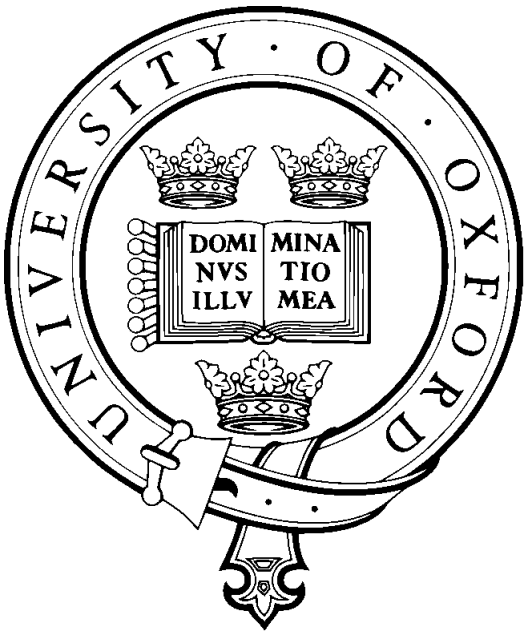
Methods of Classification of the Cardiotocogram



Alexander Paul Clibbon

St.Catherine's College

University of Oxford

Supervised by
Professor Stephen Payne
Submitted: Trinity Term,

August 18, 2016

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

Declaration

I declare that this thesis is entirely my own work, and except where otherwise stated, describes my own research.

A. P. Clibbon,

St.Catherines College

Acknowledgements

I am grateful to my supervisor Stephen Payne for his patience and for giving me the freedom to pursue my interests; to Antoniya Georgieva for her effort preparing the foundations of this work and guidance throughout. This work would not have been possible without the work of Mary Moulden and Chris Redman preparing the unique database. I'd also like to thank my colleagues and friends who've looked after me throughout the creation of this thesis.

Finally I would like to acknowledge the Research Council UK's Digital Economy Programme Grant EP/G036861/1 and the Oxford Centre for Doctoral Training in Healthcare Innovation for their support through funding.

Methods of Classification of the Cardiotocogram.

Assessing the condition of the fetus during labor is a guessing game due to its relative inaccessibility, and the simple fact that birth is a highly strenuous process for even a completely healthy fetus. The current practice of Cardiotocogram (CTG) monitoring introduced to detect signs of fetal hypoxia has not yet been demonstrated to have made a significant impact on fetal survival rates.

A range of methods to identify severe hypoxia from CTG traces have been proposed in the literature, but there are few instances in which they are directly compared. This restricts both assessment of the potential of the CTG itself and progress towards an optimal decision making process.

This Thesis compares CTG classification techniques proposed in the literature and their potential extensions. A comparison between four classifiers previously assessed - Adaboost, Artificial Neural Networks (ANN), Random Forest (RF), Support Vector Machine (SVM) - and two proposed classifiers - Bayesian ANN (BANN), Relevance Vector Machine - was conducted using a database of 7,568 cases and two open source databases. The Random Forest (RF) achieved the highest average result and is proposed as a benchmark classifier.

The proposal to use model certainty to introduce a third, unclassified, class was investigated using the BANN. An increase in the classification accuracy was demonstrated, however the proportion of cases in the unclassified class was too great to be of practical value.

The information content of time series was explored using a Hidden Markov Model (HMM). The average performance of the HMM was comparable with the performance of the benchmark with a smaller distribution across validation folds, demonstrating that time-series information provides more stable estimates of class than stationary methods.

Finally a method of system identification was implemented. Significant differences between feature trends and histograms in low pH (< 7.1) and healthy pH (≥ 7.1) cases were observed. These features were used as classifier inputs, and achieved performance similar to existing feature sets. When these features were aligned according the onset of stage 2 labour three unique trend patterns were discovered.

Abbreviations

AAWFT	Amplitude adjusted windowed Fourier transform
AB	AdaBoost / Adaptive Boosting
ACOG	The American College of Obstetrics and Gynaecology
AIC	Akaike information criterion
ANN	Artificial Neural Network
ANNC	ANN committee
ANFIS	Adaptive Neuro-Fuzzy Inference system
AR	Auto-regression
ARD	Automatic relevance determination
ARMA	Auto-regressive moving average
AUC	Area under the Receiver-Operator curve
ANOVA	Analysis of variance
BANN	Bayesian ANN
BD	Base deficit
BE	Base excess
BPM	Beats per minute
BPSRA	Bi-variate phase rectified signal averaging
CART	Classification and Regression Trees
CE	Conformité Européene
CP	Cerebral Palsy
CS	Caesarean section
CTG	Cardiotocogram
CTU-UHB	Czech Technical University and University Hospital in Brno
CV	Cross validation
ECG	Electrocardiogram
EFM	Electronic Fetal Monitoring
EHG	Electrohysterogram
EVEREST	EVEnt Rate ESTimation

FDI	Fetal distress index
fECG	fetal ECG
FHR	Fetal Heart Rate
FIGO	International Federation of Gynecology and Obstetrics.
FSE	Fetal scalp electrode
IA	Intermittent auscultation
IG	Information Gain
IH	Intrapartum hypoxia
FPO	Fetal Pulse Oximetry
GA	Genetic Algorithm
GAME-NN	Group of Adaptive Models Evolution neural network
IH	Intrapartum hypoxia
IQ	Intelligence quotient
IRF	Impulse response function
HIE	Hypoxic-Ischemic Encephalopathy
HMM	Hidden Markov Model
KDE	Kernel density estimate
K-NN	K-nearest neighbour
KS	KolmogorovSmirnov
LASSO	Least absolute shrinkage and selection operator
LTV	Long term variation
MAP	Mean arterial pressure
MDL	Minimum Description Length
MLP	Multi-layer Perceptron
MSE	Mean square error
MXE	Mean cross entropy
nICU	Neonatal Intensive Care Unit
NE	Neonatal Encephalopathy
NI	Neural index
NICE	National Institute for Health and Care Excellence

NICHD	National Institute of Child Health and Human Development
NHS	United Kingdom National Health Service
OCFMT	Oxford Centre for Fetal Monitoring Technologies
OLS	Ordinary Least Squares
PCA	Principal component analysis
PDF	Probability density function
PI	Pseudo inverse
PNN	Probabilistic Neural Network
PRSA	Phase rectified signal averaging
RELIEF	Relevance in Estimating Features
RBF	Radial Basis Function
RF	Random Forest
RVM	Relevance Vector Machine
SMOTE	Synthetic Minority Oversampling Technique
SSI	Signal stability index
STAN	ST-analysis
STV	Short Term Variability
SVD	Singular Value Decomposition
SVM	Support Vector Machine
UA	Uterine Activity
UCI	University of California, Irvine
URR	Unsupervised relative reduct
UK	United Kingdom
VAF	Variance accounted for

Glossary

Acidemia - A condition of lowered systemic pH. Fetal respiratory acidemia is a condition defined clinically as pH$<$ 7.20 and a base excess (BE) of less than -8mmol/L in any umbilical vessel, though a high risk of adverse outcome has been associated with the much lower value of pH$<$ 7.0 [6].

Acidosis - The process of lowering the systemic pH, and may arise as a result of severe hypoxemia leading to the production of lactate (metabolic acidosis) or the accumulation of CO_2 in the blood (respiratory acidosis).

Antepartum labour - The period of labour up until the start of the intrapartum period.

Apgar test - An objectively scored assessment of fetal health, performed 1, 5, and 10 minutes after birth, named after its creator.

Base Deficit/Excess - A measurement of the metabolic component of acid-base disturbances in mmol/L [78].

Caesarean section - A surgical procedure in which the fetus is removed from the mother by cutting through the maternal abdominal wall.

Cerebral Palsy - A group of disorders which may affect movement, sensation, and reasoning, and may cause seizures.

Fetal Asphyxia - A condition of reduced gaseous exchange leading to progressive hypoxemia [112].

Fetus - The unborn baby.

Hypotension - A condition of abnormally low blood pressure.

Hypoxemia - A condition of abnormally low oxygen saturation in arterial blood, a common cause of hypoxia.

Hypoxia/Anoxia - A condition where the body tissue is deprived of an adequate oxygen supply, anoxia indicating a complete deprivation.

Hypoxic-Ischemic Encephalopathy - Damage to the fetal brain caused by a lack of oxygenated blood.

Intrapartum labour - The period of labour from the end of stage 1 labour until delivery of the fetus.

pH - A measure of the acidity or alkalinity of a solution, with a value of 7 being considered neutral, lower values acidic, and higher values more alkaline.

Neonatal Encephalopathy - A syndrome of brain dysfunction affecting the fetus.

Neonate - A newly born baby, less than 4 weeks old.

Oxytocin - A medication used to artificially induce contractions in the mother.

Parity - A count of the number of births carried to gestational age by the mother. A mother with no previous pregnancies will be described as nulliparous.

Contents

Chapter 1

Introduction

1.1 Motivation

In 2014 there were 700,000 births in England and Wales [132], a number which has been rising yearly. Most of these births proceeded without incidence, however 4.3 in every 1000 low risk births experienced complications, of which 40% could be attributed to intrapartum hypoxia (IH) [20].

IH is a condition of reduced oxygen supply to the fetus which occurs to some degree in all labours as strong contractions in the final stages of delivery restrict the flow of oxygenated blood through the umbilical cord. Severe and prolonged IH leads to Neonatal Encephalopathy (NE), is strongly linked to Cerebral Palsy (CP), and is the largest single contributor to neurological impairments and mortality in intrapartum labour [151, 171].

Treating severe IH involves intervention to hasten delivery, through either Caesarean section (CS) or mechanical aid, after which the fetal oxygen supply will be restored. The likelihood of permanent damage due to IH increases the longer the condition continues. Earlier intervention will therefore improve outcomes, but must be weighed against other complications which may be caused by intervention.

To help clinicians decide whether to intervene labours may be monitored using the cardiotocogram (CTG). Uncertainties in CTG interpretation mean cases of severe IH may be missed, and along with increasingly conservative attitudes to clinical care during labour have led to a rising trend towards operative delivery. A major factor behind this changing medical attitude has been the need to reduce the cost of obstetric litigation fees related to CTG interpretation and CP [42].

A recent study of United Kingdom National Health Service (NHS) litigation costs showed that from April 1^{st} 2000 to March 31^{st} 2010 £3.1 billion was spent on claims

made to the Obstetrics and Gynaecology department, equal to the sum of all other specialities. The average claim for CTG interpretation came to £1.55 million, second to CP claims at £2.4 million [129]. Stronger, evidence based, interpretation of the CTG could help to reduce unnecessary interventions and to reduce the incidence of hypoxia related intrapartum events.

1.2 Defining the problem

Most intrapartum events, such as misalignment in the birth canal (breech presentation, shoulder dystocia etc.), uterine rupture, or maternal pre-eclampsia have readily identifiable symptoms. Identification of IH and the strongly related fetal acidosis remain difficult since the relevant physiological features - cerebral oxygen saturation and blood pressure - cannot be directly measured.

The CTG, which is a combined measurement of the fetal heart rate (FHR) and the mother's uterine activity (UA) or contraction patterns, has been successful in recognising changes in heart rate patterns due to chronic hypoxia in antepartum labour, and is routinely used for such [39].

Based on this success, it has become the most widely used method of intrapartum fetal monitoring in cases of a high risk labour or when fetal distress is suspected [130]. Unlike antepartum labour its use has not been demonstrated to reduce fetal morbidity and mortality, and instead to have slightly increased operative delivery rates [168, 3]. The poor performance of the CTG in managing IH is a result of several factors:

1. Some degree of IH will be present in all births, which some fetuses will cope with better than others. The problem is therefore identifying a threshold above which permanent damage will be caused, and identifying those cases far enough

in advance for treatment to prevent damage.

2. Signal quality is lower due to an increase in maternal and fetal movements, and unlike the antepartum stage it is not possible to wait for a higher quality sample.

In the case of antepartum assessment automated methods such as the Dawes-Redman system [39] were successful in improving the accuracy of diagnoses made using the CTG. These methods have not yet been extended successfully to the intrapartum case. Multiple commercial and research groups are investigating automation of diagnosis in the intrapartum stage to improve performance and provide an objective assessment in defence of litigation.

Progress in this area has been hindered the by poor availability of data due to the rarity of adverse incidents with complete CTG traces, and the fact that direct physiological measurements cannot be taken during labour. Therefore any transient periods of severe IH are ignored when labelling CTG traces, meaning individual patterns may be labelled incorrectly.

1.3 Objectives of this Thesis

This Thesis intends to review automated CTG classification using a uniquely large dataset, the Oxford Centre for Fetal Monitoring Technologies (OCFMT) database. The information content of the CTG is explored using the wide range of potential classifiers proposed in the literature, and a comparison between these is made to estimate an upper limit of classifier performance achieved using these methods and to establish a benchmark against which further developments may be compared.

Three proposed strategies to improve identification of IH are investigated in detail.

The first is the inclusion of confidence bounds on decisions, the second the inclusion of time-series information and the third an investigation into the use of system identification (SI) to integrates prior knowledge into the process of classification.

1.4 Overview

Chapter 2 expands on the problem defined in Section 1.2, the physiology of the fetus and the options currently available to assess fetal condition used in practice. Chapter 3 looks at the introduction of computerized methods of CTG assessment, and details the necessary steps to classification - feature extraction, selection, and classification, including details on commercially available decision support systems. Chapter 4 evaluates the classifiers proposed in the literature, and introduces several new classifiers. Chapter 5 covers the implementation of a method of system identification, and the insights into the data this method provides. Finally Chapter 6 summarises the recommendations from this investigation, and suggests possible directions for future research.

Chapter 2

The Physiology of Labour

This Chapter provides an understanding of the clinical side of labour management, the role of the CTG, and current medical practice. The stages of labour and are defined in Section 2.1. The fetal physiology is described in Section 2.2. Hypoxia, its effect on the fetus, treatment, and assessment are examined is Section 2.3. Finally the options for monitoring the fetus for signs of IH are described, starting with the focus of this Thesis, the CTG, in Section 2.4, and alternatives in Section 2.5.

2.1 The stages of labour

Labour begins with the onset of regular contractions, after which it is split into four distinct stages. Stage 1 labour begins with these regular contractions during which the cervix dilates, and ends when the cervix lies flush with the vaginal wall. In practice this is difficult and unnecessary to measure, so the second stage is assumed to begin when dilation has reached 10cm, and typically lasts from 2-20 hours.

Stage 1 labour can be further split into three sub-phases, the first being a latent phase during which the cervix slowly dilates to roughly 4cm. Admission normally occurs during this phase, which may last 2-10 hours. This is followed by a period of active dilation, where the cervix rapidly dilates at a rate of roughly 1cm/hour. Finally there is a second stage of slow dilation, accompanied by the descent of the foetus into the birth canal. During this period contractions increase in frequency, from lasting 10-40 seconds and occurring every half hour to longer contractions of 30-90 seconds every 5-10 minutes.

During Stage 2 labour, the foetus continues to descend through the birth canal and is expelled by continuous contractions which last for 30-60 seconds and occur on average every two minutes. During this stage the mother will be urged to push with

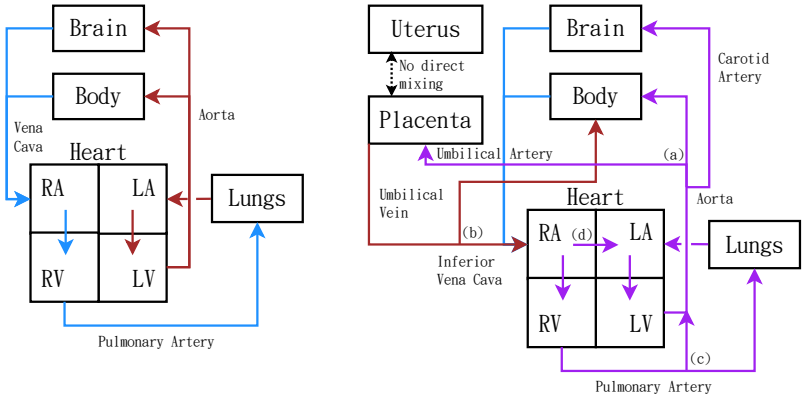
the contractions to aid with the birth. The third stage of labour begins once the fetus has been expelled, and ends with the expulsion of the placenta. The fourth and final stage concerns the hours after active labour, where uterine tone is restored.

Clinically the terms antepartum, intrapartum, and postpartum are often used alongside the stages. These terms are divided according to the clinical attention given to the mother. Intrapartum is defined as the period from the final phase of stage 1 and the arrival of frequent strong contractions through to the end of the third stage with delivery of the placenta, during which attention will be frequent or in high-risk cases constant.

Antepartum refers to the entire pregnancy preceding the intrapartum period, and is a stable period where care concerns the detection of chronic or congenital conditions through regular ultrasound surveys. Finally in the postpartum period care and attention are given towards recovery after the strain of childbirth, pain relief and psychological care. This Thesis is focused on monitoring and care delivered during intrapartum labour.

2.2 Fetal physiology

The circulatory system of the fetus is different to that found after birth as the oxygen required for metabolism is not acquired from gaseous transfer in the lungs; in fact the fetal circulation almost entirely bypasses the pulmonary circuit due to a connection between the left and right atria of the heart which will close at birth. During development the fetus is supplied with oxygenated blood via the umbilical cord, which contains a single arterial vessel transporting de-oxygenated blood, an offshoot from the iliac arteries in the pelvis of the fetus, and a single venous connection containing



(a) Adult Circulation (b) Fetal Circulation

Figure 2.1: A comparison between the adult and fetal physiology. Of particular attention are: a) the umbilical arteries which offshoot from the illiac arteries at the bottom of the descending aorta; b) the umbilical vein flowing mostly into the inferior vena cava, with a small amount shunted off to the heart and central organs; c) the connection between the pulmonary artery and the aorta; and finally d) the connection between the left and right atriums of the heart. Together these circuits preferentially supply the heart, the central organs and the brain. Image adapted from [140].

fresh blood from the placenta of which a portion is shunted directly to the inferior Vena Cava; the remainder entering at the liver. This system, shown in Figure 2.1, prioritizes distribution to the core organs and the fetal brain. The placenta lies at one end of the umbilical cord flush with the wall of the uterus; a spongy mass richly populated with blood vessels where metabolic waste products are removed and blood in the fetal circulatory system from the umbilical artery is resupplied by diffusion through a membrane - there is no direct contact between fetal and maternal blood. There is no neural regulation of the placental blood flow, but the vessels which feed it are subject to locally induced vasoconstriction [101].

Fetal blood contains a unique form of haemoglobin which has a greater affinity for oxygen than its adult counterpart, allowing for passive transport of oxygen through the

placental membrane from maternal blood, though transport of nutrients and waste are aided by active systems. At birth this will represent 60-80% of the fetal blood [167]. This however inhibits delivery of oxygen to tissues, so fetal blood has a greater haematocrit (blood cell count) to compensate, and takes advantage of the Bohr effect; the reduction of haemoglobin's oxygen binding affinity with a decrease in pH (or likewise an increase in CO_2 concentration). Together these effects increase the oxygen carrying capacity of fetus, allowing it to withstand long periods of hypoxia.

The fetal heart provides blood pressure in the fetal circuit, which at term is expected to be roughly 45mmHg [169], dropping as it passes through the placenta to roughly 20mmHg in the umbilical vein. At 40 weeks, the blood flow rate through the umbilical cord is typically 240mL/min [101], and the total fetal blood volume is estimated to be 90-115ml/kg [100], significantly higher than that in adults due to the additional volume contained within the placenta. Due to practical considerations much of this information is derived from sheep fetuses, which have very similar sizes and heart rates to humans.

The FHR is controlled by changes in the sympathetic-parasympathetic balance [183], which is governed by two systems, chemoreceptors and baroreceptors, responding to changes in pH and pressure respectively. Baroreceptors are located in the carotid artery and the aortic arch, with the latter responding only to increases in blood pressure. Baroreceptors maintain blood pressure by increasing heart rate, contractility (the strength of a heartbeat) and peripheral vasoconstriction in response to decreases in blood pressure and visa versa. Notable is the Cushing reflex, where an increase in intracranial pressure due to compression of the fetal head constricts cerebral arteries, causing bradycardia. This rapid increase in the perceived pressure gives a much faster response than that caused by occlusion of the umbilical artery alone.

There are multiple chemoreceptors in the fetus - peripheral receptors located at the aortic arch and carotid sinus, and a central receptor in the medulla [86]. Stimulation of the aortic receptor results in bradycardia and a drop in blood pressure. Stimulation of the carotid receptor may result in a range of effects, a drop or rise in heart rate, and attempts at respiration in term fetuses. A drop in fetal heart rate preserves myocardial function by reducing oxygen consumption in the heart tissue. The severity of the drop in heart rate is proportional to the severity of acidosis suffered by the fetus [199].

2.3 Intrapartum hypoxia

A hypoxic insult is a period of time where oxygen consumption of the fetus outstrips the supply. Hypoxic insults can be either severe sustained deficiencies caused by a medical event, such as cord entanglement or placental rupture, or brief repeated occurrences due to umbilical cord collapse caused by either contractions or compression due to fetal/maternal movement. The former are easy to identify but difficult to foresee or prevent, whereas the latter are suffered by all fetuses to some degree and are a normal part of labour the fetus is normally able to cope with. Detection of fetuses not coping with periodic deficiencies is the central goal of CTG analysis in this context.

In low oxygen scenarios cells begin respiring anaerobically, which both generates metabolic products and rapidly uses existing glucose supplies. Sustained anaerobic respiration will rapidly deplete glycogen supplies, causing energy dependent mechanisms, the sodium and potassium pumps, to fail. Without the charge and concentration gradients maintained by these, cell depolarisation occurs, allowing sodium and calcium to enter, and finally due to osmotic pressure enough water to cause cell lysis [13]. Thus of potentially greater importance than the supply of oxygen is the supply of glucose.

To prevent this from occurring the fetus will redistribute blood flow to the core organs and the fetal brain by peripheral vasoconstriction and shunting a larger portion of the ventricular output towards the brain. The heart rate may increase to maintain cardio-vascular output until the oxygen supply is restored allowing waste products to be removed and glycogen and oxygen supplies to be replenished. The increase in the acidity of the blood and systemic accumulation of waste products in the fetus is termed acidemia (metabolic and respiratory for lactate and CO_2 respectively).

Should the insults continue for a prolonged period of time, or the fetus fail to recover sufficiently between insults, glycogen levels may fall too low to sustain myocardial function, causing a drop in the mean arterial pressure (MAP) and a reduced glucose supply to the brain. In lamb studies hypotension, and the resultant reduction in cerebral blood flow, were identified as the critical factors in causing neuronal injury, not the accumulation of metabolic and respiratory products [79]. Every fetus will respond differently, depending on age, sex, and temperature [13], and hence there is not a well defined cut off point in any physiological measurement beyond which permanent damage will occur.

In the event that the fetuses life is deemed to be at risk a 'crash' CS will be performed. The intention of this is to deliver the baby as quickly as possible; the target time between decision and delivery is set to be < 30 minutes. [131]. In circumstances where there are non-reassuring signs and the labour has progressed far enough the process may be hastened with the use of forceps to apply additional tension in conjunction with contractions, or with a vacuum attachment (ventouse) where suction provides purchase on the fetal head, and again provides additional force to expel the fetus.

2.3.1 Outcome assessment

It is difficult to know immediately whether neurological damage has occurred in the newborn fetus and what, if any, impact it will have on its development. The only fully comprehensive answer to this question can be found through long term follow up on the newborn's developmental progress, a costly and often unnecessary process which will be carried out only when damage is suspected. Even with this information, assigning outcomes as intrapartum related or gestational is difficult. In practice simpler measures are used to assess the fetal condition upon birth to quantify the presence of severe birth hypoxia and to identify the need for further treatment. Both methods, the Apgar test and pH umbilical artery sampling are typically performed at birth.

Apgar scoring and clinical assessment

The Apgar score ranges from 1-10, and is the sum of 5 factors; Appearance (skin tone), Pulse, Grimace (the reflex to stimulation), Activity and Respiration, each scored from 0 (poor) to 2 (healthy). It is performed immediately after birth and at the 5 and 10 minute marks. Though it is in part a subjective test, originally designed to assess the need for and success of resuscitation in newborns, it is highly predictive of long term fetal outcomes such as low intelligence quotient (IQ), CP, and death [121, 81, 125].

pH & base deficit

Blood tests will be performed after delivery, and may be performed during labour on samples taken from the fetal scalp or samples from the umbilical cord after birth. Three tests are available, pH, lactate, and base deficit (BD), with each indicating the occurrence of different events.

The first, pH, is a measure of the presence of metabolites (lactic acid) and respiratory products (carbon dioxide) dissolved in the blood, and a test will typically require 18 minutes to perform [108]. The detection of acidemia is not 100% sensitive to adverse fetal condition. There is a greatly increased risk of encephalopathy at low umbilical arterial pH (< 7.0, relative risk 16.86:1) but 75% of all incidents occurred in cases with a greater pH [205], indicating that other markers must be used to quantify fetal harm at birth. Animal studies have also demonstrated that using pH as the sole outcome measure can be misleading; in lamb models with induced severe asphyxia, pH and BD recovered to normal values within 6 hours of the hypoxic insult while other physiological measures slowly returned towards baseline values over the 72 hours until the model was sacrificed [87].

It is established that the mechanisms by which the pH decreases are as important as the absolute value reached and though the likelihood of damage is much greater at lower pH values, it is not guaranteed. There is an open debate on whether a cut-off value should be used to diagnose birth asphyxia, and if so what it should be [205, 74] - a healthy value is normally considered to be > 7.26.

Lactate samples indicate the presence of lactic acid, the product of anaerobic respiration, and can be processed in as little time as 10 minutes. The test can be performed at the bedside, and with a greater success rate than pH tests due to the reduced volume of blood required ($5\,\mu l$) [46]. This is advantageous when blood tests are performed to confirm the results of CTG assessments, or to aid decisions in cases exhibiting suspicious patterns. Lactate sampling has shown greater sensitivity and specificity to low 5 minute Apgar scores and severe hypoxic ischemic encephalopathy (HIE) than pH [103] and in lamb models lactate levels in severely asphyxiated lambs remained elevated much longer than pH levels [87]. It is currently recommended as the primary method

of blood sampling for fetal condition during labour in the United Kingdom (UK) [148].

Finally base deficit measures the increase in concentration of the blood buffer base produced by the liver in response to metabolic acidosis. It is calculated from the bicarbonate concentration and the pH of a sample of blood. Intrapartum asphyxia is sometimes defined in terms of both the base deficit and the pH, required to demonstrate metabolic respiration [158], using a cut-off point of $< 12mmol/L$. Without any respiratory difficulty this value will range from $-2 < BD < 2mmol/L$.

2.3.2 Epidemiology and healthcare impact

Intrapartum asphyxia leading to neonatal-encephalopathy can cause cerebral palsy or in extreme cases death. Estimates of incidence rates of intrapartum related mortality range from 0.5 to 0.9 per 1000 births each year [132, 25], and intrapartum related disordered brain function occured in 1.6 per 1000 births [132]. Between 9-20% of all cerebral palsy is caused by intrapartum events [139, 13, 74], the remainder being caused during gestation, with an combined incidence rate of between 1.5-2.5 per 1000 births. A retrospective analysis of care in 28,486 deliveries in Sweden found that of the 161 incidents of metabolic acidosis, 49.1% were preventable, with CTG misinterpretation occurring in 19.9% of all incidents [89].

Between April 1^{st} 2000 and 31^{st} March 2010 the NHS spent £3.1 billion on litigation costs and payments on 5,087 claims made, of which 300 were made on CTG interpretations and 542 on cerebral palsy, totalling £1.7 billion in claims, though it was noted that issues of CTG interpretation arose in cases marked in other categories [129]. The same report noted that 41% of intrapartum-related CP was avoidable with proper care.

This increase in litigation and the presence of preventable cases has been a contributing factor in an increase in CS rates worldwide, rising from 20.8% in 1997 to 31.8% in 2007 in America, and 27.7% across all industrialised countries [42]. In the UK individually this rate has risen from 12% in 1990 to 28% in 2008 [128].

2.4 The cardiotocogram

Decision making during labour requires some measure of the fetal status, of which the choices are limited due to the inherent inaccessibility of the fetus and considerations for the mother's comfort. Physiological measurements are complemented with clinical information and physiological measurements from the mother and the history of treatments that have been administered during the birth such as Oxytocin or any anaesthetics.

The CTG is a continuous measurement of two signals; the FHR taken as the beat-to-beat interval, and the UA, used to identify contractions. It is available in almost all hospitals in the developed world, and though not routinely used continuously for all labour it is used for births defined as high risk (such as an underweight fetus or maternal illness), or when a non-reassuring signal has been identified in intermittent monitoring. The continuous signal is believed to be advantageous as it allows for identification of trends in the heart rate and provides access to historic patterns which can be reviewed to aid diagnosis.

A sample 20 minute digitised tracing of CTG data is shown in Figure 2.2; in clinical practice this output will be continuously printed to paper at a rate of 1-3cm/minute. The CTG is used routinely in antepartum labour to identify gestational diseases [40] and upon admission for delivery. In these situations, the monitors remain in place for

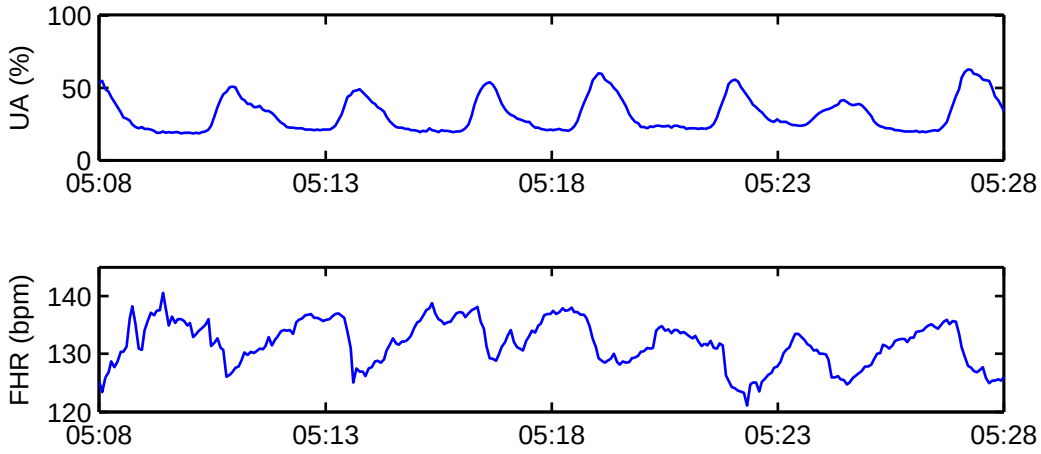


Figure 2.2: A 20-minute sample of a CTG recording. The top axis shows the uterine activity (UA), and the bottom the fetal heart rate (FHR).

30 minutes, or until sample quality criteria have been met and the sample has been deemed as reassuring.

The CTG signal can be produced using external or invasive monitors; the former being two belts strapped around the maternal abdomen, the first being a pressure transducer and the second an ultrasound probe. The pressure transducer, called a tocodynamometer estimates the pressure required to flatten an area of skin on the mother's abdomen using a circular pressure transducer held against the skin with an elasticated belt. External pressure measurement has two disadvantages - the device can be uncomfortable for the mother to wear and the measurement provided is noisy and relative. Practically it is used only to indicate the beginning and endings of contractions and trends in their relative strength.

The second measurement, the FHR, is found externally using an ultrasound probe which must be focused on the fetal heart. The shift in the frequency of the returned signal is used to estimate the flow of blood through the fetal heart. From this au-

tocorrelation methods are used to estimate the heart rate. Again there are a couple of disadvantages to external monitoring - the rate identification algorithm is vulnerable to doubling the estimated heart rate should that frequency become the dominant component [123] and a second belt must be worn to hold the monitor in place, which must remain focused on the fetal heart. Signal quality is low, and affected by maternal movements which may move the focal point away from the fetal heart, or worse onto a maternal blood vessel.

Invasive monitoring involves insertion of two probes through the cervix and into the uterus; a pressure transducer capable of measuring absolute pressures and a screw type electrode attached to the scalp to pick up the fetal electrocardiogram (ECG). On the downside, invasive monitoring can only be conducted after rupture of the amniotic membrane and the process is by its nature an invasive process. The screw electrode can become detached but the signal quality is significantly better than external methods in both stage 1 labour (7.2% of segments missing against 1.0% for internal methods) and stage 2 labour (19.0% of segments missing against 4.0%) [11].

The CTG is then assessed for signs of fetal asphyxia by identifying morphological patterns. These patterns are defined in more detail in clinical guidelines, covered in section 2.4.2. Here their physiological causes are explored. The first features identified were:

- *Accelerations* - transient rises in the FHR.
- *Decelerations* - transient decreases in the FHR, normally in response to a contraction.
- *Baseline heart rate* - the stable component of the FHR which it will return to after a transient event.

- *Short term variability* - low frequency perturbations about the baseline to 5-10 bpm caused by the fetus' auto-regulatory system.
- *Contractions* - periods of uterine activity lasting 30-120s.

Accelerations can be caused by fetal movements, and may be absent in periods of fetal sleep, hence their absence is not taken to be of clinical relevance. Accelerations can be induced by acoustically stimulating the fetus, though the inability to induce them is not an indicator of acidosis, their presence is a reassuring sign [160]. Accelerations during intrapartum labour are not expected as the fetus will typically be minimising movements to minimise the impact of hypoxia.

Contractions will typically increase the uterine pressure by 40-80mmHg, high enough to compress the umbilical vein. Smaller contractions will not compress the higher pressure umbilical artery, leading to a drop in blood pressure as blood flows out without return. Some contractions will be of low enough intensity that neither vessel will be occluded. The fetal system will respond to a contraction by dropping the heart rate to maintain pressure and to reduce myocardial work, to increase slightly in response to the decreasing pH and to return once the flow has been restored. This drop in heart rate is observed as a deceleration on the CTG. The contraction response is mediated in part by chemoreceptors [199] although prolonged restrictions lead to maintenance of the deceleration due to myocardial hypoxia [52]. The contraction-induced deceleration may be accompanied by shoulders; as the umbilical vein has the lower intraluminal pressure of the two umbilical vessels this will close first, meaning that some blood volume is lost. The drop in pressure causes a temporary increase in heart rate, observed as the leading shoulder. As the contraction ends, arterial flow will be restored first due to its higher pressure, causing another drop in volume and the second shoulder to

maintain pressure [70].

Westgate et al [199] used fetal lamb models to study the shape of the deceleration in response to progressive hypoxia by repeatedly occluding the cord at regular intervals to simulate contractions. Term lambs were fitted with monitoring equipment and a cuff fitted to occlude the umbilical cord, which was inflated and deflated in occluded to non-occluded ratios of 1:5 minutes and 1:2.5 minutes. The 1:2.5 ratio group eventually became hypotensive, with severe acidemia. The 1:5 group compensated well, with slight dips in arterial pH and base excess (BE) which were quickly restored, demonstrating the fetus' ability to survive brief repeated occlusions. The onset slopes of the observed decelerations were proportional to the magnitude of the measured chemoreceptor response, which adapts to the severity of fetal stress; developing acidosis being linked to a steeper onset slope. The magnitude of the deceleration was related to the severity of the hypoxic insult. Overshoot, a shoulder occurring only after deceleration, was also related to developing acidosis and has been further demonstrated in human CTG recordings [60].

Other aspects of the deceleration shape have not yet been explained. de Haan et al. reported on the impact of repeated occlusions on FHR recovery times and the time taken for the heart rate to return to the baseline during a deceleration [41]. The impact of 1:2.5 minute and 2:5 minute occlusion ratios were tested on FHR features, and recorded with respect to the umbilical pH, showing little difference until severe ($pH <$ 6.9) acidemia. The effect was more pronounced on the MAP, demonstrating that a careful choice of properties is important and that their discriminatory information is not uniform.

Short term variability (STV) is defined as the low amplitude wander of the FHR about the baseline with a frequency of less than 2 cycles per minute, caused by the

sympathetic/parasympathetic nervous system. In animal models heart rate variability was highly correlated with asphyxia, and depressed variability was observed following periods of asphyxia [18].

Loss of FHR variability has long been thought of as a strong indicator of neurological damage, yet studies in lamb models found reduced variability in 2/3 of severely acidotic cases, with the remaining 1/3 showing increased variability [198]. Also, not all loss of variability is cause for concern. Heart rate variability will exhibit periodic decreases in periods of fetal sleep, or may be depressed as a result of drugs administered to the mother.

Under conditions of severe fetal distress the responses of the control systems begin to fail to function at all. In these incidents the para-sympathetic system may completely collapse resulting in the loss variability and a distinctive sine-wave pattern in the baseline FHR [52]. However by this point damage is likely to be permanent or unpreventable, making identification of the pattern useless in preventing harm.

2.4.1 Criticism of the CTG

The usefulness of CTG recordings in the identification of fetal risk is a subject for debate [133], and the effectiveness of Electronic Fetal Monitoring (EFM) in comparison to the standard method of intermittent auscultation as a methodology of care for the patient has shown both benefits and costs. This first came to light in the Plymouth trial [200]. A meta-review including this trial on the impact of continuous CTG on labour outcome showed no effect on perinatal death rates, but a reduction in neonatal events, primarily a 50% decrease in neonatal seizures [168]. The use of CTG versus intermittent auscultation (IA) was recently reviewed by Alfirevic et al [3] in another

meta-analysis of 12 trials totalling 37,000 cases - some in common with Steer. This review showed no difference in the perinatal death rate, and an increase in intervention rates due to suspected fetal distress.

In review Thacker et al [173] showed a reduction in neonatal seizures when using EFM, but noted that this must be weighted against the increase in assisted deliveries. Low et al [113] showed, among others, that FHR variables and their patterns can be used as predictors, however they do suggest that supplementary tests are necessary for clinically relevant applications, as did Tasnim et al [172].

CTG has been the centre of frequent malpractice investigations centred around its interpretation due to its high sensitivity. In 71% of intrapartum asphyxia cases seeking compensation in Sweden between 1990 and 2005 signs of fetal asphyxia which could have been acted upon were found in CTG traces [14]. A second Swedish study on labours between 1994 and 2004 found evidence of CTG misinterpretation in 19.9% of cases (out of 161 in total) that resulted in metabolic acidosis. An enquiry into intrapartum deaths in the UK showed failure to act on abnormal CTG in 72% (18 total) of cases [127].

In America, part of the rise in litigation is attributable to screening programs in low-risk healthy births, such as admission CTG assessment. The sensitivity and specificity of the CTG were found to be 57% and 69% for both fetal death and CP, making it highly inappropriate as a screening programme due to the low incidence rates of each (0.5 and 2 per 1000 births) [77].

2.4.2 Guidelines for CTG interpretation

Guidelines have been developed by experienced clinicians based on current practices to bring objectivity into CTG interpretation and management. Traces are divided into separate categories depending on the medical attention that each should receive, with class boundaries being absolute values of identified morphological features.

The first of these was devised in 1958 [84], more modern examples being those proposed by the American College of Obstetricians and Gynecologists (ACOG) [5], the National Institute for Health and Care Excellence (NICE) [130, 132], the International Federation of Gynecology and Obstetrics (FIGO) - first established in 1987, though an updated version is currently being drafted [51, 7], and the National Institute of Child Health and Human Development (NICHD) [152, 115]. Each guideline defines a single class which requires intervention, a second which is reassuring of normality, and 1-3 intermediate or suspicious classes.

Each of the commonly observed CTG features described in section 2.4 has been formally defined, and in the case of the deceleration multiple shapes are described. As an example FIGO defines the different deceleration shapes that may be observed as [51]:

- *Early* - Shallow and short lasting. Indicators of fetal head compression.
- *Variable* - The most common, with a quick drop and recovery of FHR. STV is retained during the deceleration, and there is a strong relationship to the duration and intensity of the corresponding contraction. Possess a "V" shape, with a "U" being an indicator of potential complications.
- *Late* - "U" shape or reduced STV, with a slow drop and recovery in FHR, and a significant delay after the start of the corresponding contraction.

There is not agreement between guidelines on the boundaries of each category, or the number of categories that are needed in practice. For example the NICE clinical guideline [132] defines a normal CTG as having a baseline between 110-160bpm, with STV $\geq$ 5bpm, and only early decelerations. The FIGO guidelines however define normal as having a baseline between 110-150bpm, STV between 5-25bpm, and no decelerations. Further information on the discrepancies between guidelines can be found in the literature [8].

Multiple guidelines are being used currently so discrepancies can be expected between institutions, but when assessed individually interpretations made following guidelines are also not consistent. A study by Blackwell et al. [17] using the most recent NICHD guidelines demonstrated agreement between three experts in 57.7% of segments giving a kappa value across the three classes of 0.45, and an intra-observer agreement between kappa of 0.74-1. Confusion was found only between neighbouring classes, i.e. no healthy class was identified as pathological, only as suspicious. Another study using a larger 5 class system, 5 experts, and majority voting found that each expert agreed with the majority in only 57% of traces [142]. A meta-review comparing these results with studies from 1978-2010 found similar results when using other guidelines [17]. To counteract this, electronic interpretation of FHR signals was suggested, and modern guidelines from the ACOG and NICHD have been designed with computerised interpretation in mind, as will be detailed in section 3.1.

As important as unifying interpretation is deciding whether the recommendations in these guidelines correlate with fetal outcome. A study on the correlations between the categories indicated by each aspect of the FIGO criteria and fetal acidosis (defined as scalp pH$\leq$ 7.20) showed a high sensitivity (95%) but low specificity (21%) [156] for the baseline and deceleration features. This high sensitivity and low specificity agreed

with findings from previous studies [163] which showed a much greater proportion of adverse records (89%) being classified as pathological than healthy records (52%).

2.5 Alternatives/adjuncts to the CTG

2.5.1 Intermittent auscultation

IA involves listening to the fetal heartbeat using a stethoscope or hand-held Doppler ultrasound held to the mother's abdomen. It is recommended for primary assessment of the fetus upon admission, at regular time intervals throughout the birth, and after a contraction for the duration of low-risk pregnancies. The infrequent sampling and lack of wired monitors causes minimal inconvenience to the mother, but neglects potential trends and brief hypoxic events, driving the development of continuous monitoring techniques for high-risk labours. The heart rate is estimated by medical staff, and is typically taken with the maternal heart to avoid their confusion. As such training and practice is required to effectively perform it.

2.5.2 STAN monitors

STAN (ST-ANalysis) monitors, produced by Neoventa Medical [126] compute properties of the ST interval of the fetal ECG waveform, using the same electrode as the invasive CTG. Changes in the T-wave amplitude are caused by myocardial hypoxia and anaerobic metabolism in the heart muscle [52], resulting in changes to the ST interval. This effect was noted as early as the 1950's, but it was believed to be too variable and unreliable to be of use [75].

