating this information, the state durations are governed only by the self-transition probability a_{ii} . This results in a geometric distribution for the expected duration for each state [210]. This distribution monotonically decreases, resulting in the most likely state duration always being one time step. This is, intuitively, poorly suited for PCG analysis due to the expected durations of

Algorithm 4.1 Standard Viterbi algorithm

$$\delta_1[j] = \pi_j b_j(\mathbf{O}_1), \quad 1 \leq j \leq N$$

$$\psi_1[j] = 0, \quad 1 \leq j \leq N$$

for $t = 1 : T - 1$
 for $j = 1 : N$
 $\delta_{t+1}[j] = \max_{1 \leq i \leq N} \{ \delta_t[i] a_{ij} \} \times b_j(\mathbf{O}_{t+1})$
 $\psi_{t+1}[j] = \arg \max_{1 \leq i \leq N} \{ \delta_t[i] a_{ij} \}$
 end
end

$q_T^* = \arg \max_{1 \leq i \leq N} \{ \delta_T[i] \}$
for $t = T - 1 : -1 : 1$
 $q_t^* = \psi_{t+1}[q_{t+1}^*]$
end

each state lasting longer than a single time step or sample. In order to improve the duration modelling, an extra parameter is needed in the model.

Let us define the new model as $\lambda = (A, B, \pi, \mathbf{p})$, where $\mathbf{p} = \{\mathbf{p}_i[d]\}$ is the explicitly defined probability of remaining in state i for duration d (derived in Section 4.2.3). Then, modifying the Viterbi algorithm to include the duration densities, we find [99, 211]:

$$\delta_t[j] = \max_d \left\{ \max_{i \neq j} \{ \delta_{t-d}[i] a_{ij} \} \times \mathbf{p}_j[d] \times \prod_{s=0}^{d-1} b_j(\mathbf{O}_{t-s}) \right\} \quad (4.14)$$

with $1 \leq i, j \leq N$, $1 \leq t \leq T$ and $1 \leq d \leq d_{max}$ where in this case d_{max} , the maximum time expected to remain in any one state, is set to the duration of an entire heart cycle, to ensure tractability of the algorithm. As can be seen in Algorithm 4.2, the estimation of the $\delta_t[j]$, $\mathbf{D}_t[j]$ and $\psi_t[j]$ parameters is performed over values of d from 1 to d_{max} .

The observation density, $\prod_{s=0}^{d-1} b_j(\mathbf{O}_{t-s})$, is now the calculation of the probability of observing all the observations from time $t - d$ to time t in state j . Importantly, it should be noted that when incorporating duration densities in an HMM in this manner, the entries for the transition matrix, a_{ij} , in the case of a four-state heart sound model, become *unity* for the transition from one state to another, provided that it is a permissible transition (for example, S_1 to systole, or S_2 to diastole). This is because the transition between states is now only dependent on the time left in the current state, and no longer the probability of transitioning between states.

Equation 4.14 is maximised according to two arguments, i and d , which are stored in matrices $\psi_t[j]$ and $\mathbf{D}_t[j]$. The backward pass, in contrast to the standard Viterbi algorithm, now

Algorithm 4.2 The duration-dependent Viterbi algorithm

$$\delta_1[j] = \pi_j b_j(\mathbf{O}_1), \quad 1 \leq j \leq N$$

$$\psi_1[j] = 0, \quad 1 \leq j \leq N$$

$$\mathbf{D}_1[j] = 0, \quad 1 \leq j \leq N$$

for $t = 2 : T$

 for $j = 1 : N$

$$1 \leq i \leq N, 1 \leq d \leq d_{max}, t - d > 0$$

$$\delta_t[j] = \max_d \left\{ \max_{i \neq j} \{ \delta_{t-d}[i] a_{ij} \} \times \mathbf{p}_j[d] \times \prod_{s=0}^{d-1} b_j(\mathbf{O}_{t-s}) \right\}$$

$$\mathbf{D}_t[j] = \arg \max_d \left\{ \max_{i \neq j} \{ \delta_{t-d}[i] a_{ij} \} \times \mathbf{p}_j[d] \times \prod_{s=0}^{d-1} b_j(\mathbf{O}_{t-s}) \right\}$$

$$\psi_t[j] = \arg \max_{1 \leq i \leq N} \{ \delta_{t-\mathbf{D}_t[j]}[i] a_{ij} \}$$

 end

end

$$q_T^* = \arg \max_{1 \leq i \leq N} \{ \delta_T[i] \}$$

$$t = T$$

while $t > 1$

$$d^* = \mathbf{D}_t[q_t^*]$$

$$q_{t-d^*:t-1} = q_t^*$$

$$q_{t-d^*-1} = \psi_t[q_t^*]$$

$$t = t - d^*$$

end

includes stepping back to the most likely state, but also stepping back $\mathbf{D}_t[j]$ samples. This is then called a hidden semi-Markov model (HSMM)^f, as only the transition between different states is Markovian [211].

The pseudo-code for the duration-dependent Viterbi algorithm is shown in Algorithm 4.2.

An examination of Equation 4.14 highlights one of the major limitations of this algorithm. Both the transition probability, $\delta_{t-d}[i] a_{ij}$, and the observation probabilities, $\prod_{s=0}^{d-1} b_j(\mathbf{O}_{t-s})$, rely on δ and $\mathbf{O}$ values from previous samples. These previous samples must have positive time indices ($t > 0$), while the δ and $\mathbf{O}$ values are also limited by the fact that they only extend up to $t = T$. Therefore, this algorithm makes two implicit assumptions about the start and end states assigned to the observation sequence:

Firstly, the algorithm assumes that an observation sequence commences at exactly the same time as the start of a state. Therefore, the first hidden state in the observation sequence, and hence the assigned duration distribution, starts at the first observed value at $t = 1$. So, regardless of the point in the cardiac cycle at which a heart-sound recording commenced, the

^f Also known as an explicit duration, variable-duration, segmental or duration-dependent HMM [213].

algorithm attempts to assign an entire state duration from $t = 1$. For example, consider a PCG recording that commences half-way through diastole. Such a short diastolic period has a low probability given the model, as the expected duration of the diastolic state is twice the observed duration. This may lead to erroneous state assignments.

Secondly, and similarly to the first assumption, the algorithm assumes that the observation sequence ends at the exact time that a state ends. In Equation 4.14, it can be seen that the duration distributions are measured backwards from time t . Therefore, when reaching the final time-step, T , the duration of the state is assumed to end at the conclusion of the PCG recording.

These assumptions are referred to as the “simplifying assumptions” by Yu [213], and are said to be the most common form of the algorithm found in the literature. They are poorly suited for PCG analysis, as a PCG recording can start and end at any stage in a heart cycle, and such assumptions can lead to incorrect state assignment. Such limitations were identified but not addressed by Schmidt et al. [99] in their analysis of PCG recordings. Such errors would be more pronounced in shorter recordings, where the model does not have enough data to compensate for these assumptions, or in recordings of high heart rate, where the durations of systole and diastole are very similar. A further extension to the Viterbi algorithm can be used to overcome these limitations.

4.2.5.3 The extended duration-dependent Viterbi algorithm

As explained above, the duration-dependent Viterbi algorithm for decoding the most likely sequence of states when using an HSMM is limited by the fact that states are required to start and end at the start and end of the PCG signal. In order to resolve this, an extended Viterbi algorithm is proposed, which allows the duration distributions to extend beyond the beginning and end of the PCG recording. The pseudo-code for this algorithm is shown in Algorithm 4.3 on the following page.

The extended duration-dependent Viterbi algorithm overcomes the shortcomings of the previous Viterbi algorithm by allowing the durations of the states to extend past the beginning and end of the observed sequence, while only considering observations from within the PCG

Algorithm 4.3 The extended duration-dependent Viterbi algorithm

$$\delta_1[j] = \pi_j b_j(\mathbf{O}_1), \quad 1 \leq j \leq N$$

$$\psi_1[j] = 0, \quad 1 \leq j \leq N$$

$$\mathbf{D}_1[j] = 0, \quad 1 \leq j \leq N$$

for $t = 2 : T + d_{max} - 1$

 for $i, j = 1 : N$

 for $d = 1 : d_{max}$

$$start_t = t - d$$

 if $start_t < 1$

$$start_t = 1$$

 else if $start_t > T - 1$

$$start_t = T - 1$$

 end

$$end_t = t$$

 if $end_t > T$

$$end_t = T$$

 end

$$\delta_t[j] = \max_d \left\{ \max_{i \neq j} \{ \delta_{start_t}[i] a_{ij} \} \times \mathbf{p}_j[d] \times \prod_{s=start_t}^{end_t} b_j(\mathbf{O}_s) \right\}$$

$$\mathbf{D}_t[j] = \arg \max_d \left\{ \max_{i \neq j} \{ \delta_{start_t}[i] a_{ij} \} \times \mathbf{p}_j[d] \times \prod_{s=start_t}^{end_t} b_j(\mathbf{O}_s) \right\}$$

$$\psi_t[j] = \arg \max_{1 \leq i \leq N} \{ \delta_{t-\mathbf{D}_t[j]}[i] a_{ij} \}$$

 end

end

end

$$T^* = \arg \max_t \{ \delta_t[i] \} \quad 1 \leq i \leq N, \quad T \leq t \leq T + d_{max} - 1$$

$$q_{T^*}^* = \arg \max_i \{ \delta_{T^*}[i] \}$$

$$t = T^*$$

while $t > 1$

$$d^* = \mathbf{D}_t[q_t^*]$$

$$q_{t-d^*:t-1} = q_t^*$$

$$q_{t-d^*-1}^* = \psi_t[q_t^*]$$

$$t = t - d^*$$

end

for emission probability estimation. This is demonstrated in the calculation of the δ variable:

$$\delta_t[j] = \max_d \left\{ \max_{i \neq j} \{ \delta_{start_t}[i] a_{ij} \} \times \mathbf{p}_j[d] \times \prod_{s=start_t}^{end_t} b_j(\mathbf{O}_s) \right\} \quad (4.15)$$

In Equation 4.15, it can be seen that the observation probabilities are computed from $start_t$ to end_t (which are limited to $1 \leq start_t, end_t \leq T$), while the duration probability, $\mathbf{p}_j[d]$, is not limited in its duration. Therefore, this allows the inclusion of duration probabilities that may extend beyond the beginning and end of the PCG sequence, while only computing the observation probabilities from the samples within the observed sequence. This algorithm uses the equations for the “general assumption” of the forward-backward algorithm and Viterbi algorithm from Yu [213].

4.2.6 Logistic regression HSMM

A further modification to the HSMM proposed here was to incorporate logistic regression (LR) into the model in order to derive the emission or observation probability estimates. In related previous works, Gaussian or Gamma distributions were employed [99, 211]. However, the incorporation of LR into the HMM was thought to allow for greater discrimination between states, as such discriminative methods have been shown to outperform naive Bayes classification when using large datasets [214]. This is similar to the use of support vector machine-based emission probabilities as employed by other researchers [215]. This, then, can be seen as a *hybrid* between a generative and discriminative model.

Logistic regression maps predictor variables, or features, to a response variable using the logistic function $\sigma(a)$, defined by $\sigma(a) = \frac{1}{1+exp(-a)}$ [216]. More details on logistic regression can be found in Appendix B.

Using the logistic function above, the probability of a specific class or state, given the input features or observations, can be defined by:

$$P[q_t = \xi_j | \mathbf{O}_t] = \sigma(\mathbf{w}'\mathbf{O}_t) \quad (4.16)$$

where $\mathbf{w}$ are the weights of the model, applied to each input feature, and trained using iteratively re-weighted least squares optimisation on the training data.

A one-versus-all approach was implemented, training one LR model for the observations from each state in the model. Thereafter, the probability of an observation given a state, $b_j(\mathbf{O}_t|\xi_j)$, as required for the HSMM, can be found using Bayes' rule:

$$b_j(\mathbf{O}_t) = P[\mathbf{O}_t|q_t = \xi_j] = \frac{P[q_t = \xi_j|\mathbf{O}_t] \times P(\mathbf{O}_t)}{P(\xi_j)} \quad (4.17)$$

where $P(\mathbf{O}_t)$ is found from a multivariate normal distribution computed from the features from the entire training set and $P(\xi_j)$ is found from π , the initial state probability distribution. The inputs to the LR models are the signal-derived features described in Section 4.2.8.

4.2.7 Conditional random fields

Conditional random fields (CRFs) are a *discriminative* probabilistic model, used to label or segment sequential data [217]. Sutton and McCallum [218] describe CRFs as a conditional distribution, $p(\mathbf{Q}|\mathbf{O})$, with an associated graphical structure. One of the main advantages of the use of CRFs over HMMs is the fact that CRFs are not bound by the independence assumptions required by the HMM to ensure tractable inference [219] (see Section 4.2.5). CRFs have been used extensively for speech processing [217], natural language processing [220] and image segmentation [221]. Semi-Markov (or duration-dependent) CRFs have been employed for activity segmentation [222] and named-entity recognition [223]. However, the application of CRFs to biomedical data is still relatively new, with a few studies on ECG classification [224], activity recognition [225] and motor task identification from EEG data [226].

In contrast to HMMs, CRFs, due to their conditional formulation, can easily incorporate long-ranging, overlapping features. For example, CRFs can incorporate the current observation probabilities into the transition probabilities, or incorporate observations from past (as well as future) time-steps into the observation probabilities. In addition, CRFs do not suffer from the label-bias problem, a weakness exhibited by other conditional Markov models based on directed graphical models, whereby states with few outgoing transitions effectively ignore the observations in those states [217].

Using the same notation as for the HMMs above, with $\mathbf{O}$ being the observations and $\mathbf{Q}$ being the vector of state labels, q_t , over the entire sequence from time $t = 1$ to time T , then the joint

probability of a labelled sequence using a generative model such as the HMM can be written as [227]:

$$p(\mathbf{Q}, \mathbf{O}) = \prod_{t=1}^T p(q_t | q_{t-1}) p(\mathbf{O}_t | q_t) \quad (4.18)$$

In contrast, a discriminative model, such as a linear-chain CRF^g, models this relationship as a conditional distribution, $p(\mathbf{Q} | \mathbf{O})$, in the form:

$$p(\mathbf{Q} | \mathbf{O}) = \frac{1}{Z(\mathbf{O})} \prod_{t=1}^T \exp \left\{ \sum_{k=1}^K \theta_k f_k(q_t, q_{t-1}, \mathbf{O}_t) \right\} \quad (4.19)$$

where $\boldsymbol{\theta} = \{\theta_k\} \in \Re^K$ is a parameter vector and $\mathcal{F} = \{f_k(q, q', \mathbf{O}_t)\}_{k=1}^K$ is a set of K real-valued *feature functions* (which will be explained in due course), and $Z(\mathbf{O})$ is a normalisation function,

$$Z(\mathbf{O}) = \sum_{\mathbf{Q}} \prod_{t=1}^T \exp \left\{ \sum_{k=1}^K \theta_k f_k(q_t, q_{t-1}, \mathbf{O}_t) \right\} \quad (4.20)$$

From observation of Equations 4.18 and 4.19, it can be seen that an HMM can be interpreted as being a naïve Bayes classifier over multiple time points, where a naïve Bayes classifier is defined as:

$$p(y, \mathbf{x}) = p(y) \prod_{k=1}^K p(x_k | y) \quad (4.21)$$

where y in this case is a state label and $\mathbf{x}$ is a feature vector of dimension K . In contrast, a CRF can be seen to have the same form as logistic regression, written as a *log-linear model* or *softmax function*^h, but again, applied over a sequence of length T , where the conditional probability from logistic regression can be defined by:

$$p(y | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \theta_{y,0} + \sum_{j=1}^K \theta_{y,j} x_j \right\} \quad (4.22)$$

^gA linear-chain CRF is a special case where the inputs to the feature functions are limited to the current state label, q_t , and the previous state label, q_{t-1} . General CRFs incorporate arbitrary relationships between state labels, but these are not covered here.

^hSee Equation 4.104 of [216].

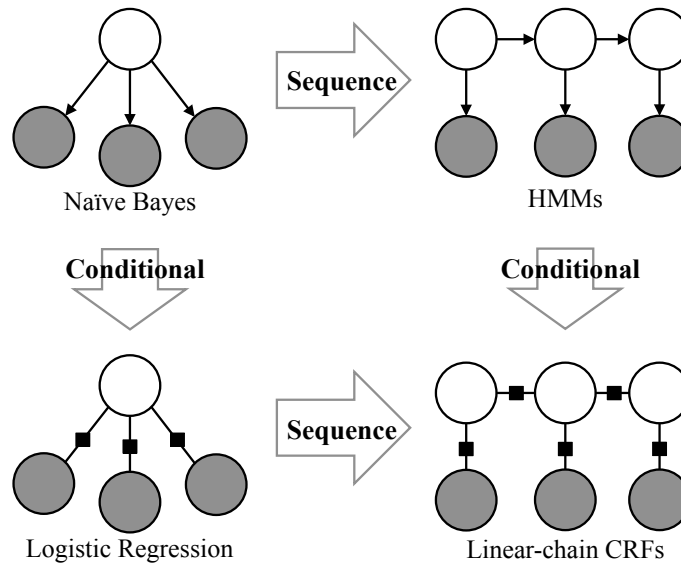


Figure 4.6 – The relationship between naïve Bayes, logistic regression, HMMs and linear-chain CRFs. White circles represent unobserved variables while grey circles represent observed features. Black arrows represent probabilities, while lines with black boxes represent feature functions. Adapted from [227].

where $Z(\mathbf{x}) = \sum_y \exp \left\{ \theta_{y,0} + \sum_{j=1}^K \theta_{y,j} x_j \right\}$ is a normalising constant, θ_y is a set of weights of the same dimension as the feature vector $\mathbf{x}$ and $\theta_{y,0}$ is a bias weight that acts like $p(y)$ in the naïve Bayes classifier. A naïve Bayes classifier and logistic regression form a *generative-discriminative pair* [214], where each classifier is the generative or discriminative implementation of the other [218]. HMMs and linear-chain CRFs are the implementation of a naïve Bayes classifier and logistic regression when applied to a sequence. This relationship is illustrated in Figure 4.6.

Feature functions, which are used in the defining equations of a linear-chain CRF (Equation 4.19), are functions that accept a current state label, a previous state labelⁱ, and a set of observations. These feature functions output real numbers, which are then weighted by the trained weights, θ_k . For example, feature functions that are analogous to the functions in an HMM are a transition feature function for the transition between each state (corresponding to each element of the transition matrix of an HMM), and a set of observation feature functions, one for each of the possible observations in each of the states (corresponding to the probability mass function for each state of a discrete observation HMM). Such CRFs, involving separate transition and observation feature functions, are often written as:

ⁱAgain, limiting the feature functions to the current and previous time-step results in a special case of linear-chain CRFs.

$$p(\mathbf{Q}|\mathbf{O}) = \frac{1}{Z(\mathbf{O})} \prod_{t=1}^T \exp \left\{ \sum_{m=1}^M \mu_m g_m(q_t, q_{t-1}) + \sum_{p=1}^P \varphi_p h_p(q_t, \mathbf{O}_t) \right\} \quad (4.23)$$

where g_m is the m^{th} transition feature function (reliant on only the current and former states, q_t and q_{t-1}), weighted by its corresponding weight vector, $\boldsymbol{\mu}$, and h_p is the observation feature function (which only takes into account the current state and observations, q_t and $\mathbf{O}_t$), also weighted by its own weight vector, $\boldsymbol{\varphi}^j$. M and P are the total number of transition and observation feature functions respectively.

It is common for such feature functions to output zero or one in cases of discrete observations. The non-zero feature functions would contribute to the model by having a non-zero product with their corresponding weights. In the case of continuous-valued observation values, the feature function outputs the observation value (rather than one or zero) with a weight trained for each observation vector of each state.

The parameters of the CRF are trained using gradient-based optimisation methods, which lead to a global optimum. If we presume that each i.i.d. training data example, i , has length T^k , then we can estimate the *conditional log likelihood* over all training examples [227]:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \ln p(\mathbf{Q}^{(i)}|\mathbf{O}^{(i)}; \boldsymbol{\theta}) \quad (4.24)$$

where N is the total number of training examples. Maximising $\ell(\boldsymbol{\theta})$ leads to the maximum likelihood estimate of the model parameters. If we substitute Equation 4.19 into Equation 4.24, we derive the following expression [227]:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{t=1}^T \sum_{k=1}^K \theta_k f_k(q_t^{(i)}, q_{t-1}^{(i)}, \mathbf{O}_t^{(i)}) - \sum_{i=1}^N \ln Z(\mathbf{O}^{(i)}) \quad (4.25)$$

If the L_2 or *Euclidian* norm of the θ parameters is used as regularisation in the model in order to reduce over-fitting^{*l*}, Equation 4.25 becomes:

^{*j*}When using graphical model nomenclature, g_m is sometimes called a *node* feature function, while h_p is called an *edge* feature function.

^{*k*}This is a simplification for the sake of deriving the equations for parameter optimisation. Data of varying length, T , can be incorporated into the equations in a straightforward fashion.

^{*l*}The L_1 is another choice for regularisation of the model, which would result in a sparser model. However, such a regulariser is not differentiable at 0, which can complicate the gradient-based optimisation of the CRF [227].

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{t=1}^T \sum_{k=1}^K \theta_k f_k(q_t^{(i)}, q_{t-1}^{(i)}, \mathbf{O}_t^{(i)}) - \sum_{i=1}^N \ln Z(\mathbf{O}^{(i)}) - \sum_{k=1}^K \frac{\theta_k^2}{2\sigma^2} \quad (4.26)$$

where σ is the parameter that controls the penalisation of large weights in the model.

As mentioned above, gradient-based optimisation is used, as the function $\ell(\boldsymbol{\theta})$ cannot be maximised in closed form. However, Equation 4.26 is convex, and therefore convergence to a global maximum can be guaranteed [219]. The partial derivatives of Equation 4.26 are:

$$\frac{\partial \ell}{\partial \theta_k} = \sum_{i=1}^N \sum_{t=1}^T f_k(q_t^{(i)}, q_{t-1}^{(i)}, \mathbf{O}_t^{(i)}) - \sum_{i=1}^N \sum_{t=1}^T \sum_{q, q'} f_k(q, q', \mathbf{O}_t^{(i)}) p(q, q' | \mathbf{O}_t^{(i)}, \boldsymbol{\theta}) - \frac{\theta_k}{\sigma^2} \quad (4.27)$$

The first term, $\sum_{i=1}^N \sum_{t=1}^T f_k(q_t^{(i)}, q_{t-1}^{(i)}, \mathbf{O}_t^{(i)})$, can be interpreted as the expectation of f_k under the *empirical distribution* of the training data - this involves summing all the values for each feature function at each time point and for each training example.

The second term, $\sum_{i=1}^N \sum_{t=1}^T \sum_{q, q'} f_k(q, q', \mathbf{O}_t^{(i)}) p(q, q' | \mathbf{O}_t^{(i)}, \boldsymbol{\theta})$, which is derived from the derivative of $Z(\mathbf{O}^{(i)})$, is the expectation of f_k under the *model distribution*. It can be interpreted as being the sum of the feature function value, $f_k(q, q', \mathbf{O}_t^{(i)})$, weighted by the probability of the current and previous state labels (q and q') given the observations and the parameter vector $\boldsymbol{\theta}$, over all possible state label pairs, times and training records. In the unregularised case, when the gradient is zero, these two terms are equal, which means that the expectations of f_k are equal under the empirical and the model distributions.

In order to perform parameter optimisation, inference is needed in order to estimate both $Z(\mathbf{O}^{(i)})$ and $p(q, q' | \mathbf{O}^{(i)}, \boldsymbol{\theta})$. These values can be estimated using the same inference routines that are commonly used for HMMs, such as the forward-backward algorithm [210]. Similarly, the Viterbi algorithm can be used to compute the most likely sequence of states, given a model and the observation values.

The forward-backward algorithm is a method of computing the probability of an observation sequence given a model, $p(\mathbf{O} | \lambda)$ [210]. A straightforward way of doing this would be to compute the probability of every possible state sequence over the observations. However, this becomes intractable for even small lengths of observations. The forward-backward algorithm performs the same computation, but employs dynamic programming to reduce computation.

Firstly, let us define $\Psi_t(j, i, O)$ as:

$$\Psi_t(j, i, O) = p(q_t = j | q_{t-1} = i) p(O = \mathbf{O}_t | q_t = j) \quad (4.28)$$

which can be interpreted as the probability of being in state j at time t , given state i at time $t - 1$, and the observation $\mathbf{O}_t$ in state j . The value $p(q_t = j | q_{t-1} = i)$ can be seen as the transition matrix in an HMM or the probabilities derived from a transition feature function in a CRF, while $p(O = \mathbf{O}_t | q_t = j)$ can be interpreted as the emission probabilities of an HMM or the observation probabilities derived from observation feature functions in a CRF. Then, let the *forward variable* be $\alpha_t[i] = p(\mathbf{O}_1, \mathbf{O}_2 \dots \mathbf{O}_t, q_t = \xi_i | \theta)$, which is the probability of being in state i at time t , with the partial observation sequence until time t , given the model parameters, θ . Then, $\alpha_t[i]$ can be solved inductively:

$$\alpha_t[j] = \sum_{i=1}^N \Psi_t(j, i, O) \alpha_{t-1}[i], \quad 2 \leq t \leq T - 1, 1 \leq j \leq N \quad (4.29)$$

For an HMM, consisting of the transition matrix, a_{ij} , and the emission probabilities, $b_j(\mathbf{O}_t)$, this could be written as:

$$\alpha_t[j] = \left\{ \sum_{i=1}^N \alpha_{t-1}[i] a_{ij} \right\} b_j(\mathbf{O}_t) \quad (4.30)$$

For a CRF, consisting of the transition feature function $\mu_{ij} g_{ij}(\xi_j, \xi_i)$, and observation feature function $\varphi_j h_j(\xi_j, \mathbf{O}_t)^m$, (see Equation 4.23), Equation 4.29 can be written as:

$$\alpha_t[j] = \left\{ \sum_{i=1}^N \alpha_{t-1}[i] \exp(\mu_{ij} g_{ij}(\xi_j, \xi_i)) \right\} \exp(\varphi_j h_j(\xi_j, \mathbf{O}_t)) \quad (4.31)$$

where $\alpha_1[j]$ is the special case for which $\alpha_1[j] = p(q_1 = j) p(O = \mathbf{O}_1 | q_1 = j)$, $1 \leq j \leq N$. The value $p(q_1 = j)$ is the initial state distribution, π_j , defining the probability of a sequence starting in state j . Importantly, $\alpha_t[j]$ refers to the *unnormalised* probability of being in state j at time t , given the partial observations up until time t .

^mWhere μ_{ij} and g_{ij} are the weighting and feature function for the specific transition from ξ_i to ξ_j , while φ_j and h_j are the specific weights for the observation vector in state j .

The normaliser in Equation 4.19, $Z(\mathbf{O})$, can be estimated by summing the final forward variable values:

$$Z(\mathbf{O}) = p(\mathbf{O}) = \sum_{i=1}^N \alpha_T[i] \quad (4.32)$$

This normaliser is equal to the probability of seeing the entire observation sequence given the model parameters, $p(\mathbf{O})$.

While the forward variable deals with the time-steps from 1 to t , the *backward variable*, $\beta_t[i]$, can be seen to compute the probability of the partial observation sequence from $t + 1$ to T , given state i at time t and the model: $\beta_t[i] = p(O_{t+1}, O_{t+2} \dots O_T | q_t = \xi_i, \theta)$. This, again, can be solved inductively:

$$\beta_t[i] = \sum_{j=1}^N \Psi_{t+1}(j, i, O) \beta_{t+1}[j] \quad (4.33)$$

where $\beta_T[i]$ is initialised as $\beta_T[i] = 1$. Again, $\beta_t[i]$ refers to the unnormalised probabilities.

The forward and backward variables can be combined to derive the probabilities needed for the re-estimation procedure of the CRF. Firstly, let us define a new variable, $\gamma_t[i] = p(q_t = \xi_i | \mathbf{O}, \theta)$, which is the probability of being in a specific state, i , at time t , given the observations and the model parameters. When using graphical model nomenclature, this is often named the *node belief*. The node belief can be computed using:

$$\gamma_t[i] = \frac{\alpha_t[i] \beta_t[i]}{\sum_i^N \alpha_t[i] \beta_t[i]} \quad (4.34)$$

Intuitively, this combines the probability of being in state ξ_i at time t given the observations from the beginning of the sequence to time t (the α variable), and the same probability when considering the observations from $t + 1$ to the end of the observation sequence (the β variable), normalised to ensure that $\sum_i^N \gamma_t[i] = 1$.

Secondly, we compute the *edge beliefs*, or the probability of being in state ξ_i at time t and state ξ_j at time $t + 1$, given the observations and the model: $\zeta_t[i, j] = p(q_t = \xi_i, q_{t+1} = \xi_j | \mathbf{O}, \theta)$.

This can be computed using:

$$\zeta_t[i, j] = \frac{\alpha_t[i] \Psi_{t+1}(j, i, O) \beta_{t+1}[j]}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t[i] \Psi_{t+1}(j, i, O) \beta_{t+1}[j]} \quad (4.35)$$

The node and edge beliefs can be used to compute the second term in Equation 4.27, $\sum_{i=1}^N \sum_{t=1}^T \sum_{q, q'} f_k(q, q', \mathbf{O}_t^{(i)}) p(q, q' | \mathbf{O}^{(i)}, \theta)$. This term gives the expectation of the k^{th} feature function, given the current values of the θ parameters. If f_k is a transition feature function, then $p(q, q' | \mathbf{O}^{(i)}, \theta)$ simply refers to the edge belief, ζ . This can be seen when considering the simplified expression for the edge belief, $\zeta_t(i, j) = p(q_t = \xi_i, q_{t+1} = \xi_j | \mathbf{O}, \theta)$. If f_k is an observation feature function, then $p(q, q' | \mathbf{O}^{(i)}, \theta)$ can be written as $p(q | \mathbf{O}^{(i)}, \theta)$, which refers to the node belief, γ . When computing $\sum_{i=1}^N \sum_{t=1}^T \sum_{q, q'} f_k(q, q', \mathbf{O}_t^{(i)}) p(q, q' | \mathbf{O}^{(i)}, \theta)$, each feature function is multiplied by its corresponding edge or node belief. This is done for each state, or state pairs when dealing with transition probabilities, and summed over the duration of each training observation sequence, and then summed for all the training sequences.

For each iteration of the algorithm, the weights, θ , are updated based on the gradient computed from Equation 4.27. Then, the forward, α , backward, β , node, γ and edge, ζ , beliefs are computed again. This is repeated until a stopping criterion on the log-likelihood, gradient values or number of iterations has been reachedⁿ. In this work, two existing CRF implementations were modified for use with PCGs [228, 229].

4.2.7.1 Duration-dependent conditional random fields

The duration-dependent CRFs used in this chapter have exactly the same formulation as above when estimating the optimal parameters for the model, but make use of the duration-dependent Viterbi algorithm, as shown in Algorithm 4.2, to decode the optimal sequence of states.

4.2.8 Feature extraction

Heart sounds are non-stationary signals, consisting of high-amplitude, short-duration, high-frequency fundamental heart sounds and clicks, as well as low-amplitude, longer-duration, lower-frequency murmurs. Therefore, information that could discriminate between the specific sounds in a PCG is present in both the time and frequency domain. While purely spectral

ⁿIn this work, the stopping criteria were the number of iterations exceeding 500 or the sum of gradient values being below 10^{-5} .

representations of the PCG are inappropriate for the identification of the S_1 and S_2 sounds, purely temporal representations of the signal are often corrupted by types of noise, which might be identified from their frequency content. For this reason, time-frequency representations are well-suited for segmentation of the fundamental heart sounds within non-stationary signals such as the PCG, whose characteristic properties evolve with time.

The segmentation algorithms evaluated in this thesis used a combination of four input features (step v in Figure 4.2). Two of these were temporal envelopes of the signal, while two were time-frequency representations of the signal. The homomorphic envelope (see below) and power spectral density features were selected, as these features were employed by Schmidt et al. [99], allowing a direct comparison of their methods with the proposed methods. The Hilbert envelope was selected in order to compare the homomorphic and Hilbert envelope extraction methods. Finally, a wavelet-based feature was used as such features have shown promise for heart sound segmentation in previous research [124, 131, 137, 188].

Before feature extraction, all recordings were downsampled to 1 kHz using a poly-phase anti-aliasing filter. The majority of the frequency content of the fundamental heart sounds (S_1 and S_2) is below 150 Hz [92] and hence the Nyquist-Shannon sampling criterion [164] is satisfied^o. The four features were then calculated, as described below.

4.2.8.1 Homomorphic envelope

The homomorphic envelope, derived from the exponentiation of the low-pass filtered natural logarithm of a signal [166, 230], has been used by numerous researchers for extracting the envelope of PCG signals [125, 128], including the state-of-the-art segmentation algorithm [99]. The principle of homomorphic filtering is to find a transformation that converts the multiplicative operation by which multiple components of a signal are combined into a simple linear operation. This, then, allows the use of a simple linear filter to remove unwanted components of the signal, after which the signal is mapped back to its original domain [166].

The homomorphic envelope is extracted by modelling the PCG as a multiplicative signal, containing the product of a slow-varying, modulating signal (the signal envelope) and high-

^oAt this stage of analysis, we want to ignore high-frequency sounds such as murmurs, in order to more accurately identify the positions of the S_1 and S_2 sounds. Therefore, an upper limit of 500 Hz for analysis is acceptable.

frequency components. Filtering out the high-frequency components yields the slow-varying envelope of the signal. If the PCG signal at time t , x_t , is modelled as:

$$x_t = a_t \times o_t, a_t > 0 \quad (4.36)$$

where a_t is the amplitude component and o_t is the oscillating component of the signal, then the product of the slow-varying and fast-varying components of the signal can be split into an addition using:

$$\ln |x_t| = \ln |a_t| + \ln |o_t| \quad (4.37)$$

This transformation from multiplication to addition allows the use of the linear filter to extract the low-frequency components of the signal. Therefore, a_t can be extracted using:

$$a_t \approx \exp(L(\ln(|x_t|))) \quad (4.38)$$

where L is a linear low-pass filter. In this instance, L , was a first-order, zero-phase Butterworth low-pass filter with a cut-off frequency of 8Hz, as used by Schmidt et al. [99].

As a result of the used of the natural logarithm operation to extract the homomorphic envelope, this transformation has a singularity at signal values of zero. However, this singularity does not exist for complex-valued signals, even if the magnitude of the complex signal is zero. Therefore, the Hilbert transform, which has the ability to construct a complex-valued signal from a real-valued signal without changing the signal characteristics significantly [166], was applied before extracting the homomorphic envelope.

4.2.8.2 Hilbert envelope

Messer et al. [136] and Kumar et al. [121] calculate the envelope of the PCG signal using the Hilbert transform. The Hilbert transform extracts the analytic signal [166], and the Hilbert envelope is calculated from the absolute value of the Hilbert transform.

The Hilbert transform for a continuous signal is defined by:

$$H(x(t)) = \pi^{-1} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau = x(t) \star \frac{1}{\pi t} \quad (4.39)$$

where $\star$ indicates the convolution operator. The complex analytic signal, $A(x(t))$, can be expressed as:

$$A(x(t)) = x(t) + jH(x(t)) = E(t)e^{j\varphi(t)} \quad (4.40)$$

with $E(t) = \sqrt{x(t)^2 + H(x(t))^2}$ and $\varphi(t) = \tan^{-1}(H(x(t))/x(t))$. Here, $E(t)$ represents the envelope of the signal, while $\varphi(t)$ can be interpreted as the instantaneous frequency [187, 136].

4.2.8.3 Wavelet envelope

Wavelets are described in detail in Appendix D. Wavelets have been extensively explored for the analysis of heart sound signals. However, there is limited agreement on the optimal wavelet to use when denoising, segmenting or classifying heart sounds. Some researchers have argued that the Morlet family is best suited for heart sound analysis [124, 231] while others have proposed the use of the Daubechies wavelet family [125, 126, 136, 137, 142, 146, 232, 233, 234]. The Symlet [235, 236] and Coiflet [237] wavelet families have also been suggested for heart sound feature extraction.

In this chapter, the choice of wavelet was determined experimentally using the labelled heart sounds computed from Section 4.2.4. This is shown as step *iv* in Figure 4.2.

The heart-sound recordings in the training set (see Section 4.2.9.1) were decomposed using the discrete wavelet transform (DWT) with various wavelet families and decomposition levels^p. The absolute value of the detail coefficients for each wavelet family and decomposition level were computed in order to exclude frequency content outside of the target wavelet range and to extract a positive-valued envelope. Then, the wavelet family and decomposition level that yielded the highest ratio of the mean detail coefficients for the S_1 and S_2 sounds compared to the mean value over other intervals across all recordings were selected for further use. This process is illustrated in Figure 4.7.

^pThese included Haar, Daubechies, symlet, Coiflet, biorthogonal and reverse biorthogonal wavelets at decomposition levels 1-10. Morlet wavelets were not used since a DWT cannot be performed using the Morlet wavelet.

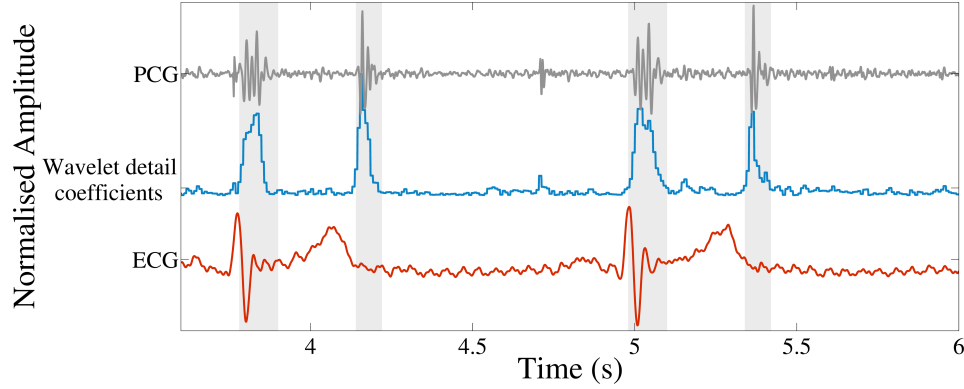


Figure 4.7 – Determination of wavelet family and decomposition level to use for wavelet feature extraction. The PCG is shown in grey, with the wavelet detail coefficients (extracted using the reverse biorthogonal 3.9 wavelet at decomposition level 3) are shown in blue. The ECG is shown in red, which was used to label the duration of the S_1 and S_2 sounds, as per Section 4.2.4 (shown as grey shaded bars). The mean value of the wavelet coefficients within the grey shaded bars was found. The ratio between this mean value and the mean value of the coefficients outside of the grey bars was found. This gives an indication of how well a wavelet accentuates the S_1 and S_2 sounds as compared to the sounds in systole and diastole. The wavelet family and decomposition level that maximised this ratio across all training recordings were selected for wavelet envelope extraction.

This ratio gives an indication of how well each wavelet discriminates between the FHSs and other regions of the heart-sound recordings. Therefore, a wavelet envelope computed using such a wavelet would hopefully provide the best overall discrimination between the FHSs and other sounds or noise for all heart-sound recordings.

4.2.8.4 Power spectral density envelope

The majority of the frequency content of the S_1 and S_2 sounds is below 150 Hz with a peak at 50 Hz [92]. Based on these frequencies, the final feature was derived from the mean power spectral density (PSD) between 40 and 60 Hz, found using overlapping windows of 0.05 s in width with 50% overlap^q. This resulted in an envelope of PSD values. The window size used ensures that the frequency content of the shortest expected fundamental heart sound (0.05 s) is covered by an analysis window. The PSD was calculated using the short-time discrete Fourier transform (DFT) after Hamming windowing.

The feature vectors for each recording were individually normalised by subtracting their mean and dividing by their standard deviation. Following Schmidt et al. [99], after feature extraction and normalisation, the resulting feature vectors were downsampled further to 50 Hz using a poly-phase anti-aliasing filter, in order to increase the speed of computation.

^qThe original PCG was downsampled to 5 kHz in this case, allowing analysis windows of 250 samples, zero padded to 256 samples.

4.2.9 Model training and evaluation

The parameters of the HSMM models and CRF were trained using the labelled PCG sequences described in Section 4.2.4, which were divided as follows.

4.2.9.1 Training and test data split

In order to avoid over-training of the model, the dataset was randomly split into equal training and test sets (step *iii* in Figure 4.2). The number of recordings for each condition (normal, murmurs relating to MVP, benign murmurs, aortic disease and miscellaneous conditions) was split evenly between the two sets. It was ensured that no recordings from any patient were in both the training and test sets.

4.2.9.2 Training

The parameters of the HSMMs (π , a_{ij} , B), the feature function weights of the CRF (μ , φ)^r and regularisation parameters of the CRF (λ) were optimised on the training set data.

In the case of the state-of-the-art segmentation method, a multivariate normal (MVN) distribution was used to model the emission probabilities when using an HSMM [99]. In this case, an MVN distribution was computed for each state by using the maximum likelihood of the input features for each state from the training recordings. This MVN distribution can be defined as $b_j(\mathbf{O}_t) \sim \mathcal{N}_{\mu_j, \Sigma_j}(\mathbf{O}_t)$, where $\mathcal{N}$ is a single or multivariate normal distribution, with μ_j and Σ_j being the respective means and covariances of the different input features for state j .

When using the logistic regression HSMM, as described in Section 4.2.6, the emission probabilities were computed using the likelihood of each state from each one-versus-all LR model. Random sub-sampling for each state was performed to ensure that there was a balanced number of samples in each class of the one-versus-all LR models.

The parameters of the CRF models were optimised using gradient descent algorithms. As stated on page 336 of Sutton and McCallum [227], gradient descent algorithms are applied “off the shelf” for CRF optimisation. Because of its success when optimising CRFs and its

^rSee Equation 4.23.

computational efficiency [218], a limited memory quasi-Newton method (L-BFGS) [238] was employed in this work.

