Computerised Analysis of Fetal Heart Rate

Liang Xu

Wadham College

Supervisor: Dr. Antoniya Georgieva and Prof. Stephen Payne

Life Science Interface Doctoral Training Centre

Department of Engineering Science

University of Oxford

This thesis is submitted to the Department of Engineering Science, University of Oxford, for the degree of Doctor of Philosophy. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

To Katherine, the love of my life

Acknowledgements

Funding for this thesis was generously provided by Clarendon Fund, jointly funded by Wadham College. Essential advice and guidance were administered by Dr. Antoniya Georgieva and Dr. Stephen Payne as my supervisors. I would like to thank their experienced, patient and very helpful guidance, without which none of this research would have happened. Also, I would like to thank Prof. Yiannis Ventikos, my college advisor for his advice and help.

I would like to thank the members and staff of Wadham, DTC, PUMMA Group and Oxford Centre for Fetal Monitoring Technologies. With your friendly help, I could focus on my research and enjoy the life of Oxford.

I would like to thank collaboration and assistance of researchers at the experimental part of John Radcliffe Hospital, especially Prof. Christopher Redman and Miss. Mary Moulden. Also, I would also like to acknowledge the use of the Oxford Supercomputing Centre (OSC) in carrying out this work.

Finally, most thanks for my fiancé Katherine, my parents, my family, my friends and again, Antoniya and Stephen. Thank you, for your love and support all along.

Abstract

This thesis presents a comprehensive work on computerised analysis of fetal heart rate (FHR) features, including feature extraction, feature selection, analysis of influencing factors and setting up/validation of a computerised decision support system.

Firstly, a novel feature – pattern readjustment – was extracted and tested. Clinical data were used to train a Support Vector Machine (SVM) to detect pattern readjustment. Then, the association of pattern readjustment and adverse labour outcome was investigated. The validation results with clinical experts show that the pattern readjustment can be accurately detected, while the study on labour outcome shows that the feature is related to fetal acidemia at birth.

Secondly, Genetic Algorithms were employed as a feature selection method to select a best subset of FHR features and to use them to predict fetal acidemia with linear and nonlinear SVM. The diagnostic power of the classifier output using selected features was tested on the total set of 7,568 cases. As the classifier output increases, there is a consistent increase of the risk of fetal acidemia.

Thirdly, an important influencing factor on FHR features - signal loss – was investigated. A bivariate model was built to estimate error based on signal loss. Validation results show that the bivariate model can accurately predict the error generated by signal loss. The influence of signal loss on labour outcome classification was also investigated.

Finally, a computerised decision support system to estimate the risk of fetal acidemia was set up based on the above studies. The system was validated using new retrospective data. Validation results show that the system is capable of predicting adverse labour outcome and providing timely decision support. It is the first time an intrapartum computerised FHR decision support system has been built and validated on this size of dataset. With further improvements, such a system could be implemented clinically in the long term.

List of Abbreviations

ACOG: American College of Obstetricians and Gynaecologists

AIC: Akaike Information Criterion

AUC: area under curve

BIC: Bayesian Information Criterion

bpm: beat per minute

BU: Baseline Un-assignable

CTG: cardiotocogram

EveREst plot: Event Rate Estimate plot

FHR: fetal heart rate

FIGO: International Federation of Gynaecology and Obstetrics

GA: Genetic Algorithms

LASSO: Least Absolute Shrinkage and Selection Operator

PPV: positive predictive value

PRSA: Phase Rectified Signal Averaging

RCOG: Royal College of Obstetricians and Gynaecologists

RF: Random Forrest

ROC curve: Receiver Operating Characteristic curve

SVM: Support Vector Machine

Contents

1 Introduction

1.1 Clinical background

During the time of labour, a baby has to deal with the stress from the contraction of the uterus, which can sometimes result in reduction of the oxygen supply. Some babies are unable to cope with this situation and suffer from suffocation i.e. birth asphyxia. The largest case study on incidence of birth asphyxia so far reported that birth asphyxia occurred in about 9.4 cases per 1,000 live term births during a four-year period (Palsdottir et al., 2007).

Birth asphyxia may lead to seizures, permanent damage in the neuro-system or even death in occasional severe conditions. It is one of the major causes of neonatal death. In the U.S. birth asphyxia was listed as the 10^{th} leading cause of neonatal death (Miniño et al., 2006). In developing countries, birth asphyxia remains a major cause of death and disability: fatality rates of birth asphyxia can be 40% or higher (Azra Haider and Bhutta, 2006). The World Health Organisation (WHO) estimated that globally between four and nine million newborns suffer birth asphyxia per year, leading to an estimated 1.2 million deaths (29% of all newborn deaths, 0.3% of all newborns) and about the same number of infants who develop severe disability (Bang et al., 2005).

Owing to improvements in obstetric care in developed countries, the death rate of birth asphyxia has reduced significantly and fewer than 0.1% newborn infants die from birth asphyxia in these countries (Badawi et al., 1998). Nevertheless, despite the decreasing death rate, birth asphyxia can still have a huge economic impact (medical expenses) for the

survivors. Birth asphyxia is the 9th most expensive medical condition treated by average hospital cost and resultant hospital charge in the U.S., the cost per case being as high as $74,942 for hospital charges, according to the report of the U.S. Agency for Healthcare Research and Quality (Levit et al., 2009). In this report, it is also noted that the most expensive medical condition - infant respiratory distress syndrome ($138,224 per case) – is linked to birth asphyxia.

Not only is birth asphyxia a major contributor to the death rate and to medical expense, it can also lead to long-term conditions such as cerebral palsy. Cerebral palsy is a group of non-progressive, non-contagious motor conditions that cause physical disability in human development, including abnormal muscle tone, urinary incontinence and intellectual disability. Cerebral palsy occurs in approximately 2 per 1,000 births, of which birth asphyxia accounts for 10-30% (Alberry et al., 2009). The long-term cost of cerebral palsy can be huge: in 2003, a study calculated the average lifetime cost for people with cerebral palsy in the U.S. at $921,000 per individual, including lost income (Honeycutt et al., 2003).

In clinical practice, to prevent birth asphyxia, it is crucial to carry out a timely intervention to assist delivery. Such interventions include Caesarean sections (incisions made through a mother's abdomen and uterus to assist delivery), forceps (use of forceps to assist grabbing, manoeuvring, or removing various things from the body) and ventouse deliveries (a vacuum device used to assist the delivery of a baby). However, such interventions may cause complications, thus they are best avoided when possible (Johanson and Menon, 1999). Therefore, timely and accurate diagnosis of birth asphyxia is essential to minimise the

damage of birth asphyxia while avoiding unnecessary complications.

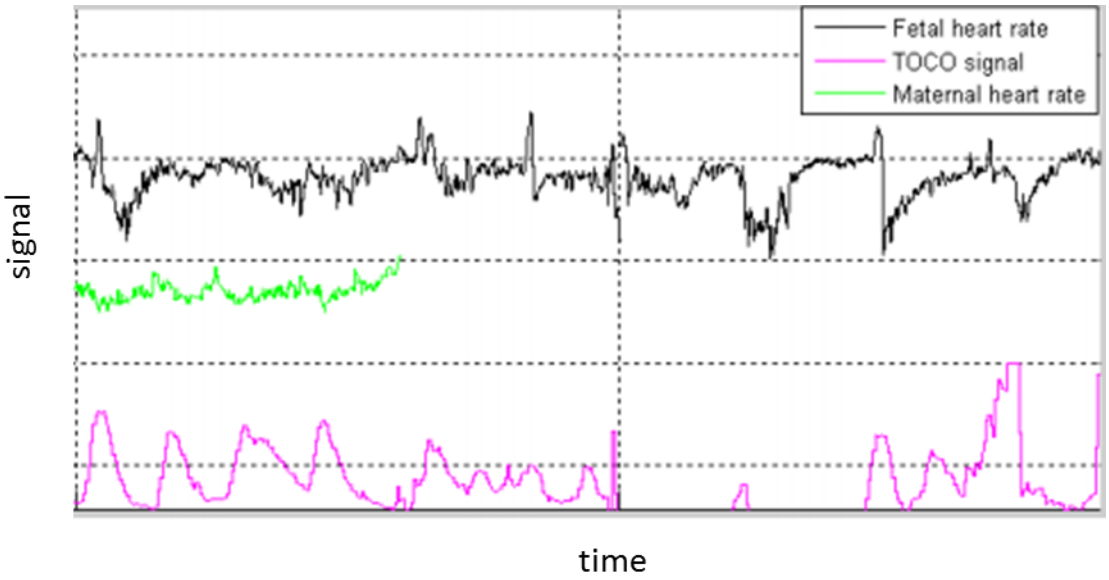
Birth asphyxia can be diagnosed with cord blood gas analyses (measuring the arterial pH) at the time of birth. With fetal acidemia (very low arterial pH) and certain clinical symptoms after birth, birth asphyxia can be diagnosed, since fetal acidemia is a critical indicator that shows the baby did not receive enough oxygen (Malin et al., 2010). Recent studies have shown that low umbilical cord arterial pH is associated with different kind of adverse outcomes, including low Apgar score - a score that measures the health condition of a baby (Apgar, 1953), seizures and other cerebral problems (Yeh et al., 2012, Georgieva et al., 2013a). Most of the time arterial pH is measured at the time of birth. To make timely diagnosis of birth asphyxia, accurate detection of developing fetal acidemia during birth is therefore necessary.

1.2 Computerised Fetal Heart Rate (FHR) Analysis

In order to monitor fetal health, the fetal heart rate (FHR) and uterine contractions are detected (by Doppler ultrasound transducer and tocodynamometer respectively), and electronically recorded during labour on a paper strip called a CTG (cardiotocogram) (Figure 1.1). CTG was invented in the 1960s (Hon, 1963) and become widely used in the 1970s (Gillmer and Combe, 1979). There were controversies about the effectiveness of CTG in the 1980s arguing that it has low predictive power for arterial pH (Steer et al., 1989), and that using the CTG could not reduce neonatal mortality (MacDonald et al., 1985). Nevertheless, a later systematic review collected evidence from large trials (18,561 births) and indicated the effectiveness of the CTG in reducing intrapartum hypoxia by 50% (Vintzileos et al., 1995). A

recent study using U.S. infant death data from 2004 (1,732,211 births) confirmed a significant association between FHR monitoring and a reduction in neonatal/infant mortality, as well as a decrease risk for low Apgar scores and neonatal seizures (Chen et al., 2011).

(A)



(B)

Figure 1.1 The fetal heart rate detection device (A) and the cardiotocogram (B).

There are various methods of measuring fetal heart rate, such as Doppler ultrasound and a spiral electrode placed on the fetal scalp (which also measures the fetal electrocardiogram) (Freeman et al., 2003). Currently most CTG traces are measured using a Doppler ultrasound transducer. The Doppler ultrasound transducer periodically emits a 1-2 MHz waveform, and the transducer also has a receiver to detect the reflected waves. The fetal heart rate can be detected by recording the waveform reflection from the moving parts of the fetal heart, which are frequency shifted with respect to the emitted waves due to the Doppler Effect.

Clinically, normal heart activity is a clinical indicator of adequate blood oxygenation, and a lack of oxygen can be reflected in abnormal FHR patterns such as an absence of baseline variability (Westgate, 2009). Many previous studies on FHR analysis have used fetal acidemia as an indicator of birth asphyxia and have found evidence that abnormal FHR patterns are

associated with fetal acidemia (Turan et al., 2007, Georgieva et al., 2013c, Czabanski et al., 2012).

However, complicated FHR patterns (an example given in Figure 1.2) are usually assessed by eye, which is tedious, error-prone and associated with very low reproducibility owing to notably high inter-observer variability with Cohen's kappa (a measurement of agreement ranging from -1 to 1) below 0.3 for over 80% of the CTG features (Chauhan et al., 2008). In addition, the cost of training the clinician to recognise FHR patterns can be high and the training may be ineffective, especially for developing countries (Costello and Manandhar, 1994). Therefore, computerised analysis for CTG patterns has become a significant and pressing need (Westgate, 2009).

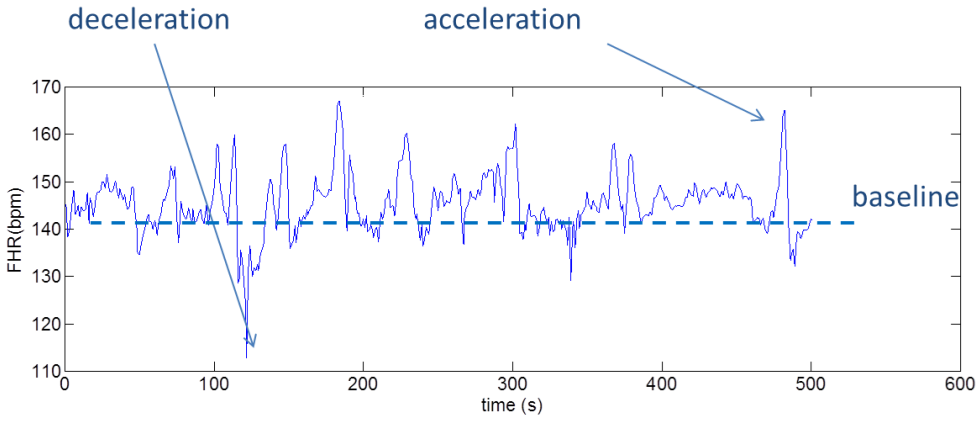


Figure 1.2 A sample of FHR features

As illustrated in Figure 1.2, in FHR analysis various patterns can be observed. These FHR features are the basic elements for FHR analysis, including both clinical patterns and statistical features. Clinical patterns such as acceleration, deceleration and baseline can be observed and used by clinicians as guidelines to evaluate the condition of a baby directly

(ACOG, 2009). On the other hand, statistical measures such as short term variability and the signal stability index can reveal the underlying condition that is difficult to observe visually (Van Laar et al., 2008, Georgieva et al., 2011a). A review of FHR features can be found in Section 2.1. Several systems have been built based on these FHR features to predict adverse labour outcomes and help clinicians in decision making of interventions, details of which can be found in Section 2.2.

Currently, the Oxford Centre for Fetal Monitoring Technologies is developing a computerised system (OxSys) to recognise unfavourable FHR patterns automatically. The object of this thesis is to develop an analysis technique that can be integrated into the OxSys system. The system aims to help to predict adverse labour outcomes, in order to help the clinicians in decision making during labour.

1.3 Summary of Thesis

The studies in this thesis focus on finding the relationship between FHR features (displayed in the intrapartum CTG) and low arterial pH (fetal acidemia, which is the major diagnostic and physical indictor of birth asphyxia). The studies include feature extraction, feature selection and building/validation of a computerised decision support system.

Chapter 2 provides an overview of literature on FHR features, computerised FHR analysis systems, FHR feature extraction and selection methods, and an overview of the database. Chapter 3 presents work on feature extraction for pattern readjustment features. Chapter 4 introduces a feature selection method, Genetic Algorithms (GA), to select a best feature subset out of the FHR features, and to use the feature subset to predict labour outcome.

Chapter 5 looks into an important parameter for the features: signal loss, where the influence of signal loss on feature values and labour outcome classification is investigated and quantified. Chapter 6 uses a new set of clinical data to validate the feature selection method in Chapter 4 and the influence of signal loss in Chapter 5. Chapter 7 gives a conclusion of the work and a discussion of future work.

2 Literature Review

2.1 Overview of FHR Features

In many computerised FHR analyses, FHR features are the basic elements for analysis and decision support. FHR features include clinically meaningful features, such as the number of accelerations and the number of decelerations, as well as features that correspond to the statistical properties of the FHR signal, such as skewness and kurtosis. An intuitive interpretation of these features is critical in assisting the clinicians in decision making, since it provides the clinicians with clear information of the clinical situation.

The FHR features can be categorised into six groups: general features, contraction features, acceleration & deceleration features, lag features, variability features and other statistically derived features. The origin and development of these groups of features are reviewed below.

2.1.1 General features

Over two decades ago, the International Federation of Gynaecology and Obstetrics (FIGO) introduced “general guidelines” on the interpretation of FHR (FIGO, 1987). These guidelines provide a standard definition of clinical features such as baseline, accelerations and decelerations, from which FHR conditions are categorised into three classes: normal, suspicious and pathological. The FIGO guidelines were the first internationally accepted guidelines for definitions, techniques and interpretation for FHR. Over 20 years later, it is still the only world-wide consensus document on FHR monitoring (Ayres-de-Campos and Bernardes, 2010), with which clinicians throughout the world are trained.

Some countries have introduced their own guidelines. In the U.K. most widely accepted are those guidelines published by the Royal College of Obstetricians and Gynaecologists (RCOG, 2001), and the National Institute of Clinical Excellence (NICE) (RCOG, 2007). In the U.S. national guidelines are published by American Congress of Obstetricians and Gynecologists (ACOG, 2009) and the National Institute of Child Health and Human Development (NICHHD, 1997).

A study in 2010 compared the differences between FIGO guidelines and the major national guidelines in the U.S. and the U.K. (Ayres-de-Campos and Bernardes, 2010). It was found that the 3-class classification system proposed by FIGO had been adopted by most of the national-wide guidelines. It was also found that there were many similarities between these guidelines, and that many ideas in the newer documents were inspired by the FIGO guidelines. The study concluded, however, that the current guidelines suffer from ambiguity in objective quantification of features and complexity that could lead to intra- and inter-observer variability. This implies that using computerised analysis could help relieve the problem of the clinical guidelines in ambiguity (by quantifying the FHR features) and complexity (by giving decision support).

The baseline is one of the most commonly used features indicating the basic level of FHR. In FIGO it is defined as the “mean level of the fetal heart rate when this is stable, accelerations and decelerations being absent, determined over a time period of 5 or 10 minutes” (FIGO, 1987). Among all guidelines, baseline is a critical indicator of fetal health condition, since it reflects the level of cardiac activity. In addition, the assigning of a baseline is also the basis

for detecting accelerations and decelerations, thus an accurate estimation of baseline is very important for FHR analysis.

Another frequently studied feature is the sinusoidal patterns in FHR, defined in FIGO as regular cyclic changes in the fetal heart rate baseline, such as a sine wave (ACOG, 2009). The feature measures the rhythm of FHR and it is found that it is associated with fetal labour compromise (Freeman et al., 2003). A recent study suggested that even though the feature is rare (0.41 per 1,000 cases), it is highly correlated with fetal anaemia (Reddy et al., 2009).

There have been approaches to simulate the clinical guidelines. In this way, new features are generated based on clinical guidelines. For instance, a feature that simulates ACOG guidelines was developed in a previous study (Georgieva et al., 2011b). This is a decision tree that classifies features into normal or adverse conditions according to the clinical guidelines of ACOG, i.e. providing in real time a three-tier grading system of the FHR decision support (normal, indeterminate, or abnormal). It was found in this study that the simulated guidelines corresponded well to the actual labour outcome and clinical management of labour (measured by arterial pH).

The above mentioned features are categorised into “general features” since they were proposed by general clinical guidelines such as FIGO and ACOG. Of course, these clinical guidelines also include other features such as contraction, acceleration and deceleration. However, since these features represent other specific aspects, they are categorised into different groups below.

2.1.2 Contraction features

Contraction features are extracted from the uterine contraction signal of the CTG. These features reflect the activity level of uterine contraction, as drastic uterine contractions can sometimes lead to birth asphyxia. For example, the number of contractions reflects the frequency of contraction, and the median of contraction durations measures the length of contraction. An important conception in contraction is the resting time. Resting time is defined as the resting period where there is no contraction. The median of the resting times and the resting/contraction time ratio are also measured. These are all important indicators of contraction activity. An illustration of contraction features is shown in Figure 2.1.

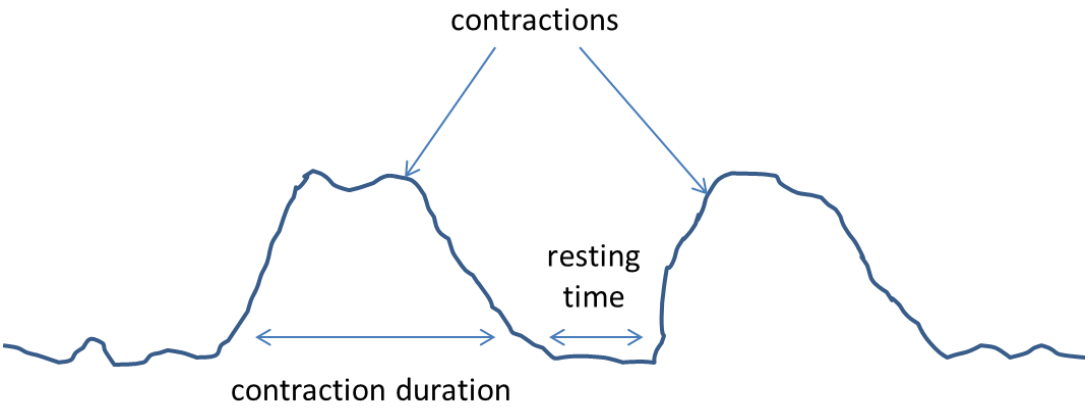


Figure 2.1 An illustration of contraction features.

According to ACOG guidelines, the influence of the contractions varies depending on the conditions of the fetus, i.e. a fetus with good oxygen supply can cope with drastic contractions better than a fetus with low oxygen supply, which can be asphyxiated by contractions at the same intensity level (ACOG, 2009). However, FIGO did give a definition of pathological uterine activity: more than 5 contractions per 10 minutes.

In 2009, a novel technique was used for automated detection of uterine contractions (Georgieva et al., 2009). Hybrids of deterministic approaches and adaptive heuristic knowledge ("meta-heuristics") were created. Parametric approaches assume that there is a certain underlying stochastic process which can be described using a small number of parameters, while non-parametric approaches explicitly estimate the covariance or the spectrum of the process without assuming that the process has any particular structure. The method used in this study is capable of assigning different thresholds to detect contractions adaptively, based on experts' knowledge of the problem, thus it is a non-parametric method. During testing, this achieved sensitivity of 87% and specificity of 75%, which is comparable to the opinion of a second clinical expert (kappa = 0.74). A true positive is defined as identifying a positive case (a contraction in this case, a fetal acidemia case in labour outcome classification) as positive, while a true negative is defined as identifying a negative case (no contraction in this case, no fetal acidemia in labour outcome classification) as negative. A false positive is defined as identifying a negative case as positive, while a false negative is defined as identifying a positive case as negative.

2.1.3 Acceleration & deceleration features

Accelerations and decelerations are the two most distinctive indicators of fetal health, with various analyses and interpretations existing in the literature (Westgate, 2009). Acceleration & deceleration features include the number, duration and amplitude. An acceleration/deceleration is defined in major guidelines as a transient increase/decrease in heart rate of 15 beats/minute or more, lasting a certain period (FIGO, 1987, ACOG, 2009).

There are other acceleration and deceleration features that are defined after FIGO. The mean time to decelerate measures the time to deceleration (from baseline to the bottom of deceleration), while the mean time to recover and the maximum time to recover measure the time recovering (from the bottom of deceleration to baseline) out of deceleration (ACOG, 2009). Quick recoveries are defined as being when the time to recovery is less than the time to decelerate, while slow recoveries are the opposite. Lost beats are defined as deviations from baseline owing to decelerations. An illustration of acceleration/deceleration features is shown in Figure 2.2

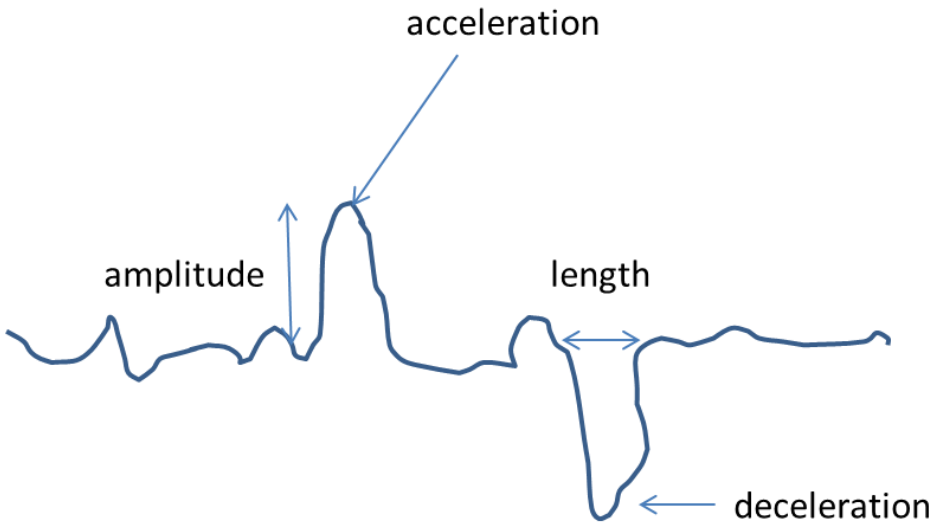


Figure 2.2 An illustration of acceleration/deceleration features.

Absence of decelerations and periodically repeated decelerations are regarded as intrapartum pathological patterns (ACOG, 2009). Later studies have found that accelerations represent active states of fetal sleep, while decelerations reflect a kind of fetal response that is generally believed to help reduce the burden of myocardial and oxygen requirements (Westgate et al., 2007).

In Baan et al. (1993), 22 fetal sheep were studied and acute hypoxemia was induced by occluding a balloon cuff around the common hypogastria artery and 151 occlusions were performed. Multiple linear regression analysis showed a ratio (fall of heart rate/fall of saturation) of averaging 2.5 $\pm$ 1.2 (bpm / % of saturation). It was concluded that the amplitude and length of decelerations are associated with the severity of hypoxia, i.e. a deep deceleration indicates that there could be a severe reduction of oxygen supply.

2.1.4 Lag features

Lag features measure the "early" and "late" decelerations and the "lag" time with respect to uterine contractions in the case of "late" decelerations (RCOG, 2001). These features describe the relationship between uterine contractions and decelerations, indicating fetal responses to contractions. "Early" and "late" are defined by the time between a contraction and the responded deceleration.

These features include the number of early and late decelerations, the number of variable decelerations, the recovery time for late decelerations, the mean of lag time, and the time ratio of deceleration and contraction. These features describe the relationship between FHR and uterus contractions. For example, a late deceleration is defined as the nadir of the deceleration occurring after the peak of the contraction, and a variable deceleration is defined as a visually apparent abrupt decrease of FHR (ACOG, 2009). An illustration of lag features can be found in Figure 2.3. Late decelerations are associated with uteroplacental insufficiency (Chandraharan and Arulkumaran, 2007).

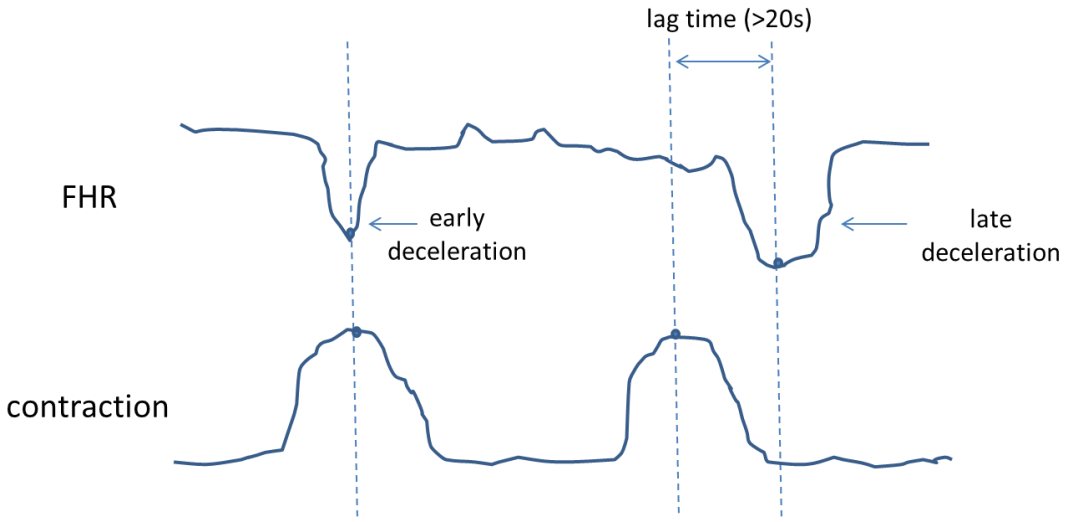


Figure 2.3 An illustration of lag features.

2.1.5 Variability features

It has been recognised that FHR variability is a critical indicator associated with labour outcome (Westgate, 2009). Variability features include the percentage of the points that have zero difference with their neighbouring points, and the residual signal (the difference between the FHR and the smoothed FHR signal). Introduced below are the variability features that have been found to be associated with fetal hypoxia.

The most widely used variability feature is Short Term Variability (STV), which is defined as being the difference of all neighbouring FHR values (FIGO, 1987). A healthy fetus should have adequate variability due to the complexity of the physiological system. A fetus with a FHR trace of low or decreasing STV would be likely to suffer from hypoxia (Barry, 2004). On the other hand, long term variability (LTV) is defined as being oscillations of fetal heart rate around its mean level in the long term (FIGO, 1987). It is often studied with STV and is also recognised as having an association with fetal hypoxia (Pardey et al., 2002).

The approximate entropy is a measurement of signal regularity, which offers another aspect of the variability i.e. the predictability of a signal. In (Dawes et al., 1992), it was found that although statistically approximate entropy offers no significant advantage over measurement of STV, it is still a prospective measurement providing a different aspect of variability.

The Signal Stability Index (SSI) is a measurement of signal stability based on the kernel density estimate (KDE) (Georgieva et al., 2011a). In a FHR window with very unstable baseline, the KDE is relatively flat and has a low peak. It has been used to detect the "baseline un-assignable" (BU) segments; further details of which can be found in Section 2.3.1.

In 2013, a study used a relatively new time series analysis method - Phase rectified signal averaging (PRSA), originally proposed in (Bauer et al., 2006) – to extract new features that measure the variability of FHR (Georgieva et al., 2013b). This new method quantifies separately the variability that occurs during increasing and decreasing parts of the signal, i.e. acceleration capacity (AC) and deceleration capacity (DC). Results of this study have shown that the PRSA feature is competitive with or outperforms previous fetal heart analysis measures such as short term variability, approximate entropy and signal stability index.

Bivariate Phase Rectified Signal Averaging (BPRSA) (Schumann et al., 2008) is an extension of the PRSA method, which can take the interaction of FHR and uterine contraction into consideration. The DC component of the PRSA method was found to be related to fetal acidemia, while the DC and AC components of the BPRSA method were found to be associated with fetal acidemia (Williams, 2012).

2.1.6 Other time series features

A time series is a sequence of data points measured at successive points in time spaced at uniform time intervals. Both the FHR signal and the uterus contraction signal from the CTG are time series signals. Time series analysis methods include time domain analysis and frequency domain analysis. Time domain analysis methods analyse information from the time domain (such as auto-correlation analysis); while frequency domain analysis methods analyse information from the frequency domain (such as spectrum analysis and wavelet transform). The variability features mentioned before are all extracted from time series analysis, thus other features extracted from time series analysis are categorised into "other time series features".

Time-series analysis methods have been used for FHR analysis in recent years, such as spectral analysis (Van Laar et al., 2008), wavelet filtering (Salamalekis et al., 2006), non-linear methods (Spilka et al., 2011) and multi-scale complexity (Helgason et al., 2011), all of whichhave also been applied to extract features from the FHR signal. Preliminary results have shown that these features are promising in the prediction of labour outcome (Salamalekis et al., 2006). The major problem of utilising these features for classification is that the size of the databases used in these studies is not large enough to give reliable classification results (smaller than 200 cases).

In a recent study, over 9 000 time-series features were extracted from each FHR time series (> 7000 deliveries), including measures of autocorrelation, entropy, distribution, and various model fits (Fulcher, 2012). 60 normal FHR windows (arterial pH > 7.2) and 59 abnormal

FHR windows (arterial pH < 7.2) were used as a training set to extract the time series features and to select those features which have a strong linear correlation with arterial pH or have a high classification performance using the linear discriminant classifier. 5 features with a strong correlation with arterial pH and 4 features with high classification performances were selected. These features include auto-mutual information, local approximate entropy and time delay embedding space. These features were then tested on a testing set of 7,221 cases and classification performances evaluated using a linear discriminant classifier. It was shown that these features have 71-74% proportion of agreement on a training set and 62%-69% on a testing set.

With regard to feature extraction, this study used methods from a broad and diverse literature on time-series analysis methods, and successfully identified and selected some of the most promising features that have high correlation with arterial pH. With regard to feature selection however, this study used only linear metrics to select FHR features, and selected each feature based only on its performance, this could result in ignoring the performance of features when used together.

2.2 Overview of Computerised FHR Analysis Systems

2.2.1 The need for computerised FHR analysis

In the late 1980s it was found in several studies that human error can be an important contributor to some of the failures of using CTG. In a case-control study (Murphy et al., 1990), it was found that 87% of babies suffering from birth asphyxia had an abnormal CTG,

but that the mean time for the clinicians to recognise the CTG abnormality was 91 to 128 minutes, which is too long to carry out an intervention. In 1994, a study of 141 cerebral palsy cases found that clinicians' failure to respond to clear signs of abnormality in the CTG occurred in 26% of cerebral palsy cases and 50% of perinatal deaths. These results suggest that if clinical interpretation and prediction of the CTG can be more timely and accurate, the benefits of CTG monitoring can be fully explored.

Therefore, to improve decision making using CTG and to minimise the damage of birth asphyxia, various approaches have been applied to build computerised FHR analysis systems since the 1980s. According to NICE guidelines for fetal monitoring, under 60% of labour conditions women should have CTG monitoring during labour (RCOG, 2007). Westgate has reviewed computerised FHR analysis and the need for further development (Westgate, 2009).

2.2.2 Antepartum systems

There are two types of CTG: antepartum CTG (before labour) and intrapartum CTG (during labour). Antepartum CTG is used in pre-labour examinations and, it usually stops when there is a steady FHR signal that shows no abnormality. Intrapartum CTG is more complex and more difficult to analyse since there are more physiological activities such as uterine contractions, which can result in more complicated FHR patterns, signal loss, artefacts and noise. On the other hand, intrapartum CTG contains important information within hours of delivery, which has been found to be associated with birth asphyxia (Murray et al., 2009).

The first commercialised antepartum FHR system was the Sonicaid System 8000 (Dawes et al., 1991). An on-line system was developed to measure the antepartum FHR in real time,

extracting and analysing five key FHR features including baseline, LTV and STV. Clinical validation and update of the system continued over the subsequent ten years, and the updated system (Sonicaid FetalCare System) and validation results (> 73,000 patients) were reported in 2002 (Pardey et al., 2002).

A primary feature of the Sonicaid FetalCare System is that it can alert the clinician when the fetal heart rate is confirmed to be normal, so that the antepartum monitoring time can be minimised. It can also detect a sinusoidal pattern and alert the clinician (Pardey et al., 2002). A study comparing the opinion of clinicians and the Sonicaid FetalCare System showed that women assessed with the system spent less time being monitored (Bracero et al., 2000). Other computerised antepartum FHR systems have been built but none have matched the large database of the Sonicaid systems (Westgate, 2009).

The FHR features extracted in the antepartum systems can also be used in intrapartum systems; but different classification criteria should be used for the intrapartum systems. This thesis is focused on intrapartum CTG analysis, thus all the CTG analysis stated below, if not specifically noted, refers to intrapartum CTG analysis.

2.2.3 Intrapartum systems

The simplest method for computerised CTG analysis is to digitise and display the CTG at a central monitoring site, which has already been offered by a number of commercial companies. However, studies have shown that CTG interpretation by eye is error-prone and associated with very low reproducibility owing to notably high inter- and intra-observer variability (Chauhan et al., 2008).

A higher level of computerised FHR analysis is to use FHR features to set alarms for abnormal ranges of parameters (baseline, variability, etc.). A typical system that extracts the FHR features and displays them on the screen for decision support is described in the literature (Jezewski et al., 2006). The system can automatically detect the FHR baseline, accelerations, decelerations and FHR variability and display them on the screen. Different parameters can be set so that the system will give alerts when some of the features have extreme values. This kind of system has reduced the complexity of visual evaluation of FHR patterns. The problem is that these extracted features still require visual assessment, which is error-prone and tedious. Also, as the FHR features are related to each other, it is difficult to make decisions based only on alerts of a single feature.

In addition to extraction and analysis of individual FHR features, the ability to integrate the information from various features and to provide accurate information is crucial in computerised FHR analysis. Computerised decision support systems should be able to assist clinicians in complex decision making. Typically they consist of a medical knowledge base, patient specific data, and an inference engine which applies the knowledge to the patient data to generate patient specific advice (Westgate, 2009).

Important developments have been made in FHR analysis methods, by using both clinical knowledge and statistical methods for decision support (Costa et al., 2009, Ayres-de-Campos and Bernardes, 2004, Devoe et al., 2000, Westgate, 2009). Different classifiers such as Artificial Neural Networks (Georgieva et al., 2013c), fuzzy inference and Lagrangian Support Vector Machines (LSVM) (Czabanski et al., 2012) have been used for the task of labour

outcome classification. Reasonable classification performances have confirmed the association between FHR patterns and labour outcome.

Integrated computerised decision support systems for intrapartum FHR analysis have been under development since the 1990s (Keith et al., 1995, Bernardes et al., 1998). An early study suggested that such a system can be potentially useful for decision support (Keith et al., 1995). This study used feature extraction methods to extract clinical features and an expert-knowledge based system to classify labour outcome. It demonstrated that their off-line computerised decision support system using CTG was highly consistent and comparable to the performance of clinical experts. This system was later named as the K2 system.

In practice, only one clinical trial has been carried out to validate the effectiveness of computerised systems. The INFANT study is a controlled clinical trial to validate whether computerised decision support systems can reduce the number of adverse labour outcomes, run in many hospitals in the U.K. The first published study of INFANT (Barber et al., 2013) investigated the effects of the technologies being used on the anxiety levels of those women, in which 469 women (234 CTG monitored only, 235 CTG monitored with decision support system) were asked to measure their anxiety levels. The study suggested that the decision-support system did not significantly (two-sided t-test, $p > 0.05$) increase overall anxiety during labour. The system being used in the study is the INFANT software produced by K2 Medical Systems. The study mentioned that baseline, variability and contractions are integrated, and that the system provides a four colour-coded “ladder of concern” for decision support.

Another recently-developed system is the Omniview-SISporto 3.5 (Ayres-de-Campos et al., 2008). The system uses a decision tree based on three-tier online alerts: attention, non-reassuring and preterminal, based on the values of FHR features (baseline, acceleration, deceleration, etc.) and fetal ECG features (ST event). A primary feature of the system is that it uses both the CTG analysis and fetal ECG (fECG) features of the latest fECG analysis system - STAN (Amer-Wåhlin and Maršál, 2011). The on-line system combines CTG features with fetal ECG features to provide visual and auditory alerts, based on features of both CTG (baseline, STV, LTV, etc.) and fetal ECG (ST event). The system uses decision trees to integrate the features and to produce alarms, which is simple and instinctive but which lacks the combination of different features and the interpretation of these alerts. This system is currently undergoing clinical evaluation. In Costa et al. (2010), 50 consecutively acquired CTG tracings were analysed by three clinicians and the system. For several FHR features such as decelerations, accelerations and contractions, the proportion of concordant identification is between 68%~87%.

While the INFANT and SISporto systems are still in clinical trials, the PeriCALM system developed in the U.S. has already been released as a commercial product. The methods and the training database of the commercialised system have not been specified, but in the report of the clinical trial (Parer and Hamilton, 2010), it was reported that 134 combinations of FHR features were used to form the classification system. In the study, a database of 769 8-minute FHR segments was used to compare the classification of the system and clinical experts, and results have shown that the system is comparable with clinical experts (kappa value averaged

0.58 with 95% confidence interval of 0.48–0.68). However, the study suffered from lack of labour outcome information, which means that the accuracy of the system is still unknown and the size of the database (769) is small since fetal acidemia is very rare (about 3% in high risk monitored population) (Parer and Hamilton, 2010),.

A comparison of these intrapartum FHR analysis systems is shown in Table 2.1.

Table 2.1 A comparison of the computerised intrapartum FHR analysis systems.

System	***Methods***	***Critical FHR features***	***Decision support***	***Application***
INFANT (K2)	Decision tree that mimic the expert's opinions	Baseline, decelerations, variability and contractions	Four colour-coded "ladder of concern"	Clinical trial
PeriCALM	134 combination of CTG features (not specified)	Baseline, baseline variability, accelerations and decelerations	Five-level classification	Commercialised
Omniview-SISporto 3.5 (SISporto + STAN)	Decision tree using Fetal ECG + CTG features	CTG (baseline, STV) and fECG features (ST event)	Visual and sound alerts	Clinical trial

Recent studies of computerised FHR analysis have focused on two directions: extraction of new features (Spilka et al., 2011, Van Laar et al., 2008) and the use of FHR features for a labour outcome classification (Georgoulas et al., 2007b, Chudáček et al., 2011, Warrick and Kearney, 2009). Similarly, the Oxford Centre for Fetal Monitoring Technologies group is still extracting new FHR features (Fulcher et al., 2012, Georgieva et al., 2013b) for the OxSys

system, while using the existing FHR features for labour outcome classification (Georgieva et al., 2013c). However, feature selection methods have not been used before to select feature subsets and integrate features, and this will be discussed in Section 2.4.

The major problem with recent reports on computerised fetal heart rate analysis is that they have datasets of small sizes (typically less than 500 FHR cases) (Georgoulas et al., 2007b, Chudáček et al., 2011, Warrick and Kearney, 2009). As the proportion of fetal acidemia is very small (usually < 5%), these datasets often contain too few adverse cases. It is extremely difficult to draw reliable conclusions using such datasets.

Another problem of labour outcome classification is clinical interpretation of the classifiers. Some studies on labour outcome classification have used multiple FHR features, but it is hard to find the importance of the features and the interpretation of classifiers (Czabanski et al., 2012). Therefore, feature selection methods need to be applied to the various features to give a ranking of the various features, as well as to give a clear feature subset for labour outcome classification.

2.3 Feature Extraction Methods

Transforming the input data into a set of features is called feature extraction (Jain et al., 2000). The performance of extracted clinical features should be evaluated through comparison with the opinions of clinical experts. On the other hand, extraction of statistical features has focused on finding those that have a high correlation with adverse labour outcome.

A comprehensive description of the general FHR feature extraction methods can be found in

Cazares (2002). Several general FHR features are extracted, such as baseline, decelerations and short-term variability. A novel open-close-smooth algorithm was used to extract the baseline. The algorithm was validated using a reference set of 12 CTG traces (6 healthy and 6 abnormal) and the baseline estimation of the algorithm was compared with those of two clinical experts. The inter-observer variance from the two experts is 0.9 bpm RMSE (root-mean-square deviation), while the variances between the algorithm and the experts are 3.9 and 5.7 bpm RMSE. Considering that the normal range of FHR is 110 to 160 bpm, this is a reasonable accurate estimation.

As for the extraction of discrete FHR features in Cazares (2002), the binary identification performed by the algorithm was compared with expert identification. For the number of decelerations (extracted based on the baseline estimation and thresholding), values of sensitivity and positive predictive value (PPV) were calculated. It was found that high values of sensitivity (above 80%) and PPV (above 60%) were achieved when compared with the identification of two clinical experts.

These feature extraction algorithms used for intrapartum FHR were then compared with the performance of the antepartum Sonicaid System 8000 (Dawes et al., 1991). It was shown that the feature extraction algorithms, when compared with clinical experts (mentioned above) were more accurate than System 8000. This study provided the feature extraction methods for general FHR features for the Oxford database. However, how to further integrate the information of these features and how to provide decision support were not investigated, and the size of the datasets used to train the features (less than 50) was rather small.

This section will focus on the review of feature extraction methods that will be used in this thesis: the pattern readjustment feature (Chapter 3), the Support Vector Machine (Chapters 3 and 4) and signal loss (Chapter 5).

2.3.1 Pattern readjustment and step detection methods

In 2011, *Georgieva et al.* developed an algorithm to detect "baseline un-assignable" (BU) segments in the FHR (Georgieva et al., 2011a). The algorithm was then validated by comparison with the opinions of clinical experts. The average agreement between the experts (kappa = 0.76) was comparable to the agreement between the method and the experts (kappa = 0.67). In a later study (Georgieva et al., 2012), 7,568 CTG traces were studied, and a consistent increase of the risk for acidemia was found with longer intervals of BU: in the last 30 min of labour, the odds ratios (with respect to baseline assignable throughout this period) increased from 1.99 (15 min un-assignable) to 4.9 (30 min un-assignable).

Based on the observation of BU, an FHR pattern that can cause un-assignable baseline was observed. The feature was named "pattern readjustment", since it reflects the readjustment of FHR from one baseline to another. As pattern readjustment reflects instantaneous jumps in the mean of a FHR signal, step detection methods were found to be suitable for feature extraction of pattern readjustment.

Step detection (or edge detection) is the process of finding abrupt changes (steps, jumps, shifts) in the mean level of signal segments (Marr and Hildreth, 1980). Step detection methods are widely applied and in multiple scientific and engineering contexts. Current step detection methods consist of *on-line* detection and *off-line* detection. If the step detection

must be performed while the signal arrives, *on-line* algorithms are usually used. On the other hand, *off-line* algorithms are applied retrospectively after signal acquisition. In the OxSys database, the FHR signal segment is analysed in each 15-minute signal window, and is therefore an *off-line* detection method.

Off-line step detection methods can be categorised into filtering methods, variational approaches, stochastic models and statistical analysis methods. Table 2.2 gives a brief comparison between the different methods. Full details of the methods can be found in the references listed in Table 2.2.

In Filtering methods, step detection is viewed as the problem of recovering a piecewise constant signal corrupted by noise, using various filters such as wavelet filters (Mallat and Hwang, 1992, Sadler and Swami, 1999). Variational approaches such as level sets (Little and Jones, 2010) aim to exploit information provided by signal variation (gradient differences) for the step detection task. Stochastic modelling methods attempt to exploit the differences in distribution between different steps using stochastic models such as Markov chains (McKinney et al., 2006). On the other hand, statistical analysis methods such as clustering or the t-test are used to minimise the itaratively cost of different classifications.

Table 2.2 A comparison of step detection methods

Methods	***Examples***	***Advantages***	***Drawbacks***
Filtering methods	Wavelet filtering (Mallat and Hwang, 1992, Sadler and Swami, 1999) Bilateral filtering (Mrazek et al., 2006)	Competitively stronger in dealing with noise. Relatively simple implementation.	Could result in loss of the underlying step owing to the overlapping of signal and noise in bandwidth.

Variational approaches	Level set (Little and Jones, 2010) Total variation de-noising (Rudin et al., 1992)	Take advantage of the sharp change detected in the signal. Relatively efficient.	Sensitive to noise. Stability and convergence dependent on parameters.
Stochastic models	Hidden Markov Models (HMM) (McKinney et al., 2006);	Relatively stronger in dealing with noise. Statistically meaningful.	Easily affected by initialisation. Easily convergence to local minima. Computational complexity.
Statistical analysis methods	K-means clustering (Macqueen, 1967); Chi-squared minimisation (Kerssemakers et al., 2006); Student's t-test (Carter and Cross, 2005)	Reveal statistical differences between different steps.	Computational complexity. Time consuming.

A systematic review of step detection and an illustration of some methods is also given in Basseville and Nikiforov (1993). The performances of different methods have been compared on artificial benchmark data (Carter et al., 2008); the authors found that many of the tested methods have similar performance when optimised, but the method based on a chi-squared optimisation procedure is the simplest.

In conclusion, step detection methods have been comprehensively developed and widely validated. It would be sensible to start with methods that are easier to implement, such as filtering methods, the fundamental examples of which are the first-order derivative filter and second order derivative filter with Gaussian smoothing.

2.3.2 The support vector machine

The support vector machine (SVM) is widely used in data analysis owing to its intuitive definition and simple implementation (Cristianini, 2000). Therefore, the linear SVM is used

to classify the pattern readjustment samples from the control samples in Chapter 3 and to classify the normal and adverse outcomes in Chapter 4.

The SVM constructs a hyperplane or a set of hyperplanes to separate the data points, in order to achieve the largest distance to the nearest training points of any class (the functional margin) (Figure 2.4). Intuitively, a larger functional margin means lower generalisation error (error when applied to new datasets) of the classifier. The principle of the linear SVM can be briefly described as follows.

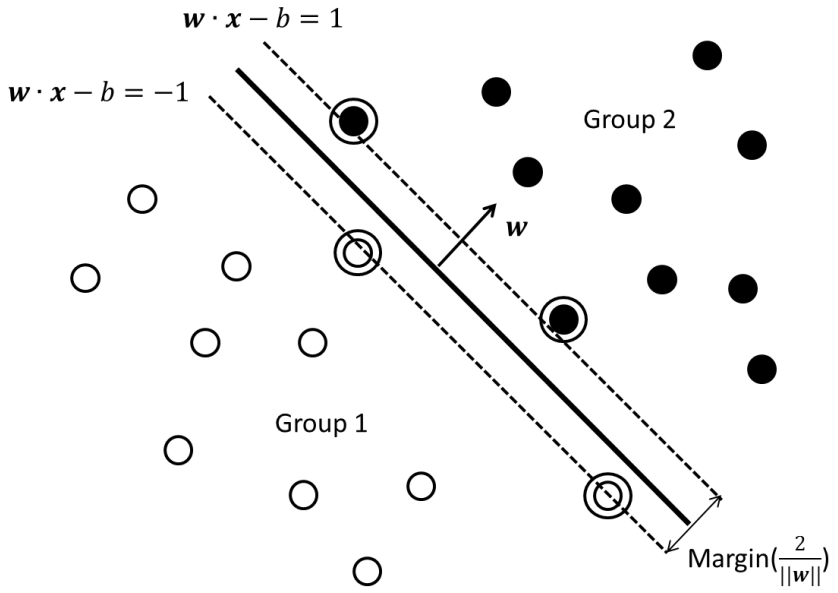


Figure 2.4 Illustration of the separation hyperplane and the support vectors (circled), adapted from Cortes and Vapnik (1995).

For a linear SVM classifier, given a set of n training data points of the form:

$$D = \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \epsilon \boldsymbol{R}^p, y_i \epsilon \{-1,1\}\}_{i=1}^{n},$$

where D is the set of data, y_i indicates the class to which the point belongs, and each $\boldsymbol{x}_i$ is a

real vector, and where the problem is to find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. Samples on the margin are called the support vectors.

The classification hyperplane can be defined as the set of points $\boldsymbol{x}$ satisfying:

$$\boldsymbol{w} \cdot \boldsymbol{x} - b = 0,$$

where $\cdot$ refers to the dot product and $\boldsymbol{w}$ is the normalised vector of the hyperplane. Therefore, the optimisation problem is to choose $\boldsymbol{w}$ and b in order to maximize the distance between the parallel hyperplanes so that they are as far apart as possible while still separating the data. These hyperplanes can be described by the equations:

$$\boldsymbol{w} \cdot \boldsymbol{x} - b = 1,$$

and

$$\boldsymbol{w} \cdot \boldsymbol{x} - b = -1,$$

If the training data are linearly separable, the two hyperplanes of the margin can be selected such that there are no points between them, and then the distance between them is thereby maximized. By using geometry, it is found that the distance between these two hyperplanes is $\frac{2}{||\boldsymbol{w}||^2}$, thus the optimisation problem is to choose $\boldsymbol{w}$ and b subject to:

For each $\boldsymbol{x_i}$, either

$$\boldsymbol{w} \cdot \boldsymbol{x_i} - b \geq 1 \text{ , for } \boldsymbol{x_i} \text{ of the first class,}$$

or

$\boldsymbol{w} \cdot \boldsymbol{x_i} - b \leq 1$, for $\boldsymbol{x_i}$ of the second class.

This can be rewritten as:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} - b) \geq 1$$

This is illustrated in Figure 2.4.

Therefore, the optimisation problem can be described as:

$$arg\ min_{\boldsymbol{w},b}\ \frac{||\boldsymbol{w}||^2}{2},\ \text{subject to}\ y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} - b) \geq 1$$

By introducing Lagrange multipliers $\boldsymbol{\alpha}$, it can be described as:

$$arg\ \min_{\boldsymbol{w},b} max_{\boldsymbol{\alpha}>0}\ \{\frac{||\boldsymbol{w}||^2}{2} - \sum_{i=1}^{n} a_i[y_i(\boldsymbol{w} \cdot \boldsymbol{x_i} - b) - 1]\}$$

This problem can be solved by quadratic programming:

$$\boldsymbol{w} = \sum_{i=1}^{n} a_i y_i\, \boldsymbol{x_i}$$

Thus the objective function can be described as:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} a_i\ - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i} \cdot \boldsymbol{x_j}$$

The original hyperplane in the SVM algorithm was linear. A way to create nonlinear classifiers by applying the kernel transform methods to maximum-margin hyperplanes was suggested (Cortes and Vapnik, 1995). The non-linear SVM is similar to the linear SVM, except that every dot product is replaced by a non-linear kernel function. In this way, the original space can be transformed into a higher dimensional space that is linearly separable.

For example, for linear SVM, the kernel

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{x_i} \cdot \boldsymbol{x_j}$$

the objective function is:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} a_i \; - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \; \sum_{i=1}^{n} a_i \; - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i} \cdot \boldsymbol{x_j}$$

while the RBF (Radial basis function) SVM used in this thesis uses the kernel

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = exp(-\gamma \parallel \boldsymbol{x_i} - \boldsymbol{x_j} \parallel^2)$$

where gamma is a parameter that decides the sensitivity of the classifier.

thus the objective function is:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} a_i \; - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \; exp(-\gamma \parallel \boldsymbol{x_i} - \boldsymbol{x_j} \parallel^2)$$

This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may be non-linear, thus the classifier can be nonlinear in the original input space. The major advantage of non-linear SVM is that the use of kernels allows its optimisation problem to be solved by standard quadratic programming methods, thus it could provide low computational complexity even in solving non-linear problems. In addition, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimising the margin are all key features of SVM (Shawe-Taylor and Cristianini, 2004).

2.3.3 The issue of signal loss

An important issue of FHR feature extraction is the signal loss. During the recording of FHR, signal loss occurs from time to time in over 90% of the FHR windows (see Section 5.1). To

the author's knowledge, the influence of signal loss has never been investigated before in other FHR studies, except for one study (Spencer et al., 1987), where two groups of 10 patients and 13 patients respectively with different FHR signal quality were tested. Fetal ECG was used as the gold standard. Statistical tests showed that for the group with higher FHR signal loss, the mean FHR variation measured by ultrasound was significantly lower than that measured by fetal ECG; while for the group with lower signal loss, the mean FHR variation measured by ultrasound had no significant statistical difference against that measured by fetal ECG. This study measured only one FHR feature with a small number of patients, so the information is insufficient to quantify the influence of signal loss.

Except for the above mentioned study, there is no other study that reports how signal loss should affect FHR feature values and FHR labour outcome classification. Therefore, it is still necessary to investigate the influence of signal loss on both the FHR features and on the labour outcome classification.

2.4 Feature Selection Methods

2.4.1 The need for FHR feature selection

In a recent study, a number of FHR features were investigated and assessed based on expert annotation (Chudáček, 2011). Although the study has its limitations due to the intra- and inter-observer variability of the experts and the size of the database, preliminary results have shown that the combination of the FIGO features and non-linear features can result in better classification performance than using the features independently. This work, along with other studies (Georgieva et al., 2013c), has shown the potential of selection and integration of these

features to provide a better predictor than using individual features.

On the other hand, the number of all possible combinations of 64 features (which is the number of features in the Oxford database) is $2^{64} = 1.8\times10^{19}$, which is too large for an exhaustive search. Therefore, feature selection methods need be applied. Georgoulas et al. (2007a) uses grammatical evolution for FHR feature selection, but the study suffers from the limited size of data with only 160 CTG traces, only 2 of which suffer from fetal acidemia. Therefore, although the classification accuracy (proportion of agreement) is as high as 90%, the limited size of data has made the significance of the results questionable.

2.4.2 Feature selection methods

The problem of feature selection is defined as follows: given a set of features, select a subset of features that gives the smallest classification error (Jain et al., 2000). Feature selection methods have been widely applied in research areas where datasets with tens or hundreds of thousands of variables are available, in order to improve the prediction performance of the predictors (Corralejo, 2011).

Various approaches such as Genetic Algorithms, forward selection and backward elimination have been applied to solve this problem. A detailed review and comparison of different methods is provided by Guyon and Elisseeff (2003). According to this study, feature selection methods can be divided into three categories, depending on how they combine the feature selection search with the construction of the classifier: wrappers, filters and embedded methods (Table 2.3).

Filters, such as Markov blanket filters (Kohavi and John, 1997), use subset selection as a

pre-processing step, regardless of the classifier. In most cases, a feature relevance score is calculated, and low-scoring features being removed by the filters. Wrappers, such as Genetic Algorithms (Mitchell, 2001), use one classifier as a "black box" to evaluate different subsets of variables. Different classifiers are "wrapped" using the feature selection method and feature subsets are evaluated by their classification performances using this classifier. Embedded methods, such as Random Forests (Breiman, 2001), apply variable selection in the process of classifier training. The search for an optimal subset of features is built into the classifier construction, thus they are usually specific to given learning machines. A comparison of these three methods is shown in Table 2.3.

Table 2.3 Comparison of major feature selection methods

Methods	*Examples*	*Advantages*	*Drawbacks*
Filter	Markov blanket filter (MBF) (Kohavi and John, 1997); Fast correlation-based feature selection (FCBF) (Yu and Liu, 2004)	Simple and fast; independent of the classifier; less computational complexity than wrapper methods;	Ignoring feature dependencies; no interaction with the classifier
Wrapper	Genetic Algorithms (Mitchell, 2001); estimation of distribution algorithms (Inza et al., 2000)	Interaction with the classifier; considering feature dependencies	Classifier dependent; risk of over fitting; intensive computational cost
Embedded	Weighted Na ïve Bayes (Eibe et al., 2003); Weight vector of SVM (Guyon et al., 2002); Random Forests (RF) (Breiman, 2001)	Interaction with the classifier; less computational complexity than wrapper methods; considering feature dependencies	Classifier dependent; designing complexity

Of the three approaches, filter methods ignore dependencies between features, which are

critical for FHR features, which can have high correlation with each other. Embedded methods such as the Random Forests (RF) tend to integrate the feature selection process into the training process, resulting in a classifier consisting of all features with different weights rather than a clear feature subset, which can be difficult for clinical interpretation. On the other hand, wrappers use a classifier as a black box to locate a specific feature that has the best performance using this classifier. This not only takes feature dependencies into consideration, but also gives a clear feature subset for clinical interpretation. Therefore, wrappers are more suitable for FHR feature selection.

2.4.3 Genetic Algorithms (GA)

Amongst wrapper methods, Genetic Algorithms (GA) are known for their competitive exploration ability. They have been developed and widely applied from the fields of finance to bioinformatics. After feature extraction, GAs can be used to select the “optimal” feature subset, in order to optimise the performance of the labour outcome predictor.

The GA method simulates genetic evolution to look for feature subsets with the best “fitness”. A brief illustration of the GA method is shown in Figure 2.5: a population of strings (the genotype of the genome) that encode candidate FHR feature subsets (individuals or phenotypes, the predictor) evolves towards better solutions. The evolution starts from a randomly generated population (with various feature subsets). In each generation, the fitness of every individual in the population is evaluated, and multiple individuals are stochastically selected from the current population based on their fitness.

As the next step, the selected individuals are modified (recombined and possibly randomly

mutated) to form a new population. The new population is then used in the next iteration of the algorithm. In this way, the performance of each generation becomes better and better, and at the end, only the best individuals (feature subsets) will be left. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population (Mitchell, 2001).