Several studies have been conducted into the efficacy of STAN monitors in dis-

criminating fetal hypoxia demonstrated by its impact on clinical outcome measures. A retrospective analysis of 563 traces by Kwee et al. [105] showed the presence of ST-events in both normal and abnormal FHR traces, supporting theories that events can also be caused by a surge of stress hormone in response to mild and compensated hypoxia. Noren [135] investigated the impact of the introduction of STAN monitoring to a Swedish hospital. Reported was a decrease in metabolic acidosis from 0.72% of cases to 0.06% in conjunction with an increase of STAN usage from 26 to 69%, which was associated with a decrease in the time between an adverse pattern being detected and clinical intervention in the birth. This result must be taken with caution as no control group was studied to take into account changes in medical practice and the increase of operative delivery rates for fetal distress. A meta-analysis of controlled trials [12] did not support the reduction in metabolic acidosis rates, but showed a decrease in operative delivery rate for non-reassuring CTG traces.

2.5.3 Fetal pulse oximetry

Fetal pulse oximetry (FPO) is a newer technology that received FDA approval in 2000, and that has showed promise as the first real-time measure of hypoxia. FPO uses a disposable sensor positioned against the fetal scalp to measure oxygen saturation.

FPO has received poor uptake in the United States (US) where it was first introduced [186] which may be due to poor performance, price of the disposable monitor, and unreliable signal quality - only reliable 57% of the time during second stage labour and 64.7% in first stage [186]. FPO has been shown to have a positive effect on operative delivery rate due to non-reassuring CTG traces with a 50% reduction [57], however in this study the total number of operative deliveries remained the same, with an in-

crease in operative deliveries for dystocia (a difficult fetal birth due to presentation or otherwise) in the test group. A meta-analysis of 6 published trials covering 7654 births by East et al. [45] showed no change in fetal outcome when using CTG + FPO against CTG alone. Whilst FPO correlates well with the systemic oxygen concentration, as explained in Section 2.3, a low systemic oxygen concentration may be due to redirection of oxygenated blood to the central organs and brain; therefore it cannot be taken as a direct measurement of cerebral perfusion and condition.

2.5.4 External electronic monitoring

Using a series of 5 electrodes placed in a circle on the maternal abdomen the Monica AN24 Abdominal monitor is able to measure both the fetal electrocardiogram (ECG) and the maternal ECG simultaneously from the electrical activity of the two hearts. Additionally the uterine activity is derived from the electrohysterogram (EHG) of the uterine muscles [120]. A small study of 74 patients showed an increased agreement with the gold standard (invasive monitoring) in number of contractions successfully detected per epoch and lower root mean square (RMS) error of the FHR estimate when compared to external CTG monitors [83, 34].

Of particular note is the performance in second stage labour, with a 71.9% agreement with the scalp electrode and a mean RMS error of 7.9 bpm against values of 61.7% and 16.1 bpm using Doppler ultrasound. By the second stage the cohort had reduced in size to 42 subjects [150]. Recording the maternal heart rate removes artefacts on the fetal FHR graph due to detection of the maternal heart rate by plotting both simultaneously. These artefacts were found to occur in 2.9% of Doppler recordings in the second stage of labour [150]. The device's uptake has been limited, but it

may find use in specific scenarios due to its increased effectiveness when monitoring cases with high body mass index [150] where conventional CTG may struggle and in offering greater comfort and mobility to the mother by removing the belt required to keep the ultrasound probe aligned and the TOCO sensor pressed against the skin.

The AN24 is not the sole system conducting external ECG monitoring. The Meridian system from Mindchild Medical (N.Andover, MA) performs a similar function without the requirement of a separate preparation to ensure good skin contact and has been approved by the FDA [1]. The system is also concerned with the accurate extraction of the fetal STAN signal [119]. A trial of 32 women measured using both external fetal ECG (fECG) and internal fECG acquired using a fetal scalp electrode (FSE) found that the RMS error between the FHR calculated from external and internal measurements was 0.36 bpm. It was also found that 89.9% of 10 second segments measured provided data of sufficient quality for ST calculation, compared with 91.2% usable segments in the FSE [33].

2.5.5 Vibro-acoustic stimulation

Vibro-acoustic stimulation can be used to agitate the fetus, inducing an increased heart rate due to fetal activity [197]. A response to stimulation is a positive indicator, but no responsiveness to stimulation is not an indicator of distress. A review of the method in 2005 [47] reported that no randomised controlled trials had been undertaken to investigate the efficacy of vibro-acoustic stimulation as an adjunct to investigation of a non-reassuring CTG, but that it is still used in practice.

2.6 Summary

There is a need to improve the accuracy of IH identification, but also the timeliness of diagnoses given the speed with which permanent injury can occur and the time taken for intervention. The CTG has not been demonstrated to reduce any primary outcome measures in practice. In the two most significant large scale meta-analyses it was found to reduce neonatal seizures by 50% in one, and to increase operative delivery for fetal distress in the other. However CTG monitoring is widely used and guidelines developed for its interpretation have demonstrated high sensitivity to fetal pH. This high sensitivity is supported by experimental work which describe correlations between multiple features of the fetal FHR and hypoxic insults.

Of the proposed alternatives to support or replace the CTG, only the STAN monitor has demonstrated a reduction in operative deliveries in a multi-centre trial. However this requires invasive monitoring. The external fECG technologies being developed in multiple independent locations are promising, with comparable data quality to the internal measurements. FPO has not been well received due to difficulties in inserting and positioning the probe. No modalities which measure the primary factor influencing damage due to IH, MAP, have been proposed.

The low specificity, and lack of early detection mean the current CTG strategy is insufficient. Research lately has focused on machine learning techniques which may be able to identify pathological patterns more successfully than the current rule based guidelines, and take advantage of large datasets which are being currently being collected.

Chapter 3

A Review of Classification Methods

An essential part of clinical systems to aid decision making from CTG traces are the methods used to identify at risk fetuses. In addition to the development of clinical guidelines significant effort has been put into evaluating the performance of machine learning techniques for CTG classification.

This Chapter reviews the techniques used to date over the entire process of classification. This includes a brief overview of feature extraction in Section 3.1, and selection in Section 3.1.1. Classifiers previously applied to identification in CTG traces are described in Section 3.2, alongside research into system identification processes. These methods are summarized and compared in Section 3.6 along with a commentary on classifier comparison. Finally commercial implementations of rule- and machine learning-based systems are described in Section 3.3.

3.1 Feature extraction

Though individual data points may be used as classifier inputs it is practical to reduce the dimensionality of time series data calculating features which represent the series. Reducing the number of input parameters helps to prevent over fitting. As a general rule it is recommended that classifiers have at least 10 samples per parameter, though this can be pushed to 3 samples for each linear parameter and 10 for each non-linear parameter.

To compensate for the non-stationary time-series features are typically calculated over a short window of 10-20 minutes within which the signal be assumed to be stationary. Several reviews in the literature include comprehensive lists of features [159, 30, 55], and the feature sets used in this Thesis are listed in appendices A and B. These features be broken up into several families:

Morphological

Computerized CTG analysis began with the introduction of rule based identification of morphological features, using the methods described in clinical guidelines. These include accelerations, contractions, decelerations, and other properties used to clinically describe the CTG.

These normally cannot be calculated automatically as the methods described in guidelines require the identification of several markers from which other features are measured, such as the FHR baseline and the start and end points of transient events.

Algorithms based on signal processing have been developed for the identification of contractions [62], decelerations and their associated overshoots and accelerations [60], and have achieved performances matching the performance of an experienced clinician. Machine learning techniques such as the Artificial Neural Network (ANN) have also been used to identify and to classify features [9, 107, 187, 98]. Automatic methods are advantageous as they are repeatable and consistent, and allow additional properties of CTG events to be calculated such as the onset and recovery gradients of decelerations, which are used extensively in the OCFMT featureset [61, 203] and also by Agrawal [2].

Time domain features

These features include Long-term variation (LTV), which has already been clinically defined but was difficult to use in practice and the STV, which can be defined and calculated in multiple different ways [26]. Statistical properties of the CTG signals, such as the mean of the FHR and skewness and kurtosis of histograms of FHR values are included in this category [30] as well as measures such as the signal stability index

(SSI) which is calculated from the peak value of the kernel density estimation of the FHR signal.

The Oxford System (OxSys) utilises properties of an STV tracker, which is defined as the mean difference between FHR values in a 1 minute window. The median of this value gives the segment STV, with other properties measuring the stability of this signal.

Frequency domain features and wavelets

Several low frequency bands of the CTG spectrum have been identified which relate to physiological events. These bands are: maternal and fetal heart rate (0.75-1.5Hz), maternal respiration (0.15-0.4Hz) and the responsive nervous system (0.04-0.15Hz) [31]. A very low frequency component has also been observed (< 0.04Hz) influenced by hormonal changes and thermoregulation. Other bands have been proposed which identify the range 0.15-0.5Hz as induced by fetal movement and maternal respiration [159]. These features have been compared directly with outcome measurements [184] and have demonstrated sensitivity to fetal distress in intra-partum labour [31].

Wavelets have been suggested for multiple medical applications due to their ability to localise frequency information in time, and direct and statistical properties of discrete wavelet coefficients have been used to successfully classify CTG data [65].

Non-linear features

The largest group of features is the non-linear features which have been demonstrated as better discriminators than linear features alone [164]. These feature heavily in automatic feature selection work conducted by Fulcher et al. [55] which have been incorporated in the OxSys featureset. These can be further split into sub-groupings

which are provided with examples:

- *Fractal dimension measurements*: one measure of complexity, which establishes how detail in the signal changes with scale and can be estimated by multiple methods such as the Higuchi dimension [99] or box-counting dimension.
- *Compression measures*: an estimate of both complexity and information content, these compare how simply signals can be represented, such as Lempel-Ziv complexity [30], or compression size with commercial algorithms [37].
- *Information measures*: again measurements of both complexity and information content, signal entropy measures have been developed for medical diagnosis by estimating the predictability of time-series. The fundamental measures are Sample entropy and approximate entropy [144, 50].
- *Adult heart rate variability measures*: these features have been developed and tested for the detection of diseases in adult hearts. Examples include NN50, a measurement of the number of beats differing by more than 50 ms in a window [30] and parameters of the phase rectified signal averaging curves [58].
- *Classifier outputs*: the output of a previous classifier such as the Neural Index [116], or the class assigned by a guideline such as the automatic implementation of the ACOG index. The output of a sinusoidal pattern detector [149] and a Support Vector Machine (SVM) classifier [203] form part of the of the OxSys featureset.
- *System identification parameters*: The magnitude and delay of impulse response functions (IRF) have been used by Warrick et al. [190].
- *Emperical Mode Decomposition signal properties*: Krupa et al. used the standard deviation of the intrinsic mode functions of the FHR as features [104, 64].

Care must also be taken especially when comparing non-linear features found using invasive methods with non-invasive methods, since the feature values have been shown to significantly vary between the two methods [72].

STAN features

These include any properties of the fetal ST waveform collected using an ECG. Since these features can only be included if the CTG is being collected using a scalp electrode their use is typically restricted to commercial systems such as the SisPorto system [35].

3.1.1 Feature selection

Given the large number of features available a subset must be selected to fulfil their purpose in reducing the complexity and dimensionality of classifier inputs. Features can be selected based upon individual performance or correlation with the target outcome [55] or a subset chosen which maximises retained information and minimises correlation between features [177]. For example the unsupervised relative reduct (URR) algorithm used by Inibari et al. iteratively removes features based on their relative dependencies [80], a measure of feature correlation.

The most widely used method, Principal Component Analysis (PCA) transforms a set of features into a new set of linearly uncorrelated features ordered such that the first new feature, or principal component has the largest possible variance. The second component has the largest possible variance in the space orthogonal to the first and so on. The data can be fully reconstructed by selecting all principal components, but the number of features can be reduced at a cost of losing some of the variance by selecting only the first N most significant components. PCA is the simplest method to use,

but has several shortcomings. The method assumes Gaussian distributions of data in each dimension, though this can be compensated for by using a kernel method prior to applying the transform.

The second major method for feature selection is the Genetic Algorithm (GA). In the GA a population of candidate sets is established, each candidate being an $1 \times N$ vector of 0's and 1's where N is the number of features being selected from, with the values indicating if it has been selected or not. Each candidate is evaluated using a fitness function which can be a direct classifier output [203] or a more complicated regularising function [137, 204]. Using this fitness function the weakest candidates (low fitness) are killed off, best performers, termed elites (highest fitness), are retained, and a next generation is found by breeding surviving members (combining parts of their genome vector) or by introducing mutations (randomly changing parts of their genome).

This process is continued until some stopping criteria have been met, giving a superior set of genes encoding which features should be selected. A similar approach termed Genetic Programming (GP) has also been used to create strong features from weaker ones by combining them using a grammar of mathematical operations [63].

Spilka et al. used the Relevance in Estimating Features (RELIEF) algorithm [166] which stochastically ranked features based on cumulative distances from each class. The Least Absolute Shrinkage and Selection Operator (LASSO) uses a regularisation parameter to minimise the sum of weights in a regression model fitting the inputs to the output. Small values for the regularisation parameter result in many weights being 0, effectively removing those parameters from consideration [204, 82].

Spilka et al. have also made use of the Group of Adaptive Models Evolution neural network (GAME-NN) which automatically discovers optimal features. This method

was combined with Information Gain (IG) and PCA using majority voting to form an optimal feature set [164].

RELIEF or LASSO retain features in their original form, whilst PCA and GP create new features, either through linear combinations or more complex transformations of the original set. In cases where it is useful to know which original features were used, or where it is useful to provide some insight into the impact of each feature selection methods which retain the original features are recommended. Alternatively, if the performance is the highest priority then non-feature-preserving methods may be recommended as the information is often arranged in a more convenient and compact manner.

Research on feature selection raises questions as to whether these studies should be limited to features from one category [66, 164] or whether features should be ranked individually [204, 30], and if so what measure should be used for their evaluation. This is particularly relevant when investigating GA and ensemble methods. For example when using linear classifiers as a fitness score for feature set optimisation the discovered set is only the maximally linearly separable set. Conversely methods featuring highly non-linear classifiers such as a Multi-layer Perceptron (MLP) are likely to be highly biased to performance on that single classifier, therefore the discovered feature set may not be assumed to be the optimal set for general use.

Feature selection should therefore be treated in a similar manner to parameter selection, and performed using only training and validation data. This requirement is not true of all results in the literature [137], which cannot then be compared to other non-biased selection methods.

3.2 Classification of the CTG

Once a suitable group of features has been selected these must be mapped onto class decisions based upon training examples already seen, the role of a classifier. The output depends on the type of classifier used and the number of target classes. Identification of IH can be formulated as a binary problem with cases classified as either positive or negative based on a threshold of the chosen outcome variable.

Alternatively this may be treated as an ordinal multi-class system typically following the normal / suspicious / pathological or the category I/II/III boundaries set up in either the FIGO or ACOG guidelines or according to clinical assessment [134, 35]. The training procedure and model parameters are briefly described for each classifier observed in the literature below.

3.2.1 Adaboost

The Adaboost algorithm developed by Freund and Schapire [54] combines multiple weak classifiers to form a single strong classifier. A weak classifier is defined as one whose performance is only slightly better than random guesses. Commonly used examples of weak classifiers are linear discriminators and decision stumps.

Intuitively the algorithm forces each successive classifier to focus on the cases which were misclassified by the previous classifiers by increasing their contributions to the error term in successive iterations. This means each individually weak classifier specialises in performing well on different subsets of the data, specifically subsets which are poorly classified by others.

Because of the emphasis given to misclassified results the method can be sensitive to inaccuracy in targets in the training data. This is especially relevant to CTG

classification, as outcome measurements are either subjective (clinical opinion, Apgar score) or are subject to variation between individuals. To counteract noise in the target variable the RobustBoost algorithm, also developed by Freund [53], can be used.

Spilka et al. [166] used boosted nearest mean classifiers to help evaluate the performance of feature selection methods which identify and rank the importance of individual CTG features. The decision to use Adaboost is appropriate here, as a highly non-linear, complex classifier would add to the already complex experimental method, which used non-linear CTG features, Synthetic Minority Oversampling TEchnique (SMOTE), and two rounds of cross validation to tune parameters and to evaluate generalised performance. For the same reason the RELIEF algorithm was used to estimate feature performance in a low dimensional space.

3.2.2 ANN

The ANN is composed of a network of nodes, with each node acting like a neuron; an activation threshold must be reached for the node to 'fire', producing an output. The output of an individual is a weighted combination of the nodes in the previous layer, passed through a non-linear transform function.

The shape of the network must be chosen to fit the problem. There are no restrictions on the shape, number of connections, or size of a network other than those imposed by the amount of available training data and the cost of computation.

Maeda et al. [117] were one of the first groups use an ANN to classify traces as at risk, suspicious or normal. They used clinically identified parameters taken at three consecutive intervals. 20 cases of 50 minutes were used to extract 800 data vectors and train the ANN. The selected network consisted of 30 input nodes, 40 nodes in a

single hidden layer, and 3 output nodes. The network was tested on a set of 26 unseen cases classifying 86% correctly. The use of an intermediate class helped in this respect, with the majority of signals being classed here. This early study provides a positive look at the potential of these systems despite the small dataset and large number of parameters in the model.

Jezewski et al. [88] extend both the number of features used to 17 and the number of individual cases to 189 patients, with a focus on the correlation between outcome measures and CTG patterns. These outcome measures including birth weight, Apgar score, umbilical arterial pH and BE, and clinician assessment. A high sensitivity for distress was demonstrated by using a committee formed of networks trained to identify each outcome measure and classifying a case as abnormal if one networks classified it as abnormal.

Improvements to feature selection and network weight initialisation have been made by combining the ANN with Grammatical Evolution (GE) algorithms and GA [178, 63]. The GE was used to select and small set of features made from combinations of an initial set of 19 features, which were taken as the inputs to a 2 layer MLP with between 5-10 hidden nodes. The weights of the nodes were initialised using a GA, and then trained using a non-linear gradient descent algorithm. In both the cases of the GA and the GE the ANN output classification performance was used as a fitness function.

The method outperformed Linear and Quadratic and K-nearest neighbour (K-NN) classifiers, using PCA for dimensionality reduction. However the impact of the GA initialisation and GE feature selection were not observed independently, so improvements cannot be attributed to either method. Selecting for maximal accuracy may detriment performance on other measures, and may only be used on a balanced dataset. Addi-

tionally the use of SMOTE, to compensate for the low occurrence rate of adverse data must be noted. Increasing the oversampling ratio increased the model accuracy and geometric mean.

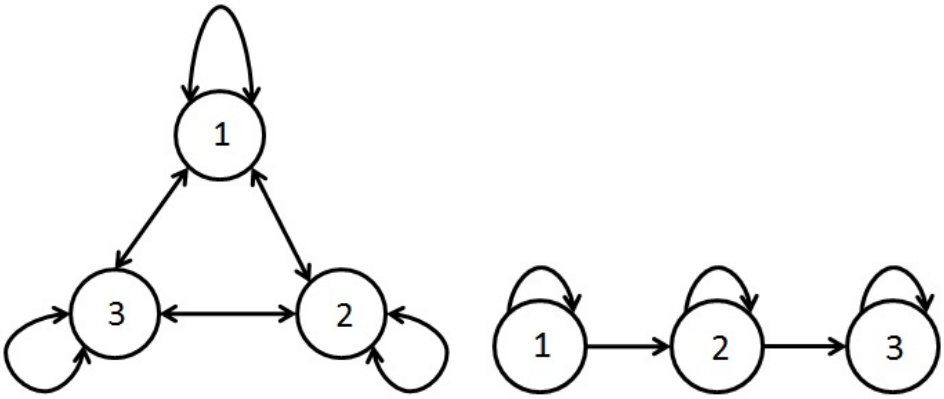
Georgieva et al. [61] used an ANN committee (ANNC) to classify FHR recordings on a database of 7568 cases, allowing a more definite abnormal pH of <7.1 to be considered for the positive class. EVEREST (EVEnt Rate ESTimation) plots demonstrated that the ANNC was a better estimator of low pH than clinical symptoms and a single network.

Many variations of the ANN exist such as the Radial Basis Function (RBF) network used by Jezewski et al. [88] which replace the activation functions of each neuron with an RBF function. The Adaptive Neuro-Fuzzy Inference System (ANFIS) combines fuzzy logic with multiple ANN's in the work by Ocak and Ertunc [138]. The Probabilistic neural network (PNN) is a type of feed-forward network which replaces the hidden node layer with a pattern layer, each node in the pattern layer representing a training sample.

3.2.3 HMM

The Hidden Markov Model (HMM) has been used extensively in the fields of automatic speech recognition [93] and gene identification [102]. The structure of the HMM is described by states, the probability of transitions between states, and the distribution of output probabilities in each state. An example of a set of three states and their associated transitions for two model types is shown in Figure 3.1.

Georgoulas et al. [69] applied the HMM to the classification of the CTG using a database of 36 recordings, each of 20 minutes in length. Features were extracted



(a) The ergodic, or fully connected model

(b) The left-right, or Bakis model with no state skips

Figure 3.1: Two sample configurations of a three state HMM, adapted from Rabiner [147].

using a 5 minute window, overlapping at 1 minute intervals to give 16 time points. 12 features were used, including frequency and time domain features. Two HMM (one each for the positive and negative case) were trained using the segmental K-means algorithm using 4-fold cross validation with test cases assigned to the class of the HMM that most likely generated them. Each model was a left-right model with no state skips allowed. A maximum accuracy of 83% was attained using a model with 7 possible states. The size of the database is small relative to the number of parameters, and the HMM performance was not compared with other methods, though the method was successful in demonstrating separability of time-series from the two classes. Since this model was not compared against single time step models the time-series cannot be proven to contain any information additional to that captured in a single time step.

3.2.4 RF

The Random Forest (RF) defined by Breiman [19] consists of a collection of decision trees trained on bagged samples of the complete dataset. Decisions are made using a majority vote of the individual trees.

Tomas et al. [177] compared RFs with a single decision tree and discriminant analysis using a heuristically selected subset of 7 features. This methodology ignores the innate feature importance investigation which can be conducted using the RF, by investigating how the performance of each tree was affected by the inclusion of each feature. In this instant RF outperformed the other methodologies with 93.7% overall accuracy on a three class (normal / suspicious / pathological) problem.

Spilka et al. [165] used RF to classify records from the Czech Technical University and University Hospital in Brno (CTU-UHB) database. 5 times 2-fold cross validation was used to evaluate algorithm parameters, though the focus of the study was on the methods used to measure label accuracy, and to combine the assessments of multiple experts using latent class analysis rather than simple majority voting, which takes into account both biases and the performance of each individual judge.

3.2.5 SVM

SVMs have received attention as widely used and simply implemented classifiers. The SVM aims to find a hyperplane that maximises the distance between the plane and the boundary to each class. The support vectors are the samples used to define the position of the plane used in calculation - those which are closest to the decision boundary and hence the most difficult to classify.

Georgoulas et al. [67] used an SVM with an RBF kernel to classify cases at risk

of metabolic acidosis, defined by pH$\leq$ 7.1, with not at risk cases defined as having pH$\geq$ 7.2 using a combination of time and frequency domain features. The SVM outperformed a K-NN classifier, which is known to be inappropriate to classify unbalanced data such as those used in this study. Wavelet features were previously classified using an SVM in a second paper [66] but these results were not compared with the time and frequency-spectra features.

Warrick [190] used an SVM to classify features of impulse response functions found representing successive windows of CTG traces. The movement of IRF properties was tracked on a 2-D plane with the SVM decision boundary overlaid to study the progression of cases as labour progressed. Two features were selected to visualise the progressions, delay, and gain, which are described in Section 5.4. The demonstrated cases were found to begin in the healthy class, and move towards the adverse class as birth approached agreeing with the implication that cases enter labour in good condition but were compromised by events during intrapartum labour.

3.3 Commercially available systems

Some of the methods above have been incorporated into monitoring devices to include some decision support software. As noted in Section 3.1 many clinical features have been successfully identified, allowing rule based guidelines to be implemented. The most successful example of a decision support system is the System8000 [39] which has been trialled in a clinical setting and is currently used to identify difficulties in antepartum labour.

However identification of abnormality in the antepartum period is regarded as the simpler task; both fetus and mother exist in a steady state without the stresses of

birth, signal quality is high and the expected response is well understood, thus any abnormality identified can be treated as suspicious. Indeed there are many similarities between identification in the intrapartum and antepartum periods, hence many techniques are expected to be consistent with both. However, no intrapartum labour decision system is widely used, though clinical trials of some are being conducted. The major currently available automatic CTG analysis solutions are described below.

SisPorto system

The SisPorto3.5 system is a hybrid method to monitor both ante- and intrapartum labour, developed in the University of Porto [161] and produced commercially by Speculum, Lisbon, Portugal [162]. The system uses classic CTG inputs and has the capability also to include STAN events, along with software to display and automatically identify and to alert staff to medical events such as prolonged decelerations. The software uses a 5-class system of alerts to classify signals, all associated with quantifiable clinical definitions (tachycardia, repetitive decelerations) and cut-off values.

The system first identifies morphological parameters based on the FIGO guidelines for FHR interpretation with a high level of accuracy [9, 10]. Similar to guidelines, classification is based upon features crossing threshold values. The number and types of features are extended, including features such as reduced variability and prolonged decelerations. Alarms are given following recommended practice, no alert, non-reassuring, very non-reassuring, and pre terminal, the last two categories sounding alerts for immediate attention.

SisPorto has been tested also in conjunction with STAN monitoring. A retrospective study of 148 consecutive cases was conducted to investigate the efficacy of including ST events in detecting acidemia, defined as pH<7.05 [35], and was shown

to increase the specificity of the system. A retrospective study of 104 cases assessed whether access to the SisPorto 3.5 increased the accuracy and reproducibility of clinician's predictions of acidemia and 5-min Apgar scores [36] versus a control group of 100 cases without access to computerised monitoring. The small number of cases involved only 3 adverse incidents, meaning that none of the results could be considered statistically significant, but demonstrated slight improvements in both measures with inter-observer agreement (measured using intraclass correlation coefficient) on umbilical arterial pH increasing from 0.43 to 0.7, and the percentage of pH estimates accurate to within pH$\pm$0.10 was 70% in the test group and 46% in the control group.

A multicentre randomised controlled trial was established in 2010 among five UK hospitals to primarily measure pH and base deficit, with secondary measure defined as scores, HIE, neonatal unit admission and medical treatment decisions and FHR signal details. The study aimed to recruit 8133 women, and eventually recruited 7730 [136]. The study found an incidence rate of metabolic acidosis of 0.4% in the experimental subset and 0.58% in the control subset, and no differences in Caesarean rates or instrumental deliveries was noted. The incidence rates of the primary outcome were too rare, however for any results to be claimed as statistically significant.

Sonicaid FetalCare

The only CE marked system in Europe is the Sonicaid FetalCare [141], based on the System 8000, developed in Oxford [39]. First, morphological features are extracted from the signal. The FetalCare system then uses a number of criteria to measure normality and to minimise the duration of monitoring, called the Dawes/Redman criteria, which are based on absolute values of morphological features of the CTG. Unique to the system is a measure to quantify when enough data have been collected

to classify the fetal CTG as re-assuring. The system has been clinically validated and has found widespread use in hospitals. However this system is used only in diagnostic ante-partum tests, not to monitor intrapartum labour.

PeriCALM

PeriCALM is a perinatal decision support system developed by researchers at the University of Toronto and commercially under LMS Medical Systems, Ltd. before a merger with PeriGen in 2009. The software records CTG traces and adds automatic annotations to them indicating unusual events and clinical feature estimates and highlighting events. Patterns are identified in accordance with the NICHD guidelines to categorize traces, and an interface for the management of suspicious (Category II) traces has been developed [32] and is provided along with the system.

The output of the system was compared with the output of 5 clinical experts on 769 8-minute segments using the NICHD 5-tier system. The agreement of the computer system with the clinicians was not statistically different from the agreement between the clinicians themselves, who achieved weighted kappa scores between 0.48-0.68, the automatic system scoring 0.52 [142]. Further validation was performed in 2013 [196] when three experts scored the assessments using the PeriCALM software as correctly identified or not. Over 2049 segments in 100 tracing variability, baseline, and acceleration agreement with clinicians was >99%, and 97% for decelerations. This study comes with the initial bias that examiners were presented with the system's results and then asked to judge whether they agreed with them, creating an artificially high proportion of agreement.

PeriGen have focused on developing a system for medical implementation, to deliver objective assessments, with the intention to reduce litigation costs and simplify

healthcare delivery. As such the focus has been towards care delivery, including shoulder dystocia and oxytocin administration, and not comparisons with fetal outcome such as pH. The system has been implemented in hospitals across America [143].

INFANT system

The INFANT (Intelligent Fetal AssessmeNT) system was developed in Plymouth Hospitals Perinatal Research Group, and has now moved under K2 medical systems [95]. The system was granted a CE Mark in 2009. INFANT categorizes CTG traces into one of four categories, raising alarms in the higher two categories. The system is a development on the original proposal [97] which used neural networks and numerical algorithms to classify important features of the signal into statements, for example 'baseline = 140bpm' becomes 'baseline normal'. An inference engine with a knowledge base represented by a series of logical rules is used to display the information and its interpretation to the clinician, recommending action to be taken. The system was assessed retrospectively on a small sample of 50 patients. These were evaluated by 17 experts and the system on two occasions, separated by at least a month [76]. The system did not recommend intervention in patients delivered with cord arterial pH>7.15, and in 11 of 12 cases delivered with pH<7.05, agreeing with at least 15 experts in each case.

A multi-centre randomised controlled study of 46,000 women is now being conducted to investigate the effect of the INFANT system in supporting decision making in labour [20]. The primary outcome studied is mortality and significant morbidity, with length of hospital stay and health service utilisation as secondary measures. 7000 patients will be randomly selected for follow up information 2 years after the study to investigate long-term outcomes, an important measure. This will provide one of

the largest databases available, and the inclusion of follow up outcome will be very important in the assessment.

To date 47,154' patients information has been collected and the study was completed in August 2013, though the study has since been extended until February 2016 [180], and no results have yet been publicly reported.

Maeda system

The approach taken by the Maeda system [118] begins with the same method as the previous systems of using algorithms to determine morphological parameters. The system then introduces two additional components. The first is an FHR score determined by points assigned to non-reassuring features found using the morphological, detected in 5 minute intervals. The fetal distress index (FDI) is introduced; a cumulative variable which counts consecutive FHR scores. 1 point was added to the FDI for an FHR score of 10-19, 2 for 20+, and various amounts for certain patterns such as severe bradycardia. In a supporting paper the FDI score was shown to correlate with the 1-minute Apgar score at birth [116]. The second addition is an ANN used to classify signals objectively as normal, intermediate or pathological. The ANN output class probability was termed the Neural Index (NI), for which a cumulative sum was also provided, updated every 5 minutes. A positive NI was found to correlate with normal outcome, and cases observed with a highly negative NI had severe acidosis.

The system was tested clinically from 1994-2011 in the Seirei-Mikatahara Hospital [118], and was shown to significantly reduce the number of perinatal deaths, although the degree of significance is not reported. There was also no change in the incidence of low Apgar score, although uptake of the comprehensive systems was not reported or discussed.

3.4 SI Models of the CTG

Though not classifiers themselves it is useful to consider the potential of models of the CTG signal in feature extraction and classification.

3.4.1 Physiological models

Physiological models offer an advantage over the traditional black box approaches as their internal values have a physiological meaning, the behaviour of which can be validated against current knowledge.

Models require accurate parameter values and a knowledge of the interactions between model components. The physiology of the maternal-fetal circulatory system, explained in Chapter 2, Section 2.2 has been modelled using an electrical circuit analogue of cardiovascular hemodynamics [182]. In the electrical circuit model the resistance, inductance, and capacitance values represent vessel resistance to flow, flow inertia, and compliance - the elastic ability of the vessel to expand and contract - respectively.

This model was used to successfully replicate early deceleration signals, and a more developed version [181] used to demonstrate replication of late decelerations induced by hypoxaemia. Couto et al. [38] used a simpler circuit analogue to the maternal-fetal circulation to simulate oxygen delivery to the fetus [38]. The model accounts for delivery through the placental membrane, and the uterine pressure dependent resistance to blood flow through the umbilical vessels.

Models of human fetal physiology cannot be validated with experimental data, however data are available for animal models which have been used for partial validation of aspects of each of the mentioned methods.

To the authors knowledge no example of a comparison between the output of a

physiological model provided with a true UA input and the true FHR recorded has been found in the literature.

3.4.2 Data driven models

Data driven models attempt to replicate observed signals by creating models which replicate the observed output when given a particular input. The simplest of these models are discrete linear time invariant models - their parameters do not change over time. These can be split into two groups, transfer functions which represent the output at some linear combination of historic inputs and outputs; and state space systems which introduce an additional vector to represent the state of the system and hence the future dynamics [111].

Jongsma and Nijhuis used the parameters of an autoregressive-moving average (ARMA) model to identify fetal behavioural states [90]. The FHR signal was split into 3-minute segments and a model with 2 autoregressive and 1 moving average parameter was calculated. The parameters of this model were used as the features of a linear discriminant classifier, which agreed with a clinical observer in 85% of cases.

Warrick et al. created a three part model of the observed FHR signal during labour, which included a stable baseline output, an autoregressive (AR) model of FHR variability and an IRF model of UA-FHR interactions [190].

The discovered values of IRF coefficients in the model have been investigated for signs of differentiation between fetuses with and without high base deficit. This includes tracking the development of features throughout labour in conjunction with a decision surface created using an SVM. Selected positive cases were observed to progress from the negative region into the positive region of the decision surface as

labour approached completion [190].

A similar approach has been used to model interaction between the UA and FHR variability by creating a new set of input signals corresponding to the instantaneous power in three frequency bands [188]. Each case was split into 20 minute windows, and a state-space model used to represent the interactions between the UA signal and each of these three frequency bands. The parameters of the IRF discovered in each window were then recorded. Cases with metabolic acidemia showed a shorter IRF delay up to 90 minutes before partum, though not enough data were available to statistically validate this difference in most windows.

This work was influenced by Romano et al. [153], who looked at variations in the power spectrum of the FHR signal in response to the UA signal. In the event of contractions the FHR demonstrated a significant response in the power in two frequency bands, 0.2-1Hz and 0.03-0.2Hz, which both persisted for roughly 60 seconds beyond the contraction, but no effort was made to identify evidence of hypoxia in the signal.

3.5 Comparison of classification results

To highlight the range and combinations of the methods possible in CTG analysis examples are presented in Table 3.1. The works are arranged alphabetically by author, listing the feature sets, dimension reduction, classifiers tested, experimental method and results recorded. Acronyms used in this Table alone are explained and listed below:

1. **Outcome measure**: The target variable for classification. AP = Apgar test (any timespan), BW = birth weight, BE/D = base excess or base deficit, CA =

clinical assessment, pH = post-partum blood pH value.

2. **Feature types**: Divided according to the feature types listed in Section 3.1. M - Morphological, T - Time domain, F - Frequency domain or Wavelets, N - Non-linear features, S - ST-analysis features.

3. **Feature reduction**: As listed in Section 3.1.1 with the inclusion of: H - heuristic selection.

4. **Classifier**: The classifiers evaluated as listed in Section 3.2. Other classifiers not already listed: DA = Discriminant analysis, LR = Logistic regression, NB = Naive Bayes, NM = Nearest mean, RB = heuristically derived rule-based classification.

5. **Experimental method**: The process used evaluate classification performance, including the number of folds the data were divided into and the method used to create divisions. CV = cross validation, RA = random fold assignment, ST = single test, performance was measured on a single test set not involved in model selection or training. CT = model was evaluated in a clinical setting. If oversampling is used it is included: SMOTE = synthetic minority oversampling technique, detailed in Chapter 4, Section 4.3.

Table 3.1: A comparison between the performance of classifiers used in literature. Not all these studies were performed on intrapartum data. *A committee of 10 neural networks was used **Used multiple time points, ***Assessed multiple pHs and content within each discovered cluster, ****Used additional features calculated based on existing features

Paper	Outcome	Featureset	Reduction	Classifiers	Method
Costa et al. [35]	pH	S, T, M	-	Omniview 3.5	CT
Elliot et al. [48]	BD	T, M	-	PeriCalm	ST
Georgieva et al. [61]	pH + CA	M, T	PCA	ANN*	10-fold RA
Georgoulas et al. [69]	pH	F	-	HMM	4-fold CV
Georgoulas et al. [68]	pH	T, F, M	-	SVM	SMOTE + 10-fold CV
Georgoulas et al. [66]	pH	T, F, M	PCA	DA, KNN, SVM	9 -fold CV
Georgoulas et al. [63]	pH	T, F, M	GE + PCA	DA, KNN, ANN	SMOTE + 10-fold CV
Gonçalves et al. [73]	pH	T, F, M, NL	H	LD	ST
Huang et al. [85]	CA	M, T	H	DT, DA, ANN	10-fold CV
Inibari et al. [80]	CA	M, T	PCA, URR	DT, ANN, PNN, RF	ST
Jezewski et al. [88]	AP, pH, BD, BW	T, M	-	RBF, ANN	50-fold RA
Karabulut et al. [96]	CA	M, TD, NL	-	NB, RBF, BN, SVM, ANN, DT, AB	10-fold CV
Keith et al. [97]	CA	M	-	RB	ST
Krupa et al. [104]	CA	NL	-	SVM	5-fold CV
Maeda et al. [117]	CA	M	-	ANN**	ST

Paper	Outcome	Featureset	Reduction	Classifiers	Method
Noguchi et al. [134]	CA, AP	M	-	ANN	ST
Ocak et al. [137]	CA	M, T	GA	SVM, ANN, ANFIS	ST
Ocak et al. [138]	CA	M, T	-	ANN, ANFIS	ST
Sahin et al. [154]	CA	M, T	-	ANN, SVM, LR, RBF, DT, RF, KNN	ST
Spilka et al. [166]	pH	M, T, F, NL	RELIEF	AB	SMOTE + 4-fold RA
Spilka et al. [164]	pH	M, T, F, NL	PCA, MI, GAME-NN	SVM, DT, NB	10-fold CV
Spilka et al. [165]	CA	M, T, F, NL	-	RF	5x2-fold CV
Sundar et al. [170]	CA	M, T	PCA	ANN	10-fold CV
Xu et al. [203]	pH	M, T, NL	GA, LASSO	SVM, RF	-
Tomas et al. [177]	CA	M, T	CF	RF, LD, DT	-
Warrick et al. [190]	BD	NL	H	SVM	10-fold CV
Yilmaz et al. [206]	CA	M, T	-	SVM	10-fold CV

Looking through the wealth of research conducted in this field in Table 3.1 several patterns arise. The first is the rise in the variety of classifier constructions and experimental methods which have been examined since the release of two open source databases from CTU-UHB and the University of California, Irvine (UCI). A second, more problematic, pattern in the literature is the difficulty in directly comparing classifiers due to the lack of benchmarking tests and established experimental procedures.

These effects can be demonstrated in the results presented in literature, and when taken together mean little useful information is being provided in many studies. With a single database research results should be comparable (excluding distinctions made between 2-class and 3-class problems). The methods of Tomas [177] and Yilmaz [206] who both used the UCI database and a 3-class outcome based on the majority voted clinical assessment are compared. Taking just the overall accuracy these results demonstrate 93.7% correct class using RF, and 91.62% correct using a SVM, which implies the SVM is the superior method. However a different experimental method was chosen in each case (10-fold cross validation vs single repeat 50% training 50% testing), and a different parameter optimisation method (Particle Swarm vs Experimentation). The larger training sets in 10-fold CV and the more granular global optimisation achieved using the particle swarm will both improve the expected performance of any classifier, invalidating the implication that one method is superior.

Sahin et al. [154] compared multiple methods using the same UCI database including the RF and SVM, but reposed the task as a 2-class problem by removing data corresponding to the suspicious label. They found RF as the better performer (accuracy 99.18% vs 98.96% for the SVM) of the two methods. This result can be compared to the result found by Ocak [137] who found, using an SVM with GA feature selection and experimentally selected model parameters, an accuracy of 99.36%,

which outperforms the previous 2-class RF result found by Sahin et al.

When Xu et al. [203] compared an SVM optimised with a genetic algorithm against RF using the OCFMT database the SVM was demonstrated as the higher performer (73.58% accuracy, kappa = 0.47 vs 72.64%, kappa = 0.45), however no details were provided on the training method or parameter selection process for the RF so any comparison between the two is impossible.

From these three experiments comparing SVM and RF classifiers, the only demonstrable results are that an SVM with one certain configuration of parameters and features selected using GA outperforms one configuration of RF on one task, and that a single configuration of RF outperforms a single SVM on a second task. Without detailed information on the selection of parameters used and the method of their implementation, results are unrepeatable and little information is obtained which can be used in future research.

Three proposals are suggested to increase the reproducibility and comparability of results. These are the selection of the correct outcome measurement, the modification of individual experimental features, and the correct selection of process complexity given the amount of available data.

Selection of an outcome measurement

Results are reported as either confusion matrices, statistical values (accuracy, sensitivity, specificity), F-Measures, Cohen's Kappa statistics, the Geometric mean, or the Area Under the Receiver-Operator Curve (AUC). Accuracy suffers in the context of heavily imbalanced datasets. In the OCFMT database only 0.82% of cases are positive with pH < 7.0, meaning 99.18% accuracy could be attained by assigning all cases as negative. Methods such as the geometric mean of the sensitivity and specificity

only partially account for this effect. ROC curves, whilst informative on relative classifier performance, are limited to binary classifiers which output a continuous class estimation without substantial modification to the original method and the classifier.

Confusion matrices are also useful measures of classifier performance especially when comparing multi-class problems as they directly output decisions made by the classifier. Cohen's Kappa is recommended as a measure of agreement which compensates for class imbalance by including an estimate of the probability of misclassification by chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed accuracy and p_e is the probability of agreement by chance. For two experts A and B $p_e = p(A = 1) \times p(B = 1) + p(A = 0) \times p(B = 0)$. In practice greater emphasis is placed on correct identification of the positive class due to the relative cost of misclassification of positive cases (risk of death or morbidity) than negative cases (unnecessary intervention), so the f-score is also an appropriate measurement of performance:

$$f = 2\frac{sens \times ppv}{sens + ppv}.$$

The recommendation is that reported results include at least the raw decisions either a confusion matrix or sensitivity specificity table along with Cohen's Kappa values for single experiments. When reporting the results of cross-validation experiments confusion matrices are impractical, hence at least the mean statistical values, Kappa, and F-measure should be reported. Related to the reporting of outcome is the selection of an outcome measurement. Jezewski et al. [88] showed that classifier performance is affected by the choice of outcome measure when using an ANN. This effect was

also demonstrated in Chapter 5 when individual CTG features are discovered to be sensitive to different outcome measures.