In order to optimise the regularisation parameter of the CRF, λ , five-fold cross-validation was employed within the training dataset. As regularisation of the CRF is not sensitive to small changes in λ (up to a factor of 10 - see page 334 of [227]), λ was varied in steps of 10^n , $n \in \{-1, 0, 1, 2, 3\}$. The λ value that resulted in the highest average F_1 score (see Section 4.2.9.3) over the five held-out sets was selected and used to train the CRF on the entire training dataset before evaluation on the separate test set.

In both HSMM models, the transition probabilities (a_{ij}) are set to zero, except for the possible transitions between successive states, which are set to 1 (see Section 4.2.5.2). The same applies to the probabilities calculated from the transition feature functions when using the CRF.

The six methods evaluated in this chapter were:

1. The MVN-based method, using a single homomorphic envelope feature as described by Schmidt et al. [99].
2. The MVN-based method, but using the Hilbert, PSD, wavelet and homomorphic features.
3. The LR-HSMM method using a single homomorphic envelope feature.
4. The LR-HSMM method using the Hilbert, PSD, wavelet and homomorphic features.
5. The duration-dependent CRF method using a single homomorphic envelope feature.
6. The duration-dependent CRF method using the Hilbert, PSD, wavelet and homomorphic features.

In addition, a comparison was made between the use of the standard Viterbi algorithm and of the extended Viterbi algorithm for each method listed above.

4.2.9.3 Model evaluation

The six segmentation methods were evaluated with respect to their ability to locate the S_1 and S_2 sounds accurately within the test set of recordings. The reference positions for the S_1 and S_2 sounds were the R-peaks and end-T-waves computed in Section 4.2.2, as shown in step *vii*

in Figure 4.2. An S_1 sound was labelled as correctly identified if the start of the segmented S_1 sound was found to be within 100 ms of the R-peak of the ECG. This tolerance is based on the recognised ECG R-peak detection tolerance of 150 ms [176], which was shortened to 100 ms as 150 ms is approximately the length of the fundamental heart sounds. Similarly, an S_2 sound was labelled as correctly segmented if the centre of this S_2 sound was found to be within 100 ms of the corresponding end-T-wave.

The performance of the segmentation algorithms was evaluated using the F_1 score, which is defined as

$$F_1 = \frac{2 \times P_+ \times Se}{P_+ + Se} = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (4.41)$$

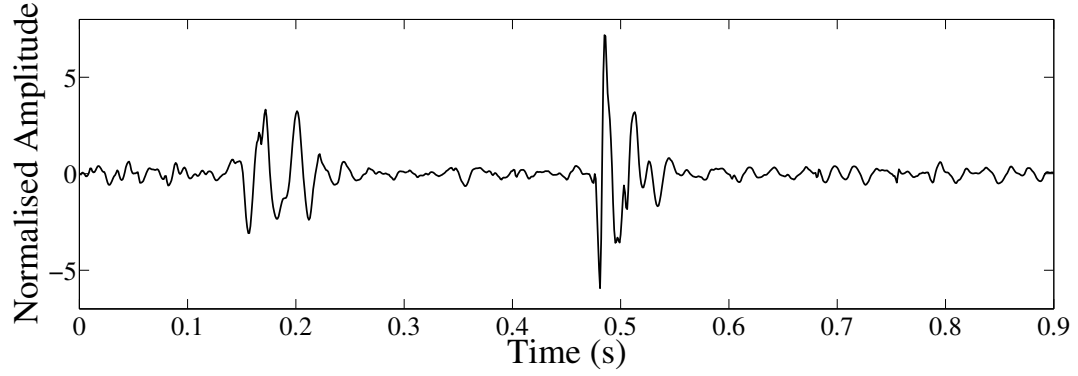
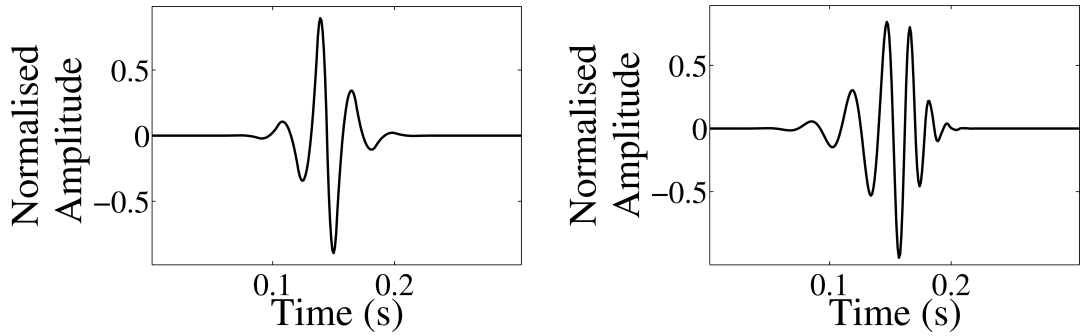
where Se is sensitivity (or recall - $Se = TP/(TP + FN)$) and P_+ is positive predictivity (or precision - $P_+ = TP/(TP + FP)$), while TP , FP and FN are the number of *true positives* (correctly identified sounds), *false positives* (extra falsely detected sounds) and *false negatives* (missed heart sounds), respectively. The F_1 score was used as it gives a single harmonic mean of Se and P_+ , which is symmetrically penalised by both FN and FP . Metrics such as accuracy ($Acc = TP/(TP + FP + FN)$) in such cases do not give an adequate representation of the results. However, Se , P_+ and Acc are reported to allow comparison to previous work.

In order to give a rigorous indication of segmentation performance, the process of splitting the data into a training and test set, the wavelet feature optimisation, the training of the HSMM and CRF and the evaluation on the test set were all repeated 100 times and the results averaged, as shown in step *vii* in Figure 4.2.

Significance testing was performed using a two-sided paired t-test on the 100 F_1 scores from the test datasets.

4.3 Results

The wavelet optimisation (see Section 4.2.8.3 and step *iv* in Figure 4.2) resulted in the reverse biorthogonal 3.9 or Daubechies 10 wavelet at decomposition level three being selected as the optimal wavelet for each of the 100 evaluation iterations. The shapes of these wavelets, as well as a typical heart cycle example, are shown in Figure 4.8.

(a) – Typical S_1 and S_2 sounds in a noisy PCG

(b) – The wavelet shapes for the reverse biorthogonal 3.9 wavelet (left) and the Daubechies 10 wavelet (right).

Figure 4.8 – Comparison of the shapes of typical S_1 and S_2 sounds with the selected reverse biorthogonal 3.9 and Daubechies 10 wavelets. Figure 4.8a shows a single heart cycle, with typical S_1 and S_2 sounds. These sounds can be seen to be similar in shape to the wavelets in Figure 4.8b. For this reason, the wavelet detail coefficients at instances of overlap between the heart sounds and the mother wavelet will be of high amplitude.

The *gross* performance results of the six algorithms under consideration on both the training and test sets, using the different features and Viterbi algorithms, are presented in Table 4.4. This table illustrates the scores for the combined S_1 and S_2 sounds, while also presenting the F_1 scores for each sound separately to give an indication of performance on the different sounds. These gross scores were calculated on a per-patient basis, summing the total number of sounds for each patient in the training and test datasets, calculating the different metrics for each patient, then averaging over patients in each of the training and test sets. The results over the 100 iterations were then averaged. The standard deviation of the average results over the 100 evaluation iterations is also shown.

The results of the state-of-the-art algorithm [99], can be seen in the first two lines of the table (algorithm 1a in Table 4.4). This algorithm achieved an average F_1 score of $86.28 \pm 1.55\%$ on the test datasets. The results of the proposed algorithms, which combined the Hilbert envelope, PSD, wavelet envelope and homomorphic envelope features with the LR-HSMM classifier and

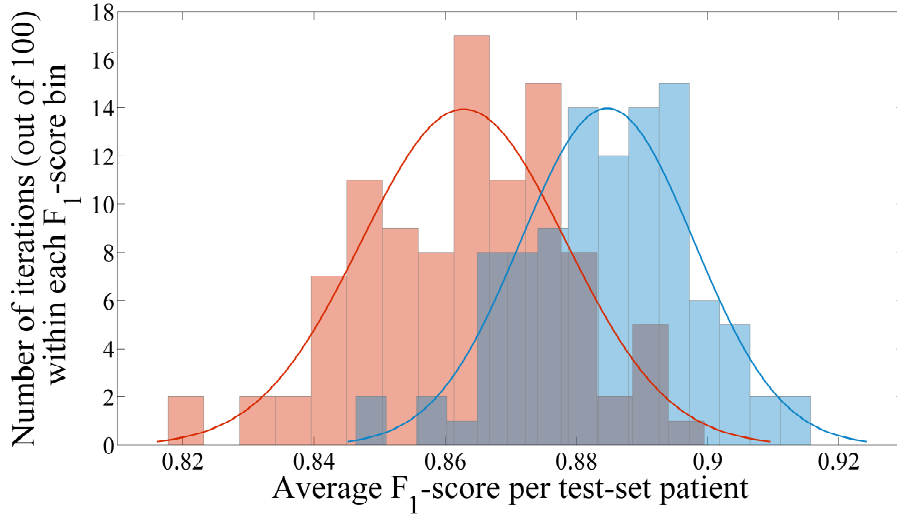


Figure 4.9 – Distributions of F_1 scores for the 100 iterations of algorithms $1a$ and $3a$ from Table 4.4. The distribution in red on the left shows the F_1 scores from the MVN-HSMM (algorithm $1a$), while the blue distribution on the right of the figure shows the F_1 scores from the LR-HSMM model (algorithm $3a$), both using the single homomorphic envelope feature. The purple region shows the overlap between the two distributions. A two-sided paired t-test showed these results to be significantly different ($p < 0.001$).

duration-dependent CRF and extended Viterbi algorithm, are shown as algorithms $4b$ and $6b$ in Table 4.4. These algorithms attained an average F_1 score of $95.63 \pm 0.85\%$ and $95.64 \pm 0.92\%$, respectively.

The incorporation of LR into the HSMM model alone resulted in a significant improvement of F_1 score between the state-of-the-art algorithm ($1a$, $86.28 \pm 1.55\%$) and the LR-homomorphic algorithm with the standard Viterbi algorithm ($3a$, $88.47 \pm 1.31\%$), ($p < 0.001$). The distributions of the 100 F_1 scores for each of these algorithms can be seen in Figure 4.9. A significant improvement was also seen between the state-of-the-art algorithm ($1a$, $86.28 \pm 1.55\%$) and the CRF model with the homomorphic feature ($5a$, $88.38 \pm 1.34\%$), ($p < 0.001$).

The introduction of the Hilbert, PSD and wavelet features, in addition to the homomorphic feature, resulted in a drop in performance when using the MVN-based HSMM ($2a$, $84.02 \pm 1.55\%$ compared to $1a$, $86.28 \pm 1.55\%$). However, the addition of these same features resulted in a significant improvement in the performance of the LR-based HSMM ($p < 0.05$, comparing the F_1 scores of algorithms $3a$ to $4a$ and $3b$ to $4b$), as well as in the performance of the CRF ($p < 0.05$, comparing the F_1 scores of algorithms $5a$ to $6a$ and $5b$ to $6b$).

The small improvement when using the CRF model with four features (algorithm $6b$) when compared to the LR-HSMM model (algorithm $4b$) was not found to be significant ($p = 0.68$).

Table 4.4 – Gross results of segmentation algorithms on train (grey) and test sets (%) using various input features, averaged across all 100 iterations. The first line shows the previous state-of-the-art. Algorithms 4b and 6b show the proposed methods, while the best results for each metric are shown in **bold**.

Algorithm	Features	Viterbi Algorithm	Se	P ₊	Ac	F ₁ ^{s1}	F ₁ ^{s2}	F ₁
1	MVN Homomorphic	Standard	85.09 ± 1.31	86.17 ± 1.27	77.49 ± 1.68	85.98 ± 1.39	85.34 ± 1.32	85.62 ± 1.29
		[23]	85.74 ± 1.57	86.85 ± 1.53	78.35 ± 2.06	86.58 ± 1.56	86.07 ± 1.70	86.28 ± 1.55
		Extended	94.29 ± 0.89	94.61 ± 0.86	90.61 ± 1.35	96.15 ± 0.85	92.73 ± 1.13	94.45 ± 0.88
b			94.66 ± 0.90	94.93 ± 0.88	91.15 ± 1.36	96.50 ± 0.84	93.06 ± 1.17	94.79 ± 0.89
2	MVN Hilbert, PSD, Wavelet, Homomorphic	Standard	83.08 ± 1.48	83.84 ± 1.45	74.05 ± 1.91	83.67 ± 1.52	83.32 ± 1.57	83.45 ± 1.47
			83.62 ± 1.57	84.45 ± 1.53	74.82 ± 1.99	84.20 ± 1.56	83.93 ± 1.72	84.02 ± 1.55
		Extended	92.74 ± 1.05	92.93 ± 1.04	87.90 ± 1.57	94.28 ± 1.07	91.37 ± 1.23	92.83 ± 1.04
b			93.10 ± 1.04	93.29 ± 1.03	88.46 ± 1.57	94.67 ± 1.02	91.70 ± 1.29	93.19 ± 1.04
3	LR Homomorphic	Standard	87.51 ± 1.29	88.72 ± 1.23	80.97 ± 1.68	87.82 ± 1.29	88.45 ± 1.38	88.10 ± 1.26
			87.87 ± 1.34	89.10 ± 1.28	81.49 ± 1.77	88.11 ± 1.36	88.91 ± 1.43	88.47 ± 1.31
		Extended	94.88 ± 0.82	95.28 ± 0.79	91.66 ± 1.25	96.70 ± 0.80	93.43 ± 1.04	95.08 ± 0.80
b			95.22 ± 0.88	95.60 ± 0.85	92.18 ± 1.35	97.00 ± 0.86	93.79 ± 1.12	95.40 ± 0.86
4	LR Hilbert, PSD, Wavelet, Homomorphic	Standard	87.58 ± 1.28	89.16 ± 1.18	81.22 ± 1.66	87.97 ± 1.32	88.80 ± 1.31	88.35 ± 1.23
			87.84 ± 1.35	89.44 ± 1.25	81.62 ± 1.76	88.18 ± 1.37	89.13 ± 1.42	88.62 ± 1.30
		Extended	95.09 ± 0.80	95.69 ± 0.75	92.11 ± 1.21	96.70 ± 0.84	94.05 ± 0.97	95.38 ± 0.78
b			95.34 ± 0.88	95.92 ± 0.83	92.52 ± 1.33	96.95 ± 0.90	94.29 ± 1.08	95.63 ± 0.85
5	CRF Homomorphic	Standard	87.67 ± 1.34	88.67 ± 1.28	81.14 ± 1.76	87.71 ± 1.33	88.66 ± 1.45	88.16 ± 1.31
			87.87 ± 1.37	88.90 ± 1.30	81.48 ± 1.80	87.88 ± 1.37	88.93 ± 1.47	88.38 ± 1.34
		Extended	94.90 ± 0.83	95.27 ± 0.80	91.68 ± 1.27	96.70 ± 0.81	93.43 ± 1.06	95.08 ± 0.81
b			95.11 ± 0.97	95.47 ± 0.91	92.03 ± 1.44	96.93 ± 0.90	93.62 ± 1.21	95.29 ± 0.94
6	CRF Hilbert, PSD, Wavelet, Homomorphic	Standard	88.34 ± 1.33	89.26 ± 1.27	82.02 ± 1.75	88.05 ± 1.33	89.61 ± 1.44	88.79 ± 1.30
			88.50 ± 1.36	89.43 ± 1.29	82.27 ± 1.79	88.15 ± 1.35	89.82 ± 1.48	88.95 ± 1.33
		Extended	95.37 ± 0.79	95.75 ± 0.75	92.46 ± 1.21	97.00 ± 0.79	94.10 ± 1.02	95.56 ± 0.77
b			95.46 ± 0.94	95.83 ± 0.90	92.62 ± 1.44	97.08 ± 0.93	94.19 ± 1.15	95.64 ± 0.92

There was a significant improvement in the F_1 score when using the extended Viterbi algorithm compared to the standard Viterbi algorithm on all combinations of features and when using the MVN-HSMM, LR-HSMM or CRF models ($p < 0.001$). This extension of the Viterbi algorithm resulted in at least a 6% improvement in F_1 scores across all algorithms.

The average performance results for algorithms 1a, 4b and 6b, averaged over *all recordings* in the training and test sets, rather than by patient, can be seen in Table 4.5. The average number of S_1 and S_2 sounds across the 100 iterations can be seen, as well as the TP , FN and FP numbers. The accuracy and F_1 score performance metrics are shown. A significant improvement in F_1 score over the state-of-the-art algorithm can be seen when using the LR-HSMM methods ($88.61 \pm 1.33\%$ to $96.49 \pm 0.64\%$) ($p < 0.001$) as well as the CRF model ($88.61 \pm 1.33\%$ to $96.54 \pm 0.70\%$) ($p < 0.001$).

While there was no significant difference between the LR-HSMM model (algorithm 4b) and the CRF model (6b) when comparing the results on a per-patient basis in Table 4.4, there was a significant difference between these methods when comparing them across all the sounds in the database, as shown in Table 4.5 ($p < 0.05$).

An illustration of the segmentation accuracy of the LR-HSMM method on a noisy test recording is shown in Figure 4.10. This patient has a large systolic murmur, while the PCG is also corrupted by various noise sources. The segmentation algorithm in this example can be seen to ignore the pathological murmurs. This is intentional at this stage, as the classification of systolic and diastolic murmurs is performed subsequently to segmentation, once the exact positions of the S_1 and S_2 sounds have been computed.

4.4 Discussion

The optimal wavelets selected for all 100 evaluation iterations (reverse biorthogonal 3.9 and Daubechies 10 at decomposition level three) agree with intuition, as their centre frequencies were between 65 and 85 Hz, as expected for the FHSs [92]. Limiting the decomposition of the PCG signal to these frequencies emphasises the S_1 and S_2 sounds, while filtering higher frequency noise due to clicks, murmurs and high-frequency noise. In addition, their shapes are similar to those of the fundamental heart sounds, as shown in Figure 4.8.

Table 4.5 – Average results of algorithms 1a, 4b and 6b from Table 4.4 when averaged across all recordings rather than per patient. These show the numbers and results averaged across all 100 iterations for the train (*grey*) and test sets. The average number of S_1 and S_2 sounds is shown ($N(S_1)$ and $N(S_2)$), as well as the numbers of true positives (TP), false negatives (FN), false positives (FP), accuracy (Acc) and F_1 score.

Algorithm	Features	Viterbi Algo- rithm	$N(S_1)$	$N(S_2)$	TP	FN	FP	Acc (%)	F_1 (%)
1a	MVN	Homomorphic	6117.07	6012.56	10613.02	1516.61	1380.26	78.56 ± 1.69	87.98 ± 1.06
		Standard	5686.93	5597.44	9945.63	1338.74	1214.19	79.58 ± 2.15	88.61 ± 1.33
4b	LR	Hilbert, PSD, Wavelet, Homomorphic	6117.07	6012.56	11649.84	479.79	410.92	92.90 ± 1.08	96.31 ± 0.58
		Extended	5686.93	5597.44	10858.53	425.84	364.4	93.22 ± 1.19	96.49 ± 0.64
6b	CRF	Hilbert, PSD, Wavelet, Homomorphic	6117.07	6012.56	11689.40	440.23	400.72	93.29 ± 1.10	96.53 ± 0.59
		Extended	5686.93	5597.44	10876.88	407.49	372.8	93.31 ± 1.32	96.54 ± 0.70

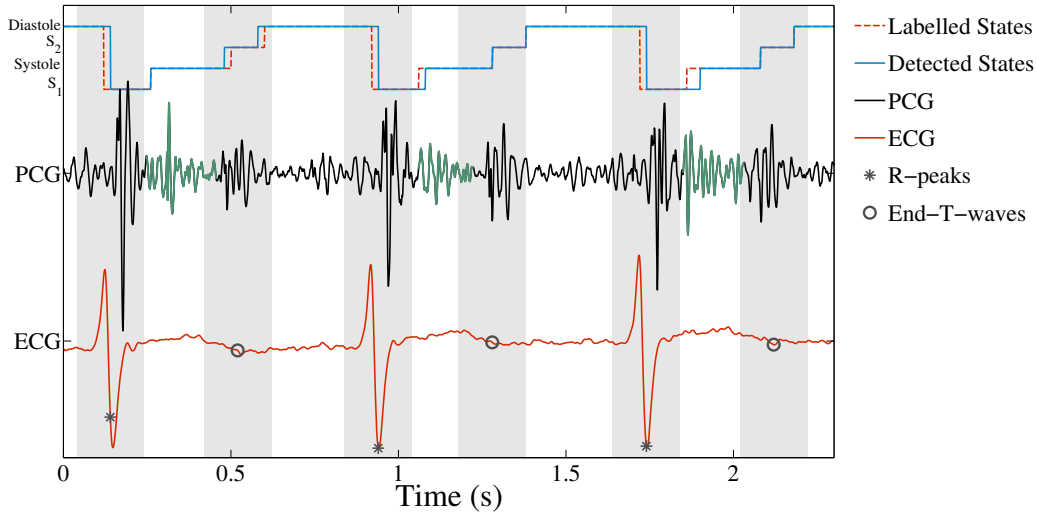


Figure 4.10 – Example of a correctly segmented, noisy, PCG using the LR-HSMM method with a clean ECG for reference. The four LR-HSMM-derived states of the heart cycle (S_1 , systole, S_2 and diastole) are shown at the top of the plot in blue, over the ECG-labelled states in red. The noisy PCG, shown in black, contains a systolic murmur with a mid-systolic click due to MVP, shown in green. The ECG is shown in red at the bottom of the plot, with the positions of the detected R-peaks (*) and the end-T-waves (o) are indicated. The grey shading indicates the 100ms tolerance window around the R-peaks and end-T-waves for detection of the start of S_1 and mid-point of S_2 . All the sounds are correctly identified, while the murmurs are not confused as being heart sounds.

The results in Tables 4.4 and 4.5, with similar results between the training and test set results when using balanced datasets, illustrate that the algorithms were not over-trained on the training data. The small values for the standard deviations in these tables indicate consistent results for the segmentation algorithms across the 100 evaluation iterations. The standard deviations can be seen to be smaller when using the extended Viterbi algorithm, indicating that the use of this algorithm results in more consistent results.

Comparing the results of the method developed by Schmidt et al. [99] with the literature review table (Tables 4.1 on page 91) and the first results table (Table 4.4 on page 133) (98.8% Se and 98.6% P_+ in Table 4.1 and 85.74% Se and 86.85% P_+ in Table 4.4) shows a decrease in Se and P_+ on our dataset when comparing test set results. This indicates that our data may be more noisy or heterogeneous, and therefore more difficult to analyse. The few pathological cases in Schmidt et al.'s [99] dataset (8% of patients as compared to 33% of patients in our dataset) may also account for this. In addition, Schmidt et al. [99] manually annotated their data and excluded data which could not be manually segmented. In contrast, the ECG was used to label heart-sound recordings automatically in our dataset. Therefore, noisy heart-sound recordings, for which manual annotation may have been impossible, may be included in our dataset.

The significant improvement in performance between the state-of-the-art algorithm ($1a$, $86.28 \pm 1.55\%$) and the LR-HSMM algorithm with the homomorphic feature and standard Viterbi algorithm ($3a$, $88.47 \pm 1.31\%$) indicates that the simple introduction of LR as opposed to the MVN-based emission probabilities significantly improves segmentation performance. This is thought to be due to the higher discrimination between states afforded by the LR model. In addition, when using a single homomorphic feature with the CRF model, a significant improvement on the state-of-the-art was seen ($5a$, $88.38 \pm 1.34\%$). Therefore, the introduction of the CRF model can be seen to improve segmentation performance when compared to the MVN-based HSMM.

The introduction of the three additional features resulted in a decrease in performance when using the MVN-based HSMM (comparing algorithm $1a$ to $2a$ and $1b$ to $2b$), probably because the MVN distribution is unable to model the higher dimensional feature space effectively. A similar drop in performance was noted when adding frequency-based features in [99]. However, the introduction of these features when using the LR-HSMM model and CRF resulted in a slight, yet significant improvement in performance in both cases. This combination of features was thought to enhance the segmentation as they contribute both information about the amplitude of the PCG and frequency information, in the form of the wavelet transform and PSD feature.

While the new segmentation models and enhanced features led to improved segmentation performance, by far the largest contributor to the improved segmentation performance was the introduction of the extended Viterbi algorithm. This modification resulted in significant improvement across all algorithms in Table 4.4 (when comparing algorithms $1a$ to $1b$, $2a$ to $2b$ and so on), where at least a 6% increase in the F_1 score can be seen. The weakness of the standard Viterbi algorithm was especially noted in short recordings with a short diastolic period before the onset of the first heart sound. In such cases, as the probability of such a short diastole at the start of the recording was low, the algorithm often assigned this period to a systolic state. Then, the first heart sound was often wrongly identified as being a second heart sound. Such an error is shown in Figure 4.11, where algorithm $4a$ (LR-HSMM with standard Viterbi algorithm) was applied to a recording with a short diastolic region before the first heart sound.

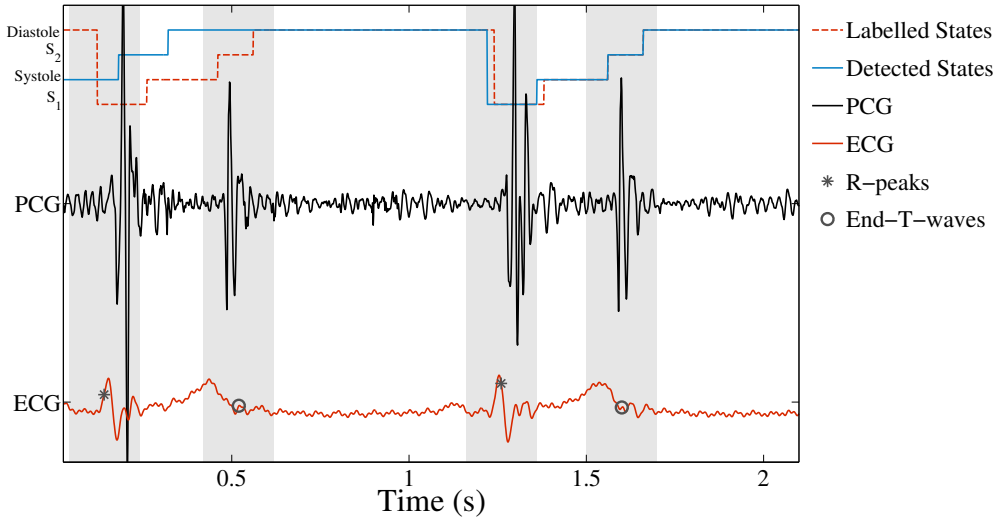


Figure 4.11 – Example of the shortcomings of the standard Viterbi algorithm. This record has a short diastolic period before the first heart sound. When using the standard Viterbi algorithm, the probability of such a short diastole is very low. Therefore, this model erroneously assigns systole to this first period, thereby incorrectly assigning S_2 to the first heart sound. The duration constraints of diastole then cause the second heart sound to be missed. The second cycle is correctly segmented.

When using the extended Viterbi algorithm, $F_1^{S_1}$ scores were higher than $F_1^{S_2}$. This can be attributed to the patients with MVP in this dataset having systolic murmurs that often mask S_2 , making it more difficult to detect. The detection of the end-T-waves, and their approximate relationship with the S_2 sound, may also contribute to lower $F_1^{S_2}$ scores.

There was an improvement in performance between the average results in Table 4.5 when compared to the gross results of Table 4.4. The lower performance of the gross averages, which are derived by averaging per patient, indicate that specific patients (which may have fewer heart sounds when compared to others in the dataset) diminish the performance of the segmentation algorithms. This may be due to these patients having recordings contaminated by noise or having murmurs that mask the positions of the fundamental heart sounds.

While the inclusion of explicit timing durations helps to improve the differentiation of heart sound-like noise from noisy heart sounds, these timing durations can also be a shortcoming of the model. These duration-dependent models work well in recordings with regular heart beats. However, they may perform poorly in recordings with ectopic beats or arrhythmia, where the durations between heart sounds may vary considerably. Such a case is shown in Figure 4.12. In this example of a very noisy PCG, a premature beat can be seen at 2.5s. Such a beat, along with noise spikes, results in erroneous segmentation.

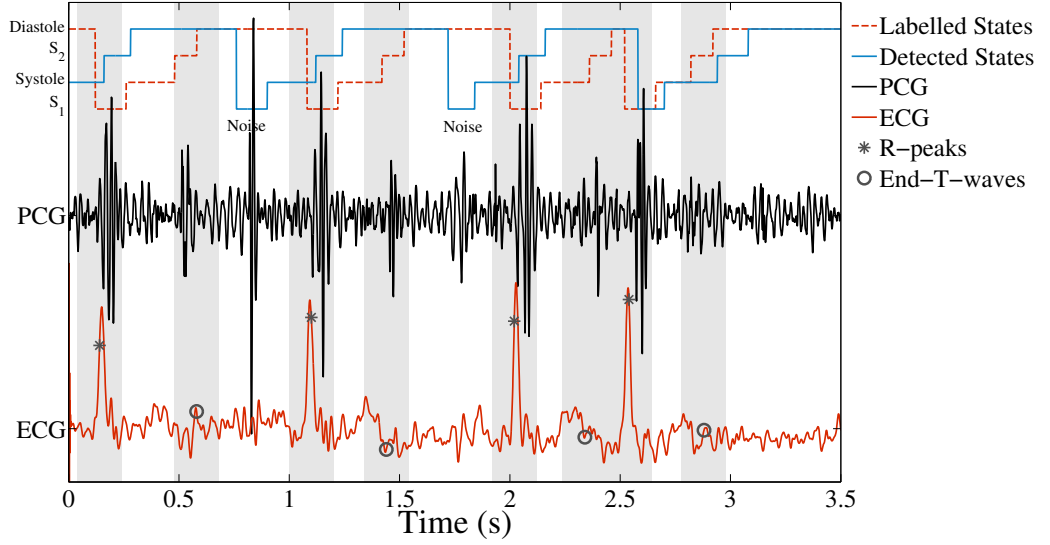


Figure 4.12 – Example of an incorrectly segmented arrhythmic PCG using the LR-HSMM method. The PCG in this case can be seen to be noisy, with a premature beat at 2.5s. The first heart sound, which is an S_1 sound, is erroneously labelled as S_2 . The noise spike at 0.75s is erroneously labelled as S_1 . The premature beat at 2.5s also causes confusion between the heart sounds between 1.75 and 3 seconds. While the ECG in this case is very noisy, the R-peaks and end-T-waves can still be seen to be correctly identified (while the first end-T-wave is slightly late).

The results presented in this chapter can be seen as a comparison between generative (MVN-HSMM), hybrid (LR-HSMM) and discriminative (CRF) probabilistic models for PCG segmentation. If only the homomorphic envelope is considered (as this leads to the best-performing generative model), the hybrid model leads to the highest segmentation performance, followed by the discriminative model. While the generative model has the lowest segmentation performance, it has other advantages. For instance, the trained generative model could be used to *generate* a likely sequence of observations. Secondly, in the absence of labels in the training data, a generative model can be used in an unsupervised manner, by using the Baum-Welch algorithm to train the model [210]. Finally, while the discriminative model outperformed the generative model in this case, it is impossible to predict which model will perform better in advance [227]. Ng and Jordan argue that generative models may be more likely to outperform discriminative models on smaller datasets [214].

4.5 Conclusion

The work presented in this chapter has investigated new methods for the segmentation of the S_1 and S_2 heart sounds from a single channel PCG recording with no external reference. These

segmentation methods have used generative, hybrid and generative probabilistic models. The major contributions of this chapter were:

- An exploration of features for heart sound segmentation, including the Hilbert envelope, a PSD envelope and a wavelet envelope, resulting from the robust selection of the wavelet family and decomposition level. These were added to the homomorphic envelope already employed by the state-of-the-art algorithm.
- The addition of LR into the HSMM, resulting in a hybrid generative-discriminative model.
- The implementation of duration-dependent CRFs for heart sound segmentation.
- The use of the extended duration-dependent Viterbi algorithm.
- The rigorous evaluation of the above approaches on one of the largest published datasets, including a comparison with the state-of-the-art segmentation algorithm.

The introduction of the LR-HSMM and duration-dependent CRF, each with the four investigated features and the extended Viterbi algorithm, resulted in the best performing segmentation algorithms, both of which showed a significant improvement on the state-of-the-art algorithm. As demonstrated on this dataset, which consisted of a large proportion of pathological recordings, these methods are able to segment the heart sounds accurately in noisy, real-world, PCGs with murmurs and other abnormal sounds.

Chapter 5

Systolic murmur classification in phonocardiograms

“What matters is achieving performance sufficient to make an impact on the world.” – K. L. Wagstaff, “Machine learning that matters”.

The final step in the automatic detection of rheumatic heart disease (RHD) in mobile phone-based heart sound recordings, as shown in Figure 1.14, is to detect the abnormal heart sounds caused by functional valve or structural heart damage. As outlined in Section 1.3.3, RHD leads to damage of the heart valves which control the flow of blood. Such damage leads to regurgitation, a leaking heart valve, or stenosis, which is the narrowing of a heart valve due to scarring. An example of a pronounced systolic murmur due to mitral valve regurgitation is shown in Figure 5.1. As outlined in Section 1.4, a low-cost mobile phone-based diagnostic tool for RHD has the potential to allow mass screening of young adults and children for RHD, facilitating the early detection of the condition, and the initiation of the necessary penicillin treatment to avoid serious heart damage.

Importantly, the presence of a murmur may not necessarily indicate pathology. As outlined in Section 1.3.4.4, innocent murmurs, which are common in children and adolescents, are sounds that are similar to pathological murmurs, but which occur due to the natural flow of blood and subsequent vibrations of the cardiac structure [90].

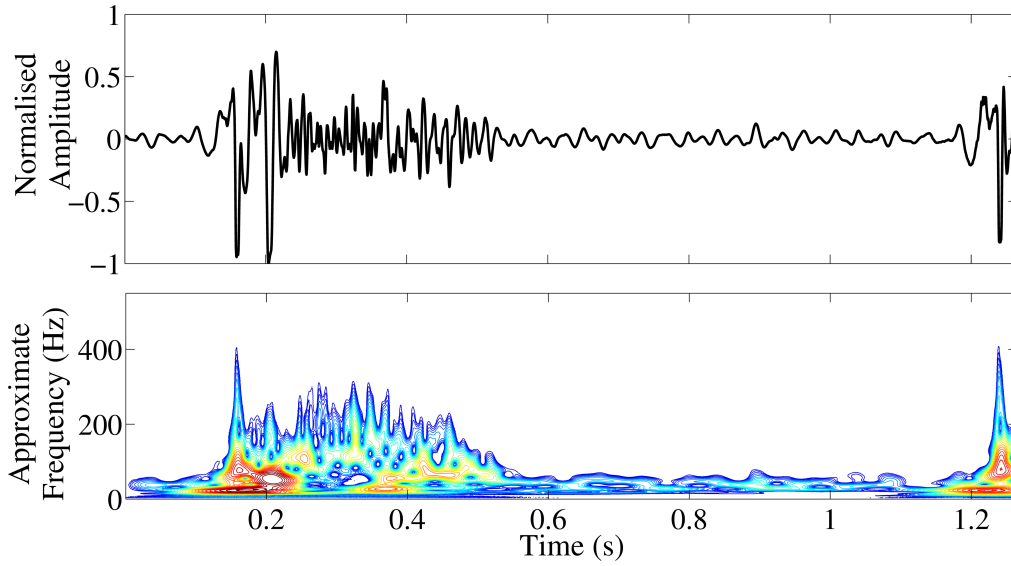


Figure 5.1 – Example of a pronounced systolic murmur due to mitral valve regurgitation. The top plot shows the PCG, while the bottom plot shows the frequency content (derived using the continuous wavelet transform, and hence, the “approximate frequency” is shown). The S_1 sound can be seen at 0.2 seconds, while the S_2 sound, at 0.4 seconds, is obscured by the high-frequency murmur, spanning S_1 and S_2 , with the main frequency content ranging between 50 and 350 Hz.

This chapter presents algorithms for discriminating between normal and diseased heart sound recordings. In particular, it focuses on the detection of systolic murmurs indicative of rheumatic heart disease (RHD) and mitral valve prolapse (MVP). Recordings from three different devices, including a mobile phone and two electronic stethoscopes, were analysed for the presence of such systolic murmurs. Time-frequency-based features, which were used to emphasise the high-frequency characteristics of pathological murmurs, were extracted from the systolic periods of the PCGs. The features were used for the automatic classification of normal and pathological heart sounds. The effects of denoising heart cycles via averaging, excluding noisy heart cycles, and heart cycle selection based on clustering were all explored.

The principal contributions of this chapter to the field of pathology classification in PCG recordings are:

- An extensive literature review of heart sound classification methods.
- The rigorous implementation of support vector machines for heart sound classification.
- The assessment of novel cycle averaging, cycle exclusion and k -means-based cycle selection techniques to improve classification accuracy.

- The use of the undecimated wavelet transform and mel-frequency cepstral coefficients, extracted from clinically-relevant segments of systole, to extract discriminating features.
- The classification of pathology in three separate real-world datasets, each being larger than the majority of the datasets presented in the literature review.
- The classification of RHD in mobile phone-based PCG recordings.

5.1 Background

The automated classification of pathology in heart sound recordings has been performed for over 50 years. The prominent works in the field of automatic classification of stethoscope-recorded heart sounds are summarised in Table 5.1. This table outlines the size of the dataset used, the features used, the classification algorithm, and important details about the evaluation of the method, such as whether the data were split into training and test sets and if this split was patient-based, such that recordings from the same patient did not end up in both training and test datasets. For brevity, only the notable studies, with rigorous validation methods (a train-test split) and sizeable datasets, are summarised in detail below, grouped by their method of classification.

5.1.1 Early classification methods

Gerbarg et al. [239] were the first researchers to attempt the automatic classification of pathology in PCGs, motivated by the need to identify children with RHD. They extracted three features based on the power of the raw PCG. A threshold-based classification algorithm was then optimised on hand-picked recordings, which achieved 91% sensitivity and 80% specificity when discriminating between 66 healthy and 21 pathological patients. Two years later, Durnin et al. [240], reported the results of “Phonocardioscan” (an automated auscultation machine). This device achieved 97.8% sensitivity (Se) and 93.1% specificity (Sp) when automatically diagnosing heart pathology in a field test of 708 (270 pathological) children. However, Klatsky et al. [241] evaluated the same diagnostic system on adults, and only achieved 75% Se and 66% Sp , highlighting that the device was not yet ready for full-scale screening.