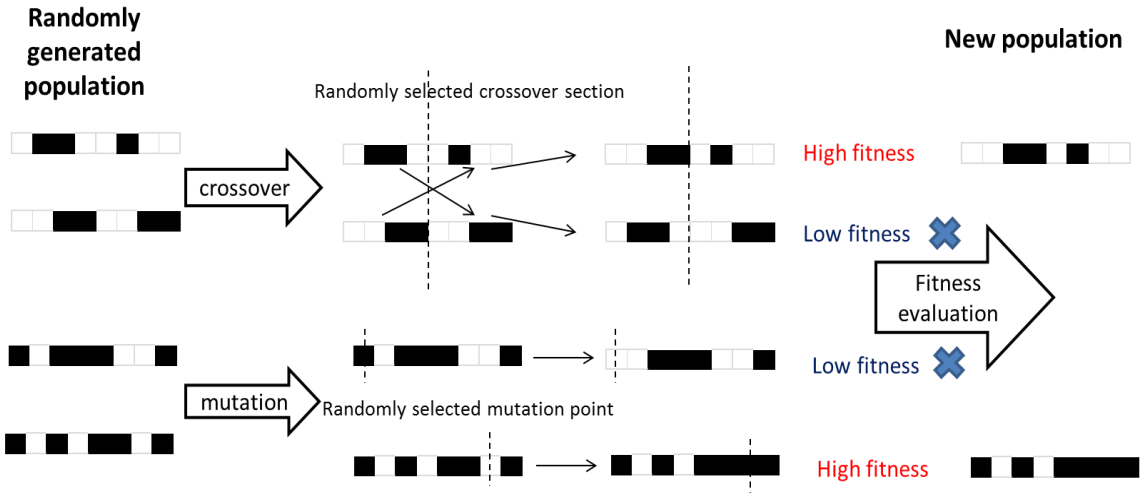


Figure 2.5 A brief illustration of the GA method.

The downside of GAs is that the search has the danger of premature convergence and thereby turning into a local optimiser. When an optimal balance with their exploitation (the ability to accurately locate the feature subset that has the best value) and exploration (the ability to explore the whole feature space to find the global optimum) abilities is found, GAs can be powerful and efficient global optimisers (Mitchell, 2001). Keeping the balance between exploitation and exploration is an important characteristic of any GA technique (Ali et al., 2005). In addition, exploration-dominated search can lead to excessive computational expense, thus adequate optimisation of the algorithm is needed to reduce the execution time of running

such programs.

Similar approaches using evolutionary methodology and Neural Networks have been applied to FHR analysis (Georgoulas et al., 2007b, Georgoulas et al., 2007a). Grammatical evolution, which is similar to a GA, was used to discriminate acidemic from normal fetuses utilising features extracted from the FHR signal. Although with a limited data size (160 FHR recordings with only 2 cases of adverse labour outcome), the study indicated some potential for evolutionary methods for FHR analysis. Therefore, further efforts are needed to solve these problems, using larger and more diverse datasets. To the author's knowledge, feature selection of FHR features has never been investigated before previously.

2.4.4 Greedy search strategies

Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) are two of the most commonly used greedy search strategies. They are very similar algorithms. The basic process of SFS is that it starts with the feature that has the best classification performance, and then it searches for the next feature that can best improve the classification performance. The searching stops when the classification performance is no longer improved by adding features. The basic process of SBS is that it starts with the whole feature set, and then it searches for the feature, assuming there is one, that can best improve the classification performance if removed. The searching stops when the classification performance is no longer better.

Not only is SFS intuitive in terms of searching for the best feature subset by starting with the best feature, it also presents the great advantage of being applicable even when the initial data

set contains more explanatory variables than needed (Blanchet et al., 2008). The major shortage of SFS lies in that it yields nested subsets of variables and thus lacks the ability to explore the full feature space. For example, two features that are useless by themselves can be useful together (Guyon and Elisseeff, 2003), which can be the case for FHR features. In SFS, it is likely that neither of such two features will be selected because they cannot improve the classification performance individually. SBS, on the other hand, is good at keeping the best feature combinations, since it only deletes the features that will reduce the classification performance, but it also tends to overfit since it generally keeps more variables than SFS.

The problems of SFS and SBS can be relieved using global search strategies as GAs, exploring fully the feature space and not missing the important feature combinations that can improve classification performance. On the other hand, wrappers such as GA are often criticised because they are sensitive to tiny changes in the classifier performance, which could result in overfitting. Coarser greedy search strategies, such as SFS and SBS, are computationally advantageous and robust against overfitting (Guyon and Elisseeff, 2003). Therefore, these greedy search strategies can be complementary to the GA, especially in dealing with overfitting issues.

2.5 Data in the OxSys System

The Oxford Centre for Fetal Monitoring Technologies and its predecessors have been studying computerised analysis of FHR signals since the early 1980s. Antepartum CTG signals have been studied and the Oxford Dawes-Redman computerised system has been developed for objective numerical recognition of important abnormal features. It is now

marketed by Huntleigh Healthcare as “Sonicaid FetalCare”. The system has been developed and validated on a large archive of antenatal traces linked to clinical outcome data.

Currently, the group is developing a computerised system (OxSys) to recognise unfavourable intrapartum FHR patterns automatically. Other groups are also developing similar systems using smaller databases and different methods (Costa et al., 2010, Westgate, 2009). Details can be found in Section 2.2.

OxSys is based on the Oxford Intrapartum FHR Database (OIFD), collected from 1993 to 2008. It comprises 127,579 digital records of FHR patterns stored non-selectively by a central archiving system (AXIS, Huntleigh Diagnostic Products, Cardiff, UK) (Georgieva and et al., 2011). The majority of the intrapartum records made during this period in the Maternity Unit of the John Radcliffe hospital in Oxford are in this archive. Two subsets of OIFD are used for different purposes (Georgieva et al., 2011a). Details are given below:

(1) OIFD1 (49,151 records): all traces were selected if they were taken during labour from a singleton pregnancy and can be linked to the clinical record of the Oxford Maternity Database (OXMAT). OxSys is being developed with this particular dataset. The data were anonymised before preparation of the final database.

(2) OIFD3 (7,568 records) is a subset of OIFD1 selected for detailed recording of both FHR recording and birth outcome. The total set of 7,568 records was pre-processed by an OxSys default routine, based on the Dawes heuristic method (Dawes et al., 1990), which identifies artefacts as abrupt signal shifts up or down followed by an equally abrupt return to baseline.

The process of labour is divided into three stages according to different physiological activities. The first stage (Stage 1) is identified as frequent regular contractions with less than 10 cm dilation of the cervix. The second stage (Stage 2) is identified as descent of the baby's head through the mother's pelvis assisted by active "pushing", with cervical dilation of 10 cm. The third stage is identified as the delivery of the placenta.

At present, OxSys extracts a total of 64 different FHR features (Table 2.4) for a sliding 15 minute window. This includes clinical or statistical features used previously (Dawes et al., 1992, Cazares, 2002, Pardey et al., 2002). There are also clinical features related to specific clinical phenomena (Cazares, 2002, Pardey et al., 2002, Georgieva et al., 2009, Georgieva et al., 2010), such as contractions, accelerations, decelerations and lag features. In addition, there are theoretical features derived using statistical methods (Georgieva et al., 2011a, Georgieva et al., 2011b, Fulcher et al., 2012, Xu et al., 2012, Xu et al., 2013, Georgieva et al., 2013b). The extraction of novel features, clinically or statistically meaningful, is still on-going to add more potentially useful features to the current database.

2.6 Summary

This chapter has provided an overview of the literature relevant to the field. The previous work on FHR features and computerised FHR analysis systems were provided in Section 2.1 and 2.2, while Section 2.3 and Section 2.4 gave a review of FHR feature extraction and feature selection methods. In conclusion, more FHR features should be extracted for further analysis of FHR, and feature selection methods should be applied to select and to integrate the FHR features for labour outcome classification and computerised decision support. The

unique large size of the OxSys database (49,151 cases) will be an advantage for these studies.

Table 2.4 List of OxSys features.

General features,

1 Baseline mean

5 Lowest dip of fetal heart rate

60 Sinusoidal feature (Reddy et al., 2009)

64 Simulation for clinical guidelines of American Congress of Obstetricians and Gynaecologists (ACOG, 2009)

Contraction features (Georgieva et al., 2009)

15 Number of contractions

16 Median of contraction duration

17 Median of resting time

18 Resting/contraction time ratio

Acceleration & Deceleration features

20 Number of accelerations

21 Mean acceleration duration

22 Mean acceleration amplitude

23 Number of decelerations

24 Mean deceleration duration

25 Maximum deceleration duration

26 Median deceleration amplitude

27 Maximum deceleration amplitude

28 Mean time to deceleration

29 Mean time to recover

30 Maximum time to recover

31 Number of quick recoveries

32 Number of slow recoveries

33 Resting/deceleration time ratio

34 Mean deceleration area

35 Total number of lost beats

36 Median onset slope

37 Median recovery slope

38 Maximal pattern readjustment after deceleration

39 Number of prolonged deceleration (≥3min)

40 Number of decelerations with right and/or left shoulder (Freeman et al., 2003)

41 Number of decelerations with only right shoulder (overshooting)

Lag features

42 Number of early decelerations

43 Number of late decelerations

44 Number of variable decelerations

45 Median recovery time after contraction end

46 Mean lag time (late decelerations only)

47 Deceleration/contraction time ratio

Variability Features

2 Percentage of zero difference between neighbour points

3 Approximate Entropy, (Dawes et al., 1992)

4 Signal Stability Index (SSI) (Georgieva et al., 2011a)

8 Long-term variability

9 SSI of the residual signal

10 Short term variability (STV)

11 SSI of STV tracker

12 Range of STV tracker

13 STV tracker trend

14 Median of the absolute derivative values of the STV tracker

19 STV (accelerations included)

61 Phase Rectified Signal Averaging (PRSA) – DC component (Georgieva et al., 2013b)

62 Bivariate Phase Rectified Signal Averaging (BPRSA) –AC component

63 BPRSA – DC component

Other time-series features (Fulcher et al., 2012, Xu et al., 2012)

6 Skewness

7 Kurtosis

48 Auto-mutual information feature

49 Ratio of standard deviation

50 Mean of local Approximate Entropy

51 STD of local Sample Entropy

52 Goodness of exponential fit

53 2-dim time delay embedding space

54 Median Absolute Deviation (MAD)

55 $(STD/mean)^2$

56 Alphabet feature

57 Pattern readjustment feature

58 Interquartile range of smoothed signal

59 STD of Gaussian filtered signal

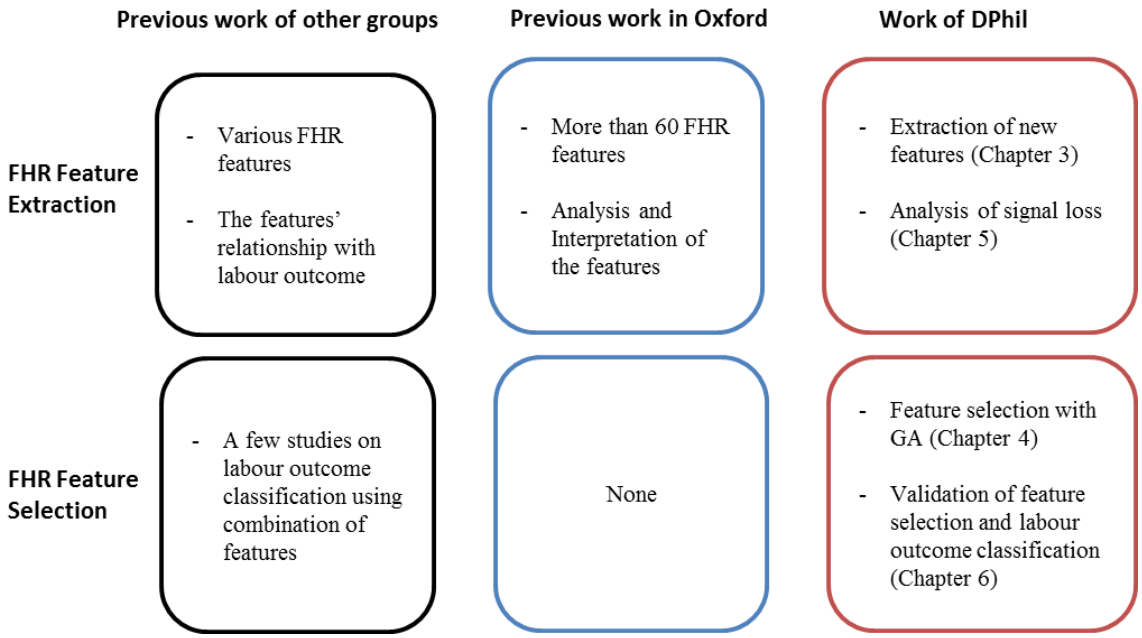


Figure 2.6 Outline of previous work and work of this DPhil.

Figure 2.6 gives an outline of the relationship between previous work and the work described by this thesis. Chapter 3 shows the previous work of feature extraction on pattern readjustment feature (Feature 57), as mentioned in Section 2.3.1. As described in Section 2.4, selection of FHR features has never been investigated previously, thus Chapter 4 introduces Genetic Algorithms (GA) to select a best feature subset out of the FHR features. Chapter 5 looks into an important issue for feature extraction that has never been investigated before (as mentioned in Section 2.3.3) - signal loss; the influence of signal loss is investigated and quantified. Chapter 6 then uses a new set of data to validate the feature selection method in Chapter 4 and the influence of signal loss in Chapter 5.

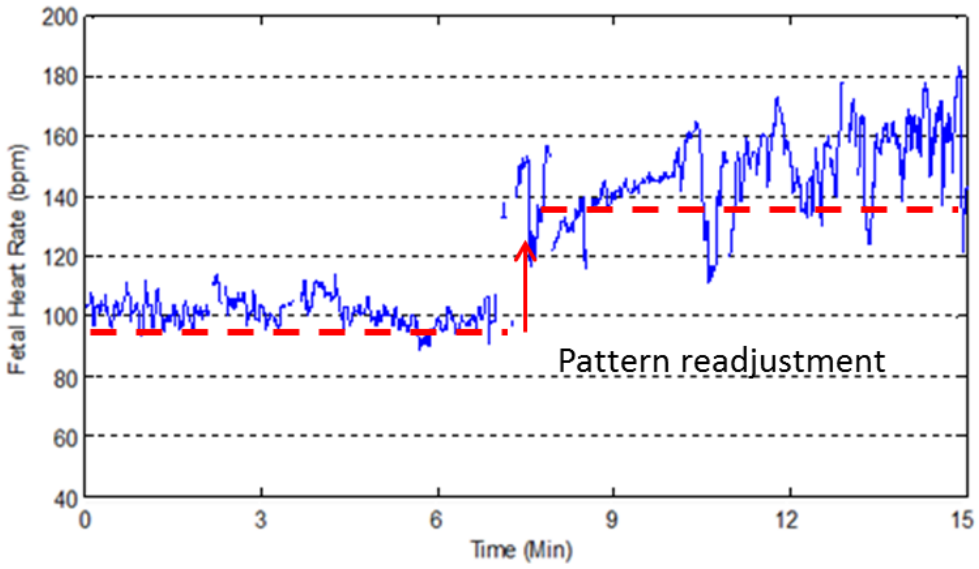
3 Detection and Analysis of Pattern Readjustment

3.1 Introduction

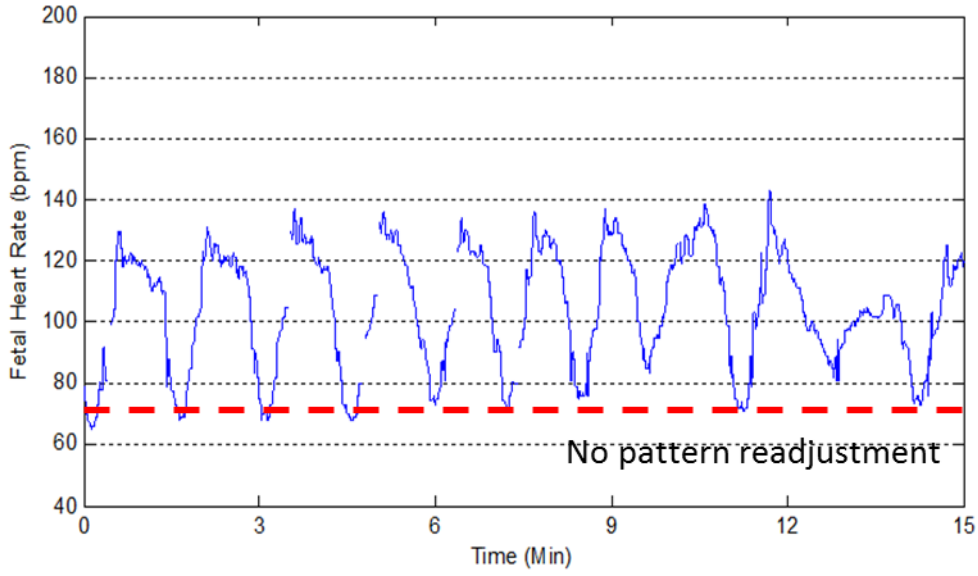
In 2011, *Georgieva et al.* developed an algorithm that detects baseline un-assignable (BU) segments, which means FHR segments without a clear baseline (Georgieva et al., 2011a). The algorithm was validated by comparison with clinical experts. The average agreement between the experts (Cohen's kappa κ = 0.76) was comparable to the agreement between the method and the experts (κ = 0.67), which revealed the potential of this method to improve the reliability of computerised analysis.

Based on observations of the BU segments, it was observed that some of the baselines are un-assignable owing to readjustments in the pattern. The detection of such "pattern readjustments" could be useful to draw attention to cases of "conversion pattern". These are patterns where "the baseline rises suddenly to a new stable rate with absent variability that persists to the end of the tracing" and these have been linked to neuronal injury of the fetus (Barry, 2004). Therefore, the hypothesis was proposed that pattern readjustment in the CTG signal is one of the abnormal patterns that are indicative of fetal compromise in labour.

"Pattern readjustment" can only be defined as changes of patterns in a single CTG window, especially referring to instantaneous jumps in the mean of a FHR signal (Figure 3.1). To the author's knowledge, the problem of FHR pattern readjustment has never been defined or investigated before.



(A) FHR segment with pattern readjustment



(B) FHR segment with no pattern readjustment

Figure 3.1 Examples of pattern readjustment in FHR segments. (A): FHR segment with pattern readjustment. The mean level of FHR has significant changes over the window. (B): FHR segment with no pattern readjustment; the FHR is fluctuating in the window but the pattern stays the

same.

The purpose of this chapter is to develop a classifier to detect FHR pattern readjustments and to assist computerised analysis in the OxSys System. An outline of the study design is shown in Figure 3.2. There are three major steps: algorithm development, validation and application. An algorithm was developed using the training samples, and its performance was compared to that of clinical experts. The FHR records of 7,568 patients were then classified into two groups: pattern readjustment and no pattern readjustment, in order to measure how often pattern readjustments occur in relation to different labour stages.

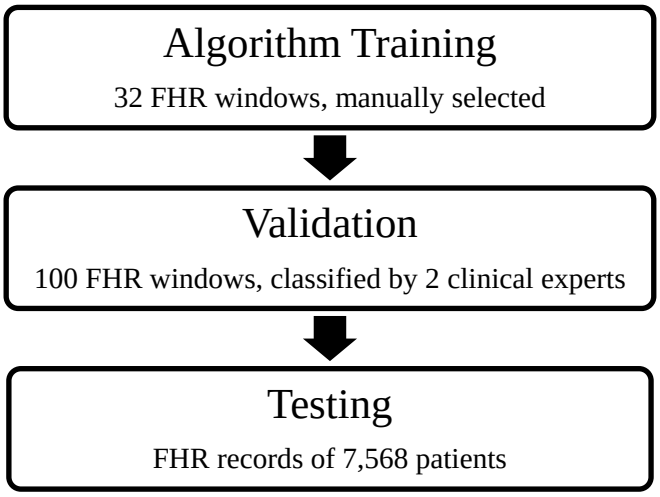


Figure 3.2 Outline of the study design

3.2 Data

The FHR signals of interest were automatically processed with a moving 15 minute window (sliding steps of 5 minutes). The data were stored with the original sampling rate of 4Hz. As there is no existing standard to detect pattern readjustments, the FHR data with and without pattern readjustment had to be selected manually. The data in the training group were selected

from a subset of OIFD1, where the cases from OIFD3 were excluded so that it could be used for testing. In this subset, 16 baseline un-assignable (BU) segments with clear pattern readjustments were manually selected to form a pattern readjustment group, together with 16 BU segments with no clear pattern readjustment to form a control group. The data were selected and noted by Liang Xu and Antoniya Georgieva, and the notation was agreed by Prof. Christopher Redman, an expert with over 30 years of experience of FHR interpretation.

It should be noted that the size of the training group (32 cases) is smaller than the size of the testing group (100 cases), which is due to the fact that in clinical situations only a small proportion of windows have a very clear sign of pattern readjustment or a very clear sign of no pattern readjustment such as those shown in Figure 3.1, thus it is difficult to identify many of the clear samples for training.

3.3 Methods

3.3.1 Pre-processing of the signal

Pattern readjustments can be regarded as instantaneous jumps in the mean of a signal, thus step detection methods can be used to detect them (Section 2.3.1). However, accelerations and decelerations can also be detected as abrupt changes in step detection. Therefore, the FHR segments must be pre-processed in order to eliminate the accelerations and decelerations, before any step detection method is applied.

To pre-process the signal, a filtering approach known as the open-close-smooth algorithm was adapted: firstly an opening filter eliminates accelerations from the original FHR signal; then a

closing filter is then applied eliminate decelerations from the open-filtered FHR signal; finally the output of the closing filter is passed through a averaging filter to smooth the signal. The algorithm developed by Cazares (2002) was integrated within OxSys for FHR baseline identification. The original algorithm was design to work with 0.25Hz signal sampling, while in this study it is applied to the original 4Hz signal, thus the 32 training windows were tested using the acceleration and deceleration detection algorithm also developed by Cazares (2002). It was found that the algorithm successfully removed all 56 accelerations and 61 decelerations in all training windows, thus the algorithm was found to be still applicable in this study. The results from this signal pre-processing are illustrated in Figure 3.3. It is found that the accelerations and decelerations are removed to estimate the baseline after the open-close-smooth algorithm.

3.3.2 Step detection filters

Two step detection filters were used to detect abrupt changes in the pre-processed signal: a first order Gaussian derivative filter and a zero crossing filter. These step detection methods were used owing to their intuitive interpretation and easy implementation.

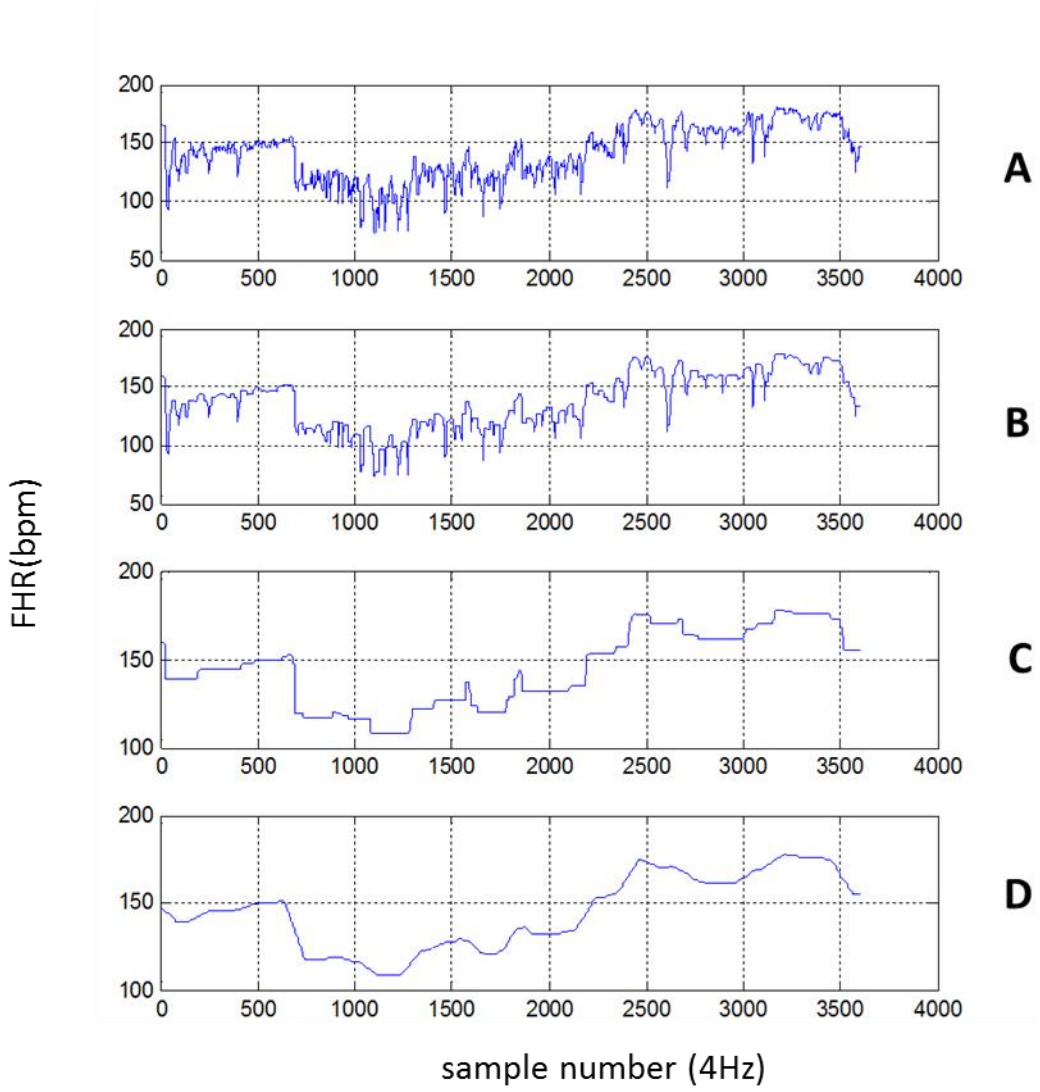


Figure 3.3 Illustration of the pre-processing steps. (A) Original FHR signal sample; (B) Opening-filtered signal of A; (C) Closing-filtered signal of B; (D) Smoothed signal of C.

The first order Gaussian derivative filter

In step detection methods, the gradient of the signal is often used as an indicator of abrupt changes. However, the simple application of differentiation to calculate the gradient can lead to amplification of the noise in the signal. To avoid this, the first order Gaussian derivative filter is often used to estimate the gradient. The original signal is convolved with the first

order derivative of the Gaussian kernel. In our cases, we use a 9 point kernel, corresponding to the first order derivative of a Gaussian function with μ = 0, sigma = 1, ranging from [-1, 1]. In this case, the original signal is convolved with the first order derivative of a Gaussian kernel, in order to detect the abrupt changes in the signal.

The zero crossing filter

The zero crossing filter is another method for detecting abrupt changes. It is frequently used owing to its simple implementation and intuitive basis. The zero crossing filter looks for places in the second order derivative of the signal where the value passes through zero, i.e. points where the second order derivative changes sign. Such points often occur at the "edge"' of the signal. In our cases, we use a 5 point kernel [0.5 0 -1 0 0.5], corresponding to the second order derivative of the signal. The original signal is convolved with the kernel, in order to detect the abrupt changes in the signal. The results from using both step detection filters are illustrated in Figure 3.4. It is shown that fluctuations of the pre-processed signal can be recorded in the filtered signal in different ways.

3.3.3 Signals, quantities and features

In order to find the statistical differences between the pattern readjustment group and the control group, different statistical quantities are considered.

The signal segments that have been tested are:

(A) Original FHR signal sample;

(B) Pre-processed FHR signal;

(C) First order Gaussian derivative filtered signal of B;

(D) Zero crossing filtered signal of B.

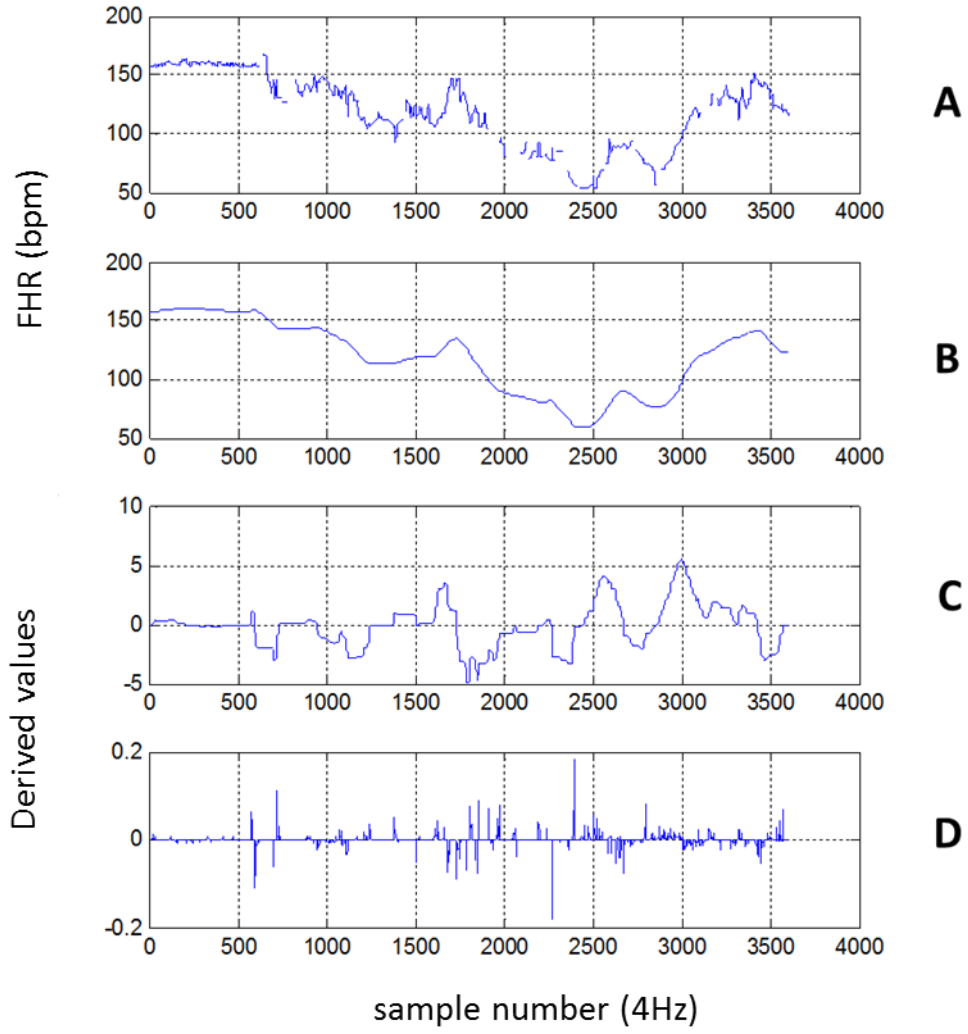


Figure 3.4 Illustration of the filtering steps. (A) Original signal sample. (B) Pre-processed signal. (C) First order Gaussian derivative filtered signal of B. (D) Zero crossing filtered signal of B.

The statistical quantities that have been considered are:

(I) Standard deviation;

(II) Interquartile range;

(III) Mean absolute deviation (MAD);

(IV) Kurtosis;

(V) Skewness.

These five quantities related to the dispersion (I, II and III) and shape (IV and V) of the data were chosen, because intuitively, these characteristics can be related to the features of pattern readjustment. Therefore, for each sample segment, there are 4×5 = 20 features to be tested. These quantities are indexed as shown in Table 3.1.

Table 3.1 Indices of the signal quantities

	(I) Standard deviation	***(II) Interquartile range***	***(III) Mean absolute deviation (MAD)***	***(IV) Kurtosis***	***(V) Skewness***
(α) Original FHR signal sample	1	2	3	4	5
(β) Pre-processed FHR signal	6	7	8	9	10
(γ) First order Gaussian derivative filtered signal of B	11	12	13	14	15
(δ) Zero crossing filtered signal of B	16	17	18	19	20

3.4 Results

3.4.1 Classification results of the SVM method

In order to evaluate the differences between the pattern readjustment group and the control group, the 20 quantities described in Table 3.1 were considered. For each pair of quantities, the values of the two specified quantities were calculated for each sample segment. This new set of data was then classified using the SVM method (see Section 2.3.2), and the kappa was also recorded.

The performance of the quantity pairs is shown in the kappa map in Figure 3.5 (A). The inputs of the SVM are the values of each quantity pairs (2-D), the target classes are pattern readjustment (PR) and no patter readjustment (Control). Three pairs of quantities obtained the best performance of kappa = 1: Quantity 7 against Quantity 11 (termed SVM1), Quantity 7 against Quantity 13 (termed SVM2), and Quantity 7 against Quantity 18 (termed SVM3). As the classification performance has reached its maximal value in 2-D, there is no need to increase the dimension of features. The classification results of the three quantity pairs are shown in Figure 3.5 (B-D). The results show that these classifiers are promising for accurate detection of pattern readjustments.

3.4.2 Classifier selection

These classifiers were tested and compared on the OIFD1 database. To investigate how these three classifiers are correlated, the classification results of the testing group that have over 90% signal quality from the three classifiers were compared in pairs. The kappa value and proportion of agreement were recorded and compared in Table 3.2. It was found that the three

classifiers are highly correlated, and that the proportion of agreement is high between each other. Therefore, the performances of the three classifiers are similar. This is expected because the three quantity pairs share one same quantity (Quantity 7).

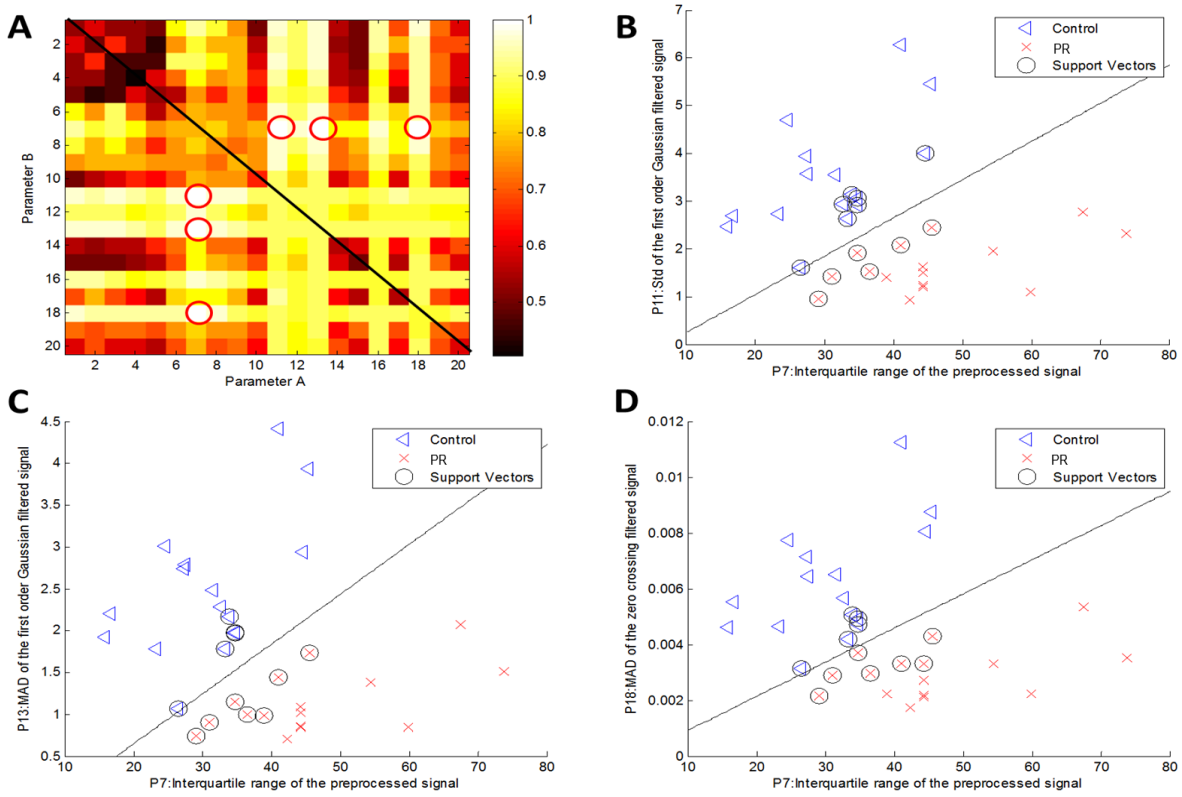


Figure 3.5 Classification results of the SVM methods (A): Kappa map of the different quantity pairs. Values of x axis and y axis refers to the index of the quantity, while different colours refer to different classification accuracies, and the red circles indicate the quantity pairs that have kappa = 1. (B) (C) (D): Classification results of the three quantity pairs with kappa = 1, in which support vectors are circled. PR: pattern readjustment and Control: no pattern readjustment.

Table 3.2 Correlation parameters between the three classifiers (in pairs).

Classifiers	*Kappa value (κ)*	*Proportion of agreement*
SVM1 & SVM2	0.88	96%
SVM2 & SVM3	0.88	96%
SVM1 & SVM3	0.85	95%

In addition, to investigate how the pattern readjustment rate (proportion of pattern readjustments in the BU segments) varies under different signal quality (the proportion of valid signal) conditions, the pattern readjustment rates of each group are compared using different signal quality threshold (minimum allowed signal quality) in Figure 3.6. It is found that, in general, the influence of the signal quality threshold (minimum proportion of valid signal) on the pattern readjustment rate is small. For SVM1, the pattern readjustment rate hardly varies in different groups. For SVM2 and SVM3, pattern readjustment segments are more likely to occur with poorer signal quality, but even with > 90% valid signal, over 20% of all BU segments can be expected to be pattern readjustment segments. In conclusion, SVM1 is less influenced by the signal quality, in comparison to SVM2 and SVM3. Therefore, SVM1 is chosen for validation, since that while the classification performances of the three classifiers are similar, SVM1 was less influenced by signal quality.

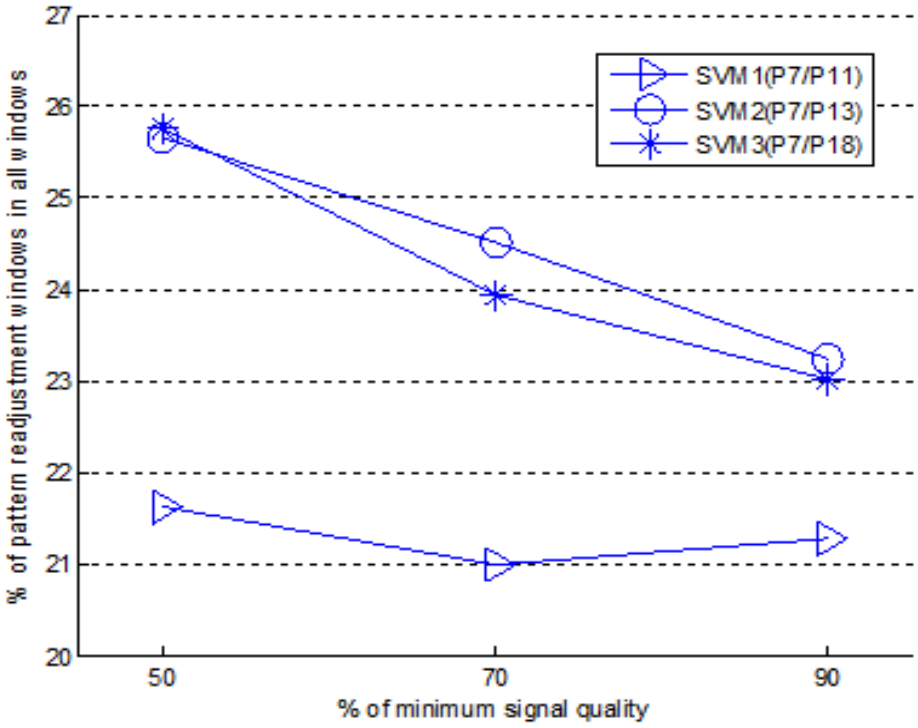


Figure 3.6 Relation between the pattern readjustment rate and the signal quality threshold of different classifiers. SVM1 is less influenced by signal quality, in comparison with the other two classifiers.

3.4.3 Validation

To validate the SVM classifiers, 100 windows of 15-minute length were selected. The segments were then reviewed by two clinical experts (Prof Christopher Redman and Ms Mary Moulden, staff members each with over 30 years' experience of FHR interpretation) blinded to the algorithm classification. Only the FHR segments within the 15-minute interval were shown to the experts, without information from before or after the observed window. The experts were then asked to classify the windows into two groups: pattern readjustment and no pattern readjustment. To compare the performance of the algorithm with the standard

interpretation in practice, no specific definition of baseline was imposed on the experts. Instead, they were asked to use their own perception based on their long experience with FHR interpretation.

Cohen's kappa coefficient (ranging between -1 and 1, higher values indicating better agreement) was used to measure the agreement between the experts. This is a statistical measure of agreement between predicted and actual results, which is considered to be a more robust measure than simple proportion agreement, since kappa takes into account the fact that agreement that can occur by chance (Cohen, 1960). By measuring the disagreement between experts, the best performance that can be expected from the system is determined, since a computerised system cannot agree with experts better than they agree with each other. The average agreement between the two experts (80% agreement, kappa = 0.61) was comparable to the agreement between the new method and experts (76% agreement, kappa = 0.52 with expert 1 and 78% agreement, kappa = 0.56 with expert 2).

The algorithm was then applied to OIFD3 (7,568 patients with FHR records and labour outcome information) to investigate the association between pattern readjustment and low arterial pH at birth (fetal acidemia). A steady increase was seen in the risk of low arterial pH with an increase in number of pattern readjustment windows (Table 3.3). In the second labour stage (active pushing), the odds ratio against occasions with no pattern readjustment window increased from 2.42, 95% confidence interval (CI) [2.09 2.76] (1 pattern readjustment window) to 6.3, 95% confidence interval [4.19 8.58] (4 pattern readjustment windows), where odds ratio refers to the odds ratio calculated against occasions with no pattern readjustment

window. This indicates that pattern readjustment is a promising feature.

Table 3.3 The risk of low arterial pH (cord blood pH < 7.05) increases with the number of pattern readjustment windows.

No. of pattern readjustment windows	First labour stage (established labour)			Second labour stage (active labour)		
	Cases with pH < 7.05 (%)	Odds ratio [95% CI]	Total No. of cases	Cases with pH < 7.05 (%)	Odds ratio [95% CI]	Total No. of cases
0	292 (4.1%)	N/A	7192	260 (3.8%)	N/A	6905
1	20 (7.5%)	1.9 [1.44 2.38]	268	44 (8.7%)	2.4 [2.09 2.76]	508
2	6 (6.6%)	1.7 [0.83 2.50]	91	10 (8.9%)	2.5 [1.84 3.17]	112
3	1 (7.6%)	1.5 [-0.45 3.60]	16	4 (10.5%)	3.0 [1.96 4.05]	38
4	0 (0%)	Not Applicable	1	1 (20.0%)	6.3 [4.19 8.58]	5

3.5 Discussion

Although there is no precise current definition of pattern readjustments, it indicates a sharp change in the FHR baseline, which can intuitively indicate a change in fetal health. In order to detect such abnormal patterns, an algorithm using the SVM method was developed. Quantity pairs were used as two-dimensional feature space for the SVM, which provided kappa = 1 on the training set.

Three SVM classifiers were compared and they showed comparable performance in detecting pattern readjustments. SVM1 was selected because it was less influenced by signal quality in comparison with the other two SVMs. The interpretation of SVM1 is intuitive: at the same level of change (reflected by the standard deviation of the first Gaussian filtered signal), the

pattern readjustment windows tend to have a higher value of interquartile range than other BU windows. This is because they tend to have more than one level of baseline, which will lead to higher interquartile range value. The classification rate of these SVMs show that the step detection model appears appropriate; it is also the first time such a model has been applied to FHR analysis.

The algorithm was used on 100 validation samples, and the opinions of two clinical experts were compared with the SVM classification. The validation results (Section 3.4.3) showed that the agreement between the two experts (kappa = 0.61) was comparable to the agreement between the experts and the SVM (kappa = 0.52 and 0.56). This validates the accuracy of detection of the algorithm.

The algorithm was then tested on 7,568 patients (OIFD3) with FHR records and labour outcome information, to investigate the association between pattern readjustment and low arterial pH at birth (an unfavourable condition of labour outcome). A steady increase was seen in the risk of low arterial pH with an increase in number of pattern readjustment windows: in the second labour stage (active pushing), the odds ratio against occasions with no pattern readjustment window increased from 2.42 [2.09 2.76] (1 pattern readjustment window) to 6.3 [4.19 8.58] (4 pattern readjustment windows). This suggests that the pattern readjustment detected here is a novel promising feature that is associated with fetal acidemia. Combined with other features, it could assist computerised FHR analysis and further clinical studies.

3.6 Conclusions

In this chapter, an observed novel feature, “pattern readjustment”, was extracted and tested. It is the first time that step detection methods have been applied to FHR analysis. The validation results with clinical experts showed that the pattern readjustment can be accurately detected. The application to 7,568 cases revealed that the feature was associated with fetal acidemia, thus it can be helpful in computerised FHR analysis and subsequent clinical studies.

4 FHR Feature Selection using Genetic Algorithms (GA)

4.1 Introduction

At present, OxSys contains 64 features. Previous studies have shown that the combination of different FHR features can perform better than using features independently (Chudáček, 2011, Georgieva et al., 2013c). These studies reveal the potential of combined features to provide a better predictor than individual features. On the other hand, the number of all possible combination of the 64 features is $2^{64} = 1.8 \times 10^{19}$, which is too large for an exhaustive search. Therefore, feature selection methods need to be applied to the task of selecting a subset of FHR features that provides the best prediction performance. To the author's knowledge, feature selection of FHR features has been investigated previously only on databases of very limited size (see Section 2.4).

Amongst the different feature selection approaches, Genetic Algorithms are known for their competitive exploration ability, i.e. the ability to explore the feature space as widely as possible. They have been found to be powerful and efficient global optimisers in various fields of data analysis (Mitchell, 2001). The reasons to select GA for FHR feature selection are summarised in Section 2.4.

The objective of this chapter is thus to apply Genetic Algorithms as a feature selection method to select a best feature subset from 64 FHR features and to integrate these best features to predict adverse labour outcome. The GA was trained on 404 cases and tested on 106 cases

(both balanced datasets) using three classifiers respectively. Regularisation methods and backward selection were used to optimise the GA algorithm. This is to the author's knowledge the first time that a feature selection method for FHR analysis has been tested on a database of this size.

4.2 Data

4.2.1 Total set of 7,568 records

To ensure that all cases are analysed at comparable stages of labour, only the last 30 minutes of Stage 2 (before birth) were examined here. The assumption of this study is that, in the last 30 minutes of Stage 2, adverse labour outcome is detectable using CTG. Therefore in this study, only CTG records taken after the onset of pushing were included. The detailed selection criteria in this study are shown in Figure 4.1. A total set of 7,568 FHR records were also selected (the OIFD3 database). The criteria are consistent with a previous study of the group using Artificial Neural Networks (ANN) for FHR classification (Georgieva et al., 2013c).

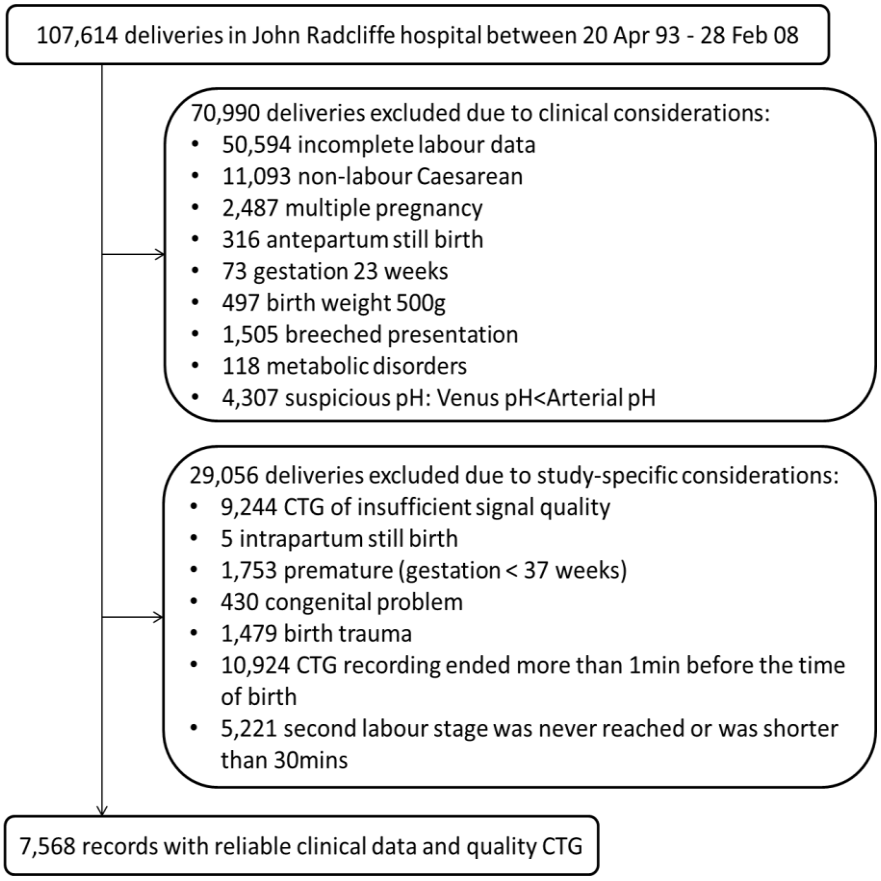


Figure 4.1 Data selection criteria of the 7,568 records that have labour outcome information.

The CTG was recorded in 15-minute windows using 5-minute sliding steps, thus in the last 30 minutes of Stage 2, there are four 15-minute windows. For each feature, the median value of the four windows was taken as a first step to reduce data complexity and to avoid extreme outliers. As there were various types of features, the scale of data can affect the performance of the classifiers. Therefore, each feature's mean was normalised to 0 and standard deviation to 1. The numerical computing and analysis package Matlab 2012b (The Mathworks, Inc.) was used for the algorithm development and all data analyses.

4.2.2 GA training set and testing set

Adverse labour outcome is defined as low arterial pH (pH < 7.05) according to previous

studies (Georgieva et al., 2013c). Normal outcome was defined as $7.27 < pH < 7.33$ and no form of compromise (959 cases). As fetal acidemia is rare, the dataset is heavily imbalanced: only 255 out of 7,568 cases have an adverse outcome (3.37%). Training a classifier using the entire dataset will be heavily biased towards recognising the healthy cases. Therefore, a balanced dataset was selected to train the classifier.

To select a balanced dataset of 50% normal outcomes and 50% adverse outcomes, 255 cases were randomly selected out of the 959 normal outcomes. Therefore, the dataset used in this study consists of 510 cases, with 255 normal outcomes and 255 adverse outcomes.

Figure 4.2 shows the distribution of the training and testing data. From the 510 cases, 404 cases (202 normal and 202 adverse cases) were randomly selected for the feature selection process using the GA. The remaining 106 cases (53 normal cases and 53 adverse cases) were used as a testing set to evaluate the performance of the features selected by GA. In addition, in each GA run, the 404 cases were further separated into a training set and a validation set (70%-30%). To avoid confusion, the data used in the GA training are referred to as the GA training set, and the separate testing set used to evaluate the performance of the GA are referred to as the testing set (Figure 4.2). The widely accepted “rule of thumb” (Van Niel et al., 2005) was followed that at least 10 training samples of each class per input feature dimension are needed. Therefore, the maximal number of features that can be selected is 14. This was incorporated as a constraint into the algorithms. As there are various types of features, the scale of the data can affect the performance of the classifiers. Therefore the features were all normalised by a zero-mean, unit-variance transformation by the mean and variance calculated

in the GA training set.

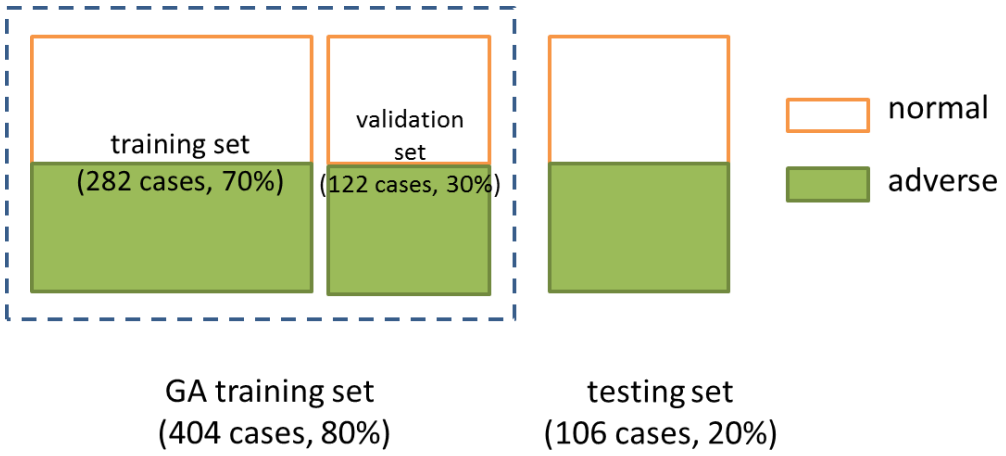


Figure 4.2 Distribution of all 510 cases used for GA.

4.2.3 Distribution analysis of arterial pH

To analyse whether the GA training set and testing set are representative of the population, the distribution of arterial pH is analysed for both sets (Figure 4.3). It is found that the distributions of the GA training set and testing set are both similar to the distribution of all 510 cases used in the study. In addition, the p-value calculated using the Kolmogorov–Smirnov test is > 0.5 between the GA training set and all 510 cases, and also > 0.5 between testing set and all 510 cases (higher p-values indicates that the two sets in the test are more likely to come from the same distribution). This indicates that there is no significant difference in distribution between GA training set/testing set and all 510 cases.

4.3 Univariate Analysis of Individual Features

To study the statistical properties of each feature individually, univariate analysis methods were used for the 255 normal cases and 255 adverse cases. All of the analysis results are listed

in Table 4.1 (if not specified), while the description and analysis are given below. It should be noted that these tests use all 510 cases, which is not the case in Section 4.4-4.7

Normality analysis: Jarque-Bera test and two-sample t-test

The Jarque-Bera test (Jarque and Bera, 1980) was used to test the normality of each feature (all 510 cases). The null hypothesis (H = 0) assumes that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution. The results indicate that most of the features (>90%) are not normally distributed at the 5% significance level (H = 1).

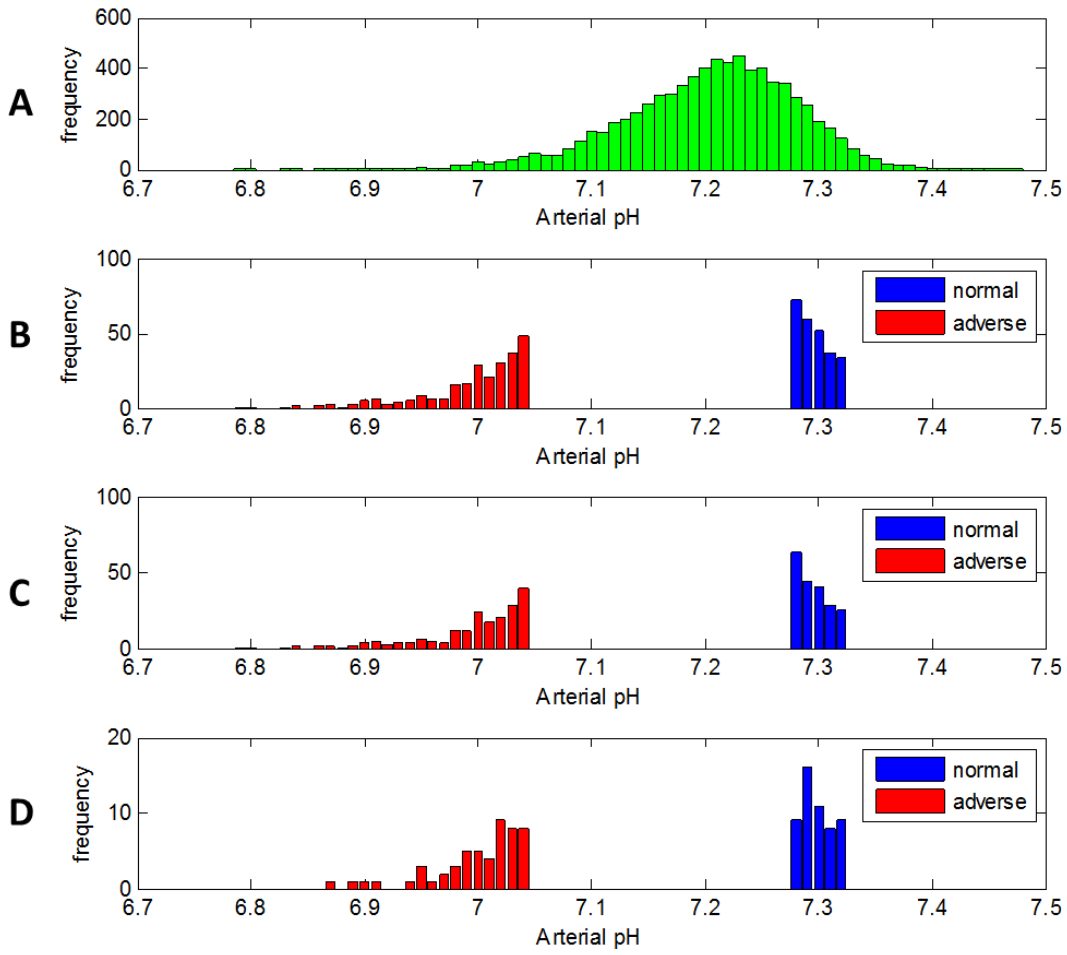


Figure 4.3 Distribution analysis of arterial pH: (A) distribution of all 7,568 cases; (B) distribution of the 510 cases used in this study; (C) distribution of the GA training set; (D) distribution of the testing set.

Table 4.1 Univariate analysis of individual feature, p-values that <0.05 are bolded; highest classification rate and AUC values are marked by red.