Experimental effect control

A second concern is the variation of multiple experimental parameters together, masking the individual efficacy of each variable. This occurs both within studies such as [137] where both classification method and parameter selection method are varied, and between studies such as [170] and [177] where classifier, parameter, and feature selection methods varied. Though other works are frequently cited in each work, there is no effort to replicate previous results and to then demonstrate new methodologies in the context of historical works.

In Table 3.1 previous classifications of the CTG are listed, demonstrating the number of available options for each of the possible parameters (feature selection, dimension reduction, classification, experimental method) and the datasets used to test the performance of these. Despite the huge range of methods which have been trialled no consensus has been reached as to optimal methods, and there has been little effort to replicate results observed by others.

Model complexity

A third concern is with the balance between the small amount of data available and the complexity of generated models. Ocak [138] provides an example of this when comparing an ANFIS against a traditional ANN using again the UCI database. The ANN used comprised of 21 nodes on an input layer, 5 and 3 nodes in 2 hidden layers, and a single output, giving 132 parameters to estimate. The ANFIS featured 13 decision rules and 21 inputs totalling 832 parameters. Under the chosen experimental

scheme the training and testing sets each consisted of only 88 positive samples against 828 negative samples, meaning nearly 10 times as many parameters were used as positive samples.

The large number of parameters brings about two problems. The first is that care must be taken to avoid over-fitting, the second is that the complexity of the ANN model is much lower than the ANFIS, so it is expected to perform less well in this specific scenario. The paper is commended for including a detailed description of the experimental method, the models used and their implementation, but does not provide strong evidence that the ANFIS will consistently outperform the ANN.

The purpose of this research has been to investigate possible methods to identify at-risk fetuses using automated methods, therefore some consensus must be eventually reached. At the same time the "no free lunch theory" for classifiers must be considered, which states that there is no universal ideal classifier [201], even in related fields.

3.6 Summary

The advent of clinical guidelines and computerized identification of features is a major step in providing an objective system for identifying at risk patterns. The number of features which can be extracted from the signal is constantly growing, and there is a lot of room to explore their significance and relations to the physiological status of the fetus. Multiple optimised sets of features have been presented in the the literature, as well as methods for selecting an optimal set in practice.

Intelligent monitoring systems have been commercially implemented, and have demonstrated an increase in the accuracy and reproducibility of clinician assessment. These systems are based upon clinical features, with further work remaining to com-

bine these with features and decisions from the classification research.

There also exist a large number of classification algorithms which have been implemented in CTG classification, a number which has grown significantly with the recent release of two open source databases. Despite this surge in research interest little progress has been made settling on an optimal solution to CTG classification.

This Thesis intends to provide insight into the relative performance of classifiers which have demonstrated potential as strong classifiers of CTG records without the potential biases of feature selection, data imbalance or model optimisation. Secondly this Thesis will explore the potential of data-driven models of the the CTG and the work by Warrick et al. to potentially provide insight on the mechanisms of fetal distress during labour and to improve classification performance.

Chapter 4

An Empirical Comparison of Classifiers

In this Chapter three experiments are performed to assess the performance of classifiers which have been suggested in the literature or proposed as appropriate for CTG classification. The data used in these experiments are described in Section 4.1. Each of the stationary (single time point) classifiers used are described in Section 4.2.

The optimal training/test data ratio is assessed in Section 4.3. Individual stationary classifiers are compared in Section 4.4. Finally in Section 4.5 the implementation and evaluation of a non-stationary classifier, the HMM, is described.

4.1 Databases

The Oxford Centre for Fetal Monitoring Technology database

The OCFMT database is uniquely large, consisting of 107,614 deliveries at the John Radcliffe hospital between 20 April 1993 - 28 February 2008. Of these records 70,990 were eliminated due to incomplete data, non-labour CS, or coming from a multiple pregnancy. 29,056 were eliminated due to poor data quality, defined as ending > 1 minute from birth, having < 30 minutes of stage 2 labour recorded, or insufficient signal quality. This left 7,568 cases in total.

Each case included FHR and UA signals at 4Hz and 0.5Hz respectively, clinical information about the patient as listed in table 4.1, umbilical arterial pH at birth, and a clinician's assessment of compromise graded from 0-3 as defined in table 4.2.

The data were pre-processed using the OxSys algorithms [62, 60]. Autocorrelation and maternal vessel artefacts identified following the Dawes heuristic; abrupt temporary increases in the FHR (> 35bpm for < 12 seconds) and decreases (> 50bpm for < 30 seconds); were removed.

For feature detection the FHR signals were then downsampled to 0.25Hz using a

Table 4.1: The clinical features provided with each case in the OCFMT database

Clinical features	Range
Gestation (weeks)	37-42
Maternal temperature (°C)	35-40
Parity	0-9
Meconium staining	-1,0,1 (clear,other,thick)
Epidural	0,1 (yes,no)
Sex	0,1 (male,female)

stepped averaging filter. The FHR baseline is assigned using the Open-Close-Smooth algorithm developed by Cazares et al. [23] and subtracted from the FHR signal. Decelerations are identified first. The FHR is smoothened using an opening filter of length 3 (12 seconds) and decelerations then identified as segments which remain below a threshold of -8bpm for at least 16 seconds [60]. Accelerations are identified using a similar rule based filter in the regions not occupied by decelerations. Contractions are identified separately using a second morphological algorithm [62].

Cases were then segmented using a 15 minute windows sliding at 5 minute intervals. In each window the 64 features listed in Appendix A are calculated using the features identified in the previous step.

The data were used to create three overlapping databases.

- *Extreme Dataset - OCFMT-E* is a balanced dataset of 124 cases divided into two classes. The adverse class contains all 62 cases with arterial pH$<$ 7.1 and severe compromise. The healthy contained 62 cases selected randomly from the 959 cases with $7.27 < pH < 7.33$ and no signs of compromise.

- *Marginal Dataset - OCFMT-M* is a balanced set of 376 cases - all 188 cases with pH$<$ 7.1 and mild, moderate, or severe compromise again matched with 188 cases randomly selected from the healthy set defined by $7.21 < pH < 7.33$

and no compromise. The marginal dataset included all cases from the extreme dataset.

- *Main Dataset - OCFMT* Contains all 7,568 cases available.

The Czech Technical University and University Hospital in Brno database

The CTU-UHB database is available online as part of the Physionet project [71]. It contains 552 records selected from 9,164 collected between 2010 and 2012 at the University Hospital in Brno using either internal or external measurements [29]. Each selected case had to conform to the following criteria:

- Singleton pregnancy.
- Gestational age > 36 weeks.
- Length of stage 2 labour ≤ 30 minutes.
- In any given 30 minute window, at least 50% of the FHR signal was present.
- No known developmental defects.

Each record was sampled at 4Hz, and was at most 90 minutes long. 46 cases were delivered by CS; the remainder being vaginal deliveries. In addition to the CTG signal multiple outcome measures were provided, with the outcomes used in this Thesis highlighted:

- Cord gas measurements - *pH*, *BD*, *BE*, and concentration of CO_2.
- Clinical assessment - Apgar Scores at 1 and 5 minutes.

Table 4.2: Definitions of compromise levels provided as one of the outcome measures in the OCFMT database

Compromise Level	Definition
3 - Severe	Neonatal death or cerebral dysfunction (including seizures & haemorrhage
2 - Moderate	Low Apgar score (<4 at 1 minute) and/or cardiac massage and/or intubation
1 - Mild	Resuscitation using a face mask
0 - None	None of the above

Clinical information including maternal risk factors, available fetal measurements at birth, and information concerning the delivery was also provided.

To extract windowed features from the CTU-UHB database the UA signal was initially down sampled to 0.25Hz using a stepped averaging filter. The same pre-processing techniques and algorithms which prepared the OCFMT database were then used to extract 64 features in 5 minute moving windows of length 15 minutes [202].

The University of California, Irvine database

The UCI dataset contains 2,126 recordings, with pre-processing and feature extraction completed. 21 features were provided including morphological, time-domain, and statistical time-domain features which are listed in Appendix B. A marker identifying one of 10 clinical patterns labelled by a consensus of experts was also included, but is not used in this Thesis.

Each record included a 3-class fetal state assessment made by majority consensus of three experts following the guidelines set out by FIGO and the NICHD. Of the 2126 samples 1,655 were identified as normal, 295 suspicious, 176 pathologic. The database is available online [109] and details of the dataset preprocessing, filtering, and feature extraction are described in related literature [9].

4.2 Classifiers evaluated

Six classifiers were selected for comparison, four based on models which had been previously evaluated in the literature (Adaboost (AB), ANN, RF, SVM). Two additional classifiers were included, the Relevance Vector Machine (RVM) and the BANN.

The implementation of each model is described below, as well as any preliminary investigations to establish model parameter importance and default parameter values conducted using the OCFMT database.

Hyperparameter optimisation

In each experiment conducted, select model parameters were optimised using either grid search where computationally viable or the Pattern Search algorithm in the Matlab Global Optimization Toolbox (Version 3.3, Mathworks inc.).

4.2.1 AdaBoost

The Adaboost algorithm developed by Freund and Schapire [54] is a meta-algorithm which combines multiple weak classifiers to form a single strong classifier. A weak classifier is defined as one whose performance is only slightly better than random guesses. Intuitively the algorithm builds a strong classifier by forcing each successive classifier to focus on the cases which were misclassified by the previous classifiers. The final class decision for an individual case, $\mathbf{x}$, is a weighted sum of N weak classifiers, $h^{(n)}$:

$$\mathbf{h}(\mathbf{x_m}) = \sum_{n=1}^{N} \alpha^n h^{(n)}(\mathbf{x_m})$$

where α^n is the weight on classifier n. The algorithm to train this strong classifier, $\mathbf{h}$, composed of N weak classifiers using dataset $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_M}]$ is run as follows:

Initialise a set of weights $\mathbf{w^{(0)}} = w_1, \ldots, w_M$, where w_m is the cost of misclassifying sample m. The superscript denotes which iteration of the method the weights belong to, and is initially 0. Each weight is initially set to $1/M$ so that they sum to unity.
For weak classifier $n = 1, \ldots N$

1. Train classifier $h^{(n)}$ to minimise the weighted error function. A new parameter $\mathbf{y}$ is introduced to simplify this calculation where:

$$y_m = \begin{cases} 1 & \text{if } h^{(n)}(\mathbf{x_m}) = t_m \\ 0 & \text{otherwise} \end{cases}$$

with t_m the target output and $\mathbf{x_m}$ the input for case m. The weighted error equation then becomes:

$$\epsilon^{(n)} = \sum_{i=1}^{M} w_m^{(n)}(y_m)$$

2. This individual classifier is then given a coefficient:

$$\alpha^{(n)} = \frac{1}{2} ln \left(\frac{1 - \epsilon^{(n)}}{\epsilon^{(n)}} \right)$$

3. Misclassification weights are reassigned so that correctly classified points lose importance, the first term, and misclassified points gain importance, the second term, whilst maintaining the condition that they sum to 1:

$$w_m^{(n+1)} = y_m \left(\frac{w_m^{(n)}}{2\epsilon^{(n)}} \right) + (1 - y_m) \left(\frac{w_m^{(n)}}{2(1 - \epsilon^{(n)})} \right)$$

Repeat from step 1 with the updated sample weights until N classifiers have been trained.

Implementation

The Adaboost algorithm was implemented using Matlab (R2014a) and the statistics and machine learning toolbox (Version 9.0. The MathWorks Inc., Natick, MA, 2015). AdaBoost is susceptible to label noise, a factor which is relevant in CTG classification as cord gas measurements are not reliable, and several outcome measures are subjectively assessed. To counteract this the robust boosting algorithm is used [53]. As mislabelled cases represent large absolute errors (they may behave strongly like their true class and will be classified as such) the margin term which the method attempts to minimize will be dominated by these cases. The robust algorithm replaces thes exponential margin function used with a sigmoid, which mitigates the contribution of incorrectly labelled samples.

This however introduces an additional tunable parameter, the target misclassification rate, p_{mc}, in addition to the number of weak classifiers to train. Default values were estimated using the OCFMT-E database in Figure 4.1 to maximise the 10-fold cross validation rate using two independent grid searches. The number of model classifiers was varied from 1 to 120, and the target misclassification proportion between 0 and 0.25. The classifier numbers were assessed first, with p_{mc} set to 0.1. A default value of 80 was selected. The target misclassification rate was then varied across the target range in increments of 0.025 with the number of classifiers set to 80. The default misclassification rate was set to $p_{mc} = 0.1$, which maximised the validation accuracy.

4.2.2 ANN

The ANN is composed of a network of nodes, with each node acting like a neuron. The output of an individual node is a weighted linear combination of nodes in the previous

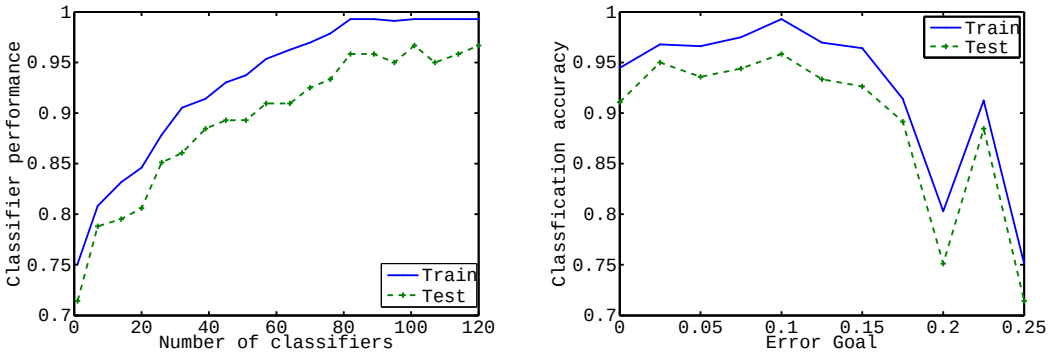


Figure 4.1: Default parameters for an AdaBoost model are created by investigating the classifier performance (defined as the proportion of cases correctly classified) on the OCFMT-E database using 10-fold cross validation. The performance did not increase beyond 80 classifiers, which is selected as the default value. The best performance was achieved with a robust error goal of 0.1.

layer, passed through a non-linear transfer function, shown in Figure 4.2. By arranging the nodes in layers, higher level representations of information can be learned. The first layer is known as the input layer, with each node representing a single feature. Nodes in the final layer are termed output nodes, and are restricted to the range $[0, 1]$ using a logistic sigmoid function, representing the probability of class membership. Layers in between are described as hidden layers as their inputs and outputs are not observed outside of the network.

The total input to each individual node is the weighted sum of the outputs from the previous layer $\mathbf{x}$. Each input z_i coming from source i to node j has its own weight, w_{ij}, therefore the contribution from node i, a_j is $z_i w_{ij}$. The bias weight b, shown in black in Figure 4.3, is included as its own node with output 1 and weight b_j equal to the bias on node j. The total is termed the activation:

$$a = \sum_{i=1}^{I} w_{ij} z_i + b_j$$

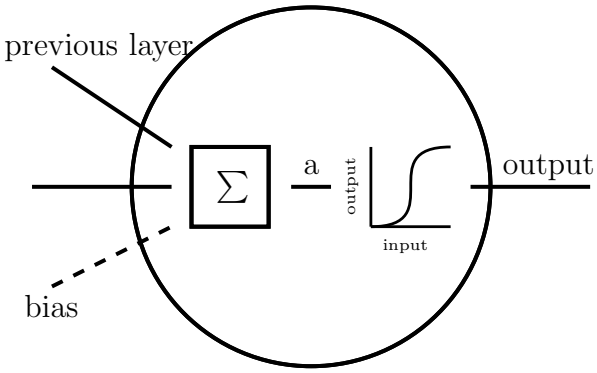


Figure 4.2: A depiction of the input/output mapping of a single node. A series of outputs are received from the previous layer, along with a bias, which are summed to give the activation, a. This activation is then transformed through the chosen function (tanh is depicted here) to produce the output of the node, z.

The output from each node, z, is a function of the activation:

$$z_k = g(a_j)$$

To optimise the weights of the connections between nodes and the biases networks are trained using the back-propagation algorithm. In the sample 2-layer network shown in Figure 4.3 the output at node k for case n can be found by forward propagating the features $\mathbf{x_n}$ through the network:

$$y_k = h\left(\sum_{j=0}^{J} w_{kj} g\left(\sum_{i=0}^{I} w_{ji}\mathbf{x_n}\right)\right)$$

where h represents the logistic sigmoid function, $i = (1...I)$ represents the first layer of I inputs, $j = (1...J)$ the second layer of J nodes and $k = (1...K)$ the final layer of nodes. w_{ij} represents the weight connecting nodes i and j.

The model error is represented using the cross-entropy error function:

$$E = -\sum_{k}(t^k ln y^k + (1 - t^k) ln(1 - y^k))$$

where t is the target output value. Model weights are adjusted to minimise the error using gradient descent. To find the error gradient at a given node the error is propagated backwards through the network. For weight w_{ij} in the example network in Figure 4.3 this process begins by defining the gradient of the error with respect to the output. Using the chain rule and the gradient of the output, y_k with respect to the output node activation a_k the error at node k, δ_k, is defined as:

$$\begin{aligned}\delta_k &= \frac{\partial E}{\partial a_k} \\ &= y_k - t_k\end{aligned}$$

The error at node j is found by summing the contributions of the errors at each node it feeds, in this case all nodes in output layer K. First the sum of the contribution from the outputs is found:

$$\begin{aligned}\frac{\partial E}{\partial z_j} &= \sum_k \frac{\partial E}{\partial a_k}\frac{\partial a_k}{\partial z_j} \\ &= \sum_k \delta_k w_{jk}\end{aligned}$$

The error at node j is found by propagating the error through the nodes activation

function

$$\begin{aligned}\delta_j &= \frac{\partial E}{\partial a_j} \\ &= \frac{\partial E}{\partial z_j}\frac{\partial z_j}{\partial a_j} \\ &= g'(a_j)\sum_k \delta_k w_{jk}\end{aligned}$$

The gradient of the error with respect to weight w_{ij} can then be defined in terms of this new error at node j in the same manner as the previous layer:

$$\frac{\partial E}{\partial w_{ij}} = \delta_j z_j$$

where z_j represents the input from node j. As this method relies on gradient descent it will converge to the nearest local minima. This can be overcome to some degree by including a momentum term, which is applied to change weights in conjunction with the learning rate and can carry the search out of a small minima. This method is also dependent on the initial model weights, therefore multiple optima may be found through repeated random initialisation and training steps.

Implementation

The ANN offers a great selection in model parameters. The number of layers, nodes in each layer, and activation functions in each layer must be selected. Parameter training is an iterative process, and the parameters of the training algorithm must also be set. These include the learning rate, L, the stopping criteria, and the maximum number of training cycles performed. The ANN were trained using the Netlab package for Matlab [124] using the scaled-conjugate gradient descent algorithm. This local opti-

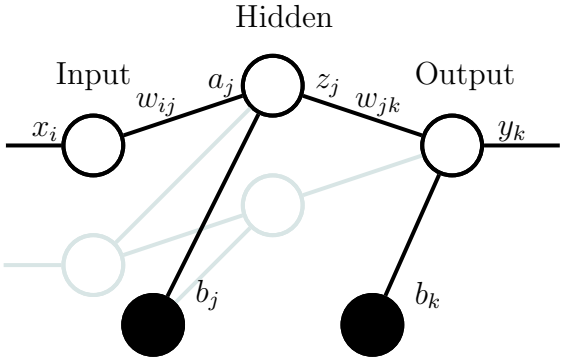


Figure 4.3: A two layer network, with one node in each layer highlighted and annotated. Bias weights are shown as black nodes, ordinary nodes are shown in white. The node outputs, z and activations, a, are shown at either end of the connecting lines, with connection weights positioned in the centre of each connecting line.

miser is sensitive to the initial weights given to each parameter, which were initialised as random numbers selected from a normal distribution with $\mu = 0$ and $\sigma^2 = 0.01$ based on default recommended in the Netlab documentation. As this algorithm does not automatically do so, data were shuffled before being presented to the learner to prevent bias in the training process.

The ANN architecture was selected based on a previous study on ANNCs by Georgieva et al. [61] consisting of a single output node with a logistic output function, and 6 input nodes. In the literature a hidden layer with 2 nodes was found to perform well in generalisation using the OCFMT-E databse, which is used as the default value. In testing the algorithm always converged within 45 iterations, which was set as an upper limit.

With the architecture and training parameters established the most significant parameter to optimise per database was the number of nodes in the hidden layer. This was optimised using a grid search between 2 and 10 nodes, with performance averaged across each fold of training data. To account for the stochastic nature of the

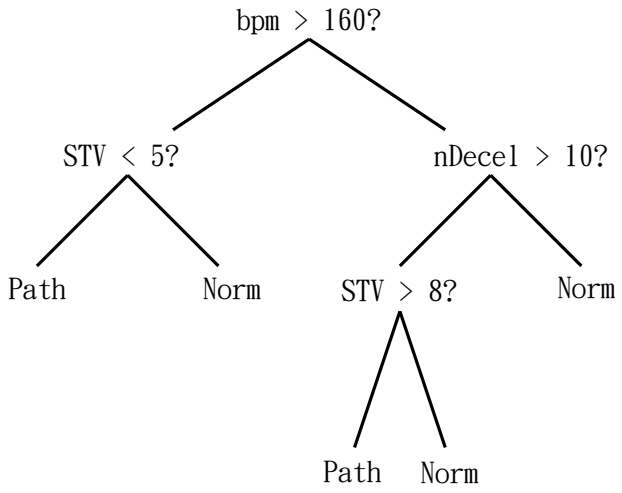


Figure 4.4: An artificial example of a single decision tree applied to CTG data. At each branch a decision is made based on the value of a chosen feature. Branches are terminated once the stopping criteria have been reached, such as Gini impurity exceeding a threshold. The terminating nodes are called leafs, and are associated with class decisions.

local optimiser each model was trained 10 times, and the best performer selected.

4.2.3 RF

To define random forests, first decision trees must be defined, along with a method for learning them. A decision tree, demonstrated in figure 4.4 classifies an incoming vector by making a series of decisions until a leaf, or termination node is reached, which will be associated with a class decision. The CART (Classification and Regression Trees) algorithm is one method to learn tree structures from training data.

Decisions are made at each node to maximise the total Gini impurity of the two subsets. Gini impurity represents how often a random sample from a set would be

misclassified if it were labelled according to the distribution of labels in a subset:

$$g = 1 - \sum_j p_j^2$$

where p_j is the proportion of class j in the dataset and g is the Gini impurity. The impurity has a maximum value when all classes occur with an equal probability, and a minimum value of 0 when there is only one class in the subset. The CART process is described in Algorithm 4.1.

Algorithm 4.1 The CART algorithm

```
while all branches are not terminated do
  for each branch do
    if termination criteria not met then
      select a random subset of features
      for each selected feature do
        find a split which minimises the subsets' Gini impurity
      end for
      create a new branch from this parent according to the feature and split giving
      the lowest overall Gini impurity
    else if termination criteria met then
      mark this branch as terminated
    end if
  end for
end while
```

The termination criteria are either the number of samples in a data-subset reaching a lower limit, or the proportion of a single class in a data-subset attaining an upper limit. The random forest is a collection of decision trees who each vote for a majority class. There are two differences between the single CART implementation and the trees grown in the random forest. Each tree is generated from the complete set of features and the full dataset by selecting a random training subset and at each node a random set of features.

Of particular importance is the selection of a new set of random features to use at each node which Breiman demonstrates minimises correlation between trees whilst maintaining discriminative strength [19]. The final decision of the total classifier is a majority vote of each tree.

Implementation

The Random Forest algorithm was implemented using the Matlab Machine Learning toolbox. Though tree structure and pruning methods are important, the parameters with the greatest influence on performance have been identified as the number of parameters selected at each tree node n_{param} and the number of decision trees in the forest n_{tree}. Two independent grid searches were used to identify default values for these parameters, shown in figure 4.5. The number of selected parameters did not influence the classification performance, therefore this value was kept at the value recommended in the literature for classification, $\sqrt{n_{feat}}$, where n_{feat} is the number of features available. The number of trees was linearly varied from 1 to 120. An accuracy of 96.7% was achieved using 32 trees, which could be increased to 98.3% with 70 trees. The former is sufficient as a default value as the simpler model is expected to perform better in generalised tasks.

4.2.4 RVM

The relevance vector machine was developed by Tipping as a method of sparsely representing the solution to classification and regression problems. The method tends to produce a more sparse representation than the SVM [175], and the samples selected as relevance vectors tend to be representatives of their class [179]. This property is

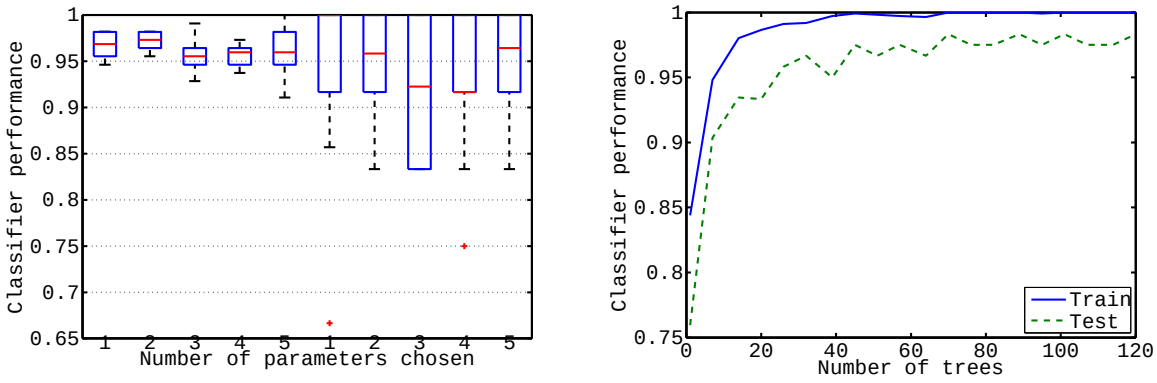


Figure 4.5: Default parameters for the Random Forest model are created by investigating the classifier performance on the OCFMT-E database using 10-fold cross validation. The number of retained parameters is plotted against the classifier performance. The first 5 boxes show the performance on the training data, the right 5 on the testing data. In both cases the values are not significantly affected by the number of retained parameters. The model accuracy levels off and is not substantially increased by using more than 30 trees.

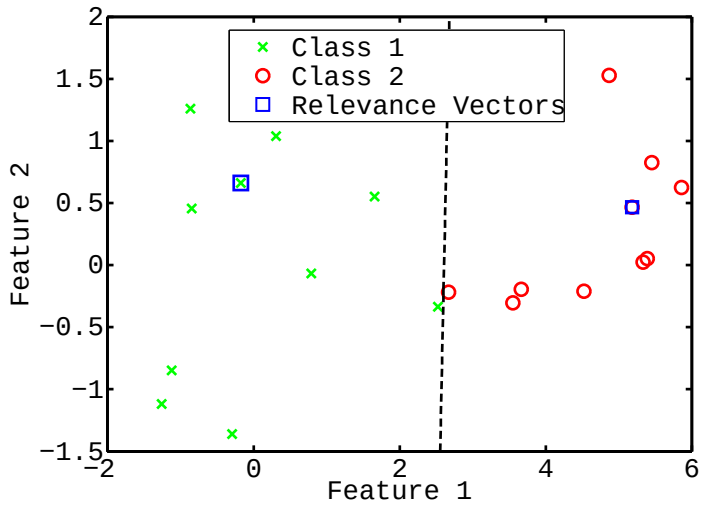
demonstrated in Figure 4.6a.

Training the RVM uses the same evidence procedure as the BANN [176] explained in detail in Section 4.2.6. A sparse vector of weights, $\mathbf{w}$, is to be found to describe the output:

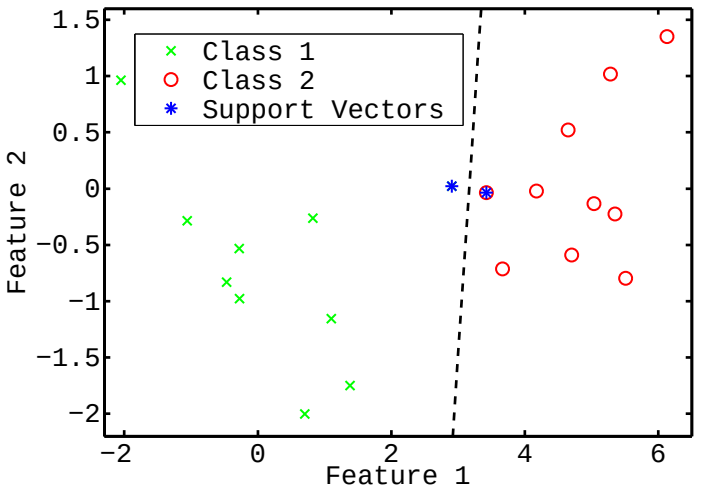
$$y = \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x})$$

where $\phi(\mathbf{x})$ is a non-linear mapping of the input $\mathbf{x}$. Assuming the model output $y(x)$ is limited using a logistic sigmoid $\sigma(y) = 1/(1 + e^{-y}$ then the posterior probability of the targets can be described by:

$$P(\mathbf{D}|\mathbf{w}) = \prod_{n=1}^{N} \sigma\left(y(\mathbf{x_n}, \mathbf{w})\right)^{t_n} \left(1 - \sigma(y(\mathbf{x_n}, \mathbf{w})\right)^{1-t_n}$$



(a) Relevance Vectors



(b) Support Vectors

Figure 4.6: An example of cases selected to define decision boundaries on a toy dataset using (a) an RVM and (b) an SVM. The support vectors, highlighted with blue asterisks, are observed as the cases closest to the decision boundary. The relevance vectors, highlighted with blue squares, are cases in the centre of their clusters, and highly representative of their class.

The weight on each training sample is defined using a prior distribution in the form:

$$P(\mathbf{w}|\alpha) = \prod_{i=0}^{M} N(0, \alpha_i^{-1})$$

where α_i is the precision, or inverse variance of the weight w_i and M is the dimensionality of $\mathbf{x}$. This distribution again acts as a regularisation parameter, penalising large weights.

To create an iterative formula to estimate the hyperparameters the evidence for the hyperparameters $p(D|\alpha)$ must be maximised. For a selected basis function ϕ the resultant formula is:

$$\alpha_i = \frac{\gamma_i}{\mathbf{w_{MP}}}$$

where:

$$\begin{aligned}
\gamma_i &= \alpha_i \Sigma_{ii} \\
\mathbf{\Sigma} &= \mathbf{\Phi^T B \Phi} + \mathbf{A} \\
\mathbf{w_{MP}} &= \mathbf{\Sigma \Phi^T t}
\end{aligned}$$

with $\mathbf{A} = diag(\alpha_0, \ldots \alpha_N)$ and $\mathbf{B} = diag(B_0, \ldots, B_N)$ with $B_n = \sigma(y(x_n))(1-\sigma(y(x_n))$ and $\mathbf{\Phi}$ is the design matrix resulting from the selected basis function.

As these variances are updated, most will tend to zero causing the relevant weight to be most probably zero. These weights are pruned, the remaining weights and their corresponding samples being the relevance vectors.

Implementation

The RVM shares many similarities with the SVM, including the use of a kernel function. The same kernel function as the SVM case, the RBF, was used again. Unlike the SVM there is no box-constraint parameter, therefore optimisation is reduced to optimising the RBF variance. The RVM model was implemented in Matlab using the Sparse Bayes toolbox [175]. The variance was optimised using the pattern search algorithm, initialised at the default value of 2.4 calculated below.

The number of relevance vectors retained and the model's performance are mapped in figure 4.7 using the OCFMT-E database to estimate strong default parameters. From this plot a default value for σ^2 of 2.4 was selected, which resulted in an average of 19.3 relevance vectors being selected and an average testing accuracy of 91.2% across all 10 folds. This model compares favourably with the SVM demonstrated in Figure 4.8 which attained an accuracy of 96.5% using $\sigma^2 = 1.32$ and $C = 10.74$, requiring an average of 75.83 support vectors.

4.2.5 SVM

The SVM aims to find a hyperplane that maximises the distance between the plane and the boundary to each class. For a linearly separable dataset $\mathbf{X} = (\mathbf{x_1}, \ldots \mathbf{x_N})$ belonging to class $\mathbf{t} = (t_1, \ldots t_N)$ the SVM finds a vector of weights $\mathbf{w}$ and constants $\mathbf{b}$ describing a line $\mathbf{w^T X} + \mathbf{b}$ whose value is ≥ 1 for one class and ≤ -1 for the other.

Additionally this line must maximise the distance between itself and the nearest points in each classes. These nearest points, defined as locations where $\mathbf{w^T X} + \mathbf{b} = \pm 1$ are called support vectors. By introducing a new variable $\mathbf{y}$ which is $+1$ for positive

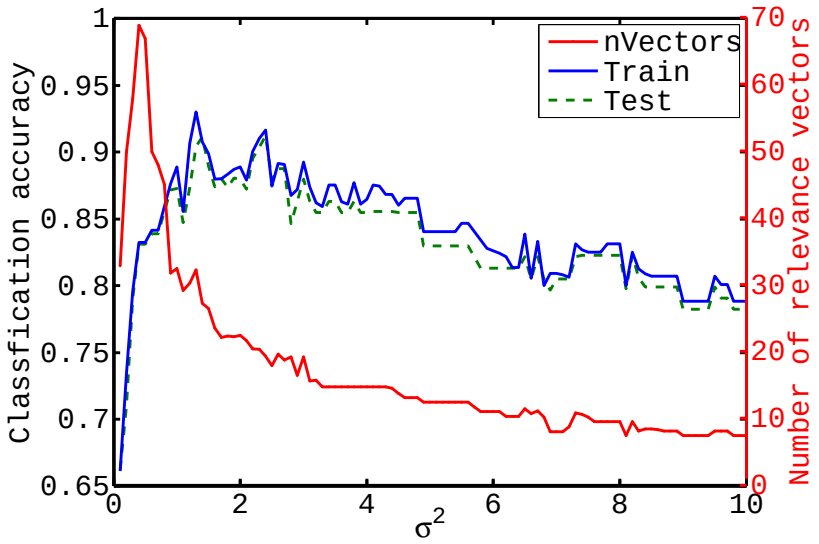


Figure 4.7: The cross-validated RVM accuracy and number of relevance vectors selected are plotted against the RBF variance. The number of relevance vectors decreases as the RBF variance increases, producing a simpler model, at the cost of classification performance.

cases and -1 for negative cases the i^{th} support vector can be re-written as:

$$y_i(\mathbf{x_i}.\mathbf{w}) - \mathbf{b}) - 1 = 0$$

Using this equation the width of the separating margin is the projection of the vector connecting the positive and negative support vectors onto a unit vector perpendicular to the dividing line, $\frac{\mathbf{w}}{\|\mathbf{w}\|}$. The result is simply:

$$\frac{2}{\|\mathbf{w}\|}$$

Maximising this margin is equivalent to minimizing $||\mathbf{w}||$ subject to the constraint that each sample must lie at least 1 away from the dividing line. For mathematical

simplicity this constrained optimisation problem is reformulated as:

$$\min f(\mathbf{x}, \mathbf{y}) \text{ s.t. } g(\mathbf{x}, \mathbf{y}) \geq 0$$

where:

$$f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{w}\|^2$$
$$g_i(\mathbf{x}, \mathbf{y}) = y_i(\mathbf{x_i}.\mathbf{w}) - \mathbf{b} - 1 \text{ for all } i$$

This can be solved by introducing a Lagrangian multiplier α and re-writing the constraint equations as a single equation:

$$L(\mathbf{x}, \mathbf{y}, \alpha) = f(\mathbf{x}, \mathbf{y}) + \alpha(g(\mathbf{x}, \mathbf{y})) \tag{4.1}$$
$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}(\alpha_i[y_i(\mathbf{w}.\mathbf{x_i} + \mathbf{b}) - 1]) \tag{4.2}$$

At this stage it is observed that for any correctly classified points, equation 4.2 is minimised by setting the corresponding α value to zero - thus they will not play a role in defining the boundary. To find the solution a point on $f(\mathbf{x}, \mathbf{y})$ must be found where its value does not increase in any direction which lies on the solution to each $g = 0$. This is done by taking the partial derivatives of equation 4.2 with respect to the weights $\mathbf{w}$ and the constant $\mathbf{b}$ to give the following pair of equations:

$$\mathbf{w} = \sum a_i y_i \mathbf{x_i} \tag{4.3}$$
$$\sum a_i y_i = 0 \tag{4.4}$$

These are substituted back into the original Lagrangian to produce a final optimisation problem which must be minimised:

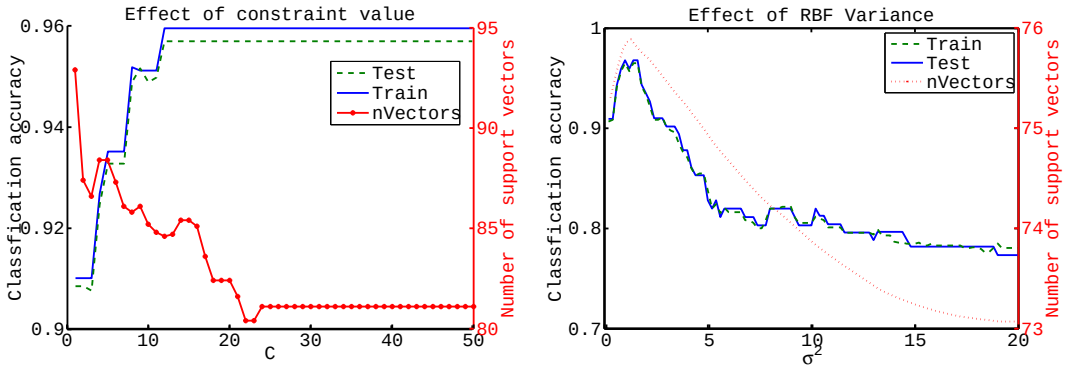
$$L(\alpha) = \sum \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}$$

Implementation

The SVM is implemented using the Matlab Statistics Toolbox version 9.0 solved using quadratic optimisation as described in Section 3.2.5. With a training method established the model selection involves choosing a kernel function to use, if any, and selecting the parameters of that function. Based on performance in the literature [66] an RBF kernel function is used, which introduces a new parameter, the variance of the function, σ^2:

$$k(x, x') = exp\left(-\frac{d(x, x')^2}{2\sigma^2}\right)$$

where $d(x, x')$ is the Euclidean distance between points x and x'. The tuning parameter is the box constraint, C, which defines the penalty for misclassification. In the literature where parameter values were recorded C and σ^2 were set to 0.3 and 2 respectively, established by trial and error [137]. In these experiments these parameters were optimised using a pattern search algorithm implemented with the Matlab Global Optimisation Toolbox version 3.2.5. Default values were established in a separate set of experiments shown in Figure 4.8 by setting one parameter to the value found in the literature and varying the other and assessing the classification performance using the OCFMT-E database. From this experiment default values for the OCFMT databases were set to $\sigma^2 = 1.32$ and $C = 10.74$, which balanced the testing accuracy and the model complexity.



(a) Box parameter

(b) RBF variance

Figure 4.8: In (a) the SVM box-parameter and corresponding number of support vectors are assessed against model accuracy. Beyond a value of 20, little change in either feature was observed, with the maximum accuracy being obtained with C = 1. In (b) the effect of RBF variance on the same two measures is plotted. The variation in the number of support vectors selected was small, with the maximum accuracy achieved with $\sigma^2 = 1.32$.

The degree of overfit in an SVM may be intuitively visualised by assessing the balance between the classification accuracy and the number of support vectors selected. The effect of each of these parameters was assessed using the cross validation error and the number of support vectors selected. The rather high number of datapoints selected as support vectors (upwards of 70 out of 112 datapoints) indicated that the data were poorly separable in the chosen projection.

4.2.6 BANN

The BANN was proposed as a method to introduce a measure of certainty into ANN classification results [61]. The BANN also allows for two further advantages over the standard ANN; firstly a greater robustness to overtraining as the model includes a regularisation function; secondly model robustness, as the uncertainty on model parameters is included within the model.

Bayesian ANN have not been found to be previously applied to the CTG classification, but have been used in the similar field of ECG classification [56]. The complete methodology is described in the literature [16, 114, 124], and is summarised below.

The BANN uses an identical model structure and annotation to the ANN. Like the ANN the model is trained by adjusting the weights of connections between nodes, however additional parameters are introduced to account for uncertainty on the estimated weights. Optimal weights are defined as those that were most likely to have generated the data; or the maximum of the posterior weight distribution $p(w|D)$, where w represents the weight parameters and D the data. Uncertainty in the weights is achieved by modelling each weight as a probability distribution shown in equation 4.5. Weights are assumed to be generally small, and their values are completely unknown before training the prior distribution assigned to each weight is a normal distribution with zero mean and variance α^{-1}.

$$p(w) = \frac{1}{Z_w(\alpha)} \exp\left(\frac{-\alpha}{2}||w||^2\right) \tag{4.5}$$

Since an optimal model will maximise the posterior weight distribution, the model error is defined as $E = -\log\left(p(w|D)\right)$. Using Bayes rule the posterior distribution can be split into the data likelihood, the weight prior, and the data prior:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} \tag{4.6}$$

The data prior $p(D)$ is a regularisation parameter, and can be neglected since only the model parameters are being adjusted. Using equation 4.6 the error term is split into two parts, the data error $-\log p(D|w)$, taken as the cross entropy error, and the

model error $-\log p(w)$, which acts as a regularisation term.

The misfit function is then defined in terms of the two errors:

$$S(\mathbf{w}) = E = E_D + \alpha E_w \tag{4.7}$$

With the hyperparameters included in the regular ANN model, the Evidence Procedure is used to determine optimal weights and hyperparameters of the model. This relies on the evidence for the hyperparameters:

$$p(D|\alpha) = \int p(D|\mathbf{w}, \alpha) p(\mathbf{w}|\mathbf{a}) d\mathbf{w}$$

From which a recursive equation for α can be formed:

$$2\alpha E_W^{MP} = W - \sum_{i=1}^{W} \frac{\alpha}{\lambda_i + \alpha}$$

where λ_i is the i^{th} eigenvalue of the Hessian $\nabla\nabla E_D$ and γ is:

$$\gamma = \sum_{i=1}^{W} \frac{1}{\lambda_i + \alpha}$$

Using this, the fact that the most probable weights are those which minimise the error, and that the error approximation holds true at the modes of the hyperparameters the evidence procedure is developed to optimise weights and parameters using the algorithm iteratively:

1. Initialize the hyperparameter and the weights. The weights are randomly initialised independently of the hyperparameter. In this study the initial bias and weight hyperparameters were set to 0.1.