Table 5.1 – Summary of previous research on heart murmur classification in PCG recordings. Ac denotes accuracy, Se sensitivity, Sp specificity

Authors	Dataset	Features	Classification Method	Acc (%)	Se (%)	Sp (%)	Separate test set	Patient-based partition	Notes on Validation
[239]	66 healthy and 21 children with murmurs due to RHD	Signal amplitude and PSD	Rule-based decision based on thresholds in 3 systolic features		91	80	✗		Hand-picked data
[240]	130 children (66 healthy) used for training, 708 children (270 with valvular heart damage) used for testing	Proprietary features-“Phonocardiogram” device	Proprietary - “Phonocardiogram” device		97.8	93.1			Unclear analysis method
[241]	456 adults (143 pathological)	Proprietary features-“Phonocardiogram” device	Proprietary - “Phonocardiogram” device		75	66			Unclear analysis method
[242]	2637 children (2583 healthy)	Rectified amplitude in 6 windows over heart cycle	Rule-based classification based on relative amplitude to S_1		96.5	92.4	✗		Algorithm optimised on 20% of dataset, included in test set
[243]	69 cycles (of unknown pathology)	8 STFT coefficients from 32 overlapping windows 128 samples from Hilbert envelope Statistical moments of wavelet features from diastole (averaged over 10 cycles) & physical examination features (age, sex, body weight, blood pressure)	Neural network (NN)	88.9-100					Unclear evaluation
[244]	112 adults (70 with CAD). 10 cycles per patient, recorded in sound-proof room	Mean 2nd level Db10 DWT within systole	Fuzzy neural network	86	85	89	✓	✓	30 (balanced) training, 82 test recordings
[146]	61 children (33 with heart disease: ASD, VSD, PS)	Moving-average filtered maximum value of PCG within 6ms windows.	MLP Neural network	93			✓	✓	31 train (14 normal), 30 test (14 normal).
[122]	61 cycles (from 4 normal and 8 pathological patients, representing 5 different pathologies)	Wavelet features (db20) - reduced to four features using covariance matrix eigenvalue analysis	Neural network	100			✓	✓	48 cycles in training set, 13 cycles in test
[245]	Unspecified size. Patients with both functional and dysfunctional native and prosthetic heart valves	6th level Sym8 wavelet packed decomposition, with 8 frequency bands. Wavelet values averaged across cycles	kNN: native valves prosthetic valves	100			✓	✓	Leave-one-out cross-validation (CV)
[235]	85 children (45 pathological). 7-12 second PCGs	Subsampled trimmed-mean spectrogram. Resulted in 81 features	Neural network	85	80	90	✓	✓	65 train (35 pathological) and 20 test
[246]	111 paediatric patients (72 pathological). Excluded noisy data	FFT of entire recording. 1Hz resolution between 0-210 Hz	Probability neural network		97.3	94.4	✓	✓	56 training (Unbalanced: 35 pathological) and 55 test (37 pathological). Subsampling of features optimised on test set
[247]	69 children (37 pathological). Hand-picked 3 cycles per recording	Median continuous wavelet transform value in mid-systole	Neural network		100	100	✓	✓	Leave-one-out CV. Optimised NN and decision threshold on entire set
[84]	193 subjects (194 recordings, 99 pathological). 15 cycles per patient. Hand-picked 184 recordings		Threshold-based classification		93	75	✗		Hand-picked data

[232]	28 teaching-grade recordings, 4 per class (6 pathology and normal). 12 cycles per patient	2nd level Db2 wavelet. Each cycle split into 32 sub-windows, with sum on squared wavelet coefficients extracted per window. Divergence analysis to select 16 best features	7-class GAL Neural network	99.4	✗		
[248]	84 hand-picked teaching-quality recordings (41 with AS and 43 with MR)	100 features (cycle duration, S_1 & S_2 peak values, mean frequency content & normalised average Shannon envelope in systole & diastole)	Decision tree	90	✗	Feature selection and classification optimised on entire dataset	
[249]	5 patients (each with a different pathology). 4 cycles per patient	Coif4 7th level wavelet decomposition (256 features per patient)	5-class Neural network	100	✗		
[250]	241 hand-picked cycles from paediatric patients (153 pathological)	FFT of entire cycle - 253 bins with 1 Hz resolution	Neural network	88	83	✓	188 train, 53 test. Unbalanced training set. Hand-picked data
[251]	109 hand-picked adult recordings (47 with aortic stenosis)	Peak frequency and duration of murmur in averaged cycles	Manual analysis		✗		No objective results reported
[252]	One hand-picked cycle from 206 patients (47 with aortic stenosis).	FFT, STFT, Wavelet transform (Mexican hat) and Shannon entropy over 12 windows per cycle	Discriminant function analysis	89.4	93.7	✗	No feature normalisation
[253]	36 hand-picked PCGS from adults (6 with mitral regurgitation, 23 with aortic stenosis). 12 cycles per recording	Time (Shannon envelope), Time-frequency (Wavelet -db2 and S-transform), eigenvalues, higher order statistics (bispectrum), state space, fractal dimension and recurrence quantification analysis-based features - 207 features averaged over cycles. Pudil's sequential floating forward selection method for feature selection.	3-class neural network	86	93	✓	Leave-one-out CV. Unbalanced dataset
[254]	257 (188 pathological) 40-45 second recordings from children	STFT-derived features within systole	Stepwise logistic regression	95	72	✗	Feature selection and classification optimised on entire dataset
[126]	163 hand-picked PCGs from paediatric patients (50 pathological murmur). 6 hand-picked cycles per patient	12 Mel-spaced frequency bins (20-420 Hz), in 10 equally spaced windows over the heart cycle	Neural network	90	94.46	✓	Leave-one-out CV. Unbalanced dataset
[125]	340 cycles from 41 recordings (124 systolic, 106 diastolic murmurs)	Same as [232]	3-class GAL neural network; 3-class MLP-BP neural network;	97		✓	Split 340 cycles into 3 sets, each of 112 cycles. 45 train, 67 test.
[108]	125 hand-picked paediatric PCGs (50 innocent vibratory murmurs, 25 innocent ejection murmurs and 50 easily confusable pathological cases). Manually selected 3 cycles per patient	STFT, with manual estimation of volume, duration and peak frequency of murmurs	Threshold-based classification	90	91	✗	
[237]	60 cycles recorded from 20 patients (10 with mitral valve prolapse)	Coif4 wavelet. Extracted 64 features per cycle	Neural network	95.5		✓	40 training cycles, 200 testing cycles. Only reported results on subset
[120]	39 patients (11 with mitral regurgitation). 30-40s PCGs	Db5 wavelet levels 6-9. Cycles merged using k NN clustering	Threshold-based classification	85.7	94.1	✗	Only used subset of data
[255]	31 adult males (14 with various pathologies). 4 cycles from 5 auscultation sites	Amplitude and frequency of S_1 , S_2 windows of systole and diastole	Neural network	85.7	94.1	✗	Feature selection and classification optimised on entire dataset. Features not normalised

[256]	41 adults (4 diastolic, 6 continuous and 11 systolic murmurs). In addition, teaching manual PCGs, leaving 1398 cycles.	Signal amplitude, STFT and MFCC	One unsupervised HMM per class	97	96	✗	
[257]	80 hand-picked cycles from 41 patients (20 for each normal, systolic, diastolic and continuous). In addition, 1381 simulated cycles of various conditions	Signal amplitude, STFT and MFCC	One unsupervised HMM per class	99.21		✗	
[258]	28 teaching-quality recording (2 per pathology). 20 cycles per PCG	Sym18 wavelet at 5 level. Systole and diastole split into 5 equally-spaced windows. Divergence analysis for feature selection	14-class neural network	95		✓	50-50 train-test split
[259]	20 teaching-quality PCGs (15 with systolic murmurs)	"Simplicity"-based features, derived from the singular value decomposition of the embedding matrix	Fuzzy clustering	80	73	100	✗
[260]	120 recordings (72 teaching-quality with valvular heart disease) of unknown duration	Mean and standard deviation of indices of windows with greatest wavelet (Db10) packet decomposition energy	Polynomial SVM	96.67	100		✓
[198]	90 paediatric subjects - 54 pathological (VSD, PDA, TOF, ASD, AS, PS). 10 seconds per recording	Frequency bands within PSD of whole recording	Neural network	95	93.33		✓
[139]	832 cycles from 104 patients (normal and 12 pathological conditions). 8 cycles per patient	Same as [232] Feature ranking using QRcp and F-ratio	13-class neural network	99.3			✓
[233]	372 hand-picked cycles (272 pathological - AR, AS and PS)	Mean and standard deviation of 5 levels of db4 wavelet decomposition over whole cycle	4-class neural network	94.4			✓
[140]	360 hand-picked cycles (50% with murmur) from 148 patients (50 normal). No echo-confirmed diagnosis	Three feature sets: a) Time-frequency features, based on Teager's energy, STFT, GT, WVD and WT b) Perceptual features based on MFCCs c) Fractal features based on the correlation dimension, largest Lyapunov exponent and Hurst exponent	Binary k NN classifier ($k=9$) (feature set c)	99.73			✗
[236]	28 teaching-quality recording (2 per pathology). 20 cycles per PCG	a) Sym18 wavelet at 5 level. Systole and diastole split into 5 equally-spaced windows b) Power of PCG in 25 equally-spaced windows	14-class neural network	99			✓
				95			✗
							50-50 train-test split.

[261]	198 teaching-quality recordings (38 normal, 41 with AS, 43 with MR, 38 with AR and 38 with MS) from unknown number of patients	Shannon energy of in windows of S_1 , systole, S_2 and diastole. Total energy in four frequency bands (50-250, 100-300, 150-350, 200-400 Hz) from systole and diastole	SVM. Binary classification: Systolic vs diastolic murmur. AS vs MR: AR vs MS:	91.43 87.5 94.74	✗	All healthy recordings from different device. No feature normalisation. Hand-picked C value for SVM ($C=\infty$). γ optimised on test set
[234]	64 patients (consisting of normal and 5 pathologies in unknown numbers). 8 cycles per recording	Same as [232].	Least squares SVM	86.72	✓	50-50 train-test split. Hand-picked C and γ SVM values. Unbalanced train-test sets
[262]	Hand-picked 201 cycles from 22 subjects (16 normal). Subset of [140]	7 different TFR methods. Extracted 53 different feature contours	k NN ($k=5$)	98	✓	10-fold CV. Feature extraction and classification optimised on entire dataset
[263]	Hand-picked 274 normal and murmur cycles from 45 (26 normal) patients. Subset of [140]	T-F features from locally stationary time-varying autoregressive model, with various dimensionality reduction methods	k NN ($k=3$)	99.0 99.56 98.45	✓	11-fold cross-validation. Feature extraction and classification optimised on entire dataset
[142]	40 patients (34 teaching-quality pathological recordings of 6 conditions with 293 cycles). Normal recordings (196 cycles) made using different device	Maximum frequency and width of frequency peak in normalised autoregressive PSD	Binary SVM	99.5 99.9	✓	✗ Training set of 56 normal and 84 pathological cycles: Unbalanced. Hand-picked C and γ and decision threshold for SVM
[264]	1031 teaching-quality cycles from 44 recordings (of 20 pathologies)	STFT-derived features, then applied discrete cosine transform. Discriminant ratio feature selection	6-class SVM	90		500 training cycles, 480 test. Hand-picked C and γ SVM values. Classified recordings, not pathologies.
[143]	57 teaching-quality recordings (various pathologies)	Same as [232], and envelope-based features (derived without segmentation)	Neural network	92	✓	✗ 10-fold CV, with test set used for both validation and testing. No feature normalisation
[265]	160 teaching-quality cycles (80 pathological consisting of 8 conditions)	Mel-frequency cepstral coefficients of order 12 (39 dimensional feature vector)	Binary SVM	97.5 98.7	✓	✗ Leave-one-out CV. Unbalanced 9-class classification. HMM and feature parameters optimised on entire dataset. Hand-picked C and γ SVM values
[266]	350 teaching-quality cycles (50 normal, 80 MR, 100 AS, 50 AR, 70 MS) from unknown number of subjects	Wavelet packet entropy in 64 bands from systole and diastole. PCA reduction to 32 features	Bayesian Network	96.94	✓	✗ 10-fold CV. All healthy recordings from different device. Unclear selection of classifier parameters
[267]	120 recordings (40 normal, 40 MS and 40 PS) of unknown duration	a) DFT of entire recording (1 Hz resolution up to 300 Hz) b) Autoregressive (Burg) PSD in 64 sample frames (up to 4 kHz)	3-class MLP neural network	95 90.48 97.44	✓	✓ 50-50 train-test split. No feature normalisation
[268]	Same as [267]	db4 level 5 wavelet decomposition. Shannon entropy over entire recording (6 features)	3-class adaptive neuro-fuzzy inference system (ANFIS)	98.33 100 95.24	✓	✓ 50-50 train-test split. No feature normalisation
[269]	Same as [267]	DFT over entire records. PCA-based feature selection	3-class HMM	97.5 95 98.8	✓	✓ 50-50 train-test split

[270]	804 cycles from 59 patients (16 normal, 19 MR, 14 AS, 10 AR)	Wavelet packet decomposition (unknown mother wavelet) over entire cycles. Rule-based node selection. Then, relative energy of each node (43 features)	4-class SVM	97.56	✗	Hand-picked C (150) and γ SVM values
[271]	72 teaching-quality recordings (6 recordings of 12 types of heart sounds: 8 pathological, 4 innocent).	24th order LPC coefficients extracted from individually segmented sounds (S_1 , S_2 , murmur etc). Modified Cuckoo search feature selection	12-class one-vs-all SVM	94.68	✓	Leave-one-out cross-validation
[272]	435 recordings (8 seconds) from 133 subjects (63 with CAD)	9 frequency types across 5 frequency bands: autoregression, MUSIC, instantaneous frequency, octave bands, Hjorth parameters, sample entropy, simplicity, spectral entropy, statistical moments. Frequency bands: 25–250 Hz; 25–1000 Hz; 50–500 Hz; 125–750 Hz; and 250–1000 Hz. 5029 features	AUC, quadratic discriminant analysis, using autoregressive features in 25–250 Hz and 25–1000 Hz	72	65.2	✓ 10-fold CV. 7 patients manually excluded due to noise
[273]	614 hand-segmented cycles from 40 children (20 of which had systolic ejection clicks)	Spectral power in five adjacent frequency bands of 50 Hz width, extracted over length of growing/sliding windows for classifiers	a) TCNN b) TDNN c) MLP NN	97.0 85.1 92.7	98.1 76.4 85.7	✓ 50-50 train-test split
[274]	80 normal recordings (unknown duration) from 42 subjects. 167 pathological recordings from 38 subjects (44 MR, 46 TI, 39 AS, 38 PS).	Wavelet packet decomposition (db10 at level 5). Reconstruction of two signals based on expected frequency content in heart sounds and murmurs. Six features: relative energy and sample entropy in heart sound and murmur signal, energy content and position of wavelet node with highest energy	5-class RBF SVM	97.17	✓	✗ 10-fold CV. Normal and pathological recordings from different devices. Hand-picked C value for SVM ($C=\infty$). γ optimised on test set
[275]	4628 teaching-quality heart cycles (444 septal defects, 626 normal, 2771 valvular defects and 787 other)	Tunable-Q wavelet transform over 10 decomposition levels ($Q=8$), then sum of average magnitude difference function (SAMDF) found for each pair of reconstructed signals	4-class least squares SVM with Morlet wavelet kernel	98.9	98.8 99.3	✓ ✗ 80% train, 20% test. Unbalanced datasets
[276]	86 recordings (30 AS, 26 innocent and 30 normal) of 10 seconds	Spectral power over varying duration frames and frequency ranges	a) 3-class growing time SVM b) 3-class SVM	86.3 83.0	89.3 84.4	✓ 50-50 train-test split

Abbreviations: AR: aortic regurgitation; AS: atrial stenosis; ASD: atrial septal defect; CAD: coronary artery disease; CV: cross-validation; DFT: discrete Fourier transform; DWT: discrete wavelet transform; FFT: fast Fourier transform; GAL: grow-and-learn; GT: Gabor transform; HMM: hidden Markov model; k NN: k nearest neighbour; MFCC: mel-frequency cepstral coefficient; MLP-BP: multi-layer perceptron back propagation; MR: mitral regurgitation; MS: mitral stenosis; PCA: principal component analysis; PDA: patent ductus arteriosis; PS: pulmonary stenosis; PSD: power spectral density; RBF: radial basis function; STFT: short-time Fourier transform; SVM: support vector machine; TDNN: time-delay neural network; TFR: time-frequency representation; TGNN: time-growing neural network; TOF: tricuspid insufficiency; VSD: ventricular septal defect; WT: wavelet transform; WVD: Wigner-Ville distribution.

Another successful early attempt to screen for heart murmurs in children was reported by Ninova et al. [242]. They diagnosed pathology based on a set of noise-sensitive rules, relating the systolic and diastolic amplitudes to the S_1 amplitude. This simple technique achieved 92.4% Sp in a test group of 2583 school children (although sensitivity was not reported on this dataset), while achieving 96.5% Se in a separate test group of 345 pathological cases.

5.1.2 Artificial neural network-based classification

Artificial neural networks (ANNs) have been the most widely used machine learning-based approach for heart sound classification. The following sections review the use of ANNs, with the relevant studies grouped by the features as the input to the ANN classifier.

5.1.2.1 Wavelet features with artificial neural networks

Alongside ANNs, the use of wavelet-based features has dominated the literature. Akay et al. [244] combined wavelet features with ANNs for the automatic detection of heart pathology, focusing on coronary artery disease (CAD). They computed four statistical moments of extracted wavelet coefficients from the diastolic period of heart cycles. These features, alongside physical characteristics (sex, age, weight, blood pressure), were fed into a fuzzy neural network, to achieve 85% Se and 89% Sp on a separate test set of 82 recordings. Similarly, Barschdorff et al. [146] classified pathological murmurs in children by extracting the mean wavelet detail coefficients from systolic periods, which were then used to train an ANN, and resulted in 93.3% accuracy on a separate test set. Three years later, Liang et al. [235] employed wavelet packed decomposition with the aim of differentiating between pathological and innocent murmurs in children. Eight nodes of the wavelet packet tree were selected automatically based on a cost function, and the cost function values served as the feature vector. With a 65-20 patient train-test split they achieved 80% Se and 90% Sp when using ANN-based classification.

In contrast to the previous methods involving wavelet feature extraction from either the systolic or diastolic periods, a number of researchers have made use of wavelet features extracted from *entire* heartbeat cycles. Ölmez and Dokur [232] calculated the sum of squared wavelet detail coefficients in 32 equally-sized windows (without regard for the position of S_2) over entire heart cycles. However, they evaluated their method without a train-test split. Other researchers

have used the same 32-element wavelet-based feature vector and ANN combination employed by Ölmez and Dokur, for example, Gupta et al. [125], Ari and Saha [139] and Yuenyong et al. [143]:

Gupta et al. [125] achieved 97.0% and 96.5% average three-class (healthy, systolic or diastolic murmur) classification accuracy when using a GAL and MLP ANN respectively with three-fold cross-validation (41 recordings), but did not divide the data by patient across the three folds. Ari and Saha [139] optimised their ANN structure on the their test set results, reporting 13-class accuracy of 99.3%^a, using 832 cycles from 104 patients. Yuenyong et al. [143] also did not divide patients and recording devices evenly across train and validation sets, but achieved 92% balanced accuracy after principal component analysis (PCA)-based dimensionality reduction. These researchers claim that segmentation was not required for their classification approach. However, for their method to work accurately, the recording of heart cycles would have to commence at the same phase in the heart cycle.

In contrast to their previous work, Dokur and Ölmez [236] extracted the sum of squared wavelet detail coefficients from five equally-spaced windows in the systolic and diastolic periods, rather than from entire heart cycles. Without dividing their data by patient in their 50-50% train-test split (in a dataset of 280 teaching-quality cycles), these researchers achieved 99% 14-class classification accuracy^b, while achieving 95% accuracy when using the raw power of the windowed PCG alone, indicating how straightforward the classification of this dataset was. The same year, these researchers compared the performance of two self-organising ANNs on the same dataset [258]. Unsupervised learning of the optimal ANN structure was performed using the same training dataset of half the cycles, with the incremental self-organising map performing best, with 95% 14-class classification accuracy.

Uğuz [268] most recently employed an ANN with wavelet-based features. This researcher employed an ANN with fuzzy logic to perform three-class classification after extracting the detail coefficients over the entire recording, then calculated the Shannon entropy of the coefficients. With a 50-50% train-test split of a dataset of 120 PCGs, they achieved 100% *Se*, 95.24% *Sp*, 98.33% three-class (healthy, mitral stenosis, pulmonary stenosis) accuracy.

^aSuch high classification accuracy over so many classes is hard to believe without severe over-training without a separate test set, or a remarkably straight-forward dataset.

^bSee previous footnote.

5.1.2.2 Other time-frequency features with artificial neural networks

Barschdorff et al. [243] reported between 88.9 and 100% classification accuracy (*Acc*) on 69 heart cycles using short time Fourier transform (STFT) and Hilbert envelope-based features as input into an ANN, with unclear validation methods. De Vos and Blanckenberg [126] hand-picked six cycles from each of 163 patients and extracted the energy in 12 frequency bins, at 10 equally-spaced time intervals over each heart cycle. These researchers employed leave-one-out cross-validation (LOOCV) to train and test an ANN, which gave 90% *Se* and 96.46% *Sp*. Visagie [255] extracted various time- and frequency-based features, and used LOOCV to test the ANN, with which he reported 85.7% *Se* and 94.1% *Sp* on a dataset of PCGs from 31 adult subjects.

Prior to their wavelet-based work, Uğuz experimented with other time-frequency features [267]. These included the discrete Fourier transform of the entire recording and the power spectral density (PSD) extracted using the Burg autoregression method from a database of 120 PCGs. An ANN was trained on 50% of the PCGs, and achieved test-set results of 90.48% *Se*, 97.44% *Sp* and 95% for a three-class classification problem (with both feature sets).

5.1.2.3 Time, frequency, and complexity-based features with artificial neural networks

Degroff et al. [247] and Bhatikar et al. [250] both extracted the FFT (253 bins with 1 Hz resolution) over three hand-picked cycles per patient. Degroff et al. [247] trained an ANN using LOOCV on a dataset of 69 children, but optimised the features and decision threshold on the test set, achieving 100% binary classification accuracy. The same researchers made use of a larger dataset of 241 paediatric patients [250], after again manually selecting the three best-quality cycles from each subject. They report 88% *Se* and 83% *Sp* when differentiating between innocent and pathological murmurs using an ANN on a separate test set (of 53 patients).

Sepehri et al. [198] automatically identified the five frequency bands that led to the greatest difference in spectral energy between pathological and normal recordings. This resulted in a choice of bands between 15 and 129 Hz – a surprising result given the expected frequency of the S_1 and S_2 sounds (24 - 144 Hz) [92]. The spectral energy in these bands served as the input

features for their ANN, trained on an unbalanced dataset of 40 recordings, which achieved 94% out-of-sample binary classification accuracy on 50 test records.

Ahlstrom et al. [253] performed an extensive review of various features for murmur classification on a small, real-world dataset of 36 patients. Most notably, they assessed various non-linear dynamics-based features that had not previously been used for murmur detection and classification. Their feature set included: the Shannon envelope; wavelet detail coefficients; the wavelet entropy; the Stockwell transform (downsampled to a 4x4 matrix); eigenvalues from a singular value decomposition (SVD) of the Stockwell transform; higher-order statistics; the variance fractal dimension; and recurrence quantification analysis. This led to a total of 207 features. Using a feature selection procedure (Pudil’s sequential floating forward selection method on the entire dataset), they selected 14 features, including various non-linear and chaos-based features. When using LOOCV (with an unbalanced dataset), they achieved 86% classification accuracy for a three-class problem, using an ANN. Their accuracy when using non-linear and chaos-based features is in contradiction with the findings of Schmidt et al. [277], who found no evidence of non-linear or chaotic behaviour in cardiovascular murmurs.

More recently, Gharehbaghi et al. [273] performed a rigorous assessment of a time growing neural network (TGNN) for the classification of systolic ejection clicks (SECs) in children. A TGNN has the advantage of growing feature windows, rather than a sliding window of a time delay neural network or the static feature window of a multi-layer perceptron ANN. They extracted the power in five adjacent frequency bands of 50 Hz width as the input feature, selecting 614 hand-segmented cycles from 40 children (20 – 20 train-test split), half of which had SECs. The various design parameters (number of windows in the TGNN and TDNN, number of hidden nodes, decision threshold) were optimised on the training dataset, with the optimised models achieving accuracies and sensitivities of 97.0 and 98.1% for the TGNN, 85.1% and 76.4% for the TDNN, and 92.7% and 85.7% for the MLP on the test set. Noise stress tests were performed for the various classifiers, with the TGNN being least sensitive to synthetic and physiological noise.

In summary, ANNs have been employed with many input features in order to discriminate between normal and various pathological heart sounds. However, many of these investigations were marred by poor methodology, often without the use of a separate test set or without

ensuring that recordings from the same patient were not in both the training and test set, almost certainly resulting in over-fitting of their model to the data. Others have hand-picked clean data in their dataset while excluding noisy data. The studies that have employed ANNs with real-world datasets and adequate train-test splits were those performed by Akay et al. [244], Barschdorff et al. [146], Liang et al. [235], De Vos and Blanckenberg [126], Visagie [255], Uğuz [267, 268], Ahlstrom et al. [253] and Gharehbaghi et al. [273].

5.1.3 Support vector machine-based classification

A number of researchers have applied support vector machines (SVMs) to the problem of heart sound classification in recent years. As in the previous section, these studies are divided according to the feature extraction method. An explanation of SVMs and their parameters is given in Appendix C.

5.1.3.1 Wavelet feature-based classification with support vector machines

Choi [260] was the first researcher to apply an SVM for the classification of heart sounds. However, this researcher’s methodology was flawed: the normal and pathological recordings were made using different devices, while the SVM parameters were heuristically chosen (C was set to a large number to not allow misclassification and a fifth-order polynomial kernel was selected). Similarly, Ari et al. [234] set their SVM parameters by trial and error when building on their previous work, by extracting the same 32 window wavelet-based feature from the entire heart cycle (originally proposed in [232] and attempted earlier by Ari et al. [139]), and then employing a least-squares SVM with a radial basis function (RBF) kernel rather than an ANN. They trained the RBF-SVM with a 50-50% train-test split (256 cycles from 32 patients in each set), and achieved 86% accuracy on their test dataset.

Many researchers who employed SVMs for classification with wavelet features failed to split the data into training and test sets, or failed to allocate patients to only either the training or test sets. For example, Safara et al. [266, 270], who also hand-crafted the SVM parameters; Zheng et al. [274], who, again, did the same; and Patidar et al. [275], who investigated the use of the tunable-Q wavelet decomposition in a large dataset of 4628 cycles, achieving 98.8% Se , 99.3% Sp , but without allocating patients uniquely to either the training or test set.

5.1.3.2 Time, frequency and other time-frequency feature-based classification with support vector machines

As with the many researchers who failed to split their data adequately or who hand-crafted their SVM parameters when using wavelet features, other researchers repeated the same mistakes with other time and frequency features. These included: Maglogiannis et al. [261], who set C to infinity, while γ was optimised on their test set; Choi and Jiang [142], who set their C and γ parameters for the RBF-SVM by trial and error while having no train-test split; Kao and Wei [264], who did not incorporate cross-validation in order to set the SVM parameters while training a different SVM for each recording, rather than pathology type; and Kwak and Kwon [265], who combined the use of HMMs for segmentation, but optimised the HMM and feature parameters on the whole dataset, while the SVM parameters were set by trial and error.

Gharehbaghi et al. [276] employed a growing-time SVM (GTSVM) (very similar to their implementation of the time growing variation of an ANN [273]) to classify pathological and innocent murmurs. However, no mention is made of the C parameter for the linear SVM they employed. The spectral power in frequency bands over varying length frames during systole was used as input features. The eight optimal frequency bands (between 0 and 1000 Hz) were selected based on the Fisher criterion in the training set, while the number of frames and initial frame size of the GTSVM were set using the classification rate on the training set. Three one-versus-all SVMs were trained, with their final outputs being the averaged class labels assigned to each 10-second recording. k -means clustering using these average class labels was then used to perform the final classification. When performing 500 iterations of a 50-50 train-test split of their dataset (30 patients with aortic stenosis, 26 with innocent murmur and 30 normal), the GTSVM achieved 86.4% Se and 89.3% Sp , while a standard linear SVM achieved 83% Se and 84.4% Sp (while a standard ANN did not reach 80% for either Se or Sp). However, the standard SVM was trained on the spectral power of the entire systolic period, yet the murmurs of interest occurred in early systole.

Redlarski et al. [271] used an SVM with linear predictive coding (LPC) coefficients to classify 12 different heart sounds. Rather than identifying pathological recordings or searching systole and diastole for signs of pathology, these researchers classified specific sounds (murmur, S_4 , split S_1) that were hand-segmented from a database of 72 teaching-quality recordings of 12

types of heart pathology. They extracted 24 LPC coefficients from each sound, then used a modified cuckoo search (MCS) optimiser, a form of particle swarm optimisation, to select features and optimise the SVM parameters. When evaluating their method, using a one-vs-all and LOOCV approach (without proper allocation of each patient to either the training or test set), they achieved 100% binary classification accuracy, and 93.23 12-class accuracy with their MCS-SVM, which outperformed ANN and standard SVM approaches.

The SVM-based literature is again dominated by studies with poor methodology. For instance, out of the above 12 studies, seven sets of authors hand-crafted their SVM parameters (Choi [260], Ari et al. [234], Zheng et al. [274], Maglogiannis et al. [261], Choi and Jiang [142], Kao and Wei [264], Kwak and Kwon [265]), with many setting their C parameter to infinity (thereby not allowing any misclassification in their training dataset. This either indicates severe over-training, or an easily separable dataset). Seven of these studies either had no train-test split or data from the same patient appeared in both the training and test sets (Safara et al. [266, 270], Zheng et al. [274], Patidar et al. [275], Choi and Jiang [142], Kao and Wei [264], Kwak and Kwon [265]). Only the studies by Redlarski et al. [271] and Gharehbaghi et al. [276] were rigorous in their methodology. However, Redlarski et al. [271] performed their analysis over hand-segmented sounds, rather than searching systole or diastole for signs of pathology, and did allocate patients independently to the train or test set, while it is unclear how Gharehbaghi et al. [276] selected their SVM parameters.

5.1.4 Classification using hidden Markov models

Hidden Markov models (HMMs), which were explained in Chapter 4, have also been employed for pathology classification in PCG recordings. Whereas the most likely sequence of states was of importance in the segmentation algorithms presented in Chapter 4, in the case of classifying pathology, the posterior probability of the PCG (or extracted features) given a trained HMM can be used to differentiate between healthy and pathological recordings.

Wang et al. [256] and Chauhan et al. [257], in their evaluation of mel-frequency cepstral coefficients (MFCCs) for pathology detection, failed to make use of a separate test set when evaluating their HMM-based classification methods.

Saraçoğlu [269] applied HMMs in an unconventional manner, by fitting an HMM to the frequency spectrum extracted from entire heart cycles. Their exact classification procedure when using the HMMs is unclear, but it is thought that they trained four HMMs, and then evaluated the posterior probability of the features given each model to classify the recordings. They optimised the HMM parameters and PCA-based feature selection on the test set to achieve 95% Se , 98.8% Sp , 97.5% accuracy on their test dataset of 60 records (20 per class), with a training set of the same size. This method performed as well as the wavelet feature-based ANN classification evaluated on the same database [268], but outperformed an ANN and k NN using frequency spectrum features.

The potential of HMMs to classify heart sounds has not yet been adequately demonstrated. The studies presented by Wang et al. [256] and Chauhan et al [257] were performed on a very clean dataset (indicated by the 93.75% classification accuracy when using the Shannon envelope alone as an input feature [257]) with a lack of train-test split, while the study by Saraçoğlu [269] used the HMM in the frequency domain, rather than the time domain, which is an atypical application.

5.1.5 Clustering-based classification

A number of researchers have made use of the k -nearest neighbours (k NN) algorithm to classify pathology in PGCs. Bentley et al. [245] showed that discrete wavelet transform features outperformed morphological features (time and frequency features from S_1 and S_2) when classifying normal and dysfunctional native and prosthetic valve replacements in a dataset of unknown size. When using LOOCV with a binary k NN classifier, these researchers achieved 100% and 87% accuracy when detecting pathology in native and prosthetic heart valves respectively.

Delgado-Trejos et al. [140], Quinceno-Manrique et al. [262] and Avendaño-Valencia et al. [263] all performed k NN-based classification with various features on portions of the same dataset. This was the same dataset analysed in [257], where a 93.75% test accuracy was achieved when using the Shannon envelope feature alone in a subset of the whole dataset (giving an indication of how pronounced the murmurs were in this dataset).

Delgado-Trejos et al. [140] experimented with three feature sets when performing binary classification of pathology. The three feature sets were: time and time-frequency features (from

various transforms), MFCCs and fractal analysis features (inspired by Ahlstrom et al. [253]). From the original dataset of 148 adult patients (with eight recordings from each) they selected only 180 cycles from each class (healthy and pathological with only systolic murmurs), indicating that approximately 96% of their data was discarded. They employed a k NN classifier ($k = 9$, set by trial and error) with 10 bootstrapping iterations in which 30% of their data was selected as a test dataset. Without the correct allocation of patients to either training or test set, classification using the fractal features performed the best with 97.17% accuracy, while the time-frequency and MFCC features achieved 94.35% and 86.85% respectively.

As with Delgado-Trejos et al. [140], Quinceno-Manrique et al. [262] compared classification accuracy when deriving features from various time-frequency representations (TFRs), on a subset of 201 cycles from the dataset used by Delgado-Trejos et al. From these TFRs, they extracted a wide range of features, including the short-time average energy, features from singular value decomposition, instantaneous frequency and MFCCs, resulting in 53 different features. The researchers employed the same 10-iteration bootstrapping method employed by Delgado-Trejos et al. The MFCC-based features performed the best with 98% accuracy (after the k NN classifier parameters were optimised on the test set).

In contrast to the previous two studies, but when analysing the same dataset, Avendaño-Valencia et al. [263] extracted only time-frequency features using a parametric locally stationary time-varying autoregressive model (LS-TVAR). A larger number of 274 cycles of each class were selected in this study (with approximately 95% of the original data excluded). The AR model order that minimised the Bayesian information criterion (BIC) for each cycle was chosen. Using a k NN classifier, these researchers achieved $99 \pm 0.06\%$ classification accuracy when using 11-fold cross validation with grid-based dimensionality reduction.

In summary, k NN-based classification shows promise, with results exceeding 95% accuracy. Bentley et al. [245] showed promising results, but did not specify the size of their dataset. Delgado-Trejos et al. [140], Quinceno-Manrique et al. [262] and Avendaño-Valencia et al. [263] all obtained very good results on their carefully-selected data. However, this is not surprising, given that a Shannon envelope feature along with an HMM classifier achieved over 93% accuracy on the same dataset. In these studies, the wavelet, time-frequency and MFCC features performed the best.

5.1.6 Other methods for heart sound classification

The sections above have discussed the most popular machine learning approaches to classifying heart sounds. However, many other techniques have also been applied, including threshold-based methods, decision trees, discriminant function analysis and logistic regression.

A number of these studies ([84, 108, 252]) are flawed because of the lack of a separate test dataset. For instance, Syed et al. [120] made use of a manually chosen portion (39 recordings out of 411) of the MAAS dataset employed in this chapter to develop a complete PCG analysis framework, including ECG-based cycle segmentation, cycle selection, wavelet decomposition, clustering of cycles and cluster selection based on frequency content. A prototypical beat was constructed from the selected cycles, while pathology classification was performed using a threshold based on the timing of frequency peaks, tuned without a train-test split. They achieved 85.7% Se and 94.1% Sp on their 39 recordings.

Pavlopoulos et al. [248] applied a decision-tree classification approach to distinguish mitral regurgitation and aortic stenosis in 84 hand-picked teaching-quality recordings (of which one source was a veterinary website). The 100 features they extracted included the mean frequency content in the systolic and diastolic regions, and the normalised average Shannon envelope in small windows. When using all 100 features (with only 67 training records) and a 80-20% train-test split, they achieved 90% classification accuracy.

El-Segaier et al. [254] developed a semi-automated method to differentiate between innocent and pathological murmurs in children. They developed a system for the automatic detection of the S_1 and S_2 sounds with an ECG for reference [278], then employed the STFT to extract the mean spectral power, the maximum intensity during systole, the mean frequency during systole, the frequency at the point of maximum intensity in the spectrum (after excluding the first 20% and last 30% of systole) and then calculated the mean and standard error of each feature over all the heart cycles in a recording. In addition, they manually measured the interval from the end of S_1 to the start of a systolic murmur, and the variation in split in the S_2 sound. When using stepwise logistic regression, they achieved 95% Se and 72% Sp on their dataset of 257 children (188 with pathological murmurs).

Schmidt et al. [272] performed a rigorous investigation into the types of frequency- and complexity-based features that may indicate coronary artery disease. They extracted nine types of features

from five frequency bands (25–250 Hz; 25–1000 Hz; 50–500 Hz; 125–750 Hz; and 250–1000 Hz) in 100ms overlapping (50%) windows in the diastolic period of 435 recordings from 133 patients (63 with coronary artery disease). The features included: AR-based pole magnitudes and angles; eigenvector-based spectral estimation; the instantaneous frequency; the power in octave bands; the sample entropy; the simplicity; the spectral entropy; the Hjorth mobility and complexity; and statistical moments. The median value of these 5029 features was found over all heartbeats. Ten-fold cross validation was employed to evaluate the features. When comparing single-feature AUROC, they found that the AR features from the 25–250 Hz band were the best-performing features (0.71), while the AR features over all frequency bands led to the highest average AUC (0.653). When using a multivariate classifier (quadratic discriminant analysis) to construct a single coronary artery disease score from multiple AR features, the mean AUROC improved to 0.73, with sensitivity of 72% and specificity of 65.2%.

It is apparent from a review of the literature that wavelet-based features have led to good classification performance when using a variety of classification techniques. Further, MFCC features, which have only been investigated in a handful of studies [140, 256, 257], have demonstrated potential to detect pathology accurately in PCG recordings [262]. For these reasons, these two types of features were investigated in this Chapter.

A number of studies have investigated the classification performance when using an SVM. However, as pointed out in this review, many of these implementations are flawed. To contribute reliable results in the field of heart sound classification, a rigorous implementation of SVM-based classification was implemented as described in this chapter.

5.2 Methods

5.2.1 Data

Three datasets, containing healthy patients and patients with pathological systolic murmurs, were analysed in this chapter. These were:

- Recordings from the MAAS dataset (see Section 2.4) of patients with diagnosed mitral valve prolapse (MVP), or those without diagnosed heart disease. This led to a total of 370 recordings (135 with MVP) from 108 patients (37 with MVP). These recordings

were taken from various auscultation positions (the majority being from the apex and tricuspid positions), with the patients in different orientations (supine, sitting, or on their left side). This dataset was difficult to analyse as many recordings were corrupted by the noise, occurring in the uncontrolled environments in which the recordings were made. Secondly, the patients with no murmur or innocent murmurs in this dataset had been referred from primary care physicians for echocardiographic screening on the suspicion that they had MVP. Therefore, the majority of the non-MVP patients were false positives, meaning that a false positive detection rate on these recordings of less than 100% would already be an improvement on the clinician diagnoses. A particularly pronounced murmur due to MVP is shown in Figure 5.1.

- Recordings from the RHD mobile phone dataset (RHD_MP – see Section 2.3) of children and young adults with latent rheumatic heart disease or their matched controls. This led to a total of 106 patients, 40 of whom had RHD (borderline or definite), one with ventricular septal defect (VSD), one with patent ductus arteriosus (PDA) and 64 controls, all confirmed via echocardiography. Of the patients with RHD, 31 had *functional valve damage*, or pathological damage to the heart structure leading to an audible murmur^c. The recordings from the apex and tricuspid positions were analysed, as these were the positions where a murmur due to RHD is most likely to be heard (see Section 1.3.4.5). This led to a total of 318 recordings (three recordings per patient). Apart from the influence of the low-cost stethoscope that was used to collect this dataset, classification of pathology in this dataset was difficult due to the *sub-clinical* RHD with which the patients were diagnosed. In the majority of the studies reviewed in Section 5.1, pathological patients were initially identified by the presence of an audible murmur, which was then confirmed using echocardiography. In contrast, this dataset consisted of pathological patients initially identified with echocardiography, regardless of the presence of an audible murmur. One of the more noticeable murmurs due to RHD in this dataset is shown in Figure 5.2, while an innocent murmur, also recorded using the mobile phone-based stethoscope, is shown in Figure 5.3.

^cImportantly, not all patients with RHD have pathological valve damage. Many of the patients in this dataset with borderline RHD had only abnormalities in the *shape* of their valves, while not having abnormal *function* of these valves. These patients belong to class A of borderline RHD, outlined in Section 2.3.

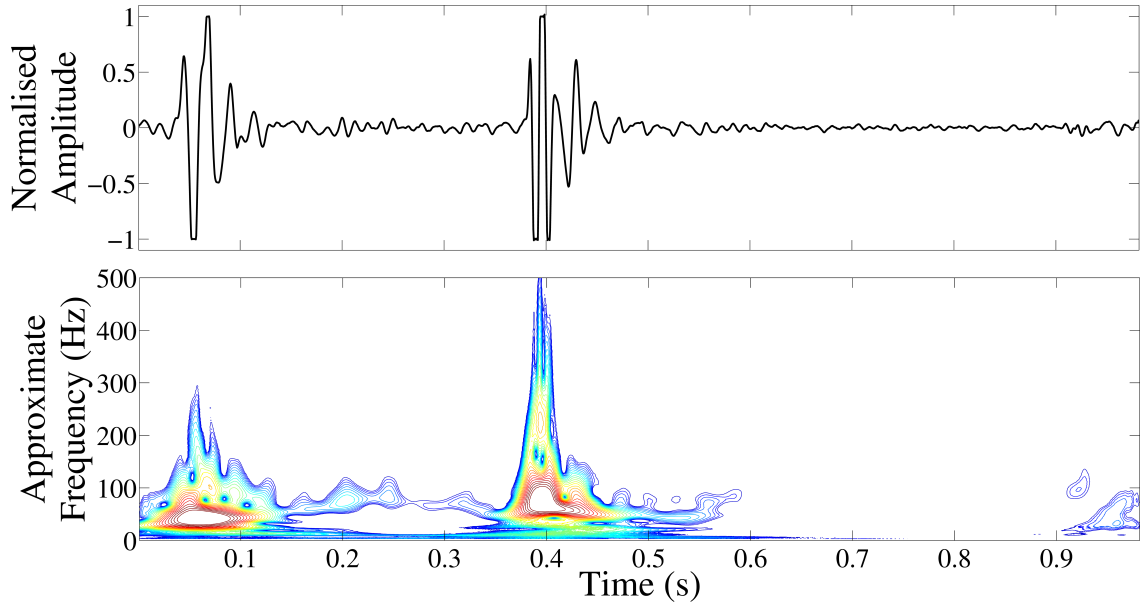


Figure 5.2 – Example of holosystolic murmur indicative of RHD recorded using the mobile phone-based stethoscope. The arcing shape of the systolic murmur, between 50 and 150 Hz, can be seen between the S_1 and S_2 sounds. This figure was generated using the continuous wavelet transform, with use of the “Morlet” wavelet. The approximate frequency is derived using the centre frequency of the wavelet.

- Recordings from the RHD electronic stethoscope dataset (RHD_ES – see Section 2.3), of the same patients in the mobile phone dataset, but with PCGs recorded using a Littmann electronic stethoscope. This dataset comprised a total of 102 patients, 38 of whom had RHD (borderline or definite), one patient with VSD and 64 controls. Of the patients with RHD, 27 had *functional valve damage*, leading to pathological regurgitation. Again, the mitral and tricuspid recordings were analysed. This led to a total of 210 recordings (with three subjects with repeated recordings) (78 with RHD, and 62 of those with functional valve damage). Again, in contrast to the majority of studies reviewed in Section 5.1, this dataset consisted of pathological patients initially identified with echocardiography, regardless of the presence of an audible murmur, making classification in this dataset difficult.

As outlined in Section 2.3, the patients recruited for the RHD datasets were also examined by a cardiologist (using manual auscultation) and Zargis Cardioscan software, which is software that automatically detects the presence of pathological murmurs by analysing the heart sound recordings from a Littmann electronic stethoscope. A comparison of the performance of these methods with respect to the automated methods developed in this chapter will be presented.

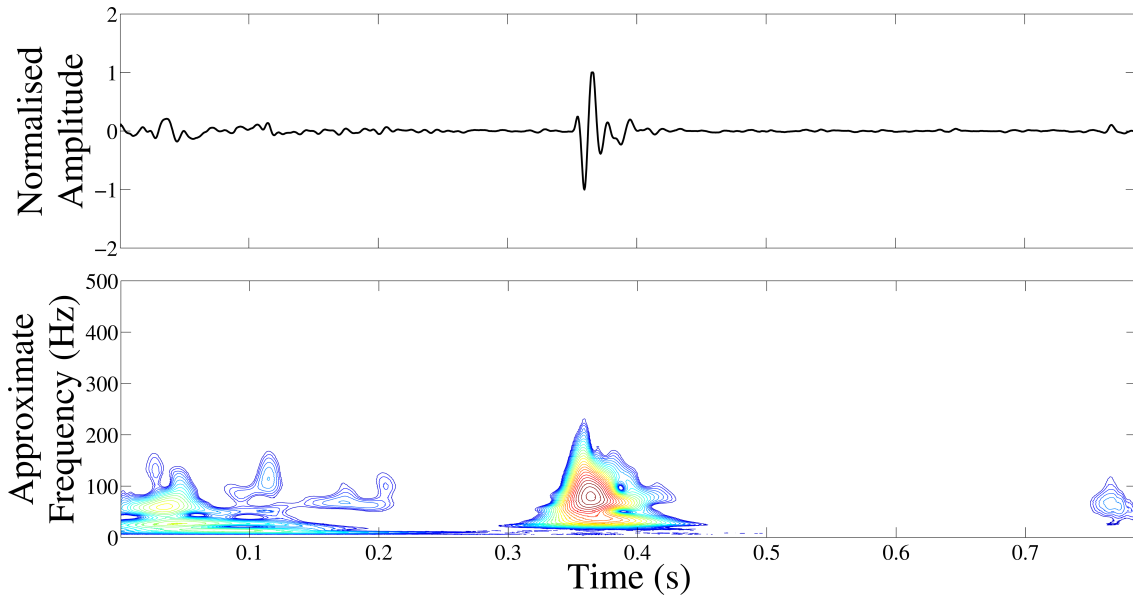


Figure 5.3 – Example of an innocent systolic murmur recorded using mobile phone-based stethoscope. This murmur is confined to early- to mid-systole, while the S_1 sound in this case is faint. The similarity between this murmur, and the first half of the pathological holosystolic murmur of Figure 5.2 can be noted.