	Feature Name	*Jarque-Bera test: H Value*	*Mann-Whitney U test: p Value*[1]	*Kolmogorov–Smirnov test: p Value*	*Linear correlation with low pH: R^2 Value*	*10-fold Linear discriminant analysis: Classification Rate: Mean/Std*		*Univariate AUC*
1.	Baseline mean	0	**0.0037**	**0.0004**	0.0218	0.53	0.06	0.602
2.	% zero diff between neighbour points	0	**0.0324**	0.0861	0.0102	0.53	0.06	0.501
3.	Approximate Entropy	1	0.3826	0.4638	0.0011	0.50	0.11	0.510
4.	Signal Stability Index (SSI)	1	**0.0000**	**0.0000**	0.1607	0.68	0.07	0.703
5.	Minimal Expected Value	1	**0.0000**	**0.0000**	0.1302	0.65	0.07	0.634
6.	Skewness	1	**0.0000**	**0.0000**	0.0291	0.6	0.05	0.639
7.	Kurtosis	1	**0.0000**	**0.0000**	0.0635	0.61	0.06	0.677
8.	Long-term variability	1	**0.0000**	**0.0000**	0.1432	0.67	0.05	0.619
9.	SSI of the residual signal	1	**0.0000**	**0.0000**	0.0392	0.60	0.05	0.560
10.	Median of STV change tracker	1	**0.0000**	**0.0000**	0.1034	0.63	0.07	0.597
11.	SSI of STV tracker	1	**0.0000**	**0.0000**	0.0490	0.63	0.06	0.566
12.	Range of STV tracker	1	**0.0008**	**0.0110**	0.0328	0.55	0.06	0.501
13.	STV tracker trend	0	**0.0387**	**0.0253**	0.0079	0.53	0.06	0.590
14.	Median of absolute derivative values	1	**0.0000**	**0.0000**	0.0949	0.63	0.04	0.582
15.	Number of contractions	1	0.4920	0.8869	0.0006	0.50	0.07	0.521
16.	Median of contraction duration	1	**0.0000**	**0.0006**	0.0278	0.56	0.07	0.593
17.	Median of resting time	1	**0.0001**	**0.0001**	0.0144	0.57	0.07	0.594
18.	Resting/contraction time ratio	1	0.2398	0.3392	0.0016	0.50	0.05	0.509
19.	Median STV tracker	1	**0.0000**	**0.0000**	0.0879	0.61	0.06	0.563
20.	Number of accelerations	1	0.7871	0.6834	0.0002	0.5	0.03	0.516
21.	Mean acceleration duration	1	**0.0365**	**0.0253**	0.0157	0.54	0.05	0.535
22.	Mean acceleration amplitude	1	**0.0001**	**0.0000**	0.0365	0.6	0.07	0.575
23.	Number of decelerations	1	**0.0000**	**0.0000**	0.0577	0.59	0.07	0.552
24.	Mean deceleration duration	1	0.9275	0.9684	0.0001	0.51	0.07	0.547
25.	Maximal deceleration duration	1	0.2537	0.4638	0.0063	0.53	0.05	0.511
26.	Median deceleration amplitude	1	**0.0036**	**0.0110**	0.0264	0.55	0.06	0.506
27.	Maximal deceleration amplitude	0	**0.0053**	**0.0061**	0.0169	0.56	0.05	0.504
28.	Mean time to decel (t1)	1	0.1035	0.1981	0.0056	0.52	0.08	0.530
29.	Mean time to recover (t2)	1	0.1380	0.2860	0.0008	0.52	0.05	0.510
30.	Maximal time to recover	1	0.6761	0.4638	0.0010	0.51	0.07	0.533

31.	Number of quick recoveries (t1 ⩾ t2)	1	**0.0000**	**0.0000**	0.0532	0.59	0.08	0.542
32.	Number of slow recoveries (t2 < t1)	1	**0.0383**	0.1981	0.0073	0.52	0.07	0.508
33.	Resting/deceleration time ratio	1	**0.0000**	**0.0001**	0.0377	0.59	0.06	0.515
34.	Mean deceleration area	1	**0.0033**	**0.0033**	0.0195	0.56	0.04	0.501
35.	Total number of lost beats	1	**0.0065**	**0.0329**	0.0182	0.53	0.06	0.500
36.	Median onset slope	1	0.2172	0.2391	0.0019	0.51	0.07	0.520
37.	Median recovery slope	1	**0.0011**	**0.0017**	0.0172	0.56	0.05	0.578
38.	Maximal pattern readjustment after	1	0.9635	0.2391	0.0179	0.51	0.09	0.588
39.	Number of prolonged decels (⩾3min)	1	0.2403	0.9684	0.0056	0.51	0.06	0.511
40.	Number of decels with right and/or left	1	**0.0000**	**0.0000**	0.1003	0.65	0.04	0.566
41.	Number of decels with only right	1	**0.0000**	**0.0000**	0.1158	0.66	0.05	0.596
42.	Number of early decelerations	1	0.1508	0.9995	0.0047	0.51	0.08	0.511
43.	Number of late decelerations	1	**0.0032**	0.0861	0.0221	0.55	0.06	0.501
44.	Number of variable decelerations	1	0.0718	0.3392	0.0058	0.54	0.09	0.540
45.	Median recovery time after con end	1	**0.0085**	0.0540	0.0169	0.54	0.08	0.504
46.	Mean lag time (late decels only)	1	**0.0081**	**0.0253**	0.0167	0.54	0.06	0.505
47.	Deceleration/contraction time ratio	1	**0.0000**	**0.0001**	0.0294	0.58	0.07	0.515
48.	Mutual information feature	1	**0.0002**	**0.0000**	0.0238	0.6	0.06	0.500
49.	Ratio of STD	1	**0.0000**	**0.0000**	0.0358	0.60	0.08	0.538
50.	Mean of local ApEn	1	**0.0000**	**0.0000**	0.1364	0.66	0.07	0.539
51.	STD of local SampEn	0	**0.0000**	**0.0000**	0.0989	0.64	0.05	0.503
52.	Goodness of exponential fit	1	**0.0000**	**0.0000**	0.1535	0.65	0.04	0.502
53.	2-dim time delay embedding space	1	**0.0000**	**0.0000**	0.1433	0.63	0.05	0.569
54.	MAD	1	**0.0000**	**0.0000**	0.1761	0.66	0.04	0.602
55.	(STD/mean)^2	1	**0.0000**	**0.0000**	0.0416	0.67	0.06	0.610
56.	Alphabet feature	1	**0.0000**	**0.0000**	0.1799	0.68	0.08	0.595
57.	SVM output of the pattern readjustment	1	**0.0083**	**0.0001**	0.0135	0.57	0.06	0.524
58.	Interquartile range of the smoothed	1	**0.0000**	**0.0000**	0.1553	0.68	0.06	0.603
59.	STD of Gaussian filter	1	**0.0000**	**0.0000**	0.1473	0.63	0.11	0.629
60.	Regmeas (Moulden's sinusoidal)	1	**0.0013**	**0.0045**	0.0181	0.53	0.06	0.542
61.	PRSA - DC	1	**0.0000**	**0.0000**	0.2037	0.70	0.06	0.615
62.	BPRSA – std_AC	1	**0.0000**	**0.0000**	0.1017	0.63	0.07	0.540
63.	BPRSA – DC	1	**0.0128**	**0.0001**	0.0134	0.53	0.06	0.539
64.	ACOG (1-normal, 2-susp, 3-pathol.)	1	0.2179	0.6834	0.0035	0.52	0.06	0.509

[1] p Values <0.05 are highlighted in bold, stating that the result of the test is statistically significant at the significance level of 0.05.

The two-sample t-test is a non-parametric statistical hypothesis test for related observations. The null hypothesis is that data in the two samples are independent random samples from normal distributions with equal means and equal but unknown variances, while low p-values reject this hypothesis and indicate that there are significant differences between the two

samples. Since the test is based on the assumption that both samples are normally distributed, it could only be applied to the five normally distributed features. The normal group and adverse group were the two samples of the t-test, and the p-values are all <0.05 for the five features, indicating that the two groups in these features are from normal distributions with different means, at the significance level of 0.05.

Mann-Whitney U-test

The Mann-Whitney U-test (Mann and Whitney, 1947) is a non-parametric statistical hypothesis test for related observations. Its null hypothesis is that for data in the two samples *x* and *y*, the differences for each couple (*x*-*y*) come from a distribution with zero difference in their median. Low p-values reject this hypothesis and indicate that there are significant differences between *x* and *y* (so that *x*-*y* is not likely to be zero). Unlike the two-sample t-test, the Mann-Whitney U-test does not have an assumption about normal distribution, so it can be applied for all features.

It is found that 49 out of 64 features have a value of $p < 0.05$ between the adverse group and normal group, which indicate that for these features, the normal group is significantly different from the adverse group.

Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test (Smirnov, 1948) is a non-parametric statistical hypothesis used to test the hypothesis that the two samples are generated from the same distribution. Low p-values reject this hypothesis and indicate that the cumulative distributions from the two samples are different from each other.

It is found that 45 out of 64 features have a value of $p < 0.05$, which indicates that for these features, the normal group are significantly differently distributed from the adverse group. In addition, the 45 features are a subset of the 49 features that have a value of $p < 0.05$ in the Mann-Whitney test, i.e. in these 45 features, the normal and adverse groups are different from each other not only in values, but also in distributions.

Linear correlation analysis: R-square value

The R-square value (or R^2 value) is used to describe the linear correlation of a set of data, i.e. how well a regression line fits it. R-square is a number between 0 and 1. An R-square near 1 indicates that a regression line fits the data well, while an R-square closer to 0 indicates that a regression line does not fit the data very well.

Figure 4.4 shows the ranking of the 64 features in terms of their R-square value with low arterial pH (0 = normal group, 1 = adverse group). It is found that the R-square values between these features and low arterial pH are generally not high: over half of the features have R-square values lower than 0.05, and the highest R-square value (Feature 61) is approximately 0.2. This indicates that none of these features has a strong linear correlation to low arterial pH individually.

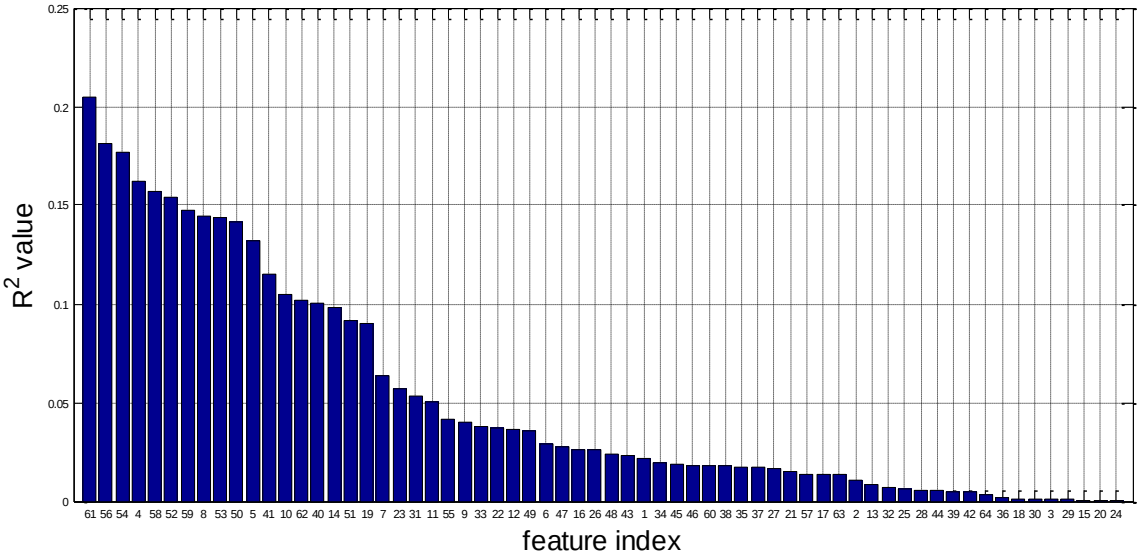


Figure 4.4 The ranking of R-square values between each feature and adverse labour outcome (low arterial pH).

The linear correlation between each feature is illustrated in terms of R-square values in Figure 4.5. It is found that some of the features have a strong linear correlation (R-square value above 0.8) to each other. This is due to the fact that some of the features were extracted using a similar clinical pattern (e.g. the acceleration features) and statistical methods (e.g. the variability features). Therefore, some of the features can provide redundant information when used together.

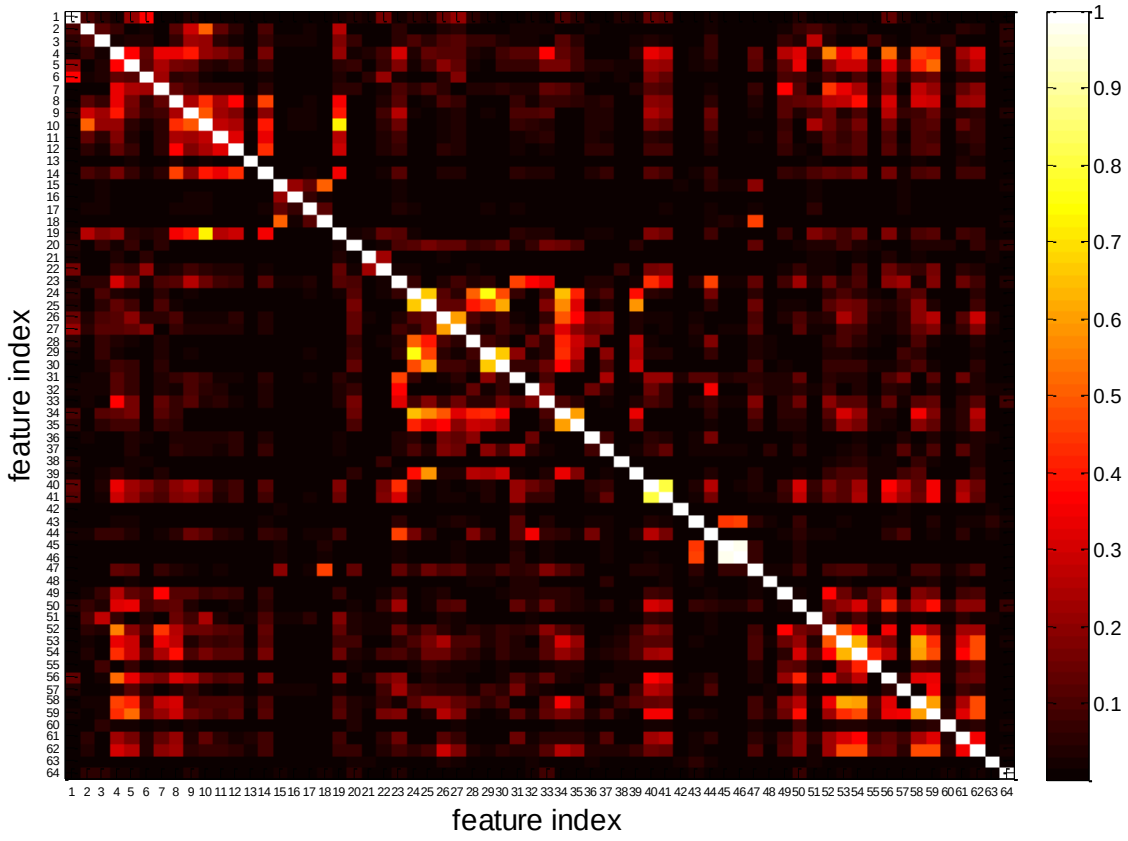


Figure 4.5 R-square values between different features.

4.4 Feature Selection using Genetic Algorithms

In this section, a GA algorithm was applied to the GA training set to select a best feature subset for three different classifiers. The classification performances of these best feature subsets were evaluated on the testing set by comparing the classifier predictions and actual labour outcomes. The performance is also compared with three other frequently used feature selection methods: Random Forests using the classification method (RF-C), Random Forests using the regression method (RF-R) (Breiman, 2001) and Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). These methods are known for their kappa, but they either cannot give a specific best set of feature (Random Forests), or are difficult to

use for interpretation (LASSO).

4.4.1 Methods

4.4.1.1 Basic framework

The framework of the method is adapted from the traditional GA method (Mitchell, 2001, Goldberg, 1989). A brief illustration of the GA implemented here is shown in Figure 4.6 (the total number of GA runs m = 100, the number of cross-validation splits n = 10). The GA method consists of five major steps:

(1) Generation of the population

In GA, the encoding method of individuals depends on the nature of the problem. Initially, the GA was proposed as a method that selected different features using binary representations (Goldberg, 1989). However, as there were more and more optimisation problems requiring searching for the optimal point in real space, most authors now use real-coding, i.e. all population points are represented as real-valued vectors (El-Mihoub et al., 2006).

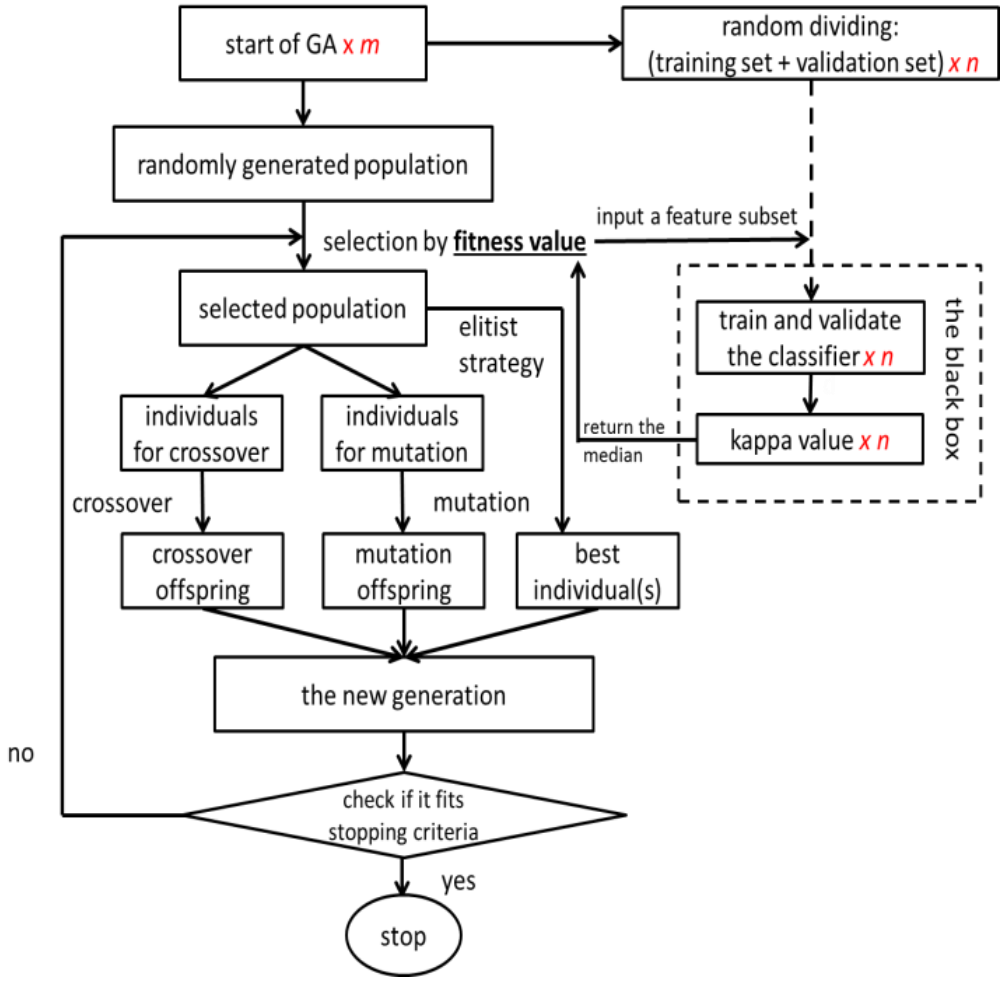


Figure 4.6 Basic framework of the GA method.

When adapted for FHR feature selection, the GA has the following framework: a population of strings (the genotype of the genome) that encode candidate FHR feature subsets (individuals) evolves towards better solutions. Therefore, binary encoding would be more appropriate, i.e. each individual of the population was encoded with a binary string, the length of which is equal to the feature set size. Each bit of these strings corresponded to one specific FHR feature. If the bit value was "1", this feature was included in the classifier; if the value was "0", it was not included in the classifier. In this way, each individual of the population represents a specific FHR feature subset.

(2) Selection based on fitness value

The evolution starts from a randomly generated initial population. For each generation, the fitness of every individual in the population is evaluated, and multiple individuals are selected from the current population based on their fitness using different selection strategies, such as roulette strategy and ranking strategy (Mitchell, 2001). The details of selection of fitness value are discussed in Section 4.4.1.2.

(3) Reproduction (crossover and mutation)

After the process of selection, the selected individuals are modified (recombined and possibly randomly mutated) to create a new population. By producing a "child" individual using the following methods of crossover and mutation, a new individual is created which typically shares many of the characteristics of its "parents". Different methods for recombination and mutation have been considered (Engelbrecht, 2002, Mitchell, 2001, Goldberg, 1989). Here one of the most common ways described in Mitchell (2001) is used, since it has been applied before on similar datasets (The heart disease dataset, from the UCI machine learning repository, with 276 cases and 13 features (Frank and Asuncion, 2010)).

In this method, a subset of the selected population is used as the “parents” for creating new individuals by crossover. In each crossover process, two individuals are randomly selected as the “parents”. The strategy used in this method is the single point crossover strategy (Mitchell, 2001). A crossover point is randomly selected, and the two individuals make a crossover with each other at the crossover point to generate two new individuals. This process continues until

the required number for individuals generated by crossover is met (Figure 4.6).

The rest of the population is used for mutation. In each mutation process, one individual is randomly selected, one or several mutation points are randomly selected, and the bits at the mutation points are flipped over to generate a new individual. This process continues until the required number for individuals generated by mutation is met.

(4) Accepting the new generation

If the selection mechanism includes an Elitist strategy, i.e. the best individual in each iteration is copied to the next one without any modification, it has been proven in (Rudolph, 1994) that the convergence of GA is guaranteed. Therefore, the Elitist strategy is adopted here.

(5) Checking the stopping criteria

The new population is then used in the next iteration of the algorithm. In most GA applications, this generational process is repeated until a termination condition is reached. In this method, the algorithm terminates when the maximum number of generations is reached, or when the best fitness no longer changes for a certain number of iterations. Details of stopping criteria are described in Section 4.4.1.4.

Parameter setting

At each run of a GA the initial population and data splits will both be different, and so the algorithm must be run multiple times to compare the results. The population size was set here to 100. A ranking strategy was used as the selection strategy, where the individuals ranking in the best 50% fitness were selected to create offspring. Single-point crossover with a

proportion of 0.8 and single-point mutation with a proportion of 0.2 was applied to every generation to generate the next population. The elitist strategy selection was set to 2, i.e. the best two individuals of the current generation were included in the next population. The maximum number of generations with the same best fitness value was set to 20. The maximum number of generations was set to 200. The number of GA runs with different initial conditions, *m*, was set to 100.

It has been tested (results not shown) on the GA training set that increasing the population size, the number of GA runs, and maximum number of generations will not affect the classification performance ($\pm$1%). The number of best individuals to keep in the elitist strategy selection, the proportion of crossover/mutation and the proportion of selection to create offspring are taken from the values recommended by Mitchell (2001).

4.4.1.2 Fitness value

In the GA, each genome (individual) is given a fitness value from a fitness function. The fitness function is crucial as it defines the selection preference for the entire GA. As one of the wrapper methods that uses the specific classifier as a black box, the fitness function of the GA reflects the performance of the classifier (Guyon and Elisseeff, 2003). In this study Cohen's kappa value (Cohen, 1960) was used as the parameter for evaluation of the classifier, as in Chapter 3, since it considers the agreement made by chance. Essentially, it is equivalent to the proportion of agreement when a dataset is balanced.

The performance of the classifier can vary depending on how the training set is drawn from

the GA training set. In order to improve the classifier's ability for generalisation, cross-validation is necessary. Before each run of the GA, the data were randomly split ten times into 70% training and 30% validation sets. The performance on each set was recorded as the kappa value comparing predictions and actual values of its testing set. The median of these ten kappa values was then recorded as the fitness value of the genome (Figure 4.6).

Since many feature subsets can have the same median value of kappa, an additional ranking strategy was applied to reduce the randomness introduced by multiple best feature subsets. (1) If the median value were the same, the number of features was compared, and the feature subset with smaller number of features would be ranked higher; (2) If the numbers of the features were the same again, then the mean values of the ten kappa values were compared.

4.4.1.3 Classifiers

Different classifiers were used in this study: linear classifiers (linear regression and linear SVM), and one non-linear classifier (RBF SVM).

(1) Linear regression*:* Linear regression is one of the simplest linear classifiers (Bishop, 2006). It is an approach to modelling the relationship between a scalar dependent variable *y* and one or more explanatory variables denoted *X*. Linear regression can be used to fit a predictive linear model to an observed data set of *y* and *X* values. After developing such a model, it can be used to make a prediction of the value of *y* given new *X* values. In the training set, adverse results are noted as 1 and normal results are noted as 0. A prediction is defined as a predicted adverse result if the output value of linear regression > 0.5; and it is defined as a predicted

normal result if the output value of linear regression ≤ 0.5.

(2) Linear SVM: The SVM is widely used in data analysis owing to its intuitive definition and simple implementation (Burges, 1998). The SVM constructs a hyperplane or a set of hyperplanes to separate data points, in order to achieve the largest distance to the nearest training points of any class. Intuitively, a larger functional margin means a lower generalisation error of the classifier. The detailed principles of the linear SVM can be found elsewhere (Cristianini, 2000). The method used here to find the separating hyperplane is the Least Squares method.

(3) RBF SVM: In this study, the Gaussian radial basis function (RBF) kernel was used as the non-linear kernel. Its feature space is a Hilbert space of infinite dimensions. Intuitively, the RBF SVM determines the category of a case based on its spatial distance to other points.

As the kernel of RBF SVM $K(\boldsymbol{x}, \boldsymbol{x}') = exp(-\gamma * \| \boldsymbol{x} - \boldsymbol{x}' \|^2)$, the choice of the gamma value (γ) in the kernel function of the RBF SVM was tested on the GA training set in terms of classification performance. 10-fold cross-validation was run on the GA training set and the mean of kappa was recorded. It was found that between 0.05 and 0.2 there was no clear difference in classification performance (kappa $\approx$ 0.49) (Figure 4.7), so the gamma value was set to 0.1. Linear SVM and RBF SVM algorithms were implemented using LIBSVM, details of which can be found elsewhere (Chang and Lin, 2011).

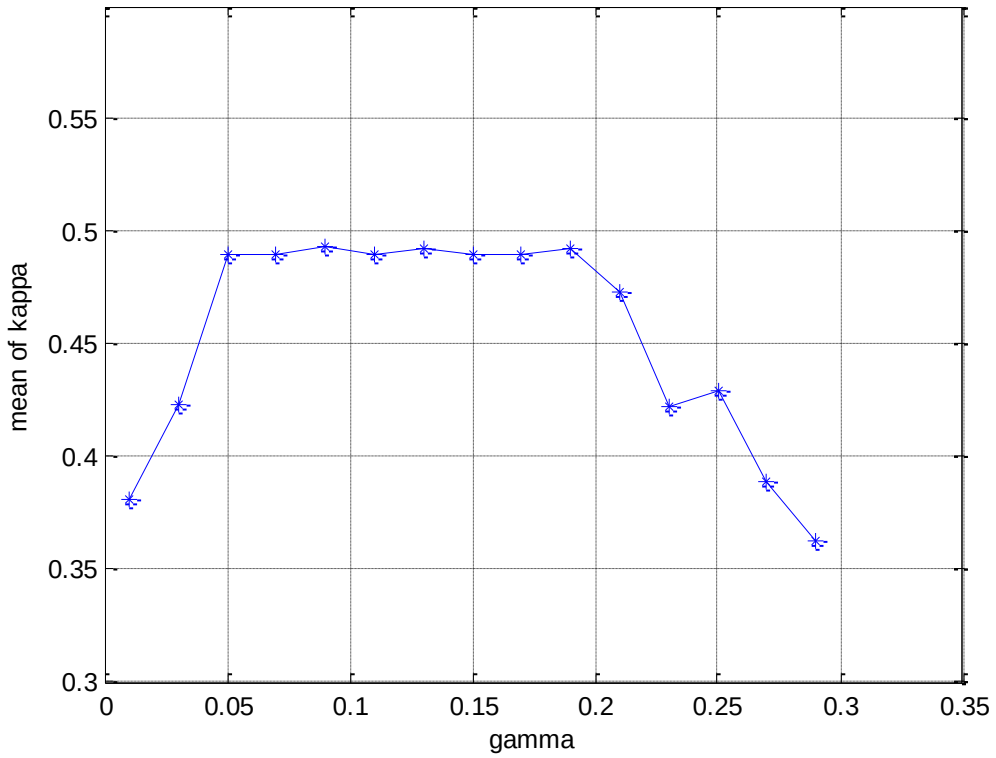


Figure 4.7 The relationship between gamma value and the mean of kappa value on a 10-fold cross validation on GA training set.

4.4.1.4 Stopping criteria and convergence of the algorithm

The decision on when to stop the GA is always about the balance of exploration of the feature space and computational cost. Specifically, in solving this FHR feature problem, the number of features and size of dataset is smaller than the high dimensional data problems. Therefore, it is more important to make sure that the highest ranking solution's fitness is reaching a plateau such that successive iterations no longer produce better results, i.e. the population has converged.

In order to find out the best indicator for the convergence of population in this problem, five indicators were studied: (1) the best fitness of the population; (2) the mean fitness of the

population; (3) the averaged mean fitness of the last five generations (the filtered fitness of the population); (4) the average L1 distance (one dimensional spatial distance) of all individuals; (5) the average L2 distance (standard deviation of the differences) of all individuals. Figure 4.8 shows a typical example of how these indicators change over the process of evolution.

It is found that the mean fitness of the population converges to the plateau value after a certain number of runs (the change from last generation to current generation < 1% of the change from the initial generation to current generation). The filtered mean fitness of the population changes more smoothly than the mean fitness of the population. The L1 distance of all individuals also converges with the best fitness of the population, but it seems to be converging to a certain level too early, even before the GA finds the best fitness point. The L2 distance of all individuals is not converging.

Therefore, the filtered mean fitness value was used in the stopping condition, since it intuitively reflects the convergence of the fitness value of the population. The filtering process (averaging) makes it smoother and thus more stable to evaluate. In addition, the L1 distance is also used in the stopping condition because it reflects the convergence of homogeneity of the population. Thus in this study, the algorithm terminated when one of the following conditions was reached:

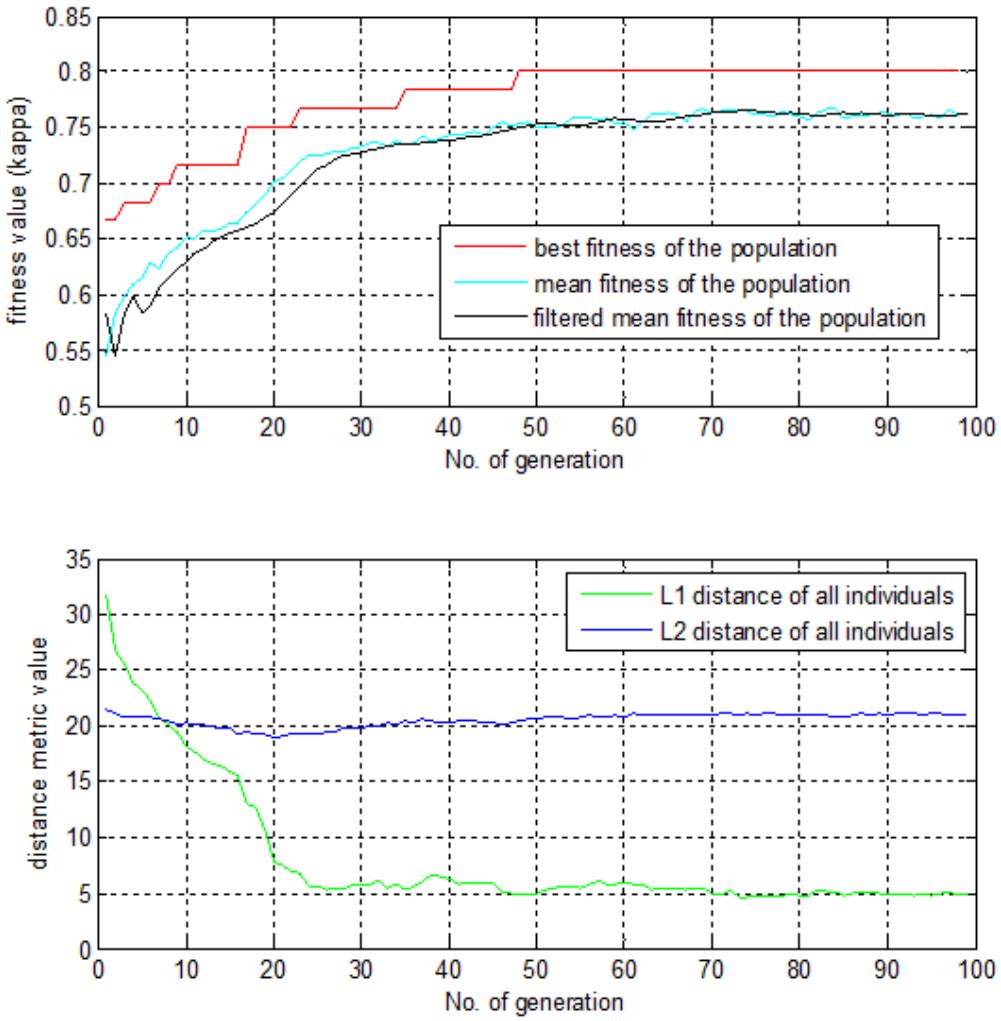


Figure 4.8 A comparison between different parameters in a single GA run.

(1) Given computational resource is exhausted, i.e. the maximum number of generations is reached or the computational time exceeds a maximum computational time.

(2) The filtered mean fitness value and the L1 distance are converging, i.e. they have not changed significantly for a certain number of generations (the change within a certain number of generations is smaller than 1% of the change from the first generation to current generation). In addition to these conditions, the best mean fitness value is no longer changing for a certain number of generations.

4.4.1.5 Controlling the number of features: regularisation using AIC and BIC

In machine learning, regularisation means introducing additional information criteria to assess and to compare the models, in order to prevent overfitting. When used in feature selection, regularisation is often introduced in a way to penalise for model complexity, such as penalty terms on the number of features. In this way, regularisation offers a relative measure for the goodness of fit, describing the trade-off between model fitness and model complexity. A regularisation method often involves a fitness measure and a complexity measure, thus it not only rewards goodness of model fitting, but also includes a penalty term which is an increasing function of the number of features.

The Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) are two commonly used regularisation methods (Bishop, 2006). For a feature subset, the AIC is:

$$AIC = 2k - 2ln(L)$$

where k is the number of features and L is the maximized value of the likelihood function ($L = R^2/N$,where R^2 is the sum of squared residue value and N is the number of cases) for the estimated model. The BIC is:

$$BIC = kln(n) - 2ln(L)$$

where k is the number of features, n is the number of cases and L is the maximized value of the likelihood function for the estimated model. Therefore, the fitness measure for AIC and BIC are similar, while BIC takes a higher penalty with increasing number of features.

The intuitive interpretation of AIC is an approach to maximise the entropy of the model, i.e. the average unpredictability in a random variable, which is equivalent to its information content. In practice, the AIC values of different models are denoted by AIC_1, AIC_2, AIC_3 ... AIC_n. Let AIC_{min} be the minimum of these values. Then $exp((AIC_{min} - AIC_i)/2)$ can be interpreted as the relative probability that the ith model minimizes the (estimated) information loss. The BIC is a method based on AIC using a Bayesian formalism.

Since AIC and BIC assess the performance of the classifier using the whole dataset instead of training-validation sets, the whole GA training set was used to calculate the log likelihood function L. Therefore, the fitness function value for a feature subset using AIC and BIC is the information criterion value of the classifier on the GA training set.

4.4.1.6 Validation strategies

Three strategies were used to evaluate the performance of the GA on the testing set:

(1) Use the best feature subset in each run of the GA and apply them independently on the testing set. The feature subsets are trained independently on the GA training set and tested on the testing set. The performance is reported here of the agreement between the prediction of these classifiers and the actual result.

(2) Use the best classifiers described above in (1) to form a voting committee. For each case in the testing set, the mode of the committee voting is taken as the prediction.

(3) Use the best classifier (that has the best fitness value) out of the 100 best classifiers, and

test it on the testing set.

For AIC or BIC, since there is only one best feature subset, the best feature subset was trained on the GA training set and tested on the testing set, and the classification performance was recorded.

4.4.2 Results

4.4.2.1 Classification performance

In each of the 100 runs of the GA, the best fitness value (on the GA training set) was recorded. These best fitness values from different classifiers are compared in Figure 4.9. It is found that the RBF SVM has slightly higher fitness values than the other two linear classifiers (Mann-Whitney U test, $p < 0.05$). This is expected since the RBF SVM used the distance from current point to normal/adverse groups instead of simple linear classification.

For the three validation strategies mentioned in Section 4.4.1.6, results are shown respectively as follows.

Strategy 1: The best classifier of each of the 100 runs of GA was applied independently on the testing set. The agreement of the output for each classifier was measured by a kappa value comparing the prediction of these classifiers and the actual result. The performances of different classifiers are compared in Figure 4.10. It is found that there is no significant difference between the three classifiers in terms of their classification performance on the testing set (Mann-Whitney U-test, $p > 0.5$), except that the RBF SVM seems to have a lower variance than the other two classifiers.

Strategy 2: The best classifiers described above were used to form a voting committee i.e. for each case the majority output (normal or adverse) of the different classifiers is taken as the output of the committee. The relationship between the number of committee members and the agreement between the output of the committee and the actual result are shown in Figure 4.11. It is found that the agreement becomes more stable with an increase in the number of committee members. The non-linear classifier still performs no better for this method. The output of all the 100 committees (agreement between the predicted and actual result) is 0.45 for linear regression, 0.47 with linear SVM, and 0.49 with RBF SVM (kappa values). It is also found that the committee output has converged and stabilized within 100 runs.

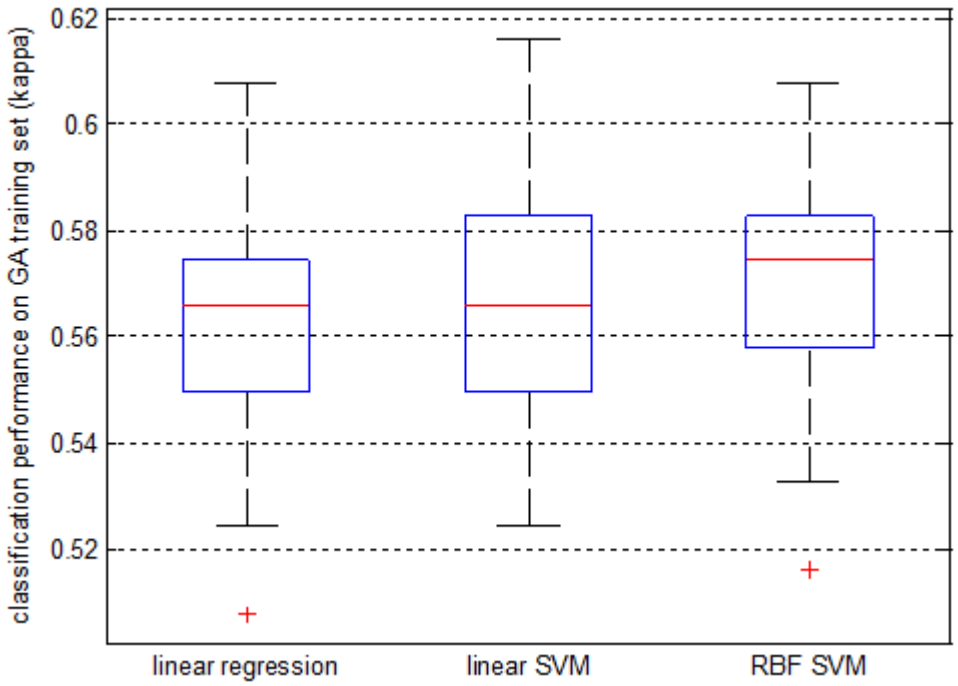


Figure 4.9 Fitness values for different classifiers over 100 runs on GA training set.

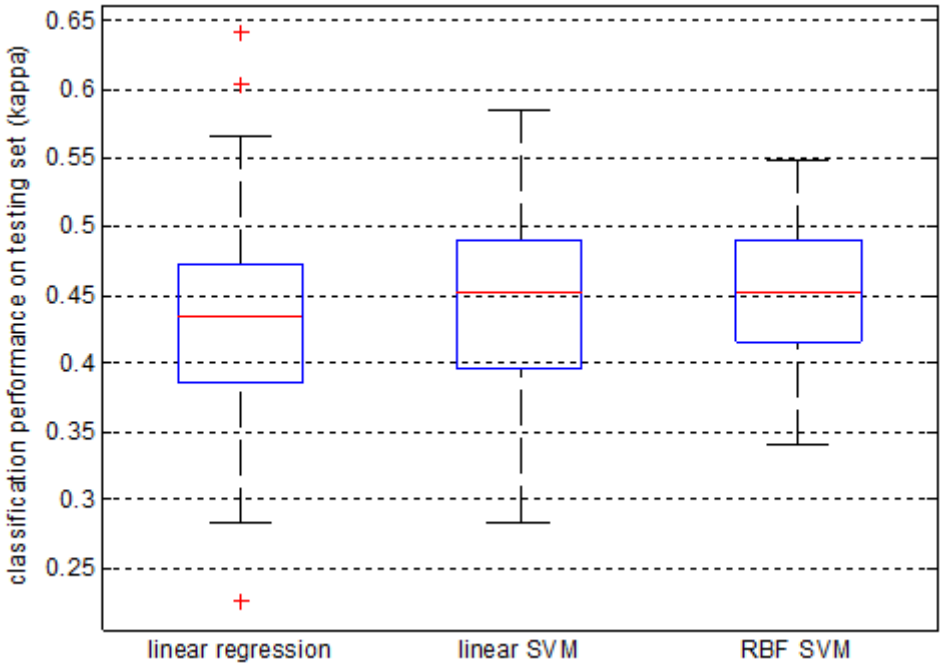


Figure 4.10 The classification performance for different classifiers over 100 runs on testing set.

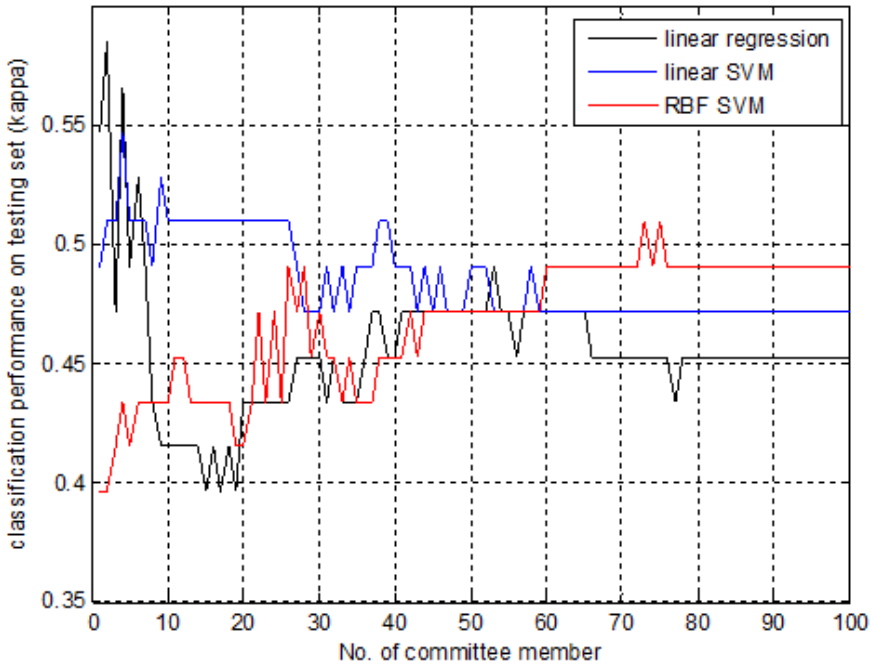


Figure 4.11 Relationship between the number of committee members and the agreement between committee output and actual result.

More statistical metrics are listed in Table 4.2. It is found that the RBF SVM has slightly better sensitivity and worse specificity than the other two classifiers. In terms of proportion of agreement, each of the three classifiers shows better performance using a selected feature subset than using all 64 features. The RBF SVM is slightly better than the other two classifiers using the (GA) selected feature subset. It is found that the classification performances of the GA are comparable to, or slightly better than, the performances of these classical feature selection methods.

The classification results were also compared with other feature selection methods such as random Forests (RF) (Breiman, 2001) and least absolute shrinkage and selection (LASSO) (Tibshirani, 1996) as shown in Table 4.2. It was found that the classification performances of the GA are comparable to, or slightly better than, the performances of these other feature selection methods.

Table 4.2 Classification performance (testing set) and comparison with other feature selection methods.

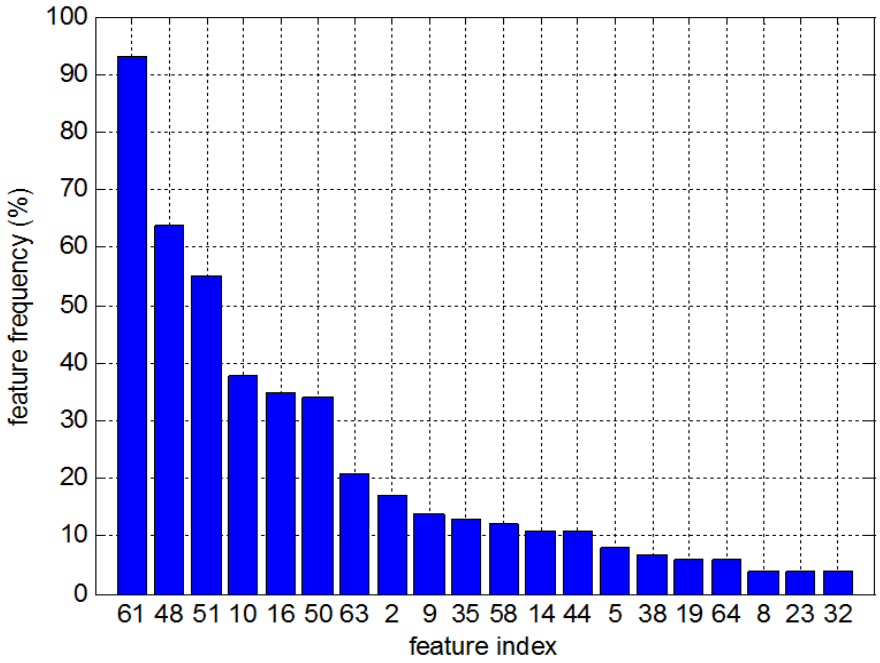
Method	***Sensitivity***	***Specificity***	***Kappa***	***Proportion of agreement***
Linear Regression	64.15%	81.13%	0.45	72.64%
Linear SVM	66.83%	81.13%	0.47	73.58%
RBF SVM	83.02%	66.03%	0.49	74.53%
RF	67.92%	77.36%	0.45	72.64%
LASSO	66.83%	78.25%	0.45	72.64%

Strategy 3: the best classifier (which has the best fitness value) out of the 100 best classifiers, was tested on the testing set. The agreement between the output of the classifier and the actual result are 0.42 for linear regression, 0.44 for linear SVM, and 0.45 with non-linear SVM. This is generally lower than those values obtained using committees.

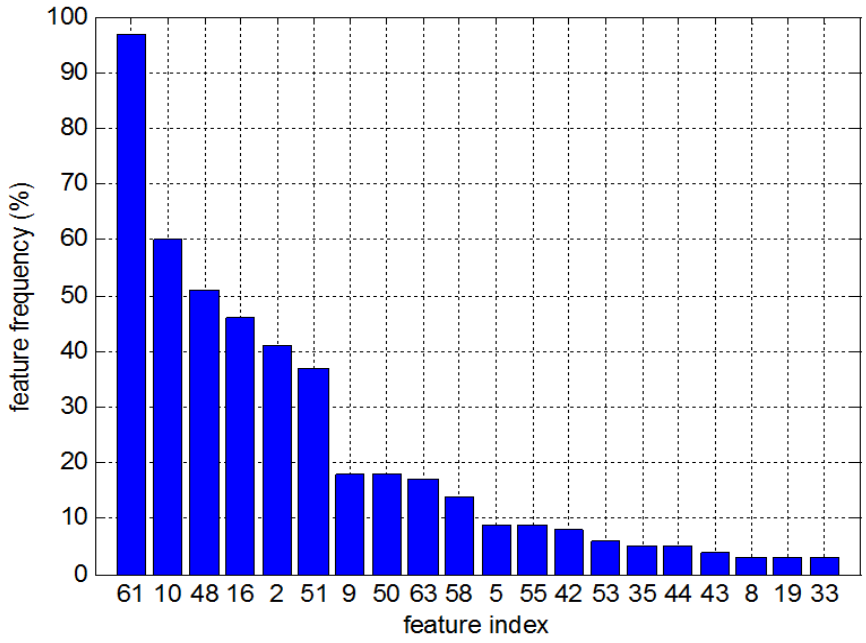
It should be noted that as there is only one value of classification performance for each classifier, it is hard to use statistical tests; thus it cannot be concluded whether the GA performs significantly better than the other methods. However, unlike the RF which uses an integrated classifier of all features, the GA gives a specific feature subset with the most important features, which is valuable for intuitive clinical interpretation.

4.4.2.2 Feature importance

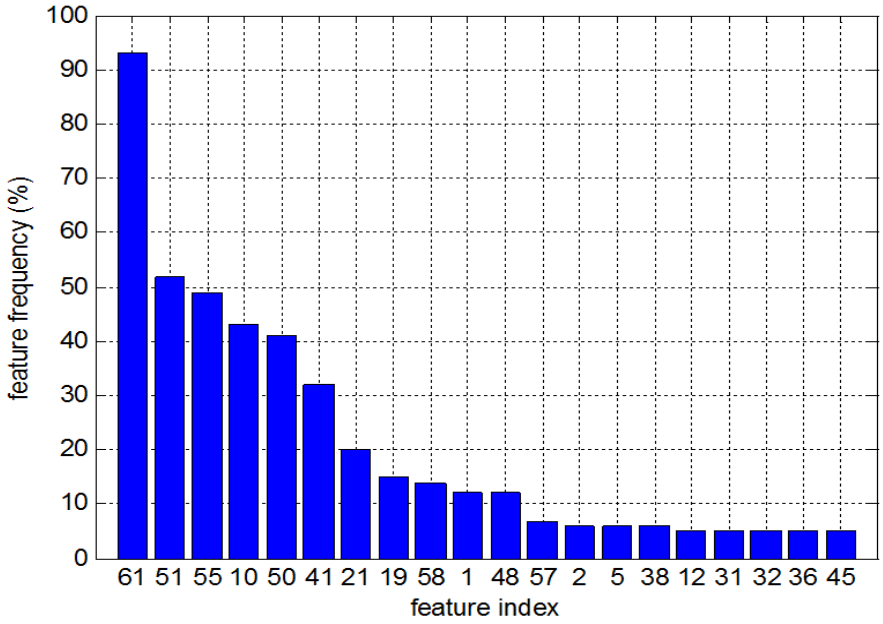
Since the GA was run 100 times with different splits of training-validation data, there are 100 sets of best features. The importance of each feature can be quantified by the frequency of the feature being selected. Figure 4.12 shows the feature importance ranking given by the different classifiers. The ranking of feature importance can be taken as a reference in selecting the best feature set. Similarities are found between the rankings for linear regression, linear SVM and non-linear SVM. For example, the most frequently selected feature (No.61: PRSA-DC) is the same for all three classifiers. This indicates that some features are important regardless of the classifier.


feature frequency (%)
100
90
80
70
60
50
40
30
20
10
0
61 48 51 10 16 50 63 2 9 35 58 14 44 5 38 19 64 8 23 32
feature index

(A)


feature frequency (%)
100
90
80
70
60
50
40
30
20
10
0
61 10 48 16 2 51 9 50 63 58 5 55 42 53 35 44 43 8 19 33
feature index

(B)



(C)

Figure 4.12 Feature frequency (top 20 frequently selected features) of: (A) Linear regression (B) Linear SVM (C) RBF SVM.

4.4.2.3 Best feature subsets selected by AIC and BIC

The best feature subsets selected by AIC and BIC are shown in Table 4.3. It is found that BIC generally performs better than AIC in selecting fewer features and has a better classification performance (kappa value). For linear regression and linear SVM, the classification performances of the BIC selected feature subsets are the same as those using committee output strategies (Section 4.4.2.1). This table shows that BIC can provide a fair classification performance while reducing the number of features. As for RBF SVM, it seems that neither AIC nor BIC has effectively penalised the number of features, since the maximum number of features is always the maximum number of features allowed in the algorithm (14). In addition,

the performances of the selected RBF classifiers are not as good as the linear classifiers.

Table 4.3 Best feature subsets selected by AIC and BIC.

Classifier		*Selected feature subset (displayed in feature index)*	*Kappa value on testing set*
Linear regression	AIC	5, 9, 10, 16, 38, 40, 41, 44, 48, 50, 51, 61, 63	0.42
	BIC	5, 9, 10, 48, 51, 61, 63	0.45
Linear SVM	AIC	4, 5, 9, 10, 16, 26, 28, 31, 38, 40, 41, 44, 48, 50, 51, 61, 63	0.41
	BIC	5, 9, 10, 16 , 44 , 48, 50, 51, 61	0.47
RBF SVM	AIC	1, 2, 3, 5, 16, 21, 22, 30, 40, 44, 45, 51, 60, 61	0.39
	BIC	1, 2, 4, 5, 13, 14, 15, 30, 44, 50, 51, 60, 61, 63	0.39

4.4.2.4 Consistency using different normal cases

The dataset consists of 255 adverse cases and 255 normal cases, in which the 255 normal cases were randomly selected from the 959 cases. To test the consistency of the GA method against different normal cases, another ten sets of 255 normal cases were drawn from the 959 cases. Linear regression, linear SVM and RBF SVM were tested respectively.

The best feature subset selected using BIC was consistent using these ten subsets. The classification performances of the 100 best classifiers on the testing set were not significantly different from each other on these ten subsets (Mann-Whitney U-test, $p > 0.6$). This indicates

that the GA method is consistent against different selection of normal cases.

4.5 Evaluation of Feature Selection Robustness Using Artificial Features

For any feature selection method, robustness is a critical performance indicator. It means that the method should perform well even if the features are contaminated. A typical way of evaluating the robustness of a method is to use datasets mixed with artificial features. In a previous study (Tuv et al., 2009), artificial features were used to remove irrelevant features. The idea is simple and intuitive: if a feature cannot perform as well as an artificially generated feature, then it is highly likely that the feature does not have good predictive power or that the classifier is having problems. Therefore, this section uses artificial features to assess the robustness of the GA method.

4.5.1 Methods

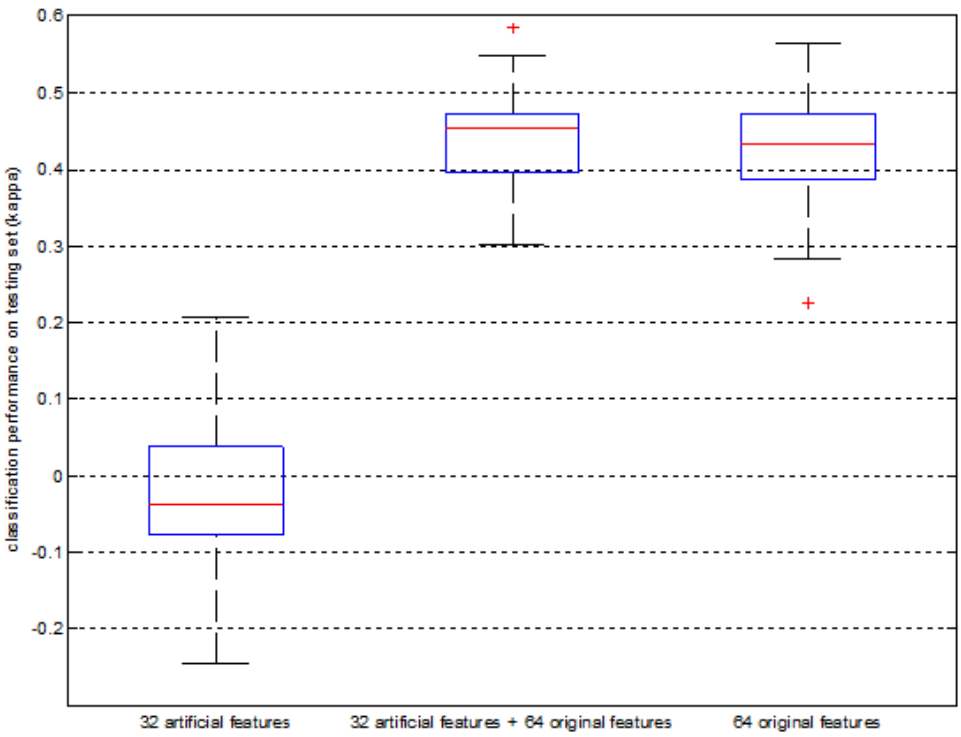
To assess the robustness of the method, 32 artificial features were generated, indexed as Feature 65-96. These features were randomly generated, normally distributed with zero mean and standard deviation equal to 1. Three datasets were then used in the GA: the 32 artificial features only, the 32 artificial features + the 64 original features, and the 64 original features only. Other properties of these datasets, including the separation of GA training set and GA testing set, are identically the same as described in Section 4.4.

The robustness of the method is evaluated in two ways: (1) the robustness of the feature selection process, i.e. the proportion of artificial features included in the best feature subsets;

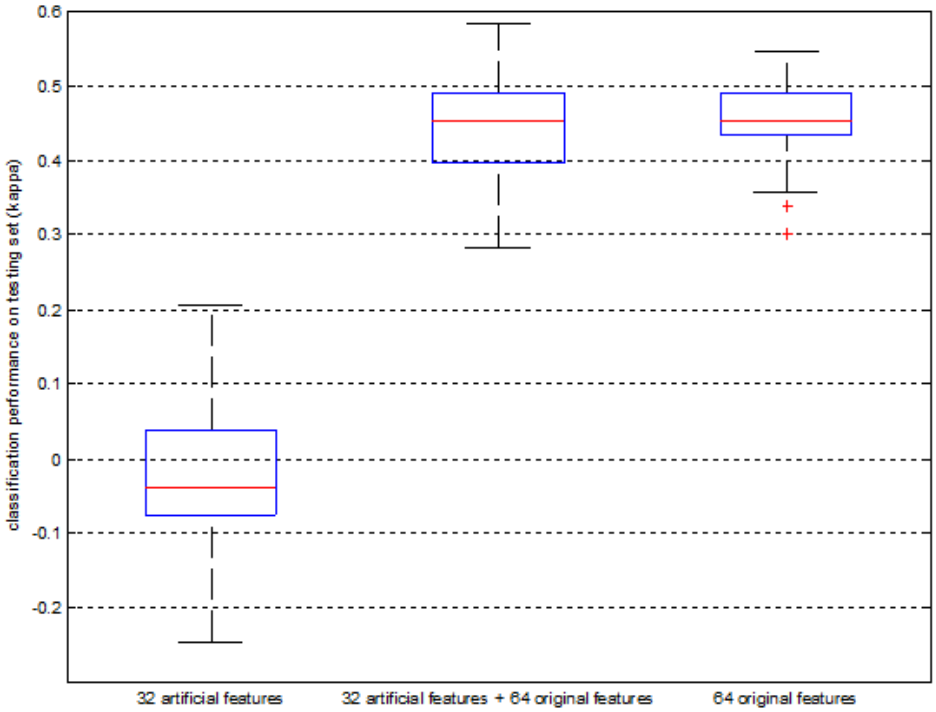
(2) the robustness of the classification performance, i.e. whether the classification performance will drop when artificial features are introduced.

4.5.2 Results

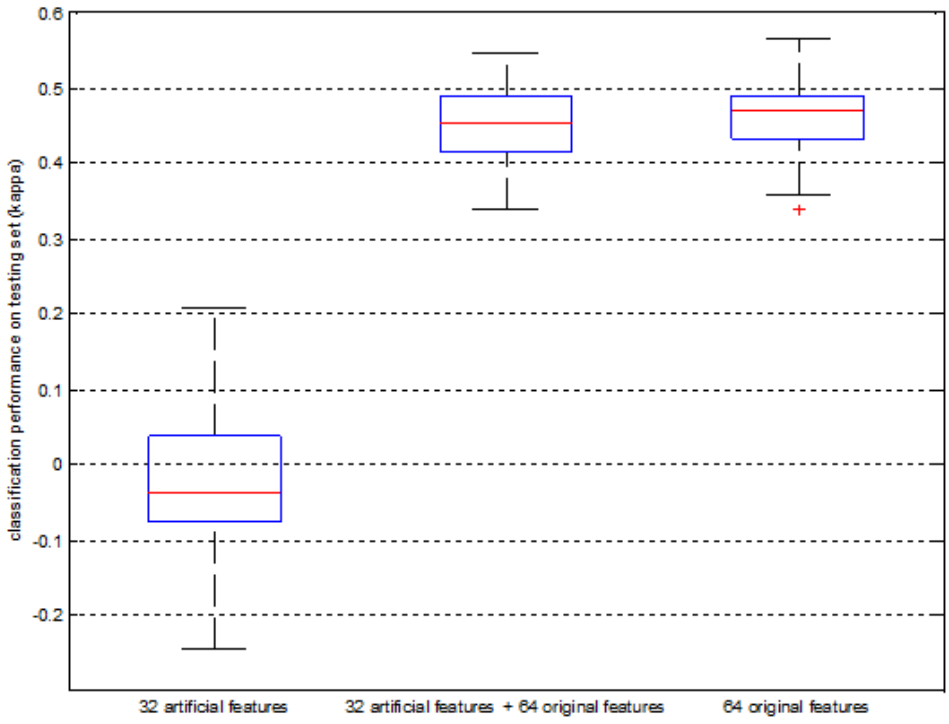
The classification performances for the testing set are compared in Figure 4.13. For all three methods, it is found that the artificial features have no predictive power. On the other hand, the combination of artificial features and original features does not significantly affect the prediction power of the linear classifiers, while the performance of the RBF SVM dropped significantly when adding artificial features (Mann-Whitney U test, $p<0.05$).