2. The networks weights are then optimised using any net optimiser which accounts for the priors, in this case the scaled conjugate gradient method.
3. The evidence is computed for the hyperparameters, which are then re-estimated. The re-estimation formula can be iterated at this stage.
4. Steps 2 and 3 are repeated until convergence.

These optimal weights and hyperparameters can then be used to make estimations about new data points by considering the posterior of the class marginalised over the weight uncertainty. This involves computing the integral:

$$P(C_1|x) = \int y(x;w)p(w|D)dw$$

The value of γ represents the number of well defined parameters, those whose weights are dictated by the data, not the model priors [15] which provide an estimate of how many weights are being effectively used. Closely related to this is the concept of automatic relevance determination (ARD), where an input parameters importance can be approximated by the variance on the corresponding input nodes weight. Inputs with a corresponding large hyperparameter are unimportant, as their weights are constrained to small values; small hyperparameters allow large weights, and the corresponding input is therefore important.

Implementation

A 2-layer feed-forward network was set up to include 6 inputs, between 1-10 hidden nodes and a single output to explore this parameters' effect on performance. Networks of more than 10 hidden nodes showed signs indicative of over-fitting (large parameter weights and poor validation performance).

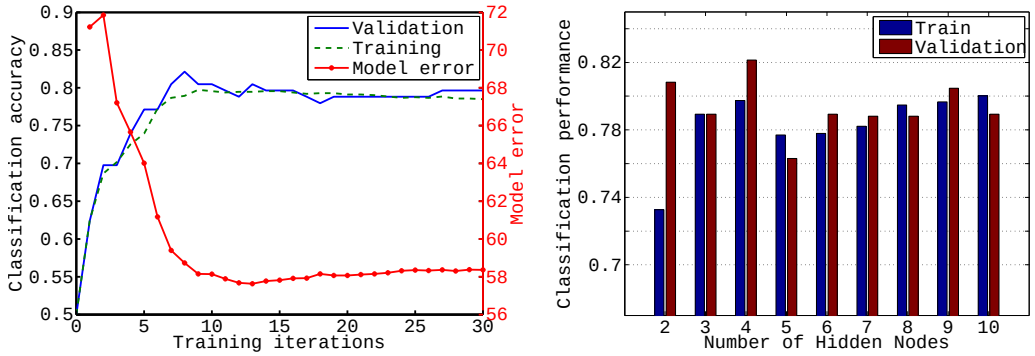
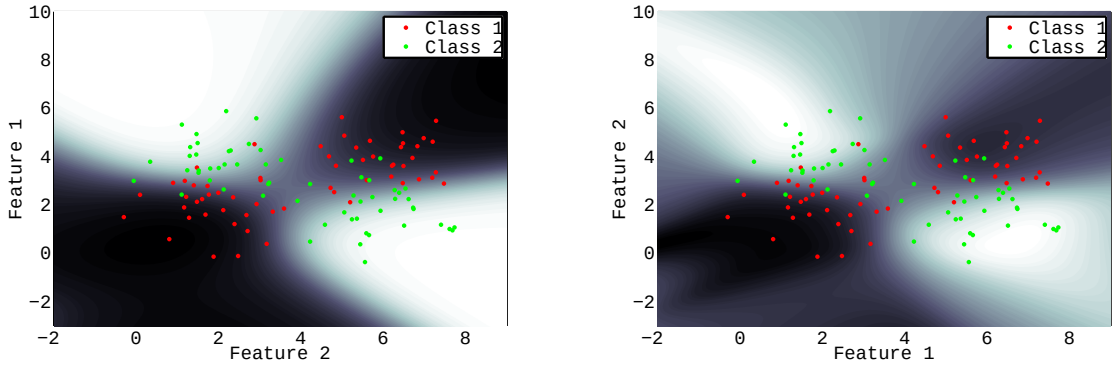


Figure 4.9: Evaluation of parameters for the BANN using the OCFMT-M dataset. In (a) the training and validation error were plotted against the number of training iterations. In (b) the model's cross-validation classification accuracy is plotted against the number of nodes in the hidden layer. The best validation performance was attained using a hidden layer with 4 nodes.

10-fold cross-validation was used to test the networks performance using dataset OCFMT-M. All priors were initially set at 0.1 with the initial networks weights random values from a zero-mean normal distribution scaled according to the number of incoming and outgoing weights. The network was then trained using the Netlab package for Matlab [124]. The resulting classification rates for 2-10 hidden nodes are shown in Figure 4.9. In implementation, a grid search over this range (2-10 hidden nodes) was performed using training data and 9-fold cross validation, with the model achieving the highest mean accuracy selected.

4.2.7 Thresholding BANN outputs

The BANN introduces marginalised class outputs, which modify the standard posterior probability of the classifier to incorporate a confidence measure dependent on the local data density. The class decision boundary remains unaffected, but the probability now changes more rapidly in regions of high training sample concentration and more



(a) Unmoderated (b) Moderated

Figure 4.10: An example of the effect of moderation on the BANN decision surface. Sample data were generated and split into two classes and used to train a BANN. The training data (points) and the resulting decision surface (color) are shown for (a) standard outputs and (b) moderated outputs. In (a) the unmoderated outputs show strong class preferences even in regions containing little data at the extremities of the image. In (b) as the decision surface moves away from the regions containing data the certainty slowly returns to uncertainty, represented as grey.

slowly in regions of low training sample concentration. Therefore decisions about test samples in regions of low data concentrations are less certain than decisions in regions with a high concentration. This effect is visualised in Figure 4.10 which shows a decision surface using moderated and unmoderated outputs. By using moderation on the classifier output, and choosing not to classify points with a low certainty (classifier outputs close to 0.5) the accuracy of the classifier can theoretically be improved.

The correlation between classifier output and concentration of the positive class can be visualised on a large dataset using an Event Rate Estimation (EveREst) plot. A BANN network trained using feature set 1 and the OCFMT-M database was used to classify all cases in dataset OCFMT, with the adverse class defined as pH < 7.1.

Figure 4.11 compares the proportion of adverse cases with the classifier output for the standard BANN outputs and the moderated BANN. At high certainties (classifier

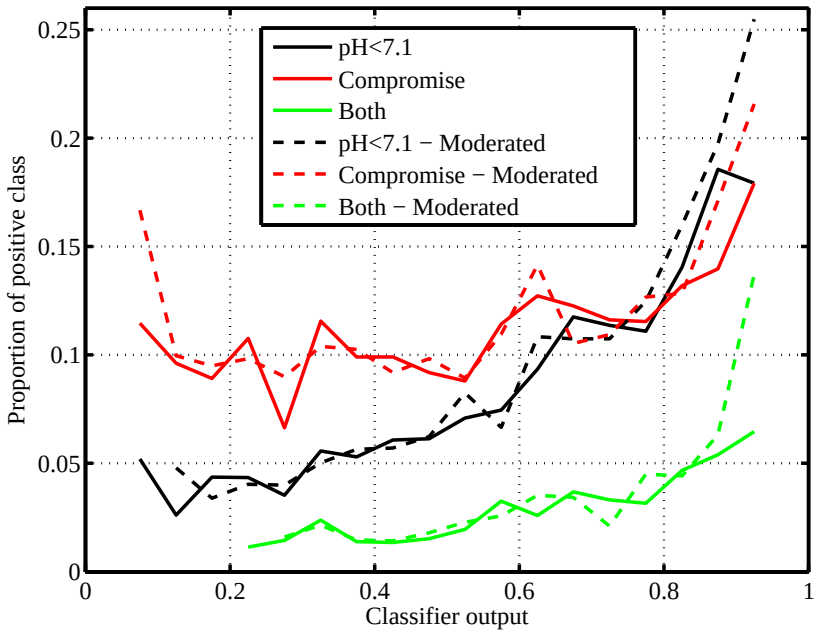


Figure 4.11: An EveREst plot showing the proportion of the positive class identified for a BANN trained using OCFMT-M and evaluated on OCFMT. The ratio of the positive is shown for standard outputs (solid line) and the moderated outputs (dashed line). At high classifier outputs, the proportion positive class is greater when using moderation.

output > 0.8) the proportion of the adverse cases is higher for each outcome measure assessed when using moderation. The number of cases falling into the higher classifier output ranges were smaller with moderated outputs. For the unmoderated BANN a total of 470, 501, and 201 cases were assigned to bins 0.8-.85, 0.85-0.9, and 0.9-0.95 with unmoderated outputs, and 475, 380, 51 with moderation. The large improvements to accuracy in the higher classifier ranges must be considered together with the loss of 75% of cases from this region.

4.2.8 Thresholding for classification

To take advantage of this effect in classification cases falling close to the marginalised decision boundary can be excluded. This effect was assessed using a BANN with dataset OCFMT-M. I third output class, uncertain, was introduced to indicate cases where there was insufficient evidence to make a decision.

Input $\mathbf{x}$ was then classified using two boundaries as: positive = $p(C_1|x) \geq t_u$, uncertain = $t_u > p(C_1|x) \geq t_l$ and negative = $t_l > p(C_1|x)$ where t_u and t_l were the upper and lower decision thresholds.

The impact of varying the decision thresholds in this experiment is plotted in Figure 4.12. Varying either boundary to its extreme maximises the sensitivity or specificity as expected, at the cost of excluding nearly all of the cases.

It was possible to achieve a modest improvement to classification scores by adjusting either threshold, which must be chosen to reflect either an acceptable proportion of cases unclassified or a desired accuracy. With no thresholding $t_u = t_l = 0.5$ the model accuracy was 64.8%. By choosing values of t_u and t_l as 0.6 and 0.4 respectively the accuracy is improved to 70.92% at the cost of choosing not to classify 77 cases (30% of the total).

In CTG classification the majority of cases fall close to the 0.5 decision threshold, making this the region of greatest data density. By eliminating this region correct decisions are potentially eliminated. To counteract this effect it is possible to examine the data density on a case by case basis.

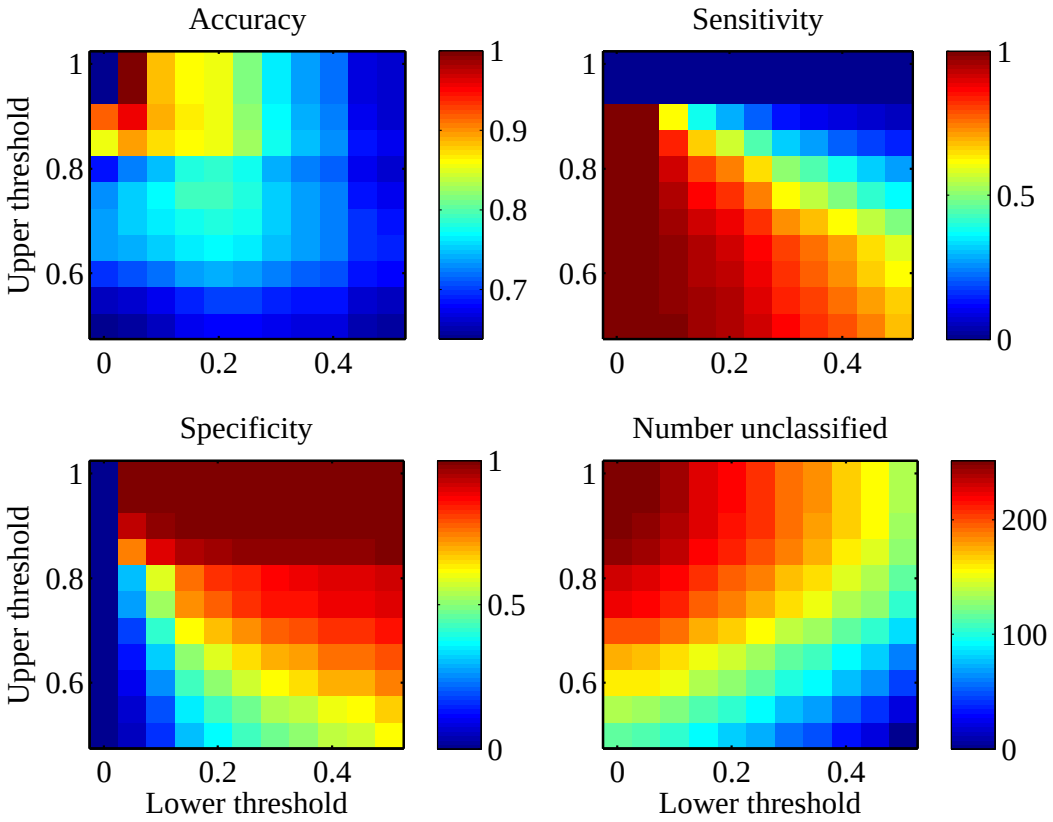


Figure 4.12: Variations in statistical measures, and the number of classes rejected as unclassified, are plotted against the lower and upper decision thresholds (x and y-axis respectively) for the BANN trained Section 4.2.6. As the thresholds are moved further away from the original decision boundary the number of cases and model accuracy both increase. A greater bias towards the upper or lower trends will improve the sensitivity and specificity respectively.

4.2.9 Calculating the individual BANN output variance

To make predictions using the BANN the posterior of the class must be found by marginalising out the uncertainty in the weights.

$$P(C_1|x) = \int y(x;w)p(w|D)dw$$

As $y(a)$ is a non-linear function, using the prediction and the approximation on the weight posterior in equation 4.7 alone is not sufficient, so the same assumption - that the distribution is normal and distributed around its mode - is used. The mode of the activation of the output node, a_{MP}, is found by forward propagating the inputs $\mathbf{x}$ through the network, and its variance $s^2 = \mathbf{g^T A^{-1} g}$ where:

$$\mathbf{g} = \left.\frac{\partial a}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w_{MP}}}$$

Using these two approximations the integral is not analytically tractable, but the resulting class posterior can be written in terms of the output activation function, $y(a)$:

$$P(C_1|x, D) = \int y(a)p(a|\mathbf{x})da$$

for which the approximated solution has been proposed,:

$$P(C_1|x, D) \approx y(\kappa(s)a_{mp})$$

where:

$$\kappa(s) = \left(1 + \frac{\pi s^2}{8}\right)^{\frac{1}{2}} \tag{4.8}$$

and a_{mp} the most probable value of the activation function found by forward propagating the inputs through the network. The output variance is then taken as s^2 in equation 4.8 and propagated through the activation function also.

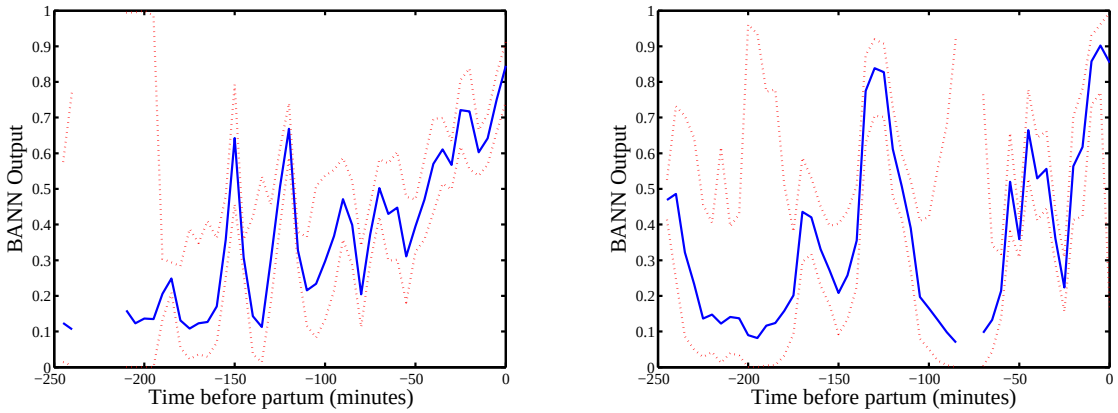
4.2.10 Classification with BANN output variance

Rather than assigning unknown classes based on threshold classifier output values, a novel measure of the unknown class was introduced based upon posterior certainty. A case was classed as positive or negative only if its upper and lower bounds did not cross the 0.5 decision threshold. To estimate the success of thresholding based on output variance the 6-node BANN used in Section 4.2.8 was initially used. It was observed that the output variance increased with model complexity, and for models with more than 5 hidden nodes less than 5% of cases lay within one standard deviation of the cut-off point.

A simpler model with 2 hidden nodes was used instead. Incorporating this step into classification increased the accuracy from 60.7% to 64.2%, a marked improvement with 35% of cases being classed as uncertain.

To visualise this variance, the 2-node model was used to produce time series plots of the neural index together with their variance, shown in Figure 4.13.

Alarm systems based on classifier outputs have proven successful at incorporating time-series information into stationary classifiers, by registering the amount of time a classifier spends above a classification threshold, and only signalling an alarm once this time has exceeded a predetermined limit [202]. It is possible to take advantage of the variance produced by the BANN model to include an estimation of certainty on the model output, which could potentially be used to enhance these alarm-based



(a) A case classified as negative

(b) An uncertain result

Figure 4.13: A BANN was used to classify data points from single windows for two sample cases from OCFMT-E. The classifier output and an estimate of its variance are plotted for each case. The final samples can be classified using the samples variance. Case 10319 (a) shows the output signal falling at least one standard deviation from the 0.5 cut-off, hence classed as negative. Case 4401 (b) shows a large standard deviation crossing the cut-off, and would be classified as uncertain.

classifiers.

However for the purposes of a classifier comparison thresholding methods were deemed as not appropriate given the cost to accuracy of using a simpler model and the large number of cases which would be labelled as uncertain in more realistic and poorly separable datasets.

4.3 Experiment 1: The effect of oversampling

CTG datasets are typically highly unbalanced due to the rarity of adverse events. This imbalance can be compensated for by adjusting the misclassification cost of the minor class or in the case of batch trained or bagged methods by selecting balanced samples. These methods are functionally equivalent to oversampling the minor class

and therefore share the same disadvantages; that the model is then highly susceptible to over training the presented samples in the minor class, which will reduce its generalised performance. This is because samples in the major class are likely more evenly spread over the parameter space.

One strategy is to under-sample the major class, however potentially useful information is lost as cases are discarded. Whilst the generalised performance of the model can be increased, the accuracy is diminished.

SMOTE

SMOTE has been introduced to address the problems of oversampling by generating new samples from existing ones. New samples are created from dataset $\mathbf{C}$ consisting of i cases with j features using the following algorithm: where N is the ratio by which

Algorithm 4.2 Create synthetic dataset $\mathbf{C}*$ using SMOTE

```
for n = 1 : N do
  for case c_i in C do
    identify M nearest neighbours using their Euclidean distance.
    select one of these M cases at random, c'
    for parameters d_j = d_1, ..., d_J in case c_i do
      d_j* = d_j + r(d_j - d'_j)
    end for
    return new case c_in* with parameters d_1*, ..., d_J*
  end for
end for
```

the minor class is being over-sampled and r is a uniformly distributed random number between 0 and 1. In this way the oversampled cases are spread throughout the region containing adverse samples, removing the problems caused by highly concentrated samples.

Each of these methods were tested on the OCFMT database to establish an ap-

propriate combination of under-sampling and oversampling/SMOTE to maximise the expected performance of classifiers.

4.3.1 Method

The positive minor class was defined as all 255 cases in the OCFMT dataset with cord gas $pH < 7.05$, the remaining 7,313 cases were considered negative. Each case consisted of a single vector of 64 features calculated from data in the final 30 minutes of labour as in Section 4.4.1. Following the experimental method used by Lawrance et al. [106] these data were divided into 6 overlapping sets containing positive:negative ratios of 1:1, 1:2, 1:4, 1:8, 1:16, and finally all available data. Each ratio was built up sequentially therefore the 1:2 group contains all cases in the 1:1 group, and 255 additional cases sampled from the negative class without replacement and so on for each ratio.

For each of these ratios data were oversampled using either repetition or SMOTE up to equality or 500% oversampling, producing 25 combinations in total. From the complete set of 64 features 6 were selected at random and used to train each stationary classifier (ANN, AB, SVM, RF, RVM). A 9-folded version of the cross validation procedure described in Section 4.4.2 was used to optimise each models hyper-parameters.

The optimised model was used to classify the complete set of 7,568 cases and the accuracy, sensitivity, and Cohen's kappa values achieved recorded. A modified form of 10-fold cross validation was used, where in cases of reduced training data each validation fold was appended to include a single fold of the remaining data. To avoid any bias introduced by feature selection this process was repeated 10 times, using a new randomly selected groups of 6 features in each iteration.

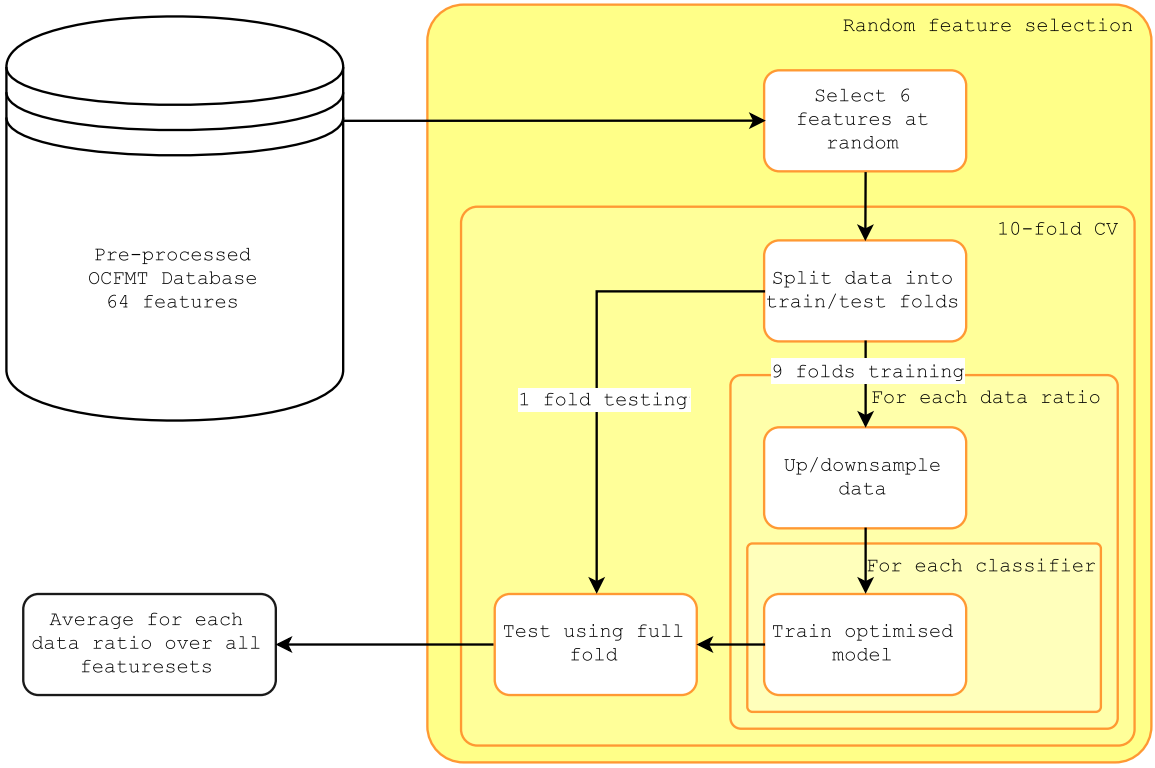


Figure 4.14: The process used to test the effect of data imbalance on classifier performance.

Table 4.3: The average performance of classifiers trained using a subset of OCFMT, under-sampled so that number of cases in each class was equal. Performance in higher ratios is plotted in Figure 4.15

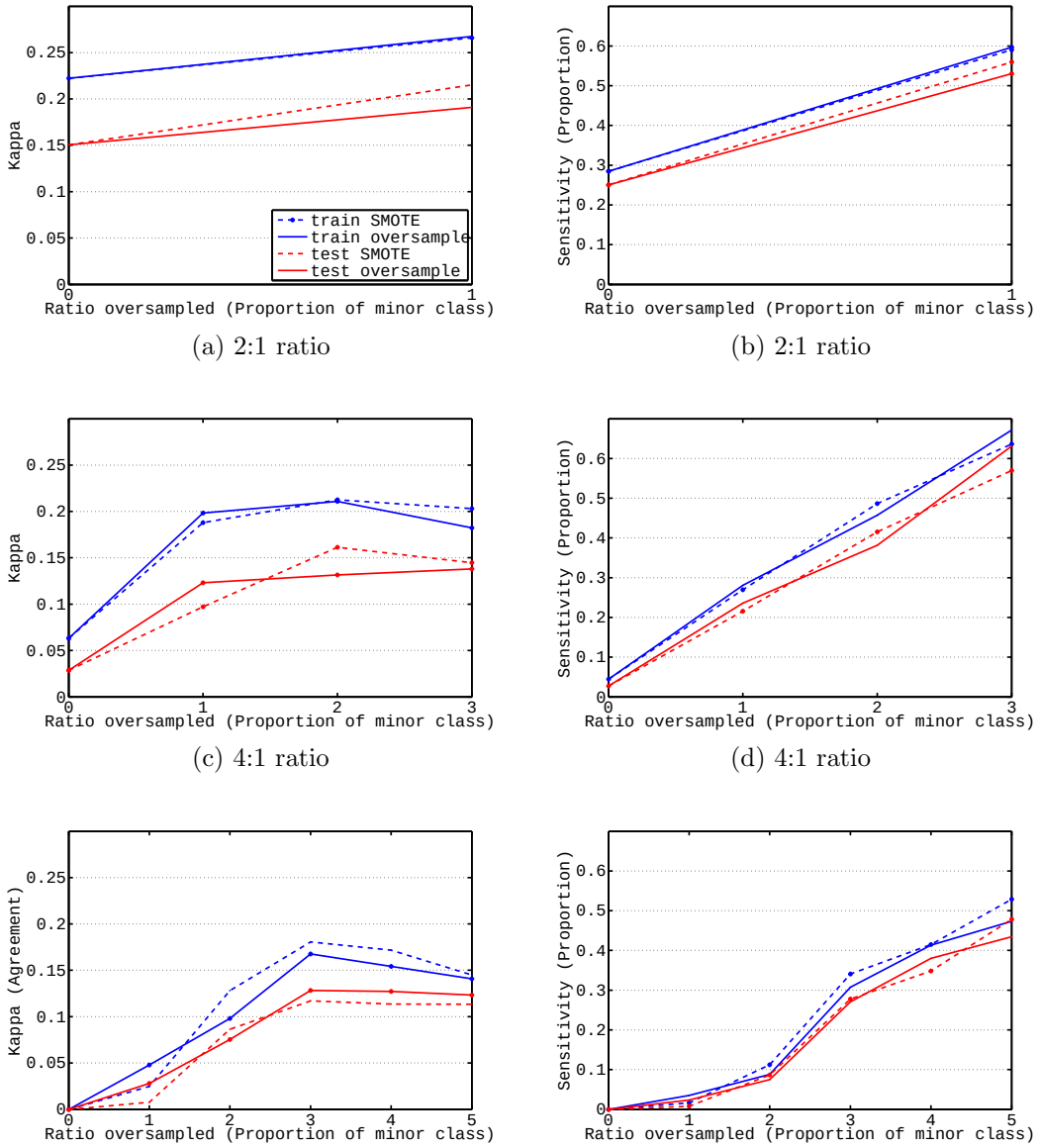
Measure	Train	Test
acc	65.27%	60.46%
sens	55.29%	50.32%
spec	75.25%	70.6%
κ	0.31	0.21

4.3.2 Results

The results using the 1:1 dataset are shown in Table 4.3, and ratios between 1:2 and 1:8 are shown in Figure 4.15. Beyond a class imbalance of 8:1 the sensitivity of the model dropped to below 10% across all classifiers, and the corresponding κ values were close to zero, therefore these results are not included. Given the average model accuracy was over 90% it is apparent that at these levels of imbalance all classes are being labelled as the major class.

The results could also be viewed per classifier to see if there was any distinction. Only the sensitivity was assessed, as the κ values tended to decrease more rapidly due to the increase in data disparity than the increase in sensitivity. Between classifiers the best performance was observed at either 2:1 ratio with 100% oversampling or at 4:1 with 300% oversampling. In both cases there was little difference between the performance using SMOTE and repetition oversampling. The sensitivity plots for the RF and the BANN are shown in Figure 4.16.

Repetition over-sampling was selected for use, being the simplest method to implement. SMOTE did appear to achieve a higher performance at high levels of oversampling and high levels of imbalance. This effect was particularly prominent in the BANN models, which achieved a higher sensitivity with higher levels of imbalance



(a) 2:1 ratio

(b) 2:1 ratio

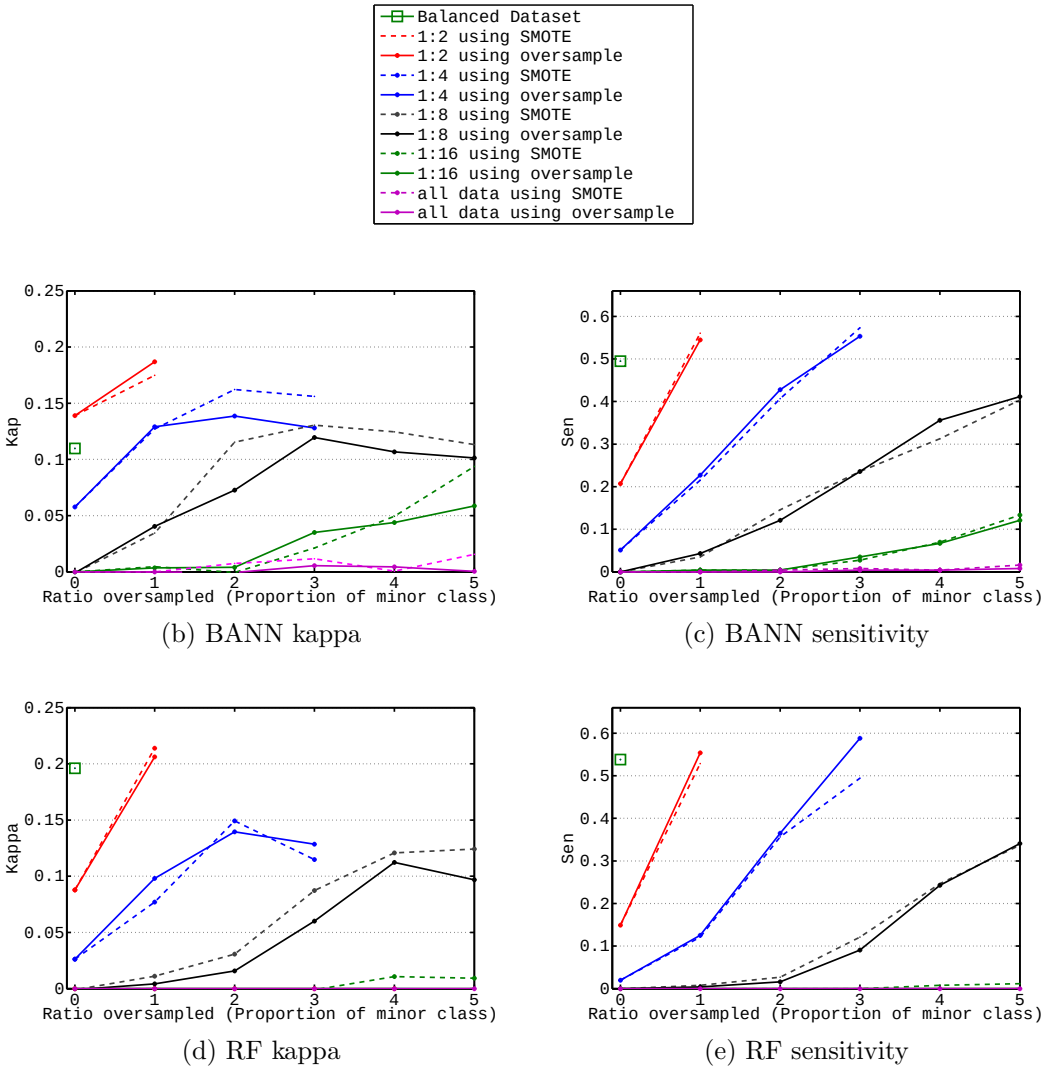
(c) 4:1 ratio

(d) 4:1 ratio

(e) 8:1 ratio

(f) 8:1 ratio

Figure 4.15: The effect of training data class imbalance on the mean κ (left column) and sensitivity (right column) achieved when each classifier was tested using the full OCFMT database. From the top down training sets were down sampled to ratios of 2:1, 4:1, and 8:1. In each plot the x-axis represents the proportion of oversampling, with 0 indicating that no oversampling was used. The maximum test kappa value was found in (a), with a ratio of 2:1 and 100% SMOTE oversampling. The maximum test sensitivity was found in (d), with a ratio of 4:1 and 300% SMOTE oversampling.



(b) BANN kappa

(c) BANN sensitivity

(d) RF kappa

(e) RF sensitivity

Figure 4.16: The effect of over- and under-sampling on the average recorded κ and sensitivity values are demonstrated for the BANN in (a) and (b), and the RF in (c) and (d). SMOTE was observed to make little difference to the performance except in scenarios of high class imbalance, or high ratios of oversampling. The two classifiers respond differently to the data imbalance, however the maximum kappa and sensitivity were attained for the same data imbalance and oversampling ratio as the group average. This suggests that these results can be applied generally to classifiers.

and a larger dataset than in the balanced case. For the general use case a conservative recommendation was taken that in cases of class imbalance the major class can be under-sampled to a ratio of 2:1, and the minor class oversampled by 100%. The maximum average sensitivity was attained at a ratio of 4:1 with 300% oversampling.

4.4 Experiment 2: Classifier performance

In this experiment the performance of 6 classifiers is compared across 3 datasets, which are further divided according to the outcome measure selected. To provide a fair comparison the hyperparameters identified as important in Section 4.2 were optimised using a grid search using the CV method described in Section 4.4.2

4.4.1 Data preparation

In the OCFMT and CTU-UHB databases data were presented as a windowed time-series of features, so only values corresponding to the final 30 minutes of labour, or the final 4 datapoints, were considered. Missing data and outliers were removed by taking the median of these points to give a single data point representing each feature in that super-window. Any cases with fewer than 3 of the final windows intact were considered invalid.

From the complete range of 64 features and 6 clinical parameters a subset of 6 clinically interesting features were chosen based on the current understanding on the fetal physiology [61]. These form *feature set 1.*

1. Signal quality
2. Mean FHR baseline

3. Signal stability index

4. Minimal expected value

5. Number of decelerations

6. Onset slope of decelerations

PCA was used to reduce the 29 UCI features down to 6. Six features were selected so ensuring equal model complexity across datasets.

For models requiring or benefiting from equal scales on input parameters each feature was normalised to have zero mean unit variance. These models were the SVM, ANN, BANN, and RVM. As the RF and AdaBoost models use decision trees these will not be affected by input scaling.

4.4.2 Experimental method

Experimental methodology when comparing classifiers (outside of CTG classification) is not well established, and reviews into the conduct of previous classifier comparisons have found flaws in the methodology in most papers [146]. The major flaws found were no use of a validation set for parameter tuning (in 40 out of 43 papers), and no comparison with the relevant base model in 33% of papers.

Salzberg [155] highlighted the need to perform all model training and adjustment in advance of seeing any test data, and also noted that caution must be taken when repeatedly performing tests on the same databases, as research is both being focused on optimising a single problem which can only guarantee to represent the population from which it was sampled, and that among repeated experiments the *multiplicity effect* causes the probability of significant results arising from chance to continue rising. Therefore all parameter adjustment will be performed here as part of the cross

validation process for each fold.

The choice of outcome measurement may bias the performance of classifiers depending on the parameters that they are optimising, for example ANN are minimising mean square error (MSE) or cross entropy (MXE) so will outperform margin maximisers such as SVM on this measure [22]. It has been demonstrated empirically [21] that different performance metrics will bias the results in favour of different classifiers on different datasets. Thus there is no way to know beforehand which outcomes will be preferred on new datasets. As such measures from several classes (accuracy, probability and ranked methods) will be used here to measure performance.

When comparing multiple classifiers, repeated measures Analysis of Variance (ANOVA) and Friedman tests can be used which maximise statistical power by taking into account the fact that some of the samples are repeated when using cross validation [185]. The Friedman test is a non-parametric equivalent to the repeated measures ANOVA, with slightly lower statistical power but better performance in cases where the assumptions of the ANOVA test, normality of the classifier outputs, is violated. These tests are used to retain or reject the hypothesis that the means of each test are equal.

If this test is rejected, then a post-hoc multiple comparison procedure can be used to test differences between pairs of tests. The Tukey test, or the non-parametric Nemenyi, test are recommended for this [145]. In this particular instance multiple classifiers are being tested over multiple datasets for which Friedman's test is recommended [43].

Based on the literature the following experimental procedure has been devised for comparisons between multiple classifiers on multiple CTG datasets based on the methodologies and suggestions found in literature [43, 145]:
For each dataset

1. Split data into 10 partitions. Retain one partition for testing, and 9 for training.

2. For each classifier:

   (a) Prepare data per classifier, and perform any feature reduction using PCA.

   (b) For each partition:

      i. Down-sample data to correct ratio if necessary.

      ii. Split training data into training and validation sets using 9-fold CV.

      iii. Optimise classifier parameters by maximising relevant performance parameter across all 9 folds. The performance parameter is the outcome measurement used for optimisation.

      iv. Using the discovered parameters, train the model on the entire partition. If the model is unstable, repeat training 10 times.

   (c) Evaluate trained models on testing sets. Report model accuracy, sensitivity, specificity, F-score, and Cohen's κ.

3. Calculate mean of each performance measure.

Use the Friedman test to evaluate differences in accuracy means. If a difference is found, use the Nemenyi test to highlight classifiers which are significantly different from other classifiers. There are several important considerations to highlight in this implementation. Firstly that the testing data are not used to optimise the model parameters meaning that an estimate is being made of the optimal parameter set. Caruana [22] demonstrates on multiple classifiers that using the testing set to optimise model parameters has a positive effect on a model's performance. Secondly the same partitions are used for each classifier here, which reduces any random variance caused by differences in the selected samples.

Table 4.4: The average performance of each classifier on each of the tested datasets using the established training protocol. For each outcome measure, its mean value over the 10 cross validation folds is reported. The values of kappa were used to calculate the best performer. The greatest kappa value achieved on each dataset is highlighted.

Dataset	Measure	AB	ANN	BANN	RF	RVM	SVM
OCFMT-E	acc	0.655	0.655	0.613	0.732	0.750	0.667
	sen	0.667	0.548	0.667	0.667	0.833	0.833
	spec	0.667	0.845	0.667	0.833	0.667	0.583
	kap	0.290	0.323	0.323	**0.452**	0.435	0.355
	fmeas	0.641	0.558	0.615	0.739	0.769	0.732
OCFMT-M	acc	0.658	0.632	0.653	0.671	0.671	0.702
	sen	0.684	0.553	0.711	0.632	0.658	0.711
	spec	0.684	0.675	0.711	0.730	0.675	0.737
	kap	0.316	0.263	0.306	0.342	0.342	**0.405**
	fmeas	0.679	0.612	0.681	0.656	0.665	0.679
OCFMT	acc	0.616	0.602	0.604	0.710	0.567	0.576
	sen	0.538	0.462	0.500	0.480	0.538	0.569
	spec	0.616	0.607	0.603	0.718	0.568	0.576
	kap	0.028	0.014	0.022	**0.048**	0.016	0.026
	fmeas	0.091	0.073	0.084	0.102	0.084	0.086
UCI	acc	0.954	0.954	0.967	0.967	0.959	0.964
	sen	0.917	0.941	0.944	0.944	0.944	0.972
	spec	0.958	0.961	0.970	0.970	0.961	0.964
	kap	0.779	0.764	0.828	**0.832**	0.792	0.824
	fmeas	0.802	0.809	0.850	0.845	0.815	0.825
CTU-pH	acc	0.627	0.682	0.682	0.691	0.645	0.609
	sen	0.667	0.464	0.667	0.667	0.619	0.583
	spec	0.663	0.717	0.694	0.714	0.633	0.612
	kap	0.129	0.094	0.163	**0.194**	0.087	0.082
	fmeas	0.269	0.241	0.267	0.282	0.234	0.249
CTU-BE	acc	0.649	0.591	0.600	0.672	0.541	0.631
	sen	0.500	0.330	0.500	0.375	0.464	0.428
	spec	0.656	0.615	0.600	0.719	0.568	0.667
	kap	0.073	0.016	0.057	0.063	0.010	**0.092**
	fmeas	0.257	0.209	0.239	0.233	0.229	0.255

Table 4.5: The ranking of each classifier according to the median value of the kappa statistic. Ties were handled by assigning each classifier the mean of the tied places.

Dataset	AB	ANN	BANN	RF	RVM	SVM
OCFMT-E	6	4.5	4.5	1	2	3
OCFMT-M	4	6	5	2.5	2.5	1
OCFMT	2	6	4	1	5	3
UCI	5	6	2	1	4	3
CTU-UHB (pH)	3	4	2	1	5	6
CTU-UHB (BE)	2	6	4	3	5	1
Mean	3.67	5.42	3.58	1.58	3.92	2.83

4.4.3 Results

The kappa statistic was found to be unsuitable for the OCFMT database as the large class imbalance resulted in a statistic around 0 for most cases. The same effect can be noted in the SVM validation kappa value, which has not been adjusted to account for the imbalance in datasets presented to the SVM. The ranks of each classifier are presented in table 4.5.

A Friedman test rejected the hypothesis these groups performance was identical with $p = 0.0166$. A post-hoc one-way Nemenyi test was used to assess whether the RF performed significantly better than the other classifiers tested. The critical difference, the difference between mean ranks considered significant, at $p = 0.05$ with 6 degrees of freedom and considering 6 datasets, was 1.392. Therefore the RF can be considered significantly superior to all methods except the SVM, with a difference in mean rank of 1.25 places.

However the lack of independence in these tests must be considered, as one dataset, OCFMT, was divided into multiple subsets, and another, CTU-UHB, was used with multiple outcome measures. With a small number of datasets, the statistical power of

these tests is also low.

The three OCFMT datasets featured identical outcome measures, taken in different ratios. Ranking was most stable between the two balanced datasets, OCFMT-E and -M. In the full dataset class imbalance effects may have had a greater effect on the rankings. The change of ranking and performance between the BE and pH measures observed in the CTU-UHB dataset may be attributed to a change in class ratio (1:6.56 and 1:8.05 respectively), though as this ratio is small it may be assumed that this difference is more likely due to differences in the outcome measure, indicating BE and pH may affect parameters in different ways.