5.2.2 Exclusion of noisy data

5.2.2.1 Generator noise

Many of the PCGs recorded from the subjects with RHD were performed in a mobile clinic, as described in Section 2.3, which ran an electric generator to power the medical equipment. This generator produced non-stationary acoustic noise which overlapped with the frequencies of interest in the PCG. This noise is displayed graphically in Figure 5.4.

Because of the spectral overlap between this generator noise and the expected frequencies of murmurs, PCG recordings of RHD patients made with generator noise were manually excluded. This led to 154 recordings being removed from the RHD_MP dataset, leaving 164 recordings free of generator noise. In the RHD_ES dataset, this led to the exclusion of 81 recordings, leaving 129 recordings free of generator noise in this dataset.

5.2.2.2 Poor-quality mobile phone-based RHD recordings

The PCGs recorded using a low-cost mobile phone-based stethoscope are often of poor quality, as identified in Chapter 3. This is due to the stethoscope's low-cost construction, lack of noise insulation, and the requirement for an air-tight seal to be made between the rim of the stethoscope and the patient's chest in order for adequate heart sounds to be recorded. For

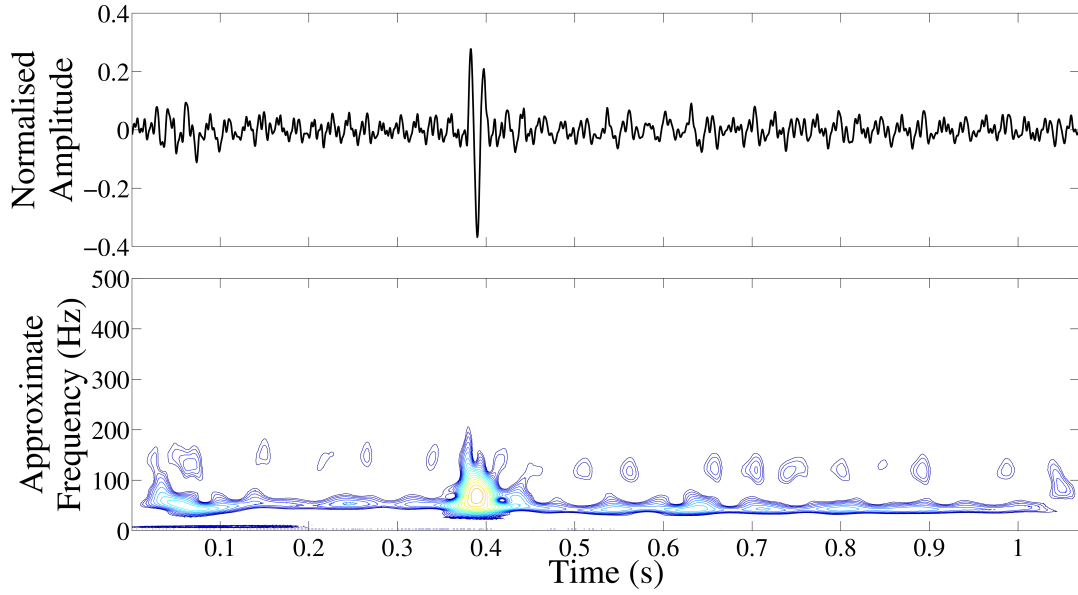


Figure 5.4 – Example of a heart cycle in the presence of the electricity generator. The raw PCG is shown at the top of the plot, while the frequency content is shown below, estimated using the continuous wavelet transform. The S_1 sound in the PCG is hidden by the noise, while the frequency content of the generator can be seen to contaminate the 50-150 Hz frequency range.

these reasons, many of the PCGs recorded using the mobile phone-based stethoscope were of lower quality, with a lack of audible heart sounds. It was necessary to identify and exclude these recordings from further analysis, as such recordings would not present any diagnostic information.

The logistic regression (LR)-based classifiers developed in Chapter 3 were employed to discriminate between good- and bad-quality mobile phone-based PCGs. In Chapter 3, a separate LR classifier was trained on each of the iPhone and Littmann datasets. However, the RHD mobile phone-based recordings in this chapter were made using a Samsung mobile phone which had a different frequency response to that of the iPhone device. Therefore, a classifier trained only on the iPhone recordings may perform poorly on recordings from a different device, which was confirmed in preliminary tests. In order to develop a new classifier which would be agnostic to recording device, both the iPhone and Littmann recordings were combined to train a signal quality classifier.

It was necessary to balance the number of good- and bad-quality recordings from both the iPhone and Littmann devices from Chapter 3, in order not to bias the classifier. Therefore, the 50 best-quality and 50 worst-quality recordings (based on their mean annotation score^d) from

^dThe mean annotation score refers to the mean of the annotations (on a scale of 1-5), provided by the five annotators in Chapter 3.

both the iPhone and Littmann datasets were selected to train the LR signal quality classifier.

The mean and standard deviation of the individual features were computed, and used to normalise the training data. The LR classifier was trained following the identical procedure as in Chapter 3, using 20-fold cross-validation in order to optimise the λ regularisation parameter, with the maximum λ value that resulted in a deviance within one standard error of the minimum deviance being selected. The selected λ value was 0.0215, with *seSQI*, *haSQI*, *mSQI*, *kSQI* and *bSQI* having non-zero β coefficients, the most influential of which were *kSQI* ($\beta = -5.33$) and *bSQI* ($\beta = -1.36$). On the training set, this led to 94.1% sensitivity and 92.52% specificity (while the difference between training and test set sensitivity and specificity using the LR classifiers in Chapter 3 was at most 2.2%).

The signal quality features were extracted from the mobile phone-based RHD recordings, and normalised by the previously computed mean and standard deviation values. With a decision threshold of 0.5 (based on balanced good- and poor-quality training data), 23 of the 164 recordings were automatically classified as poor-quality. The results of the classifier were also manually assessed, as this classifier may under-perform on a completely new dataset with a new recording device. In addition to the 23 automatically excluded recordings, 25 records were manually excluded after being aurally and visually inspected. These were excluded for the following reasons.

Six records were excluded due to loud talking. Talking was not present in the signal quality dataset used to train the classifier, as all recordings were made in a quiet hospital room. However, the RHD recordings were made in various environments, with talking often being louder than the heart sounds. Seven recordings did not have an audible S_1 sound. This was thought to be due to poor placement of the stethoscope. In addition, five of these recordings were from the mitral area from female patients. Therefore, as identified by Schmidt et al. [272], the patient's breast may impede the recording. Finally, 12 recordings were excluded due to inaudible heart sounds. The misclassification of 25 recordings out of a total of 164 (15%) is in agreement with the error rate achieved for both the iPhone and Littmann records in Chapter 3.

In the end, 116 (164-23-25) recordings remained in the RHD_MP dataset from 54 patients. Of the 116 recordings, 51 had functional valve damage (from 21 patients), while 65 were healthy recordings (from 33 patients).

As identified in Chapter 3, electronic stethoscope recordings can also be of poor quality. For this reason, and in order to present a fair comparison between the RHD_MP and RHD_ES datasets, the same poor-quality data exclusion procedure was followed for the RHD_ES dataset. The signal quality exclusion led to the automatic exclusion of 10 records (7 of which were deemed to be false positives, and re-included in the analysis), while two records were deemed to be false negatives, and manually excluded from further analysis. Therefore, 5 recordings were removed from the dataset, leaving 124 recordings. Of these, 45 recordings had functional valve damage (from 24 patients) while 79 were healthy recordings (from 42 patients).

5.2.3 Heart cycle segmentation

The heart cycles were segmented using the extended conditional random field (CRF) approach developed in Chapter 4. Of the proposed algorithms in Chapter 4, the CRF method, with an extended Viterbi algorithm and four features, was shown to outperform the other proposed methods, as well as the state-of-the-art segmentation algorithm. This model was trained on all the ECG-labelled PCG data used in Chapter 4^e.

This segmentation method requires an estimate of the heart rate over the PCG recording. The heart rate was calculated using an autocorrelation-based method [279]. The different processing steps in this method were optimised by performing a search over various wavelet denoising parameters, envelope detection schemes, five normalisation thresholds and three autocorrelation-based heart rate estimation techniques, and applied to a large training database of 585 recordings from five different devices, and a test set of 111 recordings from two more devices. The steps outlined in the best-performing algorithm were:

1. Downsample the PCG data to 1000 Hz using a polyphase anti-aliasing filter.
2. Decompose the downsampled PCG data using the discrete wavelet transform with the biorthogonal 2.8 wavelet to decomposition level three. Reconstruct the PCG waveform using the detail coefficients from this level alone, thereby denoising the signal.
3. Extract the signal envelope using the Hilbert transform (described in Section 4.2.8.2).

^eImportantly, while the CRF segmentation model was trained on the MAAS recordings with reliably identified R-peaks and end-T-waves from Chapter 4, the model was deployed on all 370 MAAS recordings in this chapter.

Table 5.2 – Segmentation performance of the CRF-based algorithm on the mobile phone-based RHD recordings (%). The sensitivity (S_e), positive predictivity (P_+), accuracy (Ac) and F_1 -score, for each sound and the combined sounds, are shown. These results are averaged over all 116 recordings in the dataset.

S_1	S_2	S_e	P_+	Ac	$F_1^{S_1}$	$F_1^{S_2}$	F_1
4076	4071	93.7	96.1	91.2	94.9	94.7	94.8

4. Normalise the envelope by subtracting the median value and dividing by a percentile value (95%) of the absolute value of the envelope.
5. Compute the autocorrelation of the envelope.
6. Locate the maximum peak in the autocorrelation within a permissible range, corresponding to heart rate limits of 40 and 140 beats per minute, based on the first and 99th percentile of expected heart rates in subjects over four years of age [173]. The heart rate was computed using the time from this peak to the origin in the autocorrelation. This method is similar to that employed by Schmidt et al. [99].

The segmentation algorithm, having been trained on electronic stethoscope data, was expected to perform well on similar data, such as the complete MAAS dataset and the electronic stethoscope recordings from the RHD_ES dataset. However, it was not known how the algorithm would fare on an unseen set of mobile phone-based recordings. Therefore, the segmentation performance of the CRF-based algorithm was assessed on the mobile phone-based recordings from the RHD_MP dataset prior to pathology classification. These recordings were hand-annotated by the author of this thesis using Audacity[®] (<http://audacityteam.org/>), an open-source software platform which allows the visualisation of the raw audio and spectrogram of the signal and labelling of the signal during playback. The positions of the S_1 and S_2 sounds were labelled in each of the mobile phone-based PCG recordings, and the same evaluation procedure for segmentation performance described in Chapter 4 was followed, with the mid-point of a CRF-identified sound taken to be correct if it was found within 0.1s of the reference annotation. The average results of the CRF-based segmentation algorithm on the unseen RHD_MP recordings are shown in Table 5.2, achieving 94.8% average F_1 score on this unseen dataset.

The CRF-based segmentation algorithm, trained on the MAAS data, was thus shown to perform similarly on the RHD_MP recordings (94.8% F_1 score) as on the MAAS data in

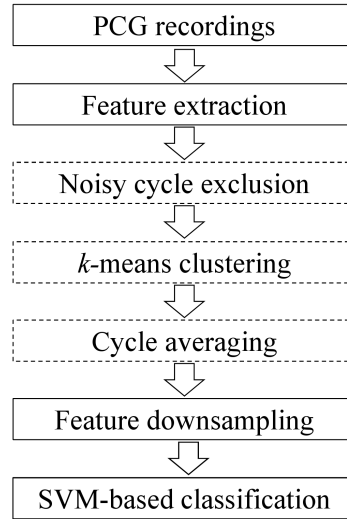


Figure 5.5 – The analysis procedures evaluated in the detection of pathology in PCG recordings. The solid boxes represent the steps performed for each classification procedure, while the dotted boxed represent the *optional* steps, with the improvement in performance with the inclusion of each of these optional steps being evaluated in this chapter.

the original evaluation reported in Chapter 4 (95.6% F_1 score). As a result, this algorithm, trained on the ECG-labelled MAAS database, was employed for segmentation of all the heart sound recordings used in this chapter.

5.2.4 Classification procedures

This chapter investigates the effect of various steps in algorithms for identifying pathology in PCG recordings. These steps include two feature extraction methods, cycle averaging, cycle exclusion, k -means clustering-based cycle selection and the downsampling of features. These steps are performed prior to classification. The analysis procedures evaluated are illustrated in Figure 5.5, while each of the different steps is described in the following sections.

5.2.5 Feature extraction

This section outlines the extraction of raw features from the PCG in order to detect pathological murmurs. Heart valve pathology leads to transient, non-stationary, high-frequency murmurs during the systolic and diastolic phases of the heart cycle [79], which are often of low amplitude. The use of purely spectral or temporal representations of the PCG as features for classification are inappropriate, as these are limited in either their lack of temporal localisation or frequency information. Therefore, in order to effectively identify pathological heart murmurs, it is vital

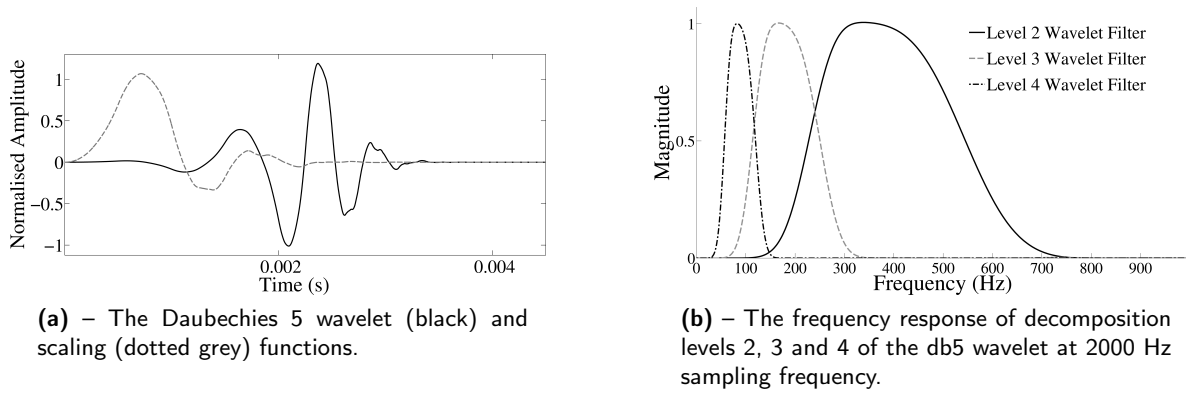


Figure 5.6 – The shape and frequency response of the Daubechies 10-tap wavelet (D10 or db5).

that the short-time spectral content of the PCG is analysed for the detection of such transient high-frequency sounds. Two such time-frequency approaches are outlined below.

5.2.5.1 Wavelet features

Wavelets, which are described in detail in Appendix D, have been employed extensively for heart sound classification (see Section 5.1). For the extraction of time-frequency features from the PCG recordings investigated in this chapter, the Daubechies 10-tap wavelet (D10 or db5) was employed along with the undecimated wavelet transform (UWT). The UWT was selected over the other wavelet transforms as it provides a wavelet decomposition at discrete decomposition levels, leading to a reduction in the feature space, while not halving the number of samples in the signal at each decomposition level. The Daubechies wavelet was selected due to its extensive application to heart sound analysis [125, 126, 136, 137, 142, 146, 232, 233, 234], while the db5 wavelet was found to work well on a portion of the MAAS dataset by other researchers [120]. This wavelet is illustrated in Figure 5.6.

Prior to wavelet feature extraction, each PCG recording was normalised by subtracting its mean and dividing by its standard deviation. All recordings were then down-sampled to 2 kHz using a poly-phase anti-aliasing filter. The frequency content of heart sounds and murmurs lies below 1 kHz, and hence the Nyquist-Shannon sampling criterion [164] was satisfied. The detail coefficients from three decomposition levels were extracted as features. These were decomposition levels two, three and four, which were selected based on their centre frequencies and the frequency responses of the high-pass filters in the UWT. These frequency responses

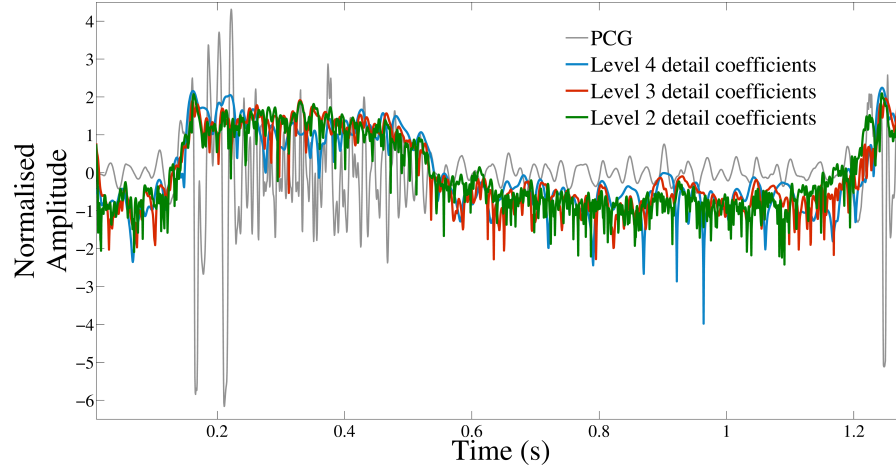


Figure 5.7 – Wavelet features derived from a single heart cycle of a patient with a severe systolic murmur. This is the same patient whose data are shown in Figure 5.1.

can be seen in Figure 5.6b, the centre frequencies being 333 Hz, 167 Hz, and 83 Hz for levels two, three and four respectively.

The frequency band covered by the second decomposition level (centre frequency of 333 Hz) is well-suited for pronounced murmurs due to MVP in PCG recordings, as these sounds have a frequency range above 200 Hz and up to 550 Hz [107]. The third decomposition level, with a centre frequency of 167 Hz, is well suited to the detection of the frequency components in the lower range of pathological murmurs. Finally, the fourth decomposition level, with a centre frequency of 83 Hz, is well-suited for the detection and differentiation of innocent murmurs from pathological ones, as innocent murmurs commonly have low frequency content (below 200 Hz), and occur in early- to mid-systole [107, 108].

These features were converted to decibels in order to emphasise the low-amplitude fluctuations (as compared to the S_1 and S_2 sounds in the signal), and subsequently normalised for each recording by subtracting their mean and dividing by their standard deviation. An example of these three features (in the same patient whose data is shown in Figure 5.1) is given in Figure 5.7.

5.2.5.2 Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCCs) have been used extensively in speech recognition. These features were originally motivated by the need to *deconvolve* the output of the vocal folds and the impulse response of the vocal tract. Such a deconvolution was performed in the

cepstral domain, where the *cepstrum* was originally defined as the squared magnitude of the inverse Fourier transform of the logarithm of the squared magnitude of the Fourier transform of a signal [148], which in the discrete case, can be written as:

$$c_n = \left| \frac{1}{N} \sum_{k=1}^N \log(|X(\omega)|^2) \cdot \exp(j \frac{2k\pi}{N} n) \right|^2 \quad (5.1)$$

where $X(\omega)$ is the discrete Fourier transform of signal $\mathbf{x} = (x_1 \dots x_N)$. In speech processing, cepstral analysis is often combined with filterbank-based partitioning of the frequency spectrum. In particular, partitioning based on the mel-frequency scale is most often employed, as such mel-spacing approximates the human auditory system's response to sounds of different frequencies. The mel-scale is defined as:

$$mel(f) = 1125 \ln(1 + f/700), 0 \leq f \leq Fs \quad (5.2)$$

where Fs is the sampling frequency. The application to cepstral analysis of the mel-frequency-based filterbank with overlapping filters gave rise to MFCCs [280]. The basic steps in the derivation of MFCCs are:

1. Window a short frame of a signal using a windowing function
2. Find the discrete Fourier transform of the windowed signal
3. Apply the mel-spaced filterbank to the amplitude spectrum, summing the energy in each filter
4. Find the logarithm of the filterbank energies
5. Find the discrete cosine transform (DCT) of the log filterbank-weighted amplitude spectrum

The MFCCs can then be written as:

$$MFCC_n = \sum_{k=1}^K E_k \cos(n_{k-0.5\pi/K}), n = 0, \dots, L \quad (5.3)$$

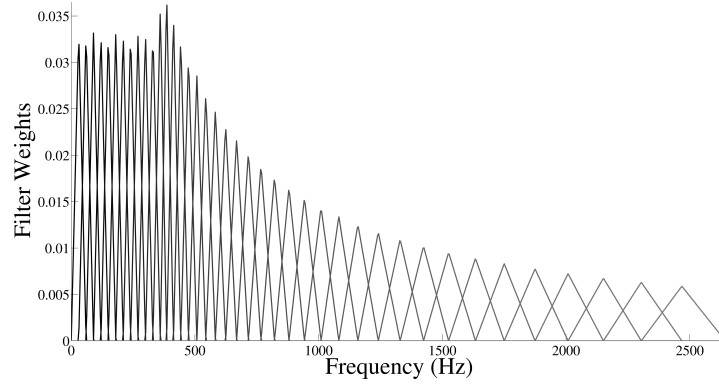


Figure 5.8 – MFCC filterbank with 13 linear filters followed by 27 logarithmically-spaced filters.

where L is the number of MFCC coefficients (typically 12-16 are used in speech recognition), and E_k is the log energy of the k^{th} filter in the filterbank, (typically $K = 20$ to 40) [281].

Prior to MFCC feature extraction, each PCG signal was normalised by subtracting its mean and dividing by its standard deviation. All recordings were then downsampled to 4 kHz using a poly-phase anti-aliasing filter. Then, frames of 256 samples, with 10 sample steps between frames, were extracted. This duration was selected as it was shorter than the shortest expected heart sound. The mean S_2 duration is 92 ms with a standard deviation of 22ms (see Section 4.2.3). Therefore, a 256-sample frame with a sampling frequency of 4 kHz, corresponding to a duration of 64 ms, is shorter than at least 84% of S_2 sounds (given that $92 - 22 = 70$ ms is one standard deviation from the mean S_2 duration). These frames were windowed using a Hamming window before finding the 256-sample discrete Fourier transform (leading to 16 Hz frequency resolution). A mel-filterbank of 40 filters with 50% overlap was applied to the amplitude spectrum. Of these 40 filters, the initial 13 were linearly spaced for greater resolution at lower frequencies of more interest in PCG analysis. The remaining 27 were logarithmically spaced. The filterbank characteristics are shown in Figure 5.8.

After finding the logarithm of the 40 filterbank energies, the 13-element discrete cosine transform (DCT) was computed. The first MFCC coefficient represents the signal energy, while the higher-order MFCC coefficients give an indication of the *shape* of the log amplitude spectrum, described in terms of the “*similarity*” in shape between the spectrum and the cosine basis functions of the DCT, described mathematically in Equation 5.3. In speech processing, it is common to discard the first MFCC coefficient and higher order MFCC coefficients. Following this, and in order to reduce the feature space, MFCC coefficients two, three and four were kept

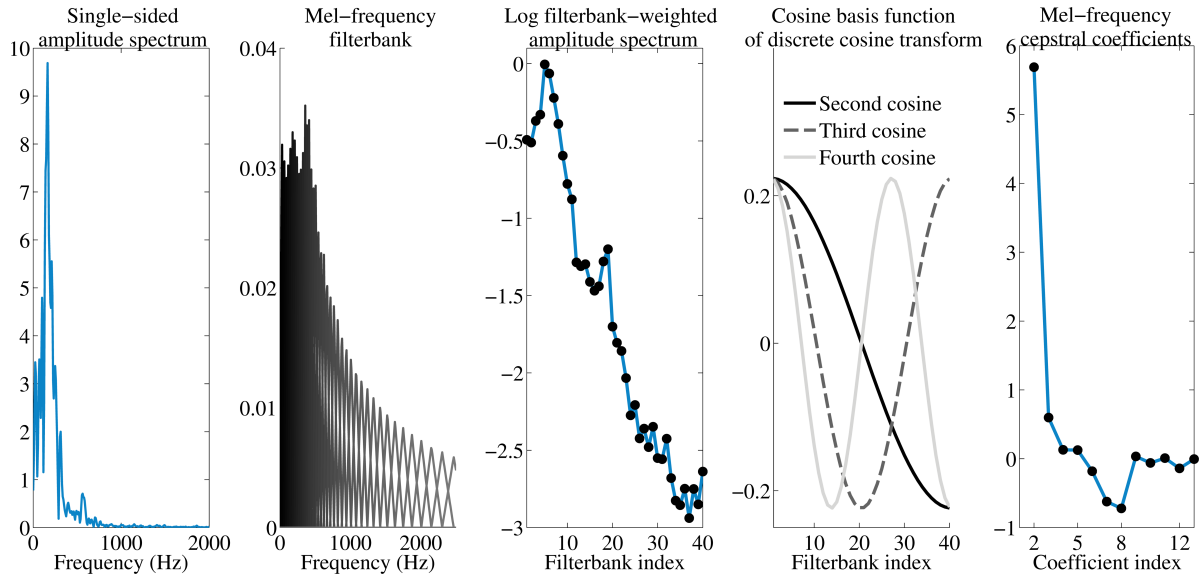


Figure 5.9 – Illustration of the steps for MFCC computation. The amplitude spectrum of a short, Hamming-windowed frame is first computed, shown in the first figure from the left. The values in the amplitude spectrum are weighted by the 40 filters in the filterbank (shown in the second figure), to produce the log filterbank-weighted amplitude spectrum (third figure). The discrete cosine transform (DCT) of these 40 values is then found, which involves finding the sum of the filterbank coefficients multiplied by different cosine basis functions (the second, third and fourth cosine basis functions are shown in the fourth figure). This ultimately leads to the *similarity* between the cosine functions and the log filterbank-weighted amplitude spectrum, which leads to the mel-frequency cepstral coefficients. This example shows the computation of MFCCs in a frame containing a murmur. The MFCC extraction for a frame without a murmur is shown in Figure 5.10.

for further analysis.

The process of MFCC computation^f for a frame containing a pathological murmur is shown in Figure 5.9. This figure illustrates the extracted amplitude spectrum, the filterbank, the log filterbank-weighted amplitude spectrum, the second, third and fourth DCT basis functions, and the final MFCC coefficients. The similarity in shape between the log filterbank-weighted amplitude spectrum and the second cosine basis function can be seen, leading to a large second MFCC coefficient value.

The extraction of the MFCCs for a frame without a murmur is shown in Figure 5.10. The lower-amplitude log filterbank-weighted amplitude spectrum across all frequencies can be seen to lead to smaller second MFCC coefficient value, and larger third and fourth coefficients.

The extracted second, third and fourth MFCC features were normalised for each recording by subtracting their mean and dividing by their standard deviation. An example of these three features (in the patient whose data are shown in Figure 5.1) is given in Figure 5.11.

^fThe MFCC implementation in the Auditory Toolbox, developed by Malcolm Slaney [282], was employed for this analysis.

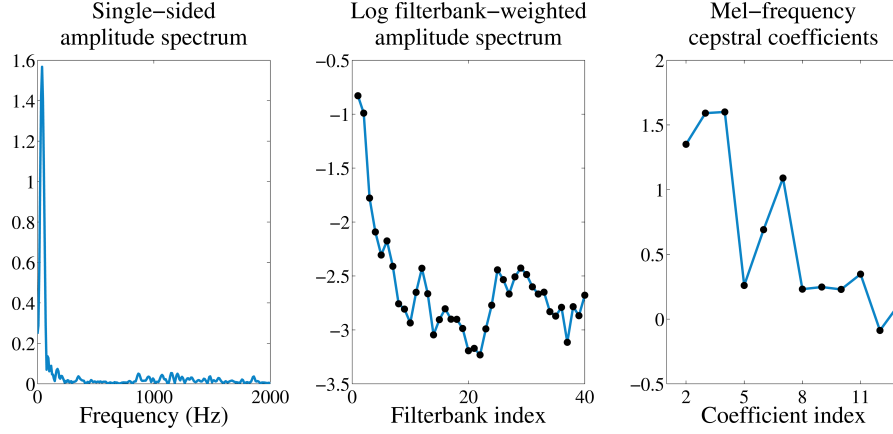


Figure 5.10 – MFCC computation in the absence of a murmur, showing the extracted amplitude spectrum, the log filterbank-weighted amplitude spectrum and the extracted MFCCs.

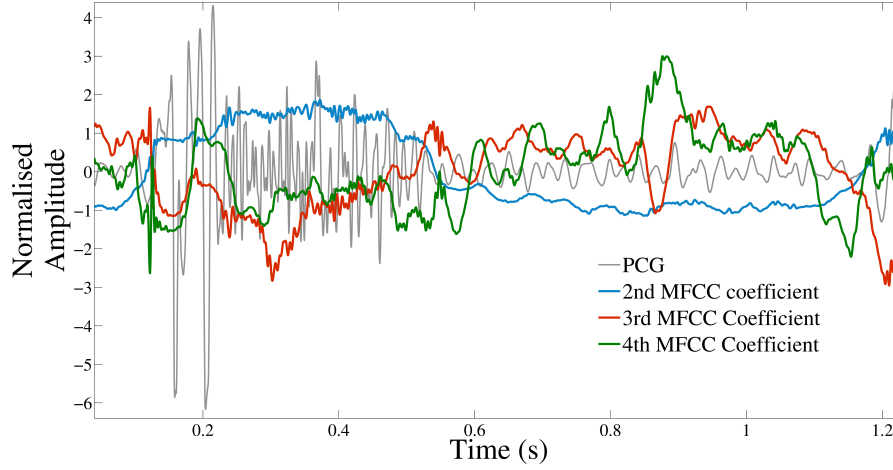


Figure 5.11 – MFCC features derived from a single heart cycle of a patient with a severe systolic murmur. These data are extracted from the patient whose data have been previously shown in Figures 5.1 and 5.7.

Following the extraction of features from each heart sound signal, the features from each systolic period were segmented using the end- S_1 and start- S_2 times derived from the segmentation algorithm employed in Section 5.2.3. Only the systolic portion of each cycle was extracted, as the murmurs in the pathological cases in each of the datasets analysed in this chapter were expected to occur during systole.

5.2.6 Cycle averaging

It is well known that the averaging of repeated signals in the time domain leads to an improved signal-to-noise ratio. A signal, x_t , composed of both the true signal, s_t , and noise, n_t , can be written as $x_t = s_t + n_t$. If this signal is repeated over N measurements, then x_t^i represents the signal's value at time t for measurement i , $i = 1, 2, 3 \dots N$. The average measurement of x_t , $\bar{x}_t$,

can be computed using:

$$\overline{x_t} = \frac{1}{N} \sum_{i=1}^N x_t^i = \frac{1}{N} \sum_{i=1}^N [s_t^i + n_t^i] \quad (5.4)$$

In the ideal case, if the signal and the noise are completely uncorrelated, the signal strength is constant in the repeated measurements and the noise is random (with zero mean and constant variance), then the signal-to-noise ratio will increase by $\sqrt{N}$, where N is the number of averaged measurements [283].

To minimise noise in the features extracted from the systolic segments of the PCG recordings, the features were averaged over all the identified cycles in the heart sound recording. This process has the intrinsic assumption that the features indicative of the events of interest, such as murmurs in the PCG recording, are present in each averaged heart cycle. While it is known that the amplitude of murmurs varies during PCG recordings (due to changes in ventricular filling time, breathing, slight changes in blood pressure etc.), the pathological valve in the heart causing the murmur is involved in each heart cycle. Hence, the presence of features indicative of such pathology can be expected to be present in each heart cycle, even if they may be of lower amplitude in some heart cycles.

Further, again as a result of the cyclical nature of the heart cycle, the same pathological valve is involved at the same point in each heart cycle. For instance, a regurgitant mitral valve closes at the commencement of systole and begins to leak when the left ventricular pressure exceeds a threshold. Pathological events therefore start at the same point in each cycle. To compensate for the natural fluctuations in systolic duration because of heart rate variability, each systolic period was resampled to the median systolic duration in the PCG recording using a polyphase anti-aliasing filter. While methods such as dynamic time warping have been employed for time alignment in PCG analysis [120, 133], it was found that such methods were noise sensitive, and often aligned segments of the systolic periods on the basis of spurious spikes in noisy recordings.

To assess the impact of such cycle averaging on the accuracy of pathology classification, two methods of classification were followed: with the first method, *each* cycle from a PCG recording is classified, and then majority voting took place using all the classifications in the PCG recording; with the second method, the averaging process over all heart cycles in a recording led to a single classification for the averaged heart cycles.

5.2.7 Noisy heart cycle rejection

Noise corruption in the PCG has an obvious detrimental effect on the identification of heart pathology. Common noise sources include motion artefacts, physiological sounds caused by breathing and digestive movements, talking and medical equipment noise (such as in a hospital or clinic setting). Such noise corruption often overlaps with the sounds of interest, such as murmurs and clicks, in the frequency domain, and therefore filtering techniques have little or no effect. Instead, prior to signal averaging, noisy heart cycles were identified and removed from the averaging process, with the aim of providing the most robust estimation of presence or absence of pathology.

The noisy cycles were identified by finding the systolic period features which had a large deviation from the median systolic period features. Firstly, the median systolic features were computed by finding the median value at each sample of the time-aligned features over all heart cycles. Then, the root-mean-square (RMS) error between the features of the median systole and each individual systole was found. In order to compute an exclusion threshold, the mean and standard deviation of the RMS errors was computed. Those systoles which had an RMS error greater than one standard deviation above the mean were excluded from further processing.

Figure 5.12 shows a heart cycle excluded during this process. The RMS error of each cycle is shown, as well as the exclusion threshold. A noisy heart cycle, with a motion artefact, can be seen to have a large RMS error, and is therefore excluded from the cycle averaging procedure.

5.2.8 Selection of cycles based on k -means clustering

As mentioned in Section 5.2.6, it is well known that the amplitude of pathological sounds may vary between the heart cycles in a PCG recording. This is due to factors such as changes in ventricular filling time, breathing and slight changes in blood pressure. In an attempt to overcome this variation, and to identify those cycles of the PCG recording which were most representative of the underlying pathology, k -means clustering, to group cycles with similar time-frequency characteristics, was introduced. This evaluation is based on the clustering approach proposed by other researchers [120].

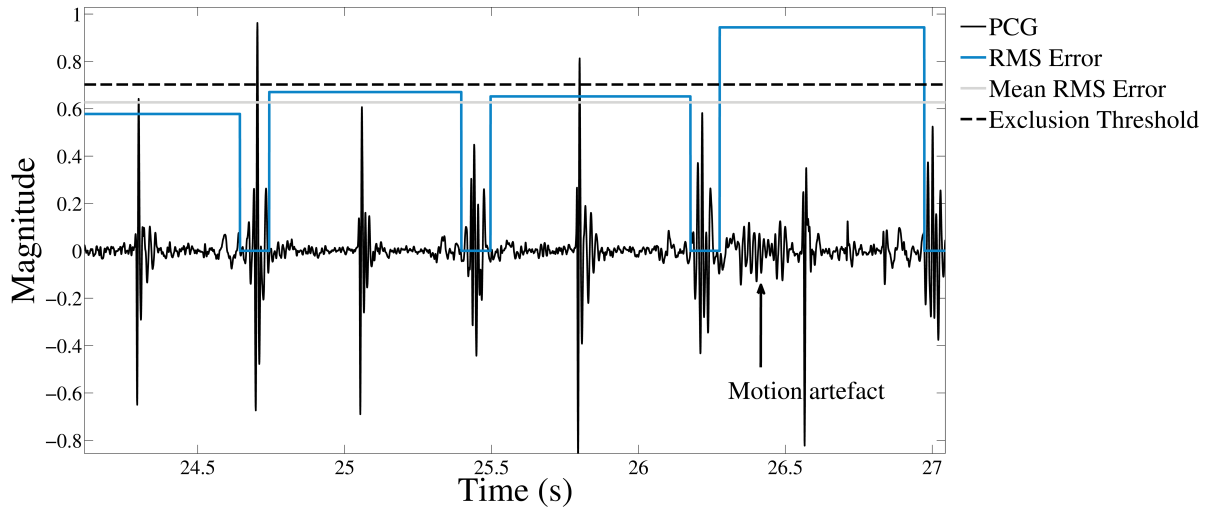


Figure 5.12 – Cycle exclusion based on RMS error from the median systole. The PCG is shown in black, while the RMS error values between the features from the median systole and each cycle are shown in blue (with zero values showing each S_1 sound). The mean RMS error value is shown as a horizontal grey line, while the exclusion threshold (mean RMS plus one standard deviation) is shown as a horizontal dotted line. The heart cycle to the right of the plot, with a labelled motion artefact, can be seen to exceed the exclusion threshold, and is subsequently removed from further analysis.

For each heart cycle, the median feature values in the second half of systole were found. The second half of systole was selected, as this avoids the inclusion of innocent murmurs, which occur during the first half of systole. Of the three wavelet features, those from decomposition levels two and three were used, as these higher-frequency features will highlight the presence of a murmur. Similarly for the MFCC features, those from coefficients three and four were used. Then, k -means clustering, with two cluster centers ($k = 2$), was performed. This is a simple unsupervised clustering technique, where the initial cluster centres are randomly placed in the feature space. Observations are then grouped according to their nearest cluster centre (based on the squared Euclidean distance), and then new cluster centres are computed by finding the centroids of the grouped observations. This computation of the position of the new cluster centre was repeated 100 times. The motivation for the use of two clusters was to group those heart cycles most representative of the underlying heart condition (if any), and those containing noise or low-amplitude murmurs in another cluster.

The cluster with the greatest membership, or greatest *cardinality*, was then identified. Then, rather than only selecting the observations (or cycles) associated with that cluster centre, the Euclidean distance of each observation from the selected cluster centre was computed. Then, those observations within one standard deviation of all the Euclidean distances from the selected cluster centre were retained, while those cycles greater than one standard deviation away from

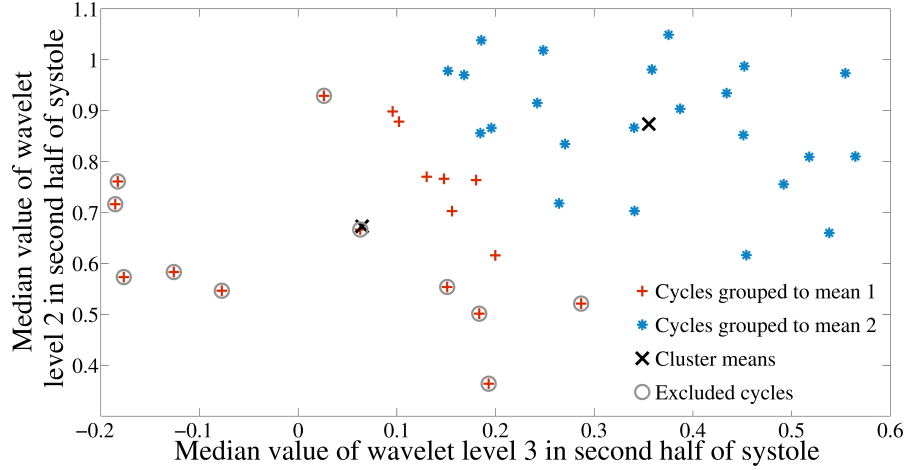


Figure 5.13 – Example of k -means clustering-based selection of heart cycles. The feature values of each heart cycle in a PCG recording can be seen, grouped by the cluster centre with the smallest squared Euclidean distance. The cluster mean in the top right of the figure has greatest membership, and is therefore selected. Those cycles with a Euclidean distance outside of one standard deviation of the distances from the selected mean are excluded, shown with grey circles. These data are from the same subject whose data are shown in Figure 5.14, with a large late-systolic murmur. The k -means clustering procedure can be seen to exclude those cycles with lower-amplitude murmurs, while selecting those with higher frequency content.

the selected cluster centre were excluded. This was carried out so that those observations (or cycles) with the greatest similarity to the majority of the cycles in a PCG recordings were kept, while those that were dissimilar were discarded. An example of such k -means clustering is shown in Figure 5.13. The impact of the inclusion of this k -means clustering-based cycle selection on the classification accuracy was evaluated on all three datasets investigated in this chapter.

5.2.9 Feature downsampling

The features extracted from each systolic interval contain a large number of samples at a sampling frequency of 2000 Hz. Even in very short systoles of 0.1s, the resulting 200 samples from each feature vector would require a large number of training examples to train a classification model effectively. To reduce the dimensionality of the feature vector, and to standardise its length, each systole was therefore reduced to three non-overlapping windows.