(A) Linear regression



(B) Linear SVM

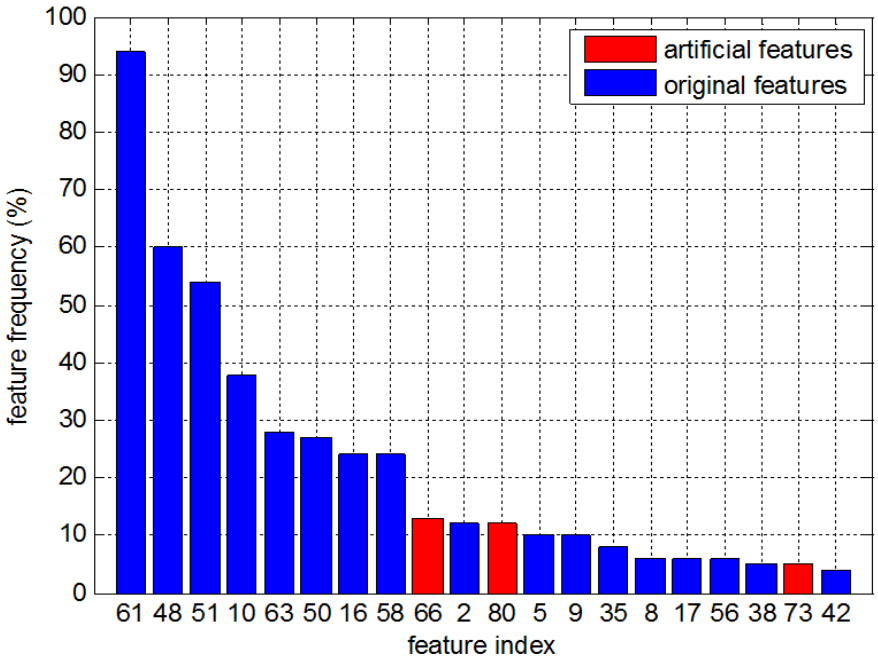


(C) RBF SVM

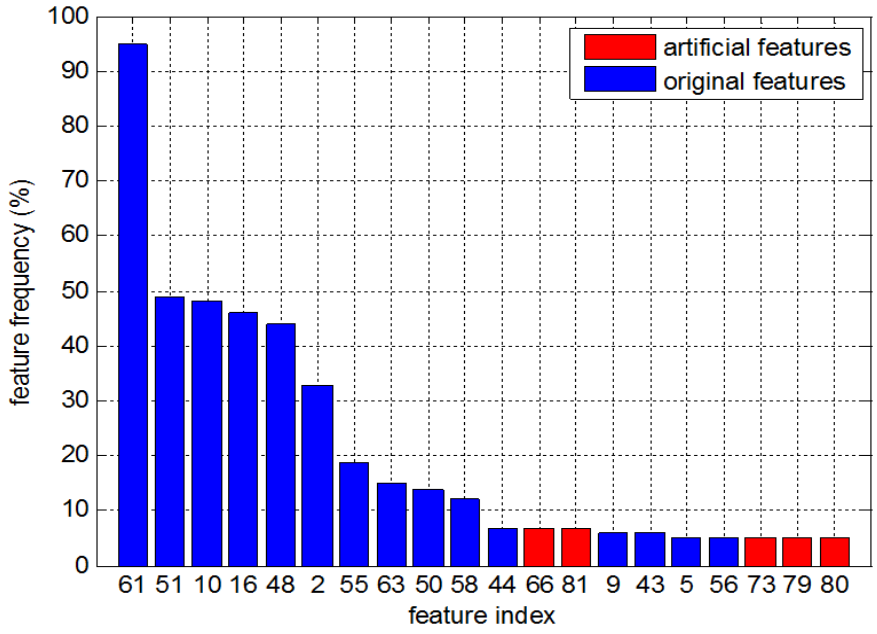
Figure 4.13 Classification performance (kappa) of: (A) Linear regression (B) Linear SVM (C) RBF

SVM on artificial features, original features and the mixture of the two.

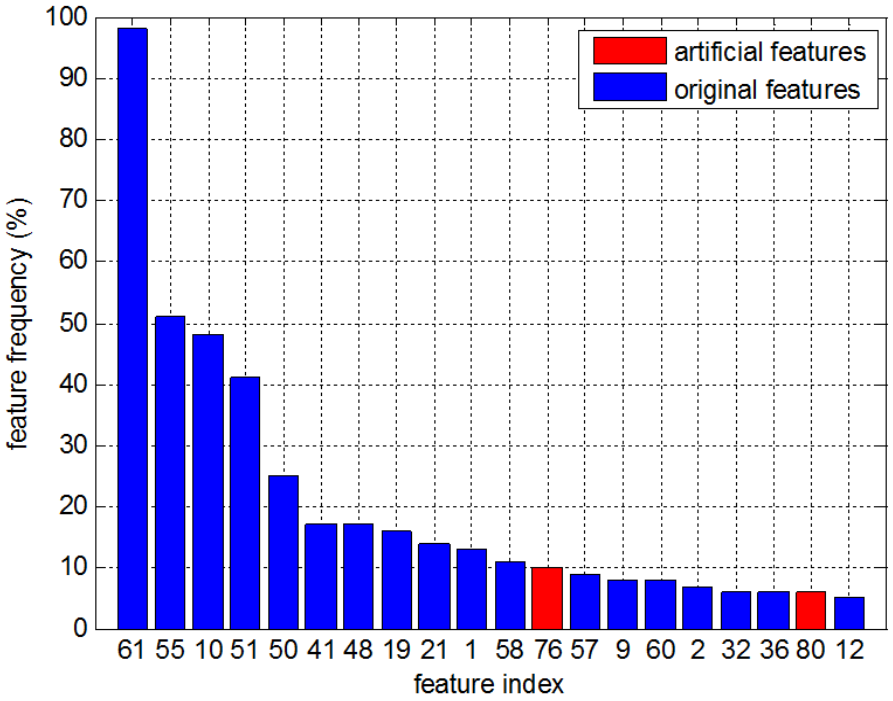
An important measurement of robustness is the proportion of artificial features being selected. The frequency with which each feature was selected is shown in Figure 4.14. For all three classifiers, it is found that some of the artificial features take a higher rank than some of the original features. This suggests that some of the original features do not have more significant predictive power than artificial features, i.e. their predictive power is likely to result from statistical randomness.



(A) Linear regression



(B) Linear SVM



(C) RBF SVM

Figure 4.14 Feature frequency (top 20 frequently selected features) of: (A) Linear regression; (B) Linear SVM; (C) RBF SVM on the mixture of actual features and artificial features.

The robustness of AIC and BIC was also tested with artificial features. It is found in Table 4.4 that AIC selected varying numbers of artificial features for the three classifiers, while BIC selected identical feature subsets with or without artificial features for the linear classifiers. On the other hand, BIC failed to retain its robustness when it comes to RBF SVM, where over 50% of the features in the feature subsets were selected as artificial features. To test the consistency of the artificial features, 10 more sets of 32 artificial features were generated and tested respectively, and the results are consistent.

Table 4.4 Best feature subsets selected by AIC and BIC, using 32 artificial features + 64 original features, artificial features is bold.

Classifier		***Selected feature subset (displayed in feature index)***	***Kappa value on testing set***
Linear regression	AIC	5, 9, 10, 16, 38, 40, 41, 50, 51, 61, 63, **66, 80, 85**	0.37
	BIC*	5, 9, 10, 48, 51, 61, 63	0.45
Linear SVM	AIC	4, 5, 9, 10, 16, 48, 50, 51, 61, 63, **85, 92, 93, 96**	0.32
	BIC*	5, 9, 10, 16, 44, 48, 50, 51, 61	0.47
RBF SVM	AIC	1, 9, 10, 16, 32, 36, 37, 60, 61, **65**, **70**, **72**, **73**, **85**, **86**, **90, 92, 93, 96**	0.35
	BIC	1, 3, 14, 21, 23, 37, 51, 61, **67, 68, 69, 75, 76, 80, 81, 83, 85, 89, 90, 96**	0.35

*: Methods that select identical feature subsets with or without artificial features.

It is found in this section that the GA with BIC is robust and effective for the two linear classifiers. However, when used with RBF SVM, less robustness is shown since the

performances of feature selection and classification both drop when artificial features are introduced. Therefore, another method has to be introduced to increase the robustness of the GA using the RBF SVM. In Section 4.6 a method is thus proposed to solve this problem.

4.6 Optimisation of Genetic Algorithms using Greedy Search Strategies

It was shown in Table 4.3 that the performance of the classifiers can give a kappa value of over 0.4 even when using fewer than 7 features. This reflects the relatively high predicting power of the best features, which indicates that it is possible to build a powerful classifier using just a small number of features. In fact, the GA with BIC using linear classifiers, which have been validated to be effective and robust, uses only 7-9 features. This indicates that there are more features than needed, thus classifiers that are more sensitive to variability such as the RBF SVM can be affected.

4.6.1 Utilisation of feature importance from GA

Feature importance can be calculated by many feature selection methods such as GAs, Random Forests and correlation-based filter methods (Guyon and Elisseeff, 2003). It is useful since it contains information about the relative importance of each feature in terms of certain feature selection methods. According to the feature importance, those less useful features can be eliminated to reduce overfitting.

An intuitive approach toward this is to start the feature selection process using these “best” features. This leads to greedy search strategies, such as Sequential Forward Selection (SFS)

and Sequential Backward Selection (SBS). These methods yield nested subsets of variables. This limits their ability to fully explore the feature space, but the most important features are kept during the moving process.

How the feature importance can be utilised in such greedy research strategies has been illustrated in several previous studies. In a study on correlation analysis of symptoms and disease, feature ranking generated by RFs was used for backward elimination: the best features were always kept during the process of SBS (Hu et al., 2009). Jiang et al. also used RFs to generate feature ranking, and the feature ranking was then used in the Sliding Window Sequential Forward Selection (SW-SFS) algorithm (Jiang et al., 2009).

In addition, as is mentioned in Section 2.4.4, greedy search strategies such as SFS and SBS can relieve the overfitting problems of GA. Therefore, it is possible that utilising the feature ranking from GA in greedy strategies will improve the performance of GA-only, SFS-only and SBS-only. To testify this hypothesis, the classification performances of GA + SFS and GA + SBS are compared with GA-only, SFS-only and SBS-only.

4.6.2 Methods

The SFS and SBS algorithms

An illustration of the SFS and SBS algorithms used here is shown in Figure 4.15. At the beginning of each algorithm, the GA training set was randomly split into a 10-fold training-validation set. Each feature subset was trained respectively on each of the ten training sets. The performance on each set was recorded as the kappa value comparing predictions and actual values for the validation set. The median of these ten kappa values was then recorded

as the classification performance of the feature subset. If the selection direction was forward, the algorithm started with the feature that performed best individually (SFS); otherwise the algorithm started with the whole feature subset (SBS).

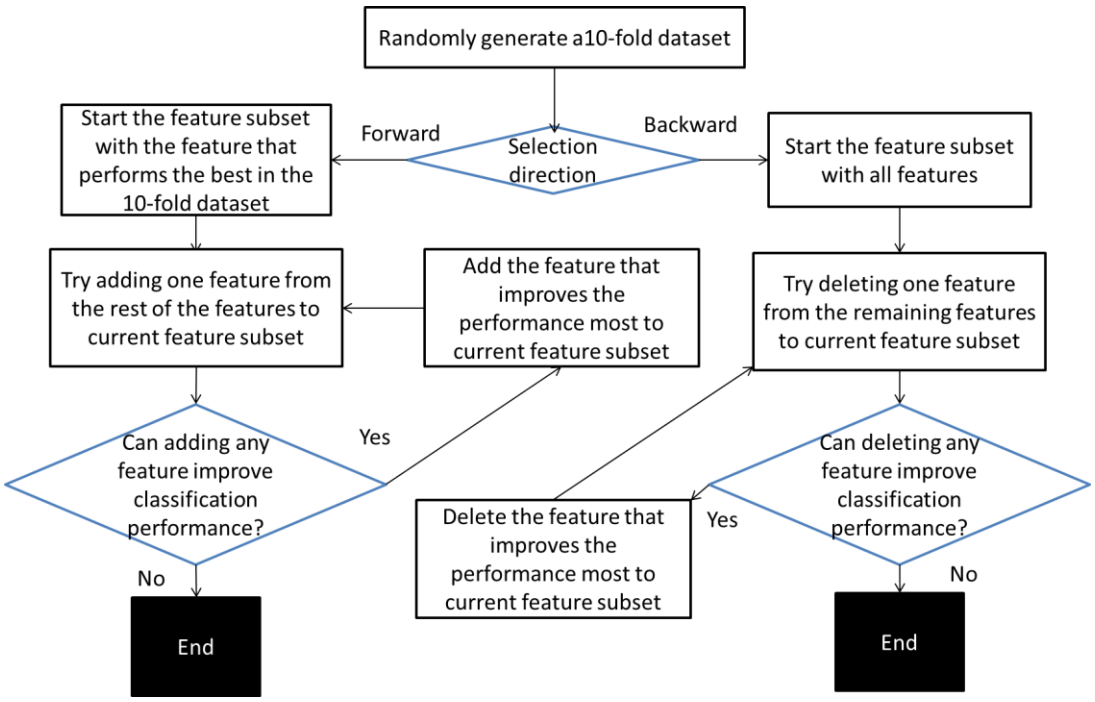


Figure 4.15 Illustration of the SFS and SBS algorithm

Choosing effective features from a GA

Since there are more features than needed, it is easy for SBS to stop at a complex model. As expected, when SBS was used alone, the output was a feature subset of 30 to 40 features, which exceeded the maximum number of features (14). To solve the premature stopping problem of SBS, the feature set had to be reduced to a group of features that have significant predictive power, these being defined as "effective features".

The feature frequency of a feature, i.e. the frequency of it being selected by the GA, reflects the predictive power of a feature using a specific classifier. If a feature could significantly

improve the classification performance when used with others, the feature will be selected effectively by the GA. In other words, if a feature is not often selected by the GA, it is not likely that this feature has significant predictive power when used with other features. Therefore, the feature frequency of GA can be used to determine which features are effective features.

It is found in Figure 4.14 that some of the artificial features have high ranking in the feature frequency table. These artificial features were introduced because wrappers like GAs are sensitive to training data. Therefore, the feature frequency of artificial features could reflect the level of a feature being selected by chance. If a feature has high predictive power, it should be selected significantly more often than artificial features. In that way, the effective features are defined as features with frequencies significantly higher than those of the artificial features.

To select the effective features, the GA was run 500 times using the 32 artificial features + 64 original features on the GA training set. The feature frequency was counted at each step, including previous runs. The selection record of each feature in each run was recorded ("0"for not being selected, "1" for being selected). The selection record of each feature was then compared with the selection record of the most frequently selected artificial feature. If the feature frequency of a feature was significantly higher than the most frequently selected artificial features of the 32 artificial features (two sample binomial test, $p < 0.05$), then it was defined as an effective feature. SFS and SBS only included these effective features. The methods were run 100 times each and the classification performances were recorded and

compared.

4.6.3 Results

Consistency of the effective features

It is assumed that when the number of samples is sufficiently large, they will be consistent against different initial conditions and different artificial feature. To test the consistency of the selection method against different initial conditions, the relationship between selected effective features and the number of GA runs was found as shown in Figure 4.16. It was found that the selection of effective feature stabilised after 150 runs. This indicates that the selection result is consistent when the number of GA runs is large enough.

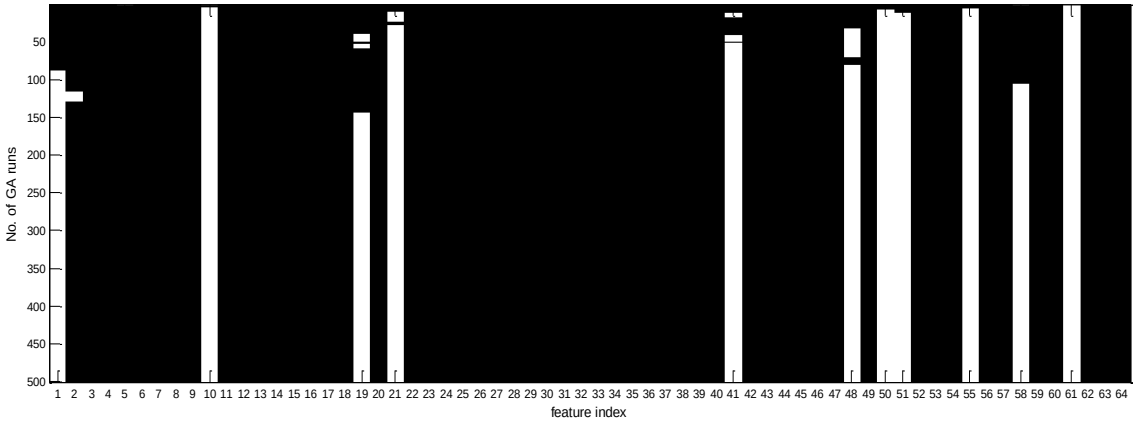


Figure 4.16 The relationship between effective features selected and the number of GA runs. A white grid indicates a specific feature being selected as an effective feature in specific GA runs.

To test the consistency of the selection method against different artificial features, another 10 artificial feature sets were randomly generated. After 500 runs, the selected effective features are identical with each other. The indexes of the selected effective features were: 1, 10, 19, 21,

41, 48, 50, 51, 55, 58, 61. This indicates that the selection method is consistent against different artificial features.

To test the robustness of the selection method, another 100 artificial features were introduced. The selection criterion is unchanged: if the feature frequency of a feature was significantly higher than the most frequently selected artificial features of the 32 artificial features (two sample binomial test, $p < 0.05$), then it was defined as an effective feature. The GA was run 500 times, and none of the artificial features was selected as an effective feature. This indicates that the selection method is robust against newly introduced artificial features.

Model selection

Five models were used: GA only (maximum number of features = 5), SFS-only, SBS-only, SFS using the GA selected effective features (GA + SFS) and SBS using the GA selected effective features (GA + SBS). SBS only was excluded from model selection since it selected excessively more features than the maximum number of features allowed. Therefore, for model selection, SFS only, GA + SFS and GA + SBS were compared. The classification performances of different methods using an RBF SVM are shown in Figure 4.17. It is found that GA + SBS has significantly higher classification performance than the other three methods, as a one-tailed t-test suggested ($p < 0.05$). Therefore GA + SBS was selected as the best method.

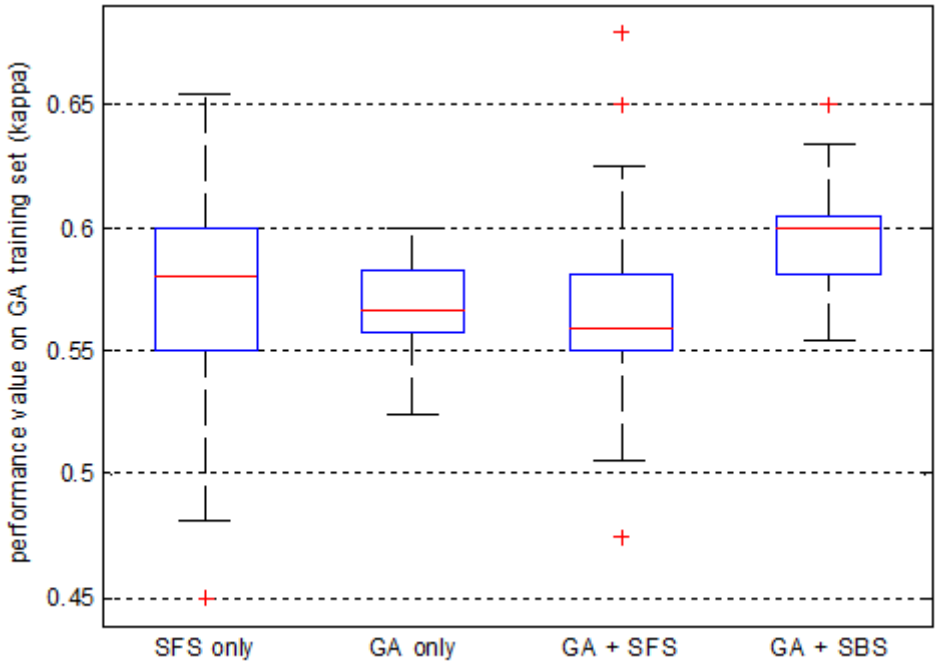


Figure 4.17 Classification performance (kappa value comparing the prediction output and actual result) of different methods on GA training set.

Validation on testing set

To validate the model selection, the classification performances of these models were tested on the testing set. It is shown in Figure 4.18 that the GA only outperforms SFS only, while the performance of GA + SFS is close to that of the GA. Moreover, the GA + SBS still has the best classification performance on the testing set, when tested by a one-tailed t-test with $p < 0.05$. This validates the hypothesis mentioned in Section 4.6.1 that using the feature ranking from GA in greedy strategies will improve the performance of GA, SFS and SBS, when used with the RBF SVM classifier.

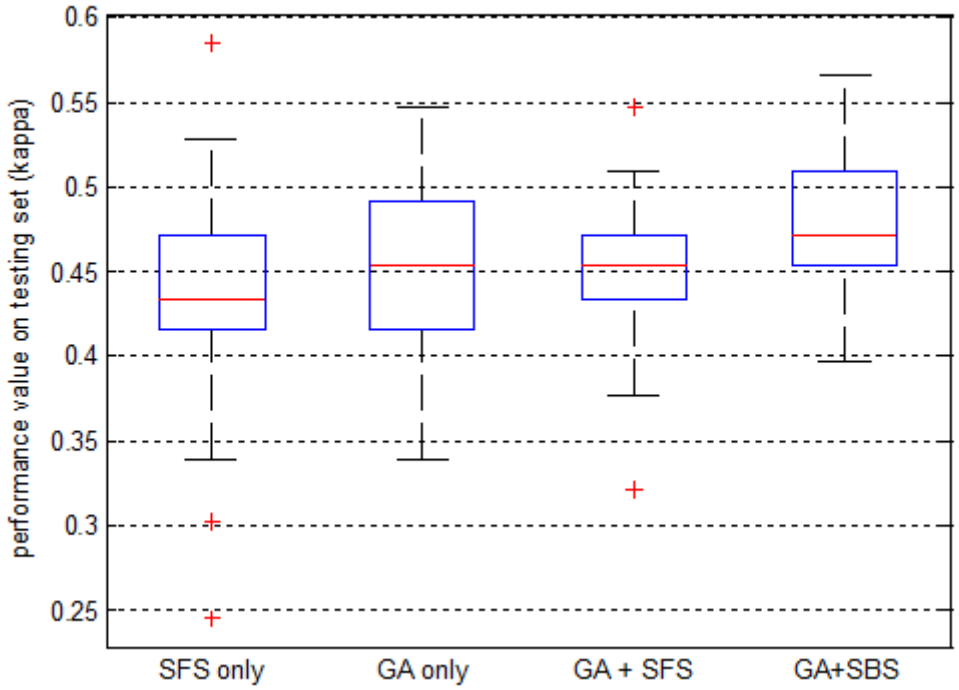


Figure 4.18 Classification performance (kappa value comparing the prediction output and actual result) of different methods on testing set.

Figure 4.19 shows the distribution of number of selected features in each feature subset using different classifiers. It is found that with a GA, over 80% of subsets select five features. GA + SBS generally selects more features than the other three methods, but is effective against overfitting and robust against artificial features. This indicates that GA + SBS provides a more effective and robust way of solving the overfitting problem.

Selection of the best feature subset

To select the best feature subset of GA + SBS using an RBF SVM, the algorithm was run 1,000 times when the feature frequency and best feature subset were stable. The feature frequency table is shown in Figure 4.20. The frequency of the most frequently selected feature

subset is listed in Table 4.5. It is found that the feature subset 1, 10, 19, 21, 48, 50, 51, 55, 61 is most frequently selected, thus it is selected as the best feature subset of GA + SBS.

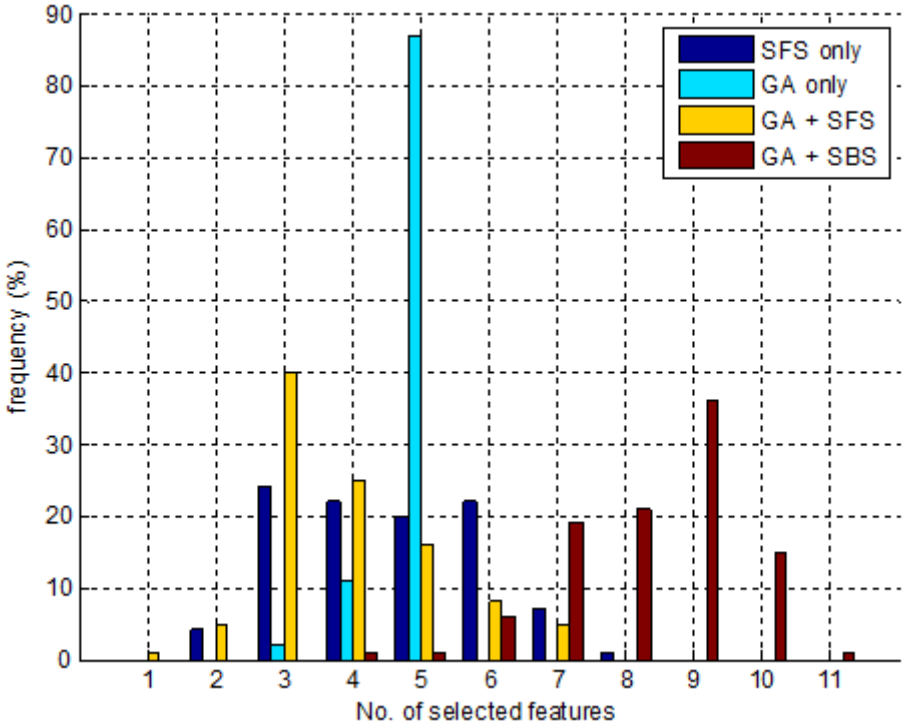


Figure 4.19 Histogram of number of selected features in each feature subset using different methods.

Table 4.5 Feature frequency of the 11 effective features in GA + SBS.

Selected feature subset (displayed in feature index)	*No. of features*	*Feature frequency in 1,000 runs*
1, 10, 19, 21, 48, 50, 51, 55, 61	9	89
10, 19, 21, 41, 48, 50, 51, 55, 58, 61	10	64
10, 19, 21, 41, 50, 51, 55, 58, 61	9	47
1, 10, 19, 21, 48, 50, 51, 58, 61	9	45
10, 19, 21, 48, 50, 51, 55, 61	8	32

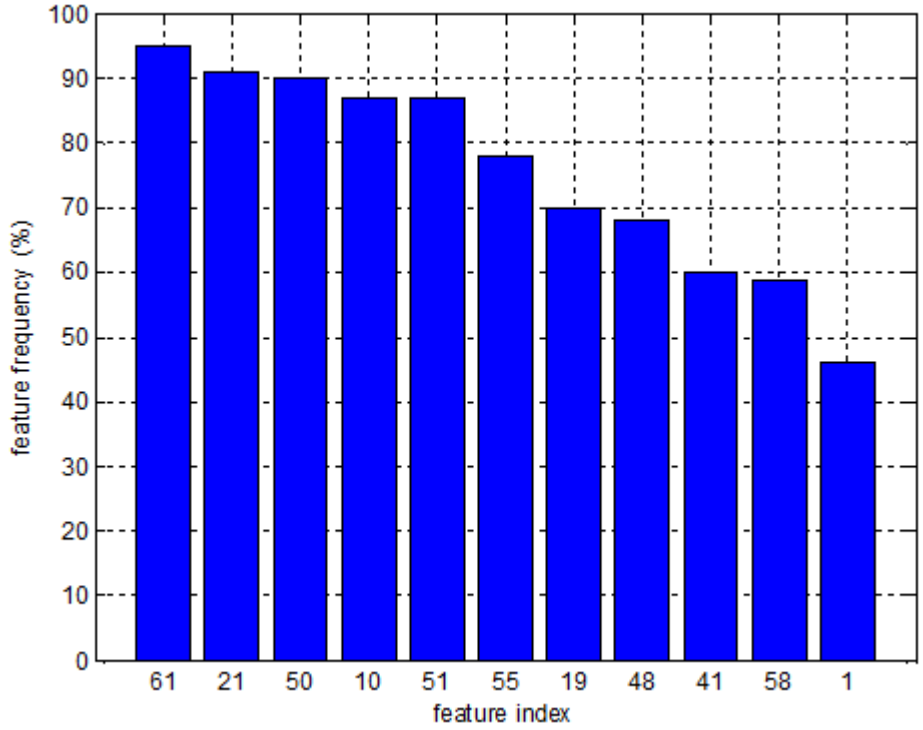


Figure 4.20 Best feature subset for different classifiers.

4.7 Evaluation of Diagnostic Power on 7,568 Cases

To evaluate the performance of the GA, it was tested on unseen data in two ways: firstly, it was tested on the balanced testing set, reporting kappa values, sensitivity and specificity, as described in Section 4.4.2; secondly, the diagnostic power of the GA selected classifier was evaluated over all of the 7,568 cases.

4.7.1 Methods

The classifier output is defined as the classifier prediction value of the best feature subset in each classifier. For linear regression and the linear SVM, the best feature subsets are those

selected using a GA with BIC. For the RBF SVM, the best feature subset is that selected using GA + SBS (Table 4.6). The classifier prediction value for linear regression is defined as the regressed prediction value. For Linear SVM and RBF SVM, the prediction value is the estimated probability of the case being as adverse outcome (Chang and Lin, 2011).

Table 4.6 Best feature subset for different classifiers

Classifier	*Selected feature subset (displayed in feature index)*	*No. of features*	*Kappa on testing set*	*Method*
Linear Regression	5, 9, 10, 48, 51, 61, 63	7	0.45	GA + BIC
Linear SVM	5, 9, 10, 16, 44, 48, 50, 51, 61	9	0.47	GA + BIC
RBF SVM	1, 10, 19, 21, 48, 50, 51, 55, 61	9	0.49	GA + SBS

Table 4.7 List of features that are selected by GA.

Feature Number	*Feature Name*
1	Baseline mean
5	Lowest dip of fetal heart rate
9	SSI of the residual signal
10	Short Term Variability (STV)
16	Median of contraction duration
19	STV (accelerations included)
21	Mean Acceleration Duration

44	Number of variable decelerations
48	Mutual information
50	Mean of local approximate entropy
51	Standard deviation (STD) of local sample entropy
55	(STD/mean)^2
61	Phase Rectified Signal Averaging (PRSA), Deceleration Capacity (DC)
63	BPRSA – DC component

To evaluate the diagnostic power of the GA selected feature subset on the whole dataset of 7,568 cases, the unbalanced nature of the dataset has to be considered. Adverse outcome cases are so rare (252 out of 7,568, 3.37%) that standard measures such as kappa value, sensitivity and specificity could disguise the predictive power of any diagnostic tool (Symonds et al., 2001). Therefore, the ROC curve, and the Event Rate Estimation (EveREst) plot was used for diagnostic power evaluation. An EveREst plot provides a better evaluation of the performance of a diagnostic classifier on a large scale over the total population of patients. In addition, it allows predictive power evaluation that is independent of thresholding points to classify the classifier output (Georgieva et al., 2013c).

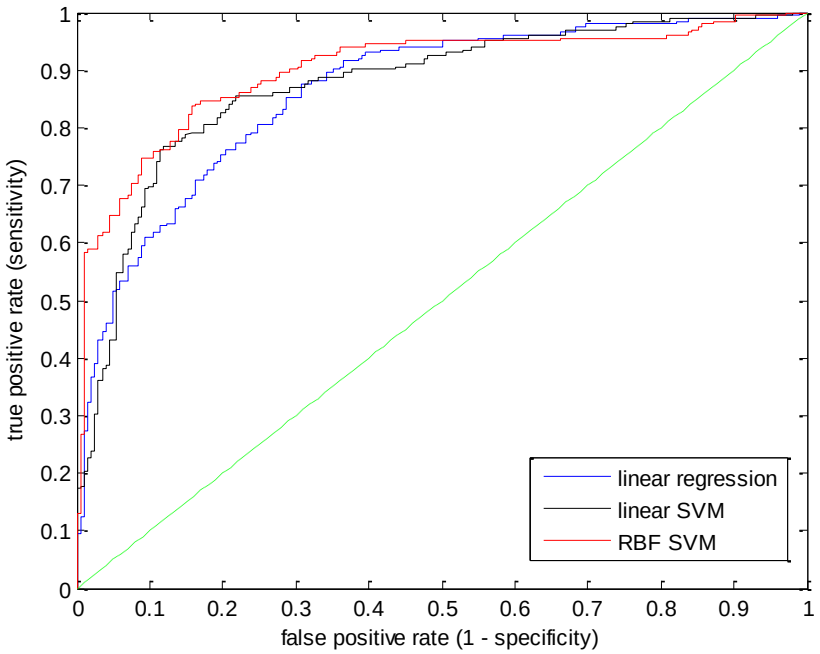
4.7.2 Results

ROC curves

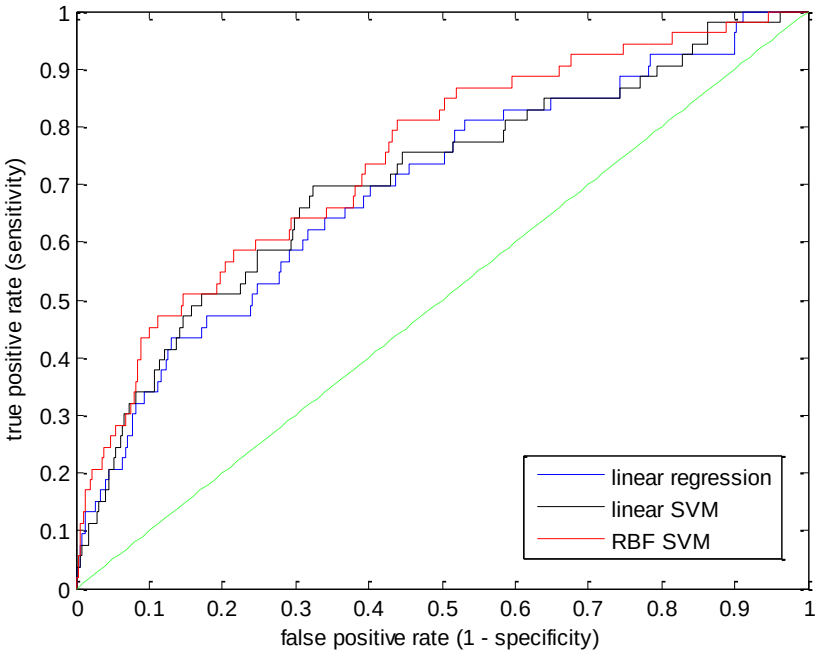
The ROC curves of different classifiers on GA training set and testing set are shown in Figure

4.21. It is found that for both ROC curves, the RBF SVM performs generally better than the other two classifiers. The performances of linear regression and the linear SVM are similar to each other. The Area Under the Curve (AUC) values were also calculated for each classifier and are displayed in Table 4.8. It was found that the RBF SVM also outperforms other two classifiers in terms of AUC values. Note that when excluding the GA training set, the dataset becomes heavily imbalanced, causing the resolution of the ROC curve to be lower.

To determine how sensitive the AUC (on the testing set) is to the selection of held-out test data, the AUC were calculated using another 100 separated GA-training sets and testing sets (80%-20%), in each of which the held-out test data are selected randomly. It was found that the AUC value of RBF SVM (0.742 ± 0.041) is significantly higher than that of linear regression (0.672 ± 0.036) and linear SVM (0.709 ± 0.044) (Mann-Whitney Utest, $p < 0.01$).



(A) GA training set



(B) Testing set

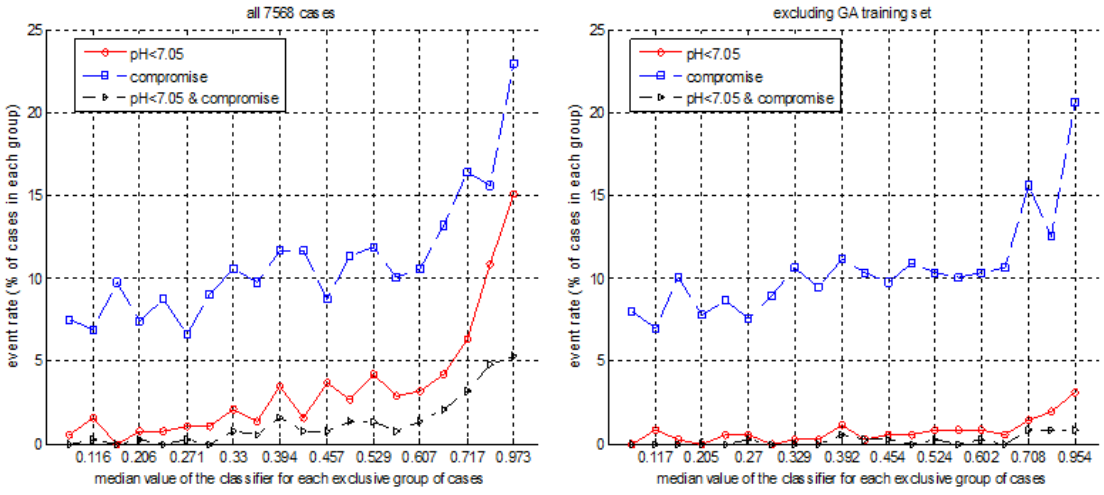
Figure 4.21 ROC curve of different classifiers for: (A) GA training set; (B) testing set.

Table 4.8 The AUC value for different classifiers.

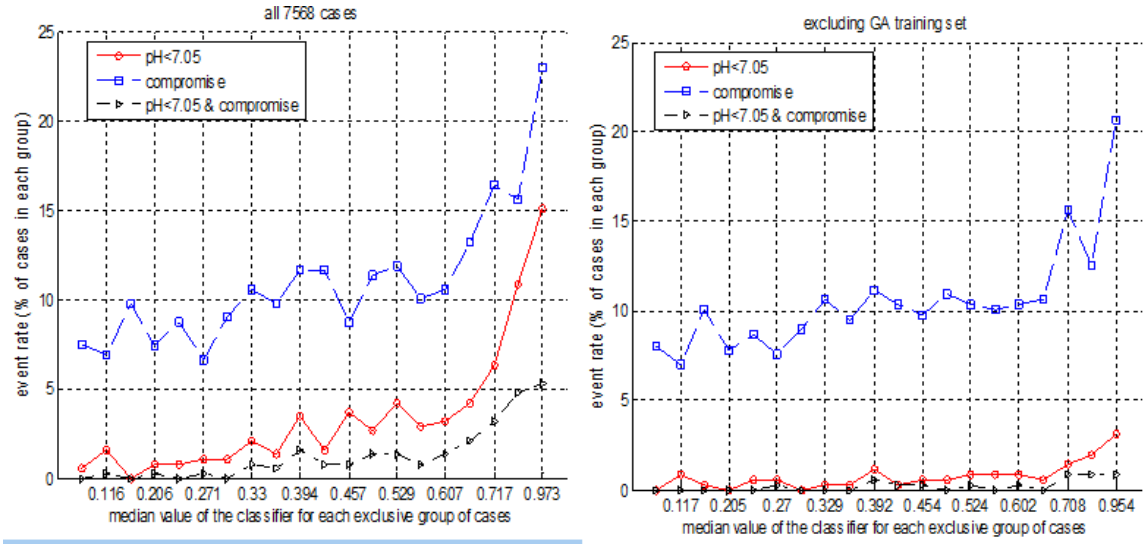
Classifier	*GA training set*	*Testing set*
Linear Regression	0.8026	0.6879
Linear SVM	0.8214	0.7103
RBF SVM	0.8623	0.7445

EveREst plots

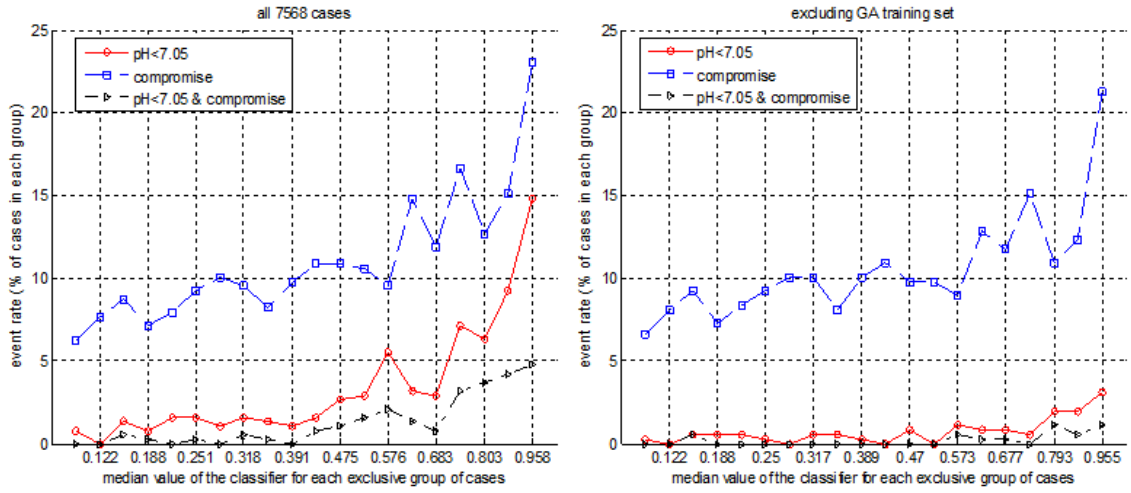
The EveREst plots of different classifiers are shown in Figure 4.22. The classifier output values for all patients are sorted in ascending order and grouped into 20 groups, each containing 5% of the cases. The x-axis is the median of each group, while the y-axis is the proportion of three types of labour outcome occurs in each group (low pH, compromise and the combination of the two). Therefore, a favourable predictor will have a zero event rate at the left beginning and have a 100% event rate at the right end. The results on two groups are represented: all 7,568 cases and all cases excluding the GA training set where the classifiers were trained (7,164 cases).



(A) Linear regression



(B) Linear SVM



(C) RBF SVM

Figure 4.22 EveREst for three events: pH < 7.05, compromise defined by clinical symptoms, and adverse outcome (combination of the two): (A) Linear regression; (B) Linear SVM; (C) RBF SVM.

The EveREst plots of the different classifiers show that event rates generally increase consistently with an increase in classifier output, whether on all 7,568 cases or on the unseen 7,164 cases. For instance, on all 7,568 cases, the event rate of low pH increases from less than 1% (in the lowest 5th centile of the classifier output) to above 10% (the lowest 95th centile of the classifier output), while the event rate of labour compromise increases from less than 8%

(in the lowest 5th centile of the classifier output) to above 15% (the highest 95th centile of the classifier output). The event rate is generally lower on unseen data, since there are only 20% of adverse cases in the unseen data. Taking this into account, the increasing trend on unseen dataset is consistent with that on all 7,568 cases. As for the unseen data, the increasing risk of low pH has a more rapid trend for the RBF SVM, while the increasing risk of compromise has a more rapid trend using the linear classifiers.

4.8 Discussion and Conclusions

The objective of this chapter was to find a best FHR feature subset using feature selection methods. Genetic Algorithms were used to select feature subsets based on three different classifiers. To the author's knowledge, this is the first time that a feature selection method has been applied on such a large scale FHR dataset (7,568 cases).

The univariate prediction power of the 64 features was evaluated individually. It was shown in Table 4.1 that 49 of 64 features have statistical significance in the normal and adverse group. However, this does not guarantee predictive power. In fact, Table 4.1 and Figure 4.4 show that the correlation between the feature values and labour outcome is rather weak. Therefore it is necessary to combine the predictive power of different features using feature selection and integration methods.

Clear and intuitive clinical interpretation is needed to assist the clinicians in decision making, thus the GA was chosen for its ability to explore the whole feature space and to give a clear best feature subset. In addition, linear regression, linear SVMs and RBF SVMs were chosen

as the classifiers for the GA. The linear classifiers were used to investigate the linear relationship between features and adverse outcome, while the RBF SVM used the spatial distance between a certain case and different outcome groups to determine to which group it is most likely each case belongs.

The whole set data of 7,568 cases is heavily imbalanced (only 3.37%, 255 cases are adverse outcomes). A balanced dataset of 510 cases (255 normal and 255 adverse) was created, where the GA was trained using 404 cases, and tested on 106 cases. Given the lack of a gold standard and the limited ability to predict labour outcome with expert knowledge (Grimes and Peipert, 2010), the classification performance on the testing set is promising (kappa value of 0.45 to 0.49 for different classifiers). This classification performance is better than previous studies using Artificial Neural Networks on similar dataset, with a kappa value of 0.28 on the testing set (Georgieva et al., 2013c). It is also comparable to other feature selection methods using the same dataset (Table 4.2).

In Table 4.1 It should be noted that some of the features have much higher classification performance than others when used individually, thus it is possible that the classification performance is just a statistical coincidence. In addition, the GA, as a fitness-based global optimisation tool, is prone to overfitting on this dataset, especially when used with a distance-based classifier such as the RBF. Methods such as regularisation and greedy strategies were used to prevent overfitting. These methods have improved the effectiveness and robustness of the GA, as is shown in Section 4.6.

Each classifier selected 7 to 9 features using the GA, the features selected by all three

classifiers being: Feature 10 (short term variability) (Dawes et al., 1992, Cazares, 2002), Feature 48 (Mutual information), Feature 51 (STD of local Sample Entropy) (Fulcher et al., 2012) and Feature 61 (Phase Rectified Signal Averaging – DC component) (Georgieva et al., 2013b). Therefore, these features appear to be useful in predicting labour outcome when used in multivariate analysis. This also indicates that these features appear to be useful in predicting labour outcome when integrated with others.

The top-ranked feature for all three classifiers was Feature 61 (PRSA-DC component). PRSA is a relatively new time series method that measures heart rate variability, and also quantifies separately the acceleration capacity (AC) and deceleration capacity (DC) (Bauer et al., 2006). In a previous study using low arterial pH as adverse outcome, it was found that the DC component compared well with or outperformed other fetal heart analysis measures such as short term variability, approximate entropy and signal stability index, the clinical interpretation of which is that the initial response to acute experimental hypoxemia or repeated asphyxia in the term fetus is an increase in FHR variability (Georgieva et al., 2013b). Also, Feature 10 (short term variability), Feature 48 (Auto mutual information) and Feature 51 (STD of local Sample Entropy) were selected using all three classifiers, all of which are different aspects of variability. This suggests that the FHR variability, as suggested clinically, has an important relationship with low arterial pH.

To evaluate the diagnostic predictive power of the GA selected classifiers, the classifier output was tested on all 7,568 cases using EveREst plots. It was found that the event rate of low pH and labour compromise both increase consistently with the increase in classifier output. This

indicates that fetal monitoring in labour can benefit from GA selected classifiers which provide objective measures of the risk of acidosis and/or labour compromise. However, to translate this work into clinical practice, extensive future work will be necessary.

It should be noted that this study takes only the last 4 windows of 15-minute length in the last 30 minutes of Stage 2. To provide a better performance, more information during the process of labour, especially time-series information, should be integrated into the classifier. There are also some additional clinical parameters that need to be studied, such as maternal infection and oxytocin augmentation (Georgieva et al., 2013c). In addition, Apgar scores should also be investigated as an indicator of labour outcome.

Clinical interpretation of the classifiers should also be considered, including the analysis of the feature selected, e.g. the clinical interpretation of the best features and the correlation between the clinical features and the statistical features. Further studies will be carried out to estimate the risk of compromise, based on the classifier prediction and its patient specific time-series trend. The next step is to apply the classifiers throughout the duration of labour, which will provide an online objective measurement of fetal health condition during different stages of labour. Further studies will be carried out to estimate the risk of compromise, based on the classifier prediction and its patient specific time-series trend.

5 Robustness of the FHR Features with respect to Signal Loss

5.1 Introduction

At present, the OxSys system contains 64 FHR features (Fulcher et al., 2012, Xu et al., 2012, Georgieva et al., 2011a, Georgieva et al., 2013c). These features reflect different aspects of FHR, and they can be used for classification of labour outcome (see Chapter 4). Signal loss in the FHR can occur for various reasons. For example, the mother is temporarily detached from the monitor; or the transducers are not getting proper contact during contractions and maternal movements; or the fetus changed his/her position. Therefore, signal loss appears in segments of different lengths, with the proportion of signal loss in a 15-minute FHR window ranging from 0% to 100%.

It is important to evaluate the effect of signal loss on the calculation of FHR features and the subsequent classification of traces. The robustness of the FHR features against signal loss has an intuitive interpretation in clinical practice: it gives a measurement of the credibility that FHR features can be trusted under a certain signal quality condition.

To the author's knowledge, the influence of signal loss on computerised FHR analysis has not been investigated before, except for one study (Spencer et al., 1987). In this study two groups of ultrasound derived FHRs were studied (10 and 13 patients respectively with different FHR signal quality: < 2% and from 2% to 33.1%). FHR derived using fetal scalp electrode (fetal ECG) was simultaneously used as the gold standard to measure the "actual FHR" under signal

loss. Statistical tests showed that for the group with more signal loss, the mean FHR variation measured by ultrasound was significantly lower than that measured by fetal ECG (t-test, $p < 0.005$); while for the first group (less signal loss), the mean FHR variation was not statistically different when ultrasound and fetal ECG were compared. It should be noted that this study measured only one FHR feature with a small number of patients.

Amongst the 3,953,878 FHR windows in the 49,151 cases of the OIFD1 database, 90.5% of the windows have some signal loss. The distribution of signal quality in the 3,953,878 FHR windows is shown in Figure 5.1. It is shown that the majority (over 80%) of signal quality lies in the range of 60%-100%. Therefore, it is worth investigating the influence of signal loss on FHR features, since it is likely that under poor signal quality conditions, the values of the features can change and this would affect the classification of labour outcome. Therefore, the aim of this chapter is to investigate the robustness of the FHR features with respect to signal loss.

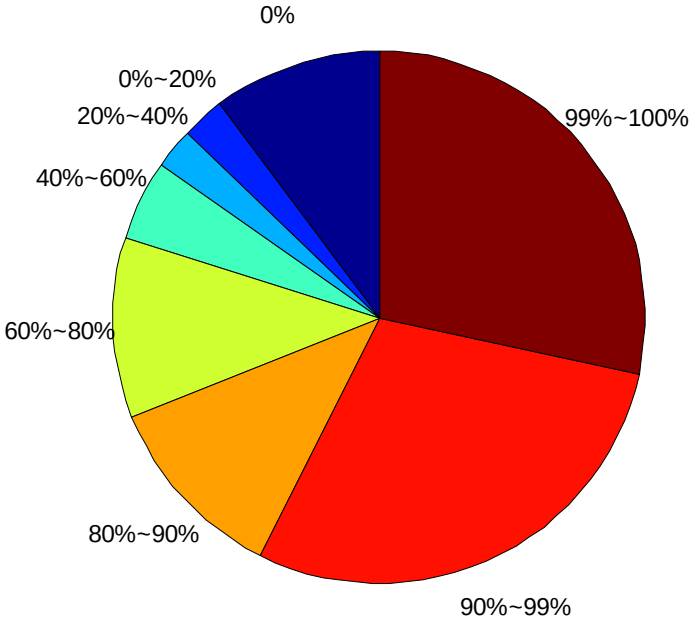


Figure 5.1. Distribution of signal quality in the 3,953,878 FHR windows

5.2 Artificial Signal Loss

In this part of the study, randomly generated artificial signal loss was used to test the robustness of each feature, and an exponential model was built to quantify the relationship between error and signal loss for each feature.

5.2.1 Data

In this section, the FHR cases studied were those in OIFD1 (all 49,151 cases) but not in OIFD3 (7,568 cases with detailed labour outcome information), since OIFD3 was left for investigating the influence of signal loss on labour outcome classification (Section 5.4). 41,583 cases were thus selected, including 1,470,483 Stage 1 windows and 349,490 Stage 2

windows. Signal quality was calculated for each FHR window, defined as the proportion of valid collected signal points. In the 41,583 cases, 144,687 Stage 1 windows have 100% signal quality, while 30,922 Stage 2 windows have 100% signal quality. 1000 Stage 1 windows and 1000 Stage 2 windows were randomly selected for the study. The selection process is illustrated in Figure 5.2. In these 2,000 windows, artificial signal loss was generated to degrade the signal quality to different levels. The 64 features were recalculated based on the degraded signal windows to examine the influence of signal loss on the feature values.

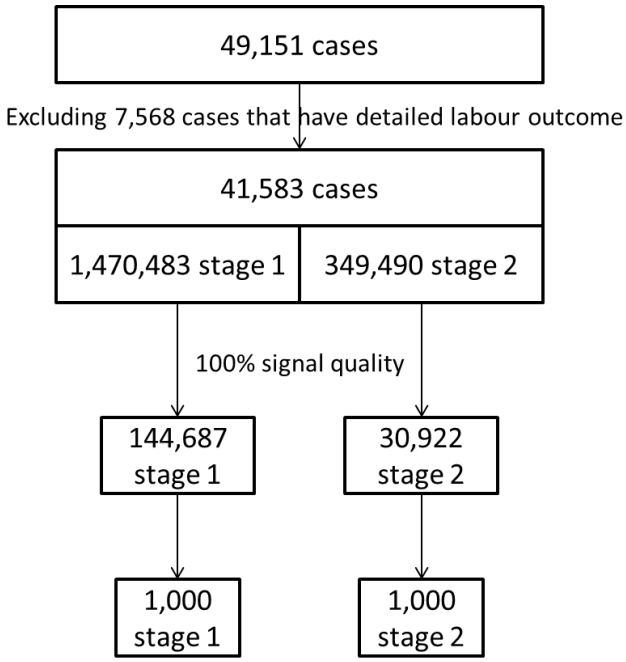


Figure 5.2 Data selection of the 2,000 FHR windows.

5.2.2 Methods

Generation of artificial signal loss

For each FHR window, artificial signal loss was randomly generated to degrade the signal quality to 10%, 20%, 30%... 90%. Two methods were used to generate signal loss, in order to

simulate different situations in FHR recording (Figure 5.3). Method 1 - Random loss points: Signal loss was generated by randomly selecting points, for instance, to generate a 90% quality window, 10% of signal points were randomly selected from the original window and removed. Method 2 - Signal loss segments: Sometimes FHR signal loss appears in large segments, therefore, this method generates signal loss as segments of loss points (removing a set of consecutive points); different numbers of signal loss segments were applied in this study: 1 segment, 2 segments and 3 segments. Intuitively, Method 1 can be taken as Method 2 with a higher number of signal loss segments.

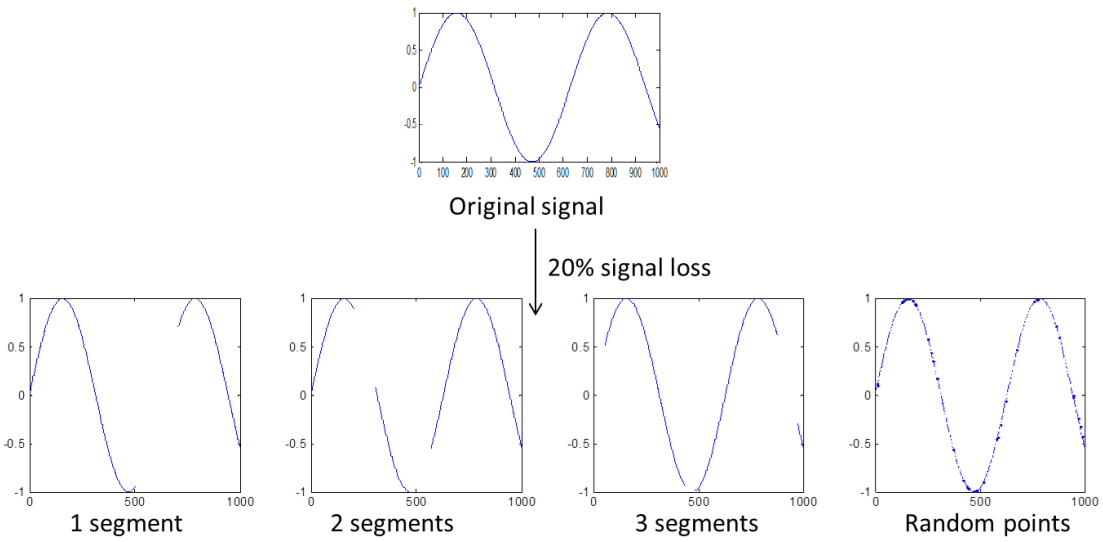


Figure 5.3 An illustration of the methods of generating artificial signal loss with 20% signal loss.

Influence of artificial signal loss

The influence of signal loss was evaluated in two ways:

1. The differences between degraded feature values and original feature values (absolute error) were calculated to show the influence of signal loss on feature values.

2. The correlation coefficient between signal loss and absolute error was calculated to compare the influence of signal loss on different features.

Statistical model between signal loss and error

To quantify the relationship between signal loss and error (absolute error), a statistical model was built. The model can be described as:

$$z = f(x, y), \tag{5-1}$$

where x is the proportion of signal loss, y is the number of signal loss segments and z is the error. Initially, the number of signal loss segments was fixed. In each specific number of segments, the parameters of the model were estimated. It was observed that with an increase in signal loss, the error grow linearly at first. On the other hand, for some features, the error stops growing when they reach a maximal level. The constraints of the model is when $x = 0, z = 0$; when $x \to \infty$, z reaches a maximal level; and when $x \to 0$, the derivative of z should be a constant. An exponential model was then built to estimate the relationship between proportion of signal loss (x) and error (z):

$$z = \lambda_1(1 - e^{-\lambda_2 x}), \tag{5-2}$$

where $\lambda_1$represents the maximal level of error, and λ_1 and λ_2 jointly decide the growth of error against signal loss. When signal loss is high enough, the error approximates the maximal level of λ_1. On the other hand, when $\lambda_2 x \to 0$, the model can be approximated using a linear model, since:

$$\frac{dz}{dx} = \lambda_1\lambda_2(e^{-\lambda_2 x}) \to \lambda_1\lambda_2 \tag{5-3}$$

For each fixed feature and number of segments, λ_1 and λ_2 (or $\lambda_1\lambda_2$ in the case of a linear model) can be estimated. A model can then be built calculating the relationship betweenλ_1, λ_2 and number of segments (y):

$$(\lambda_1, \lambda_2) = g\,(y) \tag{5-4}$$

If the parameters of (5-4) are estimated, they can be used to calculate the parameters of the exponential model in (5-2) or (5-3). The error can then be estimated using the exponential model.

5.2.3 Results

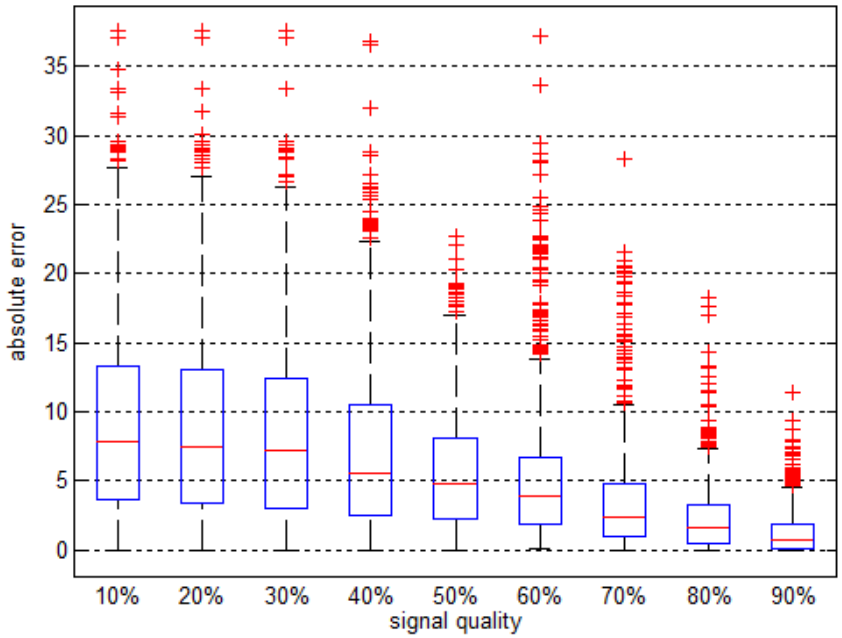
In the presentation of the results of this chapter, it should be noted that Feature 15-18 are unaffected by the signal loss of FHR. This is because they are contraction features calculated only from uterine contraction signals. There is therefore no influence of FHR signal loss on these features.