The RF was considered the algorithm most likely to perform well, with the SVM in second. Though the difference between these two classifiers was not a statistically significant this decision was supported by results in the literature on a more general set of databases [49]. The RF algorithm was thus selected for use as a benchmark against which possible improvements to CTG can be compared. The RF was particularly strong in classification of unbalanced datasets, suggesting it produced strong and accurate models of the regions occupied by positive cases. Classifiers with softer decision boundaries, the RVM and BANN, defined by their strong preference for sparse models or small weights, performed poorly in testing on large datasets.

The RF results for each dataset were compared to results in the literature. Dataset OCFMT-E was selected to provide highly separable values. In the literature Georgieva et al. [61] reported an accuracy of 83% with $\kappa = 0.66$ using an ANNC, tested using 10-fold cross validation with a 60/15/25% training / validation / testing split. In preliminary BANN model investigations in Section 4.2.6 using Leave-one-out cross validation accuracies up to 89% were recorded. The results for all classifiers on this dataset were much lower, with the best performing RF achieving an accuracy of 75%

with $\kappa = 0.452$. However in this experiment, unlike results in the literature, input data included only the 6 signal features, and no clinical information.

These two methods were used also to classify OCFMT-M. However in these cases the model trained previously on OCFMT-E was used, with OCFMT-M as the test set. In this case the ANNC reported an accuracy of 63.89%, with $\kappa = 0.28$, and the BANN achieved 64.68%. The RF, achieving 67.10% and $\kappa = 0.342$ compares favourably, however this difference is likely due to the use of 10-fold cross validation, which made a much greater proportion of data available as training information available to the classifier.

On the highly separable UCI database performance was strong for all classifiers despite the class imbalance. In a similar study Ocak et al. [137] achieved a sensitivity and specificity of 100% and 99.3% using a 50/50% train/test ratio and an SVM, with input features optimised using a GA. However the performance on the test set was used as the fitness function, which has been demonstrated to improve performance above methods using validation data alone [22]. An Adaptive-Neuro-Fuzzy-Inference System, a series of ANN incorporating fuzzy rules in the input layer, achieved a sensitivity and specificity of 97.2% and 96.6% [138]. The performance achieved by the benchmark RF algorithm, 94.4% and 97.0%, was not substantially different to these results.

The benchmark method allows a simple comparison against newly developed methods which is easily implemented and robust to outcome measure and class imbalance.

4.5 Experiment 3: Inclusions of time series information

It has been suggested that the most important aspect of CTG assessment is the evolution of features over time [199]. Due to the large variability between individuals, absolute values of CTG properties may not be appropriate, but trends may indicate pathological conditions. Several sources have looked into the trajectory of CTG features, using visualisation techniques [24] and the evolution of properties from systems identification of the CTG [190], but the performance of a classifier using time-series information has not been compared against those using absolute information. In this Section an HMM is designed and compared against the previously developed benchmark RF classifier. The HMM was selected as it has previously been applied successfully to CTG data [69].

4.5.1 HMM

The structure of the HMM is described by the states and transitions between states, with each state having a set of output probabilities called emissions. An example of a set of three states and their associated transitions for two model types is shown in Figure 3.1. The 1^{st} order Markov assumption is used here, stating that the probability of transition to next state is dependent only on the current state.

The HMM consists of a set of N states $S_1, S_2, ...S_N$, connected by state transitions. The state at time t is denoted as q_t, and the transition probability from state i to j can then be written $\alpha_{ij} = P(q_{t+1} = S_j | q_t = S_i)$. The complete set of transitions can be written in matrix form, called the transition matrix, A = $\{a_{ij}\}$.

At each state an observation is emitted, which is the only information visible to the system, hence the current state of the system as any point is 'hidden'. In the discrete case with a set of M observation symbols $V = \{v_1, v_2, ..., v_M\}$, the probability of observing symbol v_k given state S_j is written as $b_j(k) = P(v_k|q_t = S_j)$. The complete set of observation probabilities is termed the emission matrix $B = b_j(k)$. The final descriptor of the model is the initial probability vector π, the probability of starting in state i, or equivalently $\pi_i = P(q_1 = i)$. The complete set of probabilities is called $\pi = \{\pi_i\}$. These three terms are often put together to describe the complete HMM model as $\lambda = (A, B, \pi)$. To classify samples an HMM must be made for each potential class. Samples are assigned to the class which is most likely to have produced that sample.

An optimal HMM is one which maximises the probability of the training observations. There is no way analytically to solve the maximisation problem, but several alternative local optimisers have been proposed, the two major choices being the Segmental K-means Algorithm [94] and Baum-Welch re-estimation [147]. Baum-Welch re-estimation is used in this Thesis so is explained here. Firstly two concepts, the forward and backward variables, must be introduced. The forward variable $\alpha_t(q)$ is defined as the probability of the sequence up to time t finishing in state q, and is found by iterating through the observed sequence:

$$\alpha_1(q) = \pi_q b_q(O_1)$$

then for $2 \leq t < T$:

$$\alpha_{t+1}(j) = \left[\sum_{q=1}^{N} \alpha_t(q) a_{ij}\right] b_j(O_{t+1}) \tag{4.9}$$

The backward variable β represents the probability of a sequence of observations given that the current model is in state i at time t. It is found by iterating through the observed sequence, but backwards. Start at $t = T$

$$\beta_T(i) = 1\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots, O_T | i_t = i, \lambda) \tag{4.10}$$

Then a new variable is defined, $\gamma_t(i) = P(i_t = i|O, \lambda)$, the probability of being in state i at time t given the observation sequence and the model. This can be found using the forward and backward variables defined in equations 4.9 and 4.10:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}$$

A second new variable is defined, $\xi_t(i, j) = P(i_t = i, i_{t+1} = j, O_{t+1}|\gamma)$, the probability of moving from state i to state j and emitting observation O_{t+1} at time t. This is again found using the forward and backward parameters:

$$\xi_t(i, j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$

These variables can then be used to represent the following values:

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{Expected no. transitions from state i}$$
$$\sum_{t=1}^{T-1} \xi_t(i, j) = \text{Expected no. transitions from state i to state j}$$

which can be used to update the model parameters as follows:

1. The initial starting probability for state i is the normalised number of times we

expect to be in state i.

$$\pi_i' = \gamma_t(i)$$

2. The transition probability from state i to state j is the number of transitions from state i to state j divided by the number of transitions from state i.

$$a_{ij}' = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

3. The emission probability, $b_j(k)$ is the number of times the model emits observation k from state i over the number of times the model is in state i.

$$b_j(k)' = \frac{\sum_{t=1,O_t=k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)}$$

The complete set of re-estimation formulae can be written as:

$$\begin{aligned} \pi_i' &= \gamma_t(i) \\ a_{ij}' &= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \\ b_j(k)' &= \frac{\sum_{t=1,O_t=k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)} \end{aligned}$$

This algorithm is modified to accept continuous data by replacing the discrete observation probabilities with a parametric continuous distribution, in this case a normal distribution. Therefore it is assumed that input parameters also share this distribution. Using a continuous model introduces additional parameters; in this case - a mixture of M normal distributions - these are a vector of means, μ_{jm}, a covariance matrix U_{jm} and the mixture gains c_{jm}. The re-estimation formulae for these have been shown to

be [110, 92, 91]:

$$c'_{jk} = \frac{\sum_{t=1}^{T-1} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{4.11}$$

$$\mu'_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k).O_t}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{4.12}$$

$$U'_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k).(O_t - \mu_{jk})(O_t - \mu_{jk})'}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{4.13}$$

4.5.2 Data preparation

Data from OCFMT-M were prepared by normalising each feature to zero mean and unit variance. Missing data were replaced with the feature mean values, conveniently 0 after normalisation. From each case the final 20 points were selected, representing the final 110 minutes of labour, taken as close to birth as possible. This number of samples was selected to maximise the stability of the distance between the two class models, demonstrated by Dugad & Desai [44], whilst retaining the maximum amount of usable cases.

HMM were trained using Baum-Welch re-estimation, implemented using MATLAB with the HMM Toolbox [122]. In these models the emission probabilities, b, are assumed to be continuous observations with a normal distribution:

$$b_j = N(\mu_j, \mathbf{U_j})$$

for state j. Based on the recommendations in the literature the means of the emission parameters were set to the values of the parameters of random samples taken from the training set, and each element of the emission covariance matrix was set to 0.1. The transition matrix and prior state probabilities were initialised uniformly according to

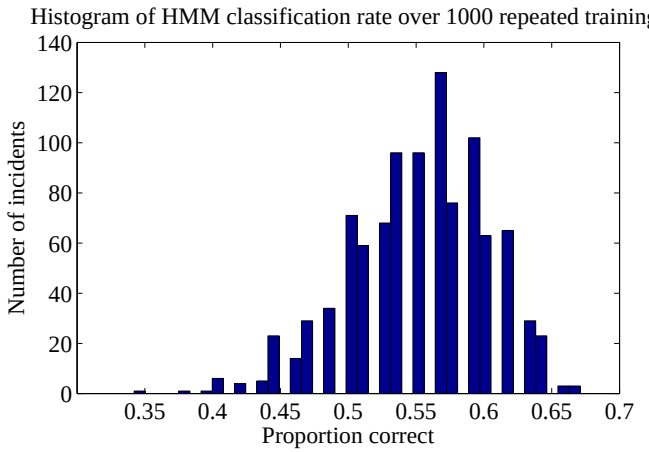


Figure 4.17: A histogram of the validation performance of an HMM using a 3-state ergodic model over 1000 iteration with random parameter initialisations. In each iteration training and testing sets were randomly selected in an 80/20% ratio. The experiment was performed using feature set 1 and the OCFMT-M dataset. The best performer achieved an accuracy of 67.05%, and the overall mean accuracy was 55.31%.

the number of states, representing the fact there was no prior knowledge about the model.

The Baum-Welch re-estimation procedure used to train these parameters is a local optimisation procedure, and was found experimentally to be highly sensitive to the initialisation of the model, demonstrated in Figure 4.17. Based on these results it was decided that selecting the best performer from 100 randomly initialised iterations of model training would be sufficient to discover a near-global optimum.

4.5.3 Feature selection

In addition to a comparison made using feature set 1 a second feature set was constructed from features which display high performance on time series. From clinical studies performed by Westgate et al. [199] it can be inferred that features chosen to perform well as absolute values will not necessarily perform identically well as time-

series.

To assess the performance individual features, each was used as the input to a simple 3-state left-right HMM. Additionally a feature was added consisting of normally distributed randomly generated numbers with identical statistical properties for each class. This feature allows each test feature to be compared to one with no discriminatory information, removing any misleading biases introduced by the training processes. Each HMM pair was trained according to the established protocol. The results for all 100 training iterations are shown in Figure 4.18, with the complete table of the features tested listed in Appendix A. The top six performing features based on the mean result, which are labelled *feature set 2*, were:

1. Feature 50: Mean of local approximate entropy
2. Feature 44: Number of variable decelerations
3. Feature 4: Signal stability index
4. Feature 61: Phase rectified signal averaging - decelerative capacity
5. Feature 47: Deceleration/contraction time ratio
6. Feature 62: Bi-variate phase rectified signal averaging - standard deviation of accelerative capacity

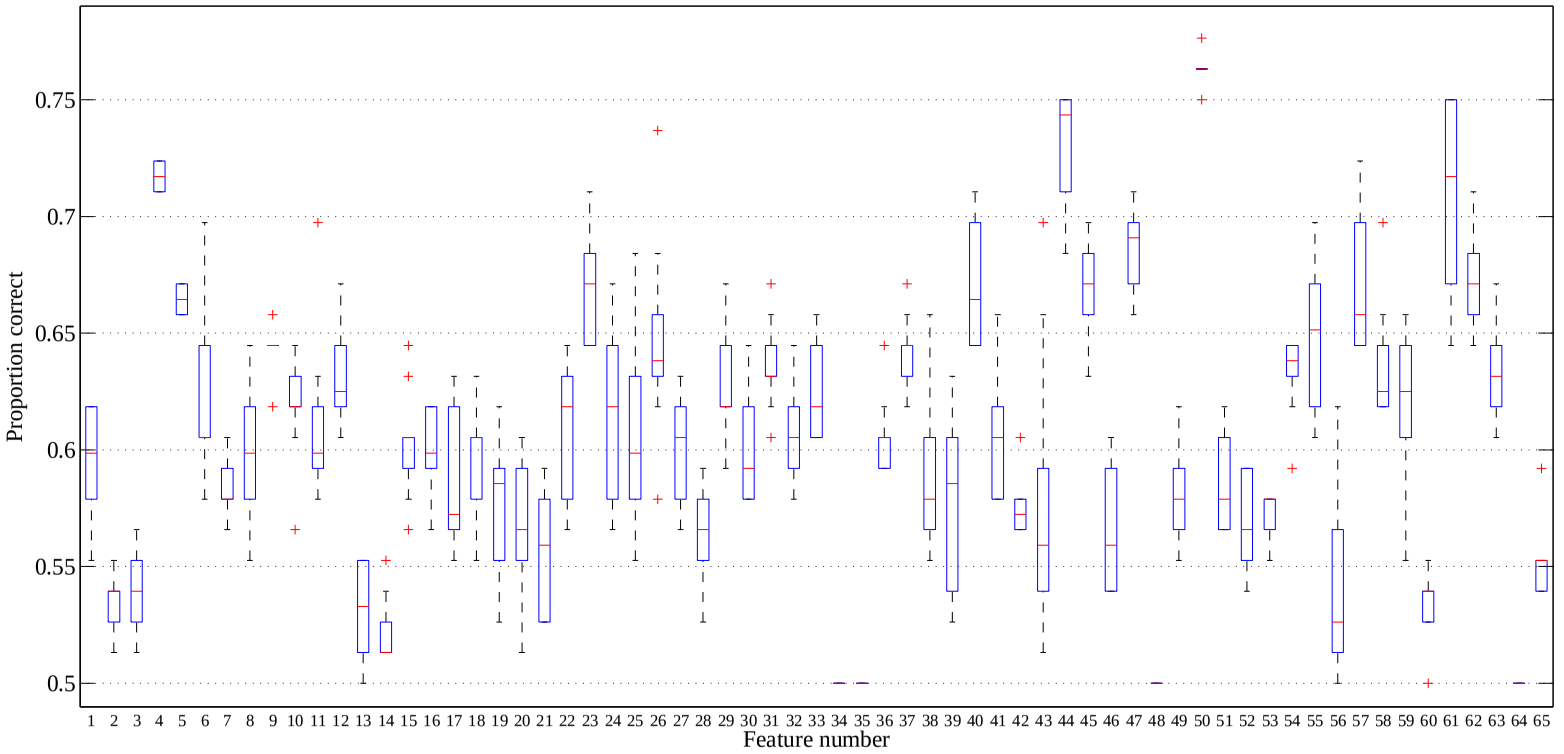


Figure 4.18: The HMM training protocol was repeated 10 times, using the 3 state left-right model with a single input feature. The resulting classification performance of the individual features is shown on a box plot.

Several features, including a test feature of random numbers scored 50% uniformly on all runs, thus being perfectly indiscriminate. This can be a result of one model dominating by having a much larger output variance, making a large range of outputs highly likely, resulting in all test cases being classified to the same class. Whilst this can be avoided with careful training or by assigning a penalty function to the model outputs, this is most often a result of poor separability of input data.

Compared to the original set of features used by Georgieva [61] the number of decelerations (feature 23), performed well with a maximum correct proportion of 0.71. However the number of variable decelerations (feature 44) had both a higher mean and a smaller distribution, and was the second best performing feature overall. The onset slope (feature 36) suggested by Westgate [199] as an important time-series feature performed averagely in testing. Interestingly the median recovery slope (feature 37) was of more importance. The best performing feature, mean of local approximate entropy had already demonstrated strong performance previously in the classification of absolute values [55].

To assess the performance increase through using an optimal set of features the training and testing procedure was repeated using the 6 best performers from the previous feature assessment step. The correct classification rate rose from 64.29% to 66.27%, although the sensitivity and specificity were less consistent, 55.56% and 80.95% respectively. Several features had a very narrow distribution, indicating that the training protocol was effective in discovering a narrow grouping around the global optimum, others had a much wider distribution, indicating a highly complex optimisation surface, suggesting that the training protocol might need to be revised if these are to be included.

4.5.4 Model selection

With an optimal feature set established, the structure of the HMM must be decided on. A hurdle in statistical classification using fetal CTG data encountered by all groups to date has been the limited number of cases, due to the rarity of adverse conditions, thus a model with fewer parameters is preferred. An HMM with an output probability density function (PDF) defined by a mixture of Gaussians with Q states, O features and M mixtures will have $Q + Q^2$ structural parameters for the prior and transfer matrix; and $(M + OQM + O^2QM)$ emission parameters for the mixture weights, means, and covariance matrices respectively. Some parameters can be tied, reducing the degrees of freedom, noting that the structural and mixture weight parameters are bound by stochastic constraints.

A left-right model with 7 hidden states has been found to achieve the highest classification rate with a balance between sensitivity and specificity [69]. A left-right transition structure was chosen for two reasons: firstly, the initial state priors are set as the model always begins in the first state, and the number of parameters in the transfer matrix is reduced; secondly, exploration of the data demonstrate that they are non-stationary, and would therefore be poorly modelled by an ergodic system.

To discover an optimal model structure for the OCFMT database, models with between 3-8 states with a left-right, or Bakis structure were tested using the OCFMT-M database and feature set 2 with no further dimensionality reduction. The accuracy, sensitivity and specificity for each structure are shown in Figure 4.19.

The most general structure, with three states, performed best in classification, achieving 77.63% correct classification and the 5-state model achieved the greatest geometric mean of sensitivity and specificity.

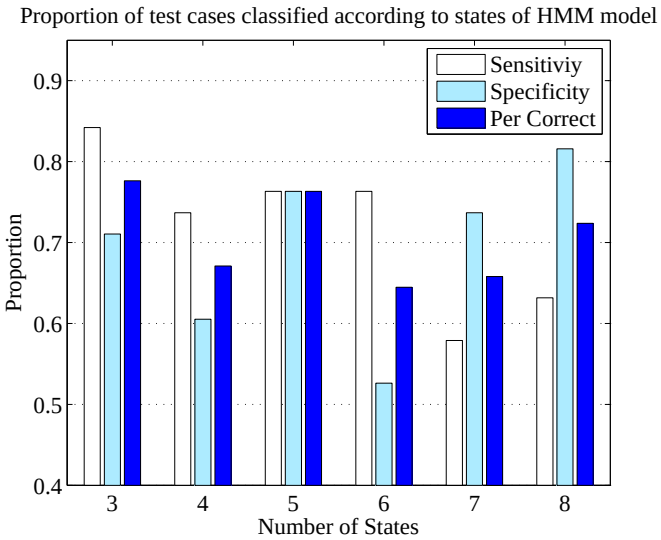
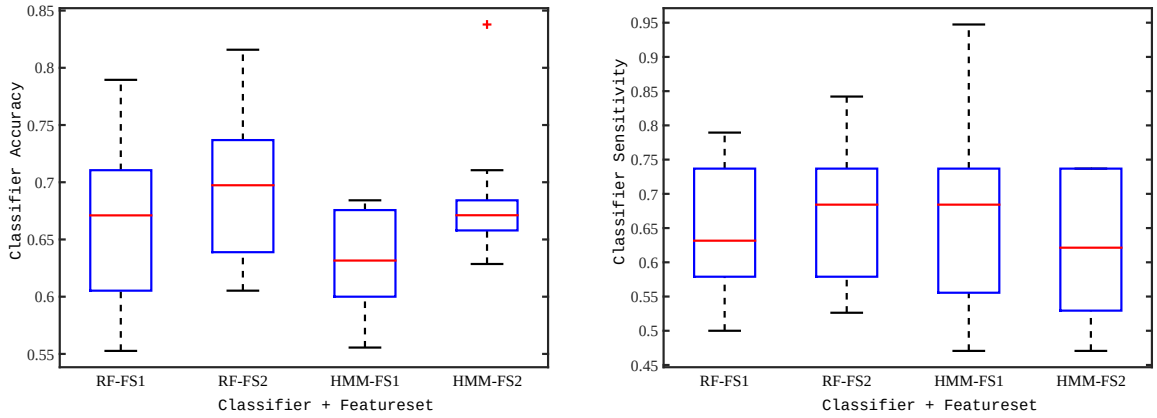


Figure 4.19: The Performance of a left-right HMM with 3-8 states was tested on the OCFMT database. A model with 3-states achieved the greatest classification accuracy, though the 5-state model showed a better balance between sensitivity and specificity.

4.5.5 Results

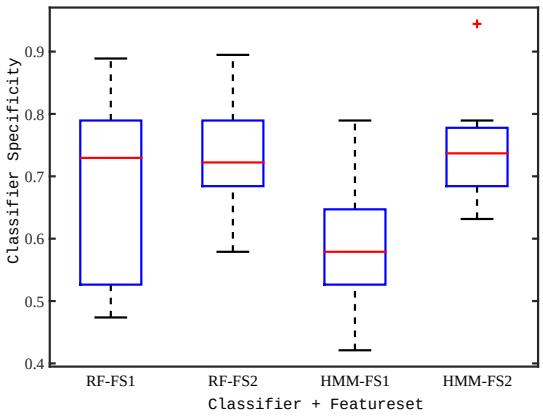
The HMM method was compared against the benchmark RF classifier using the earlier established training and testing process using both the original set of clinical features FS1, and the new set of features selecting in Section 4.5.3. The results are plotted in Figure 4.20. The performance was not significantly different from the RF performance at the $p = 0.05$ level, and the κ statistic significantly higher (one-tailed Wilcoxon rank sum $p = 0.0246$) using the new feature set. The RF also performed better using the new feature set, but not significantly so.

One interesting observation is the reduction in the variance of the classifier performance across the 10 folds, noting identical subsets were used for both classifiers. This result suggests that the output is more reliable when using time-series information, but that the overall performance does not increase. The information from the remainder of the signal can be then thought of as supporting evidence for the final segments.

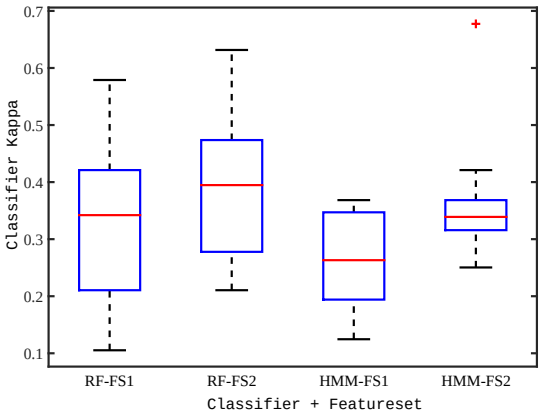


(a) Accuracy

(b) Sensitivity



(c) Specificity



(d) Kappa

Figure 4.20: The HMM is compared to the benchmark RF classifier using OCFMT-M with 10-fold cross validation using both the heuristic and time-series optimal feature sets. The accuracy of the benchmark RF model was greater for both feature sets. In (c) the time-series features significantly improved the HMM specificity. The variation in HMM accuracy, specificity, and kappa were much lower using than the RF model when using the time-series feature set.

4.6 Conclusions

A training procedure has been developed which is simple to implement and reproduce based upon recommendations in the literature. This includes a procedure to handle class imbalance in datasets and a simple method of selecting model hyper-parameters. Six classifiers were evaluated to select a strongest performer across multiple outcome measures and datasets. The RF ranking significantly higher than all methods except the SVM when evaluating performance using the κ statistic, a result which agrees with literature on classifier comparisons.

The classifiers were observed to behave differently in response to class imbalance and the oversampling methods used to correct this imbalance; in particular methods with parameter defined regularisation parameters were sensitive to the combination of oversampling and automatic parameter optimisation. It was also noted that each classifier was expected to perform well in different contexts, and though each was constrained to the same amount of input information and given the freedom to optimise parameter numbers some, for example the RF, are able to utilise large parameter sets in ways others cannot which was not accounted for in this comparison.

Interestingly in Table 4.4 the RVM performed well on the highly separable OCFMT-E dataset where other classifiers achieved poor performance in comparison to previous results in the literature. The BANN performed well on the highly separable but unbalanced UCI dataset. The performance of the RF on the unbalanced and poorly separable datasets, in particular OCFMT exceeded the other classifiers greatly. The cause of these differences may be due to the training procedure, or the type of decision surface produced, however the RF method was consistent across all datasets and outcome measures.

The main impact of including time series information appears to be decreasing the variance in the outputs, with the classification performance of the HMM being more consistent across the 10 identical folds used for training and classification than the RF benchmark. This suggests that additional information contained within the time series may be used to support decisions, but that the absolute performance of a classifier may not be improved using time series information. The method may be impacted by the number of missing windows in signals, which were replaced with mean parameter values in this thesis.

Methods which used confidence to improve accuracy by choosing not to classify some cases were able to increase performance, and can be tuned to adjust the classifier output to prefer a desired outcome in a manner similar to moving through the ROC curve. Output variance measures demonstrated a high sensitivity to the model parameters, thus a meaningful confidence bound could not be implemented without saturating the classifier output. Using a small - 0.1σ - confidence band, classification performance was improved by 3.5%, however the cost of leaving 35% of cases unclassified is considered too high for this marginal improvement.

Chapter 5

System Identification to Enhance Classification Performance

With the performance of classifiers seemingly approaching a maxima, and no realisable improvements found using time-series information or thresholding new avenues are explored. System Identification (SI) methods allow for the inclusion of current physiological knowledge about the fetus and its reponse to distress, and can potentially provide insight into the process of distress during labour.

In this Chapter SI methods which have been previously demonstrated on CTG data are used to explore the OCFMT database.

5.1 The Impulse response function

Though the CTG signal is not stationary the rate of change is slow, developing over the course of hours. As such the model can be assumed to be quasi-stationary in short windows. Non-linear models can capture higher degrees of complexity but are less stable and hence more susceptible to noise. Given the focus on very low frequency modelling and the high degree of noise a linear model is appropriate. The IRF model is a simple stationary model utilising only the model input and its history to make predictions on the current output.

A general IRF predicting the output at time t with a model order M is expressed mathematically as:

$$y(t) = \sum_{m=1}^{M} h_m u(t-m)$$

where $\mathbf{h} = [h_1, \ldots, h_M]$ is a vector of model coefficients and $u(t)$ or u_t is the input at time t. In vector form for an entire time series this becomes:

$$\mathbf{y} = \mathbf{U^T h}$$

with the regressor matrix, $\mathbf{U}$ constructed from the input thus:

$$\begin{matrix} u_1 & 0 & \dots & 0 \\ u_2 & u_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_N & u_{N-1} & \dots & u_{N-M-1} \end{matrix} \tag{5.1}$$

Assuming there are fewer model coefficients than data points the resulting system is overdetermined, and can be solved using Ordinary Least Squares (OLS). The OLS solution aims to minimise the variance of the errors between the predicted output and the true output:

$$V(h) = \frac{1}{N}\sum_{t=1}^{N}(y_n - \mathbf{u_n^T h})^2$$

which in matrix form is:

$$= (\mathbf{y} - \mathbf{U^T h})^T(\mathbf{y} - \mathbf{U^T h}) \tag{5.2}$$

where $\mathbf{u_n}$ are the regressors of the output at time n. Based upon this error measure an optimum set of parameters is the solution to $\mathbf{h_{OLS}} = arg\min V(h)$. Equation 5.2 can be differentiated analytically to give the OLS solution:

$$\mathbf{h} = (\mathbf{U^T U})^{-1}\mathbf{U^T y}.$$

The accuracy of the fitted model is measured using the percent variance accounted for (VAF), which is calculated using the error of the signal $\mathbf{e} = \mathbf{y} - \mathbf{Uh}$. The variance of

the error signal, σ_e^2, is normalised by the variance of the true output signal, σ_y^2:

$$VAF = 100\left(1 - \frac{\sigma_e^2}{\sigma_y^2}\right).$$

Properties of the IRF

The IRF can be described in terms of several properties shown in Figure 5.1. These are:

1. *Delay* - the time until the first non zero coefficient. The represents a delay in the response to any input.
2. *Memory* - the length of time from the first to last non-zero coefficients. This represents the number of input samples which are used to produce a single output estimate.
3. *Max coefficient time* - the time until the coefficient with the absolute largest value. This represents the time delay of the most influential input value.
4. *Total variation* - The sum of the absolute differences between coefficients. An approximation to the frequency content of the IRF.

$$TV = \sum_{i=1}^{M-1} |h_i - h_{i-1}|$$

5. *Gain* - The sum of the IRF coefficients. Measures the amplitude and direction of the response with respect to the size of the input.
6. *Maximum coefficient* - Absolute value of the largest coefficient. A single measure of the shape.

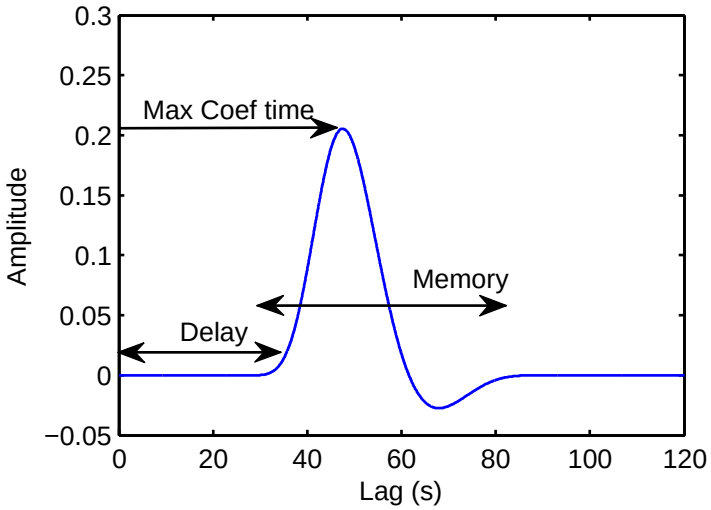


Figure 5.1: The major properties of the impulse response function. The recovery is defined as the length of the delay plus the memory or model order.

The Pseudo-Inverse

The impact of noise on the discovered IRF can be mitigated by using the Pseudo-Inverse (PI) to split the model into components and selecting the components least affected by model noise [195]. In this way the expected IRF response, smooth due to the low frequency response of the system, can be separated from the raw IRF discovered.

A model with output $\mathbf{y}$ and corrupted by noise $\mathbf{v}$ produces a noisy output $\mathbf{z} = \mathbf{y} + \mathbf{v}$; the output $\mathbf{y}$ being approximated by a transfer function $\mathbf{h}$ to give $\mathbf{z} = \mathbf{U}\mathbf{h} + \mathbf{v}$, and with the OLS estimate $\hat{\mathbf{h}}$. The Hessian matrix $\mathbf{H} = \mathbf{U}^{\mathbf{T}}\mathbf{U}$ is introduced, which has Singular Value Decomposition (SVD):

$$\mathbf{H} = \mathbf{V}\mathbf{S}\mathbf{V}^{\mathbf{T}}$$

$$\mathbf{H}^{-\mathbf{1}} = \mathbf{V}\mathbf{S}^{-\mathbf{1}}\mathbf{V}^{\mathbf{T}}$$

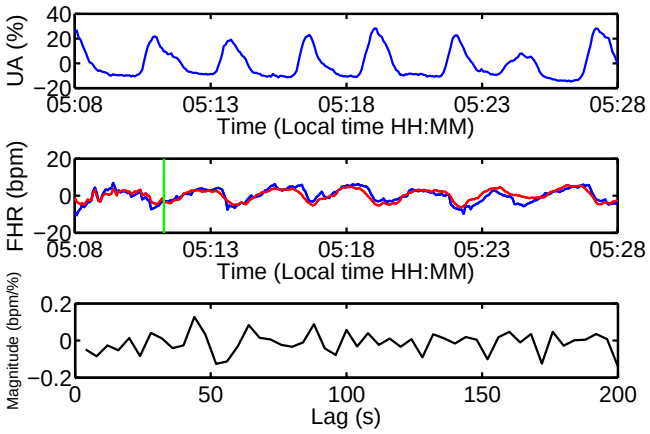


Figure 5.2: The incoming UA and FHR data, and the estimate of an IRF of order 50 discovered using OLS are shown in the top, middle, and bottom plots respectively. The red line shows the FHR estimate, with the green line representing the relative length of the IRF. The VAF for this segment was 67.7%

The original noise equation is pre-multiplied by $\mathbf{U^T}$ and two new variables are introduced, $\zeta = \mathbf{V^T h}$, the projections of the IRF onto the eigenvectors of the Hessian of the input, and $\eta = \mathbf{V^T}(\mathbf{U^T e})$, the projection of the input noise cross correlation onto the eigenvectors of the Hessian of the input:

$$\begin{aligned}\mathbf{U^T y} &= \mathbf{U^T U h} + \mathbf{U^T e} \\ &= \mathbf{VS}\zeta + \mathbf{V}\eta.\end{aligned}$$

The transfer function itself is then decomposed into M singular values.

$$\begin{aligned}\mathbf{h} &= \mathbf{V}\mathbf{S}^{-\mathbf{1}}\mathbf{V}^{\mathbf{T}}(\mathbf{V}\mathbf{S}\zeta + \mathbf{V}\eta) \\ &= \mathbf{V}\zeta + \mathbf{V}\mathbf{S}^{-\mathbf{1}}\eta \\ &= \sum_{i=1}^{M}(\zeta_i + \frac{\eta_i}{s_i})\mathbf{v_i}\end{aligned}$$

Here η is the sole term present involving the model noise. This noise is inversely scaled by the size of the eigenvalue s_i, hence the signal to noise ratio can be improved by ignoring small eigenvalues, or by retaining only the most significant singular values. The appropriate number of singular values to retain can be determined using Minimum Description Length (MDL). This combines requirements for simplicity and accuracy in model selection. The MDL of the IRF retaining m singular values is:

$$MDL(m) = \left[1 + \frac{m * P_c \log(N)}{N}\right] \sum_{i=1}^{N} [\mathbf{y}(i) - \hat{\mathbf{y}}]^2 \tag{5.3}$$

The term P_c is not normally included in the MDL equation, and is a penalty term introduced in the literature to tune the prioritisation of simpler models over more accurate models. This was set experimentally to $P_c = 4$. This operation can be computed efficiently by using the energy of the error, replacing the summation in

equation 5.3 with the energy $\xi_e = \xi_y - \xi_{\hat{y}}$, and noting that $\xi_y = \sigma_y^2$ [192].

$$\begin{aligned}\xi_{\hat{y}} &= \frac{1}{N}\sum_{t=1}^{N}\hat{\mathbf{y}}^2(t) \\ &= \hat{\mathbf{h}}^{\mathbf{T}}\phi_{\mathbf{uu}}\hat{\mathbf{h}} \\ &= \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{V}\mathbf{S}\mathbf{V}^{\mathbf{T}}\hat{\mathbf{h}} \\ &= \sum_{t=1}^{M} s_t(\mathbf{v_i^T}\hat{\mathbf{h}})^2\end{aligned}$$

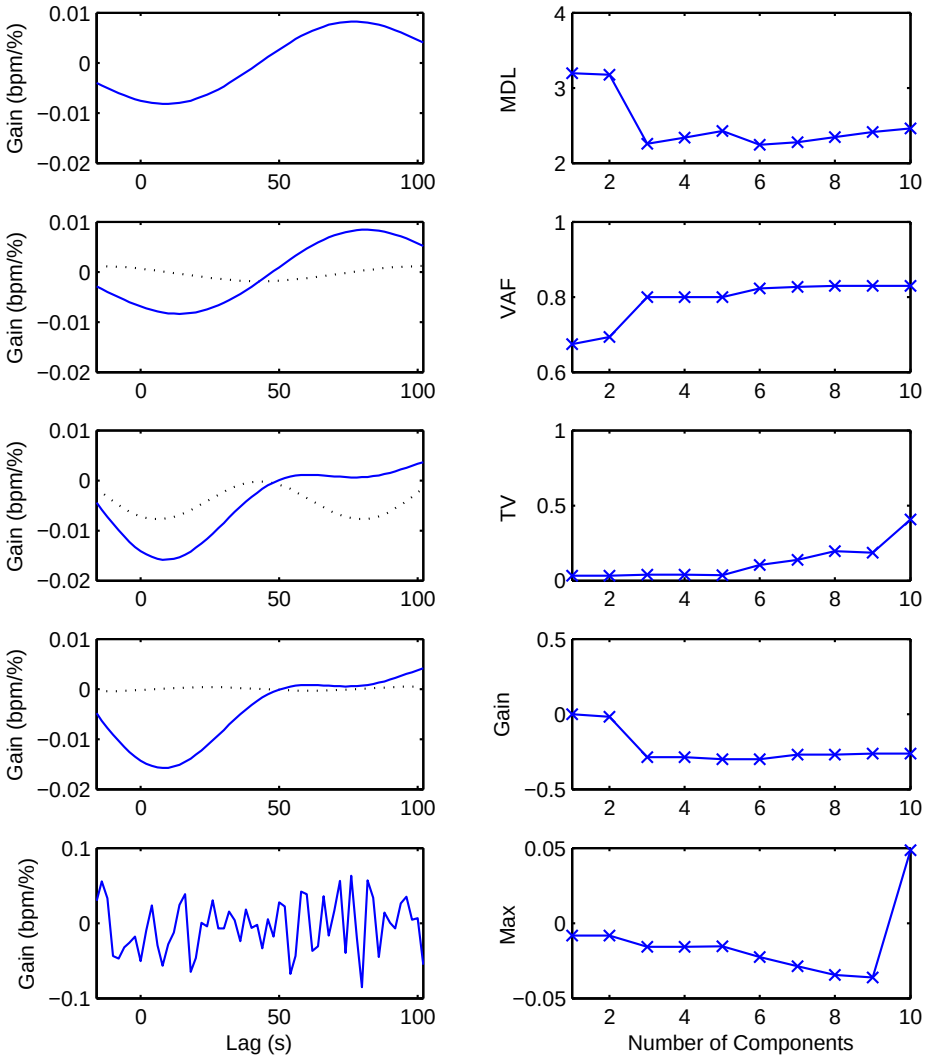
The equation for calculating the MDL is then revised to:

$$MDL(m) = \left[1 + \frac{m * P_c \log(N)}{N}\right]\left[\sigma_y^2 - \sum_{i=1}^{m} s_i(\mathbf{v_i^T}\hat{\mathbf{h}})^2\right]$$

The impact of using the PI on IRF shape on a sample of real CTG data is shown in Figure 5.3. The final figure in subplot (a) shows the original IRF discovered, which includes all components. The IRFs formed with fewer components are smoothened, at the cost of losing some information. The information lost is minimal however, as the difference in VAF in subplot (b) between the selected model and the complete model is $< 1\%$.

5.2 An algorithm for IRF identification

To implement the OLS + PI method the model parameters must be selected. The first challenge is the selection of the appropriate order for the model, termed the memory; the second is the identification of the estimated delay between the UA signal and the FHR response. To select these two parameters a method incorporating heuristic knowledge of the expected IRFs into the identification algorithm was developed by



(a) IRF Components

(b) IRF Properties

Figure 5.3: From the top down on the left are shown the four most significant components of a sample IRF, with the original noisy IRF at the bottom. Each selected component is shown individually (dotted line) together with the cumulative result (solid blue line). Figure (b) shows how the properties of the IRF vary with the number of components retained. The gain and VAF are low until the third component, at which point the MDL also reachs a nadir. Beyond 5 components the TV begins to rise, indicating an increase in noise.

Warrick et al. [191, 192, 193, 194, 189, 195]. The complete process is listed here to provide a reference detailing each step of the entire algorithm, and will be explained in the subsequent sections. Differences between the data used and method implemented in this Thesis and that implemented by Warrick et al. are highlighted in Section 5.2.5.

- Section 5.2.1 - Preprocess the data:
  - Downsample to 0.25Hz.
  - Remove artefacts.
  - Repair short segments of missing data.
  - High-pass filter to remove the baseline.
- Section 5.2.2 - Divide signal into equally sized and spaced windows:
  - Identify the maximum and minimum possible delays for window.
  - Remove windows failing a validity check.
- Section 5.2.3 - Discover IRFs in each window:
  - Find the dominant UA frequency f_{maxU} and set $1/f_{maxU}$ to be the initial estimate of memory.
  - With this memory held constant, create a set of candidate delay values by finding values where the first IRF coefficient is 0.
  - For each candidate delay, the memory is varied until the final IRF coefficient is 0. A new set of candidates is formed from all possible memory/delay pairs.
  - Select the candidate with the lowest MDL.
- Section 5.2.4 - Postprocess the candidates to ensure consistency across all windows:

- Remove IRFs which fail significance checks.
- Average the delay and recovery across all windows.
- In each window, select the candidate most like its neighbours.
- Perform a local delay and recovery search around this candidate to optimise the IRF.

5.2.1 Pre-processing

Incoming data from the OCFMT database contained FHR data at 4Hz and UA data at 0.25Hz. Any data that were missed during recording were given the value inf, and were almost entirely missed beats in the FHR signal. The FHR signal was downsampled to 0.25Hz using the inbuilt Matlab function decimate to remove high frequency components. The function includes a step to low pass filter the data to avoid aliasing effects.

Two kinds of artefacts were identified for removal. The first were segments of signal drop off, detected as segments of low variability. A segment of data was deemed as having low variability if its variance in a 15 second window dropped below $0.1bpm$ in the FHR trace or 0.1% in the UA trace. Invalid segments were then extended until the signal crossed the local baseline.

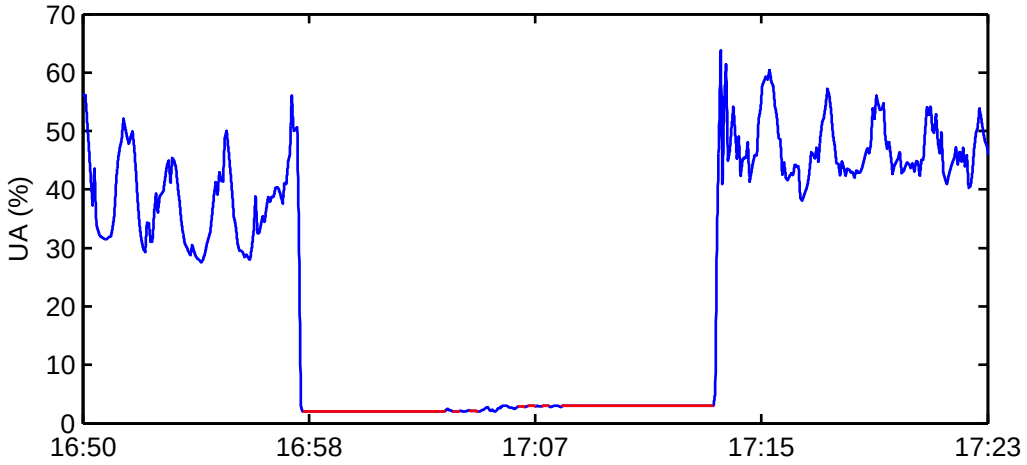
The second artefacts were short abrupt changes identified using the spike detector developed by Dawes et al. [39]. An upward change in FHR of $\geq 30bpm$ which returns to the local baseline faster than the minimum accepted acceleration length of 30 seconds, or a downward change of $\geq 30bpm$ which return faster than the minimum deceleration time, also set at 30 seconds, is labelled as a spike artefact. This detector has been developed on this database, and will not remove physiologically feasible signals from

the sample.