The use of three windows was based on the clinically-based classification of systolic murmurs. Systolic murmurs (that are not holosystolic) are classified as either early, mid or late-systolic (see Section 1.3.4.1). Therefore, a natural classification scheme would follow such categorisation, and use three equally-spaced systolic windows. The median value within these three equally-spaced non-overlapping windows, from each of the three extracted wavelet or MFCC

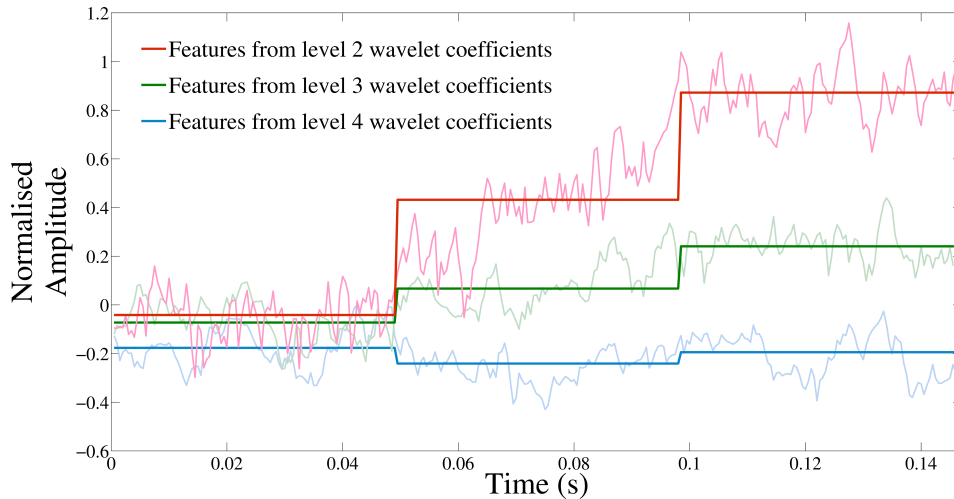


Figure 5.14 – Feature downsampling to three systolic windows of a patient with a late-systolic murmur indicative of MVP. The features derived from the systolic period of a heart cycle are shown. The median values for each window, for each wavelet coefficient vector, are shown in bold steps. The original feature vectors, resulting from averaging after cycle exclusion and k -means clustering (shown in Figure 5.13), are shown as faded lines (of the colour of the corresponding median windows).

feature vectors, then formed the nine-element feature vector for each heart cycle (if averaging was not employed) or heart sound recording (if cycle averaging was performed). An illustration of this downsampling is shown in Figure 5.14.

5.2.10 Data labelling

Binary classification was employed to differentiate between PCG recordings exhibiting functional heart damage and PCG recordings from healthy subjects, with the pathological records labelled as the positive class, while “healthy records” were labelled as the negative class. Pathology classification was performed on a record-by-record basis, rather than by grouping a particular patient’s PCG recordings. For instance, a subject with three PCG recordings would present three different instances to classify pathology, rather than one⁹.

For the MAAS dataset, binary labelling was provided by the echocardiographic labels assigned in the original dataset. In the RHD datasets, there were various stratifications within the dataset, namely definite RHD, borderline RHD and no RHD, with various sub-classes within each category. Importantly, some sub-categories of the borderline RHD diagnoses had only morphological pathology (and no functional valve damage, and hence no pathological regurgitation), while two cases with no RHD had functional heart pathology, leading to abnormal blood

⁹Importantly, all recordings from the same patient were included in the same fold of the 10-fold cross-validation discussed in Section 5.2.11, in order to avoid biasing the classification model.

flow. As the functional pathology only would be detectable using a stethoscope, cases with such heart pathology were labelled as pathological, while those cases (including the borderline RHD cases of sub-category A – see Box 1 in Section 2.3) with only morphological valve damage or no valve damage were labelled as healthy. The pathological cases included the two cases of individuals who did not have RHD, but who did have other pathological blood flow in the heart (due to VSD and PDA). The labelling of borderline cases of RHD without functional valve damage as “healthy” can be justified by the fact that it is still unknown whether such cases of *sub-clinical* RHD progress to functional valve damage later [49, 58, 64, 68].

When cycle averaging was not employed (leading to multiple feature vectors per record from each heart cycle), majority voting over the number of cycles in a record was performed to derive a final classification label for each record. If the numbers of classified pathological and healthy cycles in a particular record were equal, that record was labelled as pathological, therefore biasing the classifier towards higher sensitivity.

5.2.11 Classification methods and evaluation

A support vector machine (SVM), described fully in Appendix C, was employed as the classification algorithm in this work. SVMs have been used by many researchers for the classification of heart pathology in PCGs [142, 234, 260, 261, 264, 265, 266, 270, 271, 274, 275, 276]. However, as outlined in Section 5.1.3, the vast majority of these studies implemented an SVM classifier with poor methodology, by either tuning the SVM parameters manually or over-training their models without an adequate train-test split in their data. Of the 12 studies identified, only two are thought to have been implemented rigorously [271, 276]. However, even in these studies there is a question about the SVM parameter selection and segmentation of the original data.

This chapter presents a rigorous implementation of SVM-based PCG classification^h. In particular, an SVM with a radial basis function (RBF) kernel was selected. With such a model, there are two parameters to be optimised. These are C , which controls the penalty on errors in the SVM classification, and γ , which controls the width of the Gaussian kernel in the SVM. Both of these parameters contribute to the flexibility of the model (see Appendix C), with the optimal parameters depending on the specific classification task. As suggested by Hsu et

^hThe LIBSVM package was employed for SVM implementation [177].

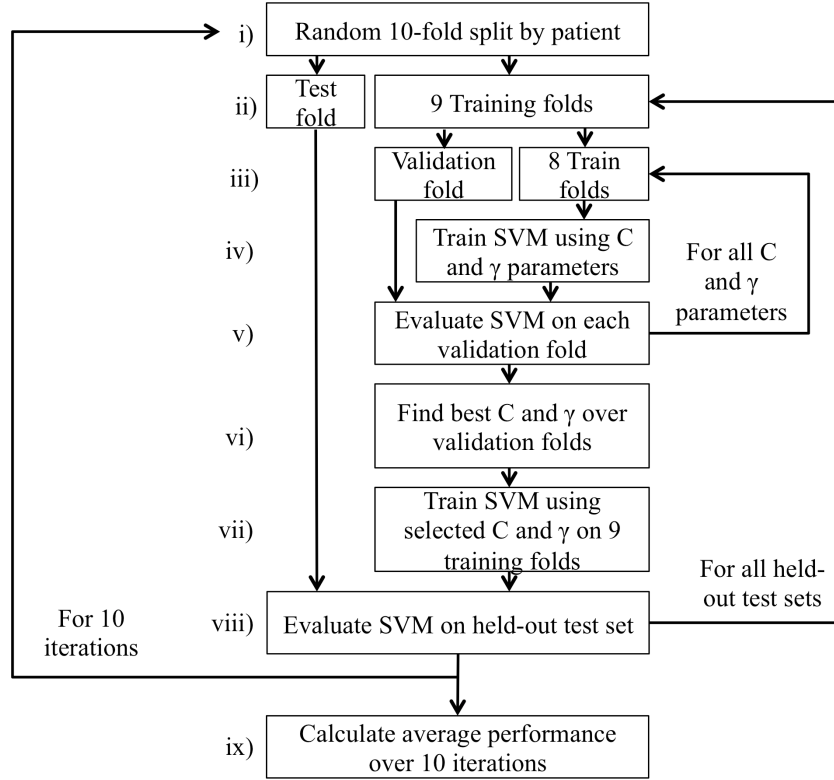


Figure 5.15 – Evaluation procedure for the SVM classification algorithms. This figure illustrates the 10-fold cross-validation, the optimisation of the SVM parameters on the validation folds, while the algorithms are evaluated on the held-out test sets. This entire procedure is repeated for 10 iterations.

al. [181], a grid-search over exponentially growing sequences of C and γ was performed, with C and γ equal to 2^n , $n \in \{-5, -4.75, -4.5, \dots, 1\}$. These ranges were determined through initial experimentation.

In order to perform a rigorous assessment of the classification algorithm and avoid over-training the model, the evaluation of the classification method was performed using 10-fold cross-validation. The folds were created so as to ensure that records from the same patient were all included in the same fold (and therefore never ended up in both the training and test sets). The patients from each class (healthy and pathological) were randomly assigned to one of ten folds. Each fold was iteratively selected as a held-out test fold. The splitting of patients into 10 folds, and the held-out test fold, are shown in steps *i* and *ii* in Figure 5.15.

For each test set fold, the grid search for the optimal SVM parameters was performed using the remaining nine folds (shown in step *ii* in Figure 5.15). One of the nine folds was used as a validation fold (step *iii* in Figure 5.15), and an SVM classifier was trained using a pair of C and γ parameters on the remaining eight training folds (step *iv*) and then evaluated on the separate validation fold. This was repeated for each validation fold (step *v*), and the average

F_1 score over all the validation folds for the values of that pair of C and γ parameters were stored.

The algorithm then looped back to step *iii* until all C and γ parameter values had been evaluated. The C and γ parameters which led to the highest average F_1 score over the validation folds were then selected (step *vi*). An SVM classifier was then trained using the selected C and γ parameter values, using the data from all nine training folds as input features (step *vii*), and evaluated on the held-out test fold (step *viii*).

The algorithm then looped back to step *ii*, and repeated the above process for a different test fold. The results from the 10 test folds were averaged in order to compute the overall classification results.

To perform a rigorous evaluation of the proposed methods across different groupings of patients within folds, the algorithm looped back to step *i*, and again randomly assigned patients to 10 different folds, and the above procedure was repeated. This was repeated 10 times. This ensured that patients were grouped in different combinations within the 10 folds. Then, the classification performance metrics were averaged across all 10 iterations (step *ix* in Figure 5.15), using *sensitivity* ($Se = TP/(TP + FN)$), *specificity* ($Sp = TN/(TN + FP)$), *balanced accuracy* ($BA = (Se + Sp)/2$) and F_1 score, which is defined as

$$F_1 = \frac{2 \times P_+ \times Se}{P_+ + Se} = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (5.5)$$

where P_+ is positive predictivity ($P_+ = TP/(TP + FP)$), and TP , FP , TN and FN are the number of *true positives* (correctly identified pathology), *false positives* (falsely detected pathology), *true negatives* (correctly identified lack of pathology) and *false negatives* (missed pathology), respectively. A metric such as *accuracy* gives a false indication of the true performance of a classifier on an imbalanced dataset. The BA and F_1 score were used instead, as these metrics give the arithmetic mean of Se and Sp and harmonic mean of Se and P_+ respectively, and give a more rigorous indication of the true performance on imbalanced datasets.

In order to compare the non-linear decision boundary formed by the SVM classifier with a simpler linear decision boundary, regularised LASSO logistic regression (LR) was employed. LASSO LR is explained in more detail in Appendix B, and has already been implemented in Chapter 3 for the classification of PCG signal quality. After the best-performing combination

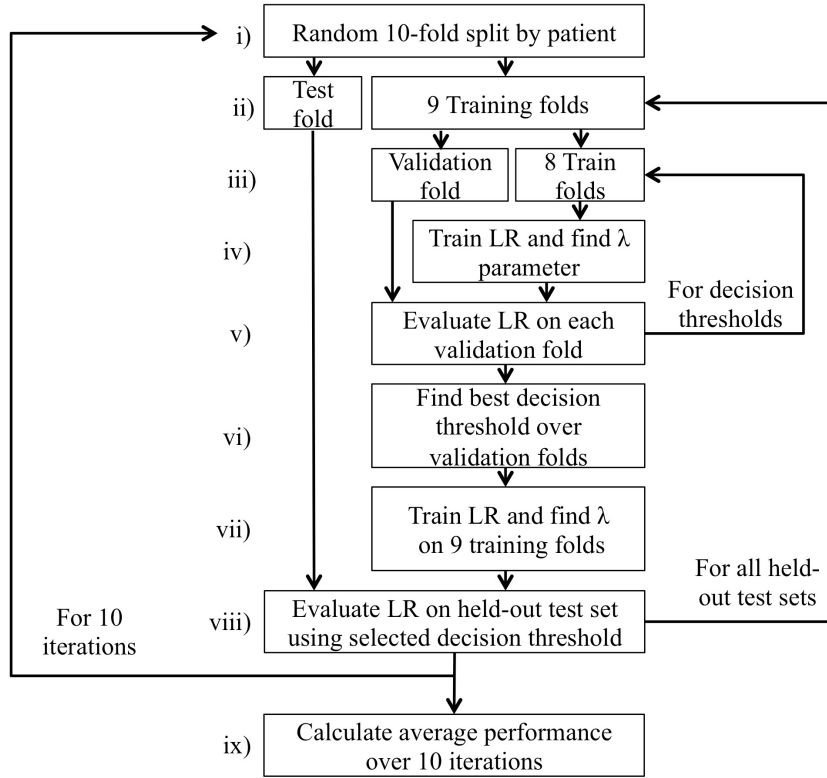


Figure 5.16 – Evaluation procedure for the LASSO LR classification algorithms. This figure illustrates the 10-fold cross-validation, the optimisation of the decision threshold on the validation folds, while the algorithms are evaluated on the held-out test sets. This entire procedure is repeated for 10 iterations.

of features and processing steps had been found for each dataset when using the SVM classifier, the same features and processing steps were evaluated with LASSO LR.

The evaluation of the LASSO LR model followed the same 10-fold cross-validation approach as for the SVM evaluation, as shown in Figure 5.16.

After randomly assigning patients to one of 10 folds (step *i* in Figure 5.16), for each held-out test fold, the remaining nine folds were used to train the LASSO LR model and to compute the decision boundary between the imbalanced classes (step *ii*). The optimal λ parameter for the LASSO LR model (which controls the degree of regularisation in the model – see Appendix B for an explanation of this variable, and Section 3.2.6.3 for an explanation of the range of λ values tested) was selected by employing cross-validation across eight of the nine training folds (step *iv* in Figure 5.16). The validation folds were used to set the decision threshold (steps *v* and *vi* – discussed in the next section). Once a decision threshold was selected, a LASSO LR model was trained on all nine training folds (step *vii*, with nine-fold cross-validation being used to set the λ parameter), and evaluated on the held-out test fold using the selected decision threshold (step *viii*). This was repeated for all 10 held-out test folds (looping back from step

viii to step *ii*). As with the SVM evaluation method, the random assignment of patients to test folds was repeated 10 times (looping back to step *i*), with the average results over all 10 iterations giving the final classification results (step *ix*).

5.2.12 Feature normalisation and class imbalance

Feature normalisation was performed by subtracting the mean and dividing by the standard deviation of all examples for that feature. The mean and standard deviation values were computed from the training set data, and the same mean and standard deviation were used to normalise the held-out folds.

Class imbalance in the training datasets when using an SVM can lead to the separating hyperplane being skewed towards the minority class [179]. This can lead to poor performance, and in the case of the datasets in this chapter where the minority class is the pathological class, poor sensitivity. Class imbalance was dealt with by using an SVM with different C values for each classⁱ. In such a model, the C parameter is modified to assign two misclassification costs: C^+ is the misclassification cost for positive class (pathological) examples, while C^- is the misclassification cost for the negative class [284]. In such a model, Equations C.8 and C.9 become:

$$\arg \min_{\mathbf{w}, b, \xi} \frac{\|\mathbf{w}\|^2}{2} + C^+ \sum_{i|y_i=+1}^{N_{y_i=+1}} \xi_i + C^- \sum_{i|y_i=-1}^{N_{y_i=-1}} \xi_i \quad (5.6)$$

$$\text{subject to :} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (5.7)$$

where $N_{y_i=k}$ is the number of training examples of class k , $k \in [-1, 1]$. In each SVM in this chapter, C^- , or the healthy class weight, was set to the optimised C parameter, while $C^+ = C \times (N_{y_i=-1}/N_{y_i=+1})$.

When using the LASSO LR model, class imbalance affects the model intercept, and therefore, a naive selection of a decision threshold value of 0.5 would lead to low sensitivity. The decision threshold needs to be optimised on a held-out validation set. A range of decision thresholds of $\{0.1, 0.125, 0.15, \dots, 0.6\}$ were evaluated on the held-out validation folds, with the decision threshold that led to the highest average F_1 score across the validation folds being selected. A

ⁱThis is also termed a *weighted SVM* [177] or an SVM with *different error costs* [179].

LASSO LR model was then trained on all nine training folds (with nine-fold cross-validation being used to set the λ parameter), with the selected decision threshold being used to evaluate the model on the held-out test fold.

5.2.13 Baseline methods for comparison

The feature extraction procedures presented in this chapter were compared to three methods from the literature. These were:

1. The sum of squared wavelet detail coefficients (db2 and second decomposition level) in 32 equally-sized windows over the entire heart cycle, as initially performed by Ölmez and Dokur [232] and repeated by a number of researchers [125, 139, 143, 234]. This method was reported to achieve 100% [143, 232, 234] and 96% [125] accuracy when detecting pathological systolic murmurs, while achieving 99.279% 13-class accuracy [139] with various classifiers. However, these studies were marred by poor validation methodology (either a lack of train-test split and/or the same patients in both training and test sets).
2. The method employed by Zheng et al. [274], who used wavelet packed decomposition using the db10 wavelet up to decomposition level 5. Two different signals were reconstructed from the wavelet packet nodes which were expected to hold frequency content from a : the fundamental heart sounds (nodes 0-5 for S_1 and S_2) and b : murmurs (nodes 6 to 24). Then, six features were extracted: the ratio of energy in signal a to the original signal; the ratio of energy in the signal b to the original signal; the energy in the node of maximum energy and the position of that node; and the sample entropy of both signals a and b . Zheng et al. [274] achieved 97.73% Se and 100% Sp when detecting mitral regurgitation (when using an RBF-SVM with hand-tuned parameters and 10-fold cross-validation).
3. The Shannon entropy of wavelet coefficients, as performed by Uğuz [268]. The wavelet detail coefficients from five levels of decomposition, and approximation coefficients from the final level, were extracted using the db4 wavelet from the entire PCG recording. Then, the Shannon entropy of each detail and approximation vector was computed. Uğuz [268] reported 100% Se , 95.24% Sp and 98.33% accuracy when detecting pathological PCGs.

Table 5.3 – Performance of the cardiologist (when using manual auscultation) and the CardioScan software when detecting RHD patients with functional valve damage as compared to the gold-standard echocardiography diagnosis (%). Se denotes sensitivity, Sp specificity, BA balanced accuracy and F_1 denotes the F_1 score.

	Se	Sp	BA	F_1
Cardiologist (with manual auscultation)	19.4	100	59.7	32.4
CardioScan (using RHD_ES recordings)	6.5	92.9	49.7	10.5

These methods were selected because of their popularity (method 1), their recent appearance in the scientific literature (2) or their reported results when evaluated with a rigorous train-test split (3). In addition, they offer a comparison between other wavelet-based techniques, and the procedures proposed in this chapter. Only the feature extraction procedures presented in these publications were repeated. The classification performance was evaluated using the 10-fold cross-validation SVM-based classification outlined above. In the case of method 1, which produced a set of features per heart cycle, the majority-voting SVM approach presented above was employed.

5.3 Results

5.3.1 Performance of cardiologist and Cardioscan software

The performance metrics of the cardiologist (when using manual auscultation) and the Zargis Cardioscan software when detecting RHD patients with functional valve damage (as compared to the gold-standard echocardiography diagnosis) are shown in Table 5.3. The cardiologist identified all 68 patients without functional valve damage correctly (100% Sp), while identifying six out of the 31 with functional valve damage (19.4% Se). This gives a balanced accuracy of 59.7%, and an F_1 score of 32.4%. The Cardioscan software, which automatically analysed four recordings per patient, achieved 92.9% Sp , falsely labelling five healthy patients (out of 70) as pathological. This software identified two subjects with RHD correctly (out of 31), leading to 6.5% Se . The BA and F_1 score were 49.7% and 10.5% respectively.

5.3.2 Results of baseline methods

The results of the three baseline comparison feature extraction methods, for each of the three datasets, are shown in Table 5.4. These results show the average and standard deviation of the

Table 5.4 – Classification results (mean $\pm$ standard deviation %) from baseline comparison feature extraction methods. The training (grey) and test (black) set results are shown for each of the three datasets. Se denotes sensitivity, Sp specificity, BA balanced accuracy and F_1 denotes the F_1 score.

Method	Dataset	Se	Sp	BA	F_1
1) Ölmez and Dokur [232]	MAAS	83.7 $\pm$ 13.9	25.5 $\pm$ 13.6	54.6 $\pm$ 0.6	25.6 $\pm$ 6.1
		82.3 $\pm$ 13.9	22.2 $\pm$ 12.9	52.3 $\pm$ 2.4	19.7 $\pm$ 6.4
	RHD_ES	61.5 $\pm$ 10.8	53.5 $\pm$ 10.6	57.5 $\pm$ 2.9	38.8 $\pm$ 5.6
		55.9 $\pm$ 12.2	35.8 $\pm$ 11.2	45.8 $\pm$ 2.1	21.9 $\pm$ 4.7
	RHD_MP	56.7 $\pm$ 7.3	67.3 $\pm$ 6.6	62.0 $\pm$ 2.7	57.9 $\pm$ 4.1
		39.8 $\pm$ 7.5	42.6 $\pm$ 7.9	41.2 $\pm$ 3.1	33.5 $\pm$ 6.0
2) Zheng et al. [274]	MAAS	49.1 $\pm$ 2.2	82.7 $\pm$ 1.8	65.9 $\pm$ 1.7	61.2 $\pm$ 1.6
		39.2 $\pm$ 2.5	71.0 $\pm$ 1.5	55.1 $\pm$ 1.5	50.9 $\pm$ 1.2
	RHD_ES	50.2 $\pm$ 10.6	75.8 $\pm$ 5.4	63.0 $\pm$ 3.5	52.0 $\pm$ 2.9
		42.1 $\pm$ 9.2	53.5 $\pm$ 10.1	47.8 $\pm$ 2.6	34.7 $\pm$ 4.2
	RHD_MP	67.9 $\pm$ 12.1	55.1 $\pm$ 12.0	61.5 $\pm$ 1.4	50.6 $\pm$ 8.3
		53.7 $\pm$ 13.9	38.4 $\pm$ 13.1	46.1 $\pm$ 3.1	31.1 $\pm$ 8.1
3) Uğuz [268]	MAAS	50.3 $\pm$ 1.1	81.1 $\pm$ 1.3	65.7 $\pm$ 0.7	60.7 $\pm$ 0.7
		44.0 $\pm$ 2.1	65.1 $\pm$ 2.6	54.6 $\pm$ 1.5	48.6 $\pm$ 1.8
	RHD_ES	25.3 $\pm$ 9.7	82.2 $\pm$ 7.7	53.7 $\pm$ 2.3	43.8 $\pm$ 4.6
		20.8 $\pm$ 8.0	69.1 $\pm$ 8.7	44.9 $\pm$ 2.1	35.2 $\pm$ 2.8
	RHD_MP	55.9 $\pm$ 16.9	54.8 $\pm$ 17.6	55.3 $\pm$ 1.6	41.6 $\pm$ 11.2
		46.2 $\pm$ 17.8	46.7 $\pm$ 19.7	46.4 $\pm$ 3.7	30.7 $\pm$ 12.2

performance metrics across the 10 iterations of the 10-fold cross-validation approach outlined in Section 5.2.11. Results for the training (grey) and test sets (black) are shown. The use of the features proposed by Zheng et al. [274] led to the best performance on the MAAS dataset, with BA of 55.1 $\pm$ 1.5% and F_1 score of 50.9 $\pm$ 1.2%. For the RHD_ES dataset, the features from the work of Ölmez and Dokur [232] achieved the highest BA (45.8 $\pm$ 2.1%), while the feature set of Uğuz [268] achieved best F_1 score (35.2 $\pm$ 2.8%). For the mobile phone dataset, RHD_MP, the feature sets of Uğuz [268] and Ölmez and Dokur [232] achieved the highest BA (46.4 $\pm$ 3.7%) and F_1 (33.5 $\pm$ 6.0%) respectively.

5.3.3 Results of proposed automatic classification algorithms

The results of the proposed automatic classification algorithms presented in this chapter are shown in Tables 5.5, 5.6 and 5.7.

Table 5.5 presents the results of the proposed algorithms on the MAAS dataset. The greatest BA (75.6 $\pm$ 1.6%) and F_1 score (68.7 $\pm$ 2.0%) were obtained using the wavelet features along with cycle averaging and cycle exclusion. This can be seen to outperform all the baseline comparison methods presented in Table 5.4. The inclusion of k -means cycle selection led

Table 5.5 – MAAS dataset average results (mean $\pm$ standard deviation %) for proposed automatic classification algorithms across the 10 iterations of 10-fold cross-validation. The training (grey) and test (black) set results are shown for the various analysis procedures, for each of the wavelet and MFCC features. Se denotes sensitivity, Sp specificity, BA balanced accuracy and F_1 denotes the F_1 score. The best-performing BA and F_1 are shown in **bold**.

Features	Averaging	Cycle exclusion	k -means	Se	Sp	BA	F_1
Wavelets	$\times$	$\times$	$\times$	96.8 $\pm$ 0.7	58.3 $\pm$ 2.7	77.5 $\pm$ 1.2	70.8 $\pm$ 1.8
				93.6 $\pm$ 2.0	47.5 $\pm$ 2.4	70.5 $\pm$ 1.6	58.2 $\pm$ 3.5
	$\checkmark$	$\times$	$\times$	78.7 $\pm$ 2.1	87.3 $\pm$ 0.7	83.0 $\pm$ 1.1	78.3 $\pm$ 1.4
				65.2 $\pm$ 2.0	80.4 $\pm$ 2.2	72.8 $\pm$ 1.4	64.6 $\pm$ 1.9
	$\checkmark$	$\checkmark$	$\times$	82.8 $\pm$ 1.8	87.9 $\pm$ 0.4	85.4 $\pm$ 1.1	81.2 $\pm$ 1.3
				71.3 $\pm$ 1.9	79.9 $\pm$ 2.0	75.6$\pm$1.6	68.7$\pm$2.0
	$\checkmark$	$\checkmark$	$\checkmark$	71.1 $\pm$ 2.4	85.9 $\pm$ 1.6	78.5 $\pm$ 0.7	72.7 $\pm$ 1.0
				63.5 $\pm$ 3.1	81.1 $\pm$ 3.3	72.3 $\pm$ 2.5	63.8 $\pm$ 3.0
	$\times$	$\times$	$\times$	98.0 $\pm$ 0.2	55.9 $\pm$ 0.5	76.9 $\pm$ 0.2	70.1 $\pm$ 0.4
				96.4 $\pm$ 0.7	45.8 $\pm$ 1.9	71.1 $\pm$ 1.1	57.2 $\pm$ 2.7
MFCCs	$\checkmark$	$\times$	$\times$	73.6 $\pm$ 2.5	87.9 $\pm$ 1.9	80.7 $\pm$ 2.0	75.5 $\pm$ 2.5
				62.2 $\pm$ 2.1	80.1 $\pm$ 1.8	71.2 $\pm$ 1.4	61.8 $\pm$ 2.2
	$\checkmark$	$\checkmark$	$\times$	77.1 $\pm$ 3.5	87.8 $\pm$ 1.6	82.4 $\pm$ 2.1	77.7 $\pm$ 2.7
				63.5 $\pm$ 2.1	79.2 $\pm$ 1.8	71.3 $\pm$ 1.5	62.0 $\pm$ 2.3
	$\checkmark$	$\checkmark$	$\checkmark$	77.2 $\pm$ 1.5	88.5 $\pm$ 1.2	82.8 $\pm$ 1.1	78.3 $\pm$ 1.5
				68.1 $\pm$ 1.0	80.2 $\pm$ 1.6	74.2 $\pm$ 1.1	66.5 $\pm$ 1.4

to a decrease in performance when using the wavelet features. The MFCC features gave similar classification performance to the wavelet features, achieving a BA and a F_1 score of 74.2 $\pm$ 1.1% and 66.5 $\pm$ 1.4% respectively when including cycle averaging, cycle exclusion and k -means clustering.

An illustration of the SVM-generated decision boundary when using two of the wavelet features with the MAAS dataset (with cycle averaging and cycle exclusion) is shown in Figure 5.17. The overlap between the two classes can be seen in the figure.

The results of the proposed classification algorithms on the RHD_ES dataset are shown in Table 5.6. Again, the training and test results across the two features and various analysis procedures are shown. The best-performing analysis procedure on this dataset combined MFCC features with cycle averaging and both cycle exclusion methods, to achieve 60.6 $\pm$ 2.1% BA and 49.8 $\pm$ 2.8% F_1 score. This outperforms the cardiologist when diagnosing the same patients (59.7% BA and 32.4% F_1 score from Table 5.3), while also outperforming the Cardioscan software (49.7% BA and 10.5% F_1 score). In addition, this method outperformed all the baseline comparison methods on the same database, as shown in Table 5.4. The wavelet features

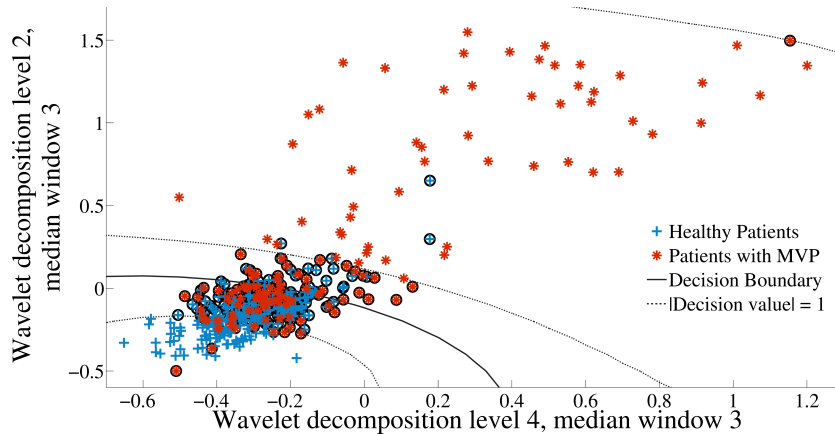


Figure 5.17 – Example of an SVM decision boundary with features derived from the MAAS dataset. The two features are the final median window from the wavelet decomposition levels 2 and 4 (after cycle averaging with cycle exclusion), selected due to the murmur of MVP most often occurring in late-systole. $C = 1$ and $\gamma = 1$ parameters were selected for illustration, and the support vectors selected are shown with black circles. The decision boundary (solid black line) and regions where the decision value was equal to 1 (dotted lines) are shown. The overlap between the two classes is clearly visible.

performed worse than the MFCC features in all analysis procedures. An example of the SVM-generated decision boundary when using two of the MFCC features with cycle averaging, cycle exclusion and k -means clustering in the RHD_ES dataset is shown in Figure 5.18. The overlap between the two classes in this classification task can again be seen.

The results of the proposed classification algorithms on the mobile phone-based RHD dataset, RHD_MP, are shown in Table 5.7. In contrast to the MAAS dataset but similarly to the RHD_ES dataset, the MFCC features led to the best average test-set BA ($74.5 \pm 2.1\%$) and F_1 score ($69.9 \pm 2.8\%$) when using cycle averaging on this dataset. This can be seen to outperform the cardiologist (59.7% BA and 32.4% F_1 score), the Cardioscan software (49.7% BA and 10.5% F_1 score), all the baseline comparison methods of Table 5.4 and the best-performing method on the RHD_ES dataset ($60.6 \pm 2.1\%$ BA and $49.8 \pm 2.8\%$ F_1 score). Cycle exclusion and k -means clustering worsened classification results when using the MFCC feature. Across all analysis procedures, the MFCC features this time outperformed the wavelet features. The best-performing wavelet feature procedure led to $61.7 \pm 2.2\%$ BA and $51.9 \pm 2.8\%$ F_1 score when including cycle averaging after cycle exclusion and k -means clustering.

An example of the SVM decision boundary when using two of the MFCC features (including cycle averaging) with the RHD_MP dataset is shown in Figure 5.19. A more discriminative division between the two classes of data, as compared to the boundaries illustrated in Figures 5.17 and 5.18, is evident from the figure.

To compare the non-linear decision boundary of an SVM to a regularised linear decision

Table 5.6 – RHD_ES dataset average results (mean $\pm$ standard deviation %) for proposed automatic classification algorithms across the 10 iterations of 10-fold cross-validation. The training (grey) and test (black) set results are shown for the various analysis procedures, for each of the wavelet and MFCC features. Se denotes sensitivity, Sp specificity, BA balanced accuracy and F_1 denotes the F_1 score. The best-performing BA and F_1 are shown in **bold**.

Features	Averaging	Cycle exclusion	k -means	Se	Sp	BA	F_1
Wavelets	$\times$	$\times$	$\times$	71.4 $\pm$ 2.5	81.9 $\pm$ 1.9	76.6 $\pm$ 1.5	70.8 $\pm$ 1.6
				56.4 $\pm$ 4.0	48.5 $\pm$ 5.6	52.4 $\pm$ 2.9	40.1 $\pm$ 4.3
	$\checkmark$	$\times$	$\times$	83.2 $\pm$ 4.0	51.8 $\pm$ 2.5	67.5 $\pm$ 1.1	61.3 $\pm$ 2.6
				66.5 $\pm$ 7.2	39.2 $\pm$ 5.2	52.9 $\pm$ 3.7	46.1 $\pm$ 4.9
	$\checkmark$	$\checkmark$	$\times$	81.0 $\pm$ 8.4	52.0 $\pm$ 3.7	66.5 $\pm$ 2.8	59.5 $\pm$ 6.5
				64.4 $\pm$ 8.3	41.1 $\pm$ 7.0	52.8 $\pm$ 3.9	44.4 $\pm$ 6.4
	$\checkmark$	$\checkmark$	$\checkmark$	79.6 $\pm$ 8.5	55.3 $\pm$ 5.0	67.4 $\pm$ 3.9	59.6 $\pm$ 6.5
				54.7 $\pm$ 6.2	42.2 $\pm$ 5.8	48.5 $\pm$ 3.4	37.2 $\pm$ 7.1
	$\times$	$\times$	$\times$	69.3 $\pm$ 2.1	74.8 $\pm$ 1.6	72.0 $\pm$ 1.3	65.6 $\pm$ 1.4
				60.7 $\pm$ 3.7	58.3 $\pm$ 2.1	59.5 $\pm$ 1.9	49.9 $\pm$ 2.8
MFCCs	$\checkmark$	$\times$	$\times$	79.1 $\pm$ 5.0	65.2 $\pm$ 2.8	72.2 $\pm$ 1.3	65.3 $\pm$ 2.8
				61.8 $\pm$ 6.8	55.4 $\pm$ 4.7	58.6 $\pm$ 2.5	49.7 $\pm$ 4.4
	$\checkmark$	$\checkmark$	$\times$	75.3 $\pm$ 5.5	66.1 $\pm$ 4.0	70.7 $\pm$ 1.1	63.5 $\pm$ 2.6
				61.4 $\pm$ 6.8	56.0 $\pm$ 7.0	58.7 $\pm$ 3.9	49.5 $\pm$ 5.0
	$\checkmark$	$\checkmark$	$\checkmark$	69.1 $\pm$ 2.7	70.1 $\pm$ 1.5	69.6 $\pm$ 0.7	62.1 $\pm$ 1.5
				57.7 $\pm$ 4.6	63.5 $\pm$ 3.9	60.6$\pm$2.1	49.8$\pm$2.8

boundary, the best-performing features from each dataset when evaluated with the SVM classifier were this time selected as the input features for the LASSO LR model. The results of the LASSO LR analysis are shown in Table 5.8. While the LASSO LR led to poor results on the RHD_ES and RHD_MP datasets (compared to the SVM method), the LASSO LR classification technique can be seen to outperform the SVM-based classification on the MAAS dataset, achieving 78.7 $\pm$ 0.9% BA and 72.8 $\pm$ 1.0 F_1 score as compared to 75.6 $\pm$ 1.6% BA and 68.7 $\pm$ 2.0 F_1 score when using the SVM (from Table 5.5). For the MAAS dataset, the features in the LASSO LR with the highest β coefficients were the third median window in the 4th wavelet decomposition level (0.16), the first median window in the 3rd wavelet decomposition level (-0.59), and the second and third median windows of the second wavelet decomposition level (1.1 and 1.0 respectively).

5.4 Discussion

This chapter has presented automatic classification algorithms for the detection of systolic heart murmurs indicative of MVP and RHD. These were compared to the classification accuracy of a

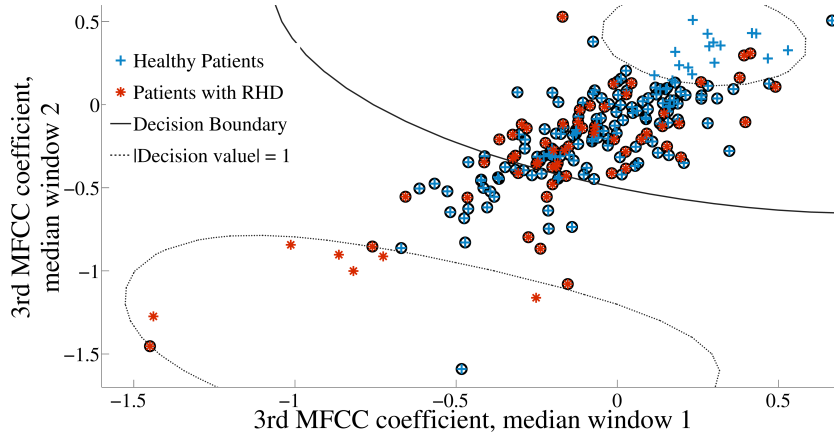


Figure 5.18 – Example of RHD_ES SVM decision boundary for MFCC features. The two features shown are the first and second median windows of the 3th MFCC coefficient, with cycle averaging, cycle exclusion and k -means clustering. $C = 1$ and $\gamma = 1$ parameters were selected for illustration, and the support vectors are shown with black circles. The decision boundary (solid black line) and regions where the decision value was equal to 1 are shown.

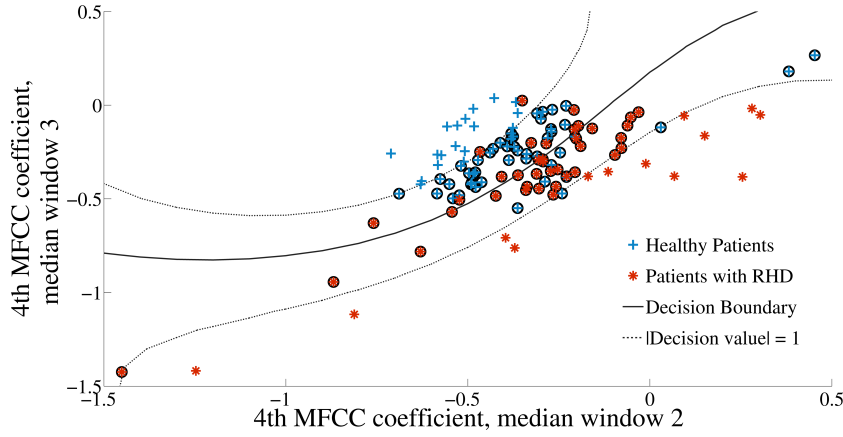


Figure 5.19 – Example of SVM decision boundary for the MFCC features derived from RHD_MP dataset. The two features shown are the second and third median windows of the 4th MFCC coefficient, with cycle averaging. $C = 1$ and $\gamma = 1$ parameters were selected for illustration, and the support vectors selected are shown with black circles. The decision boundary (solid black line) and regions where the decision value was equal to 1 are shown.

cardiologist, the commercial Zargis Cardioscan software and three feature extraction methods from the literature.

The three baseline comparison methods of feature extraction from the literature performed poorly when attempting to detect MVP in the MAAS dataset, as shown in Table 5.4. With a BA of $55.1 \pm 1.5\%$ and F_1 score of $50.9 \pm 1.2\%$, the approach by Zheng et al. [274] led to the best performance of these methods on the MAAS dataset. These results show a classification accuracy of marginally better than a random guess. The reported results of the three baseline comparison methods in the literature are far higher than those achieved on the MAAS dataset, illustrating the difficulty of classifying pathology in this set of recordings as compared to the

Table 5.7 – RHD_MP dataset average results (mean $\pm$ standard deviation %) for proposed automatic classification algorithms across the 10 iterations of 10-fold cross-validation. The training (grey) and test (black) set results are shown for the various analysis procedures, for each of the wavelet and MFCC features. Se denotes sensitivity, Sp specificity, BA balanced accuracy and F_1 denotes the F_1 score. The best-performing BA and F_1 are shown in **bold**.

Features	Averaging	Cycle exclusion	k -means	Se	Sp	BA	F_1
Wavelets	$\times$	$\times$	$\times$	90.1 $\pm$ 1.3	58.6 $\pm$ 4.4	74.4 $\pm$ 2.4	67.9 $\pm$ 3.6
				78.9 $\pm$ 4.7	31.7 $\pm$ 5.3	55.3 $\pm$ 3.4	34.0 $\pm$ 5.3
	$\checkmark$	$\times$	$\times$	55.0 $\pm$ 5.1	75.1 $\pm$ 5.3	65.1 $\pm$ 1.1	57.9 $\pm$ 3.4
				46.4 $\pm$ 8.0	71.2 $\pm$ 5.3	58.8 $\pm$ 2.2	45.5 $\pm$ 5.6
	$\checkmark$	$\checkmark$	$\times$	55.1 $\pm$ 4.6	74.6 $\pm$ 3.0	64.9 $\pm$ 1.4	56.5 $\pm$ 5.1
				48.0 $\pm$ 8.2	70.3 $\pm$ 4.1	59.1 $\pm$ 3.7	46.0 $\pm$ 7.6
	$\checkmark$	$\checkmark$	$\checkmark$	60.9 $\pm$ 3.3	74.2 $\pm$ 2.5	67.6 $\pm$ 0.6	62.5 $\pm$ 1.8
				53.7 $\pm$ 4.3	69.6 $\pm$ 4.6	61.7 $\pm$ 2.2	51.9 $\pm$ 2.8
MFCCs	$\times$	$\times$	$\times$	90.3 $\pm$ 2.4	50.2 $\pm$ 3.5	70.2 $\pm$ 1.4	60.2 $\pm$ 3.0
				83.3 $\pm$ 4.9	31.6 $\pm$ 8.3	57.4 $\pm$ 3.0	36.0 $\pm$ 7.0
	$\checkmark$	$\times$	$\times$	81.2 $\pm$ 1.6	79.3 $\pm$ 0.7	80.3 $\pm$ 0.7	78.2 $\pm$ 0.9
				74.8 $\pm$ 3.9	74.1 $\pm$ 1.9	74.5$\pm$2.1	69.9$\pm$2.8
	$\checkmark$	$\checkmark$	$\times$	81.2 $\pm$ 0.5	75.4 $\pm$ 0.7	78.3 $\pm$ 0.2	76.4 $\pm$ 0.2
				75.1 $\pm$ 3.8	69.3 $\pm$ 3.4	72.2 $\pm$ 2.7	68.0 $\pm$ 3.7
	$\checkmark$	$\checkmark$	$\checkmark$	75.8 $\pm$ 1.2	76.2 $\pm$ 0.9	76.0 $\pm$ 0.6	73.4 $\pm$ 0.7
				66.6 $\pm$ 4.3	71.1 $\pm$ 2.4	68.9 $\pm$ 2.3	62.1 $\pm$ 4.1

datasets presented so far in the literature. This difficulty is illustrated by the overlap in features shown in Figure 5.17. In addition, these poor results highlight the deficiencies in these feature extraction methods. Two of these methods extract features from the entire PCG recording ([268, 274]), while the third extracts cycle-based features but without segmentation of systole and diastole ([232]). This is unlikely to be sensitive enough for the detection of systolic murmurs in real-world, noisy recordings.