Absolute error against signal loss

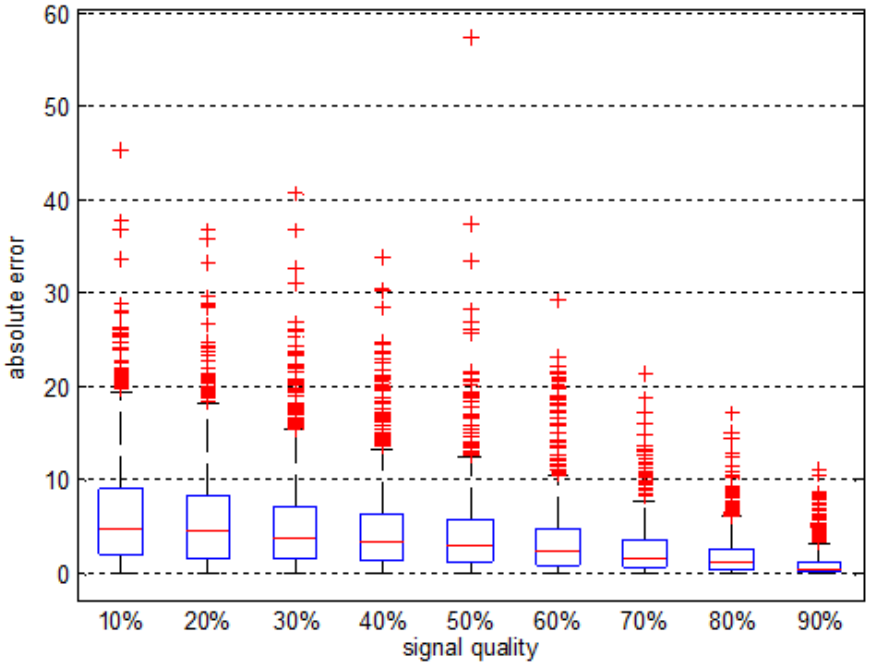
The absolute error between degraded feature values and original feature values was calculated to show the influence of signal loss on feature values. An example of this is illustrated in Figure 5.4, where the relationship between signal quality and absolute error from 1000 cases of Stage 1 for Feature1 (baseline mean) is shown.

It is shown that there is a consistent increase in absolute error with the decrease of signal quality. This is expected since intuitively the error increases with the decrease of signal quality. In addition, the absolute error decreases with the increasing number of signal loss

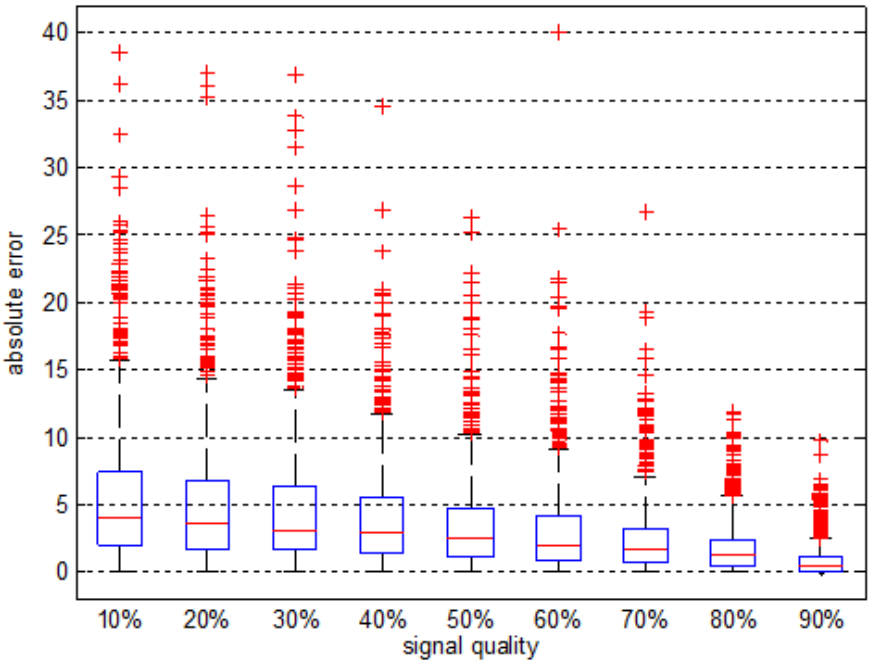
segments. This is because of the fact that in the calculation of baseline mean involves interpolation, thus one larger signal loss segment will result in higher error than many smaller segments of equivalent signal loss points.



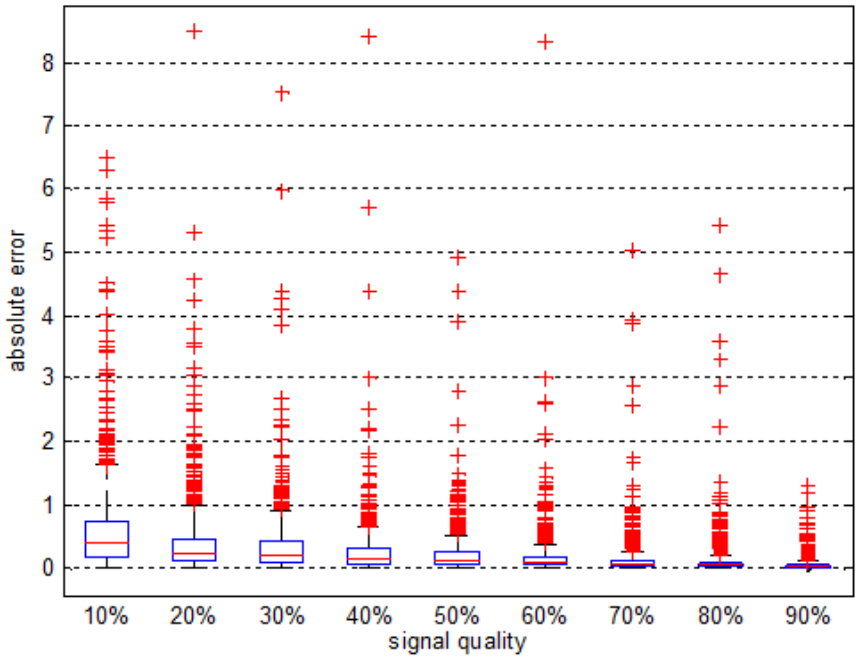
(A) 1 signal loss segment



(B) 2 signal loss segments



(C) 3 signal loss segments



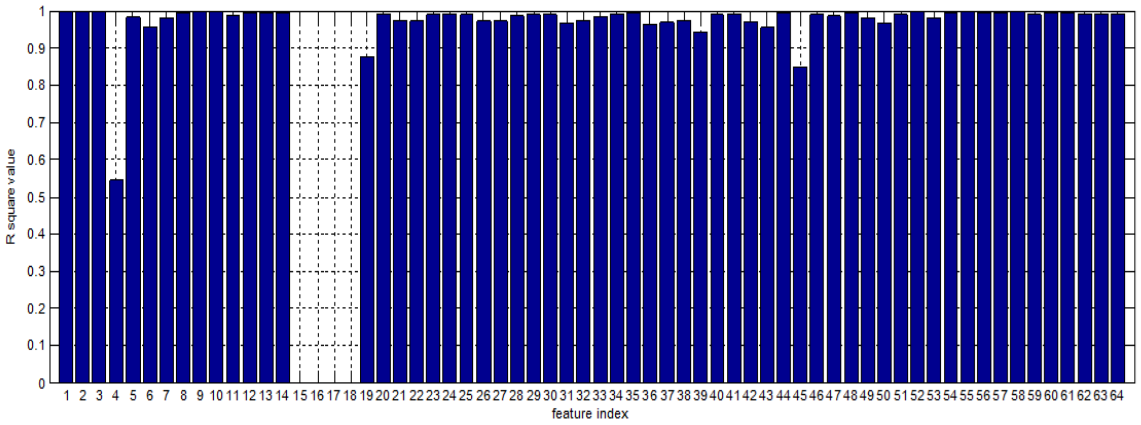
(D) Random signal loss points

Figure 5.4 Relationship between signal quality and absolute error of Stage 1, Feature1 (baseline mean) in boxplots. (A) 1 signal loss segment; (B) 2 signal loss segments; (C) 3 signal loss segments; (D) random signal loss points.

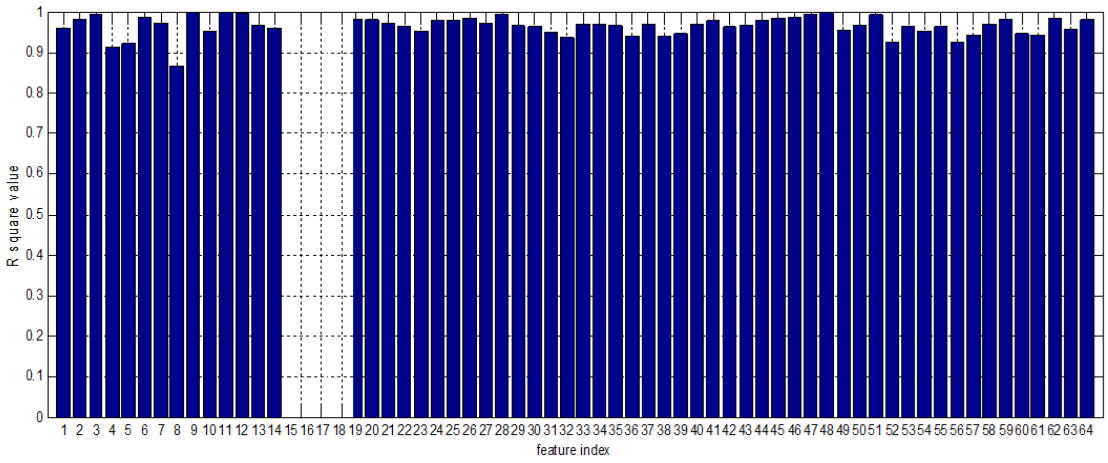
Statistical model between signal loss and absolute error

The goodness of fit for the exponential model was calculated using the R-square value. Goodness of fit for each feature is shown in Figure 5.5. The R-square values at Stage 1 for 1 signal loss segment and random missing points are shown. It is found that regardless of the methods of signal loss, for 57 out of the 60 features (the four unaffected features being excluded) the R-square values are over 0.9, indicating a high goodness of fit for the exponential model. For 2 signal loss segments, 3 signal loss segments and Stage 2 (not shown in the figure), the results are similar: over 90% of the R-square values are above 0.9, regardless of the signal loss generation method or stages.

A typical example of the exponential fitting curve is shown in Figure 5.6. It is found that the exponential model can be used to accurately describe the relationship between signal loss and error. At the beginning when the signal loss is small, the error grows linearly against signal loss; the growth rate becomes less with the increase of signal loss.

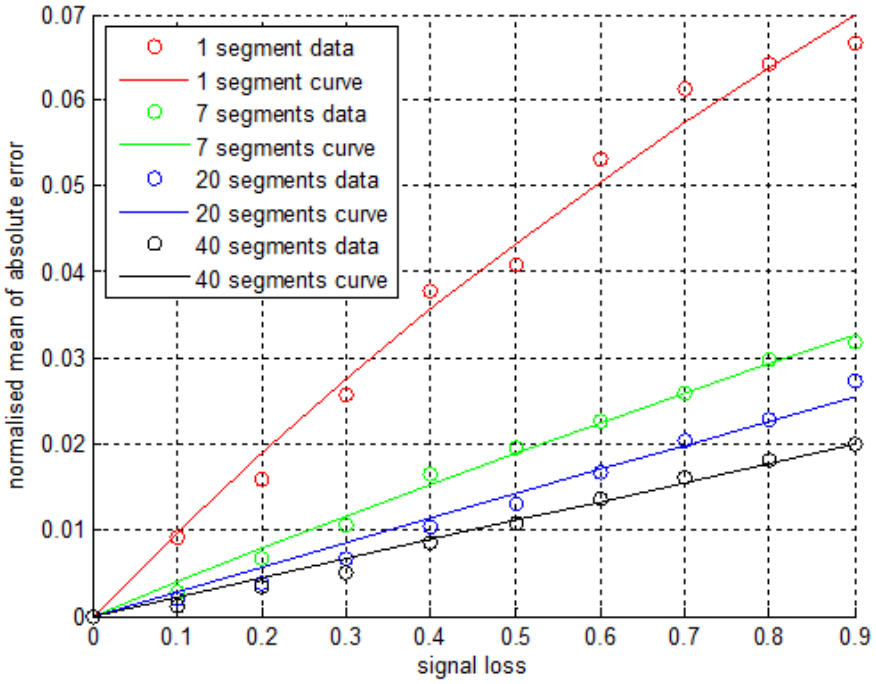


(A) 1 signal loss segment

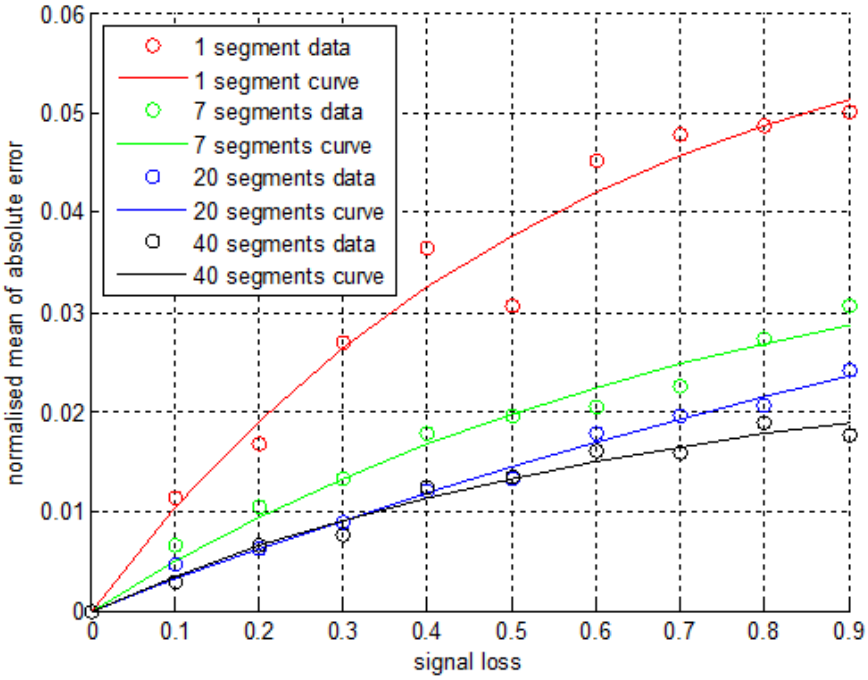


(B) Random signal loss points

Figure 5.5 The R-square value of the exponential model for each feature, Stage 1: (A) 1 signal loss segment; (B) random loss points.



(A) Relationship between signal loss and mean of absolute error



(B) Relationship between signal loss and standard deviation of absolute error

Figure 5.6 Fitting the exponential model for Feature 1 (baseline mean): (A) signal loss and mean of absolute error; (B) signal loss and standard deviation of absolute error.

5.2.4 Discussion

In Section 5.2, randomly generated artificial signal loss has been used to test the robustness of feature values for each feature. It was found that signal loss has influences in different ways for different features. In general, absolute error grows with the increase in signal loss, and the influence is higher when signal loss appears in large segments.

An exponential model, described by Equation 5-2, was built to quantify the relationship between artificially generated signal loss and absolute error for each feature. The model was validated using clinical data degraded by artificial signal loss. For 57 of the 60 features, the R-square values of the exponential model are above 0.9, indicating good agreement between the model and data. Therefore, it is concluded that the relationship between artificially

generated signal loss and absolute error can be adequately quantified by an exponential model.

However, the signal loss in clinical situations may be different from artificially generated signal loss, in terms of the length of signal loss segments, the location of signal loss and the relationship between the signal loss segments. Therefore, in Section 5.3, how to modify the model based on clinical data was investigated.

5.3 Clinical Signal Loss

In Section 5.2, artificial signal loss was generated to investigate the relationship between signal loss and error (in feature values). However, the signal loss in clinical situations could be different from artificially generated signal loss, especially in terms of length and location of the signal loss segments. Therefore, to build a robust model of error due to signal loss, the artificial signal loss has to be as similar as possible to actual clinical signal loss. In this section, signal loss was generated using randomly selected clinical templates, in order to simulate the patterns in clinical situations. The exponential model was then modified and optimised to a bivariate model using the new set of degraded signals.

5.3.1 Methods

The whole OIFD1 database was used in this study, including 3,953,878 FHR windows of 49,151 cases, in which there are 1,939,099 Stage 1 windows and 500,829 Stage 2 windows. These windows were used to form a clinical template pool to generate artificial signal loss. The 2,000 windows with 100% signal quality were identical as mentioned in Section 5.2.1.

Similarly to Section 5.2.2, in the 2,000 windows with 100% signal quality, signal loss was

generated to degrade the signal quality to different levels. The 64 OxSys features were recalculated on the degraded signal windows to examine the influence of signal loss. The difference from Section 5.2.2 is that, instead of randomly generating signal loss points or segments, the signal loss was generated using randomly sampled clinical templates.

Selection of variables for the model

To accurately describe the signal loss in a FHR window, the variables needed are:

x: Proportion of signal loss (1 variable);

y: Number of the signal loss segments (1 variable);

c: Lengths of the signal loss segments (multiple dependent variables);

d: Locations of the signal loss segments in the window (multiple dependent variables).

These variables are not independent from each other. A combination of some variables (e.g. *c* and *d*) can accurately describe the signal loss in a FHR window. To build a model of error due to signal loss, a set of variables should be selected as the controlled variables of the model, while the rest of the variables need to be simulated to be as similar as possible to those found in clinical situations.

Firstly, a set of variables (*x* and *y*) were selected as controlled variables of the model. As the number of variables increases, the condition of signal loss can be described more accurately, but the complexity of the model will also increase. This is a trade-off between prediction accuracy and complexity of the model. *x* and *y* were selected as the controlled variables in the model here, since they provide fewer parameters and a more simple way to describe the

model. On the other hand, *c* and *d* were not easy to control, since the number of variables of *c* and *d* changes with *y*, while the length and location of each signal loss segment will be limited by both *x* and *y*.

Secondly, as *x* and *y* were controlled in the model, *c* and *d* have to be simulated to be similar to values found in clinical conditions. So far there is neither a study nor any clear evidence to suggest a clear pattern of length or location of the signal loss segments. In addition, owing to the fact that it is impossible to know the exact actual FHR signal under the signal loss segments, it is difficult to simulate *c* and *d* based on their correlation with the FHR signal. Therefore, using clinical signal loss templates with the model is so far the best way possible to simulate *c* and *d* in clinical signal loss. As the number of clinical windows is large (over 3,000,000), it is easy to find many windows for each pair of parameters (*x*,*y*). Therefore, the signal loss can be generated from a large template pool (> 100 windows for each parameter pair). In this way, *x* and *y* were controlled in the model, while *c* and *d* were simulated using clinical templates.

Generation of signal loss using clinical templates

For each FHR window, signal loss was randomly generated to degrade the signal using different proportions of signal loss (*x*) of 10%, 20%, 30%... 90%. Different numbers of signal loss segments were applied for each group of degraded signal: 1, 3, 7, 10, 15, 20, 30 and 40 segments (*y*). For each parameter pair (*x*,*y*), the 2,000 windows were degraded using the parameter pair. For each set of the 2,000 windows, the clinical signal loss was simulated as follows.

Firstly, a clinical template pool was drawn from the clinical database (3,953,878 FHR windows) by selecting a subset of windows that fits the parameter set. For example, for the parameter set (50% signal loss, 5 signal loss segments), the clinical template pool consists of all the windows in the clinical database that have 50% (± 1%) signal loss and 5 signal loss segments. It was found that the size of each template pool is larger than 100 clinical cases, which was large enough to form a random sample pool.

Secondly, for each of the 1,000 windows, a window from the clinical template pool was randomly selected, and the artificial signal loss was generated exactly as the clinical signal loss in this window. In this way, the length and location of the signal loss (*c* and *d*) could be simulated as similar with clinical signal loss as possible.

The bivariate model

In this section, a bivariate model was developed to quantify the relationship between signal loss and error. This was built upon the exponential model in Section 5.2, which has one variable - proportion of signal loss (*x*). The bivariate model adds another variable to the exponential model: the number of signal loss segments (*y*).

The exponential model was the same as described in Section 5.2, which quantified the relationship between proportion of signal loss (x) and the mean of absolute error (z). The model can be described as:

$$z = \lambda_1(1 - e^{-\lambda_2 x}) \tag{5-2}$$

The clinical interpretation of the parameters is that, λ_1 represents the maximal level of error,

and λ_1 with λ_2 determines the growth of error against signal loss. Defining $\lambda_3 = \lambda_1\lambda_2$, then λ_3 represents the initial growth rate (sensitivity) of error against signal loss. λ_1 and λ_3 both show a steady growth rate against number of signal loss segments for most of the features. Intuitively, this reflects the relationship between number of segments and the growing trend of error. Therefore, it is possible to calculate λ_1 and λ_3 using the number of segments.

Assume that λ_1 and λ_3 each has a linear relationship with the number of segments, i.e.:

$$\lambda_1 = k_1 y + b_1, \quad \lambda_3 = k_2 y + b_2 \tag{5-5}$$

then error can be calculated using:

$$z = (k_1 y + b_1)(1 - e^{-\frac{k_2 y + b_2}{k_1 y + b_1} * x}) \tag{5-6}$$

where z is the mean of absolute error, x is the proportion of signal loss, and y is the number of signal loss segments. The four parameters k_1, k_2, b_1, b_2 can be estimated by multivariate non-linear regression using a set of data (x, y). As $\lambda_3 = k_2 y + b_2$ represents the initial growth rate (sensitivity) of error against signal loss, k_2 represents the initial sensitivity of error against number of signal loss segments.

As there were 8 sets of degraded signals (1, 3, 7, 10, 15, 20, 30 and 40 segments), each containing 10 degraded signals (0%, 10%, 20% ... 90% signal loss), 8×10 points were used for the bivariate non-linear regression, estimating 4 parameters (k_1, k_2, b_1, b_2). The non-linear regression toolbox of Matlab 2009b (The MathWorks Inc.) was used for the regression.

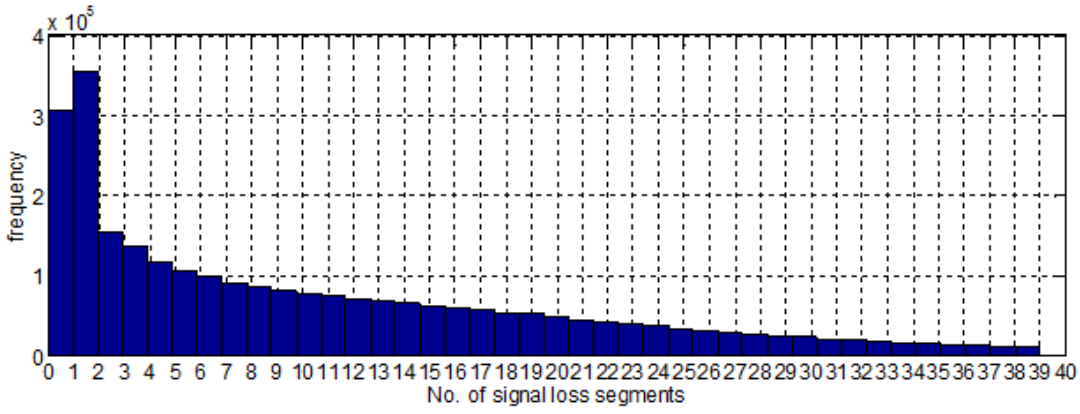
It was observed that, apart from the mean of absolute error, the standard deviation has a

similar trend against proportion of signal loss and number of signal loss segments. Therefore, another bivariate model using Equation 5-6 was developed with different parameters (k_1, k_2, b_1, b_2) to calculate the relationship between signal loss and the standard deviation of absolute error.

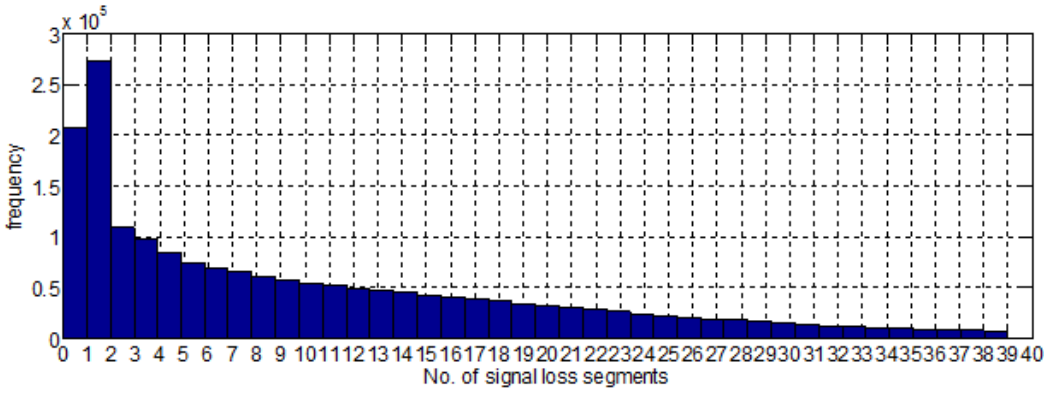
5.3.2 Results

Distribution of signal loss segments

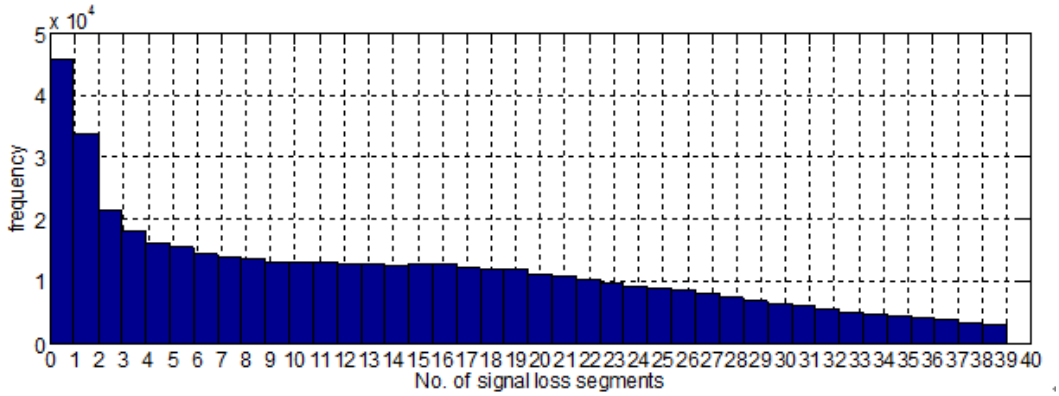
The distribution of number of signal loss segments in 3,953,878 FHR windows is shown in Figure 5.7. It is found that over 80% of the windows have fewer than 10 signal loss segments. It is also shown that in Stage 2 there are relatively more windows that have five or more segments than in Stage 1. This is expected, since in Stage 2 there are more contractions and pushing, resulting in a higher possibility of loss of contact, which leads to more signal loss segments.



(A) All stages



(B) Stage 1

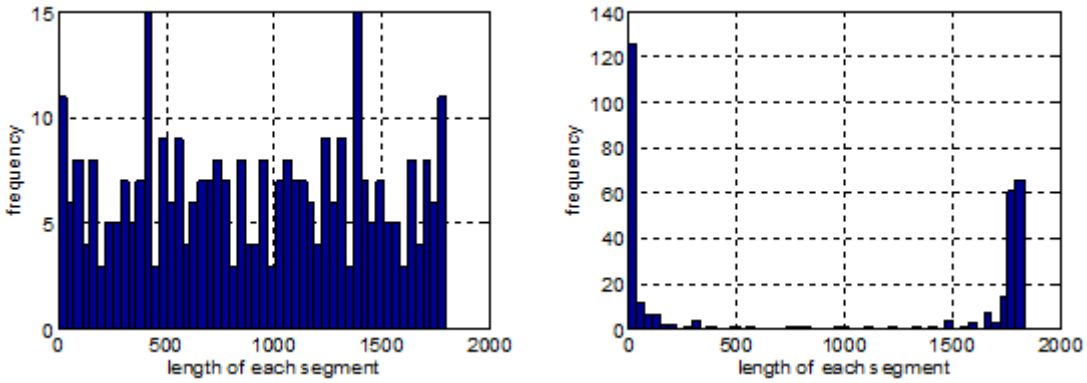


(C) Stage 2

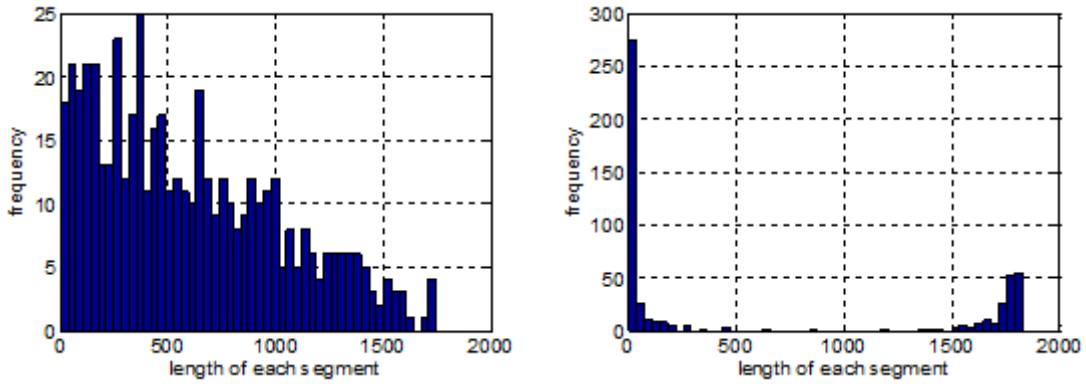
Figure 5.7 Distribution of number of signal loss segments in OIFD1.

To compare clinical signal loss with randomly generated artificial signal loss (see Section 5.2), the distribution of length of signal loss segments is shown in Figure 5.8. This shows that the length of signal loss segments in clinical situations has quite a different distribution from those of randomly generated artificial signal loss. In clinical cases, signal loss tends to appear as a few large segments instead of many small segments, as shown in Figure 5.7 (the number of signal loss segments is mostly less than 10 in a window). This can be partly caused by the clinical situations where patients need to have the ultrasound sensor removed (e.g. during urination). This also confirms that it is necessary to simulate signal loss using clinical

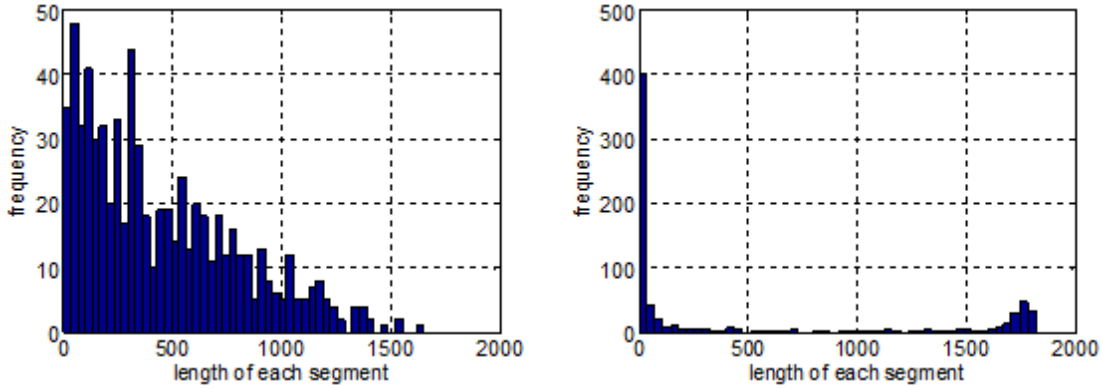
templates instead of randomly generating artificial signal loss, which can potentially introduce bias to the results.



(A) Two signal loss segments



(B) Three signal loss segments



(C) Four signal loss segments

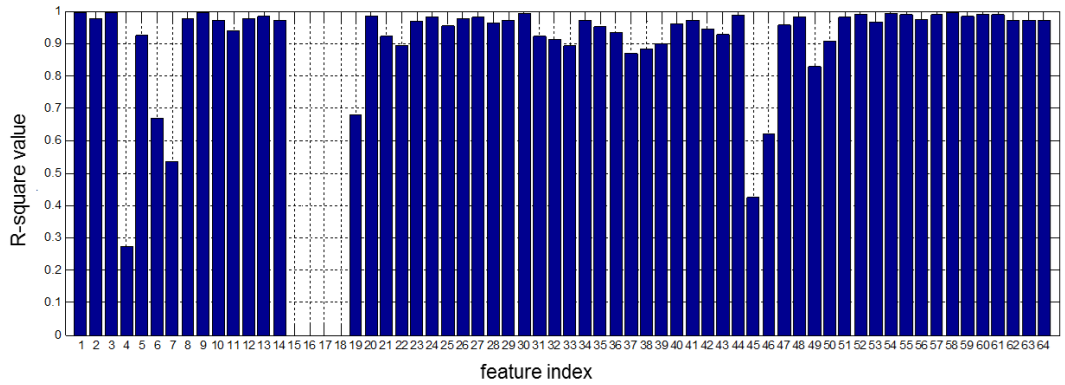
Figure 5.8 Distribution of length of signal loss segments in Stage 1 windows that have 50% (± 1%) signal quality. Left: randomly generated artificial signal loss. Right: clinical signal loss. (A) Two

signal loss segments; (B) three signal loss segments; (C) four signal loss segments.

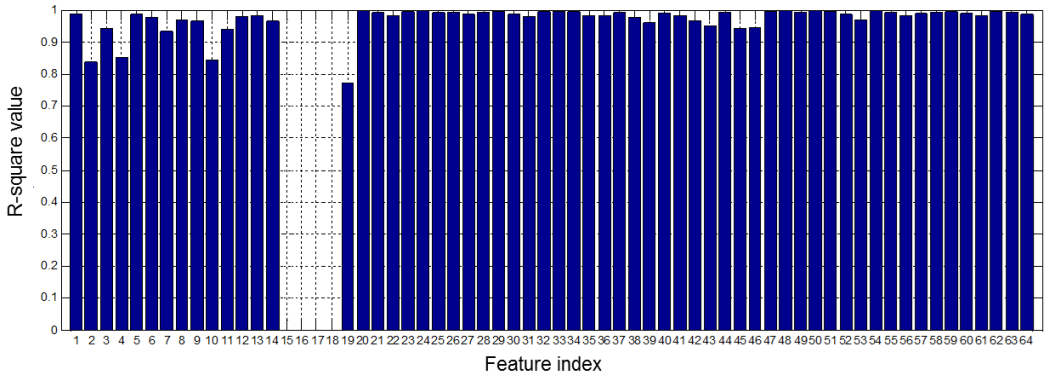
Goodness of fit: the exponential model

The bivariate model is built upon the exponential model, thus as a first step, the goodness of fit of the exponential model is validated as in Section 5.2. The procedures are exactly the same except that this time, signal loss generated from clinical templates was used instead of randomly generated signal loss. The goodness of fit for the exponential model was calculated using the R-square value.

The R-square value for each feature is shown in Figure 5.9. The situations in Stage 1 for two extreme conditions - 1 signal loss segment and 40 signal loss segments - are shown. It is found that regardless of the number of signal loss segments, for 54 of the 60 features (90%), the R-square values are over 0.8 (except for the unaffected features: Feature 15-18), indicating a high goodness of fit for the exponential model. The goodness of fit is better in the case of 40 signal loss segments, which indicates that under the same proportion of signal loss, large signal loss segments can cause more error.



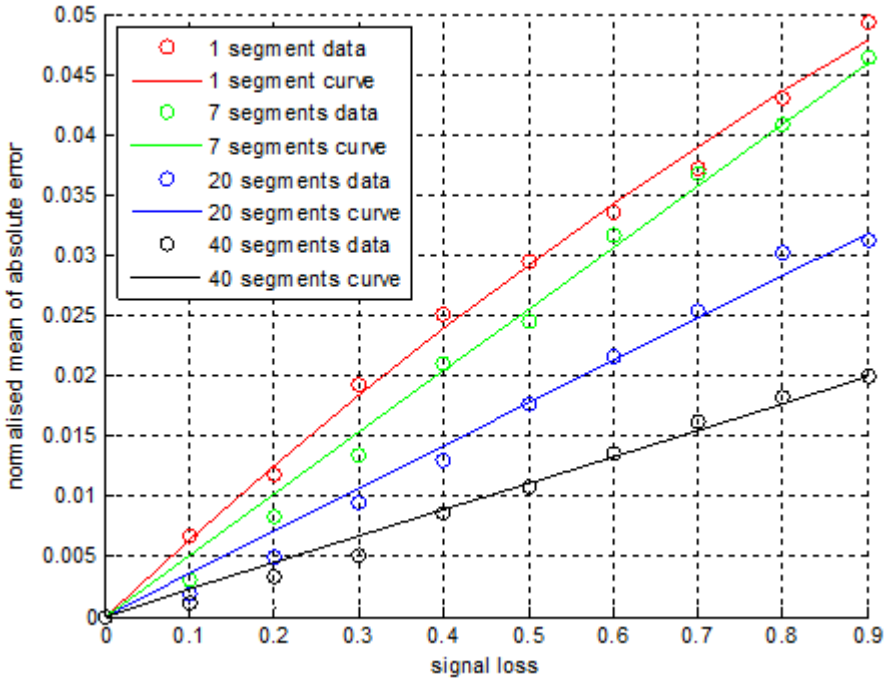
(A) 1 signal loss segment



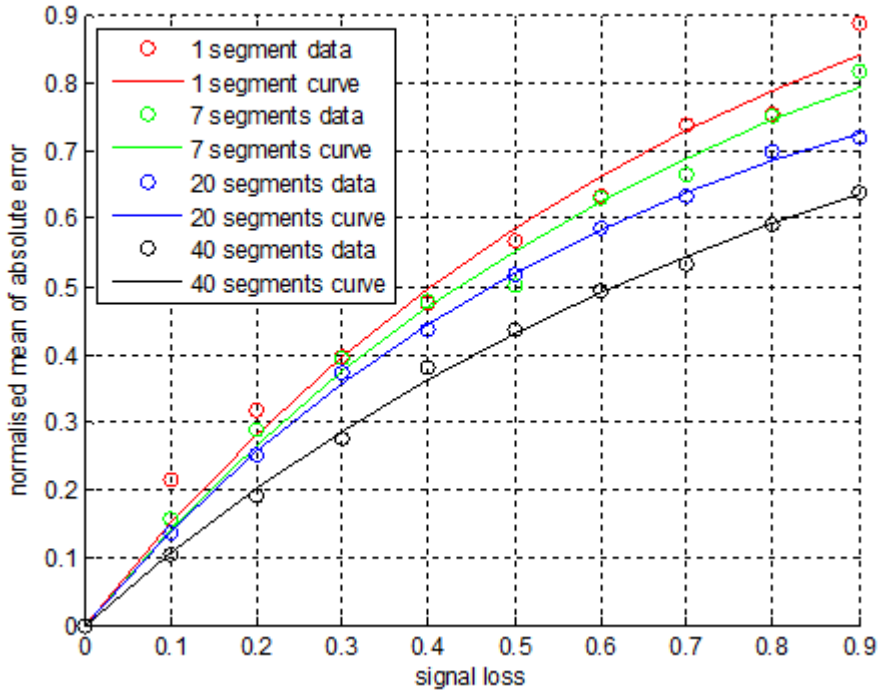
(B) 40 signal loss segments

Figure 5.9 The R-square value of the exponential model for each feature of Stage 1. (A) 1 signal loss segment; (B) 40 signal loss segments.

Two typical examples of the exponential fitting curves are shown in Figure 5.10. The mean of absolute error was normalised to compare the error between different features. It is found that the exponential can accurately describe the relationship between signal loss and error. It is also found that as the number of segments increases, the exponential model tends to be closer and closer to a linear model. Therefore, the exponential model is still applicable when simulating clinical signal loss.



(A) Feature 1



(B) Feature 20

Figure 5.10 Fitting the exponential model between signal loss and mean of absolute error: (A) Feature 1: baseline mean (continuous); (B) Feature 20: number of accelerations (discrete).

Goodness of fit: the bivariate model

For each exponential model, λ_1, λ_2 and λ_3 were calculated. The relationship between these parameters and the number of segments (y) for Feature 1 is shown in Figure 5.11. It is found that λ_3 was close to a linear relationship. Although λ_1 and λ_2 were not likely to be from a linear model, the two parameters were found to be dependent on each other. This confirms the assumption that λ_1 and λ_3 are both linearly related to the number of signal loss segments (Equation 5-5). Therefore, a bivariate model can suitably describe the relationship between the proportion of signal loss, number of signal loss segments and error. In addition, a bivariate regression that considers the correlation of λ_1 and λ_3 could be better than simply regressing the two parameters separately using two independent models.

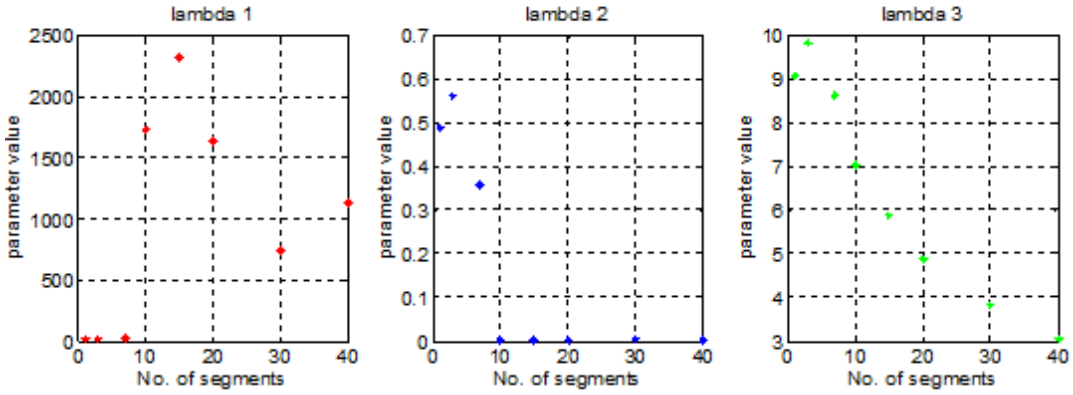


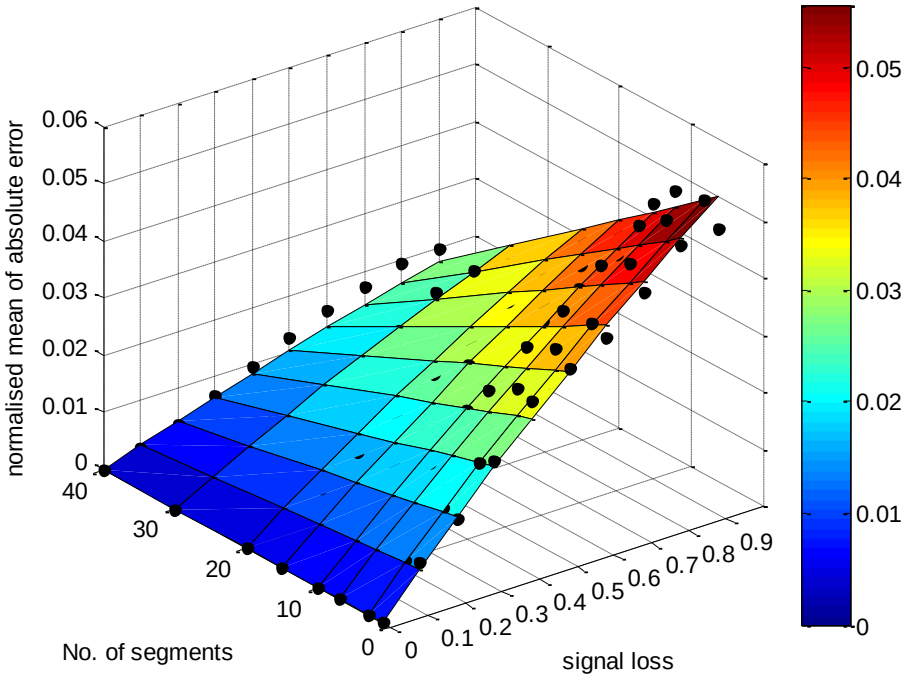
Figure 5.11 The relationship between different parameters of the exponential model and number of segments: Feature 1: baseline mean.

The goodness of fit of the bivariate model was calculated using the R-square value. It is found that the mean of the absolute error fits relatively better: for 55 out of the 60 features (> 90%, excluding the unaffected features), the R-square values are over 0.8, indicating a high

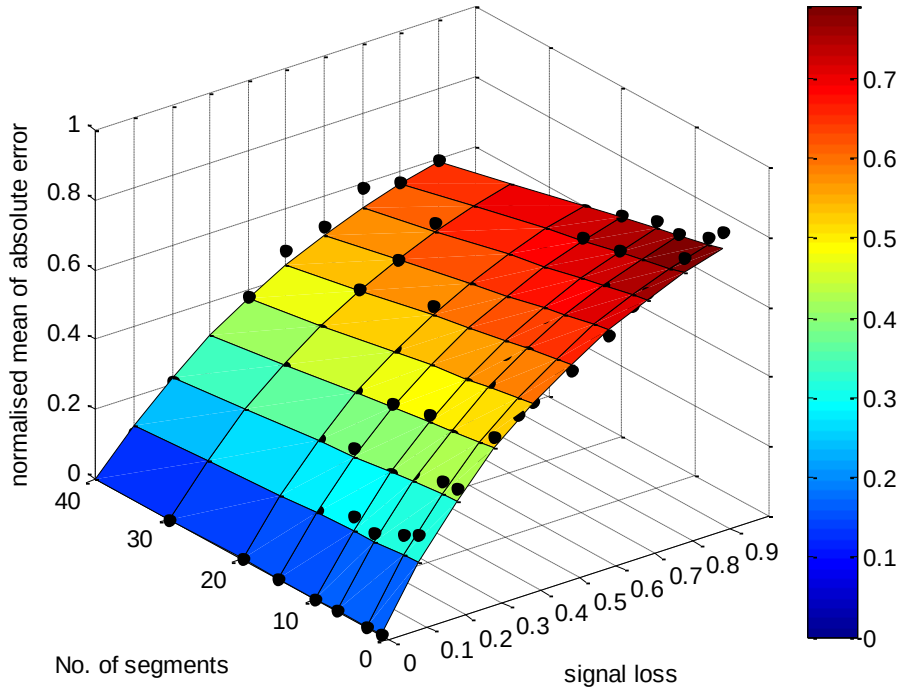
goodness of fit for the bivariate model. The standard deviation of the absolute error, on the other hand, also has a reasonable goodness of fit: for 55 out of the 60 features (over 90%, excluding the unaffected features), the R-square values are over 0.7. This indicates that for a specific feature, given the proportion of signal loss and the number of signal loss segments, the mean and standard deviation of absolute error can be estimated accurately.

On the other hand, no clear difference between Stage 1 and Stage 2 can be observed in terms of goodness of fit, which indicates that the model applies to FHR windows in both stages. The parameters for the model of Stage 1 and Stage 2 are similar. In fact, when testing Stage 2 with Stage 1 model, the goodness of fit is similar to the goodness of fit of Stage 2 (R-square values over 0.7 for over 90% of the features). Therefore the same exponential model can be used regardless of the labour stage.

Two typical examples of the fitting surface are shown in Figure 5.12. The mean of the absolute error has been normalised to compare the error between different features. These surfaces can provide an intuitive way to review the relationship between proportion of signal loss, number of signal loss segments and the mean of absolute error. It is found that the bivariate model can be used to describe the relationship between signal loss and error both for artificial signal loss and clinical signal loss.



(A) Feature 1



(B) Feature 20

Figure 5.12 Fitting the bivariate model between signal loss and mean of absolute error: (A) Feature 1: baseline mean (continuous feature); (B) Feature 20: number of accelerations (discrete feature). Black dots are data and the coloured surface is the fitted surface.

Above are two typical examples of bivariate model. The estimated error for all features is shown in Table 5.1 (for the condition of 1 signal loss segments). It provides a very direct way of comparing the robustness of different features against signal loss. For example, Feature 1 (baseline mean) has an average value of 137.65, while the mean of absolute error is about 6.39 at a signal loss of 80%, with a standard deviation of 6.59. The mean of the absolute error is relatively low compared with the average feature value, owing to the interpolation of missing signals in the algorithm.

On the other hand, Feature 20 (number of acceleration) has an average value of 2.93, while the mean error is about 2.29 at the signal loss of 80%, with a standard deviation of 2.36. The mean is relatively high with this level of average feature value, owing to the fact that acceleration can be completely covered by a large section of signal loss.

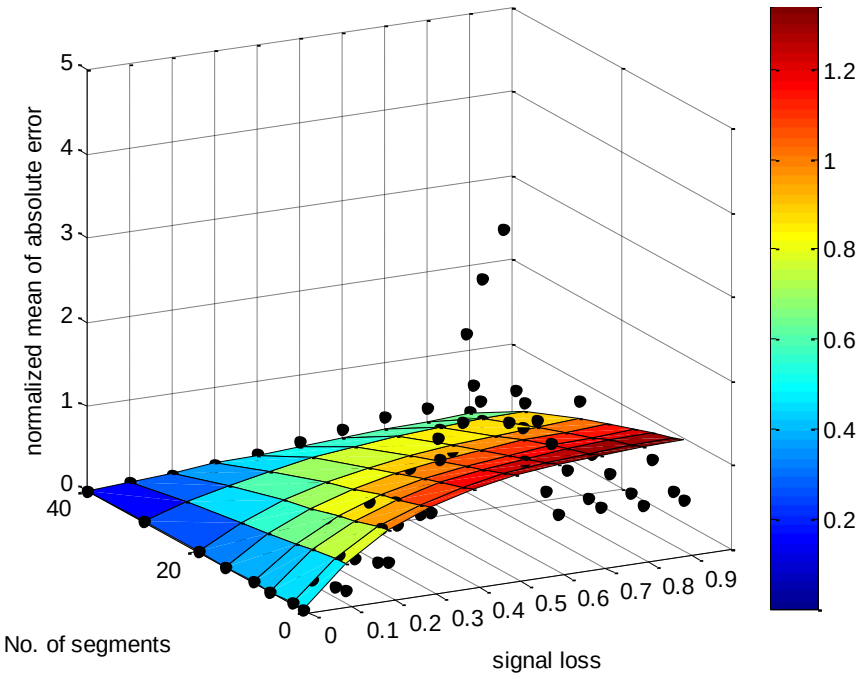
Table 5.1 Estimated error using the bivariate model for each feature. Features discussed in this section are shown in bold.

Feature	***Average feature value***	***Estimated absolute error with different signal loss***							
		80%		**60%**		**40%**		**20%**	
		Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
1	137.65	6.39	6.59	4.86	5.39	3.28	3.94	1.66	2.16
2	0.70	0.04	0.03	0.03	0.03	0.02	0.02	0.01	0.01
3	0.86	0.77	0.15	0.58	0.14	0.38	0.11	0.19	0.07
4	**0.09**	**0.12**	**0.14**	**0.11**	**0.13**	**0.10**	**0.11**	**0.07**	**0.08**
5	125.90	6.64	7.50	4.98	6.35	3.32	4.81	1.66	2.75
6	**-0.92**	**3.13**	**3.29**	**2.35**	**2.48**	**1.56**	**1.66**	**0.78**	**0.84**
7	**14.21**	**48.58**	**145.84**	**36.44**	**66.39**	**24.29**	**27.82**	**12.15**	**9.09**
8	8.93	3.75	4.41	2.81	3.63	1.87	2.66	0.94	1.47
9	0.33	0.05	0.05	0.04	0.04	0.03	0.02	0.01	0.01
10	0.40	0.09	0.12	0.07	0.09	0.04	0.06	0.02	0.03
11	4.01	2.50	3.56	1.87	2.67	1.25	1.78	0.62	0.89

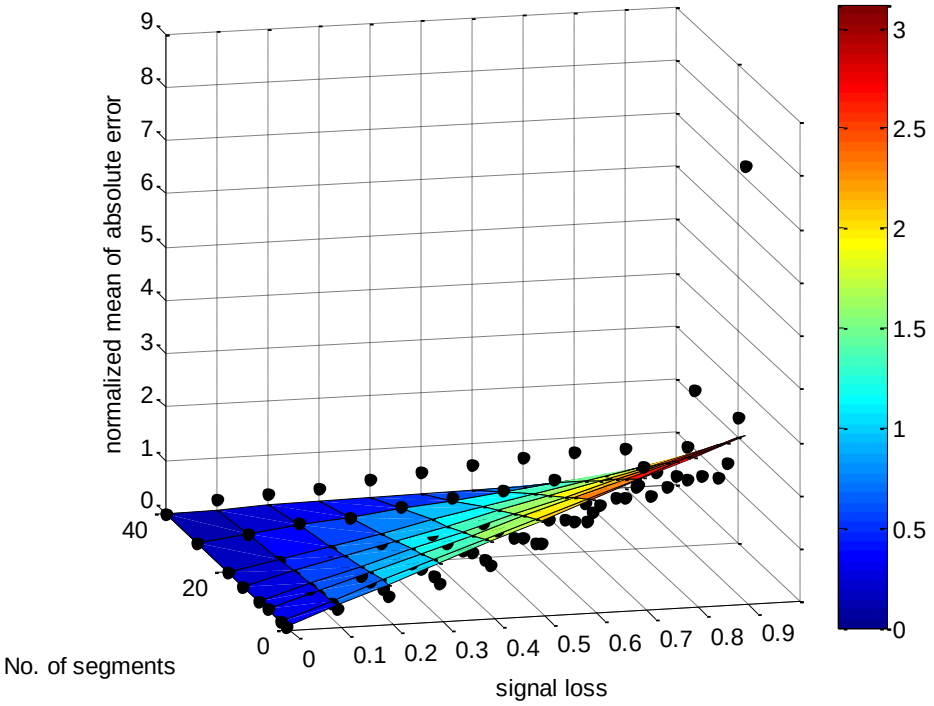
12	0.44	0.15	0.16	0.14	0.16	0.11	0.16	0.07	0.13
13	0.01	0.54	0.36	0.43	0.33	0.31	0.27	0.16	0.17
14	0.07	0.03	0.03	0.02	0.03	0.02	0.02	0.01	0.01
15	3.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	69.22	0.01	0.05	0.00	0.04	0.00	0.03	0.00	0.02
17	101.18	0.05	0.43	0.04	0.37	0.03	0.29	0.02	0.17
18	3.68	0.25	3.84	0.24	3.57	0.21	3.05	0.14	2.02
19	0.93	0.47	4.39	0.42	4.15	0.34	3.63	0.21	2.49
20	2.93	2.29	2.36	1.94	2.21	1.48	1.91	0.85	1.29
21	34.35	25.13	35.10	21.89	32.50	17.14	27.55	10.18	18.08
22	33.23	30.99	50.76	27.33	49.48	21.72	45.65	13.13	34.21
23	5.76	8.82	10.19	6.90	9.89	4.81	9.04	2.51	6.66
24	50.27	28.79	30.73	25.54	30.33	20.45	28.81	12.48	22.90
25	63.65	41.79	45.62	38.00	45.35	31.39	43.97	19.93	36.83
26	25.41	28.73	26.75	21.55	24.31	14.37	20.08	7.18	12.74
27	33.98	24.96	33.37	22.26	31.95	17.94	28.51	11.03	20.18
28	19.23	18.29	27.07	15.92	24.96	12.46	21.01	7.39	13.67
29	25.08	23.68	32.94	20.94	31.07	16.70	27.06	10.14	18.45
30	34.84	27.52	35.90	26.15	35.79	23.05	35.07	16.01	30.37
31	56.98	178.79	182.86	134.09	171.14	89.40	147.29	44.70	98.76
32	65.58	140.60	250.85	119.64	243.99	91.17	224.13	52.51	166.60
33	6.15	5.93	7.01	5.32	6.85	4.33	6.36	2.69	4.82
34	100.95	103.47	146.06	93.75	144.77	77.11	139.04	48.66	113.54
35	104.79	112.48	201.41	104.01	200.28	87.97	194.36	57.58	163.26
36	1.11	0.98	1.89	0.86	1.89	0.69	1.87	0.42	1.71
37	0.88	0.62	0.68	0.53	0.68	0.40	0.65	0.23	0.54
38	0.63	1.72	4.90	1.53	4.04	1.23	2.96	0.75	1.64
39	0.04	0.05	0.25	0.05	0.25	0.05	0.25	0.04	0.23
40	0.43	0.54	0.83	0.48	0.81	0.38	0.76	0.23	0.57
41	0.36	0.47	0.79	0.42	0.79	0.35	0.77	0.22	0.67
42	8.81	17.50	56.75	14.72	51.61	11.07	42.66	6.28	27.09
43	8.91	17.74	57.10	14.78	51.97	11.00	43.00	6.17	27.34
44	0.98	1.01	1.00	0.89	1.00	0.70	0.98	0.42	0.86
45	9.87	13.81	46.92	13.79	46.92	13.64	46.80	12.28	44.49
46	13.47	19.70	54.13	19.47	54.08	18.52	53.63	14.79	48.90
47	0.72	0.61	0.55	0.52	0.54	0.40	0.51	0.23	0.39
48	-0.14	0.53	0.56	0.52	0.54	0.50	0.49	0.40	0.36
49	49.62	117.23	95.78	103.29	91.47	82.02	81.26	49.55	57.12
50	0.53	0.43	0.28	0.37	0.27	0.29	0.24	0.17	0.17
51	1.35	1.86	2.05	1.53	2.02	1.12	1.92	0.62	1.52
52	99.53	91.02	78.62	72.04	76.48	50.73	70.27	26.82	52.25
53	0.38	0.44	0.38	0.33	0.37	0.22	0.35	0.11	0.26

54	3.07	2.30	2.33	1.76	2.27	1.20	2.09	0.62	1.57
55	1.85	4.04	4.54	3.10	4.41	2.12	4.05	1.09	3.00
56	0.87	0.48	0.29	0.36	0.29	0.24	0.27	0.12	0.21
57	5.24	6.81	7.05	5.11	6.74	3.41	5.99	1.70	4.22
58	5.32	3.49	3.88	3.01	3.75	2.33	3.39	1.37	2.46
59	1.05	0.62	0.62	0.51	0.51	0.37	0.37	0.21	0.21
60	9.89	7.23	7.45	5.42	7.18	3.61	6.47	1.81	4.65
61	1.40	0.94	1.17	0.70	1.14	0.47	1.04	0.23	0.77
62	1.09	0.89	1.07	0.76	1.07	0.58	1.04	0.33	0.88
63	-0.06	0.68	1.03	0.56	0.88	0.41	0.68	0.23	0.39
64	1.77	0.43	0.53	0.38	0.52	0.31	0.51	0.19	0.43

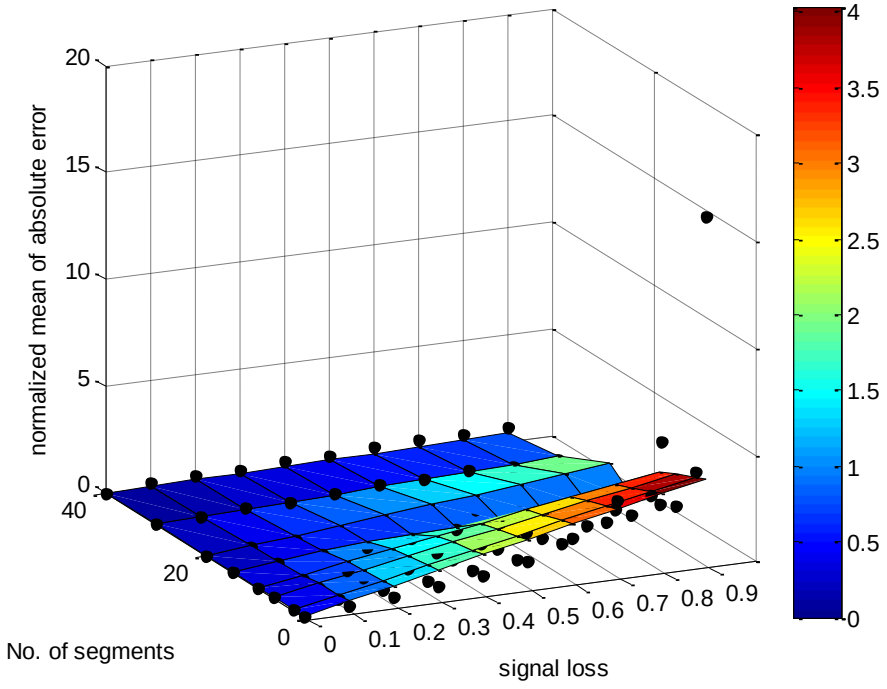
It can also be noted that there are three features that have R-square values lower than 0.7 for the mean of absolute error value: Feature 4 (Signal Stability Index), Feature 6 (Skewness) and Feature 7 (Kurtosis). These are found to be caused by outliers occurred when there is a large percentage of signal loss, as illustrated in Figure 5.13. This is expected, since when the percentage signal loss is as high as 80%-90%, the algorithms are more likely to generate outliers that have extreme values.