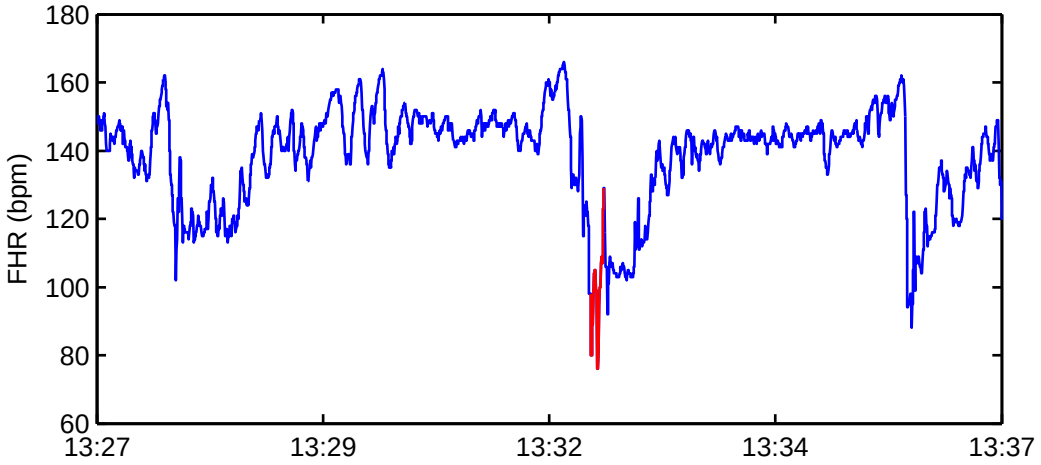
Data that were either labelled as missing in the acquisition stage or were identified as an artefact were combined under a single missing data label. To maximise the number of valid segments, attempts were made to repair the missing data by bridging small gaps. For each gap found in the data, if the gap was less than 15 seconds the missing data were bridged using linear interpolation. Gaps larger than 15 seconds were ignored, and left as missing data, which would cause the entire segment to be classified as invalid. The low variability detection and data repair steps are demonstrated in figure 5.4. The data were then filtered with a high pass filter to remove low frequency information and the signal baseline. Prior to filtering, any missing data were temporarily filled by the local signal mean to prevent the propagation of missing values. A Chebychev Type II filter was designed using Matlab, with a cut-off frequency of $f_c = 45mHz$. This second order filter was applied using the Matlab inbuilt function filtfilt to prevent signal phase distortion.

5.2.2 Signal segmentation

These signals were then split into windows of length $T_{win} = 20$ minutes, with an overlap of $T_{over} = 10$ minutes according the process in Algorithm 5.1. The window length and window step were chosen based on the original algorithm. The final window overlap was allowed to vary so that the final window always represented the final 20 minutes of each tracing. To retain regular indexing of windows for comparisons, these shifted windows were assigned to the closest index found using the regular evenly spaced indexing scheme.



(a) Low variation



(b) Missing samples

Figure 5.4: Two samples of data showing; (a) an example of a low variability artefact, highlighted with the low variability detection algorithm. These samples will be considered invalid; (b) a demonstration of small gap repair using linear interpolation. Inserted samples are highlighted.

Algorithm 5.1 Window selection

```
Initialize the cursor at sample 1
while end not reached do
  if current index is valid then
    mark index as selected, set NO-SIG to false. Move the cursor forward N_skip
    samples
  else if current index is not valid and NO-SIG is false then
    while the distance between the cursor the last previous valid window > N_min
    do
      move cursor back one place
      if Index is valid then
        mark index as selected, exit this loop, and move the cursor forward N_skip
        samples
      end if
    end while
  else if No valid indices then
    exit this loop, set NO-SIG to trueand return to the last valid index
  end if
  move cursor forward one space
end while
```

Window validation

Each window was then labelled as valid or invalid, depending on the presence of missing data. If no missing data were found, the window was assigned a maximum overshoot O_{max} and undershoot O_{min}, representing the delays, positive and negative respectively, that could be applied to the window before the signal encountered invalid data. To maximise the range of delays that could be achieved, some portion of the delayed window was allowed to consist of invalid signal. This proportion was set to 10%, based on an assessment using synthetic data. Since the IRF selection method prefers delays closer to 0, the expected delay, the impact of this was found to be minimal.

5.2.3 Heuristic IRF Identification

In all biological systems there will be a slight delay between the stimulus and the response. Normally the delay of the IRF is simply accounted for by investigating the number of insignificant coefficients at the start of the IRF. However the short intervals between contraction events, especially in the later stages of labour, mean that the response function can be misaligned and give an equally valid yet incorrect output or repeated twice at half amplitude.

The optimal model is not necessarily found at the global maximum VAF, and thus must be robust to segments which contain no system response as is common in early labour when contractions are too weak to illicit a response from the fetus. This is advantageous as to implement the method of delay, memory, and component selection in full, including validation of each IRF would involve calculating the IRF $S \times d \times m^2$ times for each window, where S is the number of surrogates used in significance testing, d, the number of delays to test, and m the maximum memory length. The error surface is smooth across values of d and m but discontinuous across components. To solve this a heuristic search method suggested by Warrick et al. [189] is used to determine the coefficients, which reduces the calculations performed to $m \times (1 + d) \times S$.

Initial memory estimation

The first stage of identification was to take an initial guess at the model order in each window. Warrick [189] found that a well performing first approximation of the memory was the mean contraction length. This was taken to be the dominant frequency in each UA window. The frequency power spectrum of the UA signal was estimated using the Yule-Walker method.

The order of this autoregressive model, k, was optimised by minimising the Akaike information criterion (AIC) defined as:

$$AIC(k) = ln[\sigma(k)^2] + (2k+1)/N$$

with N equal to the number of samples in the window and $\sigma(k)$ equal to the variance of the spectrum, as shown in Figure 5.5. The range of orders was chosen to be between 5-20% of the length of the data sample as recommended by Haykin and Kesler [157]. The initial estimate of memory was $M_{cont} = 1/f_{Max}$. If the initial estimate of the memory placed it outside of the memory range, then it was set to a default value of $120s$. The memory was limited to values between $m_{min} = 50s$ and $m_{max} = 300s$ in agreement with the values cited.

Delay Estimation

The model delay was then estimated. The memory was kept at $M = M_{cont}$ and the delay allowed to vary between $D = d_{min}, \ldots, d_{max}$. The amplitude of the first coefficient of the IRF was recorded for each delay. These values were then median filtered and quantised into 10 bins with values ranging from -0.025 to 0.025. These bins were chosen to present a high resolution around zero crossing points, though they typically also covered the entire range of coefficient values. Candidates for the delay value were then taken as the points where the quantised value first reached the 0 intercept or, if the value began at 0, first left it, as shown in Figure 5.6a. In the case where multiple candidate values were found, the two values closest to 0 were chosen as the expected delay is close to 0.

Values of $d_{max} = 60s$ and $d_{min} = -60s$ were chosen, which differ from the values

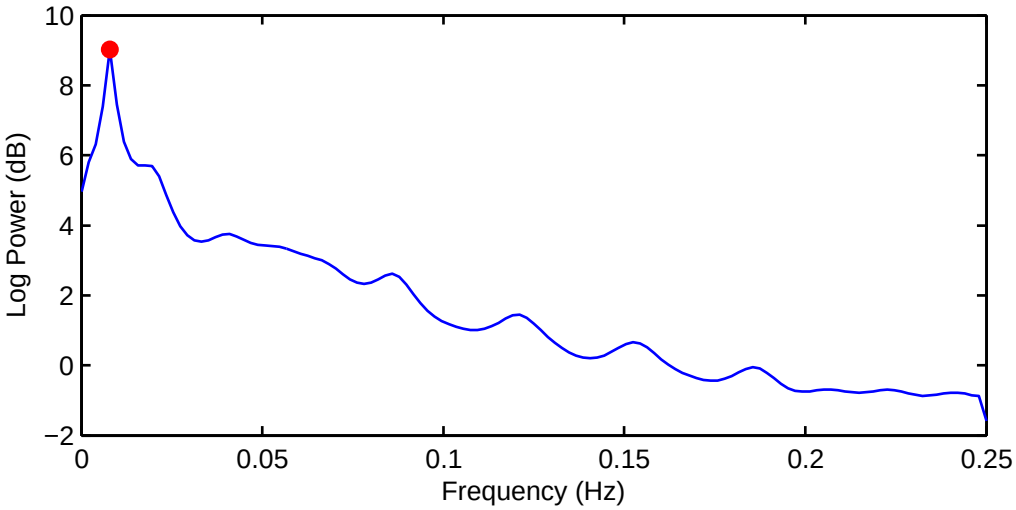
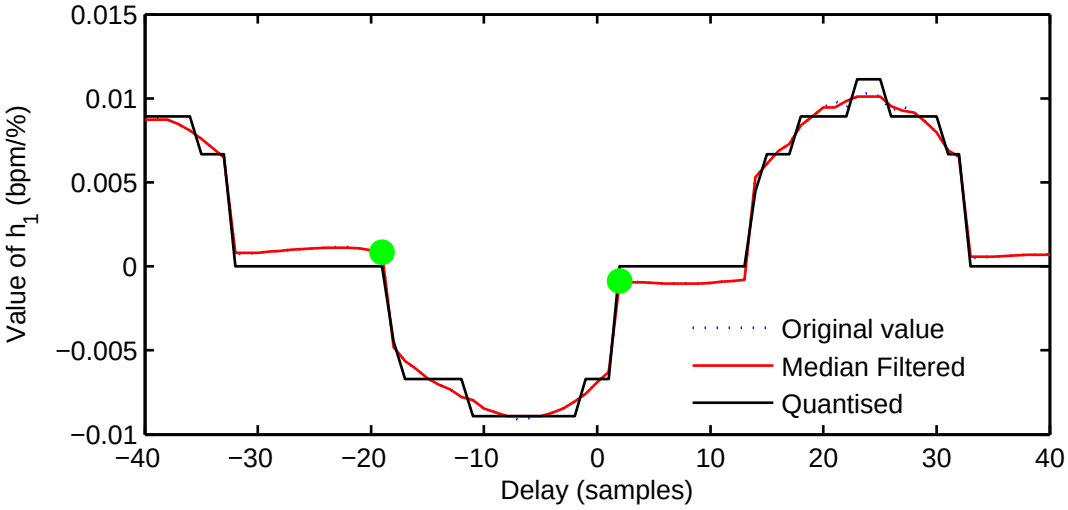


Figure 5.5: The PSD of the UA signal in a window found using the Yule-Walker method. The highlighted peak corresponds to the contraction frequency, and is used as an initial estimate of the IRF order

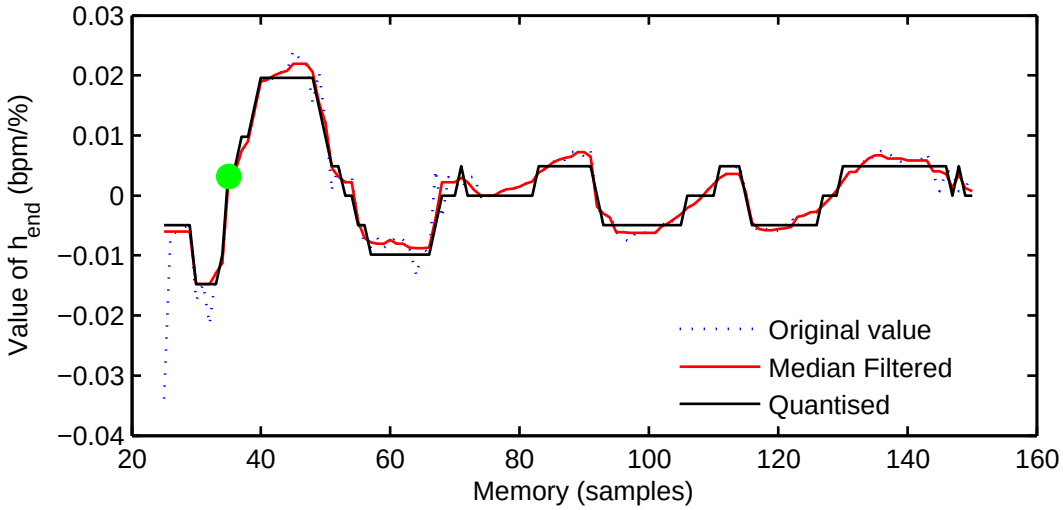
of $80s$ to $-480s$ cited in the literature [195]. The range was decreased both in favour of computational speed and because misalignment effects between the UA and FHR signals were not found to be problematic in the OCFMT database, therefore the large negative delays were highly unlikely.

Refined memory estimation

Each of the candidate delays was then taken onto the next stage. The IRF method was run again for each candidate value of d with the memory varying across the entire memory range. The final coefficients of the IRF were then median filtered and quantised into the same bins as the coefficients from the delay discovery. The ideal IRF memory was the shortest value at which the quantised value of the final coefficient was 0. If no value of m was found which satisfied this condition then the IRF discovery was said to have failed for that candidate. In the case that no candidates were successful



(a) Variation of coefficient 1



(b) Variation of final coefficient

Figure 5.6: In figure (a) the variation of the first IRF coefficient with the delay for a typical window is shown. The value can be seen oscillating about the 0 mark as it approaches and then passes the true delay. The second viable delay value, $d = -20$, is likely to be associated with the proceeding deceleration event. In (b) the variation of the final coefficient with memory is shown. The value can be seen to settle to 0 after the 100 sample mark, indicating the likely memory. Candidate values selected are highlighted in green.

the entire window is then declared invalid.

In this final step, the values of the initial coefficients were allowed to vary slightly, provided that they still abided by the rule that $h(1)$ and $h(n)$ were close to zero, defined as $h(1) \leq 0.25\ max(h)$ and $h(n) \leq 0.25\ max(h)$.

With a final set of candidates consisting of feasible delays and their associated feasible memory lengths the MDL of each candidate was also recorded, and the candidates for each window stored as triplets of the delay, memory, and MDL. Finally at this stage each candidate was checked to ensure that they met all validity criteria:

1. Candidates must sufficiently represent the system. Therefore the VAF must be above 5%.

2. The model should predict a connection between the input and the output. This response should not be disproportionate to the input given the limits of the fetal physiology. The absolute value of the gain, as defined in equation 5.4, must not be above 5 or below 5×10^{-3}.

3. The model should not attempt to fit acceleration responses, nor should the model attempt to fit any part of the previous or next contraction. The former occurs most frequently in the early stages of labour, when accelerations are frequent. The latter occurs in the final stages of labour, when contractions are frequent enough to encroach on the search range of the current contraction. Even then, the delay only allows a partial fit with a neighbouring contraction, therefore any fit in these events will be out of phase - the descending edge of a contraction is associated with a drop in FHR and the ascending edge of a contraction associated with a rise in the FHR. The resulting IRF will have a positive gain, therefore any positive gains are considered invalid.

$$Gain = \sum_{i=0}^{M-1} h_i \tag{5.4}$$

where M is the model memory, or the number of coefficients, and h_i is the ith coefficient.

5.2.4 Post-processing

Once a set of candidates has been found for each window, a post-processing step is performed to ensure that the results of each window are consistent with their neighbours and to chose a single best performing candidate. To encourage windows that start at the same time, they are matched for delay, and to ensure that they end at the same time, they are matched for recovery $r = d + m$.

Significance testing

To identify and to remove candidates which had not picked up on true dynamics, the models were tested for significance using the amplitude adjusted windowed Fourier transform (AAWFT) described by Theiler [174]. For each winning candidate, 99 surrogate datasets were generated by randomly using the AAWFT which reassigned the phases of the frequency content of the UA and FHR data for that window, an example of which is shown in figure 5.7. This maintains the frequency content of the signal but destroys any patterns between the input and output signals. The AAWFT algorithm was used as follows:

- Generate a data set of the same length as the window from N independent samples from a unit normal Gaussian distribution.
- Adjust the rank of this Gaussian series to match the rank of the windowed data.

The position of smallest value in the series will match the position of the smallest value in the window and so on.

- Window the Gaussian series to reduce the introduction of high frequency artefacts.
- Fourier transform the Gaussian series, and randomly adjust the phases of each frequency by multiplying each phase by $e^{i\theta}$ where θ is a random number picked from a uniform distribution.
- Ensure that $f(\theta) = -f(-\theta)$ by replicating the positive frequencies, multiplied by -1, on the negative side of the spectrum and then perform the inverse Fourier transform. The asymmetry ensures the result is real.
- Adjust the rank of the real data to match the rank of the new Gaussian series.

An IRF model was then fitted to each surrogate window with the number of components, delay, and memory values from the original retained, and the resulting VAF calculated. For a candidate to be considered valid its VAF must fall within the top 5% of values found. This method maximised the number of windows retained and allowed low VAF values which capture dynamics to be retained, which is investigated further in section 5.3.

Candidate selection

From the set of validated candidates an initial guess at the optimum candidate was made by selecting the candidate with the lowest MDL. Local estimates of r and d were made by median filtering the values in three windows either side of and including the current window, with the stipulation that at least 4 of the windows must be valid. The median filtered values for a window i are termed $\hat{d}_i$ and $\hat{r}_i$. Then all the potential

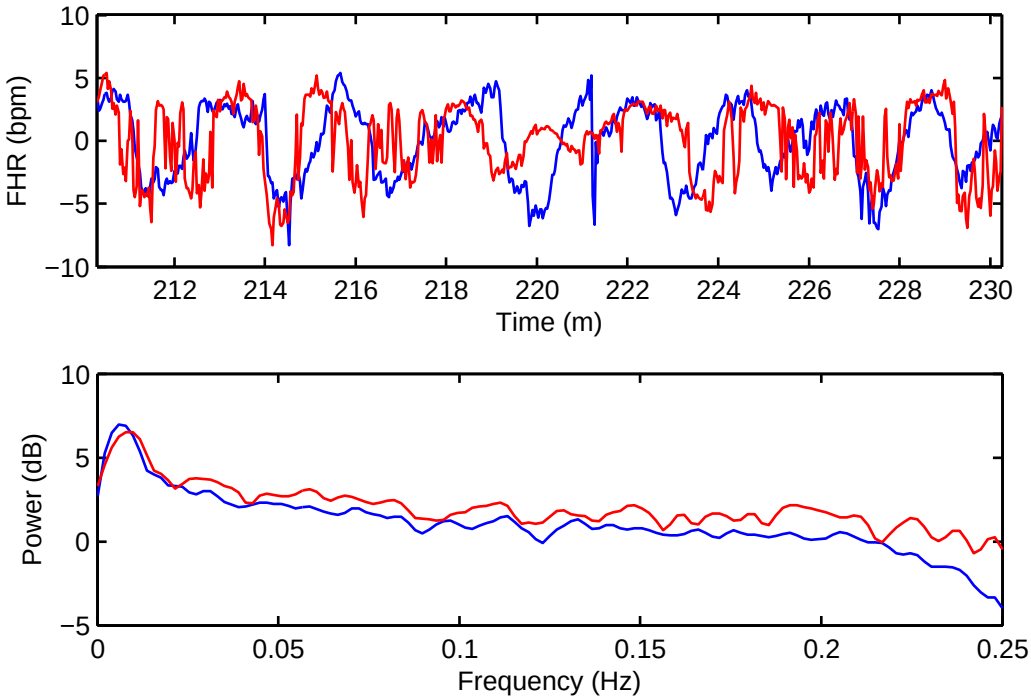


Figure 5.7: An example of generated surrogate data in red, alongside the original data sequence shown in blue. The corresponding spectra of the sequences are plotted in the lower figure. The spectra of the generated signal is not substantially different to the original.

r, d, and MDL values in a single window were normalised to have zero mean and unit variance. The Euclidean distance was found between each of the candidate in the window and the local best candidate, identified as [median(r, d), min(MDL)]. The candidate which was closest to this ideal candidate was selected as the winner, and taken on to the next stage of post-processing.

Optimisation of winning candidate

A small local search was then performed around the winning candidate to optimise the delay and memory values. The IRF identification step was performed over delay and memory values of $d_i \pm 12$ seconds and $r_i \pm 12$ seconds with a fixed number of components matching those of the winning candidate. Of all of these new fine-tuning candidates, the final IRF was chosen to be the IRF which had the smallest combined start and end values, as the low frequency response for this system is expected to begin at and to decay to zero. The impact of the algorithm up to this stage on the IRF in a single window is demonstrated in Figure 5.8, with the individual best candidate at each stage shown.

Finally any windows which produced an unusual candidate were removed by finding the standard deviation of the delays over the entire case, and removing windows whose delay lay more than 2 standard deviations away from this range. This removed instances where the recovery was large enough to allow a deceleration to be matched with a previous contraction.

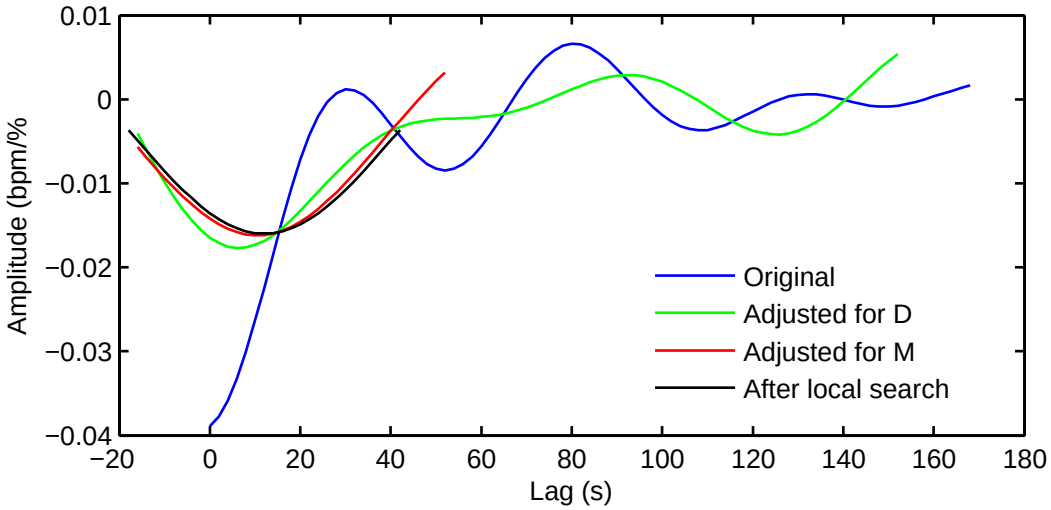


Figure 5.8: The whole process of IRF discovery, from the initial estimate of the delay, memory, and the final refinement. The original IRF is found using $m = m_{cont}$ and $d = 0$. The delay was then optimised, giving the IRF adjusted for D. Then M was adjusted with D held constant at d_{opt}, and finally a local search of both parameters gave the final result. Note that in the implemented algorithm, the process includes multiple candidates, which are not shown here.

5.2.5 Differences from the literature

This method was replicated using the 376 cases in the OCFMT-M dataset, chosen as a balanced dataset of size comparable to that used in model development. A small database allowed results to be assessed individually before the method was applied to the complete OCFMT database. The method could not be exactly replicated as described in the literature due to differences between the databases used, therefore several assumptions and changes to the method were made which are detailed below.

Dataset differences

Three datasets were described in the development of the cited method, shown in table 5.1. All cases in the literature used sampling rates of 4Hz for FHR and 1Hz for

UA signals, and were from a gestational age of at least 36 weeks with no congenital malformations unless otherwise stated. Base deficit was used as a measure of outcome, with a base deficit ≥ 12 defining the positive class, along with a clinical assessment of neurological damage. Notably some traces were acquired using internal sensors, others external. Some were recorded directly from monitoring devices and others were digitised scans of printed traces.

The OCFMT database guarantees gestational age and rejects traces from cases which exhibited congenital malformations or any other growth or development defects. These data were sampled at 4Hz for FHR signals and 0.25Hz for UA signals. All traces were acquired externally and are directly monitored outputs.

The low sampling frequency limits the upper range of frequencies that can be modelled. Several sources have demonstrated useful information content in higher frequencies [188, 159, 29], however the dynamics in question here are those between the baro- and chemo- receptors, contained in a low frequency band up to 15mHz, of which the shortest events expected are brief decelerations of at least 30 seconds in length. The results in the literature were performed at the higher sampling frequency of 1Hz, therefore all preprocessing operations and filters were modified to accept an input sampling rate of 0.25Hz. As a model of low frequency effects, this is not expected to affect the output.

All tracings used in the literature were of at least 3 hours in length. Cases in the OCFMT may be shorter than this, but not shorter than 30 minutes. The algorithm will not be affected by case length, provided that there are sufficient windows for the comparisons made in the post-processing stages.

Table 5.1: Differences between data sets in literature and in this study. The largest distinction lies in the method used to obtain recordings, with W-Databases being acquired through a combination of external or internal methods and OCFMT solely external, and the measure of outcome.

Source	Database properties
W-2013 [188]	231 recordings at least 3 hours in length. 89 normal, 142 metablic acidemia
W-2006 [191]	161 traces. 56 with BD<8, 56 with 12 $\geq$BD< 12 and 49 with BD$\geq$ 12 and compromised neurological function
W-2009 [195]	264 Traces of at least 3 hours in length. 221 Normal, BD<8mmol/L. 43 severely pathological, BD $\geq$12mmol/L and death or evidence of HIE. Acquired using both internal and external equipment.
OCFMT-M	376 traces, properties listed in Section 4.1
OCFMT	7,518 traces, properties listed in Section 4.1

Differences from the literature

Initially the method was implemented exactly as described in the literature. Table 5.2 shows the number of windows which were successful, along with the number of windows excluded at each stage. Three quarters of the windows were excluded due to detection of an artifact. Of the remainder only 3.5% of the healthy group and 2.4% of the adverse group were successfully identified. The large amount of data labelled as artefacts were caused by the original definitions of missing data and gap repair, with gaps being much more frequent and of longer duration in general in this externally acquired dataset. This had repercussions throughout the whole process as a high proportion of windows were removed due to the requirement that windows have at least 3 valid neighbours. The low number of validated windows per case meant that the outlier delay detection step was rendered useless as there were insufficient numbers to calculate the standard deviation of values.

Table 5.2: A comparison between the number of windows failing validation in each processing step using the OCFMT dataset and reported in the literature. The modified algorithm validates a much greater proportion of the data. OCFMT-M *Method not implemented, - Result not reported

Dataset	Warrick		OCFMT-M		OCFMT	
Algorithm	Warrick		Warrick		Modified	
	Nml	Adv	Nml	Adv	Nml	Adv
Artifact	8%	42%	72.6%	77.5%	23.4%	25.2%
-Low variability	-	-	1.0%	0.8%	*	*
-Failed Schmitt	-	-	2.0%	1.1%	*	*
Low gain	13%	4%	4.7%	3.9 %	33.1%	29.6%
Estimation failed	21%	9%	0.2%	0.15%	0.2%	0.3%
Insignificant	27%	15%	8.7%	8.1%	8.7%	9.9%
Outlier d	2%	2%	0%	0%	0%	0%
No neighbours	-	-	7.3%	5.9%	5.4%	6.0%
Validated	30%	28%	3.5%	2.4%	29.0%	29.0%
Total	3,978	774	9,788	10,603	160,579	4,044

To alleviate this several changes were made:

- The maximum gap size was increased to 15 samples and the Schmitt detector for artefact detection, which even with the relaxed thresholds, was still frequently labelling strong decelerations as invalid data, was replaced with the drop-off detector developed by Dawes et al. [40].
- Rather than using a regularly indexed window, windows were allowed to shift to maximise the amount of validated data.
- To prevent matching spurious patterns candidate IRFs with a gain above 5bpm/% were removed from consideration.
- To prevent IRFs selected which were out of phase when contractions were close together or the IRF was fitting accelerations IRFs with a gain below 5e-3bpm/% were removed.

- The initial value of the MDL tuning parameter, P_c, given in the literature often resulted in the selection of simple models with near-zero VAF values, as the VAF did not increase enough in proportion to the parameter penalty for higher component numbers to be selected. This is likely due to the difference in data quality due to both the method by which signals were recorded in the literature and the relaxed definitions of missing data used in this implementation, which results in generally lower VAF values. In other instances the algorithm selected complex IRFs with large weights which were fitted to signal noise. To prevent this, and to highlight the corresponding windows as invalid, the number of components considered was limited to lie in the range 2-6.

5.3 Analysis of model success

The refined algorithm was successful in producing robust and simplified IRFs for cases over the complete OCFMT database. 47,811 valid windows were discovered from a possible 164,623 windows in the final 4 hours of recording. To confirm valid windows are not being rejected the cause for window loss and the results of the significance testing are examined in more detail.

5.3.1 Causes of window loss

The variation in cause of window loss as partum is approached is shown in Figure 5.9. The high proportion of windows lost due to insufficient gain reflect contractions which did not elicit a deceleration response. This agrees with observations from CTG traces that low intensity and low duration contractions will often not cause any deceleration event. Their number decreases from 40% of windows to 25% at birth as the frequency

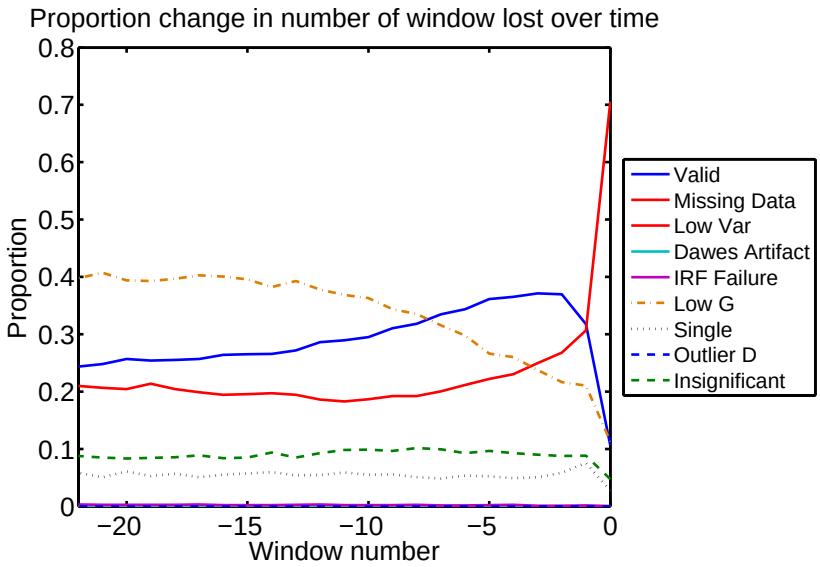


Figure 5.9: The cause of window loss changes cases approaches partum. Windows are numbered according to their distance from partum. Towards partum, the number of windows lost to insignificant gain is reduced, and the number due to missing data increased, resulting in a net increase in the proportion validated. In the final windows, the number lost to missing data leaps to 0.7, causing a large reduction in the number validated.

and intensity of contractions increase.

Low gain can also be caused in the PI step as a result of the modified weighting of the MDL equation. When a model fails to fit the low frequency signal well, but finds some spurious relationship from noise the resulting model can pass VAF requirements, but the first few selected PI components will consist of small coefficients. The associated VAF will then remain low until a high number of coefficients has been reached, meaning that under the MDL selection criteria the penalty term will outweigh the error term and a low number of coefficients will be selected.

The number of valid windows climbs slowly towards partum until the final windows where there is a significant rise in the number of windows considered invalid due to missing data. This is due to a decrease in signal quality towards partum arising from

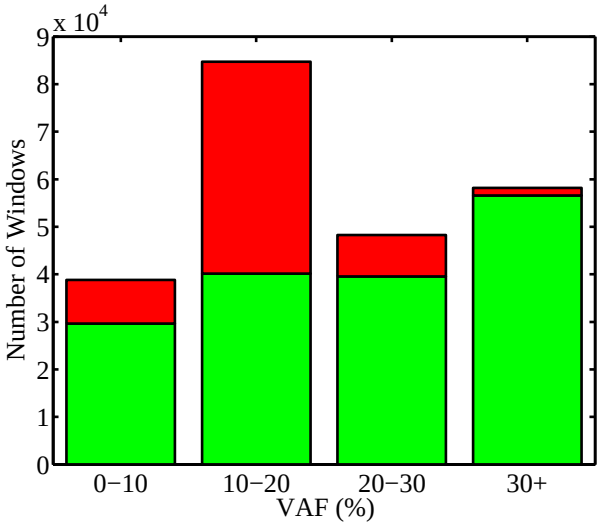
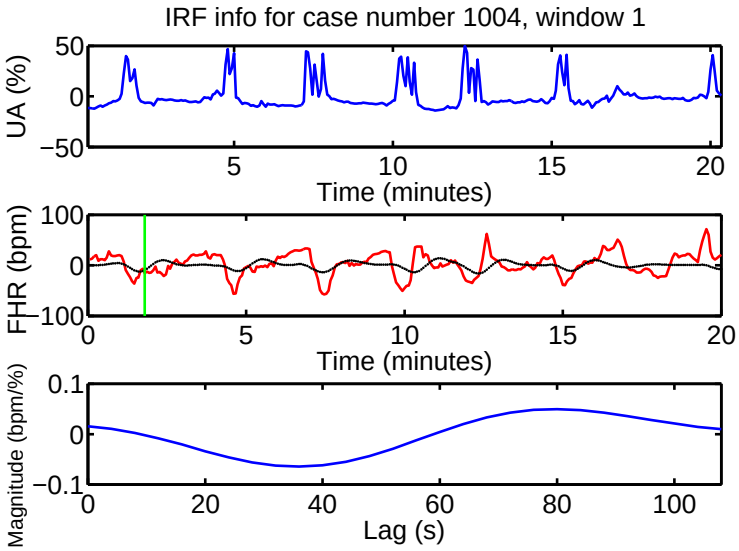


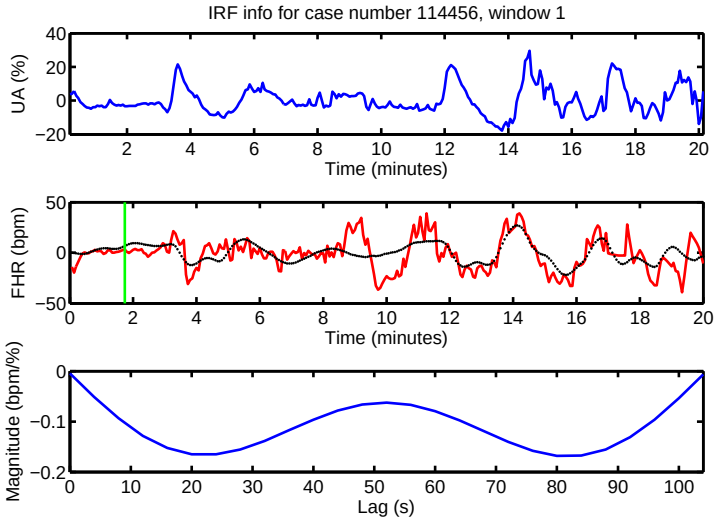
Figure 5.10: The number of windows passing the surrogate significance test was highly dependent on the VAF in that segment, demonstrated here using the OCFMT-M database. Invalid windows will pass the test in 5% of cases assuming they represent no correlation between the UA and FHR signals. VAFs below 5% were rejected by the algorithm, therefore the 0-10% bar technically represents values from 5-10%. Windows with higher VAFs were more likely to be successfully validated.

a combination of sensor, fetal, and maternal movements. The number of windows lost to isolation and failing significance tests remain constant throughout the labour until the final windows. A slight trend in the rejection of isolated windows matches the changes in validated windows, since it is less likely that a window will be isolated if the number of valid windows is increased.

The number of windows lost to significance testing, shown in Figure 5.10, was found to be highly dependent on the VAF, with a high value indicating a good model fit. In some cases models with a low VAF were accepted, indicating that a weak model was fitting a true effect, and models with a high VAF were rejected indicating a strong fit to random noise. Low VAF acceptances are in part due to the fact that to pass the significance test a model had to place in the top 5 VAF values. Therefore a model representing random dynamics will be accepted one in twenty times.



(a)



(b) The incoming data for an accepted case with high VAF

Figure 5.11: A demonstration of instances where: (a) an IRF with a low VAF was accepted by the surrogate test, and (b) a rejected window with high VAF. In (a) noise in the contraction events distorts the discovered IRF. However there is still a causal link between the UA and FHR signals. In (b) a period of UA signal loss causes the IRF to be misaligned, fitting decelerations to the current and next contraction events, allowing the first deceleration at the 10 minute mark, which has no corresponding contraction, to be fitted.

Figure 5.11a shows a window accepted despite a low VAF. The algorithm fails properly to fit a response due to noise visible as the frequent drop-outs on the UA signal, possibly caused by loss of sensor contact. However there is a strong correlation between these contractions and the associated decelerations, and the resulting IRF is retained. In figure 5.11b the window was rejected despite having a high VAF. In this window the contraction shapes are unclear and it is difficult to identify which decelerations are matched to which contraction events.

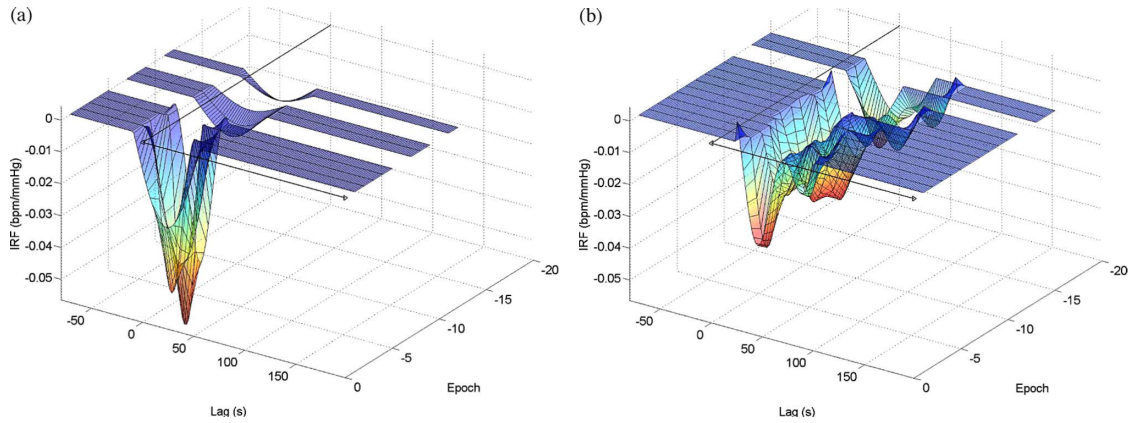
The disadvantage of this method is that one in twenty windows containing no information will be accepted as valid, and that in circumstances where low VAF windows are accepted, the actual IRF may not contain accurate information. In both cases the IRF found was of an unexpected shape - either a "W" or sinusoid.

5.3.2 Comparison with the literature

Case visualisation

Warrick et al. presented individual cases using a modified waterfall plot, shown in Figure 5.12, which present a the data in a highly compact and readable manner. They noted that the IRFs of cases presented with a large BD at birth experienced shorter delays and larger gains towards partum, whereas healthy cases presented a much more consistent pattern. Using 3-D plots can mask some information, so cases in this study are presented as 2-D surface plots. Two such surfaces are shown in Figure 5.13 for a single case with birth pH $\geq$ 7.1 and a single case with pH <s 7.1. These results demonstrate the same pattern of decreased delay and increased gain towards partum observed in the literature.

Though these plots are useful in reviewing cases, analyses made using them are



(a) A pathological case (b) A healthy case

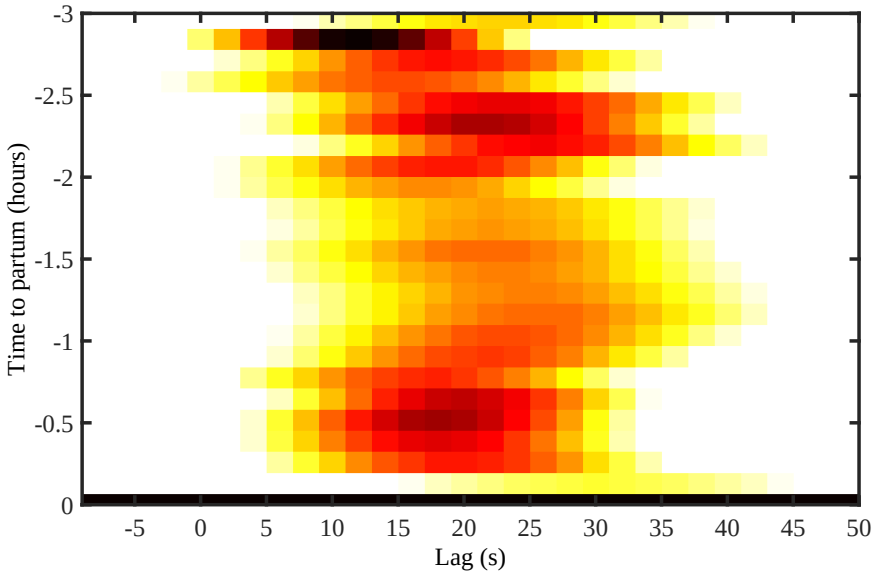
Figure 5.12: The discovered IRFs reported in the literature [195]. On the left an adverse case is shown, exhibiting larger responses and a shorter delay as birth approached. On the right a healthy example is shown, maintaining a constant response function throughout the labour.

subjective, based on trends in the delay and gain properties. The greatest distinctions made appear in the final windows, and would be hard to distinguish from healthy variation in a live environment.

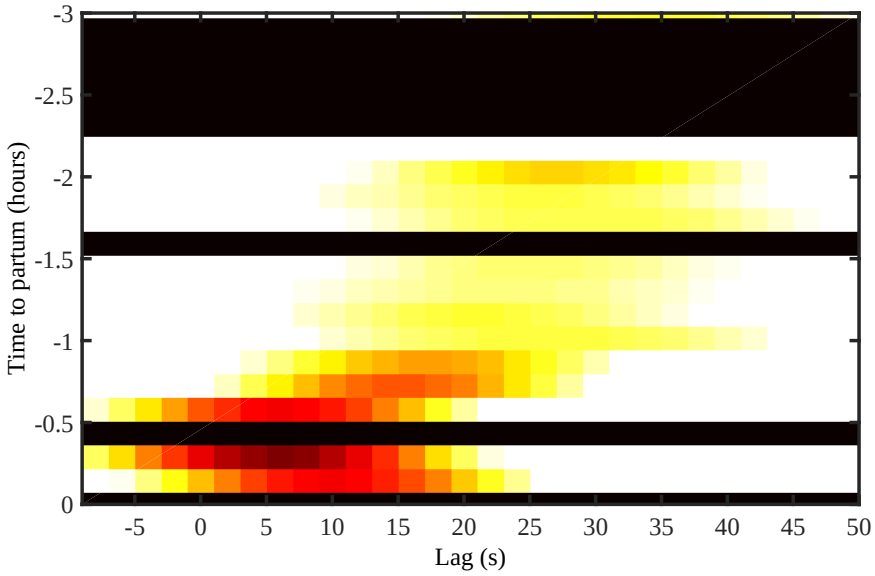
Evolution of IRF features

Warrick et al. [195] also investigated the mean and standard error of features of the IRF in the 3 hours up to partum, which are reproduced in Figure 5.15. Each window was assessed for significance differences between classes using a Kolmogorov-Smirnov (KS) test and samples passing at $p < 0.05$ are highlighted with an asterisk on the x-axis.