When analysing the MAAS dataset, all the SVM-based classification algorithms proposed in this chapter, regardless of the features used, cycle averaging, cycle exclusion and k -means clustering, outperformed all three baseline comparison methods by at least 15% BA and 8% F_1 score (shown in Table 5.5). The best results were obtained using wavelet features with cycle averaging after cycle exclusion (75.6 $\pm$ 1.6% BA and 68.7 $\pm$ 2.0% F_1 score), while k -means clustering with the wavelet features was seen to decrease performance to 72.3 $\pm$ 2.5% BA and 63.8 $\pm$ 3.0%. Conversely, k -means clustering was found to improve the classification when using MFCC features, with marginally worse results than the best-performing wavelet-based classification method (74.2 $\pm$ 1.1% BA and 66.5 $\pm$ 1.4% F_1 score).

Table 5.8 – LASSO LR classification results (mean $\pm$ standard deviation %) of best-performing algorithms on each dataset, averaged over the 10 iterations of 10-fold cross-validation. The training (grey) and test (black) set results are shown for the various datasets and analysis procedures. *BA* denotes balanced accuracy and F_1 denotes the F_1 score.

Dataset	Features	Averaging	Cycle exclusion	k -means	<i>BA</i>	F_1
MAAS	Wavelets	✓	✓	✗	80.1 $\pm$ 0.5	74.5 $\pm$ 0.5
					78.7 $\pm$ 0.9	72.8 $\pm$ 1.0
RHD_ES	MFCCs	✓	✓	✓	51.8 $\pm$ 0.8	46.5 $\pm$ 0.5
					48.6 $\pm$ 1.4	43.8 $\pm$ 1.4
RHD_MP	MFCCs	✓	✗	✗	57.9 $\pm$ 2.4	63.7 $\pm$ 0.8
					51.0 $\pm$ 4.3	56.3 $\pm$ 3.5

The small values for the standard deviation in Table 5.5 illustrate consistent results across the 10 iterations of the 10-fold cross-validation. However, discrepancies of over 10% can be seen between the training and test sets, indicating the potential presence of over-fitting of the model. In contrast, the LASSO LR model can be seen to have a maximum train-test difference of 3% on the MAAS dataset in Table 5.8, which could be due to the “conservative” selection of the λ parameter (the selected deviance is within one standard error of the minimum deviance – see Section 3.2.6.3) and the inherent feature selection in the LASSO LR model. This could explain the best overall results on the MAAS dataset being achieved by the LR model (78.7 $\pm$ 0.9 *BA* and 72.8 $\pm$ 1.0 F_1 score).

With the RHD datasets, the cardiologist and Cardioscan software were very good at identifying those patients without heart pathology (100% and 92.9% specificity respectively). However, both methodologies performed poorly when identifying those patient with valve damage (19.4% sensitivity for the cardiologist and 6.5% for Cardioscan – see Table 5.3). Similarly, all three baseline comparison methods performed poorly on both the electronic stethoscope (RHD_ES) and mobile phone-based (RHD_MP) datasets. Not one of the three baseline comparison methods achieved a *BA* above 50% and F_1 score above 40% for either the RHD_ES or RHD_MP datasets (see Table 5.4). These poor results illustrate the difficulty in identifying the *sub-clinical* RHD murmurs in this dataset. The overlap between the two classes in these datasets is illustrated in Figures 5.18 and 5.19.

The classification algorithms proposed in this chapter outperformed the cardiologist, Cardioscan and baseline comparison methods on the RHD_ES dataset (see Table 5.6). In contrast to the MAAS dataset, the MFCC features resulted in the best performance, while the wavelet fea-

tures performed poorly. The MFCC features outperformed the wavelet features on this dataset by approximately 6% in every case. Cycle averaging, cycle exclusion and k -means clustering-based cycle selection when using the MFCC features led to the best results ($60.6 \pm 2.1\%$ BA and $49.8 \pm 2.8\%$ F_1 score). The LASSO LR model performed poorly on this dataset, attaining $48.6 \pm 1.4\%$ BA and $43.8 \pm 1.4\%$ F_1 score. This could be due to the pronounced overlap between the two classes, leading to a linear decision boundary being unable to separate the two classes.

On the RHD_MP dataset, the algorithms proposed in this chapter were able to attain a BA and F_1 score of $74.5 \pm 2.1\%$ and $69.9 \pm 2.8\%$ respectively. In contrast to the MAAS dataset, but similarly to the RHD_ES dataset, this was achieved using the MFCC features. This algorithm outperformed the cardiologist, Cardioscan software, the three baseline algorithms and the best-performing RHD_ES algorithm. The separation between the two classes when using the MFCC features can be seen in Figure 5.19, with a clearer boundary between the classes as compared to the other two datasets (Figures 5.17 and 5.18). Cycle averaging alone when using the MFCC features resulted in the highest performance, while cycle exclusion and k -means clustering worsened performance (in contrast to the improvement in performance when using MFCC features with the MAAS and RHD_ES datasets). This was thought to be due to the inclusion of fewer heart cycles in cycle averaging after cycle exclusion and k -means clustering leading to a poorer representation of the true underlying pathology. As with the RHD_ES dataset, the LASSO LR model performed poorly on this dataset.

The superior performance of the MFCC features when using a mobile phone-based stethoscope is thought to be due to the frequency response of the mobile phone's sound card being optimised for speech recording and transmission. MFCC features are optimised for speech processing and automatic speech recognition. Hence the use of such features, combined with a sound card optimised for speech processing, may explain the superior performance of the MFCC features when using a mobile phone.

Cycle averaging led to an improved performance in each dataset for both the wavelet and MFCC features, highlighting its ability to exclude noise from the PCG recordings. However, cycle exclusion and k -means clustering did not always lead to improved results. These procedures led to the removal of cycles from a heart sound recording. This removal, if not performed properly, can lead to useful information being removed from the averaged heart cycles, and

Table 5.9 – Classification results (mean $\pm$ standard deviation %) when including poor-quality recordings in the RHD_ES and RHD_MP datasets. The training (grey) and test (black) set results are shown. BA denotes balanced accuracy and F_1 denotes the F_1 score.

Dataset	Features	Averaging	Cycle exclusion	k -means	BA	F_1
RHD_ES	MFCCs	✓	✓	✓	69.1 $\pm$ 0.9	61.8 $\pm$ 1.2
					57.5 $\pm$ 2.1	47.7 $\pm$ 2.5
RHD_MP	MFCCs	✓	✗	✗	75.4 $\pm$ 1.1	69.5 $\pm$ 1.4
					66.2 $\pm$ 1.6	56.3 $\pm$ 2.0

hence, classification errors. Therefore, further evaluation of cycle selection and exclusion algorithms will be necessary.

5.5 Importance of signal quality assessment

In the previous analysis, the poor-quality recordings from both the RHD_ES and RHD_MP datasets were excluded, as these were expected to have a detrimental effect on the classification performance of the proposed algorithms. To test this assumption and to assess the importance of the signal quality assessment of PCG recordings, the poor-quality recordings from each of these datasets were included, and the best-performing algorithms from the previous analysis were re-run on the larger sets of data. The results of this analysis are shown in Table 5.9.

On the RHD_ES dataset, the BA and F_1 score can be seen to drop to 57.5 $\pm$ 2.1% and 47.7 $\pm$ 2.5% from the best-performing result of 60.6 $\pm$ 2.1% BA and 49.8 $\pm$ 2.8% F_1 score without poor-quality data from Table 5.6. Similarly, on the RHD_MP dataset, the results drop to 66.2 $\pm$ 1.6% BA and 56.3 $\pm$ 2.0% F_1 score from 74.5 $\pm$ 2.1% BA and F_1 score of 69.9 $\pm$ 2.8% from the best-performing algorithm without the inclusion of the poor-quality data (shown in Table 5.7).

Such a decrease in BA and F_1 score when including poor-quality data into the datasets is expected. Due to the few (five) poor-quality recordings in the RHD_ES dataset, these results can be seen to decrease by only approximately 2.5%, while the large number of poor-quality RHD_MP recordings (48) can be seen to have a more pronounced effect on the RHD_MP classification results, leading to a drop in performance of approximately 10%.

With these poor-quality recordings included, the automated algorithms still outperform the cardiologist, Cardioscan software and all three baseline feature extraction methods. However,

Table 5.10 – Classification results (mean $\pm$ standard deviation %) when including generator-contaminated recordings in the RHD_ES and RHD_MP datasets. The training (grey) and test (black) set results are shown. BA denotes balanced accuracy and F_1 denotes the F_1 score.

Dataset	Features	Averaging	Cycle exclusion	k -means	BA	F_1
RHD_ES	MFCCs	✓	✓	✓	69.9 $\pm$ 0.7	57.6 $\pm$ 0.9
					62.2 $\pm$ 1.6	47.4 $\pm$ 2.3
RHD_MP	MFCCs	✓	✗	✗	74.0 $\pm$ 0.3	66.8 $\pm$ 0.4
					62.7 $\pm$ 1.4	51.7 $\pm$ 2.6

as indicated by the substantial drop in performance in the RHD_MP dataset classification results, the automated assessment of the signal quality of PCG recordings, especially on a low-cost stethoscope device, is essential for the accurate screening of RHD.

5.6 Impact of external noise

The recordings contaminated with generator noise were removed from the RHD_ES and RHD_MP datasets in the previous analysis. This was done as the generator noise overlapped with the expected frequencies of heart murmurs, and this was expected to effect the classification performance of the proposed algorithms. To investigate the true effect of the generator noise on the classification performance, the generator-contaminated recordings were added to the datasets and the classification algorithms re-run. The results of the previous best-performing algorithms on each of the RHD_ES and RHD_MP datasets, with the generator-contaminated recordings included, are shown in Table 5.10.

The classification performance on the RHD_ES dataset can be seen to remain approximately the same, with the BA improving by 1.6% (from 60.6 $\pm$ 2.1% in Table 5.6 to 62.2 $\pm$ 1.6%), while F_1 score decreases by 2.4% (from 49.8 $\pm$ 2.8% to 47.4 $\pm$ 2.3%). This illustrates that the proposed algorithms are able to find discriminatory information between the two classes of data in the RHD_ES dataset, even when generator noise is included.

The RHD_MP results decreased by at least 12% (from 74.5 $\pm$ 2.1% BA and 69.9 $\pm$ 2.8% F_1 score in Table 5.7 to 62.7 $\pm$ 1.4% BA and 51.7 $\pm$ 2.6% F_1 score). While this device still performs as well as the RHD_ES dataset, it can be seen that discriminatory information between the two classes is lost when including generator noise.

The smaller change in the RHD_ES dataset is thought to be due to the more rigid construction of the electronic stethoscope, which is able to suppress external noise sources. Conversely, the low-cost construction of the mobile phone-based stethoscope is thought to make these recordings more susceptible to external noise sources, and, hence, leads to a marked decrease in classification performance. Yet, in both the RHD_ES and RHD_MP datasets, with the generator-contaminated included, the proposed automated classification algorithms still outperform the cardiologist, Cardioscan software and the three baseline feature extraction methods.

5.7 Conclusion

This chapter has presented classification algorithms for the detection of MVP and RHD in single-channel PCG recordings. The main contributions of this chapter are:

- An extensive literature review of the most recent advances in pathology classification in PCG recordings, highlighting the need for a rigorous implementation of SVM-based classification.
- The rigorous implementation of support vector machines for heart sound classification, with a train-test split and with different patients in the training and test sets, and with the use of cross-validation for the selection of the SVM parameters.
- The use of the undecimated wavelet transform and mel-frequency cepstral coefficients, extracted from clinically-relevant segments of systole, for pathology detection.
- The assessment of novel cycle averaging, cycle exclusion and k -means clustering-based cycle selection techniques on classification accuracy.
- The use of the above methods for the automatic detection of MVP in a large database of 370 recordings, with best-performing results of $78.7 \pm 0.9\%$ BA and $72.8 \pm 1.0 F_1$ score when using LASSO LR with cycle averaging and cycle exclusion based on the RMS error of the extracted features from the median cycle. This detection outperformed three classification algorithms from the literature.
- The use of the above methods for the automatic detection of functional valve damage caused by RHD in a large database of electronic stethoscope recordings. The best results

on this database were $60.6 \pm 2.1\%$ BA and $49.8 \pm 2.8\%$ F_1 score, which outperformed the cardiologist, the Zargis Cardioscan software and the three baseline comparison feature extraction methods from the literature.

- The automatic detection of functional valve damage caused by RHD in mobile phone-based PCG recordings, with performance of $74.5 \pm 2.1\%$ BA and 69.9 ± 2.8 F_1 score. The detection of RHD in these mobile-phone based recordings outperformed the cardiologist, Zargis Cardioscan software, the three baseline comparison feature extraction methods and the detection of the same valve damage in electronic stethoscope recordings.
- The evaluation of the impact of poor-quality recordings and external generator noise on the classification performance. Both factors have an impact on the classification performance when analysing the RHD_ES and RHD_MP datasets, but, as expected, a more pronounced effect on classification performance was seen when using a low-cost, mobile phone-based stethoscope.

This chapter has demonstrated the accuracy of the proposed classification algorithms on three, large, real-world, datasets. Of particular note are the results on the mobile phone-based recordings. The performance of the classification algorithms on this dataset, which outperformed every other detection scheme evaluated in this chapter, highlights the potential for such a methodology to be used to detect children and young adults with latent RHD. Such a low-cost diagnostic tool could be used by a non-medically trained healthcare worker to screen those individuals at high-risk of having RHD, and refer those with identified RHD for gold-standard echocardiographic screening. However, the importance of the signal quality assessment of the RHD_MP recordings has been demonstrated. Therefore, for such a device to work effectively, feedback from automatic signal quality assessment would have to be given to the user to ensure that only good-quality recordings are analysed further, and that poor-quality recordings are re-recorded as appropriate.

Chapter 6

Discussion

“Poor Africa. No other continent has endured such an unspeakably bizarre combination of foreign thievery and foreign goodwill.” - Barbara Kingsolver, The Poisonwood Bible

This final chapter summarises the main contributions of the work presented in this thesis and how these have contributed to the field of heart sound analysis. Directions for future work are then discussed. To this end, Section 6.2 highlights the work that could be performed to build on the methods developed in this thesis, and discusses implementation issues when rolling out a mobile phone-based diagnostic tool for RHD.

6.1 Summary of work

This thesis has investigated the design of an automated mobile phone-based diagnostic tool for RHD, the aim of which is to allow non-medically trained healthcare workers, based in a low-resource setting, to identify individuals with a high probability of having the condition. Particular attention has been paid to the application of machine learning and signal processing techniques to heart sounds recorded using a low-cost stethoscope. The three broad processing steps investigated were the assessment of signal quality; the segmentation of the fundamental heart sounds; and the detection of systolic murmurs in heart sound recordings.

Heart sound recordings are susceptible to numerous noise sources (talking, stethoscope motion, digestion and breathing) that can mask any underlying pathology. Therefore, to provide robust

automatic detection of heart pathology, it is essential to assess the signal quality of recordings. Heart sounds recorded using a low-cost stethoscope by an untrained healthcare worker are even more susceptible to noise because of the lack of sound insulation; the dependance on having the correct placement of the stethoscope (ensuring an air-tight seal between the microphone and the patient’s chest); and the user’s lack of expertise, accentuating the need for automated signal quality analysis.

The automatic classification of the signal quality of heart sound recordings was investigated in Chapter 3. Five annotators rated the signal quality of 151 recordings from two different devices, and a novel assessment of the intra- and inter-rater agreement was presented. It was found that in a third of cases, the low-cost stethoscope, operated by an untrained healthcare worker, was able to record a good-quality PCG. However, when using an electronic stethoscope, an untrained healthcare worker also records poor-quality PCGs in a third of cases, highlighting the need for signal quality assessment even when using a “high-end” electronic stethoscope.

While the automatic signal quality classification of PCG recordings has been attempted before [121, 163], previous studies were handicapped by poor methodology. This thesis builds on previous work by performing a robust evaluation of nine novel signal quality indices for automatic heart sound signal quality assessment. In the binary classification of good- and poor-quality PCGs, over 80% accuracy was achieved for both the Littmann (86.5%) and mobile phone-based (82.2%) stethoscopes when using regularised logistic regression, outperforming here the more complex SVM classifier (which was found to over-fit the data). The features based on the magnitude of the first peak in the autocorrelation function and the kurtosis of the PCG signal were found to be most important in signal-quality classification for both devices.

The next major step in the processing of heart sound recordings is *segmentation*, or the detection of the exact positions of the S_1 and S_2 sounds. Knowledge of these positions allows the subsequent analysis of the segments between these sounds for the identification of pathological murmurs. Segmentation has been extensively explored in the literature, as shown by the review table in Chapter 4. Many methods have used threshold-based techniques to identify the positions of the heart sounds, which are susceptible to noise. Generative probabilistic models, such as a hidden Markov model (HMM), which take into account the expected duration of the heart states, were shown to outperform threshold-based methods [99]. However, limitations

of this study include the limited size of the dataset, the investigation of only two features (of which one was used) and the use of a standard Viterbi algorithm, which was subsequently shown in Chapter 4 to be error-prone when used with duration-dependent HMMs.

Chapter 4 evaluated various extensions to probabilistic segmentation models by introducing more features, extending the Viterbi algorithm and introducing logistic regression-based emission probability estimation. These models were evaluated on one of the largest published real-world datasets, comprising recordings from both healthy patients and patients with heart valve damage. These additions were found to significantly improve upon the results with respect to the state-of-the-art algorithm. A *discriminative* probabilistic segmentation model, a conditional random field (CRF), with an extended Viterbi algorithm, was found to be the best-performing segmentation algorithm with an average F_1 score of $95.64 \pm 0.92\%$; a substantial improvement upon the state-of-the-art algorithm ($86.28 \pm 1.55\%$).

The final processing step, discussed in Chapter 5, was the detection of pathological systolic murmurs in heart sound recordings, with the aim of developing automated algorithms to detect valve damage due to RHD. As discussed in Section 1.3, murmurs lead to high-frequency “whooshing” sounds during systole and diastole, caused by leaking heart valves. Therefore, murmur classification aims to detect these high-frequency sounds by employing appropriate feature extraction methods along with classification algorithms.

Two datasets were analysed for the purpose of murmur classification in Chapter 5. The first comprised electronic stethoscope recordings from patients with mitral valve prolapse (MVP – which results in systolic murmurs similar to those expected in cases of RHD) and healthy patients referred for clinical examination. The second dataset consisted of heart sound recordings from children and young adults diagnosed with early-stage RHD and matched controls. These recordings were made using an electronic stethoscope and a mobile phone-based stethoscope. Each subject was also diagnosed by a paediatric cardiologist using manual auscultation and by commercial automatic auscultation software (Zargis Cardioscan), while the gold-standard diagnosis was obtained using echocardiography.

In order to accentuate the frequency content of murmurs, two time-frequency features were investigated. These were (undecimated) wavelet decomposition (which has been applied extensively for murmur classification – see Section 5.1) and the extraction of mel-frequency cepstral

coefficients (MFCCs), that have previously been applied to murmur classification in a few studies.

Individual heart cycles were located by using the CRF-based method developed in Chapter 4. As many of the analysed recordings were corrupted by noise, the effect of denoising on the classification performance was also evaluated. Heart cycles were denoised using cycle averaging, which incorporated noisy cycle exclusion based on k -means clustering and an error-based metric. The classification results using these methods were compared to three feature extraction methods from the literature.

While support vector machines (SVMs) have been applied to the problem of heart murmur classification in a number of cases, no previous study could claim to be robust in its implementation of the algorithm (see Section 5.1.3). This thesis therefore performed a robust evaluation of SVMs for murmur classification, the results from which were compared to regularised logistic regression.

The three feature extraction methods from the literature performed poorly on the datasets analysed in Chapter 5, with performance, at best, slightly better than chance detection of MVP and RHD. In contrast, the proposed methods achieved over 75% *balanced accuracy* on the MVP and RHD datasets (recorded using the mobile phone), where it was found that cycle averaging improved classification results in all cases, while cycle exclusion sometimes led to improved performance. RHD detection on the mobile phone-based stethoscope outperformed that of the electronic stethoscope recordings, while also outperforming the clinician and commercial software, indicating the potential of such a device for RHD screening.

6.2 Limitations and future work

The automated methods for signal quality assessment, segmentation and murmur classification presented in this thesis perform well on the various datasets that were analysed. However, these methods do suffer from a number of limitations, which are discussed below. Directions of future research are also presented and the thesis concludes with a discussion of how to implement a low-cost, mobile phone-based screening tool for RHD.

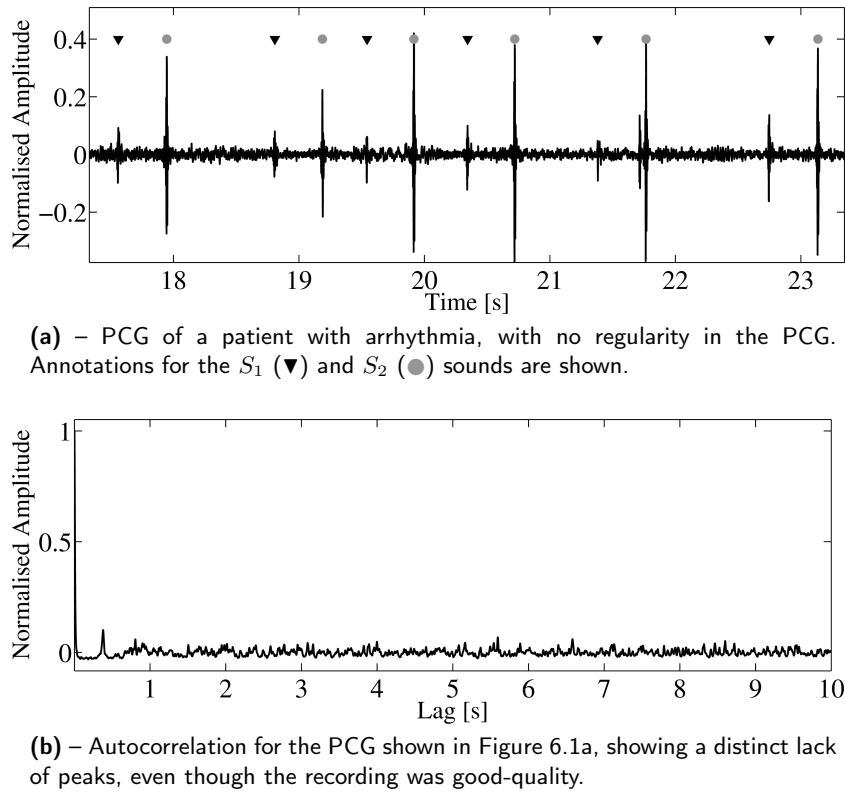


Figure 6.1 – PCG and autocorrelation of a good-quality PCG signal of an arrhythmic patient showing a lack of peaks in the autocorrelation function.

6.2.1 Signal quality classification

Many of the signal quality indices used in the classification methods presented in Chapter 3 were based on the autocorrelation of the PCG, with the intrinsic assumption that high-quality recordings lead to pronounced peaks in the resulting signal. However, this assumption is dependent on the regularity of the heart beats in the PCG recordings. While one counter-example of a recording with irregular heart rhythm and its resulting autocorrelation function was given in Figure 3.5, this recording still had some degree of periodicity, leading to clear peaks in the autocorrelation function. A good-quality PCG of an irregular heart rhythm does not have clearly-defined peaks, leading to erroneous signal quality classification. An example of such a case is shown in Figure 6.1.

There is therefore the need to evaluate the signal quality classification algorithms derived in Chapter 3 on a larger dataset of arrhythmic patients. It may be the case that the signal quality indices which were not derived from the autocorrelation (such as the kurtosis of the PCG signal, which performed well for both Littmann and iPhone devices) may still perform

well in these cases. If not, further research into robust signal quality indices would be required. A promising approach may be further investigation into the *bSQI* metric, as a similar metric has been found to work well when assessing the signal quality of arrhythmic ECG signals [162]. While a few cases of pathological murmurs were present in the datasets evaluated in Chapter 3, it is also necessary to evaluate the new algorithms on a larger dataset of cases with pathological murmurs. If these murmurs are of large amplitude, they would affect the SQIs that are not based on the autocorrelation function^a. Therefore, a robust evaluation of signal quality classification of PCG signals would require eight distinct cases: recordings with and without arrhythmia, recordings with and without pathological murmurs and then good- and poor-quality recordings of each of these cases. Furthermore, in order to make the classification agnostic to the recording device, a range of stethoscopes would need to be investigated.

6.2.2 Heart sound segmentation

The heart sound segmentation methods presented in Chapter 4 suffered from a similar shortcoming to that of the signal quality classification algorithms. While the probabilistic segmentation methods worked very well in PCGs with regular heart rhythm, their reliance on fixed duration distributions for each state made them susceptible to errors when segmenting arrhythmic recordings, or those with a large variation in heart rate. An example of such a case is shown in Figure 4.12.

A simple solution to the problem of varying heart rate could be to segment short periods of heart sound recordings, rather than the entire recording, or varying the duration distributions during the forward path of the Viterbi algorithm. However, the increase in computational complexity may not be justified when slightly wider distributions for the duration of systole and diastole may prove just as effective.

A more complicated problem is that of the segmentation of arrhythmic PCGs. The duration distributions incorporated in the segmentation models are believed to underpin their success, as these duration constraints allow the models to ignore heart sound-like noise that occurs at instances when heart sounds are not expected. However, these models then also ignore heart sounds which occur when they are not expected. A possible solution would be the integration

^aHeart sound recordings with pathological murmurs, which are regular in their rhythm, would still have large peaks in the autocorrelation function.

of signal quality estimates into the segmentation models, as employed for the segmentation of other physiological signals by Pimentel et al. [285]. These researchers employed a similar duration-dependent HMM for signal segmentation, but computed a signal quality probability for each signal. This signal quality probability was used to weight the observation probabilities in the HMM. In periods of good-quality signals, the HMM would rely more on these observations (rather than the duration distributions). Conversely, in periods of poor-signal quality, the HMM would rely more on the expected durations of each state. A similar method could be used for the segmentation of PCGs^b: in a good-quality, arrhythmic recording, the HMM could effectively ignore the duration constraints. With a poor-quality recording, the HMM could rely more on the duration constraints, which are effective in noisy recordings of regular heart rhythm. An open problem is then the segmentation of noisy arrhythmic recordings. However, in an ideal PCG analysis framework, such recordings would be detected in the signal quality assessment and repeated as appropriate.

The modelling of the duration distributions for each state could also be investigated further. In this thesis, Gaussian distributions were used to model the expected duration of each state, in order to compare the models proposed in this thesis to that of Schmidt et al. [99]. However, other distributions, or mixtures of distributions, may model these durations more effectively. One example is the case of split S_2 sounds, where a Gaussian mixture model may prove to be more accurate than a single Gaussian distribution. Furthermore, an investigation into the optimal relationship between the heart rate and systolic and diastolic durations should prove useful. For instance, researchers have modelled a linear relationship between the heart rate and time between the Q -wave in the ECG and the start of the S_2 sound [286, 287, 288]. Such modelling could be used to form more generalised duration distributions rather than the labels from a small dataset as performed by Schmidt et al. [99].

A conditional random field (CRF) was found to be the best-performing segmentation model in Chapter 4. However, this model only incorporated the feature observations from the current sample in the forward path of the Viterbi algorithm (in order to allow direct comparison to an HMM). One of the main advantages of a CRF over an HMM is its ability to incorporate features from past and future samples in the computation of the observation probabilities. With the CRF already outperforming both the standard HMM and hybrid logistic-regression

^bThis presumes the accurate assessment of signal quality even in arrhythmic PCGs

HMM in Chapter 4, it is expected that such an extension of the CRF would further improve segmentation performance.

Lastly, rather than seeing segmentation and signal quality assessment as two distinct tasks, one could perform both simultaneously. Probabilistic models such as HMMs and CRFs can give an indication of how well the model “fits” a signal by evaluating the posterior probability of the observation sequence given the model, $P(\mathbf{O}|\lambda)$ (borrowing the HMM notation). Provided that the probabilistic model provides an accurate approximation of good-quality signals, such a value would give an indication of the quality of the signal.

6.2.3 Pathology classification

A major limitation of the mobile phone-based PCG dataset analysed in Chapter 5 was the corruption of many of these recordings by generator noise. The dominant frequency components of this non-stationary noise overlapped with the expected frequencies of systolic murmurs (50-300 Hz), and therefore, linear filtering would be detrimental to pathology classification.

A straightforward solution to the issue of external noise contamination would be the simultaneous recording of both heart sounds and external noise, and the removal of the noise from the signal of interest. Indeed, modern smart phones adopt this exact procedure to remove noise from phone calls, by using a second microphone on the handset. However, at the time of data collection, it was not possible to record simultaneously from the hands-free kit microphone (used within the low-cost stethoscope) and a second microphone on the mobile phone itself, because of constraints in the mobile phone’s software application program interface (API). If such limitations could be overcome, the murmur detection capabilities of such a mobile phone-based stethoscope could be greatly enhanced.

If only a single audio channel is available, then more complex signal processing techniques could be explored. For example, speech enhancement algorithms (such as spectral subtraction [289]) could be applied to remove background noise, provided that an estimate of the noise signal alone is possible. If it is known that the pathological sounds occur in systole alone, then this estimate of the noise could be derived from the diastolic periods of a recordings.

Alternatively, a separate recording of the background noise could be performed. For instance, a “baseline noise” recording could be captured before or after the screening of a patient. This

was performed for many of the recordings from the RHD_MP dataset (see Section 2.3). This baseline noise recording was performed with the stethoscope attachment held on the abdomen of the participant, over the liver, in order to imitate the recording of sounds on the chest, but away from the heart and lungs. This was done in order to capture the background noise in the recording environment, but with minimal recording of heart and lung sounds. Spectral subtraction was then applied to the heart sound recordings, but the noise removal was found to be too severe, leading to distortion of the heart sounds and the removal of any murmur features. Further investigation into the enhancement of heart sounds using such methods is therefore needed.

6.2.4 Implementation considerations

This thesis has demonstrated the potential of a mobile phone-based screening tool for RHD. However, while the device itself might function accurately, a number of issues need further investigation before such a device could be used in practice. If we assume that: (i) the remaining issues of signal quality classification, segmentation and murmur classification mentioned above have all been dealt with; (ii) the screening tool has been evaluated using extensive field tests and gone through the process of certification; and (iii) the tool is now ready to be rolled out in the field, then three issues need to be considered before a large-scale screening programme could be undertaken.

The first is the choice of person to perform the screenings. As mentioned in Section 1.4, teachers are a natural choice as they have access to children without the need to set up a bespoke screening programme. Engel et al. [52] have identified the potential for school-based screening programmes, while the WHO also recommends school-based screening [49]. However, this raises ethical questions regarding the interaction between students and teachers (especially male teachers screening female students).

A second consideration is the mobile devices that would be used. This thesis argues that smart phones are an attractive solution, because of their processing ability and growing prevalence in developing countries. However, with multiple hand-sets available with different recording and computational capabilities, the use of personal mobile phones may prove difficult. Standard devices, provided by a central screening service, may prove more effective, but this raises

questions of funding and security of the devices and their data. Furthermore, equipping teachers in developing areas with a high-end mobile phone may put them at risk in terms of their personal safety.

Lastly, if a screened patient is detected to have RHD with a high probability, what would the next steps for that patient be? Referral to a bespoke screening service would be the ideal scenario. While the low-cost screening tool developed in this thesis may overcome some of the resource and cost barriers involved in mass screening, the set-up of a referral service, able to provide gold-standard echocardiographic diagnosis and the necessary penicillin or surgical treatment, would also be necessary to treat these RHD patients effectively.

Appendix A

Cohen's kappa

Cohen's kappa, κ , is a measure of the level of agreement between two annotators, that have each categorised N items into k mutually exclusive, nominal, categories, and corrects for the level of chance agreement between the annotators [290]. The equation for κ is given by [290]:

$$\kappa = \frac{p_o - p_c}{1 - p_c} \quad (\text{A.1})$$

p_o is the observed proportion of agreement between the annotators,

$$p_o = \frac{1}{N} \sum_{i=1}^k n_{ii} \quad (\text{A.2})$$

where n_{ij} is the number of items that were placed into the i th category by annotator 1 and the j th category by annotator 2.

p_c is the probability of chance agreement between the two annotators:

$$p_c = \frac{1}{N} \sum_{i=1}^k p_{i\bullet} \times p_{\bullet i} \quad (\text{A.3})$$

where $p_{i\bullet}$ is the proportion of items placed into category i by annotator one,

$$p_{i\bullet} = \frac{1}{N} \sum_{j=1}^k n_{ij} \quad (\text{A.4})$$

Table A.1 – Benchmarks for the levels of agreement when using Cohen’s Kappa [292].

Kappa Statistic	Strength of Agreement
< 0.00	Poor
0.00 - 0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

and $p_{\bullet i}$ is the proportion of items placed into category i by annotator two [291]. This leads to a κ value of 1 when the annotators agree completely, and a value of 0 when there is chance agreement between annotators. Less than chance agreement leads to negative values of κ .

Benchmarks for the degree of agreement when using Cohen’s kappa were proposed by Landis and Koch [292], as shown in Table A.1. Although this rating scale has been criticised because it is based on subjective judgement of the researchers, it is still widely cited [293].

As mentioned above, Cohen’s κ determines the level of agreement between two raters when classifying items according to a *nominal* scale. However, κ fails to give an adequate representation of agreement in cases for which the categories of classification are *ordinal* and can be ranked in some meaningful way, such as severity, importance or degree of signal quality. For example, in the case of the classification of disease severity, the degree of misclassification between an item placed into a “moderate” and “mild” classes by two independent annotators is less than the degree of misclassification between “severe” and “mild” classes. The standard κ calculation does not account for such degrees of misclassification, as it implicitly treats all disagreements between annotators as complete disagreement [291].

To deal with such ordinal scales, Cohen introduced the idea of a weighted kappa, κ_w [294]. This measures the level of agreement between two annotators while taking into account chance agreement, but employs weights of disagreement between classes in order to incorporate the ordinal structure of the classes. This measurement can be defined as:

$$\kappa_w = \frac{p_{ow} - p_{cw}}{1 - p_{cw}} \quad (\text{A.5})$$

where p_{ow} is now the weighted proportion of agreement between annotators,

$$p_{ow} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^k w_{ij} n_{ij} \quad (\text{A.6})$$

and p_{cw} is the weighted proportion of change agreement,

$$p_{cw} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^k w_{ij} p_{i\bullet} p_{\bullet j} \quad (\text{A.7})$$

where w_{ij} are the weights assigned to an item being categorised into classes i and j by the two annotators, respectively. Most commonly, these weights either linearly or quadratically increase as the degree of disagreement increases. For instance, in a linear weighting scheme, a disagreement by one category between two annotators would have a weight of one, a disagreement by two categories would have a weight of two, et cetera. Quadratic weights were selected in this case, as these are more lenient when it comes to partial disagreements between annotators [293], while allowing for the extension of this statistic to more than two raters, as proposed by Janson and Olsson [295].

A.1 Extending Cohen's κ to more than two annotators

Fleiss extended the general form of Cohen's κ to its use with more than two annotators, while still taking into account the agreement of these annotators by chance [296]. However, Fleiss's κ , like Cohen's original metric, is limited to nominal scales.

Cohen's κ was extended by Conger to allow for the assessment of agreement between more than three annotators while using an ordinal scale [297]. In their study, Conger et al. extended the use of Cohen's weighted κ for more than two annotators by motivating an averaging technique over all pairs of annotators. This involves finding p_{ow} and p_{cw} from Equations A.6 and A.7 for each pair of annotators, averaging these values over all the pairs, and then finding an averaged value using Equation A.5. The same method was proposed by Janson and Olsson [295], but with the explicit use of quadratic weights. For this reason, such quadratics weights were used in this study.

Appendix B

Logistic regression

Logistic regression is a probabilistic discriminative model that computes the probability that a dependent variable belongs to a discrete, categorical class [216, 298]. Therefore, although it is named *regression*, it is a classification model. Logistic regression makes use of the logistic function to map weighted predictor variables or features (which can take values from negative to positive infinity) to a Bernoulli distribution, or a probabilistic value between zero and one. The logistic function is defined as:

$$\sigma(t) = \frac{e^t}{1 + e^t} \quad (\text{B.1})$$

where t is a independent variable, predictor variable or linear combination of weighted feature values. A graph of the function is shown in Figure B.1.

If t is a linear combination of weighted features, it can be written as $t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \beta_m x_m$, where x_n is a feature value with associated weight β_n , $1 \leq n \leq m$, and β_0 is a bias weight. In this case, Equation B.1 can be rewritten as [298]:

$$\sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \beta_m x_m) = p(\mathbf{x}; \beta) = \frac{e^{\beta_0 + \beta^T \mathbf{x}}}{1 + e^{\beta_0 + \beta^T \mathbf{x}}} \quad (\text{B.2})$$

where $p(\mathbf{x}; \beta)$ is the logistic function-derived probability given the feature vector $\mathbf{x}$ and coefficient vector β . Therefore, in the case of binary classification, $p(\mathbf{x}; \beta)$ can be interpreted as the conditional probability of a *positive* class label, $p(y = 1 | \mathbf{x}; \beta)$, whereas, the posterior probability of a negative class (with a class label of zero) can be defined as $p(y = 0 | \mathbf{x}; \beta) = 1 - p(\mathbf{x}; \beta)$. As the input to the logistic function is a linear combination of weighted predictor variables,

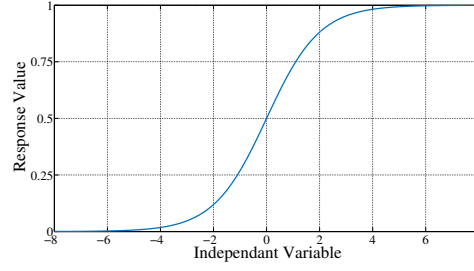


Figure B.1 – The response of the logistic function, which maps independent variables values between negative and positive infinity to values between zero and one.

logistic regression is termed a *generalised linear model*, while the logistic function in such a model is called the *link function*.

The β coefficients are computed using maximum likelihood estimation using the conditional log-likelihood of the labelled classes of a set of training data. If y_i and $\mathbf{x}_i$ are the the label (either zero or one) and feature vector of training example i , then the probability of all the labels, given the coefficients, or the likelihood function, can be computed using [216, 298]:

$$\ell(\beta) = p(\mathbf{y}|\beta) = \prod_{i=1}^N p(\mathbf{x}_i; \beta)^{y_i} (1 - p(\mathbf{x}_i; \beta))^{1-y_i} \quad (\text{B.3})$$

where N is the total number of training examples. Then, the log-likelihood of the coefficients can be defined as [180]:

$$L(\beta) = \ln(\ell(\beta)) = \sum_{i=1}^N \{y_i \ln(p(\mathbf{x}_i; \beta)) + (1 - y_i) \ln(1 - p(\mathbf{x}_i; \beta))\} \quad (\text{B.4})$$

$$= \sum_{i=1}^N \left\{ y_i (\beta_0 + \beta^T \mathbf{x}_i) - \ln(1 + e^{\beta_0 + \beta^T \mathbf{x}_i}) \right\} \quad (\text{B.5})$$

where Equation B.2 has been substituted into Equation B.4 to produce Equation B.5. The coefficients which maximise the log-likelihood are then found by finding where the first derivative is equal to zero:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^N \mathbf{x}_i (y_i - p(\mathbf{x}_i; \beta)) = 0 \quad (\text{B.6})$$

Equation B.6 can be interpreted as the error between the actual class label and the model-predicted label, $y_i - p(\mathbf{x}_i; \beta)$, multiplied by the feature vector, $\mathbf{x}_i$. There is no closed-form solution to Equation B.6. However, it is a convex function, and therefore a unique optimal

point can be found using gradient-based optimisation methods, such as iteratively reweighted least squares [216].

In order to avoid over-fitting of the model, especially in the case of small datasets or a large number of predictor variables, regularisation can be used. Regularisation refers to a penalty on the complexity of a model, thereby reducing over-fitting and improving generalisability to new data.

In the case of logistic regression, a regularisation technique called LASSO (least absolute shrinkage and selection operator) is often implemented, where the log-likelihood function includes an additional sum over the coefficients in the model [180]:

$$\max_{\beta_0, \beta} \left\{ \sum_{i=1}^N \left[y_i(\beta_0 + \beta^T \mathbf{x}_i) - \ln(1 + e^{\beta_0 + \beta^T \mathbf{x}_i}) \right] - \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (\text{B.7})$$

where p is the total number of predictor variables in the model and $\lambda \geq 0$ is a complexity parameter: the larger the value of λ , the greater the amount of penalisation on the coefficient terms. Larger values of λ may reduce some coefficients to zero, thereby performing feature selection. The complexity parameter is often set using cross validation. This function remains concave, and thus solvable using the same optimisation methods mentioned previously.

The β values that maximise Equation B.7 are called the *maximum likelihood estimates*, and are denoted by $\hat{\beta}$. In order to assess the *goodness of fit* of the logistic regression model, the *deviance* can be computed. The deviance summarises the likelihood of the fitted model, and is computed using [298]:

$$D = -2 \ln[(\text{likelihood of fitted model})] \quad (\text{B.8})$$

$$= -2 \sum_{i=1}^N \left[y_i \ln\left(\frac{p(\mathbf{x}_i; \hat{\beta})}{y_i}\right) + (1 - y_i) \ln\left(\frac{1 - p(\mathbf{x}_i; \hat{\beta})}{1 - y_i}\right) \right] \quad (\text{B.9})$$

The deviance is used to compute the optimal complexity parameter value, λ , as is shown in Figure 3.11.