(A) Feature 4



(B) Feature 6



(C) Feature 7

Figure 5.13 Outliers: features that have R-square values lower than 0.7: (A) Feature 4 (Signal Stability Index); (B). Feature 6 (Skewness); (C). Feature 7 (Kurtosis).

Two parameters, k_2 and λ_3, were compared between different features to evaluate the initial sensitivity of error against signal loss and number of signal loss segments. As mentioned in Section 5.2, k_2 represents the initial sensitivity of error against number of signal loss segments, and when $y = 1$, $\lambda_3 = k_2 + b_2$ represents the initial sensitivity of error against signal loss. The comparison of these two parameters is shown in Table 5.2. It was found that different features have various initial sensitivity values against signal loss and number of signal loss segments. For instance, Feature 13 (STV tracker trend) has the highest initial sensitivity against signal loss, indicating that this feature is most likely to be affected by a small amount of signal loss. This is owing to the current interpolation method, which interpolates the missing signal using a single value, resulting in a very low STV during signal loss.

In addition, some features, such as Feature 48 (Mutual information feature) have a high initial sensitivity against the number of signal loss segments, indicating that these features are more likely to be affected by the number of signal loss segments for the same quantity of signal loss. This is due to the calculation method for mutual information that uses a sliding window to calculate mutual information, thus under the same signal loss proportion, the more signal segments, the greater the error.

Table 5.2 Top 5 features for initial sensitivity of mean for absolute error against signal loss and signal loss segments.

Ranking	*Sensitivity against signal loss*		*Sensitivity against signal loss segments*	
	Feature name	**Sensitivity (percentage of error/percentage of signal loss)**	**Feature name**	**Sensitivity (percentage of error/number of segments)**
1	No.13 STV tracker trend	72.4	No. 13 STV tracker trend	0.73
2	No. 48 Mutual information	25.4	No. 48 Median recovery time	0.38
3	No. 63 BPRSA - DC	19.5	No. 45 Median recovery time	0.25
4	No. 45 Median recovery time	15.4	No. 39 Number of prolonged decelerations	0.21
5	No. 39 Number of prolonged decelerations	11.7	No. 46 Mean lag time	0.14

As the signal loss was simulated using randomly selected clinical templates, it is necessary to validate whether these models are still valid using different samplings in the clinical template pool. To validate this, 10 new sets of data (2,000 cases) were regenerated using different random drawings of signal loss from the clinical template pool. The models were applied to these 10 new sets of data and the goodness of fit was measured for each feature. The goodness

of fit was found to be similar for all the 10 new sets of data (for over 90% of the features, the R-square values are found to be above 0.8 for the mean and above 0.7 for the standard deviation of absolute error). This is as expected, since both the size of data (2,000 cases) and the clinical sample pool (more than 100 cases for each parameter pair) is large enough to reduce the variance in regressing the model. Therefore, the models are robust with different drawings of data and the results are reproducible.

5.3.3 Discussion

To estimate the error generated by signal loss using a model, the balance between model complexity and prediction accuracy needs to be found. In this section, to build a simple, intuitive model, two variables were selected: proportion of signal loss (x) and number of signal loss segments (y). The proportion of signal loss was selected because it is the most important indicator of signal quality, while the number of signal loss segments was selected since it is an indicator of the spread of signal loss.

Once the variables of the model were selected, which decides the complexity of the model, other parameters needed to be simulated to be as similar to the actual clinical situation as possible to improve the prediction accuracy. The length and location of signal loss segments were difficult to describe and simulate, thus clinical FHR windows were used as a template pool to generate artificial signal loss. Owing to the fact that it is impossible to recover the actual signal under the signal loss, using templates is the best possible way to simulate clinical signal loss. Since the size of the template pool is large, it should give a reliably high number of representative clinical signal loss patterns. This has been confirmed using 10 sets

of different drawings with the results remaining solid for each set.

An exponential model was built for the relationship between (proportion of) signal loss and (mean and standard deviation of) absolute error. The intuitive explanation of the model is that the absolute error grows with the increasing proportion of signal loss, and stops growing when it reaches a maximum level. This is intuitive both statistically and clinically. The two parameters of the model (λ_1 and λ_3) represent the maximum level of error and the initial growth rate (sensitivity) of error against signal loss respectively. Validation data suggest that this model has a high goodness of fit (R-square value > 0.9 for over 90% of the features).

It is found that there is a consistent linear relationship between the number of signal loss segments and the two parameters of the exponential model, thus a bivariate model was built on the exponential model to quantify the relationship between absolute error and proportion of signal loss/ number of signal loss segments (x,y). The assumption of the model is that y has a linear relationship with both λ_1 and λ_3 in the exponential model.

The bivariate model was validated using clinical data degraded by simulated clinical signal loss. For over 90% of the features, the R-square values of the exponential model are found to be above 0.8 for the mean of absolute error, and above 0.7 for the standard deviation of absolute error. This indicates that the goodness of fit of the model is solid. Therefore, it can be concluded that the relationship between proportion of signal loss, number of signal loss segments and absolute error can be adequately described by bivariate models.

As a result, for each feature, given the proportion of signal loss and the number of signal loss

segments, the distribution of error can be measured using mean and absolute error. These two parameters were chosen since the quantile of discrete features can be less accurate to measure in a model. In this way, the influence of signal loss on each FHR feature can be accurately quantified.

In the model, it is found that signal loss has influence in various ways. In general, absolute error grows with the increase in signal loss. For some features, the influence is higher when signal loss appears in large segments (e.g. Feature 1 baseline mean), for some others, the influence is higher when signal loss appears in small segments (e.g. Feature 2 percentage of zero difference between neighbour points). This is owing to the different calculation methods of various features. This influence can be quantified using the bivariate models.

The influences of signal loss amongst different features have been compared in two ways: initial sensitivity against signal loss and initial sensitivity against number of signal loss segments. It is found that some of the features are more sensitive than others. This is because they were calculated using different methods. Some of the calculation methods are more likely to be affected by signal loss, e.g. the acceleration and deceleration features, for which it is possible for the signal loss part to cover the acceleration. Some of the calculation methods are less sensitive to signal loss, e.g. the baseline mean for which interpolation can be used to minimize the influence of signal loss.

The major limitation of this study lies in the simulation of clinical signal loss using clinical templates, which ignores the correlation between signal loss and FHR. Nevertheless, using clinical templates is currently the best way possible to simulate signal loss given the fact that

it is impossible to find out the actual signal. Further studies should focus on finding the relationship between FHR and signal loss. Other methods, such as fetal ECG, could be used as a gold standard to find out the actual signal under the FHR signal loss in the future.

5.4 Influence of Signal Loss on FHR Labour Outcome Classification

In Section 5.3, signal loss was generated based on the clinical dataset to investigate the relationship between signal loss and error in FHR feature values. A bivariate model was built and validated to quantify the relationship between signal loss and error. Based on this study, it would also be valuable to know how signal loss will affect the classification of labour outcome using FHR features. Therefore, in this section, 510 FHR cases with labour outcome information were used to test the influence of signal loss on FHR labour outcome classification. FHR windows of different signal quality were tested for their classification performance on labour outcome using Genetic Algorithms. In this way, the influence of signal loss on labour outcome prediction can be estimated and quantified.

5.4.1 Data

To ensure that all cases are analysed at comparable stages of labour, only the last FHR window (last 15 minutes of Stage 2) before birth was examined for each case. The assumption of this study is that, in the last 15 minutes of second labour stage, adverse labour outcome is detectable using FHR. The selection criteria were consistent with a previous study of the Oxford Centre for Fetal Monitoring Technologies group using Artificial Neural Networks (ANN) for FHR classification: a total set of 7,568 FHR records (OIFD3) were thus selected

(Georgieva et al., 2013c). The dataset is the same as described in Section 4.2.

The distribution of signal quality in these 510 cases is shown in Figure 5.14. The 510 cases were divided into five groups based on signal quality: 50%-60%, 60%-70%, 70%-80%, 80%-90%, 90%-100%. It is shown that the distribution of the signal quality is similar across different signal quality groups, except that in the lower signal quality group the proportion of adverse outcomes appears to be higher. This is probably owing to the fact that labour with adverse outcomes tends to occur with more drastic clinical situations and more clinical interventions, which can also result in more signal loss.

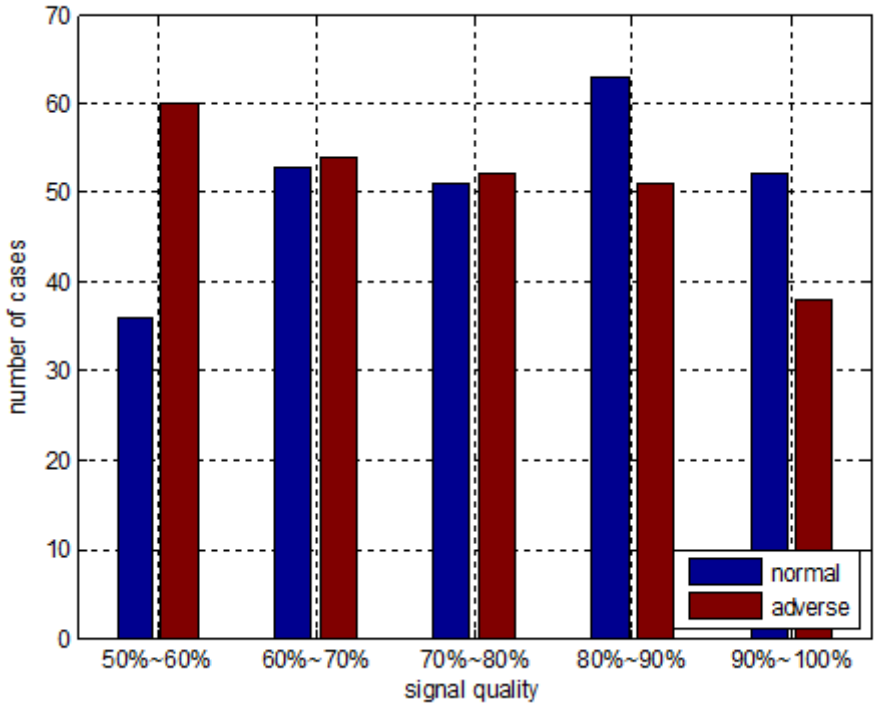


Figure 5.14 Distribution of signal quality in the 510 FHR cases.

5.4.2 Methods

The study consists of three steps:

A. Investigate the influence of signal loss on the predictive power of each feature, to see which features are more robust against signal loss.

B. Investigate the influence of signal loss on the feature ranking, i.e. degrade the signal and see how feature ranking (from GA) changes.

C. Investigate the influence of signal loss on classification, i.e. degrade the signal to different levels, and see how GA will select different feature subsets, and how well the classification performances of these feature subsets will be for different levels of signal quality.

Five questions were thus proposed for these three steps. An outline of this section is shown in Figure 5.15. The study is structured under these five questions corresponding to the five black boxes: (1) How robust is the univariate predictive power of each feature against signal quality? (2) How will the feature ranking change against signal quality? (3) Is the feature subset still robust under low signal quality conditions? (4) Under what signal quality condition can the data be used for feature selection? (5) Is the classifier still robust under low signal quality conditions?

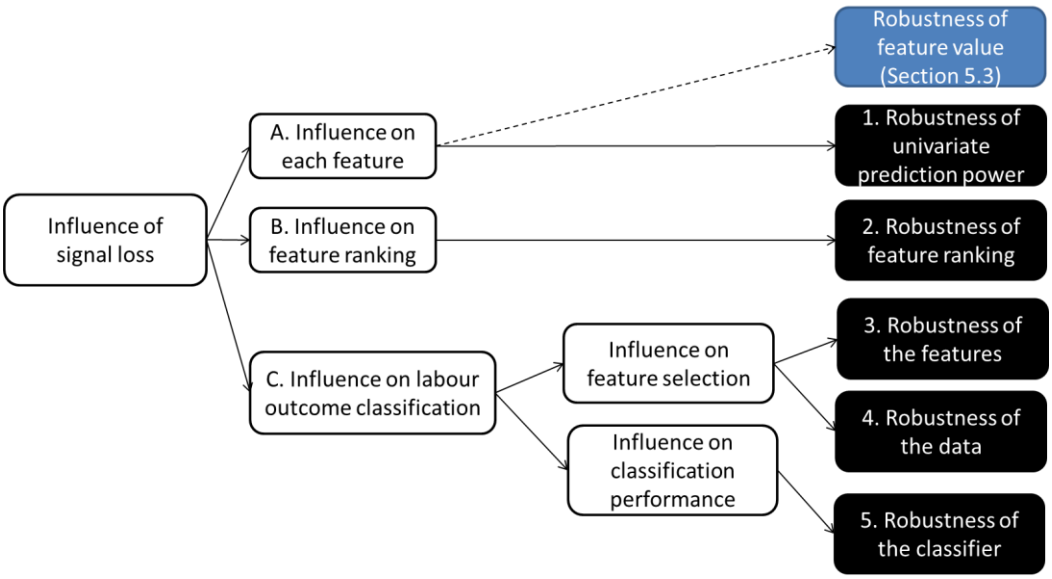


Figure 5.15 Outline of Section 5.4.

Robustness of univariate prediction power

The influence of signal loss on the univariate prediction power of individual features was evaluated using the ROC curve. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. It is a balanced measurement of prediction power of a univariate classifier. In this case, each feature is used as a univariate classifier, and the ROC curve is calculated and plotted for each signal quality group of each feature. The AUC value was also calculated to quantify the predictive power. To test the trend of the AUC against signal quality to see if there is a significant trend in univariate prediction power against signal loss, the two sided Mann-Kendall Tau trend test (Gilbert, 1987) was used.

Robustness of feature ranking

The influence of signal loss on feature ranking was evaluated using Genetic Algorithms. For the GA, three classifiers were used: linear regression, linear Support Vector Machine (linear

SVM) and Radial Basis Function Support Vector Machine (RBF SVM). This is consistent with the previous study on feature selection using Genetic Algorithms in Chapter 4.

The method and parameters of GA used here are also consistent with Chapter 4. For each signal quality group, the entire data of the group were used as the training set. The GA was trained 100 times on this training set, i.e. 100 feature selection processes were conducted and 100 best feature subsets were selected. The feature importance was measured using the feature frequency, i.e. the frequency at which a feature was selected in these 100 best feature subsets.

Robustness of the features

To evaluate the robustness of the feature selection process, the features selected in the four lower signal quality groups (50%-60%, 60%-70%, 70%-80%, 80%-90%) are compared with the features selected using the 90%-100% group (used as a gold standard) in terms of classification performance.

The features selected by each group were compared with features selected using the gold standard. These feature subsets were trained and compared in each signal quality group using cross-validation. This was to find out whether using feature subsets selected from each signal quality group will be better than using the same feature subset selected by the gold standard.

For each group of data, a five-fold cross validation was used to compare the classification performance of the features selected using data from this group and the features selected using data from the gold standard. In each cross-validation, a feature subset was selected by GA using the 80% training set (known as self-trained GA). The performance of this feature

subset was then compared with the performance of the feature subset selected using the gold standard on the 20% testing set.

Robustness of the data

The classification performances of features selected in each signal quality group were tested in other groups. This was to find out how robust is the data in each group, i.e. under what signal quality condition can the data be robust enough for feature selection. In this section, the classifier trained using GA in each of the five groups was tested on other groups. The classification performances were then compared to investigate whether a classifier was still valid in other groups.

Robustness of the classifier

To test the robustness of the classifier, features were selected and classifiers were trained using the data in the 90%-100% group. The sensitivity and specificity were examined respectively to give a clear explanation of how classification performance changed. In this case, using unbalanced data to train a classifier can result in a good performance in proportion of agreement but a biased sensitivity and specificity. As there are 52 normal cases and 38 adverse cases, to avoid the bias of an unbalanced dataset, 38 normal cases and 38 adverse cases were randomly selected to form a balanced dataset of 76 cases.

Robustness of these classifiers was evaluated using the data with lower signal quality: the 76 cases were degraded into signal quality groups of 10%, 20%, 30%... 90% using the degradation method based on clinical templates mentioned previously in Section 5.3. The classifier (trained using the original 76 cases) was evaluated respectively in these groups to

find out how classification performance changes with signal quality. The degradation process was repeated 100 times to explain the variance of different templates of the database.

5.4.3 Results

Robustness of univariate prediction power

For each feature, the ROC curve was calculated, plotted and examined. A typical example is shown in Figure 5.16. In Chapter 4, the predictive power for individual features was not so clear (more than 70% of the features have an AUC value less than 0.6). Therefore, as expected, the influence on univariate prediction power of most features is not clear, since most features have only small predictive power individually. It is thus more important to investigate the influence of signal loss on combined feature subsets.

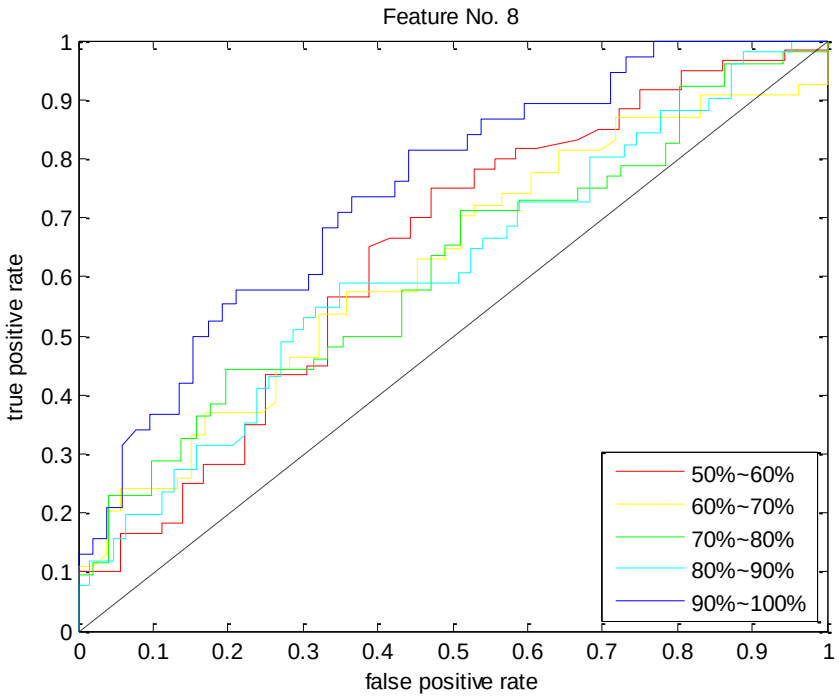


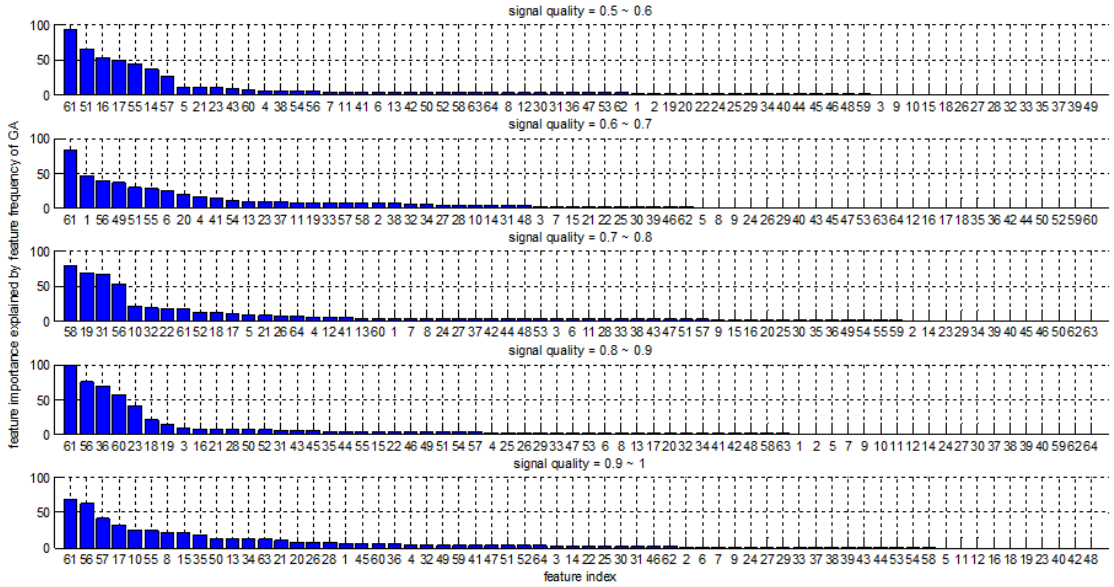
Figure 5.16 ROC curve of individual feature in labour outcome classification: Feature 8 (long-term variability).

The AUC values for different signal quality groups are listed in Appendix Table 1. Also listed are the *p*-values for the two sided Mann-Kendall Tau non-parametric trend test. It was found that there is no significant trend ($p < 0.05$) of AUC values against signal quality for individual features.

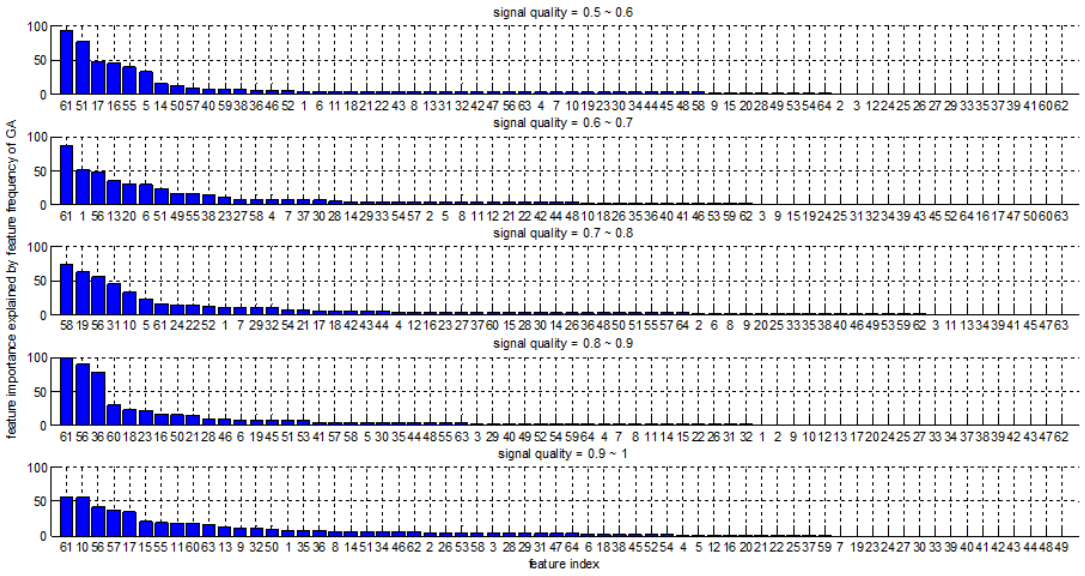
Robustness of feature ranking

The feature ranking under different signal quality conditions is shown in Figure 5.17. It is shown that for all three classifiers, the feature rankings are different amongst different various signal quality conditions. However, the features most frequently selected are consistently found over all values signal quality (when signal loss < 50%). For example, Feature 61 (Phase Rectify Signal Averaging - DC component) stays at the highest ranking for most conditions.

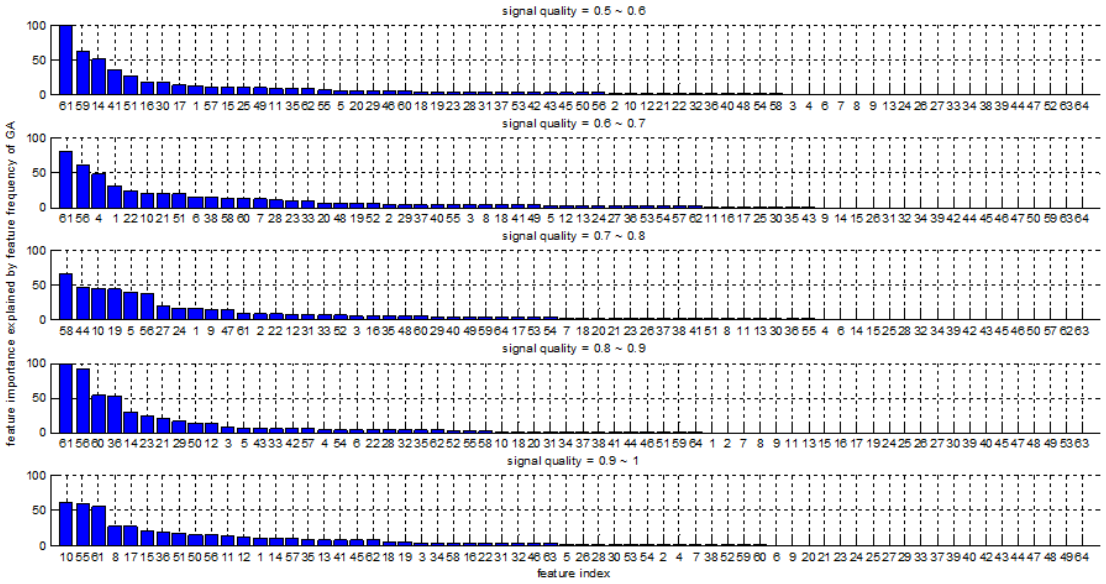
In addition, the features that appear in the top 10 for at least four groups out of five are: Feature 61 and Feature 56 (Alphabet feature) for all three classifiers. This indicates that it is likely that these most important features will not be heavily influenced when signal loss < 50%. Referring back to Chapter 4, this is also found to be the case with feature ranking when all cases with signal quality above 50% were selected (Figure 4.12).



(A) Linear regression



(B) Linear SVM



(C) RBF SVM

Figure 5.17 Feature importance using feature frequency in the GA.

As described in Chapter 4, adding BIC and backward elimination to prevent overfitting, one feature subset will be determined for each signal quality group. The feature subsets selected by GA in each group are listed in Table 5.3. It should be noted that for each classifier, there are several features selected consistently regardless of the signal quality conditions. For instance, for all three classifiers, Feature 61 was selected in most conditions (4 out of 5 at least). Apart from that, for Linear SVM, Feature 51 (standard deviation of local sample entropy) was selected in most conditions, as it is for Feature 55 in RBF SVM. This indicates that some features (shown in bold) are robust against signal quality when the signal loss is < 50%.

Table 5.3 Feature subsets selected by different signal quality groups. Features listed are using

feature index. Features selected in at least five out of six subsets are shown in bold.

Signal Quality Group	Linear regression	Linear SVM	RBF SVM
50%-60%	[14,17,51,57,**61**]	[13,14,15,16,17,29,36,38,50,**51**,59,**61**]	[2,7,16,25,34,37,38,41,46,**55**,59,**61**]
60%-70%	[1,20,54,55,**61**]	[1,28,**51**,57,**61**]	[1,6,9,10,20,37,**55**,58,**61**,63]
70%-80%	[56,58,60]	[1,6,9,10,19,31,32,**51**,57]	[2,11,16,40,44,45,51,**55**]
80%-90%	[18,21,36,56,**61**]	[20,23,35,36,43,**51**,60,**61**]	[10,13,14,30,37,41,50,56,57,**61**,63]
90%-100%	[10,50,55,57,**61**]	[10,11,34,35,**51**,57,**61**]	[1,10,19, 48,50,**55**,**61**]
All 510 cases	[5,9,10,48,51,**61**,63]	[5,9,10,16,44,48,50, **51**, **61**]	[1,10,19,21,48,50,51,**55**,**61**]

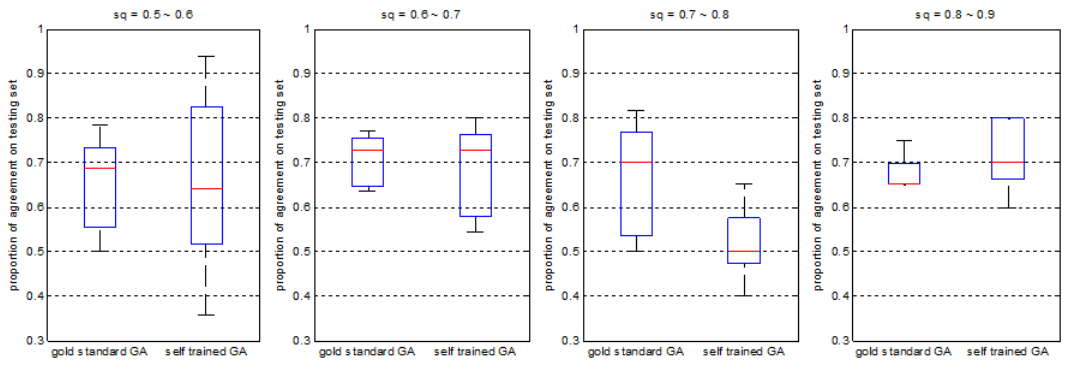
It is also noted that, in each feature subset selected, there are always some features whose feature values are found to be robust against signal loss. According to Table 5.1, features such as Feature 1, 5, 9, 10, 14, 16 are relatively robust against signal loss (less than 20% mean absolute error of when signal loss is 40%). Many of the FHR features have a high value of correlation with each other as stated in Figure 4.5. Therefore, under lower signal quality conditions, the features that are more robust against signal loss can provide similar information for features that are less robust against signal loss.

For example, Feature 61 (PRSA – DC component) has a correlation coefficient with Feature 5 (Minimum expected value) of 0.28, and they were both selected in all 510 cases of linear regression and linear SVM. It is likely that in lower signal quality conditions when the prediction power of Feature 61 was compromised, Feature 5 would provide similar information for the classifier. This could result in a certain level of robustness of the classifier against signal quality, as will be discussed later.

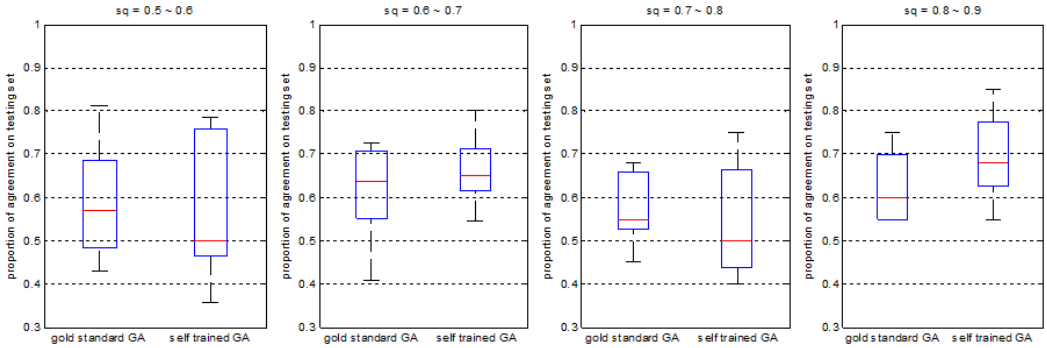
Robustness of the features

For all three classifiers, different feature subsets were selected under various signal quality conditions (Table 5.3). Therefore, it is worthwhile considering whether the different subsets should be used under different signal quality conditions, or the same feature subset should be used for all signal quality conditions. The robustness of selected features against signal quality was evaluated within the four groups (50%-60%, 60%-70%, 70%-80%, 80%-90%) respectively. In a five-fold cross validation of each group, the features selected by GA in the 90%-100% group (gold standard GA) were compared with the features selected by GA using the training group of cross validation (self-trained GA).

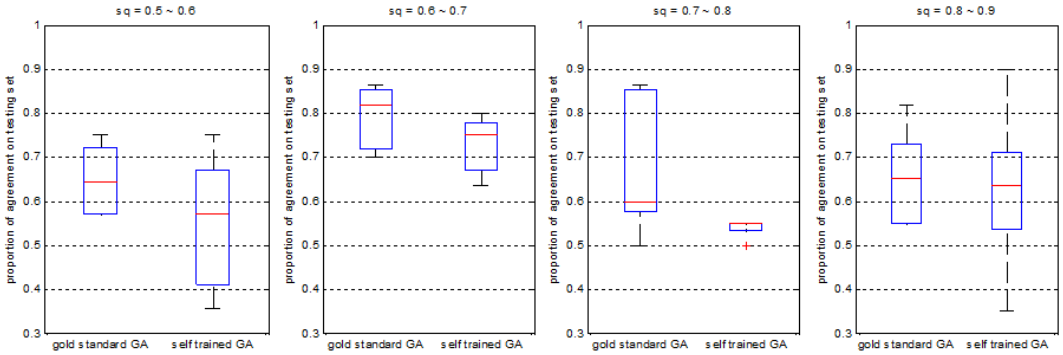
The classification performance was evaluated by the proportion of agreement in the five testing sets of each group. The classification performances for all three classifiers are shown in Figure 5.18. It was found that for all three classifiers, there is no clear difference between the classification performance between the self-trained GA and the gold standard GA in general. This indicates that the predictive powers of the self-trained GA and the gold standard are similar, thus there is no need to use different feature subsets for different signal quality conditions (under the condition that signal loss < 50%).



(A) Linear regression



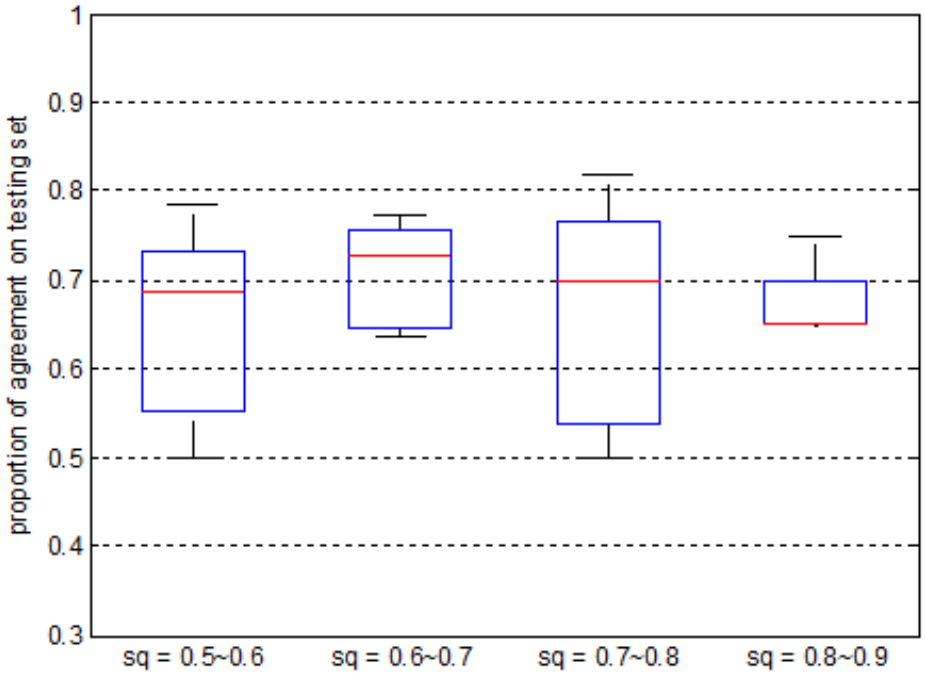
(B) Linear SVM



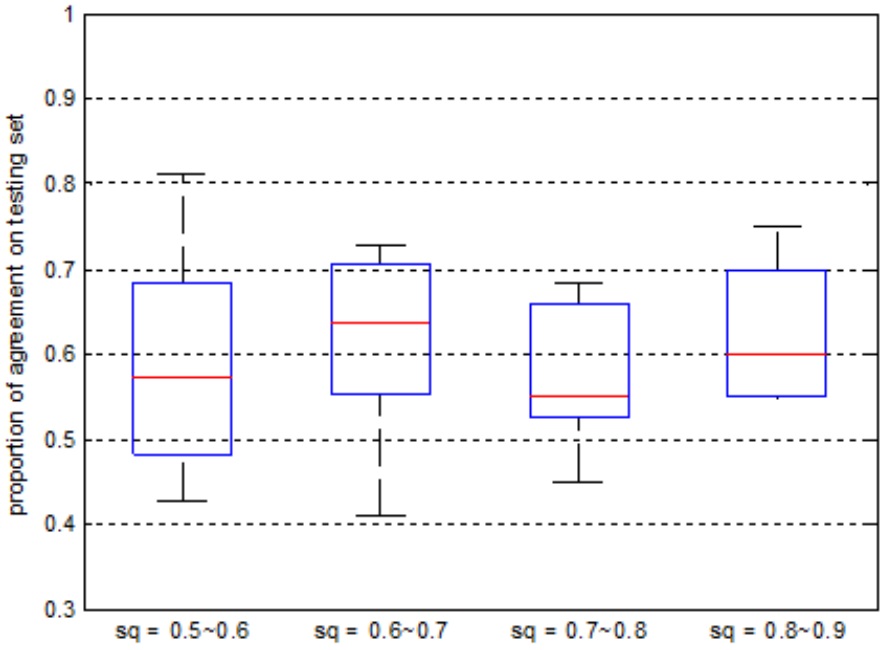
(C) RBF SVM

Figure 5.18 Comparison between gold standard GA and self-trained GA in each signal quality group: (A) Linear regression; (B) Linear SVM; (C) RBF SVM. (sq = signal quality)

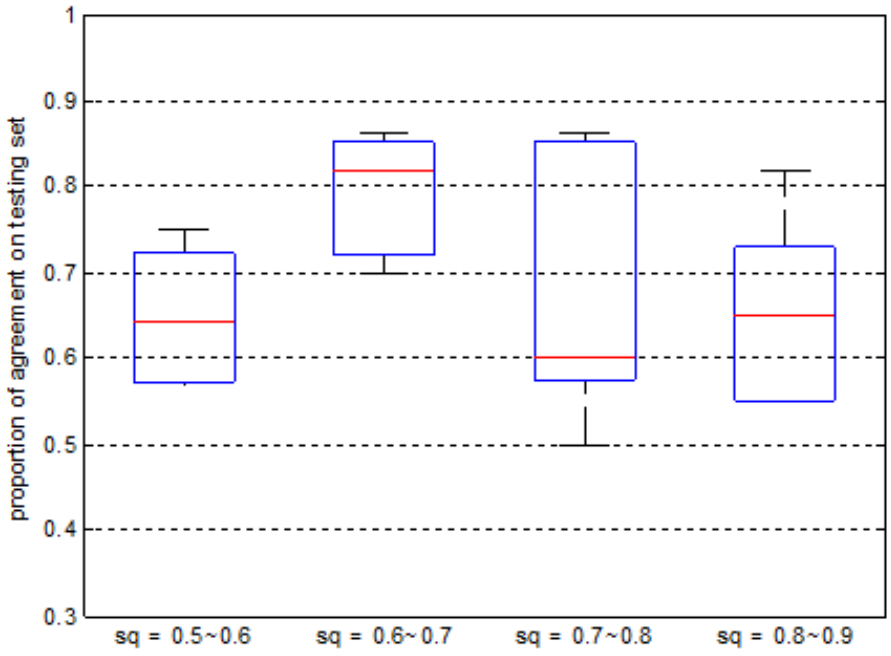
In addition, comparing the performance of the gold standard GA in different signal quality groups (Figure 5.19), it is found that, using the features selected by the gold standard, there is no clear trend (two-sided Mann-Kendall Tau non-parametric trend test, $p > 0.5$) for different signal quality groups in terms of classification performance (under the condition that signal loss < 50%). This indicates that under the condition that signal loss < 50%, both the features selected and the classifier trained by the gold standard are reasonably robust against signal loss.



(A) Linear regression



(B) Linear SVM



(C) RBF SVM

Figure 5.19 Comparison between classification performances of different signal quality groups: (A) Linear regression; (B) Linear SVM; (C) RBF SVM. (sq: signal quality)

Robustness of the data

To evaluate the robustness of data against signal quality, features were selected by the GA in the five signal quality groups respectively, and the features selected were trained and tested using the data from the other four groups. Classification performances are compared in Figure 5.20. It is found that in general, there is no clear trend that the classification performance increases with the increase in signal quality, according to the two sided Mann-Kendall Tau non-parametric trend test (p-values are 0.2207, 0.8065 and 0.0864 for linear regression, linear SVM and RBF SVM). On the other hand, the classification performances mostly lie in the range of 60%-70% proportion of agreement. This indicates that the data are reasonably robust for feature selection when signal loss is < 50%.

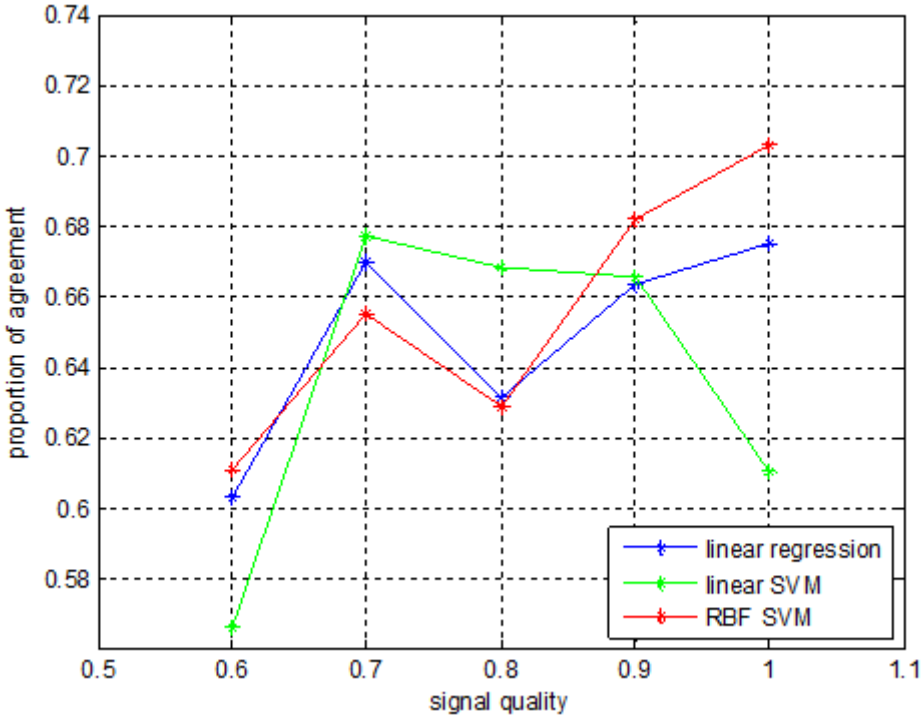
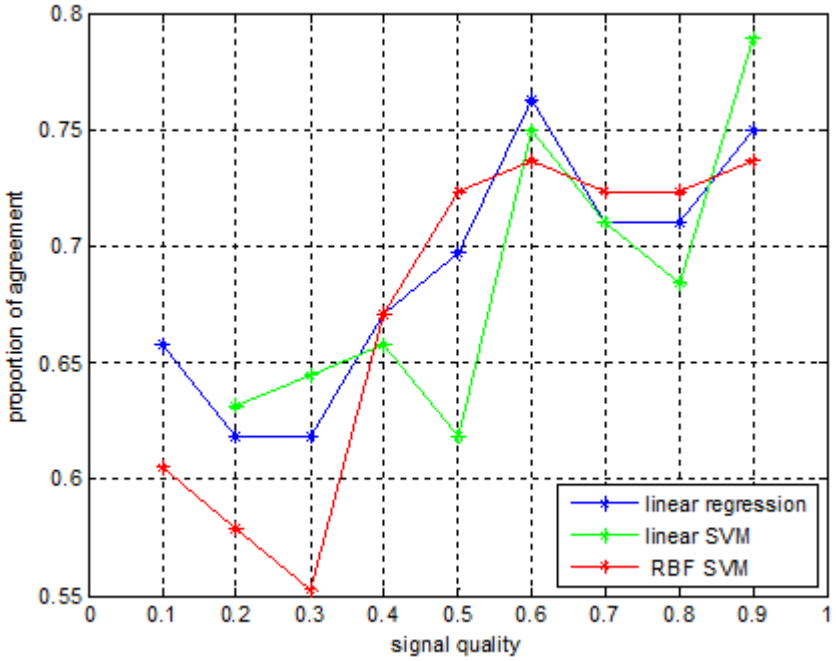


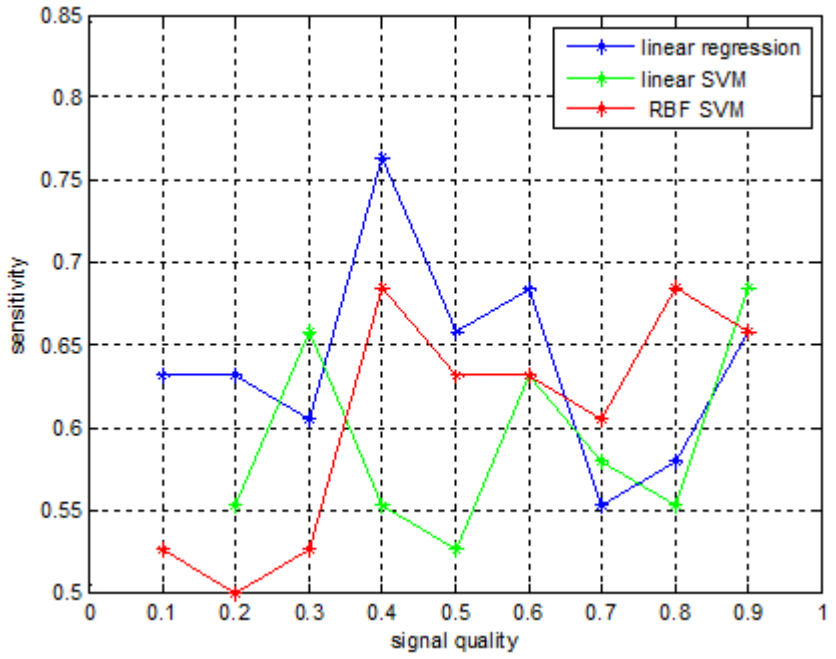
Figure 5.20 Classification performances of classifiers trained by each signal quality group and tested on other four groups.

Robustness of the classifier

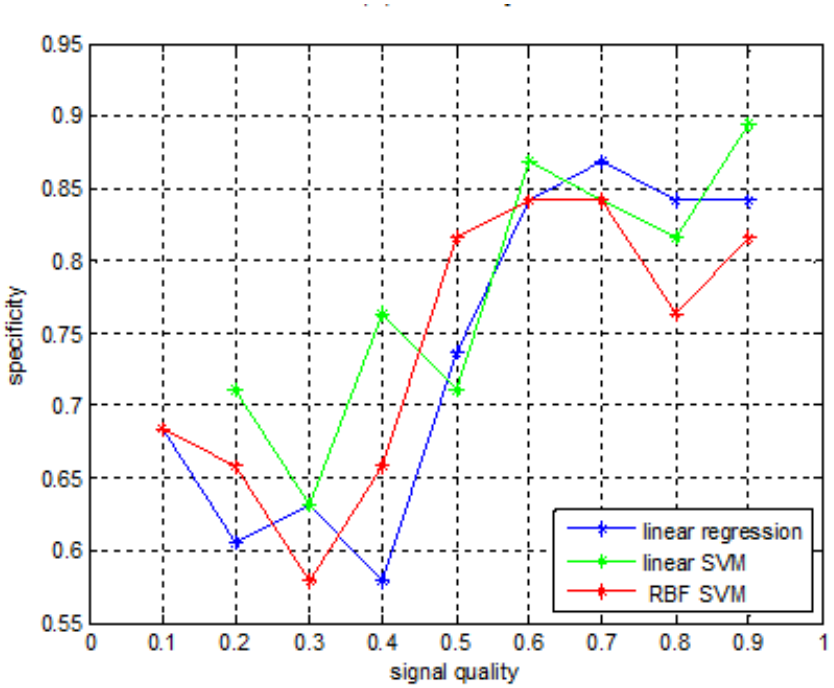
The robustness of the classifier against signal loss was evaluated using the degraded 90%-100% group data (76 cases). The data were degraded to 10%, 20%, 30%...90% groups and the classifier trained by original data (using GA selected features) was tested on each group. The classification performance against signal quality is shown in Figure 5.21.



(A) Proportion of agreement



(B) Sensitivity



(C) Specificity

Figure 5.21 Robustness of the classifier against signal quality, tested on degraded (based on clinical dataset) data in terms of: (A) proportion of agreement; (B) sensitivity; (C) specificity. It is observed that when the signal quality is 0.1, the linear SVM will output NaN values, which is owing to that Feature 35 (Total number of lost beats) outputs extreme values.

It is shown in Figure 5.21 (A) that, in general, classification performance increases against signal quality for all three classifiers (two-sided Mann-Kendall Tau non-parametric trend test, p = 0.0153, 0.0165, 0.0246 for linear regression, linear SVM and RBF SVM respectively), especially when signal loss is more than 50%. When signal loss is lower than 50% (signal quality over 50%), however, the change in classification performance is not significant (two-sided Mann-Kendall Tau non-parametric trend test, p = 0.6134, 0.4624, 0.7728 for linear regression, linear SVM and RBF SVM respectively).

It is also shown in Figure 5.21 that the increase of classification performance mostly lies in

the increase of specificity, i.e. the true negative rate or the ability not to output false positives. The trend of increase in sensitivity against signal quality is not significant (two-sided Mann-Kendall Tau non-parametric trend test, $p > 0.5$). In addition, the increase in sensitivity against signal quality is also not significant (two-sided Mann-Kendall Tau non-parametric trend test, $p > 0.5$) when signal loss is < 50%, which is consistent with the proportion of agreement.

The degradation process was repeated 100 times to ensure that the trend does not result from different templates of the database. The two-sided Mann-Kendall Tau non-parametric trend test results were the same for all 100 tests: the trend of proportion of agreement and sensitivity against signal quality is significant with signal quality of all range, while when the signal loss is < 50%, the trend of proportion of agreement and sensitivity against signal quality is not significant. A boxplot reflecting the variance in the 100 tests is shown in Figure 5.22. It is found that despite the variance, the increasing trend in error against signal quality is not clear when signal loss is < 50% (signal quality > 50%), while when signal loss is over 50%, the trend is much more clear. This indicates that 50% can be a reasonable threshold for data selection.

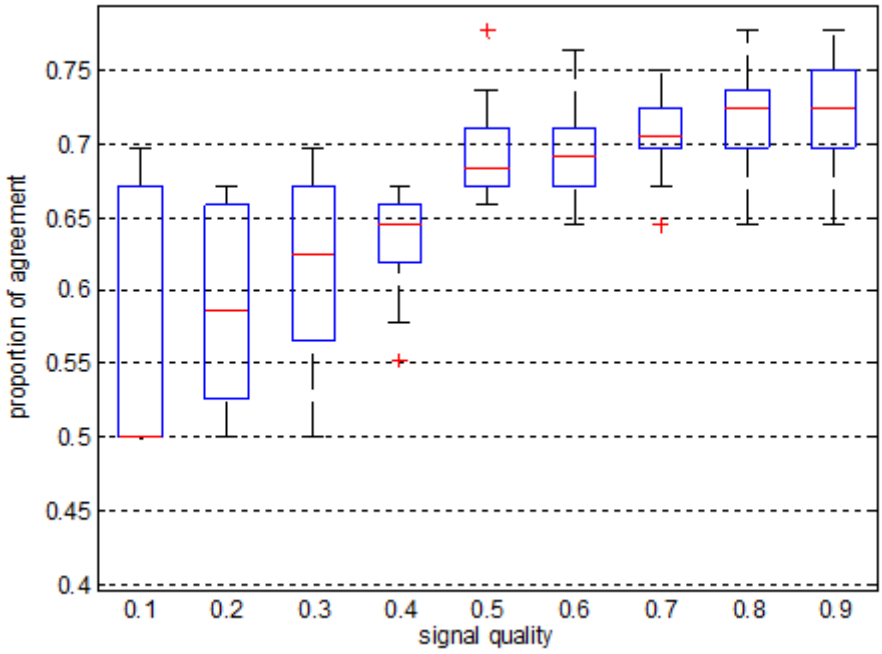


Figure 5.22 Robustness of the classifier against signal quality, tested on degraded data in terms of proportion of agreement using RBF SVM.

It is found that the mean of classification rate (proportion of agreement) can be approximated using another exponential model:

$$z = \lambda_1(1 - e^{-\lambda_2 x}),$$

where z is the mean of absolute error and x is the proportion of signal loss. The goodness of fit of the model is shown in Figure 5.23. The R-square value of the model is 0.9763, indicating a good fit of the model. Therefore, provided the proportion of signal loss, the classification error can then be estimated.

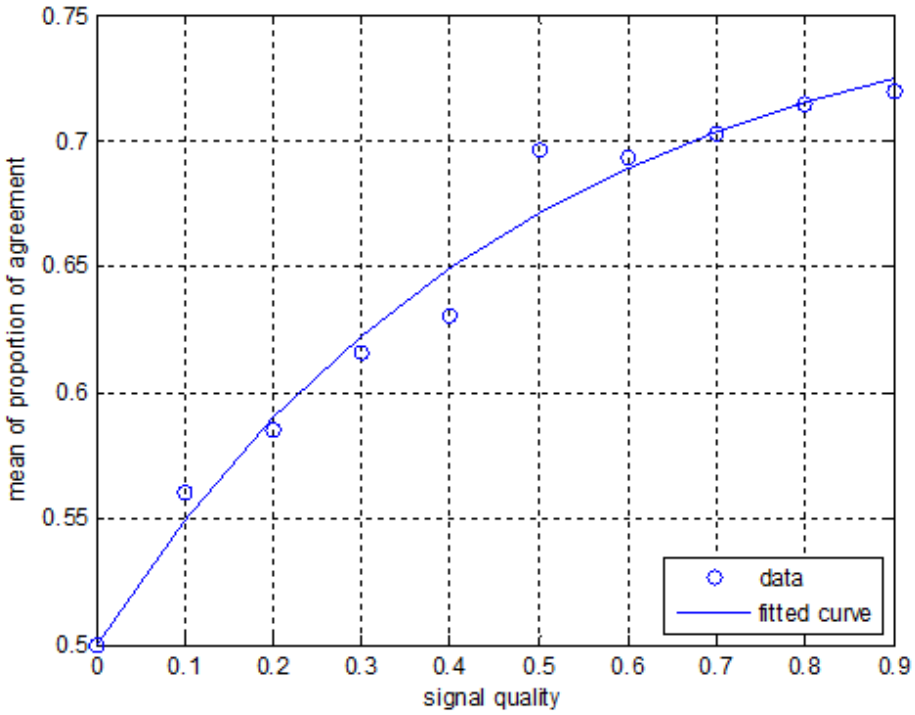


Figure 5.23 The exponential model to describe the relationship between signal quality and mean of proportion of agreement.

5.4.4 Discussion

As shown in Figure 5.15, the study presented here answers five questions: (1) How robust is the univariate predictive power of each feature against signal quality? (2) How will the feature ranking change against signal quality? (3) Is the feature subset still robust under low signal quality conditions? (4) Under what signal quality condition can the data be used for feature selection? (5) Is the classifier still robust under low signal quality conditions?

For (1), as the univariate predictive power of individual features is quite weak (Appendix Table 1), the influence of signal quality on an individual feature's univariate predictive power is not clear. The distribution of error against signal quality can be estimated by the bivariate

model described in Section 5.3.

For (2), the feature importance for each signal quality group was evaluated. It was shown that while the feature importance changes with signal quality, Feature 61 (PRSA – DC component) (Georgieva et al., 2013b) was always selected. Other features that measure variability (e.g. Feature 8 - long term variability) were not selected as often. This indicates that PRSA could be a more effective method to extract the time-series information of heart rate variability that is related to labour outcome. Also, for different classifiers, some of the features remain at the highest ranking. This indicates that some features are robust against signal loss when signal loss is < 50%.

Some features were selected differently in various groups, which could be due to several reasons. Firstly, the sizes of these groups were rather small (about 100 cases each), which could cause variances in feature selection, selecting some features by chance. Secondly, as shown in the previous GA chapter (Figure 4.5) some features have high correlation coefficients, which indicates that come features can have similar information to others, thus in some cases one feature was selected, in other cases another with similar information was selected.

For (3), to evaluate the robustness of features selected by GA, the features selected by the 90%-100% group (gold standard) were used in other groups and compared with the features selected in these groups. It is found that there is no clear difference between the gold standard features and the self-trained features in each group. This indicates that there is no need to select different feature subsets for different signal quality groups, when the signal loss is <

50%. This is consistent with findings in (4) and (5) below.

For (4), to evaluate the robustness of data used in feature selection, the features selected by each signal quality group were tested in other groups. It is shown that in general, the classification performance increases with signal quality, but that there is no clear difference between the five groups. This indicates that the FHR cases with signal quality > 50% (signal loss < 50%) can be included in the feature selection process, and the greater the signal quality, the more accurate the classifier.

For (5), to test how signal loss influences the classifiers, the 90%-100% group was degraded into nine groups with different signal quality. It is shown that in general, the classification performance increases with signal quality, but when signal loss < 50% the increase is modest. This indicates that the classifiers are reasonably robust when signal loss < 50%, thus the classifiers are not likely to be accurate when applied to FHR cases with lower signal quality. In addition, the increase in classification lies mainly in the increase of specificity. This is probably owing to the fact that in lower signal quality cases there are more large error, thus more cases will be classified into adverse cases owing to extreme values, resulting in reporting of more false negative and less specificity.

Another exponential model was used to quantify the relationship between signal quality and classification rate. It should be noted that this model is different to the exponential model in Section 5.2. The model has a good value of fitness with an R-square value of 0.9763. In this way, an estimated error of the classification can be given in relation to the particular signal loss.

In conclusion, the feature selection method, the classifier and the data are robust when the signal loss is < 50% (over 50% signal quality). Therefore, in future studies, only FHR records with less than 50% signal loss should be included in the feature selection process, and the classifier should only be used on data with less than 50% signal loss (i.e. over 50% signal quality).

The major limitation of this study is that the data size of each group is small (about 100 cases each), since in the last few hours before delivery, it is difficult to find cases with 90%-100% signal quality. This might have caused the variance in feature selection and classification. In other words, features can have been selected by chance in smaller datasets. In addition, this study is based on the adverse outcome of low arterial pH only, how signal loss will influence labour outcome classification in terms of other adverse outcomes is yet to be investigated. Further studies should focus on using bigger datasets with more cases to optimise the feature selection process and the classifier, as well as investigating more adverse labour outcomes for a more comprehensive understanding of the clinical situation.