These results were compared with results found using the OCFMT database in Figure 5.14.. The mean and standard error of each feature was plotted across all windows and data in each window tested using a KS test, for differences between the healthy and adverse data. Windows passing this test at $p < 0.05$ were highlighted



(a) A pathological case



(b) A healthy case

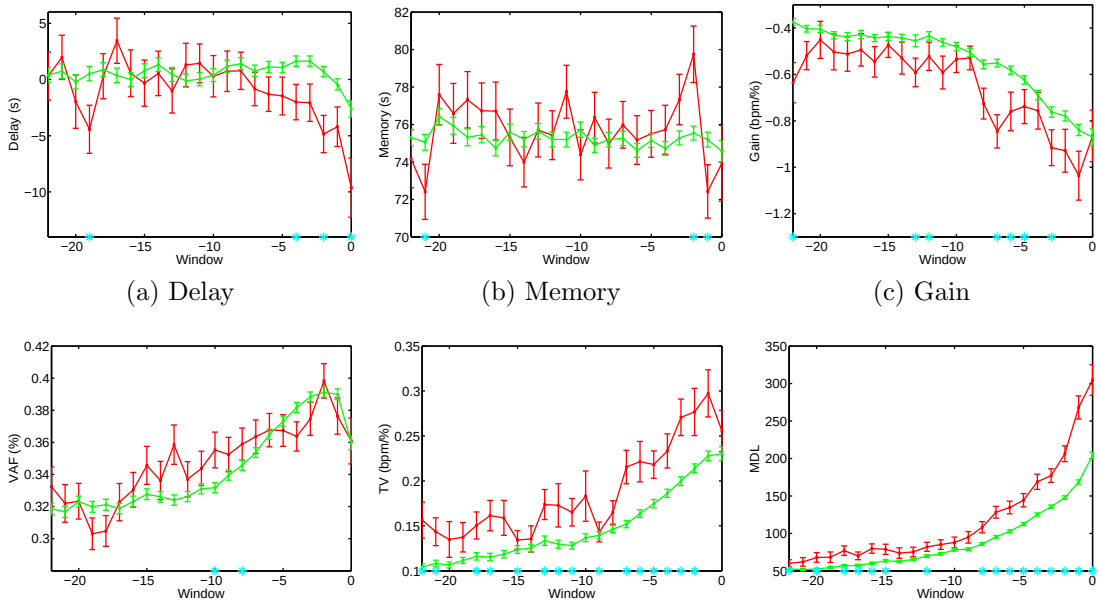
Figure 5.13: The discovered IRFs for selected cases from the Oxford Database are shown as surface intensity plots. Agreeing with the literature, the pathological case shows an increase in response intensity and a shorter delay towards partum, where the healthy cases IRF remained stable.

using an asterisk. To provide larger numbers in the adverse class the data were divided into groups with pH$<$ 7.1 and pH$\geq$ 7.1 giving 626 and 6,942 negative and positive cases respectively.

The VAF values identified in the two trials were similar, being in the range 32-40% in this study and 20-40% in the literature. An upwards trend was observed on both healthy VAF datasets from window -10 onwards, though the baseline rate in this study was higher, at 33% against 27%. The distinction between the positive and negative cases were not significant in this study, and not significant in windows beyond the 7^{th} in the literature.

A gradual shift across the entire birth which was visible across all features except for the IRF memory. This shift was divided into two segments. Features in the first segment, between windows 11-22, are roughly stable or slowly evolving. In the second segment the change is much more rapid. The average second stage of labour roughly lasts 3 hours, and marks a change in the dynamics of labour with increased contraction duration and intensity. During this stage the mother will be encouraged to push in time with contractions, so this stage will also include additional effort from the mother. Therefore the shift in trajectory observed may be correlated with the start of the second stage of labour, which is examined further in Section 5.4.2. Warrick et al. [195] found healthy windows to have a delay between -10 and -20 seconds, which remained steady throughout the entire labour. This study noted a decline in delay towards partum, being especially notable in the final three windows. However as shown in Section 5.4 pH was sensitive to delay, but not to clinical output. This may indicate that a high base deficit also does not effect delay.

Values of gain were in general larger in this study, ranging from -0.4 to -0.85bpm/% against -0.1 to -0.3bpm/% for healthy cases identified by Warrick et al. in Figure 5.15.



(a) Delay (b) Memory (c) Gain

(d) VAF (e) TV (f) MDL

Figure 5.14: The evolution of IRF properties for cases with pH $\geq$ 7.1, shown in green, and $<$ 7.1, shown in red. Features included are (a) delay, (b) memory, (c) gain, (d) VAF, (e) TV, and (f) MDL. Birth occurs at the end of window 0, with the start of window 22, 4 hours prior. Standard error bars are plotted, and windows which passed a Kolmogorov-Smirmov significance test at $p < 0.05$ are highlighted with an asterisk. Significant differences between windows were observed for the TV and MDL values in particular in later windows. The delay values became lower for adverse cases as partum approached.

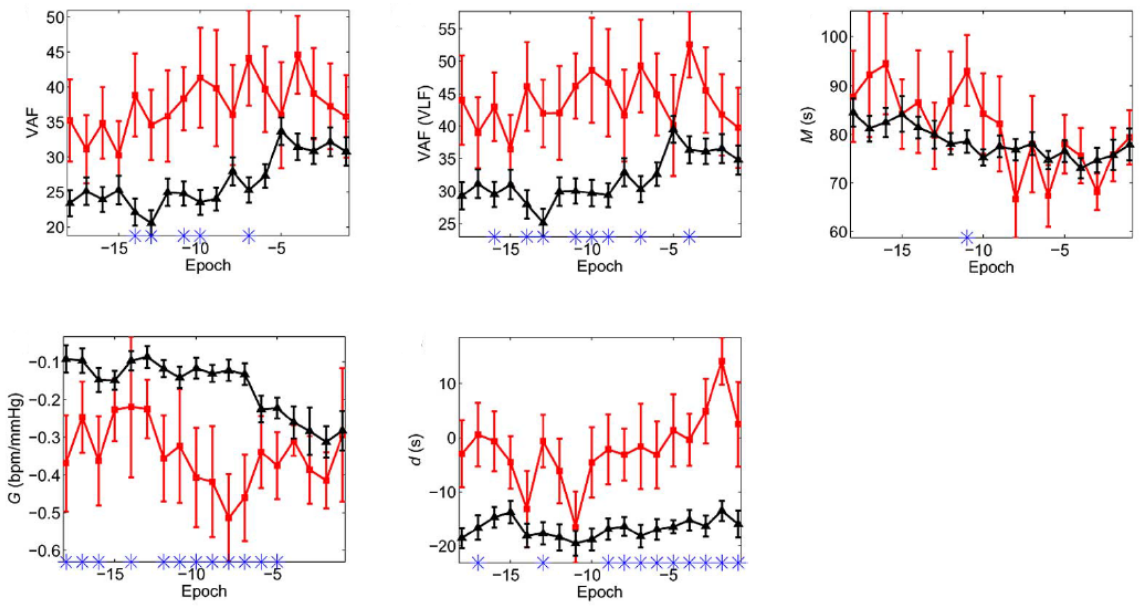


Figure 5.15: The progression of 5 IRF parameters reported by Warrick et al. [195]. The parameters in the top row are, from left to right, the VAF, the VAF calculated on a low pass filtered version of the FHR data, and the Memory. On the bottom row are Gain and Delay. For each window, standard error bars are also shown. Parameters found to be significantly different between the classes using a KS significance test with $p < 0.05$ are highlighted with an asterisk. Throughout the majority of labour significant differences were observed between the classes, in particular the delay. The VAF was greater for adverse cases, indicating that their responses were easier to model.

The adverse cases exhibited a lower gain overall throughout the entire course of the labour, though not significantly so until windows -7 to -4. In both studies, an increase in the gain was observed in adverse cases only, and occurred between windows 10 and 5, though the effect was less clear in this study.

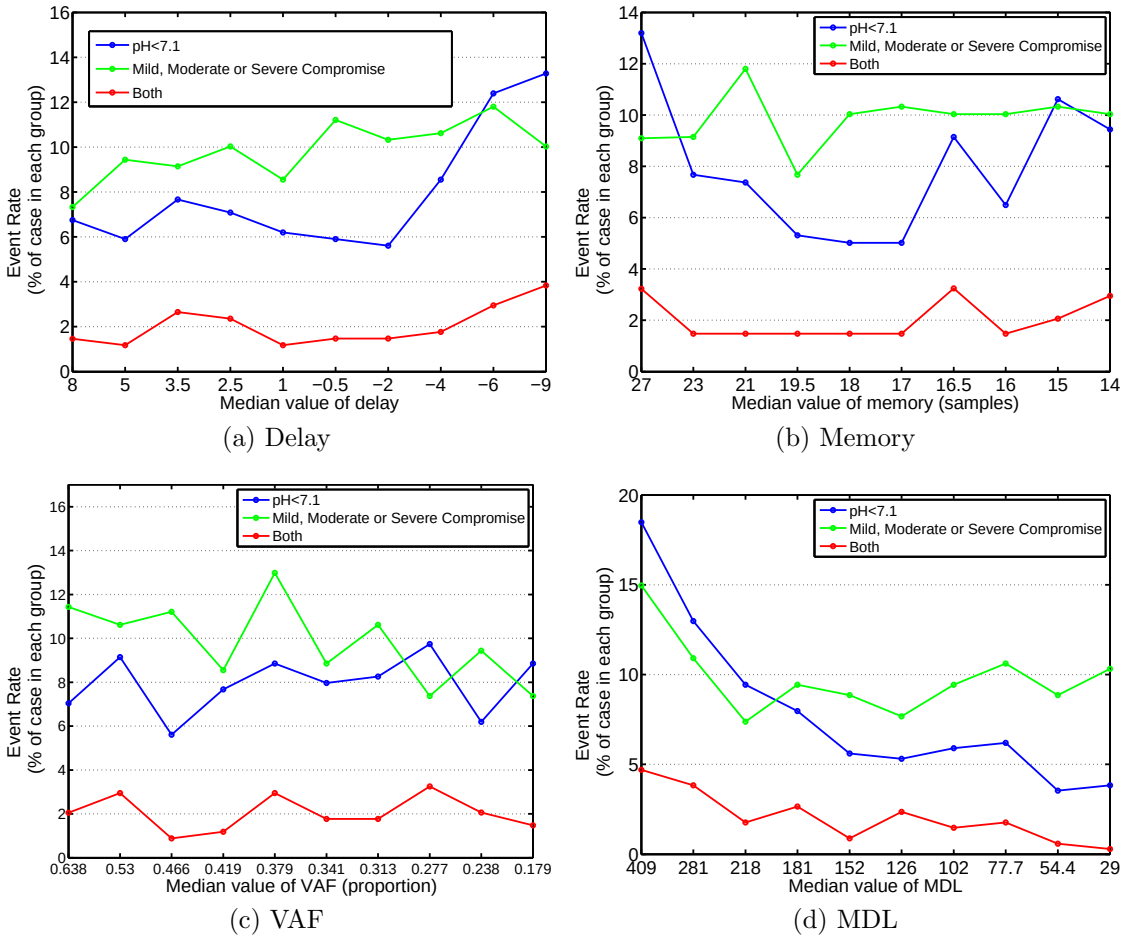
This represents the time period 120-60 minutes before birth, shortly after the expected start of the second stage of labour. This also supports the suggestion that the timing of stage 2 labour is an important predictor. The two newly introduced features, the TV and MDL, produced the greatest number of significantly different window. The variability was more than double the baseline value at partum, and the MDL appeared to increase exponentially towards birth.

5.4 IRF properties as features

To assess which features may be of greatest impact EveREst plots were used to visualise the variations of the proportions of each class with each of the features. Of these features delay, memory, VAF, and MDL are shown in Figure 5.16. The remaining distributions were similar to the VAF distribution and are not shown.

In previous studies the delay was shown to be greater (more negative) for pathological cases [195]. This same effect is observed in these results with 14% of cases having a delay < -9 samples being born with pH<7.1, versus 6% for those with values ≥ -2. This effect is more strongly pronounced for the pH value than for the clinical assessment.

In the literature the VAF was shown typically to be higher for cases with a base deficit $\geq 12mmol/L$ than for those with a deficit $\leq 8mmol/L$, and the delay was in general longer for these pathological cases [195]. It was theorised that a weakened



(a) Delay (b) Memory

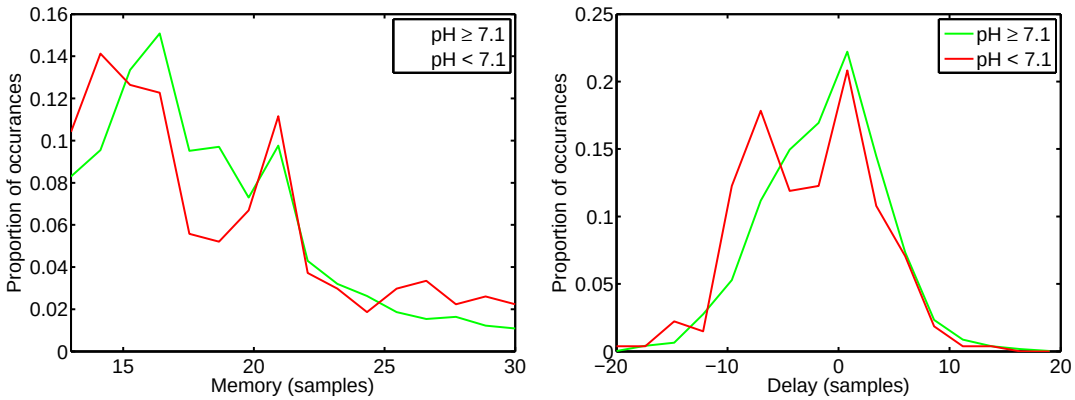
(c) VAF (d) MDL

Figure 5.16: EveREst plots of features of the IRF. Three outcome classes were identified based on indicators of hypoxia; low pH, clinical assessment, and cases where both indicators were present. The median value of the final three windows was selected for each case, which represented the final 40 minutes of labour. More negative delay values were associated with an increase in low pH incidence rates. Extremely short or long model orders showed an increase in low pH incidence, but not with compromise. The MDL showed the strongest response, with both compromise and low pH having increased incidence rates at high MDL values.

system will more closely track the input as a healthy system required a smaller compensatory response, causing adverse cases to show a higher VAF [195]. The VAF values in this study were weakly correlated with the clinical outcome, and showed no correlation with the pH value. This result indicates that the different measures of clinical outcome are sensitive to different features of the signal, in this case the success of the model fit.

The memory shows the opposite effect: there is little change in the proportion of adverse clinical outcomes in each bin but the pH value changes. This effect was missed previously as it was assumed that the two groups would be linearly separable, and have different means. Unique among all of these plots, the memory appears to show a safe region, with memories between 17 and 19.5 samples having a 5% risk and those on the extreme values up to 13%. The models created may be falling into groups either side of the normal or may simply have a much greater spread.

The most significant distinction between classes was found in the MDL in Figure 5.16. At the top end of values the likelihood of belonging to the adverse classes rises exponentially. The feature is strongly discriminatory for both clinical assessment and pH value, and hence performs well on cases defined as pathological. The MDL is strongly linked to the model's performance, but also the complexity of the incoming signal. In Section 5.4.1 this is further investigated and compared to the other complexity measures found using OxSys such as entropy, which have been found to perform well as a feature. A histogram of memory values in the final three windows is shown in Figure 5.17(a). The memory appears to split into three groups, each with a normal distribution, though only two are visible in the healthy data. The first low pH grouping peaks at 14 samples, compared with 16.5 for the healthy group. The central group for both lies at 21 samples, and the third low pH group lies above 26 samples.



(a) Memory (b) Delay

Figure 5.17: Histograms of (a): the median model memory, and (b): the median model delay found in the final 3 windows using the SI algorithm and the OCFMT Database. In each plot healthy subjects are plotted in green, and those pathological in red. Three clusters of memory are observed, with the lower cluster having a lower peak for adverse cases. Adverse cases showed two peaks of delay values, indicating there are two groups of responses, some of which have the same delay as healthy cases, the second roughly 10 seconds faster.

A shorter memory indicates a shorter duration response, or in this case, a deceleration which finishes sooner after a contraction.

In Figure 5.17 (b) peaks in the mean of the delay histogram shows two distinct groups appearing. The first group has a centre at -7, the second at 0 along with the centre of the distribution of normal cases. This potentially represents two groups of adverse cases; those where the hypoxia has caused the response to a contraction to slow or flatten, and those behaving similarly to healthy cases.

5.4.1 Correlation with existing parameters

To compare performance of the method against existing parameters in the OCFMT database for each case, the method was run with the window length adjusted to 15 minutes, and the step size to 5 minutes. A requirement of new features is that they

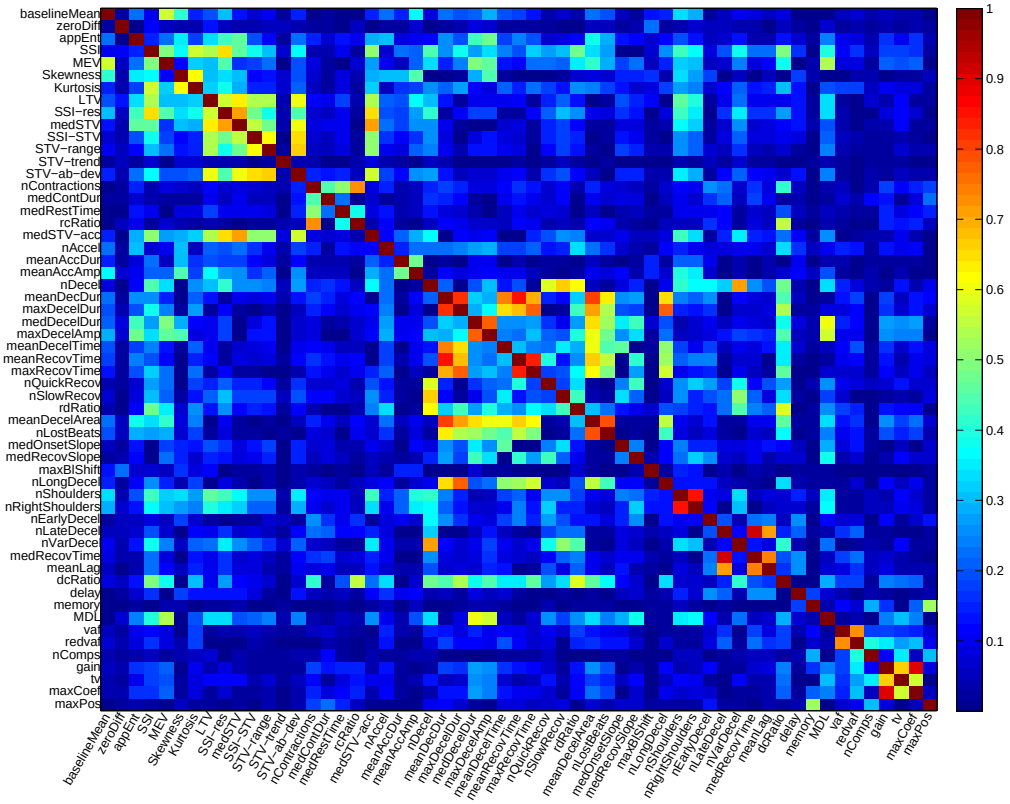


Figure 5.18: The correlation between features including only cases which had valid values for all features, totalling 1,941 cases. Properties of the IRF were weakly correlated with the existing features, though there were stronger correlations between IRF features.

do not overlap with existing features, which was checked by finding the correlation between each pair of features as shown in Figure 5.18.

To maximise the number of valid samples the median of the value for the final 4 windows of each case was found. After removing cases lost to insufficient data for the SI method, or a lack of accelerations (and hence acceleration features), the number of cases used in the correlation comparison was 1,941.

As expected there was a strong correlation (0.7293) between VAF and the reduced

VAF, between gain and total variation (0.6706), and between gain and the maximum coefficient (0.8916). There was a notable correlation between the IRF length and the position of the maximum (0.5285). In comparison to the existing features, outside of the MDL no correlation exceeded 0.2739. The MDL however showed some relation to the maximum deceleration amplitude (0.5949), the deceleration duration (0.6499) and the minimum expected value (0.5620).

The immediate implication is that the MDL is related to the size of the deceleration, and that a larger deceleration requires a more complicated model. The MDL is also related to the recovery slope, and the presence of shouldering in the deceleration and less so to the beats lost or deceleration area.

That these new features are at best weakly correlated with existing features suggests that they are capturing information not yet recorded in any of the existing features, or at least strongly by any single existing feature. This indicates that they may be used to improve absolute classifier performance in some combination with the existing feature sets.

5.4.2 Correct choice of alignment

The results in Section 5.3.2 suggest that the stage of labour impacted the evolution of the IRFs and their properties. To better visualise the impact of stage on these features the case windows were aligned around the transition between stages 1 and 2.

In Figure 5.19 the features are shifted so that the onset of stage 2 labour occurs at index 0, with any shifts greater than 23 grouped together into the 23 bin, this group including only 3% of cases. When lined up in this fashion the trends in IRF features become linear.

The number of values in each window bin drop towards the extremities of the index which represent extremely long or short second stages as these are understandably less common. The number of validated windows in the adverse group peaks at 204, and windows from either groups with less than 50 samples were removed from consideration when fitting regressions.

Over most features, such as gain and VAF, the trajectories of the trends were similar for both classes. For two of the features however, the trajectories of the features were significantly different, delay and MDL.

Again the MDL is the most discriminative feature, with a good fit ($r^2 = 0.917$ and 0.942 for healthy and adverse classes respectively) and visibly diverging lines. In delay a stronger fit was found for low pH cases than for healthy cases ($r^2 = 0.625$ and 0.283).

Cases born with a healthy pH (≥ 7.1) showed little change in delay, indicating that they continued to respond rapidly to contractions unlike low pH cases whose response was gradually becoming slower potentially indicating a loss in function of the quick responding baroreceptor reflex. High delay IRFs can be matched to late contractions, an indicator of fetal distress.

The increase in gain suggests that the response to contractions intensifies throughout the average labour. The baroreceptor head compression response is proportional to the contraction intensity, though systemic blood pressure compensation has a cut off threshold once the umbilical cord is occluded by the contraction pressure. More likely is that this downward trend represents the accumulation of metabolic and anaerobic products in the blood, causing stronger chemoresponses with each contraction.

Research into the association between the length of second stage labour and fetal outcome has not focused on pH levels at birth, but instead on perinatal outcome. Studies of nulliparous women found either no association [27] or slight association [4]

between nICU admission rates and labour length. In multiparous women both studies [28, 4]found an increased risk of low 5-minute Apgar scores and neonatal Intensive Care Unit (nICU) admission with increasing labour length from 2 to 5 hours. These studies included all forms of neonatal morbidity. Allen et al. [4] noted that some part of the increased fetal lactate may be due to maternal effort in the second stage of labour.

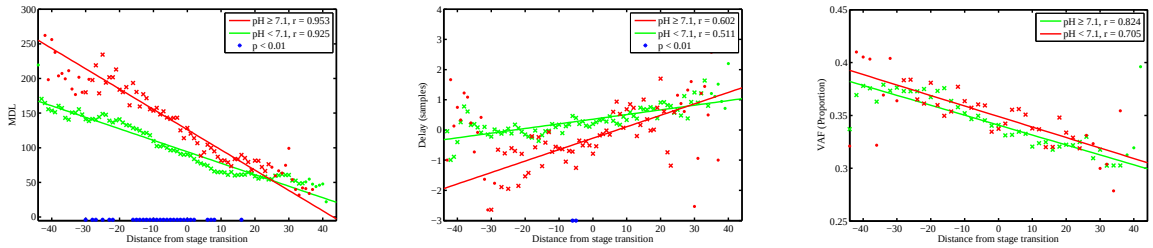
That a large negative delay, high gain, and high MDL were all positively correlated with low pH at birth agrees with the interpretation both that a prolonged labour carries a greater risk to the fetus, and that the fetal condition deteriorates over the course of labour.

5.4.3 Individual regressions

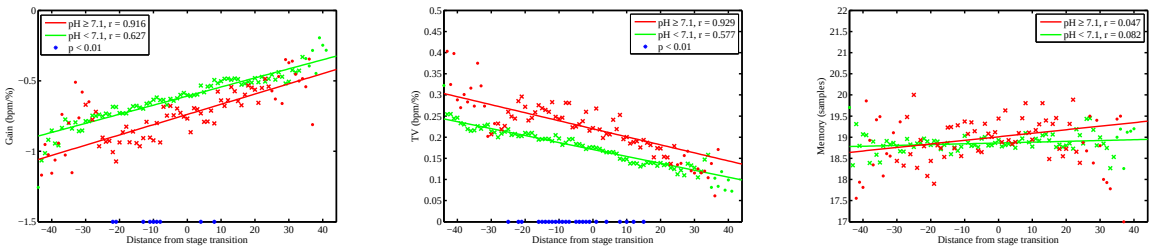
To assess whether the linear trends observed were linear for individual cases or only that the average response formed a linear trend regressions were fitted to each case. Cases with fewer than 4 valid windows either side of the stage transition were excluded.

This formed a new set of features: the trends, offsets, and r^2 values associated regressions fitted to each of the SI parameters. Looking at individual results for the MDL feature, they consist of one of three patterns. A linear trend such as that shown in Figure 5.20a, an exponential trend such as that shown in Figure 5.20c, or no visible pattern such as that shown in Figure 5.20b. To identify exponential trends, a linear regression was fitted to the logarithm of the features, which were first normalised to lie in the range [eps 1], where eps is the minimum spacing between floating point numbers in MATLAB. This was chosen as a small number with a valid log.

For each case a linear and exponential regression was found, and the r^2 values



(a) MDL (b) Delay (c) VAF



(d) Gain (e) TV (f) Memory

Figure 5.19: The features of the IRF were aligned about the transition from stage 1 to stage 2 labour, with -ve windows indicating the period after stage 2. Regression lines were fit to the data in each class highlighting the differences in progression for each. Only means calculated from more than 50 samples were used in regression estimation, and are represented as crosses. Windows with less than 50 samples are shown as filled circles. In cases where then difference between the trends was significant at $p < 0.01$ the p-value was reported. Trends were observed as stronly linear, with high r^2 scores. The greatest differences in trend trajectory can be seen for the MDL (a) and Delay values (b). The gain was continually lower for adverse cases, and the TV higher, indicating larger, less smooth reponses.

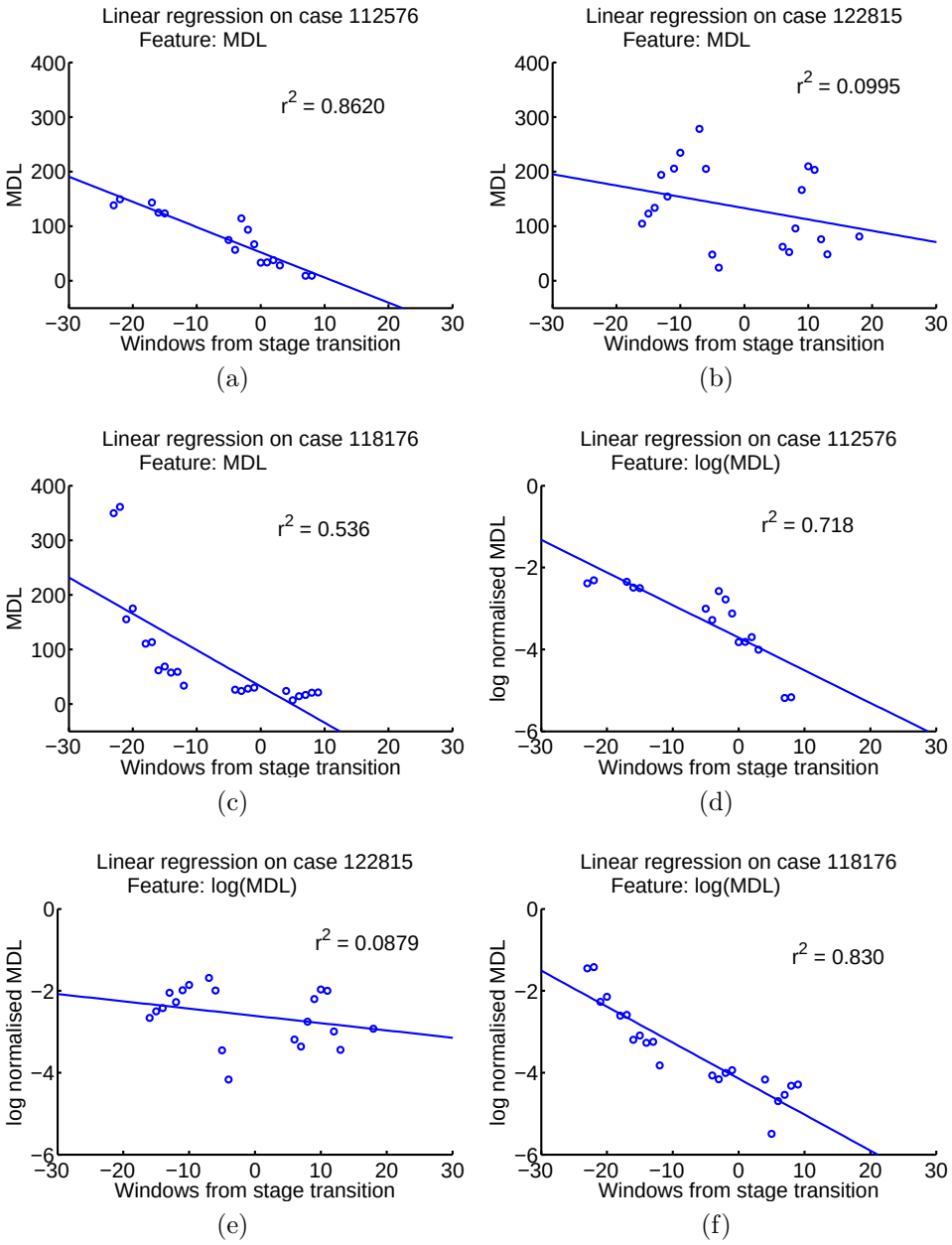


Figure 5.20: Individual regressions for feature MDL in Figures 5.20a to 5.20c and normalised log MDL in Figure 5.20d to 5.20f are shown on three different cases, a case identified as displaying a linear pattern on the left, one showing no discernible pattern in the centre, and an exponential trend on the right. For each regression the r^2 score is also shown.

Table 5.3: The number of cases belonging to each of the three observed trends, and the proportion of each group which were identified as low pH (< 7.1). The exponential group had roughly double the proportion of low pH cases as the linear and patternless groups. The total number of validated cases was 1,992, of which 169 were of low pH.

r^2 Cut-off		0.5			0.7	
Trend	Linear	Exp	None	Linear	Exp	None
Count	543	228	1221	265	86	1641
% of total	27.3%	11.5%	61.3%	13.3%	4.27%	82.38%
Low pH	8.66%	13.6%	7.45%	9.81%	15.12%	7.92%

calculated. If neither r^2 value exceeded a threshold value, r^2_{min}, that cases was labelled as having no visible trend. If the r^2 value were higher for the linear regression the case was labelled as linear, and when the r^2 for the exponential fit was higher the case was labelled exponential. The results, using two different values of r^2 are shown in Table 5.3. A higher proportion of low pH cases showed an exponential trend than linear or patternless, with increased sensitivity at higher r^2 values. The step change pattern of the exponential group was the expected response observed as described earlier in section 5.4.2.

Groups which showed no pattern were the most common, occurring in 61.3% of cases, which may indicate a fluctuation in the condition of the fetus, that the method has failed to find a significant model, or even that noise which passed through the preprocessing process is affecting the results.

The coefficients found for both fits, and a quadratic fit, were tested to see if differences could be found between the distributions of positive and negative cases. This was achieved using a two-sample KS test. The resulting p-value for each test is shown in table 5.4, with the most significant values for each feature highlighted. Coefficients from the quadratic fit were not used further in this study, as they were observed to

be highly unstable due to the low numbers of data available. A more accurate fit was expected using a second-order term, however the increased separation between classes was not. Without a sufficient number of windows present in most cases, it was decided non-linear trends would not be investigated.

Table 5.4: The p-values for KS tests performed on the distributions of each coefficient modelled using the three regression techniques described in Section 5.4.3. The most significant linear parameter for each feature is highlighted. In the case the most significant parameter overall was part of the quadratic regression this value is presented in italics.

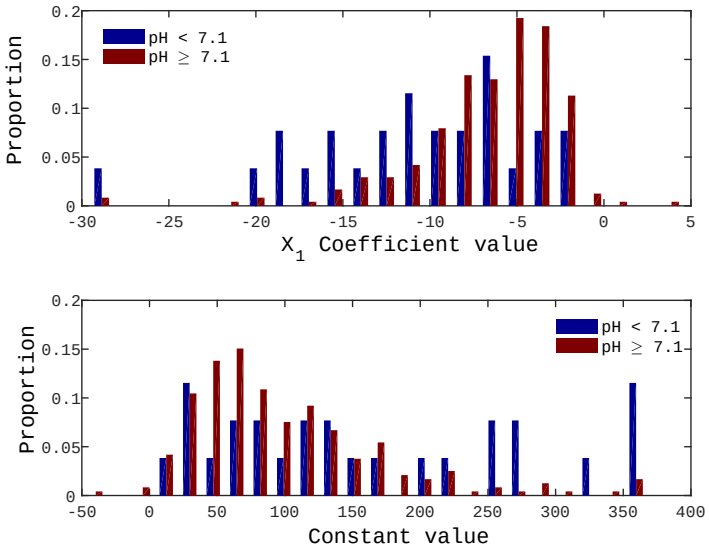
Regression		Delay	Memory	MDL	VAF	RedVaf	comps	Gain	TV	MaxCoef	MaxPos
Linear	c	0.236	0.218	0.002	0.144	0.943	0.743	0.035	0.015	0.023	**0.177**
	x	**0.014**	**0.138**	3.66e-6	**0.013**	**0.302**	0.686	0.006	**7.36e-6**	**0.003**	0.323
	r^2	0.257	0.381	0.047	0.666	0.696	0.827	0.573	0.479	0.975	0.436
Quadratic	c	0.635	0.170	0.054	*0.004*	*0.106*	*0.046*	0.448	0.344	0.293	0.120
	x	0.236	*0.095*	0.007	0.244	0.541	0.491	0.092	0.071	0.066	*0.041*
	x^2	*0.001*	0.176	*4.75e-7*	0.005	0.162	0.462	0.005	2.32e-4	0.017	0.611
	r^2	0.281	0.535	0.008	0.473	0.550	0.897	0.663	0.696	0.642	0.202
Exponential	c	0.168	0.408	0.715	0.288	0.800	0.161	0.029	0.932	0.023	0.592
	x	0.015	0.378	**1.04e-6**	0.028	0.880	**0.156**	**0.004**	1.15e-5	0.003	0.791
	r^2	0.198	0.321	0.124	0.611	0.626	0.567	0.806	0.447	0.998	0.752

The distributions of the regression parameters were plotted for both linear and exponential fits for the most distinctive feature, the MDL, in Figure 5.21. By this stage the number of cases in each group was reduced to 543 linear and 228 exponential cases.

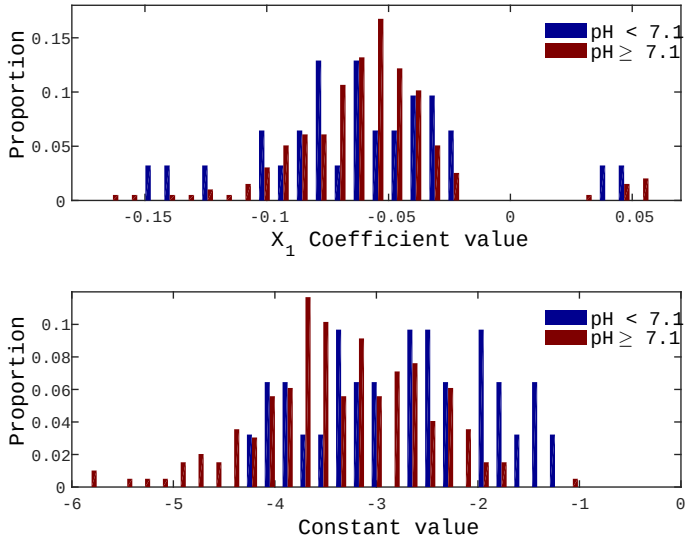
5.4.4 Classification performance

With multiple properties of IRFs and IRF trends demonstrating significantly different values between classes the next logical step is to assess these features as classifier inputs. Two new sets of 6 features were formed. The first, feature set 3, consists of the properties of identified IRFs for the window closest to partum. If the final available window occured more than 30 minutes from partum, the case was declared invalid. These were the gain, delay, VAF, MDL, TV, and maxCoef of IRFs calculated as defined in Section 5.2.3. These values were successfully calculated for 1,941 cases, and are compared against the equivalent cases using feature set 1. The two feature sets were compared using an RF, trained and optimised using the method developed in Chapter 4. The mean results across 10-validation folds are reported in Table 5.5. A second set, feature set 4 was constructed using the most significant IRF parameter trends identified in Table 5.4:

- Delay linear fit X coefficient.
- MDL exponential fit X coefficient.
- VAF linear fit X coefficient.
- Gain exponential fit X coefficient.
- TV linear fit X coefficient.



(a) Histograms of linear coefficients.



(b) Histograms of exponential fit coefficients.

Figure 5.21: When only cases which had achieved an $r^2 \geq 0.5$ were selected, the distribution of parameters values became more distinct between classes. This is demonstrated using the MDL feature. In the top plot, the coefficients of the linear fits are shown, coloured by class. In the bottom plot, the coefficients of the exponential fit are shown. For the constant of the exponential fit parameter in particular a strong distinction between values is seen.

Table 5.5: The median classification results across 10-folds using the benchmark RF classifier and two featuresets. Feature set 1 are 6 clinical parameters selected previously, feature set 3 being a heuristic selection of SI parameters listed in Section 5.4. This experiment was conducted using the OCFMT database with positive cases defined as having pH < 7.1.

Measure	Feature set 1	Feature set 3
Accuracy	0.721	0.700
Sensitivity	0.432	0.400
Specificity	0.737	0.722
Kappa	0.085	0.046
F-meas	0.192	0.162

Table 5.6: The median classification results across 10-folds using the benchmark RF classifier and two featuresets. Feature set 1 are 6 clinical parameters selected previously, feature set 4 were the parameters of the feature regressions defined in Section 5.4.3. This experiment was conducted using the validated cases from the OCFMT database, with positive cases defined as having pH < 7.1. Of the original 7,538 cases 1,992 were validated at this stage.

Measure	Feature set 1	Feature set 4
Accuracy	0.714	0.700
Sensitivity	0.471	0.454
Specificity	0.745	0.725
Kappa	0.087	0.072
F-meas	0.208	0.194

- Max Value linear fit X coefficient.

This new feature set was again compared to feature set 1 using only the cases which were successfully modelled by the SI method and additionally contained at least 5 valid windows to calculate trend fits. The resultant set contained 1,992 cases. Again the benchmark RF classifier was trained an optimised using the method in Chapter 4. The results are presented in Table 5.6. The new feature set performs almost as well as the existing set, even including that fact that in most cases no meaningful trend was

found, and improves upon the performance of feature set 3. Potentially, this result could be improved by selecting only cases which achieved a strong fit to one of the linear regressions.

5.5 Conclusions

A method of SI was successfully implemented, and the discovered IRF features used to classify cases in the OCFMT database. The results demonstrated significant differences in these features between classes, which were captured using the established benchmark classification method. These features were not strongly correlated with existing features indicating they may potentially be used to improve the performance of existing methods. As individual datasets these feature did not outperform existing features, though they have been demonstrated in the literature to outperform a feature set constructed using Heart Rate Variability and the FHR baseline [194].

This style of investigation would not have been possible without access to a database of this size. Investigations into class differences are hampered still by the small quantity of positive cases, which resulted in a more liberal measure, pH < 7.1 being used to define the positive class.

The linear trends observed when signals were aligned according to stage, and the disappearance of the dip in the in the gain feature were surprising, but explained as the total effect of three combined patterns - cases with no pattern, which represented the majority, the linear cases, and those with an exponential offshoot after the onset of stage 2 labour. The high number of classes showing no visible pattern suggests that these cases were inadequately modelled, as it is expected that nearby IRFs are similar.

Linear cases can be interpreted as easier births, which did not respond to the

increased pressure of stage 2 labour. Cases with exponential trends appear to be not coping with labour, with increasingly severe responses after an initial delay, both in terms of deceleration strength, captured by the gain feature, but becoming increasingly non-linear, captured by the MDL feature.

Using the most discriminative fit parameters to classify results showed a performance that matched the performance of the established clinical feature set, however by this point a substantial number of cases (74%) had been removed for either failing the SI identification process (57%) or for having too few windows for trend fitting (17%). Within the trend subgroups the fit features showed different distributions for the two classes, however by this stage the dataset had been diminished to 771 cases.

These methods have several shortfalls which must be addressed if they are to ever be considered in the decision making process. The first is the large number of cases for which no IRF could successfully be found, even with relaxed restrictions on window quality. The second is the dependence of trend analyses on the timing of the second stage of labour. In practice this would mean no analysis could be performed until the onset of stage two labour. Whilst intervention in this stage is common, any method dependent on trends will have a very short operational window.