Appendix C

Support vector machines

Support vector machines (SVMs), originally proposed by Cortes and Vapnik [299], are a class of supervised, non-parametric, non-probabilistic binary classifiers [180, 216, 300, 301, 302]. They are *maximum margin* classifiers, aiming to find the separating hyperplane between two classes of data that maximises the minimum distance between the classes. Such a large margin, intuitively, and due to statistical learning theory, improves the generalisability of this model to unseen data [180, 301].

Let us initially consider a linear SVM used to classify N objects in the dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where y_i is one of two possible labels, $y_i \in \{-1, 1\}$, and $\mathbf{x}_i$ refers to a feature vector, of dimension M where $\mathbf{x}_i[j]$ refers to feature j of $\mathbf{x}_i$. Then, any *discriminant function* which is used to separate two sets of data can take the form

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{C.1}$$

where the points $\mathbf{x}$ which satisfy $\mathbf{w} \cdot \mathbf{x} + b = 0$ lie on the separating hyperplane. Here, $\mathbf{w}$ refers to a weight vector which is normal to the hyperplane, and the scalar b is called the bias, where $b/\|\mathbf{w}\|$ gives the perpendicular distance of the hyperplane from the origin [300]. Therefore, the discriminant function divides the original space in two, with the *sign* of $f(\mathbf{x})$ dictating which side of the hyperplane a point sits. This *decision boundary* is linear, as it is defined by a linear combination of the input features. An example of such a hyperplane can be seen in Figure C.1.

Given the data points in Figure C.1, a number of different separating hyperplanes could be

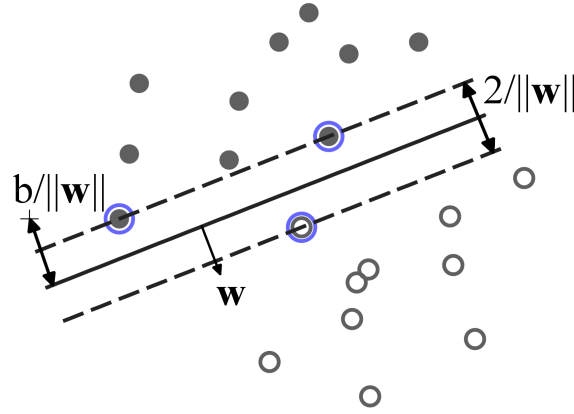


Figure C.1 – An example of a large margin separating hyperplane between two sets of linearly separable data in 2D. The solid (●) and hollow (○) discs represent the two data classes. The solid line represents the separating hyperplane, while the dotted lines represent the boundaries of the margin between the classes. The support vectors are circled in blue. The black cross (+) shows the origin. The arrows represent the perpendicular distance from the hyperplane to the origin ($b/\|\mathbf{w}\|$), the weight vector $\mathbf{w}$ which is normal to the hyperplane and the margin width ($2/\|\mathbf{w}\|$).

drawn to perfectly separate the two classes. However, as mentioned above, an SVM is a large margin classifier. Therefore, the *optimal* hyperplane is the one which maximises the margin between the two sets of data. It can be shown that the minimum perpendicular distance from the separating hyperplane to element of either class is $1/\|\mathbf{w}\|$, with the distance between classes, or margin width, being $2/\|\mathbf{w}\|$.

Therefore, in a perfectly separable case, the $\mathbf{w}$ and b parameters are found by minimising $\|\mathbf{w}\|$ (which corresponds to maximising the margin) while ensuring the every example lies on the correct side of the separating hyperplane. This optimisation can be written algebraically as:

$$\arg \min_{\mathbf{w}, b} \frac{\|\mathbf{w}\|^2}{2} \quad (\text{C.2})$$

$$\text{subject to : } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \text{ for } i = 1 \dots N \quad (\text{C.3})$$

where Equation C.3 ensures that each example is correctly classified while the margin is maximised. This can be rewritten as

$$L_P = \frac{\|\mathbf{w}\|^2}{2} - \sum_{i=1}^N \alpha_i y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + \sum_{i=1}^N \alpha_i \quad (\text{C.4})$$

after introducing the Lagrange multipliers, α_i . Now, the minimisation of the Lagrangian, L_P , with respect to $\mathbf{w}$ and b is a convex optimisation problem. If we set the derivatives of

Equation C.4 to zero, one can prove that

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \quad (\text{C.5})$$

$$\sum_{i=1}^N \alpha_i y_i = 0, \alpha_i \geq 0 \quad (\text{C.6})$$

which, if substituted back into Equation C.4, can then be used to write the *primal* optimisation problem in its *dual* formulation:

$$L_D = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (\text{C.7})$$

The maximum margin hyperplane is then found by maximising L_D , subject to the constraints in Equations C.5 and C.6, and $\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0$. Therefore, if $\alpha_i > 0$, then $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$, and $\mathbf{x}_i$ lies on the boundary of the margin. If $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1$, then $\mathbf{x}_i$ is not on the boundary of the margin and $\alpha_i = 0$.

The training points for which $\alpha_i > 0$, which lie on the boundaries of the margin, are then called the *support vectors*, as shown in Figure C.1. Therefore, only the support vectors are required to formulate the separating hyperplane. All other training examples could be discarded, and the hyperplane would take exactly the same form. For that reason, an SVM is called a *sparse* model.

After training an SVM, classification of a new example, $\mathbf{x}_j$, is performed by finding $y_j = \text{sign}(\mathbf{w} \cdot \mathbf{x}_j + b)$.

In the case of non-linearly separable data, the above algorithm will find no solution, as no straight line could be drawn through the data that would still satisfy all the conditions. In such cases, and in order to further increase the margin between classes, it is beneficial to allow the SVM to allow some examples to lie within the margin, or to be misclassified. This is achieved by introducing *slack variables*, ξ_i , into the model. If we reformulate Equations C.2 and C.3 as

$$\arg \min_{\mathbf{w}, b, \xi} \frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^N \xi_i \quad (\text{C.8})$$

$$\text{subject to :} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (\text{C.9})$$

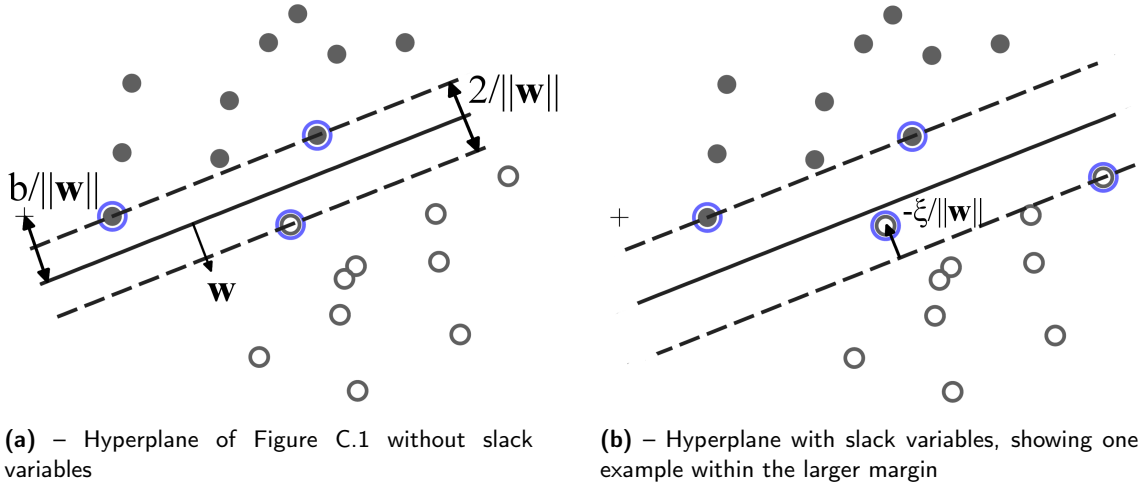


Figure C.2 – Comparison of linear hyperplanes with and without slack variables

then examples are allowed to be within the margin between classes or misclassified. The value of ξ_i is the proportional amount (in proportion to the margin width) by which the prediction, is on the wrong side of its margin, defined by $\xi_i = 0$ for data points on the correct side of their margin boundary, and $\xi_i = |y_i - f(\mathbf{x}_i)|$ for others [216]. The term $C \sum_{i=1}^N \xi_i$ penalises the use of slack variables, where $C > 0$ is a user-selected parameter with a high value corresponding to a higher penalty on errors. When using slack variables, the dual optimisation problem of Equation C.7 remains unchanged, while it is subject to $\sum_{i=1}^N \alpha_i y_i$, $0 \leq \alpha_i \leq C = 0$. A comparison between the original separating hyperplane in Figure C.1 and the same dataset, but with the use of slack variables, is shown in Figure C.2.

A key outcome of using Lagrange multipliers to formulate the optimisation problem (Equation C.7) is that the training data only appears as a dot product between vectors. It is this property that allows the extension of the linear decision boundaries to non-linear cases [300]. It is intuitive that non-linear decision boundaries may outperform linear decision boundaries in some cases^a. The process of deriving a non-linear decision boundary involves the mapping of the original data to a (possibly infinite dimensional) *feature space*, using a function $\phi(\mathbf{x})$. Then, the discriminant function becomes

$$f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b \quad (\text{C.10})$$

^aFor example, imagine one class of data in 2D completely surrounded by a second class. In this case, a linear decision boundary would be ineffectual at discriminating between the two classes.

which, given that $\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$, can be written as

$$f(\mathbf{x}) = \sum_{i=1}^N \alpha_i y_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b \quad (\text{C.11})$$

which remains linear in the feature space, but may be non-linear in the original input space if $\phi(\mathbf{x})$ is a non-linear function. If we define a *kernel function* as

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}') \quad (\text{C.12})$$

then we would only need to use $k(\mathbf{x}, \mathbf{x}')$ in the SVM algorithm, without ever explicitly knowing what $\phi(\mathbf{x})$ is. This is called the *kernel trick*. When computing the class of an unseen test example, $\mathbf{x}_j$, again, we can use $k(\mathbf{x}, \mathbf{x}')$ in the discriminant function:

$$y_j = \text{sign}\left(\sum_{i=1}^{N_s} \alpha_i y_i k(\mathbf{s}_i, \mathbf{x}_j) + b\right) \quad (\text{C.13})$$

where N_s is the number of support vectors, $\mathbf{s}_i$. One example of a widely used kernel, and the classifier used in this work, is the *Gaussian, squared exponential* or *radial basis function* (RBF) kernel, defined by

$$k_{\gamma}^{\text{Gaussian}} = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2), \quad \gamma = \frac{1}{2\sigma^2} \quad (\text{C.14})$$

where γ is called the *bandwidth*, and is a parameter which controls the width of the Gaussian, by controlling the size of the standard deviation term, $1/2\sigma^2$, and therefore plays a role in controlling the flexibility of the resulting classifier. This kernel has high values within a region around $\mathbf{x}'$, which decrease with increased distance from $\mathbf{x}'$. Therefore, the discriminant function can be interpreted as being a sum of Gaussian “bumps” around each support vector [301]. Examples of the non-linear decision boundaries from such a model, with varying parameter values, are shown in Figure C.3.

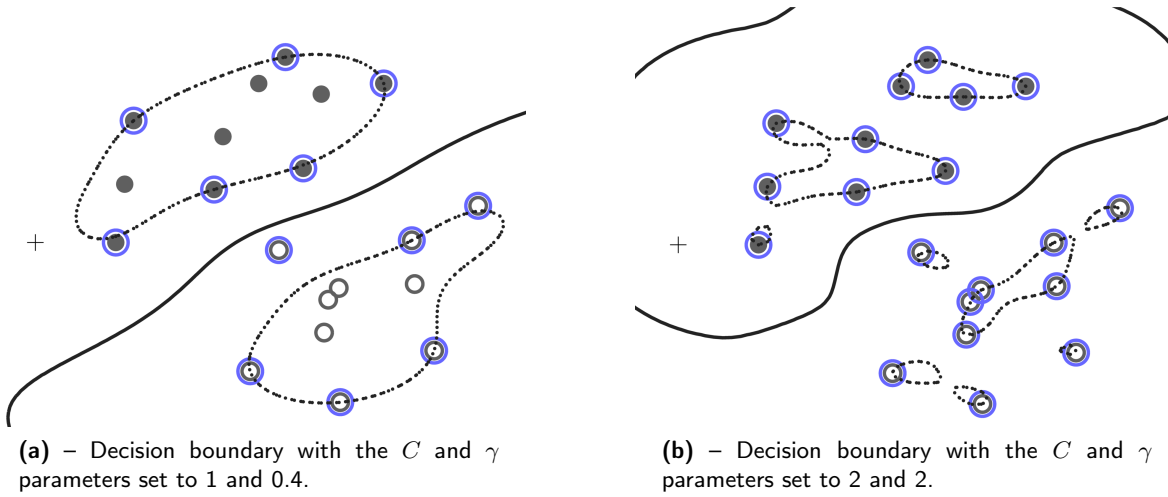


Figure C.3 – Non-linear decision boundaries generated using an SVM with a Gaussian kernel, showing the effect of varying the C and γ parameters. The support vectors are shown with additional blue circles. The decision boundary, where the discriminant function is equal to zero, is shown as a solid black line. The regions where the decision values were 1 and -1 are shown with dotted lines. The stricter classifier, resulting from greater values of C and γ can be seen in Figure C.3b.

Appendix D

Wavelet decomposition

Wavelet features lend themselves very well to PCG analysis, due to their effective time-frequency decomposition of signals. In contrast to the short-time Fourier transform (STFT), the wavelet transform makes use of varying time-frequency resolution windows, thereby allowing better time resolution at high frequencies and better frequency resolution at lower frequencies. In addition, the wavelet transform does not suffer from the cross-terms which hamper the use of the the Wigner-Ville transform (WVT) (while compromising on time-frequency resolution), while providing greater time-frequency resolution than the Choi-Williams or psuedo-WVT [303].

Similarly to the STFT, the wavelet transform employs a windowing function to analyse short segments of a signal. However, in contrast to the STFT, the windowing function can be *scaled* as well as *translated*, leading to variable time-frequency resolution. This window function is called a *wavelet*. A wavelet, Ψ , is a function that [304]:

- $\Psi \in \mathbf{L}^2(\mathbb{R})$, where $\mathbf{L}^2$ represents the space of “square integrable” or finite energy functions
- has zero average ($\int_{-\infty}^{+\infty} \Psi(t)dt = 0$)
- is normalised ($\|\Psi\| = 1$)
- satisfies an *admissibility condition*, which ensures that a linear combination of a set of scaled and dilated wavelets can represent any function in $\mathbf{L}^2(\mathbb{R})$.

These conditions ensure the wave-like form of the wavelet, while ensuring compact support and localisation in the time domain. The *continuous*^a wavelet transform (CWT) is then given by:

$$C(a, \tau) = \int \frac{1}{\sqrt{a}} \Psi^*\left(\frac{t - \tau}{a}\right) x(t) dt \quad (\text{D.1})$$

where Ψ^* is the complex conjugate of the *mother wavelet*, Ψ , scaled by a factor a and translated by τ , before computing its correlation with the signal $x(t)$. The $1/\sqrt{a}$ term ensures the wavelet remains normalised. The name “continuous” is derived from the use of *continuous* values for a and τ ($a, \tau \in \mathbb{R}, a > 0$).

Therefore, the values of $C(a, \tau)$, termed the CWT coefficients, then give a measure of similarity between the scaled wavelet and the signal at time τ [305]. As the wavelet has finite support, such a measure is localised in time, while the pseudo-frequency of the scaled wavelet can be computed using the centre frequency of the mother wavelet and the scaling factor. Under smaller scales, the wavelet offers high temporal resolution (with poorer spectral resolution), while offering higher spectral resolution at larger scales (with the draw back of poorer temporal resolution). This trade-off between the temporal and spectral resolution is defined by the *uncertainty principal*, which governs the lower bound on the joint time-frequency resolution [304, 305].

The CWT produces large, redundant, features, as it is calculated by continuously shifting a continuously scalable function over a signal. In order to reduce the feature size of the CWT, the *discrete* wavelet transform (DWT) computes the wavelet transform using dyadic scales and time-shifts with discretely sampled wavelets to produce, under certain constraints on the choice of wavelet, an orthogonal wavelet set, which allows the perfect reconstruction of the input signal [305]. The scale parameter is restricted to $a = 2^k, k \in \mathbb{Z}$, while the time-shifts are restricted to integer multiples of the scale, $\tau = 2^k l, l \in \mathbb{Z}$. Therefore, the DWT is then given by:

$$D(a, \tau) = \int \frac{1}{\sqrt{2^k}} \Psi^*\left(\frac{t - 2^k l}{2^k}\right) x(t) dt \quad (\text{D.2})$$

which can be viewed as a subsampled version of the CWT. The DWT is computed efficiently using a series of quadrature mirror filters, or a *filter bank*. These filters split the signal into a

^a“Continuous” in this sense does not refer to a continuous-time signal, but rather to a particular form of the wavelet transform. The same applies to the “discrete” wavelet transform, to be defined later in this section. Both the continuous and discrete wavelet transforms are defined for both continuous- and discrete-time signals.

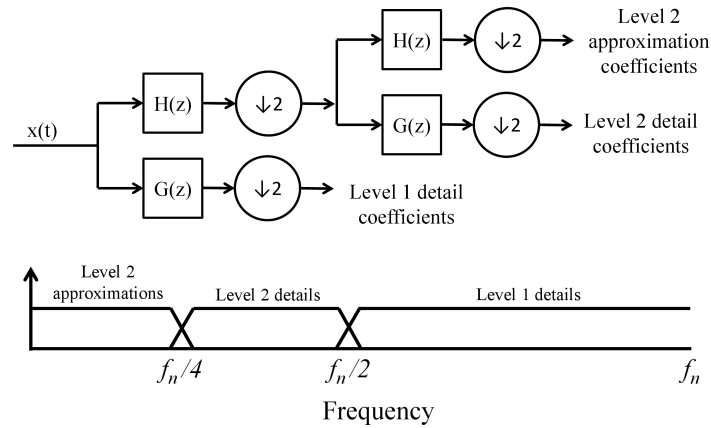


Figure D.1 – A two-level wavelet filter bank, employed when calculating the discrete wavelet transform. The high- and low-pass filters can be seen ($G(z)$ and $H(z)$), with the signal being subsampled by two after the filtering operation. The detail and approximation coefficients at each level are shown. The associated division of the frequency domain by the high- and low-pass filters is shown below, with f_n showing the Nyquist frequency, or half the sampling frequency.

low-pass channel, using a *scaling* filter, $H(z)$, and a high-pass channel using a *wavelet* filter, $G(z)$. The outputs of these filters are then subsampled by a factor of two by only keeping their even elements (as half of their frequency content has been removed).

The low-pass channel can be recursively split, using the same two filters, leading to *decomposition levels*, with *approximation* and *detail* coefficients produced from the low- and high-pass filters respectively, which represent the correlation between the signal and wavelet. The subsampling of the signal after filtering at each level effectively leads to an expansion of each subsequent filter with respect to the original signal, a doubling of the wavelet scale, and hence a halving of the centre frequency and bandwidth of the filter [305]. This process is illustrated in Figure D.1.

The *undecimated* wavelet transform (UWT)^b can be used to avoid the subsampling of the original signal at each decomposition level introduced by the DWT [307]. This leads to a signal with an equal number of samples as the original signal at each decomposition level. This is achieved by removing the downsampling operations in the DWT, and rather upsampling the filter coefficients in each low- and high-pass filter [304]. This is achieved by inserting $2^{j-1} - 1$ zeros between the original non-zero filter coefficients, where j is the decomposition level, in a

^bThe UWT is also known as the *algorithme à trous*, stationary wavelet transform, redundant discrete wavelet transform, overcomplete discrete wavelet transform, shift invariant discrete wavelet transform and discrete wavelet frames [306].

procedure originally referred to as the *algorithme à trous* [308]^c.

The equation for the UWT takes the form:

$$U(a, \tau) = \int \frac{1}{\sqrt{2^k}} \Psi^*\left(\frac{t - \tau}{2^k}\right) x(t) dt \quad (\text{D.3})$$

where k remains an integer, $a = 2^k$, $k \in \mathbb{Z}$, while τ takes the form from the CWT, allowing any real number, $\tau \in \mathbb{R}$. Therefore, $U(a, \tau)$ are the coefficients at scale $a = 2^k$ and time shift τ . This results in a translation invariant transform, such that a circular shift of the original signal will result in an identical shift of the wavelet coefficients. Therefore, unlike the DWT, this transform results in a redundant decomposition of the original signal, with the wavelet set not forming an orthogonal basis.

^cTrous means “holes” in French, with the insertion of zeros or holes into the filters giving rise to the name of the algorithm.

Bibliography

- [1] A. van Heerden, M. Tomlinson, and L. Swartz, "Point of care in your pocket: a research agenda for the field of m-health." *Bulletin of the World Health Organization*, vol. 90, no. 5, pp. 393–4, may 2012.
- [2] Vital Wave Consulting, "mHealth for development: the opportunity of mobile technology for healthcare in the developing world. United Nations Foundation and Vodafone Foundation Technology Partnership Report," Washington, D.C. and Berkshire, UK, Tech. Rep., 2009.
- [3] World Health Organization, "mHealth: New horizons for health through mobile technologies," Tech. Rep., 2011.
- [4] W. Nilsen, S. Kumar, A. Shar, C. Varoquiers, T. Wiley, W. T. Riley, M. Pavel, and A. Atienza, "Advancing the Science of mHealth," *Journal of Health Communication*, vol. 17, no. sup1, pp. 5–10, may 2012.
- [5] M. Tomlinson, M. J. Rotheram-Borus, L. Swartz, and A. C. Tsai, "Scaling Up mHealth: Where Is the Evidence?" *PLoS Medicine*, vol. 10, no. 2, pp. 1–5, feb 2013.
- [6] J. Kahn, J. Yang, and J. Kahn, "'Mobile' Health Needs And Opportunities In Developing Countries," *Health Affairs*, vol. 2, no. 2, pp. 252–258, feb 2010.
- [7] R. Richards-Kortum and M. Oden, "Devices for low-resource health care," *Science*, vol. 342, no. 6162, pp. 1055–1057, nov 2013.
- [8] L. Pritchett and L. H. Summers, "Wealthier is Healthier," *The Journal of Human Resources*, vol. 31, no. 4, pp. 841–868, jan 1996.
- [9] World Health Organisation, "The World Health Report: Health Systems Financing: The path to Universal Coverage," Tech. Rep., 2010.
- [10] R. A. Malkin, "Design of health care technologies for the developing world." *Annual Review of Biomedical Engineering*, vol. 9, pp. 567–87, jan 2007.
- [11] J. P. Smith, "Healthy bodies and thick wallets: the dual relation between health and economic status." *The Journal of Economic Perspectives : a Journal of the American Economic Association*, vol. 13, no. 2, pp. 144–66, jan 1999.
- [12] A. Boutayeb, "The double burden of communicable and non-communicable diseases in developing countries," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 100, no. 3, pp. 191–199, jul 2006.
- [13] N. Banatvala and L. Donaldson, "Chronic diseases in developing countries," *The Lancet*, vol. 370, no. 9605, pp. 2076–8, dec 2007.
- [14] A. Beratarrechea, A. G. Lee, J. M. Willner, E. Jahangir, A. Ciapponi, and A. Rubinstein, "The impact of mobile health interventions on chronic disease outcomes in developing countries: a systematic review." *Telemedicine Journal and E-health: the Official Journal of the American Telemedicine Association*, vol. 20, no. 1, pp. 75–82, jan 2014.
- [15] S. C. E. Anyangwe and C. Mtonga, "Inequities in the global health workforce: the greatest impediment to health in sub-Saharan Africa." *International Journal of Environmental Research and Public Health*, vol. 4, no. 2, pp. 93–100, jun 2007.

- [16] Y. Kinfu, "The health worker shortage in Africa: are enough physicians and nurses being trained?" *Bulletin of the World Health Organization*, vol. 87, no. 3, pp. 225–230, mar 2009.
- [17] R. Scheffler, J. Liu, Y. Kinfu, and M. Dal Poz, "Forecasting the global shortage of physicians: an economic-and needs-based approach," *Bulletin of the World Health Organization*, vol. 86, no. 7, pp. 516–523, jul 2008.
- [18] "Working together for health," The World Health Organization, Geneva, Switzerland, Tech. Rep., 2006.
- [19] *Pocket World in Figures 2013*. The Economist, Profile Books Limited, 2013.
- [20] World Health Organisation, "Global Health Workforce Statistics," Geneva, Switzerland, Tech. Rep., 2013.
- [21] C. P. Waegemann, "mHealth: the next generation of telemedicine?" *Telemedicine Journal and E-health: the Official Journal of the American Telemedicine Association*, vol. 16, no. 1, pp. 23–5, jan 2010.
- [22] The World Bank, *Information and Communications for Development 2012: Maximising Mobile*, Washington, D.C., 2012.
- [23] The International Telecommunication Union, "Measuring the Information Society," Geneva, Switzerland, Tech. Rep., 2012. [Online]. Available: http://www.itu.int/ITU-D/ict/publications/idi/material/2012/MIS2012{_}without{_}Annex{_}4.pdf
- [24] C. S. Hall, E. Fottrell, S. Wilkinson, and P. Byass, "Assessing the impact of mHealth interventions in low- and middle-income countries - what has been shown to work?" *Global Health Action*, vol. 7, no. 25606, pp. 1–12, oct 2014.
- [25] P. Mechael, "Towards the development of an mHealth strategy. WHO mHealth Review," The World Health Organization, Tech. Rep. August, 2007.
- [26] C. Free, G. Phillips, L. Watson, L. Galli, L. Felix, P. Edwards, V. Patel, and A. Haines, "The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis." *PLoS Medicine*, vol. 10, no. 1, pp. 1–26, jan 2013.
- [27] Informa Telecoms and Media, "Africa Telecoms Outlook," Tech. Rep., 2014.
- [28] F. Collins, "How to Fulfill the True Promise of mHealth," *Scientific American*, vol. 307, no. 1, pp. 16–16, 2012.
- [29] D. Estrin and I. Sim, "Open mHealth Architecture: An Engine for Health Care Innovation," *Science*, vol. 330, no. 6005, pp. 759–760, nov 2010.
- [30] A. Chib, M. H. van Velthoven, and J. Car, "mHealth Adoption in Low-Resource Environments: A Review of the Use of Mobile Healthcare in Developing Countries." *Journal of Health Communication*, vol. 20, no. 1, pp. 4–34, mar 2014.
- [31] J. Car, I. Gurol-Urganci, T. de Jongh, V. Vodopivec-Jamsek, and R. Atun, "Mobile phone messaging reminders for attendance at healthcare appointments." *Cochrane database of systematic reviews*, vol. 7, jan 2012.
- [32] R. Guy, J. Hocking, H. Wand, S. Stott, H. Ali, and J. Kaldor, "How effective are short message service reminders at increasing clinic attendance? A meta-analysis and systematic review." *Health Services Research*, vol. 47, no. 2, pp. 614–32, apr 2012.
- [33] C. Free, G. Phillips, L. Galli, L. Watson, L. Felix, P. Edwards, V. Patel, and A. Haines, "The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review." *PLoS Medicine*, vol. 10, no. 1, pp. 1–45, jan 2013.

- [34] R. T. Lester, P. Ritvo, E. J. Mills, A. Kariri, S. Karanja, M. H. Chung, W. Jack, J. Habyarimana, M. Sadatsafavi, M. Najafzadeh, C. Marra, B. Estambale, E. Ngugi, T. B. Ball, L. Thabane, L. J. Gelmon, J. Kimani, M. Ackers, and F. Plummer, "Effects of a mobile phone short message service on antiretroviral treatment adherence in Kenya (WelTel Kenya1): a randomised trial." *The Lancet*, vol. 376, no. 9755, pp. 1838–45, nov 2010.
- [35] C. Pop-Eleches, H. Thirumurthy, J. P. Habyarimana, J. G. Zivin, M. P. Goldstein, D. de Walque, L. McKeen, J. Haberer, S. Kimaiyo, J. Sidle, D. Ngare, and D. R. Bangsberg, "Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: a randomized controlled trial of text message reminders," *Aids*, vol. 25, no. 6, pp. 825–834, mar 2013.
- [36] K. Bobrow, T. Brennan, D. Springer, N. S. Levitt, B. Rayner, M. Namane, L.-M. Yu, L. Tarassenko, and A. Farmer, "Efficacy of a text messaging (SMS) based intervention for adults with hypertension: protocol for the StAR (SMS Text-message Adherence support trial) randomised controlled trial." *BMC Public Health*, vol. 14, no. 28, pp. 1–9, jan 2014.
- [37] A. Labrique, L. Vasudevan, L. W. Chang, and G. Mehl, "H₂O for mHealth: More "y" or "o" on the horizon?" *International Journal of Medical Informatics*, vol. 82, no. 5, pp. 467–469, may 2013.
- [38] T. Chen, K. Kuan, L. Celi, and G. Clifford, "Intelligent HeartSound Diagnostics on a Cellphone using a Hands-free Kit," in *AAAI Spring Symposium on Artificial Intelligence for Development*, Stanford University, 2010, pp. 26–31.
- [39] K. L. Kuan, "A framework for automated heart and lung sound analysis using a mobile telemedicine platform," Masters Thesis, Massachusetts Institute of Technology, 2010.
- [40] P. Bentley, "iStethoscope Pro." [Online]. Available: <http://www.peterjbentley.com/istethoscopepro.html>
- [41] G. Nordehn and S. Stunic, "iAuscultate," 2009. [Online]. Available: <http://iauscultate.com/iAuscultate.html>
- [42] Thinklabs Medical, "Thinklabs Stethoscope App," 2015. [Online]. Available: <http://www.thinklabs.com/{#}!apps-software/c11s0>
- [43] StratoScientific, "StethIO," 2015. [Online]. Available: <http://www.stethio.com/>
- [44] CliniCloud Inc., "CliniCloud," 2015. [Online]. Available: <https://clinicloud.com/>
- [45] Eko Devices, "Eko Core," 2015. [Online]. Available: <https://ekodevices.com/>
- [46] E. Marijon, M. Mirabel, D. S. Celermajer, and X. Jouven, "Rheumatic heart disease." *The Lancet*, vol. 379, no. 9819, pp. 953–64, mar 2012.
- [47] J. R. Carapetis, M. McDonald, and N. J. Wilson, "Acute rheumatic fever," *The Lancet*, vol. 366, no. 9480, pp. 155–68, jul 2005.
- [48] M. Mirabel, K. Narayanan, X. Jouven, and E. Marijon, "Prevention of Acute Rheumatic Fever and Rheumatic Heart Disease," *Circulation*, vol. 130, no. 5, pp. e35–e37, jul 2014.
- [49] K. Roberts, S. Colquhoun, A. Steer, B. Reményi, and J. Carapetis, "Screening for rheumatic heart disease: current approaches and controversies." *Nature Reviews. Cardiology*, vol. 10, no. 1, pp. 1–10, nov 2012.
- [50] J. R. Carapetis, A. C. Steer, E. K. Mulholland, and M. Weber, "The global burden of group A streptococcal diseases." *The Lancet Infectious Diseases*, vol. 5, no. 11, pp. 685–94, nov 2005.
- [51] M. J. McLaren, M. Markowitz, and M. Gerber, "Rheumatic heart disease in developing countries: the consequence of inadequate prevention." *Annals of Internal Medicine*, vol. 120, no. 3, pp. 243–5, mar 1994.

- [52] M. E. Engel, A. Haileamlak, L. J. Zühlke, C. E. Lemmer, S. Nkepu, M. van de Wall, W. Daniel, M. Shung King, and B. M. Mayosi, "Prevalence of rheumatic heart disease in 4720 asymptomatic scholars from South Africa and Ethiopia," *Heart*, vol. 101, pp. 1389–1394, jun 2015.
- [53] A. S. Dajani, E. Ayoub, F. Z. Bierman, A. L. Bisno, F. W. Denny, D. T. Durack, P. Ferrieri, M. Freed, M. Gerber, E. L. Kaplan, A. W. Karchmer, M. Markowitz, S. H. Rahimtoola, S. T. Shulman, G. Stollerman, M. Takahashi, A. Taranta, K. A. Taubert, and W. Wilson, "Guidelines for the diagnosis of rheumatic fever: Jones criteria, updated 1992," *Journal of the American Medical Association*, vol. 268, no. 15, pp. 2069–2073, oct 1992.
- [54] M. K. Park, *Park's pediatric cardiology for practitioners*, 6th ed. Philadelphia, PA, USA: Elsevier - Health Sciences Division, 2014.
- [55] B. Longo-Mbenza, M. Bayekula, R. Ngiyulu, V. E. Kintoki, N. F. Bikangi, K. V. Seghers, L. E. Lukoki, M. F. Mandundu, M. Manzanza, and Y. Nlandu, "Survey of rheumatic heart disease in school children of Kinshasa town." *International Journal of Cardiology*, vol. 63, no. 3, pp. 287–94, feb 1998.
- [56] G. H. Stollerman, "Rheumatic fever in the 21st century." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, vol. 33, no. 6, pp. 806–14, sep 2001.
- [57] J. R. Carapetis, "Rheumatic heart disease in developing countries." *The New England Journal of Medicine*, vol. 357, no. 5, pp. 439–441, aug 2007.
- [58] M. Rothenbühler, C. J. O'Sullivan, S. Stortecky, G. G. Stefanini, E. Spitzer, J. Estill, N. R. Shrestha, O. Keiser, P. Jüni, and T. Pilgrim, "Active surveillance for rheumatic heart disease in endemic regions: a systematic review and meta-analysis of prevalence among children and adolescents," *The Lancet Global Health*, vol. 2, no. 12, pp. e717–e726, dec 2014.
- [59] L. Zühlke, M. Mirabel, and E. Marijon, "Congenital heart disease and rheumatic heart disease in Africa: recent advances and current priorities." *Heart (British Cardiac Society)*, vol. 99, no. 21, pp. 1554–61, nov 2013.
- [60] L. Zühlke, M. E. Engel, D. Watkins, and B. M. Mayosi, "Incidence, prevalence and outcomes of rheumatic heart disease in South Africa: a systematic review of contemporary studies." *To appear in the International Journal of Cardiology*, 2015.
- [61] E. Marijon, P. Ou, D. S. Celermajer, B. Ferreira, A. O. Mocumbi, D. Jani, C. Paquet, S. Jacob, D. Sidi, and X. Jouven, "Prevalence of rheumatic heart disease detected by echocardiographic screening." *The New England Journal of Medicine*, vol. 357, no. 5, pp. 470–6, aug 2007.
- [62] C. J. L. Murray, T. Vos, R. Lozano, *et al.*, "Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010," *The Lancet*, vol. 380, pp. 2197–2223, dec 2012.
- [63] L. J. Zühlke, M. E. Engel, G. Karthikeyan, S. Rangarajan, P. Mackie, B. Cupido, K. Mauff, S. Islam, A. Joachim, R. Daniels, V. Francis, S. Ogendo, B. Gitura, C. Mondo, E. Okello, P. Lwabi, M. M. Al-Kebsi, C. Hugo-Hamman, S. S. Sheta, A. Haileamlak, W. Daniel, D. Y. Goshu, S. G. Abdissa, a. G. Desta, B. A. Shasho, D. M. Begna, A. ElSayed, A. S. Ibrahim, J. Musuku, F. Bode-Thomas, B. N. Okeahialam, O. Ige, C. Sutton, R. Misra, A. Abul Fadl, N. Kennedy, A. Damasceno, M. Sani, O. S. Ogah, T. Olunuga, H. H. M. Elhassan, A. O. Mocumbi, A. M. Adeoye, P. Mntla, D. Ojji, J. Mucumbitsi, K. Teo, S. Yusuf, and B. M. Mayosi, "Characteristics, complications, and gaps in evidence-based interventions in rheumatic heart disease: the Global Rheumatic Heart Disease Registry (the REMEDY study)," *European Heart Journal*, vol. 36, no. 18, pp. 1115–1122, nov 2014.
- [64] B. Reményi, N. Wilson, A. Steer, B. Ferreira, J. Kado, K. Kumar, J. Lawrenson, G. Maguire, E. Marijon, M. Mirabel, A. O. Mocumbi, C. Mota, J. Paar, A. Saxena, J. Scheel, J. Stirling, S. Viali, V. I. Balekundri, G. Wheaton, L. Zühlke, and J. Carapetis,

- “World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease - an evidence-based guideline,” *Nature Reviews Cardiology*, vol. 9, no. 5, pp. 297–309, may 2012.
- [65] J. R. Carapetis, “Pediatric rheumatic heart disease in the developing world: echocardiographic versus clinical screening,” *Nature clinical practice. Cardiovascular medicine*, vol. 5, no. 2, pp. 74–5, feb 2008.
- [66] M. de Melo Barbosa, M. do Carmo Pereira Nunes, and R. Muller, “Rheumatic heart disease: A neglected heart disease,” in *Prevention of Cardiovascular Diseases: From current evidence to clinical practice*, 1st ed., J. P. Andrade, D. K. Arnett, and F. J. Pinto, Eds. Atheneu and Springer, 2015, ch. 15, pp. 143–157.
- [67] E. Marijon, P. Ou, D. S. Celermajer, B. Ferreira, A. O. Mocumbi, D. Sidi, and X. Jouven, “Echocardiographic screening for rheumatic heart disease,” *Bulletin of the World Health Organization*, vol. 86, no. 2, feb 2008.
- [68] L. Zühlke and B. M. Mayosi, “Echocardiographic screening for subclinical rheumatic heart disease remains a research tool pending studies of impact on prognosis,” *Current Cardiology Reports*, vol. 15, no. 3, p. 343, mar 2013.
- [69] M. Mirabel, D. S. Celermajer, B. Ferreira, M. Tafflet, M.-C. Perier, N. Karam, A.-O. Mocumbi, D. N. Jani, D. Sidi, X. Jouven, and E. Marijon, “Screening for rheumatic heart disease: evaluation of a simplified echocardiography-based approach,” *European Heart Journal: Cardiovascular Imaging*, vol. 13, no. 12, pp. 1024–9, dec 2012.
- [70] Williams Medical, “GE Healthcare VScan Handheld Ultrasound Imaging System,” 2015. [Online]. Available: <http://www.wms.co.uk/Ultrasound/Ultrasound{ }Equipment{ }and{ }Accessories/GE{ }Healthcare{ }VScan{ }Handheld{ }Ultrasound{ }Imaging{ }System>
- [71] A. O. H. Mocumbi, “The challenges of cardiac surgery for African children,” *Cardiovascular Journal of Africa*, vol. 23, no. 3, pp. 165–7, apr 2012.
- [72] M. Mirabel, A. Grimaldi, J. Freers, X. Jouven, and E. Marijon, “Access to cardiac surgery in sub-Saharan Africa,” *The Lancet*, vol. 385, no. 9968, p. 606, feb 2015.
- [73] J. Manyemba and B. M. Mayosi, “Intramuscular penicillin is more effective than oral penicillin in secondary prevention of rheumatic fever - A systematic review,” *South African Medical Journal*, vol. 93, no. 3, pp. 212–218, mar 2003.
- [74] A. C. Steer and J. R. Carapetis, “Prevention and treatment of rheumatic heart disease in the developing world,” *Nature Reviews. Cardiology*, vol. 6, no. 11, pp. 689–98, nov 2009.
- [75] “Rheumatic Fever and Rheumatic Heart Disease: Report of a WHO Expert Consultation,” World Health Organization, Geneva, Tech. Rep., 2001.
- [76] L. Sörnmo and P. Laguna, “The electrocardiogram - A brief background,” in *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. Amsterdam: Academic Press, 2005, ch. 6, p. 411.
- [77] “Unlabelled diagram of the heart.” [Online]. Available: <http://www.clipartbest.com/clipart-xTgojLrnc>
- [78] A. K. Abbas and R. Bassam, *Phonocardiography Signal Processing*. Morgan & Claypool Publishers, 2009.
- [79] A. Leatham, *Auscultation of the heart and phonocardiography*. Churchill Livingstone, 1975.
- [80] J. M. Felner, “The first heart sound,” in *Clinical Methods: The History, Physical, and Laboratory Examinations.*, 3rd ed., H. Walker, W. Hall, and J. Hurst, Eds. Boston: Butterworths, 1990, ch. 22, pp. 117–121.