5.5 Conclusions

In this chapter, the robustness of the FHR features with respect to signal loss was investigated in three ways. As a first step, artificial signal loss was used to test the robustness of feature values for each feature, and an exponential model was used to quantify the relationship between signal loss and error (Section 5.2). The goodness of fit (R-square > 0.9 for over 90% of the features) has shown that the exponential model can accurately estimate the error of randomly generated signal loss, given the proportion of signal loss. Secondly, the signal loss

was generated using templates from the clinical database to simulate clinical signal loss, and a bivariate model was built to quantify the relationship between signal loss and error (Section 5.3). The model also yields an accurate estimation (R-square > 0.8 for over 90% of the features) of error. Thirdly, the influence of signal loss on labour outcome classification was tested using Genetic Algorithms and three classifiers (Section 5.5). It was found that in general, while signal quality is > 50%, the feature selection method with the FHR data is robust for feature selection. Future studies could focus on using techniques such as fetal ECG to find out the correlation between signal loss and FHR signal values, and to analyse the pattern of signal loss.

6 Setting Up and Validation of a Computerised Decision Support System

6.1 Introduction

The ultimate goal of FHR feature extraction and selection is to build a decision support system that helps clinicians monitor fetal health. In this way, timely and accurate intervention can be carried out to prevent birth asphyxia while avoiding unnecessary complications. A detailed review of existing FHR decision support systems can be found in Section 2.2.

In Chapter 4, Genetic Algorithms were used to select the FHR feature subset which best predicts labour outcome and GA + RBF SVM was found to be the most powerful classifier. In Chapter 5, the influence of signal loss on labour outcome classification was studied. However, there are two limitations of these studies. First, since only the last four windows (last 30 minutes) before delivery were analysed, whether or not the predictor can be used at an earlier time in labour has not been investigated. Secondly, it has not been investigated yet how to estimate the risk of fetal acidemia and to give decision support based on the predictor. Based on the studies of Chapter 4 and Chapter 5, a computerised decision support system was set up. Unlike the previous studies, the entire time period of labour will be considered in the building and validation of the system, and the risk of fetal acidemia will be estimated by the system during labour. This chapter describes the process of building a computerised decision support

system (Section 6.2) and validation of the system using new retrospective data on an internal dataset (Section 6.3) and an external dataset (Section 6.4).

6.2 Setting up the System

6.2.1 Data

The OIFD3 (7,568 cases) dataset used for GA from Chapter 4 was used to build the decision support system, selected from the database of records from 1993 to July 2008 (Section 2.5). Only cases with more than 30 minutes of Stage 2 were selected (detailed selection criteria shown in Figure 4.1). An adverse case was defined as fetal acidemia (umbilical arterial pH $<$ 7.05) (Georgieva et al., 2013a), and a normal case was defined as pH $\geq$ 7.05. As fetal acidemia is rare clinically, the dataset is heavily unbalanced: only 255 cases are adverse, and 7,313 cases are normal.

The GA + RBF SVM algorithm in Chapter 4 was used as the predictor. It should be noted that it was trained on a balanced dataset of 404 deliveries from the 7,568 cases. The feature subset used is listed in Table 6.1 below (for descriptions of the features see Section 2.1). Therefore, this chapter will not use the 7,568 cases to validate the predictor, but will only use them to set up system parameters, i.e. further to train the system.

Table 6.1 List of features selected by the GA.

Feature Number	***Feature Name***
1	Baseline mean
10	Short Term Variability (STV)

19	STV (accelerations included)
21	Mean Acceleration Duration
48	Mutual information
50	Mean of local approximate entropy
51	Standard deviation (STD) of local sample entropy
55	(STD/mean)^2
61	Phase Rectified Signal Averaging (PRSA), Deceleration Capacity (DC)

6.2.2 Analysis of the predictor

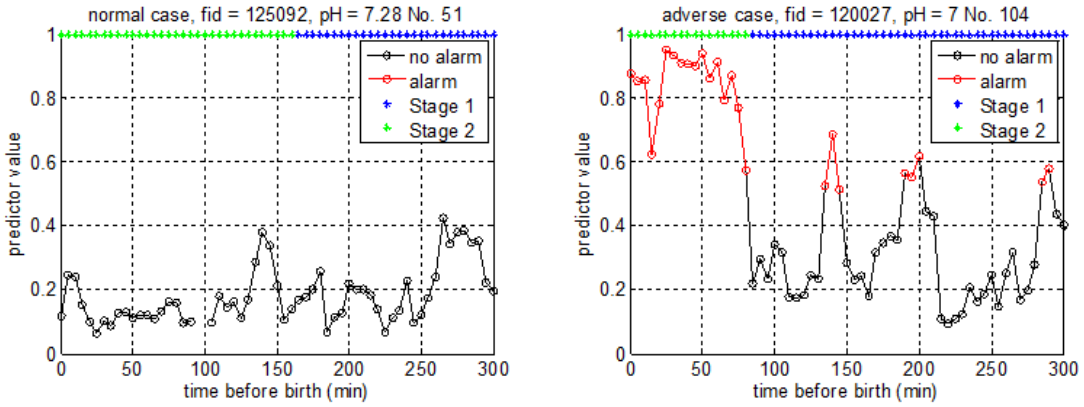
The predictor was applied on the entire dataset of the 7,568 cases. For each FHR window, a predictor value (from 0 to 1) was calculated by the RBF SVM using the GA selected features. If the predictor value is higher than or equal to the alarm threshold (temporarily set to be as 0.5, a discussion of the alarm threshold can be found in Section 6.2.5), the window will be reported as an alarm; otherwise it will be reported as no alarm. Since the lengths of the labour are different for different women, the numbers of alarms in the last four windows (last 30 minutes of labour) were compared between normal cases and adverse cases. It is shown in Table 6.2 that the proportion of adverse cases increases from 1.16% of no alarm to 7.86% of four alarms. It is also noted that 47.5% of the adverse cases have four alarms, while 39.7% of the normal cases have no alarms. This indicates that these alarms can be an effective way of recognising adverse labour outcome.

Table 6.2 Comparison of normal and adverse cases in terms of number (percentage from total) of

alarms in the last four windows of labour.

Number of alarms in last four windows	*0*	*1*	*2*	*3*	*4*	*Total*
Normal cases	2903 (39.7%)	1175 (16.1%)	892 (12.2%)	925 (12.7%)	1418 (19.4%)	7,313 (100%)
Adverse cases	34 (13.3%)	21 (8.2%)	42 (16.5%)	37 (14.5%)	121 (47.5%)	255 (100%)
Proportion of adverse cases	**1.16%**	**1.76%**	**4.5%**	**3.85%**	**7.86%**	**3.37%**

The change of the predictor value against time was then examined. This was to observe how predictor values change at different periods of labour. A few samples were randomly selected and are shown in Figure 6.1, predictor values ≥ 0.5 are defined as alarms marked in red, only windows with $\geq$ 50% signal quality are included (alarm threshold selected as in Chapter 4, signal quality threshold selected as Chapter 5). It is shown that in these samples that adverse cases tend to have more alarms than normal cases, especially in the last few windows before birth.



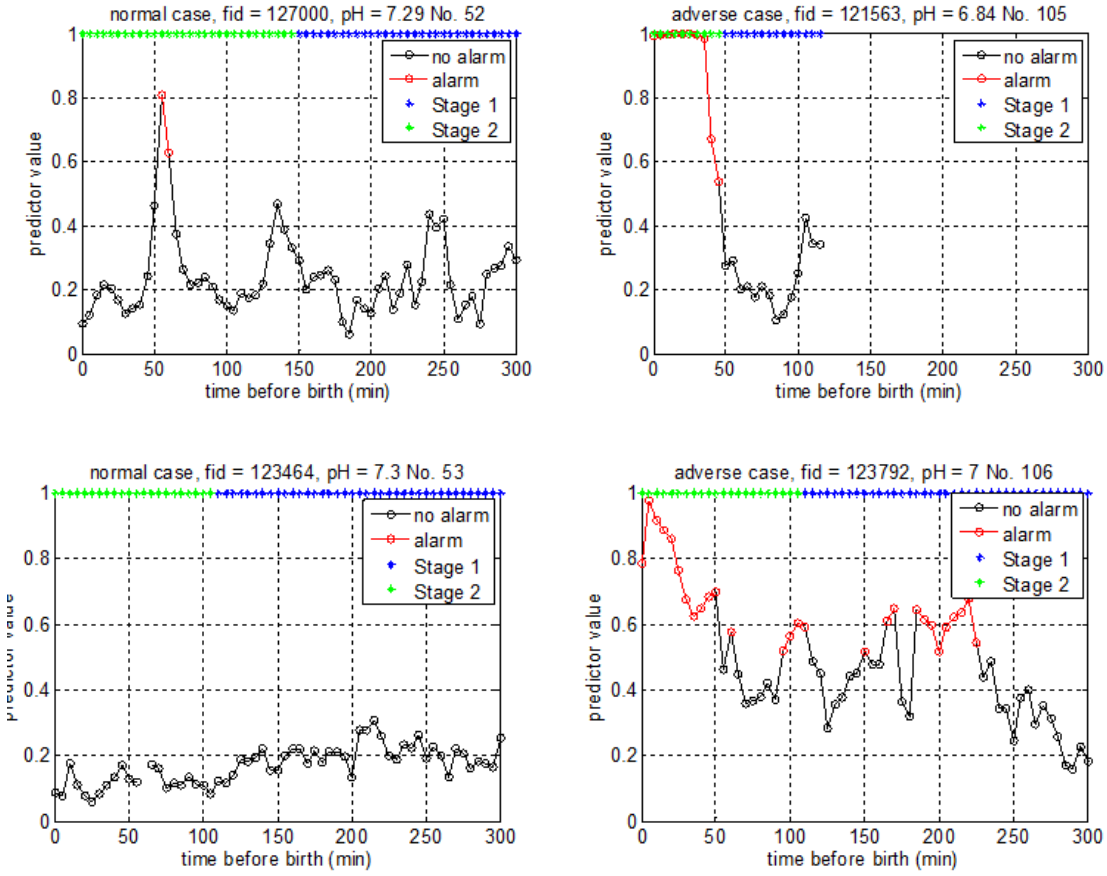


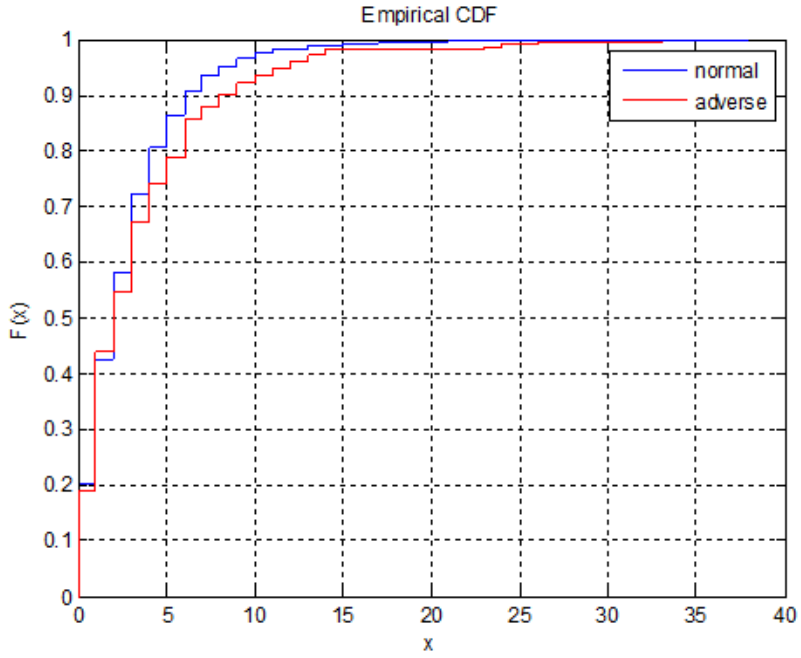
Figure 6.1 Randomly selected samples from the 106 cases, predictor value ≥ 0.5 are marked with an alarm shown in red. Left: normal cases, Right: adverse cases.

Based on Figure 6.1, several factors can be proposed as signs of an adverse labour outcome: high number of alarms, high number of consecutive alarms, and high fraction of alarms. Since the length of the CTG trace varies among different cases, with the same fraction of alarms, a longer CTG trace tends to have more alarms than a shorter one; while with the same number of alarms, a longer labour tends to have a lower fraction of alarms than a shorter one. This makes it difficult to compare different cases. Large numbers of consecutive alarms, on the other hand, can exist in either long or short labours. Clinically, this seems to indicate an adverse labour condition lasting for a long time. Therefore, it is proposed that cases with high

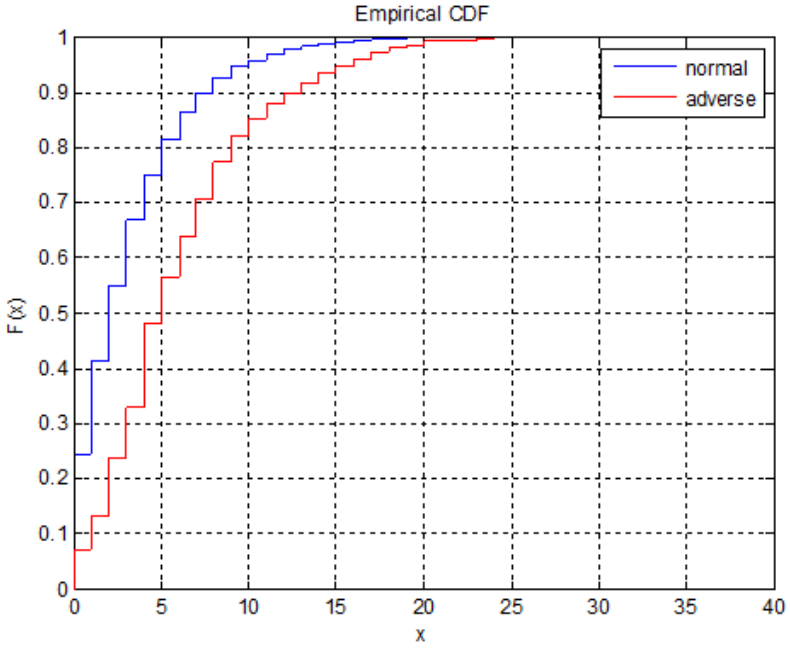
numbers of consecutive alarms are more likely to have adverse labour outcomes, thus the number of maximum consecutive alarms was used as an indicator to compare the difference between normal cases and adverse cases.

6.2.3 Analysis of maximum consecutive alarms: Stage 1

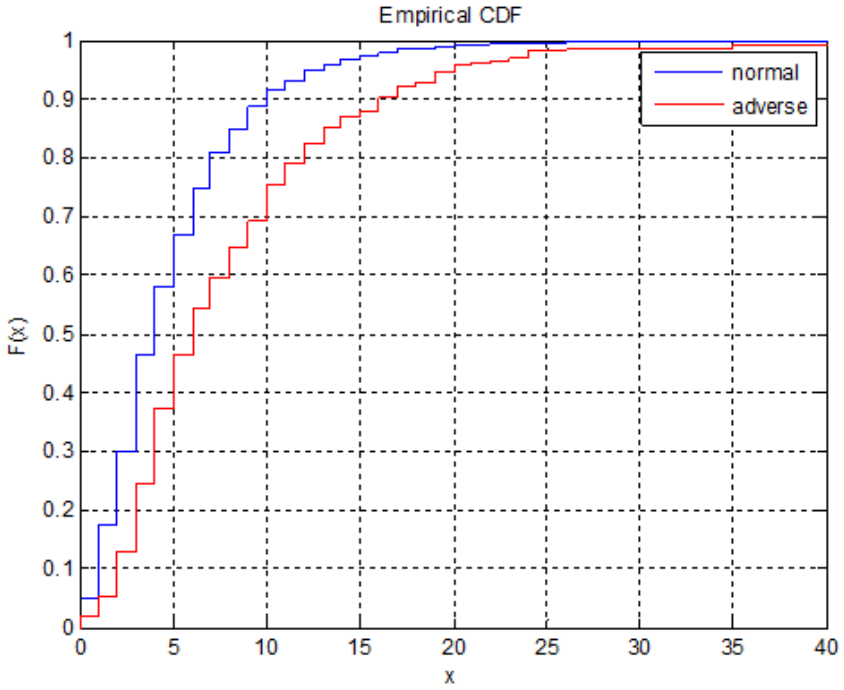
The distribution of maximum consecutive alarms was analysed in both the normal group and the adverse group. The empirical cumulative distribution function (CDF) is plotted in Figure 6.2. It is found that in Stage 2, the maximum consecutive alarms distribution is quite different between normal cases and adverse cases. For example, 25% of normal cases have no alarms while fewer than 10% of adverse cases have no alarms. Also, over 40% of adverse cases have five or more consecutive alarms, while for normal cases the proportion is less than 20%. On the other hand, in Stage 1 the distribution of normal cases and adverse cases are similar, indicating that in Stage 1 the maximum (number of) consecutive alarms has little predictive power. A two-sample Kolmogorov-Smirnov test (null hypothesis is two samples are from the same distribution, for details see Section 4.3) shows that for Stage 2, $p < 0.0001$, indicating that distribution of the normal group and the adverse group are statistically significantly different. On the other hand, for Stage 1, $p > 0.1$. This suggests that the difference in the distributions (between the normal and adverse cases) in both stages shown in Figure 6.2 (C) is mainly due to the difference in the distribution for Stage 2. Thus maximum consecutive alarms can only be predictive in Stage 2.



(A) Stage 1



(B) Stage 2

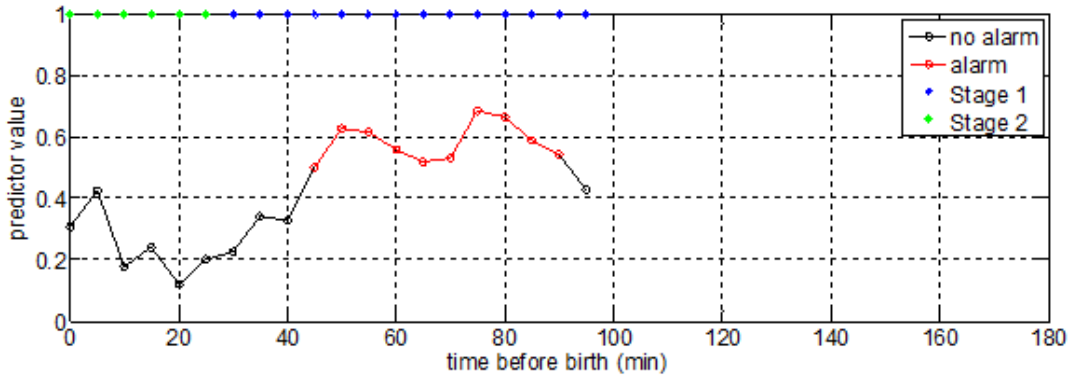


(C) Both stages

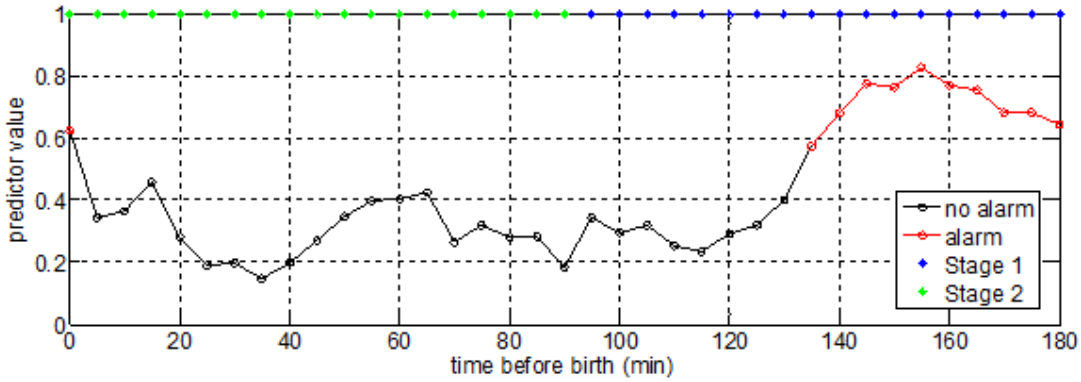
Figure 6.2 Empirical CDF of maximum consecutive alarms in normal group/adverse group. (A) Stage 1; (B) Stage 2; (C) both stages. x: maximum consecutive alarms; F(x): CDF of x.

To investigate why maximum consecutive alarms is poorly predictive in Stage 1, a few samples are shown in Figure 6.3. These are representative cases with large maximum consecutive alarms in Stage 1 but no fetal acidemia (A) (B), and cases with no alarms in Stage 1, but fetal acidemia (C) (D). It is shown in Figure 6.3 (A) (B) that a high number of consecutive alarms were shown in Stage 1 but later in labour, the predictor values have decreased in Stage 2. This suggests that some clinical measures (maternal hydration and/or oxygenation; reduction or stopping of oxytocin drip, etc.) might have been carried out in Stage 1 due to the observed abnormal FHR. These then resulted in an improved condition of the fetus in Stage 2, but since there is no digital recordings of such clinical measures in the

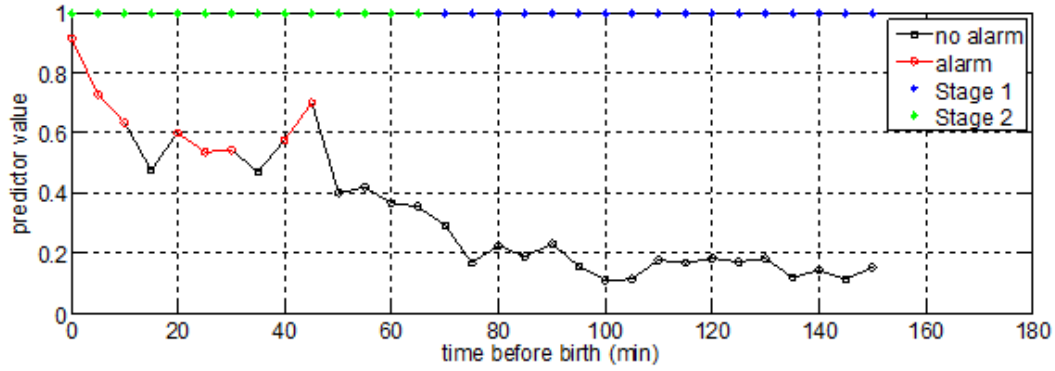
database, it is difficult to find out if this hypothesis is valid. In the other representative examples, Figure 6.3 (C) (D), no alarms were shown in Stage 1 but later in Stage 2, the predictor values have increased and the fetus had fetal acidemia. This suggests that another reason that Stage 1 alarms are not predictive is because some fetuses did not have problems until Stage 2. It can be speculated that this is often the case as Stage 2 is known to be a period of increased stress for the baby (Armstrong and Stenson, 2007).



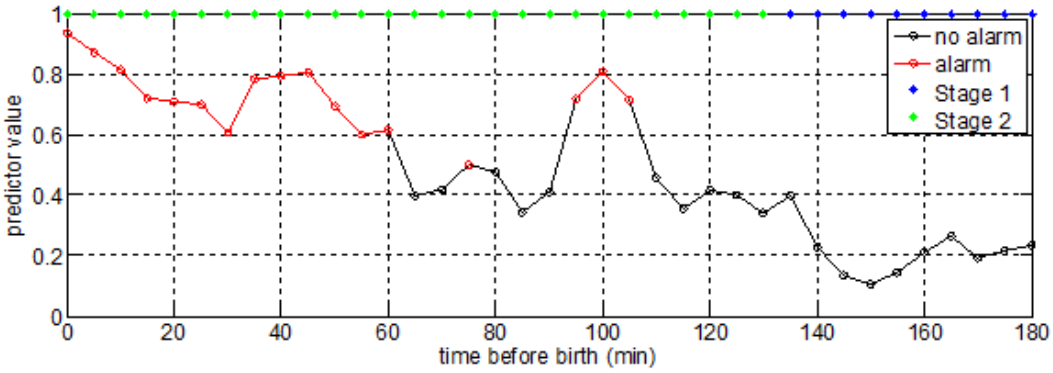
(A)



(B)



(C)



(D)

Figure 6.3 Samples of (A) (B) false positive: maximum consecutive alarms > 6 in Stage 1 but no fetal acidemia; (C) (D) false negative: no alarms in Stage 1 but with fetal acidemia.

Therefore, only maximum consecutive alarms in Stage 2 is used for prediction in this Chapter. All maximum consecutive alarms mentioned below in this chapter, if not specified, refer to maximum consecutive alarms in Stage 2.

6.2.4 Analysis of maximum consecutive alarms: Stage 2

Thresholding can be used on maximum consecutive alarms to predict labour outcome. With an increase in the threshold, the specificity will increase and the sensitivity will decrease. To illustrate the trade-off between sensitivity and specificity, different thresholds were tested on

their classification performance is shown in Figure 6.4.

It is found that when the threshold is too high (predicting almost every case as normal), even though the overall proportion of agreement is high, the sensitivity can become very low. This is expected because the prevalence of acidotic (adverse) cases is very low (3.37%). Different thresholds can be used for different requirements of sensitivity and specificity. PPV (positive predictive value) is also reported to calculate the proportion of actual adverse cases in the predicted-to-be-adverse cases, and the PPV is always found to be less than 0.3 since the dataset is heavily unbalanced. Figure 6.5 also shows the ROC curve for maximum consecutive alarms, the AUC for which is 0.71. This indicates that maximum consecutive alarms can be a good indicator of labour condition.

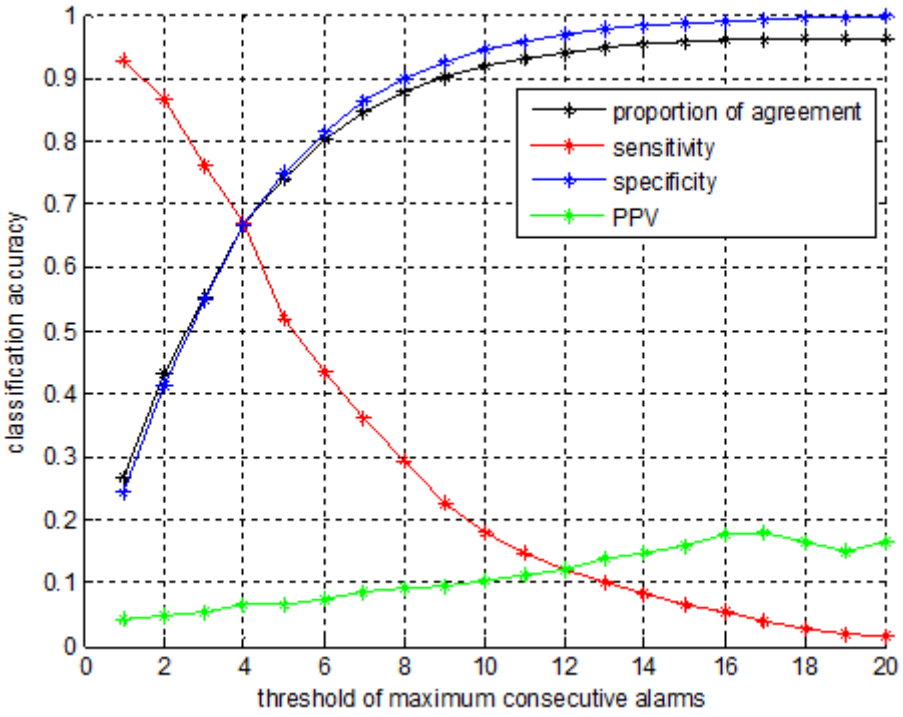


Figure 6.4 The relationship between kappa and the threshold of maximum consecutive alarms.

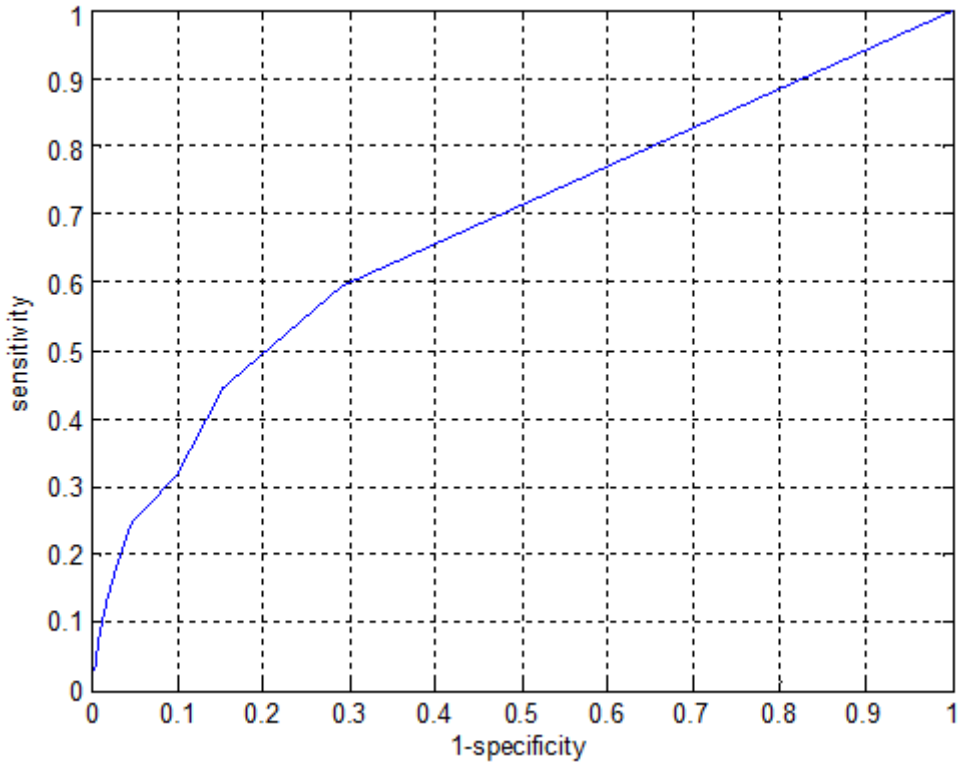


Figure 6.5 The ROC curve of maximum consecutive alarms.

6.2.5 Selection of thresholds

To give decision support effectively, two thresholds have to be decided: alarm threshold, to decide when to give an alarm based on predictor value; and maximum consecutive alarms threshold, to decide when to intervene based on the number of maximum consecutive alarms. In Section 6.2.4, the alarm threshold was temporarily set to 0.5 to balance sensitivity and specificity. However, in clinical situations, as fetal acidemia is very rare (about 3%) and intervention can cause complications, specificity is usually higher than sensitivity. Therefore, different thresholds have to be selected to give reliable decision support during labour.

As reported in a previous study (Georgieva et al., 2012), the intervention rate due to fetal distress was about 20% in the OIFD3 dataset. Therefore the should recommend intervention

in under 20% of cases. As both the alarm threshold and the maximum consecutive alarms threshold increase, the intervention rate will decrease, the sensitivity will decrease, and the positive predictive value (PPV) will increase. These are the most important measures of classification performance in this case, since the task is to identify more adverse cases with fewer interventions.

Table 6.3 shows the values of these measures under different thresholds. The favourable values are shown in bold (intervention rate $\leq$10%, sensitivity $\geq$30%, PPV$\geq$10% and specificity$\geq$95%). It is found that when the alarm threshold is 0.8 and the maximum consecutive alarm threshold is 4, the lowest intervention rate of 5.4% and best PPV of 15.6% can be achieved, with a reasonable sensitivity of 25.1%. This means the system will be able to predict 25.1% of fetal acidemia cases while causing only a 5.4% intervention rate. From the cases with intervention, 15.6% would be correctly identified cases with the potential to develop acidemia (true positive).

Table 6.3 Value of various classification measures at different thresholds. The favourable values are shown in bold, and the best value of each column is marked with red.

Alarm threshold	***Maximum consecutive alarm threshold***	***Intervention rate***	***Sensitivity***	***PPV***	***Specificity***
0.5	6	19.4%	**43.5%**	7.6%	81.4%
	7	14.3%	**36.1%**	8.5%	86.4%
	8	10.7%	29.4%	9.3%	90.0%
	9	**8.0%**	22.7%	9.6%	92.5%

0.6	5	15.0%	**39.2%**	8.8%	85.8%
	6	10.6%	**34.5%**	**11.0%**	90.2%
	7	**7.3%**	25.5%	**11.8%**	93.3%
	8	**5.4%**	20.0%	**12.5%**	95.1%
0.7	3	20.0%	**47.8%**	8.1%	81.0%
	4	12.0%	**37.3%**	**10.5%**	88.9%
	5	**8.0%**	**31.0%**	**13.0%**	92.8%
	6	**5.6%**	23.1%	**13.8%**	**95.0%**
0.8	2	16.4%	**44.7%**	9.2%	84.6%
	3	10.6%	**31.4%**	**10.0%**	90.2%
	4	**5.4%**	25.1%	**15.6%**	**95.3%**
0.9	1	15.5%	**41.6%**	9.0%	85.4%
	2	**7.2%**	26.7%	**12.4%**	93.5%

It should be noted that after the change of alarm threshold to 0.8, the classification performance of maximum consecutive alarm has changed (Figure 6.6). Compared to the classification performance when the alarm threshold was 0.5 (Figure 6.4), it is found that the AUC has decreased from 0.71 to 0.68, since sensitivity and specificity are no longer balanced. Therefore, the alarm threshold was set to 0.8 and the maximum consecutive alarm threshold was set to 4 – the equivalent of 30 minutes abnormal CTG trace. A computerised decision support system can be set up as follows:

(1) For each 15-minute window, calculate the value of the predictor, check whether signal quality $<$ 0.5, if yes, report "inadequate signal quality" (signal quality threshold selected from Chapter 5). When the signal quality is adequate ($\geq$ 0.5), if the predictor value $\geq$ 0.8, report as an alarm to indicate abnormal fetal condition, if the predictor value $<$ 0.8, report as

no alarm.

(2) As shown in Table 6.3, the risk of adverse outcome (fetal acidemia) can be estimated. When ≥ 4 consecutive alarms (30 minutes) appeared in Stage 2, report "fetal acidemia risk > 15.6% and intervention recommended".

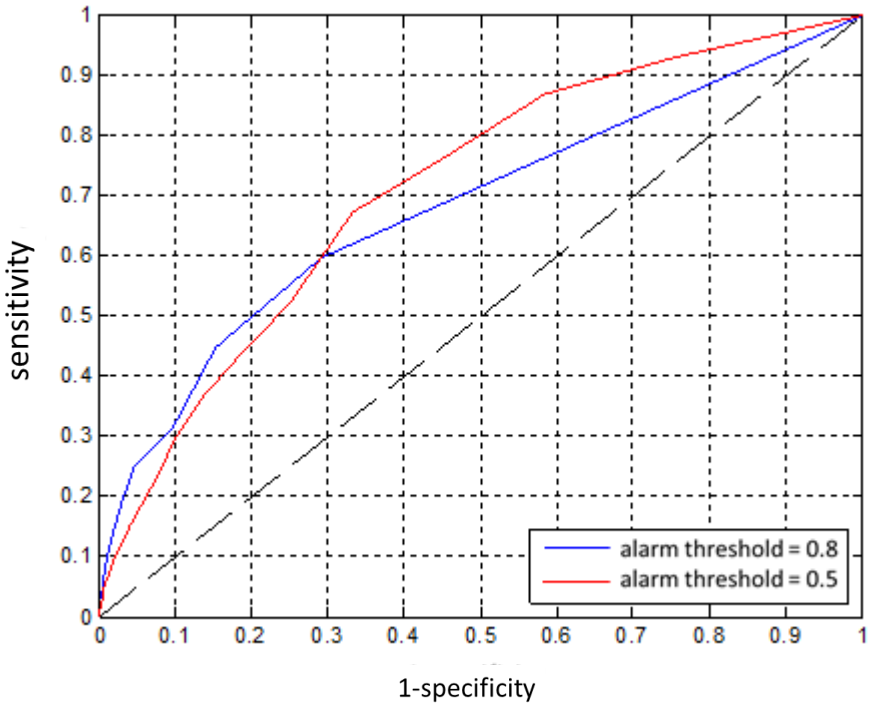


Figure 6.6 ROC curve of maximum consecutive alarms after the alarm threshold was readjusted to 0.8.

6.2.6 Discussion

The dataset used to set up the system is heavily unbalanced with only 255 adverse cases (pH < 7.05) out of 7,568 cases (3.37%). This is because fetal acidemia (low arterial pH at birth) is a rare clinical event. Therefore, the prediction accuracy should be reported very carefully. For example, if a classifier predicts all cases as normal, the proportion of agreement will still be over 95%. Sensitivity, specificity, PPV and AUC should also be reported to illustrate the

various aspects of classification performance. In particular, as the decision to intervene is at the cost of clinical complications, more true positives with fewer negatives would be favourable. Therefore, the PPV is especially important for decision support.

A window with predictor value higher or equal to the alarm threshold will be reported as an alarm. The analyses of a few samples in Figure 6.1 has illustrated that the change of the prediction value can be a straight and instinctive way of showing the change of labour condition in different time periods of labour. Generally, an adverse case will have more alarms than a normal case. This can be helpful for clinicians to monitor the condition of fetal health. The intermediate zone between alarm and no alarm is worth studying in the future. Showing a predictor value provides a qualitative way of monitoring labour condition. To provide a quantitative prediction of risk, several variables can be used: number of alarms, fraction of alarms and number of maximum consecutive alarms. Maximum consecutive alarms was selected as the indicator to predict labour outcome.

However, it was shown in Figure 6.2 that the maximum consecutive alarms in normal cases and adverse cases were distributed differently only in Stage 2; in Stage 1 the distribution of the two groups are much more similar. The possible reasons for this are: (1) the predictive power of the predictor is weak in Stage 1 since the classifier is trained on Stage 2 traces; (2) in Stage 1 clinicians have more time to intervene at the end of labour (Caesarean, forceps or ventouse delivery) or to take clinical measures during labour (maternal hydration and/or oxygenation; reduction or stopping of oxytocin drip, etc.), thus some cases with an adverse labour condition can sometimes recover after measurements, resulting in the reporting of

many alarms in Stage 1 but healthy labour outcome; (3) Stage 2 is increasingly stressful for the fetus and it is more likely that it becomes asphyxiated.

For reason (1), it is known that different criteria of the FHR should be used in different stages of labour (Clark et al., 2013), thus there the predictive power of the predictor trained using Stage 2 will decrease in Stage 1. Also, reasons (2) and (3) are likely to contribute to the situation, as illustrated in Figure 6.3. However, since there is no digital information of clinical measures in the database, there is no way to evaluate the influence of such clinical measures during labour. This information should be noted in future studies. In this study only maximum consecutive alarms in Stage 2 were investigated.

Two major parameters (alarm threshold and maximum consecutive alarms threshold) of the system were optimised here for the best performance of maximum consecutive alarms. The constraint of optimisation was to have an intervention rate of no more than 20%. The criteria for optimisation were to use fewer interventions (lower intervention rate) and to recognise more adverse cases (higher sensitivity, higher PPV). The alarm threshold was set to 0.8 and the maximum consecutive alarms threshold was set to 4 to achieve a balanced combination of sensitivity, intervention rate and PPV (Table 6.3).

The computerised system provides two methods of decision support: to report alarm and to quantify the risk of adverse outcome (fetal acidemia) based on maximum consecutive alarms. The GA + RBF SVM predictor trained in Chapter 4 is used, and a high value of the predictor is reported as an alarm. Although it should be noted that an alarm only indicates the increasing risk of abnormal labour outcome, the alarm itself can be a useful complement to

the current systems that can only report abnormal features (see Section 2.2). More importantly, estimating the risk of fetal acidemia can help clinicians to obtain a quantitative prediction for decision making. No previous study has had a database large enough to estimate the actual risk of fetal acidemia.

The major issue of this system is that it uses retrospective data to set up. The information of clinical measures was never entered electronically and is lost for the entire database. Therefore, it is highly possible that in some cases clinical measures have affected the labour outcome. In that case, the risk of fetal acidemia reported by the system would be underestimated. Additionally, the classifier trained using Stage 2 was found to be not predictive in Stage 1. Further studies should focus on applying the predictor to data that have more information on clinical measures, and using Stage 1 windows to train another predictor for Stage 1.

6.3 Validation of the System: Internal Database

After setting up the system using the 7,568 cases (OIFD3 database, see Figure 4.1), the system was validated using a new retrospective dataset of 4,385 cases in Oxford. The system was validated in two aspects: (1) the predictor power of maximum consecutive alarms on the dataset (Section 6.3.2); (2) the risk estimation of fetal acidemia and the decision support it provides (Section 6.3.3).

6.3.1 Data

A new database (the Centrale database) was used to validate the system. The data were collected from Aug 2008 to Mar 2011, with selection criteria shown in Figure 6.7. 4,385 cases

were selected as the validation dataset. The selection criteria were similar to those of the OIFD3 database (7,568 cases, see Figure 4.1), based on which the system was set up. The major differences were that in OIFD3 cases with congenital problem and birth trauma are excluded; CTG records of insufficient signal quality (less than 3 windows in last 30 minutes have over 50% signal quality) are excluded; and CTG recordings ending more than 1 minute before the time of birth are excluded.

An adverse case was defined as fetal acidemia (umbilical arterial pH $<$ 7.05) (Georgieva et al., 2013a), and a normal case defined as pH $\geq$ 7.05. The dataset was also unbalanced, with only 111 (2.53%) out of 4,385 cases having fetal acidemia. Note that this proportion is less than 3.37% in the OIFD3 database (7,568 cases collected from 1993 to 2008, see Figure 4.1), which could result either from progress in timely interventions over the years or the different selection criteria for the databases.

One major reason could be that Centrale includes cases where the recording ended more than 1 minute before the time of birth, which could result in some cases that could have been picked up by the system not being picked up since a part of the CTG is missing. Additionally, in the Centrale database metabolic disorders, congenital problems and birth traumas have not been excluded. These are more risky groups but with CTGs that would not look like standard developing acidemia. Birth trauma and congenital malformations and metabolic diseases could not be excluded from Centrale at this stage, as further information for these needs to be gathered from a separate electronic archive.

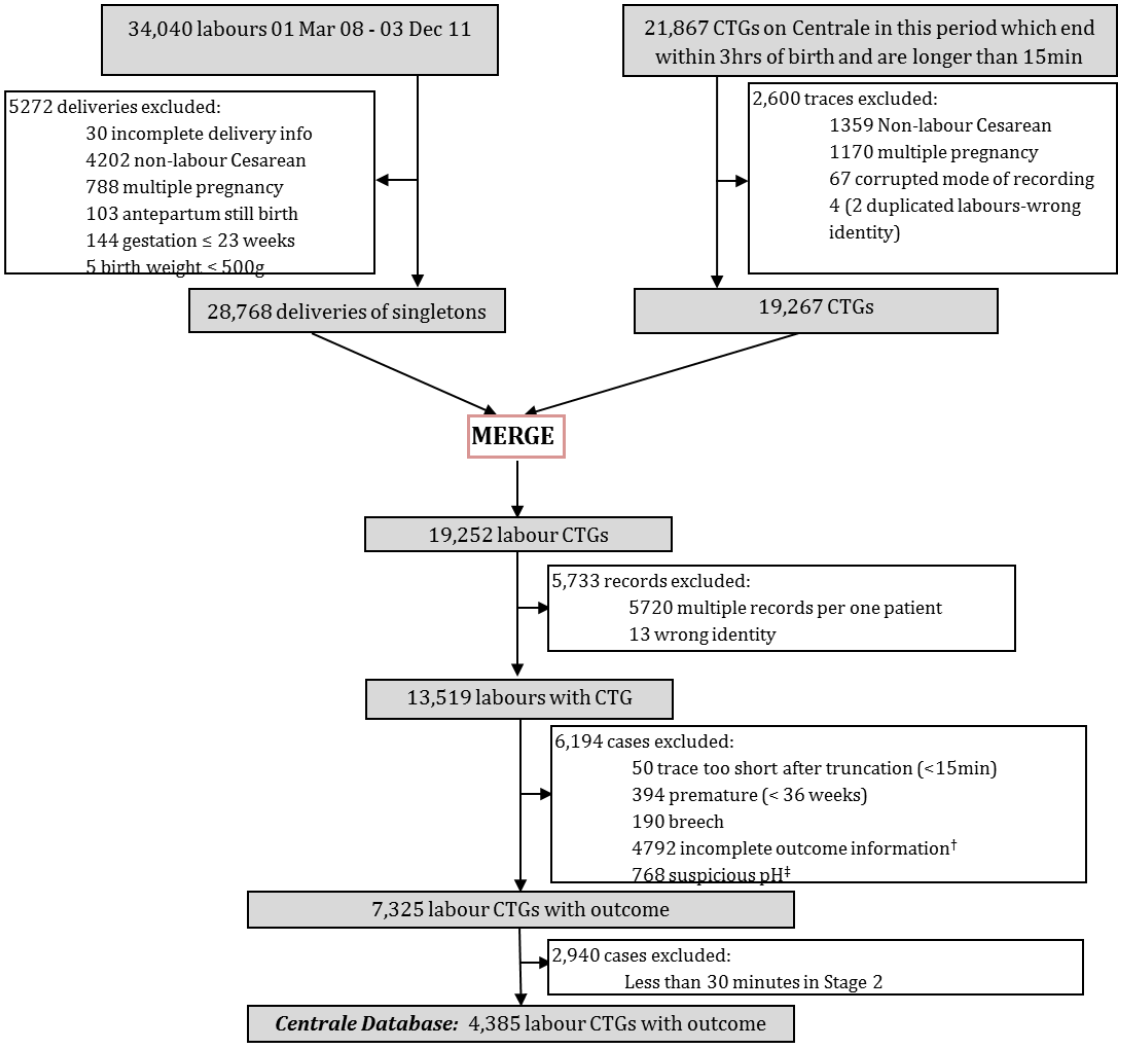


Figure 6.7 Selection criteria of the Centrale database from Aug 2008 to Mar 2011, figure courtesy of Antoniya Georgieva. †Missing arterial pH or venous pH ‡Venous pH – Arterial pH < 0.01

The dataset has digital information for interventions (Caesarean birth, forceps or ventouse delivery) but no digital information for clinical measures during labour (maternal hydration and/or oxygenation; reduction or stopping of oxytocin drip, etc.). There are 16% cases with interventions due to fetal distress (observed from CTG), and 10% cases are recorded as interventions due to other reasons (prolonged second stage, obstructed labour, multiple birth, etc.). This study will focus on the interventions due to fetal distress because these are the

interventions that happened due to an abnormal CTG.

6.3.2 Validation of predictive power

The predictive power of the maximum consecutive alarms was validated using the ROC curve (Figure 6.8). The AUC dropped from 0.68 in the OIFD3 (7,568 cases) to 0.63. Despite the reduction, this validates that maximum consecutive alarms (in Stage 2) is still a good indicator of labour condition.

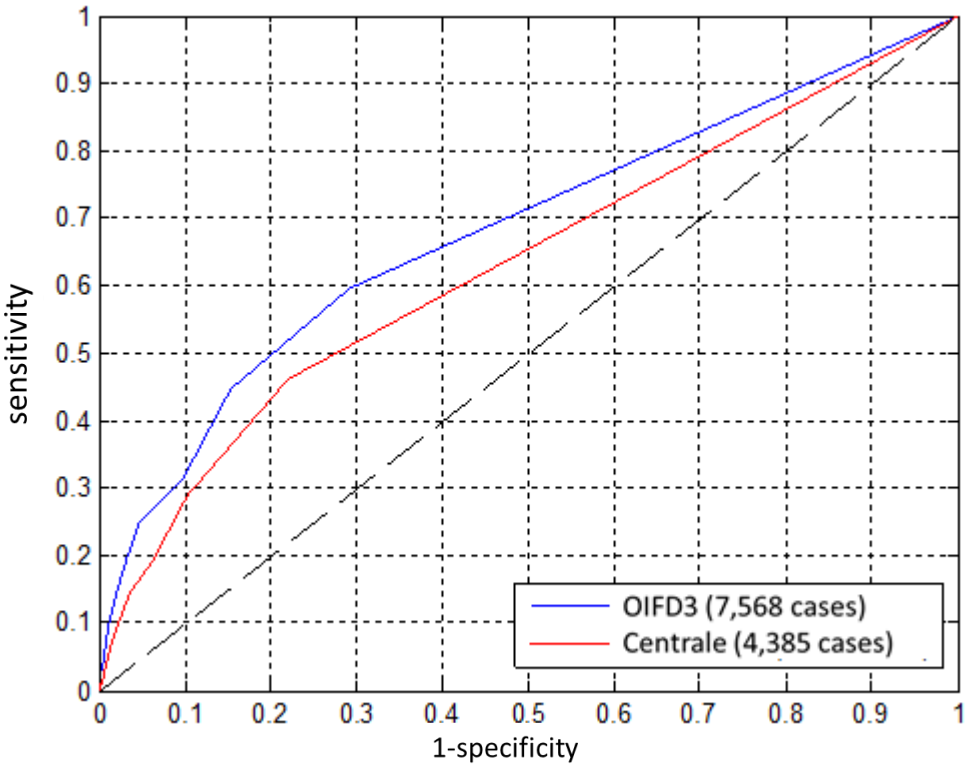


Figure 6.8 ROC curve of maximum consecutive alarms on the 4,385 cases.

The alarm threshold and maximum consecutive alarm threshold have also been validated (Table 6.4). It is found that an alarm threshold of 0.8 with a maximum consecutive alarm threshold of 4 still has the lowest intervention rate, the highest PPV and specificity. This indicates that the selection of both the alarm threshold and the maximum consecutive

threshold is reasonable.

Table 6.4 Value of classification measures at different thresholds on the validation set (4,385 cases). The favourable values (intervention rate ≤ 10%, sensitivity ≥ 30%, PPV ≥ 8% and specificity≥95%) are shown in bold, the best value of each column being marked with red.

Alarm threshold	*Maximum consecutive alarm threshold*	*Intervention rate*	*Sensitivity*	*PPV*	*Specificity*
0.5	6	15.7%	**35.1%**	5.7%	84.8%
	7	11.6%	27.9%	6.1%	88.8%
	8	8.2%	18.9%	5.8%	92.1%
	9	5.9%	14.4%	6.2%	94.4%
0.6	5	12.5%	**33.3%**	6.8%	88.1%
	6	8.0%	24.3%	7.7%	92.5%
	7	5.5%	18.9%	**8.7%**	94.8%
	8	**4.0%**	13.5%	8.6%	**96.3%**
0.7	3	14.1%	**30.6%**	5.5%	86.3%
	4	9.1%	24.3%	6.8%	91.3%
	5	5.9%	18.9%	**8.1%**	94.4%
	6	**3.9%**	14.4%	**9.3%**	**96.4%**
0.8	2	10.8%	28.8%	6.7%	89.6%
	3	6.5%	18.9%	7.4%	93.8%

	4	**3.9%**	14.4%	**9.5%**	**96.4%**
0.9	1	8.6%	23.4%	6.9%	91.8%
	2	**4.0%**	13.5%	8.6%	**96.3%**

6.3.3 Validation of risk estimation

In Section 6.2.4, the risk of adverse outcome (fetal acidemia) was estimated based on the number of maximum consecutive alarms: when more than four consecutive alarms (alarms lasting 30 minutes or more) appear in Stage 2, it is reported that "fetal acidemia risk > 15.6% and intervention recommended". To validate the risk estimation, the estimation criteria were then applied to the new dataset.

In the 4,385 cases, 169 cases have $\geq$ 4 maximum consecutive alarms (3.9% intervention rate), of which 16 cases (9.5% PPV) have fetal acidemia, which covers 14.4% of all the 111 cases with fetal acidemia (14.4% sensitivity). The actual risk of fetal acidemia (9.5%) in the recommend-intervention cases is less than the predicted 15.6% and the sensitivity decreases from 25.1% to 14.4%. However, considering that the proportion of fetal acidemia in the Centrale database (2.37%) is less than the OIFD3 (3.37%) and that the intervention rate in the Centrale database (3.9%) is lower than the OIFD3 (5.4%), the risk estimation is reasonable.

Clinically, the time point of abnormality (fetal acidemia) identification is critical. As abnormality was identified using four consecutive alarms, the time needed for identification is 30 minutes. The time point of identification among the 16 cases ranges from 5 minutes before delivery to 75 minutes before delivery. As the number of identified cases is small, not much

can be learned from the distribution of the time points. The average identification time point (26.6 minutes), on the other hand, can be useful for future studies. This means that intervention can be done nearly half an hour before actual delivery time. Timely intervention can be carried out for these cases nearly half an hour before delivery to prevent fetal acidemia. This illustrates the ability of the system to provide timely decision support.

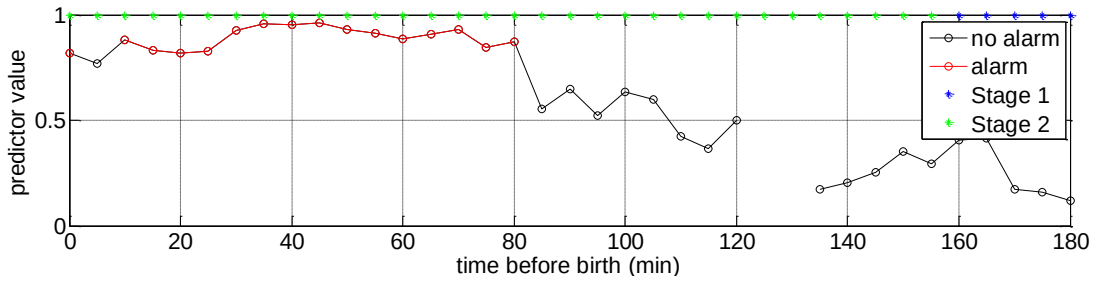
Details of delivery information can be found in Table 6.5. In these 16 cases ("true positives"), 10 cases were spontaneously delivered, two cases had Caesarean birth, and four cases had operative vaginal deliveries (forceps or ventouse delivery). If timely intervention had been carried out in the 10 spontaneous births, their condition would be potentially better. In the 153 cases with a recommendation to intervene but with no fetal acidemia ("false positives"), only 48 cases (31.4%) were spontaneously delivered, and 61 cases (39.9%) had an intervention due to fetal distress. This indicates that about 40% of these "false positives" may actually be "true positives", since intervention might have relieved the asphyxia condition. For the 48 spontaneously delivered cases, it is also possible that clinical measures had been taken to relieve the issue, but since there is no such digital data, it is difficult to find out.

Table 6.5 Delivery information of the Centrale dataset.

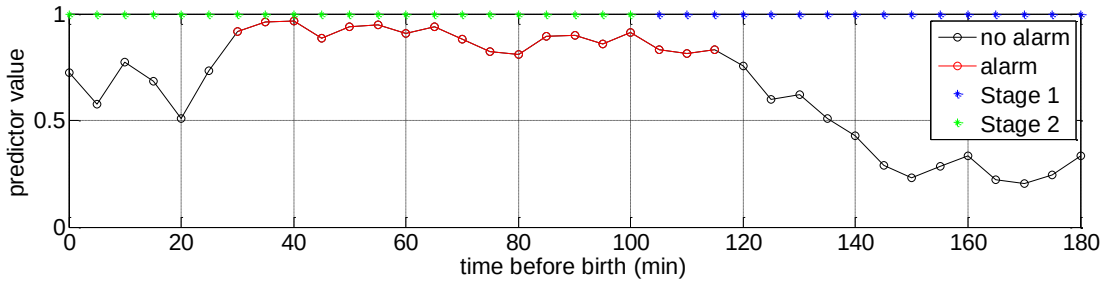
	Spontaneous delivery	***Delivery due to feta distress***	***Delivery due to other reasons***	***Total***
True positive	10 (62.5%)	3 (18.8%)	3 (18.8%)	16 (100%)
False positive	48 (31.4%)	61 (39.9%)	44 (28.8%)	153 (100%)
True negative	1643 (39.9%)	916 (22.2%)	1562 (37.9%)	4121 (100%)

False Negative	30 (31.6%)	23 (24.2%)	42 (44.2%)	95 (100%)

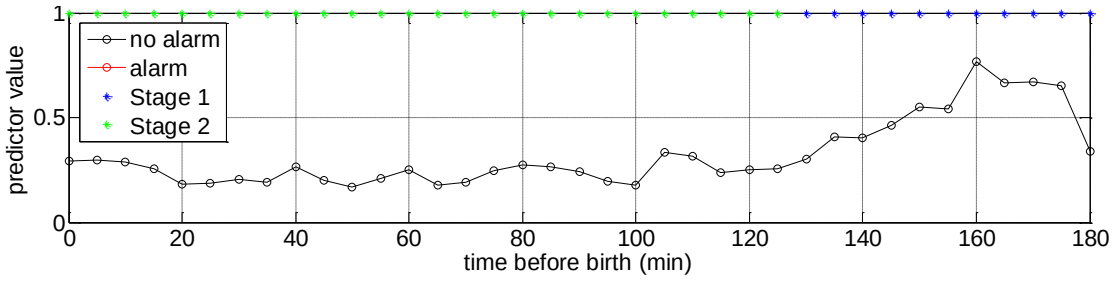
Figure 6.9 shows a few representative sample traces. Figure 6.9 (A) shows a true positive: 15 maximum consecutive alarms were found in Stage 2, four consecutive alarms were found over 60 minutes before delivery, and the fetus suffered from fetal acidemia and was delivered spontaneously. If interventions had been carried out as recommended (an hour before delivery), the situation could have been improved. Figure 6.9 (B) shows a false positive: 15 consecutive alarms were found in Stage 2 before delivery, and the fetus was delivered with forceps and has no fetal acidemia. As suggested in Section 6.2.6, there is no digital information of clinical measures in the database, thus it is highly possible that in some cases clinical measures or interventions have been carried out to prevent birth asphyxia. For this fetus, there is a decrease of predictor values in the last 30 minutes of labour, which is likely to result from clinical measures. In such cases, many "false positives" might actually be true positives. Figure 6.9 (C) shows a spontaneously delivered true negative: no maximum consecutive alarms and no fetal acidemia. This case is a typical illustration of a normal FHR trace. Figure 6.9 (D) shows a ventouse delivered false negative: no alarm was shown but fetal acidemia was reported. This case is an illustration of the abnormality that cannot be identified yet by the classifier. This shows a limitation of the system, which will need to be tackled in future studies.



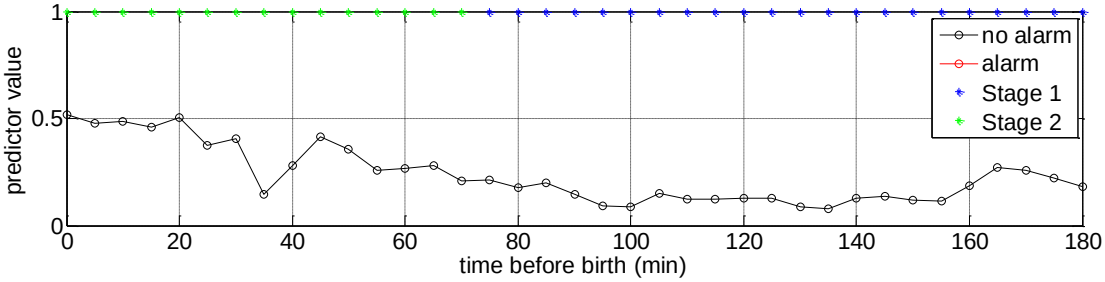
(A) True positive: 15 maximum consecutive alarms, pH = 6.96, spontaneous delivery



(B) False positive: 15 maximum consecutive alarms, pH = 7.3, forceps for other reasons



(C) True negative: no maximum consecutive alarms, pH = 7.22, spontaneous delivery



(D) False negative: no maximum consecutive alarms, pH = 7.04, ventouse for fetal distress

Figure 6.9 Case studies of the decision support system (Centrale dataset).

6.3.4 Discussion

As discussed in Section 2.2, other existing computerised FHR systems have quite different methods and databases from this system. Details of validation results have not been shown in previous publications. Therefore, it is difficult to compare this system with other systems. However, the largest dataset used for system validation in previous publications contains only 769 cases (Parer and Hamilton, 2010). Therefore, this is the first time that a system has been validated on a database of this size. The system proposed here can give alarms when abnormal FHR conditions are found, while estimating the risk of fetal acidemia and giving timely decision support. It is also the first time that the recommended intervention time point has been reported. More importantly, most existing systems are only validated comparing the classification to the opinions of clinical experts (Keith et al., 1995, Ayres-de-Campos et al., 2008, Parer and Hamilton, 2010, Czabanski et al., 2012).

The prediction power of maximum consecutive alarms was validated on a new retrospective dataset of 4,385 cases. The AUC dropped to 0.63 compared with 0.68 of the OIFD3 database on which the system was set up. The risk estimation of fetal acidemia was also validated. Of the 4,385 cases, 3.9% (169 cases) had $\geq$ 4 maximum consecutive alarms, of which 16 (9.5%) suffered from fetal acidemia. This is less than the predicted risk of 15.6%. Despite the decrease, this validates prediction power and the risk estimation ability of the system.

If interventions had been carried out on all cases that have more than four consecutive alarms, with only 3.9% cases recommended for intervention, 14.4% (16 out of 111) fetal acidemia cases can be covered. This validates the effectiveness of the system. A few case studies shown

in Figure 6.9 have illustrated that in the cases that were reported as no fetal acidemia, clinical measures might have been carried out during labour (but this is not known since there are no digital data on this).

The major benefit of the system is that it could identify adverse situations in a timely manner long before delivery. In order to identify a fetus as at risk (identification time), the system requires only 30 minutes. In the 16 recognised fetal acidemia cases, the time recommend to intervene is nearly half an hour before delivery on average, and for some cases the intervention time can be as long as over an hour before delivery. These timely interventions could bring about a reasonable reduction of birth asphyxia.

The major limitation of the system shown here in validation is its sensitivity. The identified 16 adverse cases cover 14.4% of all 111 cases that suffered from fetal acidemia. Although this is good considering that the intervention rate is only 3.9% and that about 30% “false positives” can actually be “true positives”, it still needs to be improved. As shown in the case studies in Figure 6.9, some of the adverse cases cannot be recognised by the system. Since fetal acidemia is rare in the population and interventions have a very high cost and risk of complications, the system only focuses on the prediction of highest (maximum consecutive alarms) cases. This will inevitably lead to the missing of many other adverse cases (decrease of sensitivity). Further efforts should be made to improve the sensitivity of the predictor whilst keeping the intervention rate low.

Another major limitation of the validation is the size of the dataset: although it is currently the largest dataset worldwide, only 169 cases in it had fetal acidemia. Also, the system as yet does

not utilise the information in Stage 1, which omits about 40% of the records that have inadequate or zero information in Stage 2. As already discussed in Section 6.2.6, digital information of clinical measures should be collected to enable prediction in Stage 1, and another predictor for Stage 1 should be trained. This could not only improve the sensitivity of the system, but also advance the time point of prediction

Further studies should focus on improving the prediction accuracy by extracting more predictive features and optimising the feature selection methods. Efforts should also be made to use a larger dataset with information about clinical measures, to train and validate the system, and to utilise information from Stage 1.

6.4 Validation of the System: External Database

The system was also validated using an external dataset of 552 cases. This dataset was made available as an open database by *Chudáček et al.* (Chudáček et al., 2014). It is the first open database for CTG. This allows our results presented here to be referenced in future studies by other research groups. The system was validated in two aspects: (1) the predictor power of maximum consecutive alarms on the dataset; (2) the risk estimation of fetal acidemia and the decision support it provides.

6.4.1 Data

The CTU-UHB Intrapartum Cardiotocography Database is one of the open databases contained in PhysioNet (Goldberger et al., 2000). It comes from the Czech Technical University (CTU) in Prague and the University Hospital in Brno (UHB), and contains 552 CTG recordings, selected from 9164 recordings collected between 2010 and 2012 at UHB.

The CTU-UHB database is the first open-access database for research on intrapartum CTG signal processing and analysis. This is the first study using it for validation of a computerised decision support system.

Details of the data selection criteria can be found in the literature (Chudáček et al., 2014). Similar to the OxSys databases (OIFD3, Centrale, etc.), the CTG data is sampled at 4Hz. The major differences from the OxSys databases that might affect this study are: (1) the data were collected using the STAN S21 and S31(Neoventa Medical, Mölndal, Sweden) and Avalon FM40 and FM50 (Philips Healthcare, Andover, MA) fetal monitors, whereas in the OxSys databases the data were collected using other systems (Hewlett-Packard 8040A, Phillips Series 50 XM, Sonicaid FM7, etc.); (2) the CTG recordings in the CTU-UTB database start no more than 90 minutes before actual delivery, and each is at most 90 minutes long to keep recordings easily comparable, while for the Oxford databases the entire labour period are recorded; (3) the information of delivery in CTU-UTB only consists of two categories – vaginal and Caesarean, without the reason for intervention.

In the 552 cases considered here, 40 cases (7.2%) have fetal acidemia. It should be noted that this is higher than for the OIFD3 and the Centrale databases. 46 cases (8.3%) have Caesarean delivery and other 506 cases have vaginal delivery (including spontaneous delivery, forceps delivery and ventouse delivery). As the cases in the database include only 90 minutes before delivery at most, the CTG traces are either from Stage 2 or very close to Stage 2. Therefore, the system using maximum consecutive alarms in Stage 2 can be applied.

The nine GA selected FHR features (Table 6.1) were extracted using the same method as for

the OxSys databases (Section 2.5). The RBF SVM classifier trained in Chapter 4 was used. The methods of reporting alarm and risk estimation were the same as in Section 6.2.

6.4.2 Validation results and discussion

The prediction power of maximum consecutive alarms was validated using the ROC curve (Figure 6.10). The AUC dropped from 0.68 in the OIFD3 dataset to 0.65 in the CTU-UHB dataset, which is comparable to 0.63 on the Centrale dataset. Again, this validates that maximum consecutive alarms is a good indicator of labour condition.

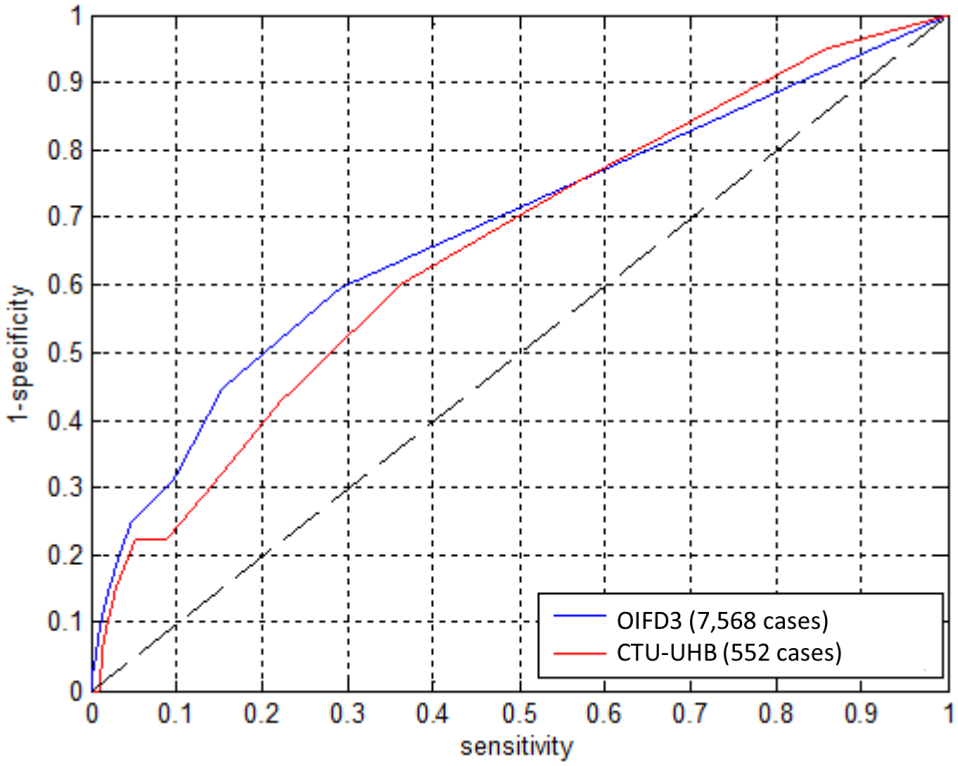


Figure 6.10 ROC curve of maximum consecutive alarms on the 552 cases.

The alarm threshold and maximum consecutive alarm threshold were also validated (Table 6.6). It was found that an alarm threshold of 0.8 with a maximum consecutive alarm threshold

of 4 still has a low intervention rate (23.6%), the highest PPV (13.1%), a high sensitivity (42.5%) and a high specificity (77.9%). This indicates that the selection of both the alarm threshold and the maximum consecutive threshold is also reasonable on the CTU-UHB dataset.

Table 6.6 Value of various classification measures at different thresholds on the CTU-UHB dataset (552 cases). The favourable values (intervention rate≤25%, sensitivity≥40%, PPV≥12% and specificity≥70%) are shown in bold, the best value of each column being marked with red.

Alarm threshold	*Maximum consecutive alarm threshold*	*Intervention rate*	*Sensitivity*	*PPV*	*Specificity*
0.5	6	41.1%	**47.5%**	8.4%	59.4%
	7	31.5%	**45.0%**	10.3%	69.5%
	8	**23.7%**	**40.0%**	**12.2%**	**77.5%**
	9	**18.5%**	27.5%	10.8%	**82.2%**
0.6	5	38.8%	**47.5%**	8.9%	61.9%
	6	27.5%	**40.0%**	10.5%	**73.4%**
	7	**21.7%**	37.5%	**12.5%**	**79.5%**
	8	**15.6%**	27.5%	**12.8%**	**85.4%**
0.7	3	53.6%	65.0%	8.8%	47.3%
	4	37.0%	**52.5%**	10.3%	64.3%
	5	29.2%	**45.0%**	11.2%	**72.1%**
	6	**17.9%**	30.0%	**12.1%**	**83.0%**

0.8	2	58.0%	**75.0%**	9.4%	43.4%
	3	38.0%	**60.0%**	11.4%	63.7%
	4	**23.6%**	**42.5%**	**13.1%**	**77.9%**
0.9	1	74.1%	**82.5%**	8.1%	26.6%
	2	37.9%	**55.0%**	10.5%	63.5%

In Section 6.2.4, the risk of adverse outcome (fetal acidemia) was estimated based on the number of maximum consecutive alarms: when more than four consecutive alarms (30 minutes) appeared in Stage 2, it was reported that "fetal acidemia risk > 15.6% and intervention recommended". To validate the risk estimation, the estimation criteria were then applied to the CTU-UHB dataset.

Of the 552 cases, 130 cases have ≥ 4 maximum consecutive alarms (23.6% intervention rate), of which 17 cases (13.1% PPV) have fetal acidemia, which covers 42.5% of all 40 cases with fetal acidemia (42.5% sensitivity). The actual risk of fetal acidemia (13.1%) is comparable to the estimated risk (15.6%).

Results from the OIFD3, the Centrale and the CTU-UHB datasets are compared in Table 6.7. The major difference between results on the CTU-UHB and Centrale dataset is the sensitivity and intervention rate. The sensitivity on the CTU-UHB dataset is 42.5%, while on the Centrale dataset is 14.4%; the intervention rate on the CTU-UHB dataset is 23.6%, while on the Centrale dataset it is only 3.9%. This is likely due to the different selection criteria between the two datasets and the different ratio between adverse and healthy cases (3.37% vs

7.25%). For the Centrale dataset, the average length of Stage 2 is 113.9 minutes; while for the 552 cases in the CTU-UHB dataset, the average length before delivery is 84.5 minutes, and the length of Stage 2 is $\leq$ 30 minutes. Therefore, cases in the CTU-UHB database have a very short Stage 2 before delivery, which can be related to more abnormal FHR patterns and thus bring about a higher sensitivity and intervention rate.

Table 6.7 Comparison between results of the OIFD3, the Centrale and the CTU-UHB datasets.

Database	***OIFD3***	***Centrale***	***CTU-UHB***
Number of cases	7,568	4,385	552
Number of fetal acidemia	255 (3.37%)	111 (2.53%)	40 (7.25%)
Intervention rate proposed by the system	5.4%	3.9%	23.6%
Clinical intervention rate due to fetal distress	20.0%	16.0%	N/A
Sensitivity	52.1%	14.4%	42.5%
PPV	15.6%	9.5%	13.1%
Specificity	95.3%	96.4%	77.9%
Intervention time before delivery (minute)	34.2	26.6	33.2

The time point of identification was also investigated. As abnormality was identified using four consecutive alarms, the length of time used for identification of abnormality is 30 minutes. The time point of identification among the 17 cases ranges from 15 minutes before

delivery to 55 minutes before delivery, the average being 33.2 minutes before delivery. Even with the traces limited to less than 90 minutes, the system can still identify abnormality over half an hour before birth. This again validates the ability of the system to provide timely decision support.

Although the lengths of CTG traces used in CTU-UHB (Stage 1 and Stage 2, 84.5 minutes on average) are shorter than those used in Centrale (Stage 2 only, 113.9 minutes on average), the average identification time point of CTU-UHB (33.2 minutes before delivery) is earlier than that of the Centrale dataset (26.6 minutes before delivery). This is likely to also result from the fact that CTU-UHB uses only information from the last 90 minutes before delivery and cases that have very short Stage 2 (no more than 30 minutes), which as discussed, can cause more abnormal FHR patterns and thus bring about more alarms and an earlier time for identification.

The major benefits and problems of the system have been discussed in Section 6.3.4. The validation results on this external dataset show that the system is consistently effective even when using datasets that have a different data collection system and data selection criteria. The CTU-UHB database is the first open-access intrapartum CTG database, and this is the first computerised decision support system validated using the database. This validation result is thus an important reference for future studies.

6.5 Conclusions

In this chapter, a computerised decision support system was set up based on the classifier of Chapter 4 and the analysis of Chapter 5. The system was set up and optimised using the

OIFD3 database (7,568 cases). The system can give an alarm when an abnormal predictor value is observed, and the risk of fetal acidemia can be estimated based on consecutive alarms during Stage 2.

The system was validated using a new retrospective internal dataset of 4,385 cases (part of the Centrale dataset). The validation results showed that the system is capable of estimating the risk of fetal acidemia. With only a 3.9% intervention rate, the system can identify 14.4% of fetal acidemia cases. The system can also provide timely decision support nearly half an hour before delivery on average.

The system was then validated using an external dataset of 552 cases (the CTU-UHB database). These validation results showed that the system is capable of estimating the risk of fetal acidemia on databases using different data collection systems and selection criteria. With only a 23.6% intervention rate, the system can identify 42.5% of fetal acidemia cases. This shows the robustness of the system when using different datasets. This is the first time that a computerised decision support system has been validated on an open-access intrapartum CTG dataset.

This is also the first time that an intrapartum computerised FHR decision support system has been built and validated on a database of this size. It is also the first system to report the time point of intervention. Further studies should focus on improving the prediction power of the system (especially in Stage 1), as well as using a larger dataset to train and validate the system.

Implementation of the system clinically is a realistic prospect in the longer term. After

optimisation of the system as mentioned above, further efforts should be made to integrate the system into clinical practice. After further validation of the system and clinical trials, the future application of the system in clinical situations can be expected.

7 Conclusions

7.1 Summary of Thesis

This thesis has presented work on computerised analysis of fetal heart rate features, including feature extraction, feature selection and building a computerised decision support system. A novel feature - pattern readjustment - was extracted in Chapter 3. This feature, along with other features (64 features in total), was used for feature selection to find a best feature subset in Chapter 4. An important factor – signal loss - was studied for its influence on FHR features in Chapter 5. All of these studies were used in setting up a computerised system in Chapter 6, which has also been validated using new retrospective data.

In Chapter 3 a novel feature – pattern readjustment was extracted and tested. Clinical data were used to train a SVM classifier to detect pattern readjustment, and the association between pattern readjustment and adverse labour outcome was investigated. The validation results, when compared to clinical experts, showed that pattern readjustment can be accurately detected, and the study on labour outcome showed that the feature is related to fetal acidemia. This is the first time that step detection methods have been applied on FHR analysis (Xu et al., 2012).

In Chapter 4 Genetic Algorithms were used as feature selection methods to select a best subset of FHR features. Three classifiers were used in the GA and backward elimination was used for optimisation. A feature subset of seven to nine features was selected for each classifier. Fair classification performance was shown on a balanced 510-case dataset, with

kappa values ranging from 0.45 to 0.49 when calculated using different classifiers. The diagnostic power of the classifier output using selected features was tested on the total set of 7,568 cases. As the classifier output increases, there was a consistent increase in the risk of fetal acidemia. Based on these results, it can be concluded that GA can be successfully applied to FHR features to integrate and optimise their predictive power. This is the first time that a feature selection method has been applied on a database of this size.

In Chapter 5 the influence of signal loss in FHR analysis was investigated. A bivariate model was built to quantify the relationship between signal loss and error for each feature. Validation results showed that the bivariate model can accurately predict the error generated by signal loss. The influence of signal loss on labour outcome classification was also investigated; the classifier trained in Chapter 4 was used. It is found that in general, when signal quality is over 50%, the FHR data are reasonably robust for labour outcome prediction. A shortened version of Chapter 5 is in preparation for submission as a journal manuscript.

In Chapter 6 a computerised decision support system was set up based on the classifier of Chapter 4 and the analysis of Chapter 5. The system can give an alarm when an abnormal predictor value is observed. The risk of fetal acidemia can be estimated based on consecutive alarms. The system was validated using new retrospective data: one internal dataset and one external dataset. Validation results showed that the system is capable of predicting adverse labour outcome and providing timely decision support (about half an hour before delivery) on different datasets. This is the first time that an intrapartum computerised FHR decision support system has been built and validated on this size of actual labour outcome data, and the

first time that the intervention time point has been reported for a system. It is also the first system that has been validated on an open-access intrapartum FHR dataset.

The conclusions of this thesis based upon the results of these studies are thus:

1. FHR features such as pattern readjustment are useful in analysis of the fetal condition and prediction of labour outcome.

2. Feature selection methods such as GA can be used to select a best subset of FHR features on a specific classifier, upon which the performance of labour outcome prediction can be better.

3. Signal loss has an important influence on the value of FHR features and labour outcome prediction, the error caused by which can be predicted using proper models.

4. A computerised FHR analysis system can offer decision support for clinicians by reporting abnormal labour condition and estimation of the risk of fetal acidemia.

7.2 Limitations and Future Work

In this thesis, adverse outcome was defined as fetal acidemia (low arterial pH), since it is one of the necessary conditions for diagnosing birth asphyxia. There are also clinical parameters that need to be studied, such as maternal infection and oxytocin augmentation. This could be integrated as another database of FHR features in further studies. The labour outcome classification in Chapter 4, the prediction of classification error based on signal loss in Chapter 5 and the decision support system in Chapter 6 can all be studied in terms of different forms of labour outcome and integrated together.

The labour outcome prediction in Chapter 4 only uses the last 4 15-minute long windows from the last 30 minutes of Stage 2. To provide a better performance, more information during the process of labour, especially time-series information, should be integrated into the classifier. Clinical interpretation of the classifiers should also be considered, including the analysis of the features selected, e.g. the clinical interpretation of the best features and the correlation between the clinical features and the statistical features. Further studies will need to be carried out to estimate the risk of compromise based on the classifier prediction and its patient specific time-series trend.

The major limitation of the signal loss and error model in Chapter 5 is the simulation of clinical signal loss using clinical templates, which ignores the relationship between signal loss and FHR. Further studies should focus on investigating the relationship between FHR and signal loss. Other methods, such as fetal ECG, could be used as a gold standard to find out the actual signal under the FHR signal loss, as was used in (Spencer et al. (1987)).

The major limitation of the signal loss and labour outcome classification analysis in Chapter 5 is that the data size of each signal quality group is small (about 100 cases each). This can lead to variance in feature selection and classification - features can have been selected by chance in smaller datasets. Further studies should focus on using bigger datasets with more cases to optimise both the feature selection process and the classifier.

The major limitation of the labour outcome prediction system in Chapter 6 is that the database lacks information on clinical measures. Further studies should focus on improvements on all aspects of the system, including obtaining digital information about clinical measures during

labour (especially in Stage 1), improvements for the labour outcome prediction and prediction of various different forms of abnormal labour condition. The study of time-series information could also be very important to help improve the prediction performance.

Implementation of the system clinically can be a realistic prospect in the longer term. After optimisation of the system as mentioned above, further efforts should be made to integrate the system into the current OxSys computerised FHR analysis systems. After further validation of the system and clinical trials, the future of the system applied to clinical situations can be expected.

Appendix

Table 1 AUC values for different signal quality groups and the p-values of the two sided Mann-Kendall Tau non-parametric trend test.

Feature Index	Signal Quality Group						P-value
	50%-60%	60%-70%	70%-80%	80%-90%	90%-100%	All cases	
1	0.501	0.696	0.638	0.582	0.548	0.602	0.806
2	0.568	0.606	0.504	0.528	0.635	0.501	0.806
3	0.538	0.520	0.569	0.624	0.623	0.510	0.779
4	0.623	0.762	0.680	0.697	0.749	0.703	0.538
5	0.681	0.593	0.695	0.560	0.614	0.634	0.806
6	0.606	0.741	0.577	0.648	0.635	0.639	0.806
7	0.612	0.749	0.672	0.683	0.698	0.677	0.538
8	0.605	0.615	0.606	0.602	0.736	0.619	0.806
9	0.562	0.546	0.514	0.539	0.655	0.560	0.806
10	0.629	0.543	0.580	0.533	0.784	0.597	0.806
11	0.709	0.631	0.509	0.552	0.604	0.566	0.538
12	0.629	0.603	0.560	0.566	0.511	0.501	0.914
13	0.569	0.553	0.565	0.680	0.665	0.590	0.538
14	0.663	0.571	0.510	0.517	0.722	0.582	0.806
15	0.531	0.524	0.508	0.538	0.611	0.521	0.538
16	0.664	0.501	0.602	0.648	0.566	0.593	0.806
17	0.662	0.539	0.601	0.537	0.725	0.594	0.806
18	0.578	0.514	0.550	0.519	0.616	0.509	0.806
19	0.500	0.562	0.514	0.606	0.675	0.563	0.914
20	0.556	0.537	0.511	0.513	0.557	0.516	0.806
21	0.611	0.517	0.533	0.552	0.584	0.535	0.806
22	0.649	0.546	0.607	0.510	0.547	0.575	0.538
23	0.646	0.525	0.535	0.503	0.614	0.552	0.806
24	0.530	0.518	0.649	0.546	0.535	0.547	0.806
25	0.606	0.534	0.530	0.553	0.514	0.511	0.779
26	0.548	0.585	0.515	0.547	0.527	0.506	0.538
27	0.562	0.586	0.514	0.575	0.518	0.504	0.806
28	0.582	0.571	0.529	0.601	0.560	0.530	0.806
29	0.598	0.524	0.533	0.529	0.550	0.510	0.806
30	0.570	0.516	0.562	0.512	0.566	0.533	0.806
31	0.609	0.504	0.500	0.513	0.595	0.542	0.806
32	0.594	0.604	0.558	0.544	0.515	0.508	0.914

33	0.548	0.575	0.513	0.505	0.627	0.515	0.806
34	0.575	0.558	0.515	0.550	0.528	0.501	0.779
35	0.600	0.505	0.511	0.565	0.507	0.500	0.806
36	0.522	0.549	0.556	0.519	0.556	0.520	0.806
37	0.619	0.604	0.528	0.592	0.535	0.578	0.779
38	0.521	0.676	0.565	0.529	0.666	0.588	0.806
39	0.515	0.510	0.527	0.533	0.520	0.511	0.538
40	0.555	0.645	0.563	0.618	0.627	0.566	0.538
41	0.537	0.678	0.508	0.648	0.609	0.596	0.806
42	0.549	0.514	0.515	0.537	0.573	0.511	0.538
43	0.507	0.534	0.531	0.501	0.546	0.501	0.806
44	0.579	0.554	0.544	0.563	0.588	0.540	0.806
45	0.503	0.507	0.500	0.552	0.521	0.504	0.538
46	0.503	0.503	0.502	0.550	0.516	0.505	0.538
47	0.530	0.503	0.549	0.551	0.513	0.515	0.806
48	0.546	0.509	0.519	0.547	0.510	0.500	0.806
49	0.509	0.546	0.520	0.574	0.528	0.538	0.538
50	0.523	0.511	0.509	0.533	0.653	0.539	0.538
51	0.611	0.520	0.539	0.579	0.586	0.503	0.806
52	0.522	0.628	0.510	0.503	0.588	0.502	0.806
53	0.567	0.535	0.563	0.527	0.624	0.569	0.806
54	0.504	0.676	0.561	0.670	0.627	0.602	0.806
55	0.683	0.577	0.639	0.512	0.610	0.610	0.538
56	0.627	0.517	0.611	0.594	0.592	0.595	0.538
57	0.598	0.571	0.527	0.596	0.507	0.524	0.779
58	0.514	0.660	0.571	0.665	0.622	0.603	0.538
59	0.698	0.619	0.626	0.599	0.600	0.629	0.779
60	0.583	0.559	0.610	0.535	0.623	0.542	0.806
61	0.532	0.676	0.536	0.719	0.630	0.615	0.538
62	0.546	0.595	0.534	0.602	0.583	0.540	0.806
63	0.604	0.533	0.513	0.534	0.587	0.539	0.806
64	0.533	0.518	0.527	0.519	0.515	0.509	0.221

Table 2 Major Matlab functions coded by Liang Xu

Chapter	***Function Name***	***Function Usage***
3	Edge_detector.m	Use edge detectors to classify pattern readjustment

3	validation_samples110909.m multi_classifier11902.m	Validate the pattern readjustment classification results
3	phrelateddirect.m	Test the pattern readjustment on its relationship with pH
3	Isbs.m	Extract pattern readjustment features
4	Ga_main_v17.m Ga_general_v17.m	GA toolbox
4	Lr_bic_func.m Svmlib_bic_func.m Rbflib_bic_func.m	Apply BIC on the three classfiers: linear regression, linear SVM and RBF SVM (SVMs are from libSVM)
4	Sfs_v10.m Sbs_v10.m	Use SFS and SBS for feature selection
5	Degrade.m	Randomly generate signal loss
5	Degrade_clinical.m	Generate signal loss based on clinical templates
5	Extract_no_chunks.m	Calculate the number of signal segments
5	Clinical_degrade_test.m Sq_tesing.m Model_fitting.m	Test the performance of the bivariate model on clinically degraded signal
6	Max_consecutive.m	Calculate the number of max consecutive alarms

6	predictor_quantification_7568.m	Test the predictor on different datasets

Table 3 Mathematical definition/calculation method of each feature.

Index	***Feature Name***	***Mathematical Definition/Calculation Method***
1	Baseline mean	The mean of FHR excluding accelerations and decelerations. The calculation is identical to the Cazares baseline assignment methodology using morphological filters (Cazares, 2002).
2	Percentage of zero difference between neighbour points	The percentage of points in a window that have zero different with neighbouring points.
3	Approximate entropy	This feature is implemented with the same method and parameters as in Dawes et al. (1992)
4	Signal stability index (SSI)	The peak of the value of kernel density estimation of the FHR window, parameters set as in Georgieva et al. (2011a)
5	Minimal expected value	The 5th centile of the 4Hz FHR signal.
6	Skewness	The skewness of the FHR window.
7	Kurtosis	The kurtosis of the FHR window
8	Long-term variability	The mean of differences of the highest peak and the lowest trough within subsequent 1 minute windows where at least 30% of the values are valid.
9	SSI of the residual signal	The SSI of the residual signal. The residual signal is defined as the difference between the FHR and the smoothed FHR signal (averaging window of 30 values, working with the 4Hz signal).
10	Median of the short-term variability (STV) tracker	The STV tracker is defined as the mean absolute difference of all neighbouring FHR values (STV) within a moving 1 min window (sliding step is 0.5 min). In each 1min segment the FHR values that belong to decelerations or accelerations are excluded from the calculation. The STV is not calculated in a 1 min window if more than 70% of the FHR values in the window are absent due to decelerations, accelerations or signal loss.
11	SSI of the STV tracker	The SSI of the STV tracker.
12	Range of the STV tracker	The difference of the 95th and 5th centiles of the STV tracker values.

13	Median of the absolute derivative values of the STV tracker	The median of the absolute derivative value of the STV tracker.
14	STV tracker trend	The correlation coefficient of the baseline and time.
15	Number of contractions	This feature is implemented with the same method and parameters as in Georgieva et al. (2009).
16	Median of contraction duration	The median of the duration of contraction.
17	Median of resting time without contractions	The median of resting time (the part in FHR without contractions).
18	Resting/contraction time ratio	The ratio of resting time against contraction time.
19	Median STV tracker	The median of the STV tracker.
20	Number of accelerations	Accelerations: Transient increases in heart rate of 15 bpm or more and lasting 15 seconds or more. The calculation is identical to the Cazares acceleration detection methodology using morphological filters (Cazares, 2002).
21	Mean acceleration duration	The mean of the duration of acceleration.
22	Median of acceleration amplitude	The median of the amplitude (the difference between the peak of acceleration and baseline) of acceleration.
23	Number of decelerations	Deccelerations: Transient decreases in heart rate of 15 bpm or more and lasting 10 seconds or more. The calculation is identical to the Cazares deceleration detection methodology using morphological filters (Cazares, 2002).
24	Mean deceleration duration	The mean of the duration of decelerations.
25	Maximum deceleration duration	The maximum of the duration of decelerations.
26	Median of deceleration amplitude	The median of the amplitude of decelerations.
27	Maximum deceleration amplitude	The maximum of the amplitude of decelerations.
28	Mean time to decelerate	The mean of time to decelerate (time from the beginning of a deceleration to the minimal of the deceleration)
29	Mean time to recover	The mean of time to recover (time from the minimal of the deceleration to the end of a deceleration).
30	Maximum time to recover	The maximum of time to recover.
31	Number of quick recoveries	The number of quick recoveries (Number of decelerations that have Feature 28 $\geq$ Feature 29).

32	Number of slow recoveries	The number of slow recoveries (Number of decelerations that have Feature 28 < Feature 29).
33	Resting time/deceleration time ratio	The ratio of resting time against deceleration time.
34	Mean deceleration area	The area in the CTG covered by a deceleration (calculated by integration).
35	Total number of lost beats in the widow	The total number of lost beats (the deviations from baseline due to decelerations).
36	Median of the onset slope	The median of (amplitude of deceleration/time to decelerate).
37	Median of the recovery slope	The median of (amplitude of deceleration/time to recover).
38	Maximum baseline shift after a deceleration	The maximum value of baseline shift (The difference of mean baseline value in the last 2 minutes proceeding the deceleration and the 2 minutes after the deceleration are over 15 bpm).
39	Number of prolonged decelerations	The number of prolonged decelerations (length of deceleration over 3 minutes).
40	Number of decelerations with at least one shoulder	Number of decelerations with at least one shoulder .A left shoulder is defined as acceleration that ends less than 12sec before the onset of a deceleration. Similarly, a right shoulder (overshooting) is defined as acceleration that starts in less than 12sec after the end of a deceleration.
41	Number of decelerations with a right shoulder	Number of decelerations with at least one shoulder.
42	Number of early decelerations	Number of early decelerations. An early deceleration is defined as the start of acceleration $\in$ [-4,32] seconds later than the corresponding contraction and end of acceleration < 12 seconds.
43	Number of late decelerations	Number of late decelerations. An late deceleration is defined as the start of acceleration $\in$ [-8,60] seconds later than the corresponding contraction and end of acceleration < $\in$ [12,120] seconds.
44	Number of variable decelerations	The number of deceleration that are neither early decelerations nor late decelerations.
45	Median recovery time after contraction end	The median of the length of the overlapping part of decelerations and the end of corresponding contractions.
46	Mean lag time	The mean of lag time. Lag time is defined as the time between the start of a deceleration and the start of the corresponding contraction
47	Deceleration/contraction	The ratio of the total time of deceleration against the

	time ratio	total time of contraction in a FHR window
48	Mutual information feature	These features are time series features implemented with the same method and parameters as in (Fulcher, 2012)
49	Ratio of STD	
50	Mean of local ApEn	
51	Std of local ApEn	
52	Goodness of exponential fit	
53	2-dim time delay embedding space	
54	Max absolute deviation	
55	(STD/mean)^2	
56	Alphabet feature	
57	Pattern readjustment	Whether the FHR window is pattern readjustment (1) or is not pattern readjustment (0), as described in Chapter 3.
58	Interquartile of smoothed signal	Interquartile of the smoothed FHR window, as described in Chapter 3.
59	STD of Gaussian filter	Std of the first order Gaussian derivative filtered FHR window, as described in Chapter 3.
60	Sinusoidal pattern	This feature is implemented with the same method and parameters as in Pardey et al. (2002).
61	PRSA – DC	The deceleration component of the PRSA, calculated using the same method and parameter in Georgieva et al. (2013b).
62	BPRSA – std_AC	The standard deviation of acceleration component of the BPRSA, calculated using the same method and parameter in Williams (2012).
63	BPRSA – DC	The deceleration component of the BPRSA, calculated using the same method and parameter in Williams (2012).
64	ACOG	Simulation of ACOG guideline, 1- normal, 2 – suspicious, 3 – pathological, calculated using the same method and parameter in(Georgieva et al., 2011b) .

Bibliography

ACOG 2009. ACOG Practice Bulletin No. 106: Intrapartum Fetal Heart Rate Monitoring: Nomenclature, Interpretation, and General Management Principles. *Obstetrics & Gynecology,* 114**,** 192-202 10.1097/AOG.0b013e3181aef106.

AKAIKE, H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on,* 19**,** 716-723.

ALBERRY, M., FUENTE, S. & SOOTHILL, P. 2009. Prediction of asphyxia with fetal gas analysis. *Levene MI, Chervenak FA (eds) Fetal and Neonatal Neurology and Neurosurgery, 4th edn.* Churchill Livingstone.

ALI, M. M., KHOMPATRAPORN, C. & ZABINSKY, Z. B. 2005. A Numerical Evaluation of Several Stochastic Algorithms on Selected Continuous Global Optimization Test Problems. *Journal of Global Optimization,* 31**,** 635-672.

AMER-W HLIN, I. & MARŠ L, K. ST analysis of fetal electrocardiography in labor. Seminars in Fetal and Neonatal Medicine, 2011. Elsevier, 29-35.

APGAR, V. 1953. A proposal for a new method of evaluation of the newborn. *Curr Res Anaesth,* 32**,** 260-267.

ARMSTRONG, L. & STENSON, B. J. 2007. Use of umbilical cord blood gas analysis in the assessment of the newborn. *Arch. Dis. Child. Fetal Neonatal Ed.,* 92**,** F430.

AYRES-DE-CAMPOS, D. & BERNARDES, J. 2004. Comparison of fetal heart rate baseline estimation by SisPorto 2.01 and a consensus of clinicians. *Eur. J. Obstet. Gynecol. Reprod. Biol.,* 117**,** 174.

AYRES-DE-CAMPOS, D. & BERNARDES, J. 2010. Twenty-five years after the FIGO guidelines for the use of fetal monitoring: Time for a simplified approach? *International Journal of Gynecology & Obstetrics,* 110**,** 1-6.

AYRES-DE-CAMPOS, D., SOUSA, P., COSTA, A. & BERNARDES, J. 2008. Omniview-SisPorto® 3.5 – a central fetal monitoring station with online alerts based on computerized cardiotocogram+ST event analysis. *Journal of Perinatal Medicine.*

AZRA HAIDER, B. & BHUTTA, Z. A. 2006. Birth asphyxia in developing countries: current status and public health implications. *Current Problems in Pediatric and Adolescent Health Care,* 36**,** 178-188.

BAAN, J. J., BOEKKOOI, P., TEITEL, D. & RUDOLPH, A. 1993. Heart rate fall during acute hypoxemia: a measure of chemoreceptor response in fetal sheep. *Journal of developmental physiology,* 19**,** 105.

BADAWI, N., KURINCZUK, J. J., KEOGH, J. M., ALESSANDRI, L. M., O'SULLIVAN, F., BURTON, P. R., PEMBERTON, P. J. & STANLEY, F. J. 1998. Intrapartum risk factors for newborn encephalopathy: the Western Australian case-control study. *Bmj,* 317**,** 1554-1558.

BANG, A. T., BANG, R. A., BAITULE, S. B., REDDY, H. M. & DESHMUKH, M. D. 2005. Management of birth asphyxia in home deliveries in rural Gadchiroli: the effect of two types of birth attendants and of resuscitating with mouth-to-mouth, tube-mask or bag–mask. *Journal of Perinatology,* 25**,** S82-S91.

BARBER, V., LINSELL, L., LOCOCK, L., POWELL, L., SHAKESHAFT, C., LEAN, K. & BROCKLEHURST, J. P. 2013. Electronic fetal monitoring during labour and anxiety levels in women taking part in a RCT. *British Journal of Midwifery,* 21**,** 394-403.

BARRY, S. S. 2004. The CTG and the timing and mechanism of fetal neurological injuries. *Best Practice & Research Clinical Obstetrics & Gynaecology,* 18**,** 437-456.

BASSEVILLE, M. L. & NIKIFOROV, I. 1993. *Detection of Abrupt Changes - Theory and Application*, Prentice Hall, Inc.

BAUER, A., KANTELHARDT, J. W., BARTHEL, P., SCHNEIDER, R., M KIKALLIO, T., ULM, K., HNATKOVA, K., SCH MIG, A., HUIKURI, H. & BUNDE, A. 2006. Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study. *The lancet,* 367**,** 1674-1681.

BERNARDES, J., AYRES-DE-CAMPOS, D., COSTA-PEREIRA, A., PEREIRA-LEITE, L. & GARRIDO, A. 1998. Objective computerized fetal heart rate analysis. *International Journal of Gynecology & Obstetrics,* 62**,** 141-147.

BISHOP, C. 2006. *Pattern Recognition and Machine Learning*, Berlin: Springer.

BLANCHET, F. G., LEGENDRE, P. & BORCARD, D. 2008. FORWARD SELECTION OF EXPLANATORY VARIABLES. *Ecology,* 89**,** 2623-2632.

BRACERO, L. A., ROSHANFEKR, D. & BYRNE, D. W. 2000. Analysis of antepartum fetal heart rate tracing by physician and computer. *Journal of Maternal-Fetal and Neonatal Medicine,* 9**,** 181-185.

BREIMAN, L. 2001. Random Forests. *Machine Learning,* 45**,** 5-32.

BURGES, C. J. C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* 2**,** 121-167.

CARTER, B. C., VERSHININ, M. & GROSS, S. P. 2008. A Comparison of Step-Detection Methods: How Well Can You Do? *Biophysical Journal,* 94**,** 306-319.

CARTER, N. J. & CROSS, R. A. 2005. Mechanics of the kinesin step. *Nature,* 435**,** 308-312.

CAZARES, S. M. 2002. *Automated identification of abnormal patterns in the intrapartum cardiotocogram.* Doctor of Philosophy, University of Oxford.

CHANDRAHARAN, E. & ARULKUMARAN, S. 2007. Prevention of birth asphyxia: responding appropriately to cardiotocograph (CTG) traces. *Best Practice & Research Clinical Obstetrics & Gynaecology,* 21**,** 609-624.

CHANG, C.-C. & LIN, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.,* 2**,** 1-27.

CHAUHAN, S. P., KLAUSER, C. K., WOODRING, T. C., SANDERSON, M., MAGANN, E. F. & MORRISON, J. C. 2008. Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability. *American Journal of Obstetrics and Gynecology,* 199**,** 623.e1-623.e5.

CHEN, H.-Y., CHAUHAN, S. P., ANANTH, C. V., VINTZILEOS, A. M. & ABUHAMAD, A. Z. 2011. Electronic fetal heart rate monitoring and its relationship to neonatal and infant mortality in the United States. *American journal of obstetrics and gynecology,* 204**,** 491. e1-491. e10.

CHUD ČEK, V., BURŠA, M., JANKŮ, P., HUPTYCH, M. & LHOTSK , L. 2014. Open access intrapartum CTG database. *BMC pregnancy and childbirth,* 14**,** 16.

CHUD ČEK, V., SPILKA, J., JANKŮ, P., KOUCK , M., LHOTSK , L. & HUPTYCH, M. 2011. Automatic evaluation of intrapartum fetal heart rate recordings: a comprehensive analysis of useful features. *Physiological measurement,* 32**,** 1347.

CHUD ČEK, V. A. 2011. Assessment of features for automatic CTG analysis based on expert annotation. *33rd Annual International Conference of the IEEE EMBS.* Boston, Massachusetts USA,.

CLARK, S. L., NAGEOTTE, M. P., GARITE, T. J., FREEMAN, R. K., MILLER, D. A., SIMPSON, K. R., BELFORT, M. A., DILDY, G. A., PARER, J. T., BERKOWITZ, R. L., D'ALTON, M., ROUSE, D. J., GILSTRAP, L. C., VINTZILEOS, A. M., VAN DORSTEN, J. P., BOEHM, F. H., MILLER, L. A. & HANKINS, G. D. V. 2013. Intrapartum management of category II fetal heart rate tracings: towards standardization of care. *American journal of obstetrics and gynecology,* 209**,** 89-97.

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psych. Meas.,* 20**,** 37.

CORRALEJO, R. 2011. Feature Selection using a Genetic Algorithm in a Motor Imagerybased Brain Computer Interface. *33rd Annual International Conference of the IEEE EMBS.* Boston, Massachusetts USA,.

CORTES, C. & VAPNIK, V. 1995. Support-vector networks. *Machine Learning,* 20**,** 273-297.

COSTA, A., AYRES DE-CAMPOS, D., COSTA, F., SANTOS, C. & BERNARDES, J. 2009. Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals. *Am. J. Obstset. Gynecol.,* 201**,** 464.e1.

COSTA, M. A., AYRES-DE-CAMPOS, D., MACHADO, A. P., SANTOS, C. C. & BERNARDES, J. 2010. Comparison of a computer system evaluation of intrapartum cardiotocographic events and a consensus of clinicians. *J. Perinat. Med.,* 38**,** 191.

COSTELLO, A. & MANANDHAR, D. S. 1994. Perinatal asphyxia in less developed countries. *Archives of Disease in Childhood-Fetal and Neonatal Edition,* 71**,** F1-F3.

CRISTIANINI, N., AND SHAWE-TAYLOR, J 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods.,* Cambridge

University Press.
CZABANSKI, R., JEZEWSKI, J., MATONIA, A. & JEZEWSKI, M. 2012. Computerized analysis of fetal heart rate signals as the predictor of neonatal acidemia. *Expert Systems with Applications,* 39**,** 11846-11860.
DAWES, G., MOULDEN, M., SHEIL, O. & REDMAN, C. 1992. Approximate entropy, a statistic of regularity, applied to fetal heart rate data before and during labor. *Obstetrics and gynecology,* 80**,** 763.
DAWES, G. S., MOULDEN, M. & REDMAN, C. W. 1991. System 8000: computerized antenatal FHR analysis. *Journal of Perinatal Medicine-Official Journal of the WAPM,* 19**,** 47-51.
DAWES, G. S., MOULDEN, M. & REDMAN, C. W. G. 1990. Limitations of antenatal fetal heart rate monitors. *Am. J. Obstet. Gynecol.,* 162**,** 170.
DEVOE, L., GOLDE, S., KILMAN, Y., MORTON, D., SHEA, K. & WALLER, J. 2000. A comparison of visual analyses of intrapartum fetal heart rate tracings according to the new National Institute of Child Health and Human Development guidelines with computer analyses by an automated fetal heart rate monitoring system. *American Journal of Obstetrics and Gynecology,* 183**,** 361-366.
EIBE, F., MARK, H. & BERNHARD, P. 2003. Locally weighted naive Bayes. Morgan Kaufmann.
EL-MIHOUB, T. A., HOPGOOD, A. A., NOLLE, L. & BATTERSBY, A. 2006. Hybrid genetic algorithms: a review. *Engineering Letters,* 13**,** 124-137137.
ENGELBRECHT, A. P. 2002. *Computational Intelligence: An Introduction*, John Wiley & Sons, Ltd.
FIGO 1987. Guidelines for the Use of Fetal Monitoring. *International Journal of Gynecology & Obstetrics,* 25**,** 159-167.
FRANK, A. & ASUNCION, A. 2010. UCI machine learning repository.
FREEMAN, R. K., GARITE, T. J. & NAGEOTTE, M. P. 2003. *Fetal Heart Rate Monitoring,* Lippincott Williams & Wilkins.
FULCHER, B. 2012. *Highly comparative time-series analysis.* DPhil, University of Oxford.
FULCHER, B., GEORGIEVA, A., REDMAN, C. & JONES, N. Highly comparative fetal heart rate analysis. Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, 2012. IEEE, 3135-3138.
GEORGIEVA, A., MOULDEN, M. & REDMAN, C. W. G. 2013a. Umbilical cord gases in relation to the neonatal condition: the EveREst plot. *European Journal of Obstetrics & Gynecology and Reproductive Biology,* 168**,** 155-160.
GEORGIEVA, A., PAPAGEORGHIOU, A. T., PAYNE, S. J., MOULDEN, M. & REDMAN, C. W. G. 2013b. Phase rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidaemia at birth. *Brit J Obstet Gynaecol,* In Press.
GEORGIEVA, A., PAYNE, S., MOULDEN, M. & REDMAN, C. Automated fetal heart rate analysis in labor: decelerations and overshoots. AIP Conference

Proceedings, 2010. 255.

GEORGIEVA, A., PAYNE, S., MOULDEN, M. & REDMAN, C. 2011a. Computerized fetal heart rate analysis in labor: detection of intervals with un-assignable baseline. *Physiological Measurement,* 32**,** 1549.

GEORGIEVA, A., PAYNE, S., MOULDEN, M. & REDMAN, C. 2012. Relation of fetal heart rate signals with unassignable baseline to poor neonatal state at birth. *Medical and Biological Engineering and Computing,* 50**,** 717-725.

GEORGIEVA, A., PAYNE, S., MOULDEN, M. & REDMAN, C. 2013c. Artificial neural networks applied to fetal monitoring in labour. *Neural Computing & Applications,* 22**,** 85-93.

GEORGIEVA, A., PAYNE, S. & REDMAN, C. W. G. 2009. Computerised electronic foetal heart rate monitoring in labour: automated contraction identification. *Medical and Biological Engineering and Computing,* 47**,** 1315-1320.

GEORGIEVA, A., PAYNE, S. J., MOULDEN, M. & REDMAN, C. W. G. Computerized intrapartum electronic fetal monitoring: Analysis of the decision to deliver for fetal distress. Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE, Aug. 30 2011-Sept. 3 2011 2011b. 5888-5891.

GEORGOULAS, G., GAVRILIS, D., TSOULOS, I. G., STYLIOS, C., BERNARDES, J. & GROUMPOS, P. P. 2007a. Novel approach for fetal heart rate classification introducing grammatical evolution. *Biomedical Signal Processing and Control,* 2**,** 69-79.

GEORGOULAS, G., STYLIOS, C., CHUDACEK, V., MACAS, M., BEMARDES, J. & LHOTSKA, L. 2007b. Classification of Fetal Heart Rate Signals Based on Features Selected Using the Binary Particle Swarm Algorithm. *In:* KIM, S. I. & SUH, T. S. (eds.) *World Congress on Medical Physics and Biomedical Engineering 2006, Vol 14, Pts 1-6.* Berlin: Springer-Verlag Berlin.

GILBERT, R. O. 1987. *Statistical methods for environmental pollution monitoring*, Wiley. com.

GILLMER, M. & COMBE, D. 1979. Intrapartum fetal monitoring practice in the United Kingdom. *BJOG: An International Journal of Obstetrics & Gynaecology,* 86**,** 753-758.

GOLDBERG, D. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Professional.

GOLDBERGER, A. L., AMARAL, L. A., GLASS, L., HAUSDORFF, J. M., IVANOV, P. C., MARK, R. G., MIETUS, J. E., MOODY, G. B., PENG, C.-K. & STANLEY, H. E. 2000. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation,* 101**,** e215-e220.

GRIMES, D. A. & PEIPERT, J. F. 2010. Electronic fetal monitoring as a public health screening program: the arithmetic of failure. *Obstetrics & Gynecology,* 116**,** 1397.

GUYON, I. & ELISSEEFF, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research,* 3.

GUYON, I., WESTON, J., BARNHILL, S. & VAPNIK, V. 2002. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning,* 46**,** 389-422.

HELGASON, H., ABRY, P., GONCALVES, P., GHARIB, C. & GAUCHERAND, P. 2011. Adaptive Multiscale Complexity Analysis of Fetal Heart Rate. *IEEE Transactions on Biomedical Engineering,* 58.

HON, E. H. 1963. The classification of fetal heart rate I. A working classification. *Obstetrics & Gynecology,* 22**,** 137-146.

HONEYCUTT, A. A., GROSSE, S. D., DUNLAP, L. J., SCHENDEL, D. E., CHEN, H., BRANN, E. & AL HOMSI, G. 2003. Economic costs of mental retardation, cerebral palsy, hearing loss, and vision impairment. *Research in Social Science and Disability,* 3**,** 207-228.

HU, X. Q., CUI, M. & CHEN, B. 2009. *Feature Selection Based on Random Forest and Application in Correlation Analysis of Symptom and Disease,* New York, Ieee.

INZA, I., LARRA AGA, P., ETXEBERRIA, R. & SIERRA, B. 2000. Feature Subset Selection by Bayesian network-based optimization. *Artificial Intelligence,* 123**,** 157-184.

JAIN, A. K., DUIN, R. P. W. & JIANCHANG, M. 2000. Statistical pattern recognition: a review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 22**,** 4-37.

JARQUE, C. M. & BERA, A. K. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters,* 6**,** 255-259.

JEZEWSKI, J., WROBEL, J., HOROBAL, K., KUPKA, T. & MATONIA, A. Centralised fetal monitoring system with hardware-based data flow control. Advances in Medical, Signal and Information Processing, 2006. MEDSIP 2006. IET 3rd International Conference On, 2006. IET, 1-4.

JIANG, R., TANG, W. W., WU, X. B. & FU, W. H. 2009. A random forest approach to the detection of epistatic interactions in case-control studies. *Bmc Bioinformatics,* 10.

JOHANSON, R. & MENON, V. 1999. Vacuum extraction versus forceps for assisted vaginal delivery. *Cochrane Database of Systematic Reviews,* 2.

KEITH, R. D. F., BECKLEY, S., GARIBALDI, J. M., WESTGATE, J. A., IFEACHOR, E. C. & GREENE, K. R. 1995. A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. *BJOG: An International Journal of Obstetrics & Gynaecology,* 102**,** 688-700.

KERSSEMAKERS, J. W. J., LAURA MUNTEANU, E., LAAN, L., NOETZEL, T. L., JANSON, M. E. & DOGTEROM, M. 2006. Assembly dynamics of microtubules at molecular resolution. *Nature,* 442**,** 709-712.

KOHAVI, R. & JOHN, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence,* 97**,** 273-324.

LEVIT, K., WIER, L., STRANGES, E., RYAN, K. & ELIXHAUSER, A. 2009.

HCUP facts and figures: statistics on hospital-based care in the United States, 2007. *Rockville, MD: Agency for Healthcare Research and Quality*.

LITTLE, M. A. & JONES, N. S. 2010. Generalized Methods and Solvers for Noise Removal from Piecewise Constant Signals. *ArXiv e-prints* [Online], 1012. Available: http://adsabs.harvard.edu/abs/2010arXiv1012.5095L [Accessed December 1, 2010].

MACDONALD, D., GRANT, A., SHERIDAN-PEREIRA, M., BOYLAN, P. & CHALMERS, I. 1985. The Dublin randomized controlled trial of intrapartum fetal heart rate monitoring. *Am J Obstet Gynecol,* 152**,** 524-539.

MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967,* 1**,** 281-297.

MALIN, G. L., MORRIS, R. K. & KHAN, K. S. 2010. Strength of association between umbilical cord pH and perinatal and long term outcomes: systematic review and meta-analysis. BMJ Publishing Group Ltd.

MALLAT, S. & HWANG, W. L. 1992. Singularity detection and processing with wavelets. *Information Theory, IEEE Transactions on,* 38**,** 617-643.

MANN, H. B. & WHITNEY, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics,* 18**,** 50-60.

MARR, D. & HILDRETH, E. 1980. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences,* 207**,** 187-217.

MCKINNEY, S. A., JOO, C. & HA, T. 2006. Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. *Biophysical Journal,* 91**,** 1941-1951.

MINI O, A. M., HERON, M. P. & SMITH, B. L. 2006. Deaths: preliminary data for 2004. *National vital statistics reports,* 54**,** 1-49.

MITCHELL, M. 2001. *An Introduction to Genetic Algorithms,* Cambridge, MA, MIT Press.

MRAZEK, P., WEICKERT, J. & BRUHN, A. 2006. On Robust Estimation and smoothing with Spatial and Tonal Kernels. *In:* KLETTE, R., KOZERA, R., NOAKES, L. & WEICKERT, J. (eds.) *Geometric Properties for Incomplete data.* Springer Netherlands.

MURPHY, K. W., JOHNSON, P., MOORCRAFT, J., PATTINSON, R., RUSSELL, V. & TURNBULL, A. 1990. Birth asphyxia and the intrapartum cardiotocograph. *BJOG: An International Journal of Obstetrics & Gynaecology,* 97**,** 470-479.

MURRAY, D. M., O'RIORDAN, M. N., HORGAN, R., BOYLAN, G., HIGGINS, J. R. & RYAN, C. A. 2009. Fetal Heart Rate Patterns in Neonatal Hypoxic-Ischemic Encephalopathy: Relationship with Early Cerebral Activity and Neurodevelopmental Outcome. *Amer J Perinatol,* 26**,** 605,612.

NICHHD 1997. Electronic Fetal Heart Rate Monitoring: Research Guidelines for Interpretation. *J Obstet Gynecol Neonatal Nurs,* 26**,** 635-640.

PALSDOTTIR, K., DAGBJARTSSON, A., THORKELSSON, T. & HARDARDOTTIR, H. 2007. Birth asphyxia and hypoxic ischemic

encephalopathy, incidence and obstetric risk factors]. *Læknablaðið,* 93**,** 595.

PARDEY, J., MOULDEN, M. & REDMAN, C. W. G. 2002. A computer system for the numerical analysis of nonstress tests. *American Journal of Obstetrics and Gynecology,* 186**,** 1095-1103.

PARER, J. T. & HAMILTON, E. F. 2010. Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation. *Am J Obstet Gynecol,* 203**,** 451.e1-7.

RCOG 2001. The use of electronic fetal monitoring: Evidence-based clinical guideline. *Royal College of Obstetricians and Gynaecologists,* 8.

RCOG 2007. National Institute for Health and Clinical Excellence: Guidance. *Intrapartum Care: Care of healthy women and their babies during childbirth.* London: RCOG Press

National Collaborating Centre for Women's and Children's Health.

REDDY, A., MOULDEN, M. & REDMAN, C. W. 2009. Antepartum high-frequency fetal heart rate sinusoidal rhythm: computerized detection and fetal anemia. *American journal of obstetrics and gynecology,* 200**,** 407. e1-407. e6.

RUDIN, L. I., OSHER, S. & FATEMI, E. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena,* 60**,** 259-268.

RUDOLPH, G. 1994. Convergence analysis of canonical genetic algorithms. *Neural Networks, IEEE Transactions on,* 5**,** 96-101.

SADLER, B. M. & SWAMI, A. 1999. Analysis of multiscale products for step detection and estimation. *Information Theory, IEEE Transactions on,* 45**,** 1043-1051.

SALAMALEKIS, E., HINTIPAS, E., SALLOUM, I., VASIOS, G., LOGHIS, C., VITORATOS, N., CHRELIAS, C. & CREATSAS, G. 2006. Computerized analysis of fetal heart rate variability using the matching pursuit technique as an indicator of fetal hypoxia during labor. *Journal of Maternal-Fetal and Neonatal Medicine,* 19**,** 165-169.

SCHUMANN, A. Y., KANTELHARDT, J. W., BAUER, A. & SCHMIDT, G. 2008. Bivariate phase-rectified signal averaging. *Physica A: Statistical Mechanics and its Applications,* 387**,** 5091-5100.

SCHWARZ, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics,* 6**,** 461-464.

SMIRNOV, N. 1948. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics,* 19**,** 279-281.

SPENCER, J. A. D., BELCHER, R. & DAWES, G. S. 1987. The influence of signal loss on the comparison between computer analyses of the fetal heart rate in labour using pulsed doppler ultrasound (with autocorrelation) and simultaneous scalp electrocardiogram. *European Journal of Obstetrics & Gynecology and Reproductive Biology,* 25**,** 29-34.

SPILKA, J., CHUD ČEK, V., KOUCK , M., LHOTSK , L., HUPTYCH, M., JANKŮ, P., GEORGOULAS, G. & STYLIOS, C. 2011. Using nonlinear features for fetal heart rate classification. *Biomedical Signal Processing and Control*.

STEER, P., EIGBE, F., LISSAUER, T. & BEARD, R. 1989. Interrelationships among

abnormal cardiotocograms in labor, meconium staining of the amniotic fluid, arterial cord blood pH, and Apgar scores. *Obstetrics & Gynecology,* 74**,** 715-721.

SYMONDS, E. M., SAHOTA, D. & CHANG, A. 2001. *Fetal electrocardiography*, World Scientific Publishing Company.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 267-288.

TURAN, S., TURAN, O., BERG, C., MOYANO, D., BHIDE, A., BOWER, S., THILAGANATHAN, B., GEMBRUCH, U., NICOLAIDES, K. & HARMAN, C. 2007. Computerized fetal heart rate analysis, Doppler ultrasound and biophysical profile score in the prediction of acid‐base status of growth‐restricted fetuses. *Ultrasound in Obstetrics & Gynecology,* 30**,** 750-756.

TUV, E., BORISOV, A., RUNGER, G. & TORKKOLA, K. 2009. Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. *J. Mach. Learn. Res.,* 10**,** 1341-1366.

VAN LAAR, J. O. E. H., PORATH, M. M., PETERS, C. H. L. & OEI, S. G. 2008. Spectral analysis of fetal heart rate variability for fetal surveillance: review of the literature. *Acta Obstetricia et Gynecologica Scandinavica,* 87**,** 300-306.

VAN NIEL, T. G., MCVICAR, T. R. & DATT, B. 2005. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. *Remote Sensing of Environment,* 98**,** 468-480.

VINTZILEOS, A. M., NOCHIMSON, D. J., GUZMAN, E. R., KNUPPEL, R. A., LAKE, M. & SCHIFRIN, B. S. 1995. Intrapartum electronic fetal heart rate monitoring versus intermittent auscultation: a meta-analysis. *Obstetrics & Gynecology,* 85**,** 149-155.

WARRICK, H. E. F. P. D. & KEARNEY, R. E. 2009. Identification of the dynamic relationship between intrapartum uterine pressure and fetal heart rate for normal and hypoxic fetuses. *IEEE Trans. Biomed. Eng.,* 56**,** 1587.

WESTGATE, J. 2009. Computerizing the Cardiotocogram (CTG). *Medical Informatics in Obstetrics and Gynecology.* Parry, D., & Parry, E. (Eds.).

WESTGATE, J. A., WIBBENS, B., BENNET, L., WASSINK, G., PARER, J. T. & GUNN, A. J. 2007. The intrapartum deceleration in center stage: a physiologic approach to the interpretation of fetal heart rate changes in labor. *American journal of obstetrics and gynecology,* 197**,** 236. e1-236. e11.

WILLIAMS, J. 2012. *The Application of Phase-Rectified Signal Averaging to Intrapartum Cardiotocography Signals to Predict Fetal Acidemia.* Master of Engineering, University of Oxford.

XU, L., GEORGIEVA, A., PAYNE, S. J., MOULDEN, M. & REDMAN, C. W. G. 2012. Detection and Analysis of Pattern Readjustment in Fetal Heart Rate Signal. *MEDSIP 12.* Liverpool, UK.

XU, L., GEORGIEVA, A., PAYNE, S. J., MOULDEN, M. & REDMAN, C. W. G. 2013. Feature Selection for Computerized Fetal Heart Rate Analysis using Genetic Algorithms. *IEEE EMBS 2013.* Osaka, Japan.

YEH, P., EMARY, K. & IMPEY, L. 2012. The relationship between umbilical cord arterial pH and serious adverse neonatal outcome: analysis of 51 519 consecutive validated samples. *Brit J Obstet Gynaecol*, 119**,** 824-831.

YU, L. & LIU, H. 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res.*, 5**,** 1205-1224.