- [81] S. Mangione and L. Z. Nieman, "Cardiac auscultatory skills among internal medicine and family practice trainees: A comparison of diagnostic proficiency," *Journal of the American Medical Association*, vol. 278, no. 9, pp. 717–22, sep 1997.
- [82] P. R. A. Gaskin, S. E. Owens, N. S. Talner, S. P. Sanders, and J. S. Li, "Clinical Auscultation Skills in Pediatric Residents," *Pediatrics*, vol. 105, no. 6, pp. 1184–1187, jun 2000.
- [83] J. M. Vukanovic-Criley, S. Criley, C. M. Warde, J. R. Boker, L. Guevara-Matheus, W. H. Churchill, W. P. Nelson, and J. M. Criley, "Competency in cardiac examination skills in medical students, trainees, physicians, and faculty: a multicenter study." *Archives of Internal Medicine*, vol. 166, no. 6, pp. 610–616, mar 2006.
- [84] W. R. Thompson, C. S. Hayek, C. Tuchinda, J. K. Telford, and J. S. Lombardo, "Automated cardiac auscultation for detection of pathologic heart murmurs." *Pediatric Cardiology*, vol. 22, no. 5, pp. 373–9, sep 2001.
- [85] M. E. Tavel, "Cardiac auscultation: a glorious past—and it does have a future!" *Circulation*, vol. 113, no. 9, pp. 1255–9, mar 2006.
- [86] R. Watrous, W. Thompson, and S. Ackerman, "The impact of computer-assisted auscultation on physician referrals of asymptomatic patients with heart murmurs," *Clinical Cardiology*, vol. 83, pp. 79–83, feb 2008.
- [87] L. Zühlke, L. Myer, and B. M. Mayosi, "The promise of computer-assisted auscultation in screening for structural heart disease and clinical teaching." *Cardiovascular Journal of Africa*, vol. 23, no. 7, pp. 405–8, feb 2012.
- [88] G. Douglas, E. F. Nicol, and C. Robertson, *Macleod's Clinical Examination*, twelfth ed. Elsevier Health Sciences, 2009.
- [89] A. G. Tilkian and M. B. Conover, *Understanding heart sounds and murmurs: with an introduction to lung sounds*, 4th ed. Saunders, 2001.
- [90] A. N. Pelech, "The physiology of cardiac auscultation," *Pediatric Clinics of North America*, vol. 51, no. 6, pp. 1515–1535, dec 2004.
- [91] J. M. Felner, "The second heart sound," in *Clinical Methods: The History, Physical, and Laboratory Examinations*, 3rd ed., H. Walker, W. Hall, and J. Hurst, Eds. Boston: Butterworths, 1990, ch. 23.
- [92] P. J. Arnott, G. W. Pfeiffer, and M. E. Tavel, "Spectral analysis of heart sounds: Relations between some physical characteristics and frequency spectra of first and second heart sounds in normals and hypertensives," *Journal of Biomedical Engineering*, vol. 6, pp. 121–128, apr 1984.
- [93] E. Braun-Menendez, "The heart sounds in normal and pathological conditions," *The Lancet*, vol. 232, no. 6005, pp. 761–767, oct 1938.
- [94] M. McKee, "Heart sounds in normal children," *American Heart Journal*, vol. 16, no. 1, pp. 79–87, jul 1938.
- [95] —, "Heart sounds and murmurs in children with rheumatic heart disease," *American Heart Journal*, vol. 16, no. 1, pp. 88–98, jul 1938.
- [96] H. Sacks and R. Roberts, "Heart sounds in young adults," *American Heart Journal*, vol. 18, no. 3, pp. 303–307, sep 1939.
- [97] A. A. Luisada, F. Mendoza, and M. M. Alimurung, "The duration of normal heart sounds." *British Heart Journal*, vol. 11, no. 1, pp. 41–7, jan 1949.
- [98] C. Aravanis, L. Feigen, and A. Luisada, "Durations and intervals of normal heart sounds in man," *American Heart Journal*, vol. 82, no. 2, pp. 187–192, aug 1971.
- [99] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model." *Physiological Measurement*, vol. 31, no. 4, pp. 513–29, apr 2010.

- [100] H. Nazeran, "Phonocardiography: Recording of Heart Sounds and Murmurs," in *The Physiological Measurement Handbook*, J. G. Webster, Ed. New York, NY, USA: CRC Press, 2014, ch. 5.
- [101] M. A. Alpert, "Systolic Murmurs," in *Clinical Methods: The History, Physical, and Laboratory Examinations.*, 3rd ed., H. Walker, W. Hall, and J. Hurst, Eds. Boston: Butterworths, apr 1990, ch. 26, pp. 131–137.
- [102] E. Hayek, C. N. Gring, and B. P. Griffin, "Mitral valve prolapse," *The Lancet*, vol. 365, pp. 507–518, feb 2005.
- [103] I. S. Crawley, "Diastolic murmurs," in *Clinical Methods: The History, Physical, and Laboratory Examinations.*, 3rd ed., H. Walker, W. Hall, and J. Hurst, Eds. Boston: Butterworths, 1990, ch. 27, pp. 138–141.
- [104] T. Biancaniello, "Innocent murmurs." *Circulation*, vol. 111, no. 3, pp. e20–2, jan 2005.
- [105] G. F. Still, *Common Disorders and Diseases of Childhood*. London: Hodder and Stoughton, 1909.
- [106] N. Advani, S. Menahem, and J. L. Wilkinson, "The diagnosis of innocent murmurs in childhood." *Cardiology in the Young*, vol. 10, no. 4, pp. 340–2, oct 2000.
- [107] A.-L. Noponen, S. Lukkarinen, A. Angerla, K. Sikio, and R. Sepponen, "How to recognize the innocent vibratory murmur," in *Computers in Cardiology*, vol. 27, sep 2000, pp. 561–564.
- [108] A.-L. Noponen, S. Lukkarinen, A. Angerla, and R. Sepponen, "Phono-spectrographic analysis of heart murmur in children." *BMC Pediatrics*, vol. 7, no. 23, pp. 1–10, jan 2007.
- [109] M. K. Park, *Acquired heart disease*, 6th ed. Philadelphia, PA, USA: Elsevier - Health Sciences Division, 2014.
- [110] J. R. Carapetis, "The stark reality of rheumatic heart disease," *European Heart Journal*, vol. 36, pp. 1070–1073, jan 2015.
- [111] G. Marascio and P. A. Modesti, "Current trends and perspectives for automated screening of cardiac murmurs," *Heart Asia*, vol. 5, no. 1, pp. 213–218, aug 2013.
- [112] Zodiac Stainless Products Company, "Stainless Steel Egg Cups," 2015. [Online]. Available: <http://www.zodiacspco.co.uk/products/731-stainless-steel-4-egg-cups-carded.aspx>
- [113] GSM Arena Team, "Apple iPhone 3GS review: Same clothes, new feel," 2009. [Online]. Available: <http://www.gsmarena.com/apple{ }iphone{ }3gs-review-369p6.php{#}aq>
- [114] —, "Nokia 3500 classic review: Classically simple," 2008. [Online]. Available: <http://www.gsmarena.com/nokia{ }3500{ }classic-review-238.php>
- [115] L. Zühlke, M. Engel, S. Nkepu, and B. Mayosi, "Evaluation of a focused protocol for hand-held echocardiography and computer-assisted auscultation in detecting latent rheumatic heart disease in scholars," *To appear in: Cardiology in the Young*, 2015.
- [116] GSM Arena Team, "Samsung I9100 Galaxy S II review: Brightest star," 2011. [Online]. Available: <http://www.gsmarena.com/samsung{ }i9100{ }galaxy{ }s{ }ii-review-597.php>
- [117] GE Healthcare, "The Vscan family," 2015. [Online]. Available: <http://www3.gehealthcare.com/en/products/categories/ultrasound/vscan{ }family>
- [118] Philips, "CX50 CompactXtreme Ultrasound System," 2014. [Online]. Available: <http://www.healthcare.philips.com/main/products/ultrasound/systems/cx50/>
- [119] Z. Syed, "MIT Automated Auscultation System," Masters Thesis, Massachusetts Institute of Technology, 2003.

- [120] Z. Syed, D. Leeds, D. Curtis, F. Nesta, R. Levine, and J. Guttag, "A framework for the analysis of acoustical cardiac signals," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 4, pp. 651–662, apr 2007.
- [121] D. Kumar, P. Carvalho, M. Antunes, R. P. Paiva, and J. Henriques, "Noise detection during heart sound recording using periodicity signatures," *Physiological Measurement*, vol. 32, no. 5, pp. 599–618, may 2011.
- [122] I. Cathers, "Neural network assisted cardiac auscultation," *Artificial Intelligence in Medicine*, vol. 7, no. 1, pp. 53–66, feb 1995.
- [123] L. Gamero and R. Watrous, "Detection of the first and second heart sound using probabilistic models," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Cancun, Mexico: IEEE, aug 2003, pp. 2877–2880.
- [124] T. Oskiper and R. Watrous, "Detection of the first heart sound using a time-delay neural network," in *Computers in Cardiology*. Memphis, TN, USA: IEEE, sep 2002, pp. 537–540.
- [125] C. Gupta, R. Palaniappan, S. Swaminathan, and S. Krishnan, "Neural network classification of homomorphic segmented heart sounds," *Applied Soft Computing*, vol. 7, no. 1, pp. 286–297, jan 2007.
- [126] J. P. de Vos and M. M. Blanckenberg, "Automated pediatric cardiac auscultation," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 2, pp. 244–52, feb 2007.
- [127] J. Martinez-Alajarin and R. Ruiz-Merino, "Efficient method for events detection in phonocardiographic signals," *Proceedings of SPIE - Microtechnologies for the New Millennium*, pp. 398–409, jun 2005.
- [128] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelopogram and self-organizing probabilistic model," in *Computers in Cardiology*. Lyon, France: IEEE, sep 2005, pp. 957–960.
- [129] A. Ricke, R. Povinelli, and M. Johnson, "Automatic segmentation of heart sound signals using hidden Markov models," *Computers in Cardiology*, pp. 953–956, sep 2005.
- [130] A. Moukadem, A. Dieterlen, and C. Brandt, "Shannon Entropy based on the S-Transform Spectrogram applied on the classification of heart sounds," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, Canada: IEEE, may 2013, pp. 704–708.
- [131] D. Kumar, P. Carvalho, M. Antunes, J. Henriques, L. Eugenio, R. Schmidt, and J. Habetha, "Detection of S1 and S2 heart sounds by high frequency signatures," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, New York, NY, USA, oct 2006, pp. 1410–6.
- [132] A. S. Paul, E. A. Wan, and A. T. Nelson, "Noise reduction for heart sounds using a modified minimum-mean squared error estimator with ECG gating," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, New York, New York, USA, 2006, pp. 3385–3390.
- [133] H. Tang, T. Li, and T. Qiu, "Noise and disturbance reduction for heart sounds in cycle-frequency domain based on nonlinear time scaling," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 2, pp. 325–333, oct 2010.
- [134] H. Tang, T. Li, Y. Park, and T. Qiu, "Separation of heart sound signal from noise in joint cycle frequency-time-frequency domains based on fuzzy detection," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2438–2447, oct 2010.
- [135] L. Hall, J. Maple, J. Agzarian, and D. Abbott, "Sensor system for heart sound biomonitor," *Microelectronics Journal*, vol. 31, no. 7, pp. 583–592, jul 2000.
- [136] S. R. Messer, J. Agzarian, and D. Abbott, "Optimal wavelet denoising for phonocardiograms," *Microelectronics Journal*, vol. 32, no. 12, pp. 931–941, dec 2001.

- [137] H. Liang, L. Sakari, and H. Iiro, "A heart sound segmentation algorithm using wavelet decomposition and reconstruction," in *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4, Chicago, IL, USA, oct 1997, pp. 1630–1633.
- [138] H. Liang and I. Hartimo, "A heart sound feature extraction algorithm based on wavelet decomposition and reconstruction," in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, vol. 3, no. 3. Ieee, oct 1998, pp. 1539–1542.
- [139] S. Ari and G. Saha, "In search of an optimization technique for Artificial Neural Network to classify abnormal heart sounds," *Applied Soft Computing*, vol. 9, no. 1, pp. 330–340, jan 2009.
- [140] E. Delgado-Trejos, A. F. Quiceno-Manrique, J. I. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez, "Digital auscultation analysis for heart murmur detection." *Annals of Biomedical Engineering*, vol. 37, no. 2, pp. 337–53, feb 2009.
- [141] T.-H. Chen, P.-Y. Guo, S.-X. Xing, and Y. Zheng, "A New Method of Sorting of Heart Sound Signal Based on Wavelet Transform and Parameter Model Method," in *4th International Conference on Bioinformatics and Biomedical Engineering*. Chengdu, China: IEEE, jun 2010, pp. 1–4.
- [142] S. Choi and Z. Jiang, "Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique." *Computers in Biology and Medicine*, vol. 40, no. 1, pp. 8–20, jan 2010.
- [143] S. Yuenyong, A. Nishihara, W. Kongprawechnon, and K. Tungpimolrut, "A framework for automatic heart sound analysis without segmentation." *Biomedical Engineering Online*, vol. 10, no. 1, p. 13, jan 2011.
- [144] D. Gradolewski and G. Redlarski, "Wavelet-based denoising method for real phonocardiography signal recorded by mobile devices in noisy environment." *Computers in Biology and Medicine*, vol. 52, pp. 119–29, sep 2014.
- [145] A. Almasi, M. Bagher Shamsollahi, and L. Senhadji, "Bayesian denoising framework of phonocardiogram based on a new dynamical model," *IRBM*, vol. 34, no. 3, pp. 214–225, jun 2013.
- [146] D. Barschdorff, U. Femmer, and E. Trowitzsch, "Automatic phonocardiogram signal analysis in infants based on wavelet transforms and artificial neural networks," in *Computers in Cardiology*, Vienna, Austria, sep 1995, pp. 753–756.
- [147] F. Beritelli and A. Spadaccini, "Heart sounds quality analysis for automatic cardiac biometry applications," in *Proceedings of the 1st IEEE International Workshop on Information Forensics and Security*, London, UK, dec 2009, pp. 61–65.
- [148] B. Bogert, M. Healy, and J. Tukey, "The quefrency alanalysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Time Series Analysis*, M. Rosenblatt, Ed., 1963, ch. 15, pp. 209–243.
- [149] T. Li, H. Tang, T. Qiu, and Y. Park, "Best subsequence selection of heart sound recording based on degree of sound periodicity," *Electronics Letters*, vol. 47, no. 15, p. 841, jul 2011.
- [150] T. Li, T. Qiu, and H. Tang, "Optimum heart sound signal selection based on the cyclostationary property," *Computers in Biology and Medicine*, vol. 43, no. 6, pp. 607–612, mar 2013.
- [151] G. Moody and R. Mark, "QRS morphology representation and noise estimation using the Karhunen-Loeve transform," in *Computers in Cardiology*, Jerusalem, Isreal, sep 1989, pp. 269 – 272.
- [152] W. Zong, G. B. Moody, and R. G. Mark, "Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure," *Medical and Biological Engineering and Computing*, vol. 42, no. 5, pp. 698–706, sep 2004.

- [153] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter." *Physiological Measurement*, vol. 29, no. 1, pp. 15–32, jan 2008.
- [154] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals." *Physiological Measurement*, vol. 33, no. 9, pp. 1491–501, sep 2012.
- [155] V. Monasterio, F. Burgess, and G. D. Clifford, "Robust classification of neonatal apnoea-related desaturations." *Physiological Measurement*, vol. 33, no. 9, pp. 1503–16, sep 2012.
- [156] S. J. Redmond, N. H. Lovell, J. Basilakis, and B. G. Celler, "ECG quality measures in telecare monitoring." in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vancouver, Canada: IEEE, aug 2008, pp. 2869–72.
- [157] S. J. Redmond, Y. Xie, D. Chang, J. Basilakis, and N. H. Lovell, "Electrocardiogram signal quality measures for unsupervised telehealth environments." *Physiological Measurement*, vol. 33, no. 9, pp. 1517–33, sep 2012.
- [158] I. Silva, G. B. Moody, and L. Celi, "Improving the quality of ECGs collected using mobile phones: The PhysioNet/Computing in Cardiology Challenge 2011," in *Computing in Cardiology*. Hangzhou, China: IEEE, sep 2011, pp. 273–276.
- [159] H. Xia, G. A. Garcia, J. C. McBride, A. Sullivan, T. D. Bock, J. Bains, D. C. Wortham, and X. Zhao, "Computer algorithms for evaluating the quality of ECGs in real time," *Computers in Cardiology*, vol. 38, pp. 369–372, sep 2011.
- [160] H. Xia, G. A. Garcia, J. Bains, D. C. Wortham, and X. Zhao, "Matrix of regularity for improving the quality of ECGs." *Physiological Measurement*, vol. 33, no. 9, pp. 1535–48, sep 2012.
- [161] G. D. Clifford, J. Behar, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms." *Physiological Measurement*, vol. 33, no. 9, pp. 1419–33, sep 2012.
- [162] J. Behar, J. Oster, Q. Li, and G. D. Clifford, "ECG signal quality during arrhythmia and its application to false alarm reduction." *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 6, pp. 1660–6, jun 2013.
- [163] H. Naseri and M. R. Homaeinezhad, "Computerized quality assessment of phonocardiogram signal measurement-acquisition parameters." *Journal of Medical Engineering & Technology*, vol. 36, no. 6, pp. 308–18, aug 2012.
- [164] C. Shannon, "Communication In The Presence Of Noise," *Proceedings of the IEEE*, vol. 86, no. 2, pp. 447–457, feb 1998.
- [165] A. Sarkady, R. Clark, and R. Williams, "Computer analysis techniques for phonocardiogram diagnosis," *Computers and Biomedical Research*, vol. 9, no. 4, pp. 349–363, aug 1976.
- [166] I. Rezek and S. Roberts, "Envelope Extraction via Complex Homomorphic Filtering. Technical Report TR-98-9," Imperial College, London, Tech. Rep., 1998.
- [167] S. J. Orfanidis, *Introduction to Signal Processing*, ser. Prentice Hall Signal Processing Series. Prentice Hall, 1996.
- [168] G. Box, G. M. Jenkins, and G. Reinsel, *Time Series Analysis: Forecasting & Control (3rd Edition)*. Prentice Hall, 1994.
- [169] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th ed. Wiley, 2006.
- [170] J. J. Richman and J. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–49, jun 2000.

- [171] T. A. Kuusela, T. T. Jartti, K. U. O. Tahvanainen, and T. J. Kaila, "Nonlinear methods of biosignal analysis in assessing terbutaline-induced heart rate and blood pressure changes." *American Journal of Heart and Circulatory Physiology*, vol. 282, no. 2, pp. H773–83, feb 2002.
- [172] M. Costa, A. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of biological signals," *Physical Review E*, vol. 71, no. 2, pp. 1–18, feb 2005.
- [173] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Maconochie, L. Tarassenko, and D. Mant, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies," *Lancet*, vol. 377, no. 9770, pp. 1011–8, mar 2011.
- [174] L. T. DeCarlo, "On the meaning and use of kurtosis." *Psychological Methods*, vol. 2, no. 3, pp. 292–307, jun 1997.
- [175] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelopogram," in *Computers in Cardiology*, vol. 24, Lund, Sweden, sep 1997, pp. 105–108.
- [176] American National Standards Institute, "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms: ANSI/AAMI EC57," 2012.
- [177] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, apr 2011.
- [178] Q. Wei and R. L. Dunbrack, "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," *PLoS ONE*, vol. 8, no. 7, jul 2013.
- [179] R. Batuwita and V. Palade, *Class Imbalance Learning Methods for Support Vector Machines*. Wiley, 2013, ch. 6.
- [180] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer-Verlag, 2009.
- [181] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Taipei, Tech. Rep., 2010. [Online]. Available: <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>
- [182] Q. Liu, C. Chen, Y. Zhang, and Z. Hu, "Feature selection for support vector machines with RBF kernel," *Artificial Intelligence Review*, vol. 36, no. 2, pp. 99–115, aug 2011.
- [183] M. Manikandan, S. Dandapat, and M. Sabarimalai Manikandan, "Wavelet energy based compression of phonocardiogram (PCG) signal for telecardiology," *International Conference on Information and Communication Technology in Electrical Sciences*, pp. 650–654, dec 2007.
- [184] V. Nigam and R. Priemer, "Assessing heart dynamics to estimate durations of heart sounds." *Physiological Measurement*, vol. 26, no. 6, pp. 1005–18, dec 2005.
- [185] S. Ari, P. Kumar, and G. Saha, "A robust heart sound segmentation algorithm for commonly occurring heart valve diseases." *Journal of Medical Engineering & Technology*, vol. 32, no. 6, pp. 456–65, jan 2008.
- [186] J. Vepa, P. Tolay, and A. Jain, "Segmentation of heart sounds using simplicity features and timing information," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, mar 2008, pp. 469–472.
- [187] S. Choi and Z. Jiang, "Comparison of envelope extraction algorithms for cardiac sound signal segmentation," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1056–1069, feb 2008.

- [188] S. Rajan, E. Budd, M. Stevenson, and R. Doraiswami, "Unsupervised and uncued segmentation of the fundamental heart sounds in phonocardiograms using a time-scale representation." in *International Conference of the IEEE Engineering in Medicine and Biology Society*, New York City, New York, USA, aug 2006, pp. 3732–5.
- [189] A. A. Sepehri, A. Gharehbaghi, T. Dutoit, A. Kocharian, and A. Kiani, "A novel method for pediatric heart sound segmentation without using the ECG." *Computer Methods and Programs in Biomedicine*, vol. 99, no. 1, pp. 43–8, jul 2010.
- [190] Z. Yan, Z. Jiang, A. Miyamoto, and Y. Wei, "The moment segmentation analysis of heart sound pattern." *Computer Methods and Programs in Biomedicine*, vol. 98, no. 2, pp. 140–50, may 2010.
- [191] S. Sun, Z. Jiang, H. Wang, and Y. Fang, "Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified Hilbert transform," *Computer Methods and Programs in Biomedicine*, vol. 114, no. 3, pp. 219–230, may 2014.
- [192] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, "A robust heart sounds segmentation module based on S-transform," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 273–281, may 2013.
- [193] H. Tang, T. Li, T. Qiu, and Y. Park, "Segmentation of heart sounds based on dynamic clustering," *Biomedical Signal Processing and Control*, vol. 7, no. 5, pp. 509–516, sep 2012.
- [194] H. Naseri and M. R. Homaeinezhad, "Detection and Boundary Identification of Phonocardiogram Sounds Using an Expert Frequency-Energy Based Metric," *Annals of Biomedical Engineering*, vol. 41, no. 2, pp. 279–92, feb 2013.
- [195] V. N. Varghees and K. Ramachandran, "A novel heart sound activity detection framework for automated heart sound analysis," *Biomedical Signal Processing and Control*, vol. 13, pp. 174–188, sep 2014.
- [196] C. D. Papadaniil and L. J. Hadjileontiadis, "Efficient heart sound segmentation and extraction using ensemble empirical mode decomposition and kurtosis features." *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1138–52, jul 2014.
- [197] Z. Jiang and S. Choi, "A cardiac sound characteristic waveform method for in-home heart disorder monitoring with electric stethoscope," *Expert Systems with Applications*, vol. 31, no. 2, pp. 286–298, aug 2006.
- [198] A. Sepehri, J. Hancq, T. Dutoit, A. Gharehbaghi, A. Kocharian, and A. Kiani, "Computerized screening of children congenital heart diseases." *Computer methods and programs in biomedicine*, vol. 92, no. 2, pp. 186–92, nov 2008.
- [199] G. M. Friesen, T. C. Jannett, M. Jadallah, S. L. Yates, S. R. Quint, and H. T. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms." *IEEE Transactions on Biomedical Engineering*, vol. 37, no. 1, pp. 85–98, jan 1990.
- [200] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A Wavelet-Based ECG Delineator: Evaluation on Standard Databases," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 570–581, apr 2004.
- [201] C. Li, C. Zheng, and C. Tai, "Detection of ECG characteristic points using wavelet transforms," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 1, pp. 21–28, jan 1995.
- [202] A. Illanes-Manriquez and Q. Zhang, "An algorithm for QRS onset and offset detection in single lead electrocardiogram records," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Lyon, France, aug 2007, pp. 541 – 544.
- [203] J. Behar, A. Johnson, G. D. Clifford, and J. Oster, "A comparison of single channel fetal ecg extraction methods," *Annals of Biomedical Engineering*, vol. 42, no. 6, pp. 1340–1353, jun 2014.

- [204] J. Behar, J. Oster, and G. D. Clifford, "Combining and Benchmarking Methods of Foetal ECG Extraction Without Maternal or Scalp Electrode Data," *Physiological Measurement*, vol. 35, no. 8, pp. 1569–1589, aug 2014.
- [205] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–6, mar 1985.
- [206] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–E220, 2000.
- [207] P. Laguna, R. Jané, and P. Caminal, "Automatic detection of wave boundaries in multilead ECG signals: validation with the CSE database," *Computers and Biomedical Research*, vol. 27, no. 1, pp. 45–60, feb 1994.
- [208] Q. Zhang, A. I. Manriquez, C. Medigue, Y. Papelier, and M. Sorine, "An algorithm for robust and efficient location of T-wave ends in electrocardiograms," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12 (Pt 1), pp. 2544–52, dec 2006.
- [209] C. R. Vázquez-Seisdedos, J. E. Neto, E. J. Marañón Reyes, A. Klautau, and R. C. Limão de Oliveira, "New approach for T-wave end detection on electrocardiogram: performance in noisy conditions," *Biomedical Engineering Online*, vol. 10, no. 1, pp. 77–87, jan 2011.
- [210] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, feb 1989.
- [211] N. P. Hughes, "Probabilistic Models for Automated ECG Interval Analysis," DPhil Thesis, University of Oxford, 2006.
- [212] K. P. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," Ph.D. thesis, University of California, Berkeley, 2002.
- [213] S.-Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, feb 2010.
- [214] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Advances in Neural Information Processing Systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14. Cambridge, MA: MIT Press, 2001, pp. 841–848.
- [215] F. Marzbanrad, Y. Kimura, K. Funamoto, R. Sugibayashi, M. Endo, T. Ito, M. Palaniswami, and A. H. Khandoker, "Automated Estimation of Fetal Cardiac Timing Events From Doppler Ultrasound Signal Using Hybrid Models," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1169–1177, jul 2014.
- [216] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [217] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, Williamstown, MA, USA, jun 2001, pp. 406–417.
- [218] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. Cambridge, MA, USA: MIT Press, 2006, ch. 4, pp. 93–128.
- [219] H. Wallach, "Conditional random fields: An introduction," Department of Computer and Information Science, University of Pennsylvania, Tech. Rep., 2004. [Online]. Available: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1011&context=cis_reports
- [220] F. Sha and F. Pereira, "Shallow Parsing with Conditional Random Fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 134–141.

- [221] X. He, R. Zemel, and M. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. Washington, DC, USA: IEEE, jun 2004, pp. 695–702.
- [222] L. T. Vinh, S. Lee, H. X. Le, H. Q. Ngo, H. I. Kim, M. Han, and Y.-K. Lee, "Semi-Markov conditional random fields for accelerometer-based activity recognition," *Applied Intelligence*, vol. 35, no. 2, pp. 226–241, mar 2010.
- [223] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," in *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds. Whistler, BC, Canada: MIT Press, dec 2005, pp. 1185–1192.
- [224] G. De Lannoy, D. François, J. Delbeke, and M. Verleysen, "Weighted conditional random fields for supervised interpatient heartbeat classification," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 1, pp. 241–247, 2012.
- [225] H. L. Chieu, W. S. Lee, and L. P. Kaelbling, "Activity Recognition from Physiological Data using Conditional Random Fields," in *Singapore-MIT Alliance Symposium*, Singapore, jan 2006.
- [226] J. F. D. Saa and M. Cetin, "Hidden conditional random fields for classification of imaginary motor tasks from EEG data," in *19th European Signal Processing Conference*. Barcelona, Spain: IEEE, sep 2011, pp. 1377–1381.
- [227] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields," *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2011.
- [228] M. Schmidt and K. Swersky, "crfChain: Matlab code for chain-structured conditional random fields with categorical features." 2008. [Online]. Available: <http://www.cs.ubc.ca/~schmidtm/Software/crfChain.html>
- [229] L. van der Maaten, "Hidden-Unit Conditional Random Fields," 2011. [Online]. Available: <http://cseweb.ucsd.edu/~lvdmaaten/hucrf/Hidden-Unit{ }Conditional{ }Random{ }Fields.html>
- [230] S. W. Smith, "Audio Processing," in *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd ed. San Diego, California: California Technical Publishing, 1999, ch. 22, p. 351.
- [231] B. Ergen, Y. Tatar, and H. O. Gulcur, "Time-frequency analysis of phonocardiogram signals using wavelet transform: a comparative study," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 15, no. 4, pp. 371–81, jan 2012.
- [232] T. Ölmez and Z. Dokur, "Classification of heart sounds using an artificial neural network," *Pattern Recognition Letters*, vol. 24, no. 1-3, pp. 617–629, jan 2003.
- [233] S. Babaei and A. Geranmayeh, "Heart sound reproduction based on neural network classification of cardiac valve disorders using wavelet transforms of PCG signals." *Computers in Biology and Medicine*, vol. 39, no. 1, pp. 8–15, jan 2009.
- [234] S. Ari, K. Hembram, and G. Saha, "Detection of cardiac abnormality from PCG signal using LMS based least square SVM classifier," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8019–8026, dec 2010.
- [235] H. Liang and I. Hartimo, "A feature extraction algorithm based on wavelet packet decomposition for heart sound signals," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Pittsburgh, PA, USA, oct 1998, pp. 93–96.
- [236] Z. Dokur and T. Ölmez, "Feature determination for heart sounds based on divergence analysis," *Digital Signal Processing*, vol. 19, no. 3, pp. 521–531, may 2009.

- [237] R. K. Sinha, Y. Aggarwal, and B. N. Das, "Backpropagation Artificial Neural Network Classifier to Detect Changes in Heart Sound due to Mitral Valve Regurgitation," *Journal of Medical Systems*, vol. 31, no. 3, pp. 205–209, may 2007.
- [238] M. Schmidt, "minFunc," 2005. [Online]. Available: <http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html>
- [239] D. S. Gerbarg, A. Taranta, M. Spagnuolo, and J. J. Hoffer, "Computer analysis of phonocardiograms," *Progress in Cardiovascular Diseases*, vol. 5, no. 4, pp. 393–405, jan 1963.
- [240] R. E. Durnin, H. F. Glassner, C. O. Jorgensen, and D. C. Fyler, "New approach to heart sound screening in children by analog-digital circuitry," *Public Health Reports*, vol. 80, no. 9, pp. 761–770, sep 1965.
- [241] A. L. Klatsky, A. A. Bolomey, and A. B. Siegelau, "Auscultation of the adult heart by machine," *The New England Journal of Medicine*, vol. 279, no. 5, pp. 230–234, aug 1968.
- [242] P. Ninova, I. Dascalov, and M. Dimitrova, "Automated phonocardiographic screening for heart disease in children," *Cardiology*, vol. 63, pp. 5–13, 1978.
- [243] D. Barschdorff, A. Bothe, and U. Rengshausen, "Heart sound analysis using neural and statistical classifiers: a comparison," in *Computers in Cardiology*, Jerusalem, Isreal, sep 1989, pp. 415–418.
- [244] Y. M. Akay, M. Akay, W. Welkowitz, and J. Kostis., "Noninvasive Detection of Coronary Artery Disease," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 13, no. 5, pp. 761–764, nov 1994.
- [245] P. M. Bentley, P. M. Grant, and J. T. McDonnell, "Time-frequency and time-scale techniques for the classification of native and bioprosthetic heart valve sounds," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 1, pp. 125–8, jan 1998.
- [246] T. Leung, P. White, W. Collis, E. Brown, and A. Salmon, "Classification of heart sounds using time-frequency method and artificial neural networks," in *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2, Chicago, IL, USA, jul 2000, pp. 988–991.
- [247] C. C. G. Degroff, S. Bhatikar, J. Hertzberg, R. Shandas, L. Valdes-cruz, and R. R. L. R. Mahajan, "Artificial Neural Network Based Method of Screening Heart Murmurs in Children," *Circulation*, vol. 103, no. 22, pp. 2711–2716, jun 2001.
- [248] S. Pavlopoulos, A. C. H. Stasis, and E. N. Loukis, "A decision tree-based method for the differential diagnosis of Aortic Stenosis from Mitral Regurgitation using heart sounds," *Biomedical Engineering Online*, vol. 3, no. 1, p. 21, jun 2004.
- [249] T. Reed, N. E. Reed, and P. Fritzson, "Heart sound analysis for symptom detection and computer-aided diagnosis," *Simulation Modelling Practice and Theory*, vol. 12, no. 2, pp. 129–146, may 2004.
- [250] S. R. Bhatikar, C. DeGroff, and R. L. Mahajan, "A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics," *Artificial Intelligence in Medicine*, vol. 33, no. 3, pp. 251–60, mar 2005.
- [251] M. E. Tavel and H. Katz, "Usefulness of a new sound spectral averaging technique to distinguish an innocent systolic murmur from that of aortic stenosis," *American Journal of Cardiology*, vol. 95, no. 7, pp. 902–904, apr 2005.
- [252] A. Voss, A. Mix, and T. Hübner, "Diagnosing aortic valve stenosis by parameter extraction of heart sound signals," *Annals of Biomedical Engineering*, vol. 33, no. 9, pp. 1167–1174, sep 2005.
- [253] C. Ahlstrom, P. Hult, P. Rask, J.-E. Karlsson, E. Nylander, U. Dahlström, and P. Ask, "Feature extraction for systolic heart murmur classification," *Annals of Biomedical Engineering*, vol. 34, no. 11, pp. 1666–77, nov 2006.

- [254] M. El-Segaier, E. Pesonen, S. Lukkarinen, K. Peters, L. Sörnmo, and R. Sepponen, "Detection of cardiac pathology: Time intervals and spectral analysis," *Acta Paediatrica, International Journal of Paediatrics*, vol. 96, no. 7, pp. 1036–1042, 2007.
- [255] C. Visagie, "Screening for abnormal heart sounds and murmurs by implementing neural networks," Masters thesis, University of Stellenbosch, 2007.
- [256] P. Wang, C. S. Lim, S. Chauhan, J. Y. Foo, and V. Anantharaman, "Phonocardiographic signal analysis method using a modified hidden Markov model." *Annals of Biomedical Engineering*, vol. 35, no. 3, pp. 367–74, mar 2007.
- [257] S. Chauhan, P. Wang, C. Sing Lim, and V. Anantharaman, "A computer-aided MFCC-based HMM system for automatic auscultation." *Computers in Biology and Medicine*, vol. 38, no. 2, pp. 221–33, feb 2008.
- [258] Z. Dokur and T. Ölmez, "Heart sound classification using wavelet transform and incremental self-organizing map," *Digital Signal Processing*, vol. 18, no. 6, pp. 951–959, nov 2008.
- [259] V. Nigam and R. Priemer, "A simplicity-based fuzzy clustering approach for detection and extraction of murmurs from the phonocardiogram." *Physiological measurement*, vol. 29, no. 1, pp. 33–47, jan 2008.
- [260] S. Choi, "Detection of valvular heart disorders using wavelet packet decomposition and support vector machine," *Expert Systems with Applications*, vol. 35, no. 4, pp. 1679–1687, nov 2008.
- [261] I. Maglogiannis, E. Loukis, E. Zafiropoulos, and A. Stasis, "Support Vectors Machine-based identification of heart valve diseases using heart sounds." *Computer Methods and Programs in Biomedicine*, vol. 95, no. 1, pp. 47–61, jul 2009.
- [262] A. F. Quiceno-Manrique, J. I. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez, "Selection of dynamic features based on time-frequency representations for heart murmur detection from phonocardiographic signals." *Annals of Biomedical Engineering*, vol. 38, no. 1, pp. 118–37, jan 2010.
- [263] L. D. Avendaño-Valencia, J. I. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez, "Feature extraction from parametric time-frequency representations for heart murmur detection." *Annals of Biomedical Engineering*, vol. 38, no. 8, pp. 2716–32, aug 2010.
- [264] W.-C. Kao and C.-C. Wei, "Automatic phonocardiograph signal analysis for detecting heart valve disorders," *Expert Systems with Applications*, vol. 38, no. 6, pp. 6458–6468, jun 2011.
- [265] C. Kwak and O.-W. Kwon, "Cardiac disorder classification by heart sound signals using murmur likelihood and hidden Markov model state likelihood," *IET Signal Processing*, vol. 6, no. 4, p. 326, jan 2012.
- [266] F. Safara, S. Doraisamy, A. Azman, A. Jantan, and S. Ranga, "Wavelet Packet Entropy for Heart Murmurs Classification," *Advances in Bioinformatics*, no. 25, pp. 1–6, jan 2012.
- [267] H. UÇŞuz, "A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases," *Journal of Medical Systems*, vol. 36, no. 1, pp. 61–72, 2012.
- [268] H. Uguz, "Adaptive neuro-fuzzy inference system for diagnosis of the heart valve diseases using wavelet transform with entropy," *Neural Computing and Applications*, vol. 21, no. 7, pp. 1617–1628, apr 2012.
- [269] R. Saracoglu, "Hidden Markov model-based classification of heart valve disease with PCA for dimension reduction," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 7, pp. 1523–1528, oct 2012.

- [270] F. Safara, S. Doraisamy, A. Azman, A. Jantan, and A. R. Abdullah Ramaiah, "Multi-level basis selection of wavelet packet decomposition tree for heart sound classification." *Computers in Biology and Medicine*, vol. 43, no. 10, pp. 1407–14, oct 2013.
- [271] G. Redlarski, D. Gradolewski, and A. Palkowski, "A system for heart sounds classification." *PloS one*, vol. 9, no. 11, pp. 1–12, jan 2014.
- [272] S. Schmidt, C. Holst-Hansen, J. Hansen, E. Toft, and J. Struijk, "Acoustic features for the identification of coronary artery disease," *To appear in IEEE Transactions on Biomedical Engineering*, 2015.
- [273] A. Gharehbaghi, T. Dutoit, P. Ask, and L. Sörnmo, "Detection of systolic ejection click using time growing neural network," *Medical Engineering and Physics*, vol. 36, no. 4, pp. 477–483, 2014.
- [274] Y. Zheng, X. Guo, and X. Ding, "A novel hybrid energy fraction and entropy-based approach for systolic heart murmurs identification," *Expert Systems With Applications*, vol. 42, no. 5, pp. 2710–2721, nov 2015.
- [275] S. Patidar, R. B. Pachori, and N. Garg, "Automatic diagnosis of septal defects based on tunable-Q wavelet transform of cardiac sound signals," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3315–3326, 2015.
- [276] A. Gharehbaghi, M. Borga, B. J. Sjöberg, and P. Ask, "A novel method for discrimination between innocent and pathological heart murmurs," *Medical Engineering & Physics*, vol. 37, no. 7, pp. 674–682, 2015.
- [277] S. Schmidt, M. Graebe, E. Toft, and J. Struijk, "No evidence of nonlinear or chaotic behavior of cardiovascular murmurs," *Biomedical Signal Processing and Control*, vol. 6, no. 2, pp. 157–163, apr 2011.
- [278] M. El-Segaier, O. Lilja, S. Lukkarinen, L. Sörnmo, R. Sepponen, and E. Pesonen, "Computer-based detection and analysis of heart sound and murmur," *Annals of Biomedical Engineering*, vol. 33, no. 7, pp. 937–942, jul 2005.
- [279] D. B. Springer, T. Brennan, J. Hitzeroth, B. M. Mayosi, L. Tarassenko, and G. D. Clifford, "Robust Heart Rate Estimation from Noisy Phonocardiograms," in *Computing in Cardiology*. Cambridge, MA, USA: IEEE, sep 2014, pp. 613–616.
- [280] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," pp. 374–388, 1976. [Online]. Available: http://www.haskins.yale.edu/sr/SR047/SR047_{_}07.pdf
- [281] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, aug 1980.
- [282] M. Slaney, "Auditory Toolbox," Tech. Rep., 1998. [Online]. Available: <https://engineering.purdue.edu/~malcolm/interval/1998-010/>
- [283] W. van Drongelen, "Signal Averaging," in *Signal Processing for Neuroscientists*, 1st ed. London, UK: Academic Press, 2007, ch. 4, pp. 55–70.
- [284] K. Veropoulos, C. Campbell, N. Cristianini, and Others, "Controlling the sensitivity of support vector machines," *Proceedings of the international joint conference on artificial intelligence*, pp. 55–60, 1999.
- [285] M. A. F. Pimentel, M. D. Santos, D. B. Springer, and G. D. Clifford, "Heart beat detection in multimodal physiological data using a hidden semi-Markov model and signal quality indices," *Physiological Measurement*, vol. 36, no. 8, pp. 1717–1727, aug 2015.
- [286] A. M. Weissler, W. S. Harris, and C. D. Schoenfeld, "Systolic Time Intervals in Heart Failure in Man," *Circulation*, vol. 37, no. 2, pp. 149–159, feb 1968.

- [287] R. P. Lewis, S. E. Rittgers, W. F. Froester, and H. Boudoulas, "A critical review of the systolic time intervals," *Circulation*, vol. 56, no. 2, pp. 146–158, aug 1977.
- [288] H. Boudoulas, P. Geleris, R. P. Lewis, and S. Rittgers, "Linear relationship between electrical systole, mechanical systole, and heart rate." *Chest*, vol. 80, no. 5, pp. 613–617, nov 1981.
- [289] P. C. Loizou, "Spectral-Subtractive Algorithms," in *Speech Enhancement: Theory and Practice*, 2nd ed. London, UK: CRC Press - Taylor and Francis, 2013, ch. 5, pp. 93–136.
- [290] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [291] M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha, "Beyond kappa: A review of interrater agreement measures," *Canadian Journal of Statistics*, vol. 27, no. 1, pp. 3–23, mar 1999.
- [292] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data." *Biometrics*, vol. 33, no. 1, pp. 159–174, mar 1977.
- [293] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.*, 4th ed. Advanced Analytics, LLC, 2014.
- [294] J. Cohen, "Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit." *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, oct 1968.
- [295] H. Janson and U. Olsson, "A Measure of Agreement for Interval or Nominal Multivariate Observations," *Educational and Psychological Measurement*, vol. 61, no. 2, pp. 277–289, 2001.
- [296] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.
- [297] A. J. Conger, "Integration and Generalization of Kappas for Multiple Raters," *Psychological Bulletin*, vol. 88, no. 2, pp. 322–328, sep 1980.
- [298] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2013.
- [299] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, sep 1995.
- [300] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.
- [301] A. Ben-Hur, C. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support Vector Machines and Kernels for Computational Biology," *PLoS Computational Biology*, vol. 4, no. 10, pp. 1–26, oct 2008.
- [302] K. P. Murphy, *Machine learning: a probabilistic perspective*. Cambridge, MA, USA: MIT Press, 2012.
- [303] G. D. Clifford, F. Azuaje, and P. McSharpy, "Linear Filtering Methods," in *Advanced methods and tools for ECG data analysis*. Artech House, 2006, ch. 5, pp. 135–170.
- [304] S. Mallat and C. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Burlington, MA, USA: Academic Press, 1999.
- [305] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, vol. 8, no. 4, pp. 14–38, oct 1991.
- [306] J. Fowler, "The Redundant Discrete Wavelet Transform and Additive Noise," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 629–632, sep 2005.
- [307] M. Shensa, "The discrete wavelet transform: wedding the a trous and Mallat algorithms," *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464–2482, oct 1992.

-
- [308] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, “A real-time algorithm for signal analysis with the help of the wavelet transform,” in *Wavelets, Time-Frequency Methods and Phase Space*, ser. Inverse Problems and Theoretical Imaging, J.-M. Combes, A. Grossmann, and P. Tchamitchian, Eds. Springer Berlin Heidelberg, 1990, pp. 286 – 297.