Chapter 6

Conclusions and Discussion

6.1 Summary

In this work classification and system identification methods observed in the literature were successfully implemented and compared using the OCFMT dataset. The findings and contributions of interest from this investigation are summarized below:

- **The impact of training data class imbalance** was explored in Section 4.3 and its impact on individual classifier performance demonstrated. An optimal approach to data rebalancing to maximise the expected Cohen's kappa value was found to be 2:1 (healthy:adverse) with 100% oversampling or 4:1 with 300% oversampling to maximise sensitivity. Little difference was observed between using SMOTE and re-sampling.
- **A strong general classifier of CTG data** was discovered in Section 4.4 by assessing the performance of six methods across five datasets. The RF was significantly better than all methods except the SVM, and is recommended as a benchmark method against which future developments should be evaluated.
- **The impact of thresholding based on classifier output** was investigated in Section 4.2.6. Two potential methods to threshold class decisions were developed using absolute values and individual cases variance. Both methods demonstrated an improvement to the classifier accuracy, at the cost of leaving roughly a third of cases unclassified.
- **The information contained in the time-series** was investigated using an HMM. The average performance of the HMM was comparable with the performance of the benchmark classification method, but with a smaller distribution, demonstrating that using time-series information may provide more stable estimates of class than stationary methods.

- **A method of SI was implemented on externally collected data** in Chapter 5 using the OCFMT database. Cases identified as adverse were observed as following a different trend through labour, with stronger and less delayed responses as birth approached.
- **The importance of the timing of the second stage** was demonstrated in Section 5.4.2. When cases were retrospectively aligned according the onset of stage two labour significant differences in the trends followed by healthy and adverse cases were observed. Trends in individual case parameters were found and divided into three subgroups, linear, exponential, and patternless. When divided in this fashion adverse cases were much more likely to display an exponential trend.
- **SI features were evaluated as classifier inputs** in Section 5.4.4. These features were found to not be strongly correlated with existing features, achieving a classification performance comparable to existing feature sets.

6.2 Future Work

Though a substantial section of the process of classification was assessed when comparing methodologies in this Thesis, isolating the effect of classifier selection alone a difficult task. Two sections of the classification process were ignored, hyperparameter optimisation and the selection of features. Though the selected feature set was chosen based on an heuristic assessment the method does run some risk of bias towards classifiers favoured on these features. Assessing and the impact of feature selection and selection techniques on each classifiers will be a second important step in developing a stronger benchmark.

The impact of including time series information was promising, though the HMM model may be inappropriate for CTG feature classification. The HMM assumes distinct states with different means and variances, not the slowly evolving features observed in CTG time series. Time series models which do not make these same assumptions about the data may be more successful.

New classification features derived from model IRF parameters were introduced to encapsulate time series information, and were found to not be strongly correlated with any existing features. However for a substantial number of cases these features could not be found, which limits the potential application of the method. Alternatives to the CTG which are being developed must be considered, most notably abdominally acquired fECG which may soon provide signal quality comparable to internally acquired data. This, and other technologies will likely have a major impact on the future of fetal assessment during labour.

Appendices

Appendix A

List of OxSys features

These features are extracted using the OxSys algorithms. Details of calculation of the STV tracker can be found in Section 3.1 and the SSI in the literature [59]. Features are listed according to their index in the software, with their type category defined as **M** - morphological, **T** - time domain, **F** - frequency domain and **N** - non-linear.

Standard features

1. **M** Baseline mean
2. **M** % Zero difference between neighbours
3. **N** Approximate entropy
4. **T** Signal stability index (SSI)
5. **M** Minimal expected value
6. **T** Skewness of SSI Kernel density estimate (KDE)
7. **T** Kurtosis of SSI KDE
8. **M** Long-term variability
9. **T** SSI of the residual signal
10. **T** Median of STV change tracker
11. **T** SSI of STV tracker

12. **T** Range of STV tracker
13. **T** STV tracker trend
14. **T** Median of the absolute derivative values of the STV tracker

Contraction Features

15. **M** Number of contractions
16. **M** Median of contraction duration
17. **M** Resting/contraction time ratio

Acceleration features

19. **T** Median STV tracker including accelerations
20. **M** Number of accelerations
21. **M** Mean acceleration duration
22. **M** Mean acceleration amplitude

Deceleration Features

23. **M** Number of decelerations
24. **M** Mean deceleration duration
25. **M** Maximal deceleration duration
26. **M** Median deceleration duration
27. **M** Maximal deceleration amplitude
28. **M** Mean time to deceleration nadir (t_1)
29. **M** Mean time to recovery (t_2)
30. **M** Maximal time to recovery
31. **M** Number of quick recoveries ($t_1 \geq t_2$)
32. **M** Number of slow recoveries ($t_2 > t_1$)
33. **M** Resting/deceleration time ratio

34. **M** Mean deceleration area
35. **M** Total number of beats lost
36. **M** Median onset slope
37. **M** Median recovery slope
38. **M** Maximal baseline shift after decel
39. **M** Number of prolonged decelerations ($\geq 3minutes$)

Overshoot Features

40. **M** Number of decelerations with right and/or left shoulders
41. **M** Number of decelerations with only right shoulders

Lag Features

42. **M** Number of early decelerations
43. **M** Number of late decelerations
44. **M** Number of variable decelerations
45. **M** Median recovery time after contraction end
46. **M** Median lag time (for late decelerations only)
47. **M** Deceleration/contraction ratio

Identified by Fulcher et al. [55]

48. **N** Mutual information
49. **N** Ratio of STD
50. **N** Mean of local approximate entropy
51. **N** STD of local sample entropy
52. **N** Goodness of exponential fit
53. **N** 2-dim time delay embedding space
54. **N** MAD

55. **N** (STD/Mean) squared
56. **N** Alphabet feature

Identified by Xu et al. [202]

57. **N** SVM output
58. **T** Interquartile smoothed signal
59. **N** STD of Gaussian filter

Sinusoidal FHR pattern classifiers [149]

60. **N** Regmeas - Sinusoidal pattern classifier output

Phase Rectified Signal Averaging (PRSA) features [58]

61. **N** PRSA - Decelerative component
62. **N** Bivariate PRSA (BPRSA) - Standard deviation of accelerative component
63. **N** BPRSA - Decelerative component

CTG guideline classes [5]

64. **N** ACOG class (1-3 = normal/suspicious/pathological)

Appendix B

List of UCI Database Features

The UCI database does not provide raw CTG tracings, but a single pre-calculated feature vector for each case. Cases are labelled according to a three-tiered clinician assessment as normal, suspicious, or pathological. Details of the features can be found in the literature [9]. Features are labelled with the same class codes as the OxSys featureset in Appendix A.

1. **M** FHR baseline (bpm)
2. **M** Number of accelerations per second
3. **M** Number of fetal movements per second
4. **M** Number of uterine contractions per second
5. **M** Number of light decelerations per second
6. **M** Number of severe decelerations per second
7. **M** Number of prolongued decelerations per second
8. **M** Percentage of time with abnormal short term variability
9. **M** Mean value of short term variability
10. **M** Percentage of time with abnormal long term variability
11. **M** Mean value of long term variability

12. **T** Width of FHR histogram
13. **T** Minimum of FHR histogram
14. **T** Maximum of FHR histogram
15. **T** Number of histogram peaks
16. **T** Number of histogram zeros
17. **T** FHR histogram mode
18. **T** FHR histogram mean
19. **T** FHR histogram median
20. **T** FHR histogram variance
21. **T** FHR histogram tendency
22. **N** CLASS - FHR pattern class code

Bibliography

[1] J. Adam. The future of fetal monitoring. *Reviews in Obstetrics and Gynecology*, 5:e132–6, 2012.

[2] S. K. Agrawal, F. Doucette, R. Gratton, B. Richardson, and R. Gagnon. Intrapartum computerized fetal heart rate parameters and metabolic acidosis at birth. *Obstet. Gynecol.*, 102(4):731–8, 2003.

[3] Z. Alfirevic, D. Devane, and G. M. L. Gyte. Continuous cardiotocography CTG as a form of electronic fetal monitoring EFM for fetal assessment during labour. *Cochrane Database of Systematic Reviews*, 19(3), 2006. CD006066.

[4] Victoria M. Allen, Thomas F. Baskett, Colleen M. OConnell, Dolores McKeen, and Alexander C. Allen. Maternal and perinatal outcomes with increasing duration of the second stage of labor. *Obstet. Gynecol.*, 113(6):1248–58, 2009.

[5] American College of Obstetrics and Gynecologists. ACOG practice bulletin no. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. *Obstet. Gynecol.*, 114:192–202, 2009.

[6] L. Armstrong and B. Stenson. Use of umbilical cord blood gas analysis in the assessment of the newborn. *Archives of Disease in Childhood: Fetal and Neonatal Edition*, 92(6):F430–4, 2007.

[7] D. Ayres-de Campos. The need for an international consensus on CTG interpretation. Presentation, March 2015. 1st Signal Processing and Monitoring (SPAM) in Labour Workshop.

[8] D. Ayres-de Campos and J. Bernandes. Twenty-five years after the FIGO guidelines for the use of fetal monitoring: Time for a simplified approach? *Int. J. Gynaecol. Obstet.*, 110(1):1–6, 2010.

[9] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de Sa, and L. Pereira-Leite. SisPorto 2.0: A program for automated analysis of cardiotocograms. *J. Matern. Fetal Med.*, 9(5):311–8, 2000.

[10] D. Ayres-de Campos, P. Sousa, A. Costa, and J. Bernardes. Omniview-SisPorto 3.5 - A central fetal monitoring station with online alerts based on computerized cardiotocogram + ST event analysis. *J. Perinat. Med.*, 36(3):260–4, 2008.

[11] P. C. A. M. Bakker, G. J. Colenbrander, A. A. Verstraeten, and H. P. van Geijn. The quality of intrapartum fetal heart rate monitoring. *Eur. J. Obstet. Gynecol. Reprod. Biol.*, 116(1):22–7, 2004.

[12] J. H. Becker, L. Bax, I. Amer-Wahlin, K. Ojala, C. Vayssiere, M. E. Westerhuis, B. W. Mol, G. H. Visser, K. Maršál, A. Kwee, and K. G. Moons. ST analysis of the fetal electrocardiogram in intrapartum fetal monitoring: A meta-analysis. *Obstet. Gynecol.*, 119(1):145–54, Jan 2012.

[13] L. Bennet, J. Westgate, P. D. Gluckman, and A. J. Gunn. Fetal responses to asphyxia. In D. K. Stevenson, W. E. Benitz, and P. Sunshine, editors, *Fetal and neonatal brain injury*, pages 83–110. Cambridge University Press, 3rd edition, 2003.

[14] S. Berglund, C. Grunewald, H. Pettersson, and S. Cnattinguis. Severe asphyxia due to delivery-related malpractice in Sweden 1990-2005. *Br. J. Obstet. Gynaecol.*, 115(3):316–23, 2008.

[15] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.

[16] C. M. Bishop. *Pattern recognition and machine learning*. Springer Scientific, 2009.

[17] S. C. Blackwell, W. A. Grobman, L. Antoniewicz, M. Hutchinson, and C. Gyamfi Bannerman. Interobserver and intraobserver reliability of the NICHD 3-tier fetal heart rate interpretation system. *Am. J. Obstet. Gynecol.*, 205(4):1–5, Oct 2011.

[18] A. Boardman, F. S. Schlindwein, N. V. Thakor, T. Kimura, and R. G. Geocadin. Detection of asphyxia using heart rate variability. *Med. Biol. Eng. Comput.*, 40(6):618–24, 2002.

[19] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[20] P. Brocklehurst, P. Hardy, J. Hollowell, L. Linsell, A. Macfarlane, C. McCourt, N. Marlow, A. Miller, M. Newburn, S. Petrou, et al. Perinatal and maternal

outcomes by planned place of birth for healthy women with low risk pregnancies: the birthplace in England national prospective cohort study. *BMJ: British Medical Journal*, 343(7840), 2011.

[21] R. Caruana and A. Niculescu-Mizil. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 69–78. ACM, 2004.

[22] R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–8. ACM, 2006.

[23] S. Cazares. *Automated Identification of abnormal patterns in the intrapartum cardiotocogram*. Ph.D. Thesis, University of Oxford, 2002.

[24] S. Cazares, L. Tarassenko, L. Impey, M. Moulden, and C. W. G. Redman. Automated identification of abnormal cardiotocograms using neural network visualization techniques. In *Proceedings of the 23rd Annual International Conference of the IEEE on Engineering in Medicine and Biology Society, 2001.*, volume 2, pages 1629–1632, 2001.

[25] Centre for Maternal and Child Enquiries. Perinatal mortality 2009. www.publichealth.hscni.net/sites/default/files/Perinatal%20Mortality%202009.pdf, 2009. Accessed 20/08/15.

[26] M. Cesarelli, M. Romano, and P. Bifulco. Comparison of short term variability indexes in cardiotocographic foetal monitoring. *Comput. Biol. Med.*, 39(2):106–18, 2009.

[27] Yvonne W Cheng, Linda M Hopkins, and Aaron B Caughey. How long is too long: Does a prolonged second stage of labor in nulliparous women affect maternal and neonatal outcomes? *Am. J. Obstet. Gynecol.*, 191(3):933–8, 2004.

[28] Yvonne W Cheng, Linda M Hopkins, Russell K Laros, and Aaron B Caughey. Duration of the second stage of labor in multiparous women: Maternal and neonatal outcomes. *Am. J. Obstet. Gynecol.*, 196(6):585–e1, 2007.

[29] V. Chudáček, J. Spilka, M. Bură, P. Janku, L. Hruban, M. Huptych, and L. Lhotska. Open access intrapartum CTG database. *BMC Pregnancy and Childbirth*, 14(16), 2014.

[30] V. Chudáček, J. Spilka, P. Janku, M. Koucky, and L. Lhotska. Automatic evaluation of intraprtum fetal heart rate recordings: A comprehensive analysis of useful features. *Physiol. Meas.*, 32(8):1347–60, 2011.

[31] D. Y. Chung, Y. B. Sim, K. T. Park, S. H. Yi, J. C. Shin, and S. P. Kim. Spectral analysis of fetal heart rate variability as a predictor of intrapartum fetal distress. *Int. J. Gynaecol. Obstet.*, 73(2):109–16, May 2001.

[32] S. L. Clark, M. P. Nageotte, T. J. Garite, R. K. Freeman, D. A. Miller, K. R. Simpson, M. A. Belfort, G. A. Dildy, J. T. Parer, and R. L. et al. Berkowitz. Intrapartum management of category II fetal heart rate tracings: Towards standardization of care. *Am. J. Obstet. Gynecol.*, 209(2):89–97, 2013.

[33] G. Clifford, R. Sameni, J. Ward, J. Robinson, and A. J. Wolfberg. Clinically accurate fetal ECG parameters acquired from maternal abdominal sensors. *Am. J. Obstet. Gynecol.*, 205(1):47.e1–5, 2011.

[34] W. R. Cohen, S. Ommani, S. Hassan, F. G. Mirza, M. Solomon, R. Brown, B. S. Schifrin, J. Himsworth, and B. R. Hayes-Gill. Accuracy and reliability of fetal heart rate monitoring using maternal abdominal surface electrodes. *Acta Obstetrica et Gynaecologica Scandinavica*, 91(11):1306–13, 2012.

[35] A. Costa, D. Ayres-de Campos, F. Costa, C. Santos, and J. Bernardes. Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals. *Am. J. Obstet. Gynecol.*, 201(5):1–6, Nov 2009.

[36] A. Costa, C. Santos, D. Ayres-de Campos, C. Costa, and J. Bernardes. Access to computerised analysis of intrapartum cardiotocographs improves clinicians' prediction of newborn umbilical artery blood pH. *Br. J. Obstet. Gynaecol.*, 117(10):1288–93, Sep 2010.

[37] C. Costa-Sandos, J. Bernandes, P. M. B. Vitanyi, and L. Antunes. Clustering fetal heart tracings by compression. In *19th IEEE Symposium on Computer-Based Medical Systems*, 2006.

[38] Pedro M. Sá Couto, Willem L. van Meurs, Jo ao F. Bernardes, Joaquim P. Marques de Sá, and Jane A. Goodwin. Mathematical model for educational simulation of the oxygen delivery to the fetus. *Control Engineering Practice*, 10(1):59–66, 2002.

[39] G. S. Dawes, M. Moulden, and C. W. G. Redman. System 8000: computerized antenatal FHR analysis. *J. Perinat. Med.*, 19(1-2):47–51, 1991.

[40] N. W. Dawes, G. S. Dawes, M. Moulden, and C. W. Redman. Fetal heart rate patterns in term labor vary with sex, gestational age, epidural analgesia, and fetal weight. *Am. J. Obstet. Gynecol.*, 180(1):181–7, Jan 1999.

[41] H. H. de Haan, A. J. Gunn, and P. D. Gluckman. Fetal heart rate changes do not relflect cardiovascular deterioration during brief repeated umbilical cord occlusions in near-term fetal lambs. *Am. J. Obstet. Gynecol.*, 1(1):8–17, 1997.

[42] E. Declercq, R. Young, H. Cabral, and J. Ecker. Is a rising ceasarean delivery rate inevitable? Trends in industrialized countries, 1987 to 2007. *Birth*, 38(2):99–104, 2011.

[43] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[44] R. Dugad and U.B. Desai. A tutorial on hidden Markov models. Technical report, Signal Processing and Artificial Neural Networks Laboratory, Department of Electrical Engineering, Indian Institute of Technology, Bombay, 1996.

[45] C. E. East, L. Begg, and P. B. Colditz. Fetal pulse oximetry for fetal assessment in labor. *Cochrane Database of Systematic Reviews*, 2, 2007. CD004075.

[46] C. E. East, L. R. Leader, P. Sheehan, N. E. Henshall, and P. B. Colditz. Intrapartum fetal scalp lactate sampling for fetal assessment in the presence of a non-reassuring fetal heart rate trace. *Cochrane Database of Systematic Reviews*, 17(3), 2010. CD004075.

[47] C. E. East, R. Smyth, L. R. Leader, N. E. Henshall, P. B. Colditz, and K. H. Tan. Vibroacoustic stimulation for fetal assessment in labour in the presence of a nonreassuring fetal heart rate trace. *Cochrane Database of Systematic Reviews*, 18, 2005.

[48] C. Elliot, P. Warrick, E. M. Graham, and E. Hamilton. Graded classification of fetal heart rate tracings: association with neonatal metabolic acidosis and neurologic morbidity. *Am. J. Obstet. Gynecol.*, 202(3):258.e1–8, 2010.

[49] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–81, 2014.

[50] M. Ferrario, M. G. Signorini, G. Maganes, and S. Cerutti. Comparison of entropy-based regularity estimators: Application to the fetal heart rate signal for the identification of fetal distress. *IEEE Transactions on Biomedical Engineering*, 53(1):119–25, 2006.

[51] FIGO subcommittee on Standards in Perinatal Medicine. Guidelines for the use of fetal monitoring. *Int. J. Gynaecol. Obstet.*, 1987.

[52] R. K. Freeman, T. J. Garite, and M. P. Nageotte. *Fetal heart rate monitoring.* Lippincott, Williams, and Wilkins, 3rd edition, 2003.

[53] Y. Freund. A more robust boosting algorithm. *arXiv preprint*, 2009.

[54] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[55] B. D. Fulcher, A. E. Georgieva, C. W. Redman, and N. S. Jones. Highly comparative fetal heart rate analysis. In *Conf Proc IEEE Eng Med Biol Soc*, pages 3135–8, 2012.

[56] D. Gao, M. Madden, D. Chambers, and G. Lyons. Bayesian ANN classifier for ECG arrhythmia diagnostic system: A comparison study. In *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, volume 4, pages 2383–2388 vol. 4, July 2005.

[57] T. J. Garite, G. A. Dildy, H. McNamura, M. P. Nageotte, F. H. Boehm, E. H. Dellinger, and R. A. Knuppal. A multicenter controlled trial of fetal pulse oximetry in the intrapartum management of non-reassuring fetal heart rate patterns. *Am. J. Obstet. Gynecol.*, 183(5):1049–58, 2000.

[58] A. Georgieva, A. T. Papageorghiou, S. J. Payne, M. Moulden, and C. W. G. Redman. Phase rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidaemia at birth. *Br. J. Obstet. Gynaecol.*, 128(7):889–94, 2014.

[59] A. Georgieva, S. J. Payne, M. Moulden, and C. W. Redman. Computerized fetal heart rate analysis in labor: detection of segments with uncertain baseline. *Physiol Meas*, 32(10), 2011.

[60] A. Georgieva, S. J. Payne, M. Moulden, and C. W. G. Redman. Automated fetal heart rate analysis in labour: Decelerations and overshoots. In *Proceedings of the 36th international conference on the application of mathematics in engineering and economics*, Sozopol, Bulgaria, 2010.

[61] A. Georgieva, S. J. Payne, M. Moulden, and C. W. G. Redman. Artificial neural networks applied to fetal monitoring in labour. *Neural Computing and Application*, 22(1):85–93, Jan 2011.

[62] A. Georgieva, S. J. Payne, and C. W. G. Redman. Computerized electronic fetal heart rate monitoring in labor: Automated contraction identification. *Medical Biological Engineering Computation*, 47(12):1315–1320, 2009.

[63] G. Georgoulas, D. Gavrils, I. G. Tsoulos, C. Stylios, J. Bernandes, and P. P. Groumpos. Novel approach for fetal heart rate classification introducting grammatical evolution. *Biomedical Signal Processing and Control*, 2(2):69–79, 2007.

[64] G. Georgoulas, C. Stylios, and P. Groumpos. Integrated approach for classification of cardiotocograms based on independent component analysis and neural networks. In *Proceedings of 11th IEEE Mediterranean conference on Control and Automation*, 2003.

[65] G. Georgoulas, C. Stylios, and P. Groumpos. Invesitgation and comparison of different scale dependent features for fetal heart rate classification. In *Proc 16th IFAC World Congress*, Prague, Czech Republic, 2005.

[66] G. Georgoulas, C. Stylios, and P. Groumpos. Feature extraction and classification of fetal heart rate using wavelet analysis and support vector machines. *International Journal on Artificial Intelligence Tools*, 15(3):411–32, 2006.

[67] G. Georgoulas, C. Stylios, and P. Groumpos. Predicting the risk on metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. *IEEE Transactions on Biomedical Engineering*, 53(5):875–84, 2006.

[68] G. Georgoulas, C. Stylios, E. Papageorgiou, and P. Groumpos. Tuning support vector machines via particle swarm optimization for the classification of fetal

heart rate signals. In *7th Biosignal Conference*, pages 169–71, Brno, Czech Republic, 2006.

[69] G.G. Georgoulas, C.D. Stylios, G. Nokas, and P. P. Groumpos. Classification of fetal heart rate during labour using hidden Markov models. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 3, pages 2471–2475, July 2004.

[70] D. Gibb and S. Arulkumaran. *Fetal monitoring in practice*. Churchill Livingstone, 3rd edition, 2008.

[71] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13).

[72] H. Gonçalves, A. P. Rocha, D. Ayres-de Campos, and J. Bernandes. Internal versus external intrapartum foetal heart rate monitoring: The effect on linear and nonlinear parameters. *Physiol. Meas.*, 27(3):307–19, 2006.

[73] H. Gonçalves, A. P. Rocha, D. Ayres-de Campos, and J. Bernandes. Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery. *Med. Biol. Eng. Comput.*, 44:847–55, 2006.

[74] E. M. Graham, K. A. Ruis, A. L. Lartman, F. J. Northington, and H. E. Fox. A systematic review of the role of intrapartum hypoxia-ischemia in the causation of neonatal encephalopathy. *Am. J. Obstet. Gynecol.*, 199(6):587–95, 2008.

[75] K. R. Greene. The ECG waveform. *Baillieres Clin. Obstet. Gynaecol.*, 1(1):131–55, 1987.

[76] K. R. Greene. Intelligent fetal heart rate computer systems in intrapartum surveillance. *Curr. Opin. Obstet. Gynecol.*, 8(2):123–7, Apr 1996.

[77] D. A. Grimes and J. F. Peipert. Electronic fetal monitoring as a public health screening program: The arithmetic of failure. *Obstet. Gynecol.*, 116(6):1397–400, 2010.

[78] A. W. Grogono. Acid-base tutorial. www.acid-base.com, 2010. Accessed 20/08/15.

[79] A. J. Gunn, J. T. Parer, E. C. Mallard, C. E. Williams, and P. D. Gluckman. Cerebral histologic and electrocorticographic changes after asphyxia in fetal sheep. *Pediatr. Res.*, 31(5):486–91, 1992.

[80] H. Hannah Inbarani, P.K. Nizar Banu, and AhmadTaher Azar. Feature selection using swarm-based relative reduct technique for fetal heart rate. *Neural Computing and Applications*, 25(3-4):793–806, 2014.

[81] D. J. Harrington, C. W. Redman, M. Moulden, and C. E. Greenwood. The long-term outcome in surviving infants with Apgar zero at 10 minutes: A systematic review of the literature and hospital based cohort. *Am. J. Obstet. Gynecol.*, 196(5):463.e1–5, 2007.

[82] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer, 10th edition, 2013.

[83] B. R. Hayes-Gill, S. Hassan, F. G. Mirza, S. Ommani, J. Himsworth, M. Solomon, R. Brown, B. S. Schifrin, and W. R. Cohen. Accuracy and reliability of uterine contraction identification using abdominal surface electrodes. *Clinical Medicine Insights: Women's Health*, 5:65–75, 2012.

[84] E. H. Hon. The electronic evaluation of the fetal heart rate: Preliminary report. *Obstet. Gynecol. Surv.*, 13(5):654–56, 1958.

[85] M. Huang and Y. Hsu. Fetal distress prediction using discriminant analysis, decision tree, and artificial neural network. *Journal of Biomedical Science and Engineering*, 5(9):526–33, 2012.

[86] K. Hunt. *Fetal heart monitoring: Principles and practices.* Association of Women's Health, Obstetric and Neonatal Nurses, 1993.

[87] T. Ikeda, Y. Murata, E. J. Quilligan, J. T. Parer, I. M. Theunissen, P. Cifuentes, S. Doi, and S. D. Park. Fetal heart rate patterns in postasphyxiated fetal lambs with brain damage. *Am. J. Obstet. Gynecol.*, 179(5):1329–37, Nov 1998.

[88] M. Jezewski, J. Wrobel, P. Labaj, J. Leski, N. Henzel, K. Horoba, and J. Jezewski. Some practical remarks on neural networks approach to fetal cardiotocograms classification. In *Proc. 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society EMBS 2007*, pages 5170–5173, 2007.

[89] M. Johnsson, S. Norden-Lindeburg, I. Ostlund, and U. Hanson. Metabolic acidosis at birth and suboptimal care - Illustration of the gap between knowledge and clinical practice. *Br. J. Obstet. Gynaecol.*, 116(11):1453–60, 2009.

[90] H. W. Jongsma and J.G. Nijhuis. Classification of fetal and neonatal heart rate patterns in relation to behavioural states. *Eur J Obstet Gynecol Reprod Biol*, 21(5-6):293–9, 1986.

[91] B. Juang. Maximum likelihood estimation for mixture multi-variate stochastic observations of Markov chains. *AT & T Techical Journal*, 64(6):1235–49, 1985.

[92] B. Juang, S. Levinson, and M. Sondhi. Maximum likelihood esimation for multivariate mixture observations of Markov chains. *IEEE Transactions on Information Theory*, 32(2):307–9, 1986.

[93] B. H. Juang and L. R. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–72, 1991.

[94] B.H. Juang and L.R. Rabiner. The segmental K-Means algorithm for estimating the parameters of hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38:1639–41, 1990.

[95] K2 Medical Systems. Intelligent fetal assessment. www.k2ms.com/iFMTS-INFANT.html. Accessed August 2012.

[96] E. M. Karabulut and T. Ibrikci. Analysis of cardiotocogram data for fetal distress: Determination by decision tree based adaptive boosting approach. *Journal of Computer and Communications*, 2:32–7, 2014.

[97] R. D. Keith and K. R. Greene. Development, evaluation and validation of an intelligent system for the management of labour. *Baillieres Clin. Obstet. Gynaecol.*, 8(3):583–605, Sep 1994.

[98] R. D. Keith, J. Westgate, E. C. Ifeachor, and K. R. Greene. Suitability of artificial neural networks for feature extraction from cardiotocogram during labour. *Med. Biol. Eng. Comput.*, 32(4 Suppl):S51–57, Jul 1994.

[99] A. Kikuchi, N. Unno, T. Horikoshi, T. Shimizu, S. Kozuma, and Y. Taketani. Changes in fractal features of fetal heart rate during pregnancy. *Early Hum. Dev.*, 81(8):655–61, 2005.

[100] T. Kiserud. Physiology of the fetal cirulation. *Semin. Fetal. Neonatal Med.*, 10:493–503, 2005.

[101] T. Kiserud and G. Acharya. The fetal circulation. *Prenatal Diagnosis*, 24(13):1049–59, 2004. PMID 15614842.

[102] A. Krogh, I. S. Mian, and D. Haussler. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Research*, 22(22):4768–79, 1994.

[103] K. Kruger, B. Hallberg, M. Blennow, M. Kublickas, and M. Westgren. Predictive value of fetal scalp blood lactate concentration and pH as markers of neurologic disability. *Am. J. Obstet. Gynecol.*, 181(5):1072–8, 1999.

[104] N. Krupa, M. A. Ma, E. Zahedi, S. Ahmed, and F. M. Hassan. Antepartum fetal heart rate feature extraction and classification using empirical mode decomposition and support vector machine. *Biomed. Eng. Online*, 10(1):6, 2011.

[105] A. Kwee, A. H. S. Dekkers, H. P. J. van Wijk, C. W. van der Hoorn-van den Beld, and G. H. A. Visser. Occurance of ST-changes recorded with the STAN S21 monitor during normal and abnormal fetal heart rate patterns during labour. *Eur. J. Obstet. Gynecol. Reprod. Biol.*, 135(1):28–34, 2007.

[106] J. Lan, M.Y. Hu, and E. Patuwo. Neural network classifiers with uneven misclassification costs and imbalanced group sizes. In K. D. Lawrance, S. Kudyba, and R. K. Klimberg, editors, *Data mining methods and applications*, pages 61–82. CRC Press, 2007.

[107] A. Lee, C. Ulbricht, and G. Dorffner. Application of artificial neural networks for detection of abnormal fetal heart rate pattern: A comparison with conventional algorithms. *J. Obstet. Gynaecol.*, 19(5):482–5, Sep 1999.

[108] K. Leslie and S. Arulkumaran. Intrapartum fetal surveillance. *Obstetric Gynecology and Reproductive Medicine*, 21(3):59–63, 2011.

[109] M. Lichman. UCI machine learning repository. archive.ics.uci.edu/ml, 2013. Accessed 20/08/15.

[110] L. A. Liporace. Maximum likelihood estimation for multi-variate mixture observations of Markov sources. *IEEE Transactions on Information Theory*, 28(5):729–34, 1982.

[111] L. Ljung. *System identification: Theory for the user*. Prentice Hall PTR, 2nd edition, 1999.

[112] J. A. Low. Intrapartum fetal asphyxia: Definition, diagnosis, and class. *Am. J. Obstet. Gynecol.*, 176(5):957–9, 1997.

[113] J. A. Low, R. Victory, and J. E. Derrick. Predictive value of electronic fetal monitoring for intrapartum fetal asphyxia with metabolic acidosis. *Obstet. Gynecol.*, 93(2):285–91, 1999.

[114] D.M.C. MacKay. *Information theory, inference, and learning Algorithms*. Cambridge University Press, 2003.

[115] G. A. Macones, G. D. Hankins, C. Y. Spong, J. Hauth, and T. Moore. The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: Update on definitions, interpretation, and research guidelines. *Obstet. Gynecol.*, 112(3):661–6, Sep 2008.

[116] K. Maeda, Y. Noguchi, F. Matsumoto, and T. Nagasawa. Quantitive fetal heart rate evaluation without pattern classification: FHR score and artificial neural network analysis. *Network*, 12(3-4):127–41, 2010.

[117] K. Maeda, M. Utsu, A. Makio, M. Serizawa, Y. Noguchi, T. Hamada, K. Mariko, and F. Matsumoto. Neural network computer analysis of fetal heart rate. *Journal of Maternal Fetal Investigation*, 8(4):163–71, Dec 1998.

[118] K. Maeda, M. Utsu, Y. Noguchi, F. Matsumoto, and T. Nagasawa. Central computerized automatic fetal heart rate diagnosis with a rapid and direct alarm system. *The Open Medical Devices Journal*, 4:28–33, 2012.

[119] Mindchild Medical, Inc. Mindchild company webpage. www.mindchild.com/index.html, 2015. Accessed 20/08/15.

[120] Monica Healthcare, Ltd. Company webpage. www.monicahealthcare.com, 2015. Accessed 15/09/15.

[121] D. Moster, R. T. Lie, L. M. Irgens, T. Bjerkedal, and T. Markestad. The association of Apgar score with subsequent death and cerebral palsy: A population-based study in term infants. *J. Pediatr.*, 138(6):798–803, 2001.

[122] K. Muprhy. Hidden Markov model (HMM) toolbox. Online, 1998. Accessed 20/08/12.

[123] M. Murray. *Antepartum and intrapartum fetal monitoring*. Springer Publishing Company, 3rd edition, 2006.

[124] Ian Nabney. *Netlab: Algorithms for pattern recognition*. Springer Science & Business Media, 2002.

[125] G. Natarajan, S. Shankaran, A. R. Laptook, C. M. Bann, S. A. McDonald, A. Das, R. D. Higgins, S. R. Hintz, and B. R. Vohr. Apgar scores at 10 min and outcomes at 6-7 years following hypoxic-ischemic encephalopathy. *Arch. Dis. Child. Fetal Neonatal Ed.*, 98(6):F473–9, 2013.

[126] Neoventa Medical. STAN S31 product page. www.neoventa.com/products/stan, August 2012. Accessed 20/08/15.

[127] NHS. Confidential enquiry into intrapartum related deaths, 2010. Accessed 28/07/2015.

[128] NHS. NHS report on the Caesarean section, August 2012. Accessed 29/07/2015.

[129] NHS. Ten years of maternity claims: An analysis of NHS litigation authority data, October 2012. Accessed 12/06/15.

[130] NICE. NICE clinical guideline 55: intrapartum care: care of healthy women and their babies during childbirth. Online, 2007. Accessed 06/05/15.

[131] NICE. NICE Clinical Guideline 132: Caesarean section. Online, 2011. Accessed 05/06/15.

[132] NICE. NICE clinical guideline 190: Intrapartum care: care of healthy women and their babies during childbirth. Appendix A: Adverse outcomes. Online, 2014. Accessed 17/08/15.

[133] P. V. Nielsen, B. Stigsby, C. Nickelsen, and J. Nim. Intra- and inter-observer variability in the assessment of intrapartum cardiotocograms. *Acta Obstetricia et Gynecologica Scandinavica*, 66(5):421–4, 1987.

[134] Y. Noguchi, F. Matsumoto, K. Maeda, and T. Nagasawa. Neural network analysis and evaluation of the fetal heart rate. *Algorithms*, 2(1):19–30, 2009.

[135] H. Noren and A. Carlsson. Reduced prevalence of metabolic acidosis at birth: An analysis of established STAN usage in the total population of deliveries in a Swedish district hospital. *Am. J. Obstet. Gynecol.*, 202(6):546.e1–e7, 2010.

[136] I. Nunes, D. Ayres-de Campos, A. Ugwumadu, P. Amin, P. Banfield, A. Nicoll, S. Cunningham, P. Sousa, C. Costa-Santos, and J. Bernandes. FM-ALERT: a randomised clinical trial of intrapartum fetal monitoring with computer analysis and alerts versus previously available monitoring. In 2^{nd} *European Congress on Intrapartum Care*, Porto, Lisbon, 2015.

[137] H. Ocak. A medical decision support system based on support vector machines and the genetic algorithm for the evaluation of fetal well-being. *J. Med. Syst.*, 37(2):1–9, 2013.

[138] H. Ocak and H. M. Ertunc. Prediction of fetal state from the cardiotocogram recordings using adaptive neuro-fuzzy inference systems. *Neural Computing and Applications*, 23(6):1583–9, 2013.

[139] E. Odding, M. E. Roebroeck, and H. J. Stam. The epidemiology of cerebral palsy: Incidence, impairments and risk factors. *Disabil. Rehabil.*, 28(4):183–91, 2006.

[140] OpenStax College. Fetal development. http://cnx.org/contents/29b65785-b2c5-4bec-9101-944f05f3c6af@3/Fetal-Development. Accessed 10/08/15.

[141] J. Pardey, M. Moulden, and C. W. Redman. A computer system for the numerical analysis of nonstress tests. *Am. J. Obstet. Gynecol.*, 186(5):1095–103, May 2002.

[142] J. T. Parer and E. F. Hamilton. Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation. *Am. J. Obstet. Gynecol.*, 203(5):451.e1–7, 2010.

[143] Perigen. Perigen company website. www.perigen.com. Accessed 01/09/15.

[144] S. M. Pincus and R. R. Viscarello. Approximate entropy: a regularity measure for fetal heart rate analysis. *Obstet. Gynecol.*, 79:249–55, 1992.

[145] Joaquın Pizarro, Elisa Guerrero, and Pedro L Galindo. Multiple comparison procedures applied to model selection. *Neurocomputing*, 48(1):155–73, 2002.

[146] L. Prechelt. A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. *Neural Networks*, 9(3):457–62, 1996.

[147] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–86, Feb 1989.

[148] RCOG. Is it time for UK obstetricians to accept fetal scalp lactate as an alternative to scalp pH? (Scientific Impact Paper No.47). Online, 2015.

[149] A. Reddy, M. Moulden, and C. W. Redman. Antepartum high-frequency fetal heart rate sinusoidal rhythm: computerized detection and fetal anemia. *Am. J. Obstet. Gynecol.*, 200:407.e1–6, 2009.

[150] J. Reinhard, B. R. Hayes-Gill, S. Schiermeier, H. Hatzmann, T. M. Heinrich, and F. Louwen. Intrapartum heart rate ambiguity: A comparison of cardiotocogram and abdominal fetal electrocardiogram with maternal electrocardiogram. *Gynecol. Obstet. Invest.*, 75:101–8, 2013.

[151] J. M. Rennie, C. F. Hagmann, and N. J. Robertson. Outcome after instrapartum hypoxic ischemia at term. *Semin. Fetal. Neonatal Med.*, 12(5):398–407, 2007.

[152] B. Robinson. A review of NICHD standardized nomenclature for cardiotocography: The importance of speaking a common language when describing electronic fetal monitoring. *Rev. Obstet. Gynecol.*, 1(2):56–60, 2008.

[153] M. Romano, P. Bifulco, M. Cesarelli, M. Sansone, and M. Bracale. Fetal heart rate power spectrum response to uterine contraction. *Medical Engineering and Computing*, 44(3):188–201, 2006.

[154] H. Sahin and A. Subasi. Classification of the cardiotocogram data for anticipation of fetal risks using machine leaning techniques. *Applied Soft Computing*, 33:231–8, 2015.

[155] S. L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–28, 1997.

[156] S. Schiermeier, A. Pildner von Steinburg, A. Thieme, J. Reinhard, M. Daumer, M. Scholz, W. Hatzmann, and K. T. M. Schneider. Sensitivity and specificity of intrapartum computerised FIGO criteria for cardiotocography and fetal scalp pH during labour: Multicentre, observational study. *Br. J. Obstet. Gynaecol.*, 115(12):1557–63, 2008.

[157] F. S. Schlindwein and D. H. Evans. Autoregressive spectral analysis as an alternative to fast Fourier transform analysis of Doppler ultrasound signals. In *Diagnostic vascular ultrasound*. Edward Arnold, London, 1992.

[158] M. I. Shevell. The "Bermuda triangle" of neonatal neurology: Cerebral palsy, neonatal encephalopathy, and intrapartum asphyxia. *Semin. Pediatr. Neurol.*, 11(1):24–30, 2004.

[159] M. G. Signorini, G. Maganes, S. Cerutti, and D. Arduini. Linear and nonlinear parameters for the analysis of fetal heart rate signals from cardiotocographic recordings. *IEEE Transactions on Biomedical Engineering*, 50(3):365–74, 2003.

[160] C. V. Smith, H. N. Nguyen, J. P. Phelan, and R. H. Paul. Intrapartum assessment of fetal well-being: A comparison of fetal acoustic stimulation with acid-base determinations. *Am. J. Obstet. Gynecol.*, 155(4):726–8, Oct 1986.

[161] Speculum S.A. Sisporto project webpage. www.sisporto.med.up.pt. Accessed 01/09/15.

[162] Speculum S.A. Omniview-sisporto webpage. www.omniview.eu, 2015. Accessed 01/09/15.

[163] J. A. D. Spencer, N. Badawi, P. Burton, J. Keogh, P. Pemberton, and F. Stanley. The intrapartum CTG prior to neonatal encephalopathy at term: A case-control study. *British Journal of Obstetrics and Gynaecology*, 104(1):25–8, 1997.

[164] J. Spilka, V. Chudáček, M. Koucky, L. Lhotska, M. Huptych, P. Janku, G. Georgoulas, and C. Stylios. Using nonlinear features for fetal heart rate classification. *Biomedical Signal Processing and Control*, 7(4):350–7, 2012.

[165] J. Spilka, G. Georgoulas, P. Karvelis, V. Chudáček, C. Stylios, and L. Lhotska. Discriminating normal from abnormal pregnancy cases using an automated FHR evaluation method. In A. Likas, K Blekas, and D. Kalles, editors, *Proceedings of the 8th Hellenic Conference on AI, SETN 2014*, volume 8445 of *Lecture Notes in Computer Science*, pages 521–31. Springer International, 2014.

[166] J. Spilka, G. Georgoulas, P. Karvelis, V. P. Oikonomou, V. Chudáček, C. Stylios, L. Lhotska, and P. Janku. Automatic evaluation of FHR recordings from CTU-UHB CTG database. In *Information Technology in Bio- and Medical Informatics: Lecture notes in Computer Science*, volume 8060, pages 47–61. Springer Publishing Company, 2013.

[167] G. Stamatoyannopoulos. Molecular and cellular basis of hemoglobin switching. In *Disorders of hemoglobin: genetics, pathophysiology, and clinical management*, pages 131–45. Cambridge University Press, 2001.

[168] P. J. Steer. Has electronic fetal monitoring made a difference? *Semin. Fetal. Neonatal Med.*, 13(1):2–7, 2008.

[169] P. C. Struijk, V. J. Mathews, T. Loupas, P. A. Stewart, E. B. Clark, E. A. Steegers, and J. W. Wladimiroff. Blood pressure estimation in the human fe-

tal descending aorta. *Ultrasound Obstet. Gynecol.*, 32(5):672–81, 2008. PMID: 18816497.

[170] C. Sundar, M. Chitradevi, and G. Geetharamani. Classification of cardiotocogram data using neural network based machine learning technique. *International Journal of Computer Applications*, 47(14), 2012.

[171] P. Sunshine. Perinatal asphyxia: An overview. In D. K. Stevenson and W. E. Benitz, editors, *Fetal and neonatal brain injury: mechanisms, management, and risks of practice*, chapter 1, pages 3–29. Cambridge University Press, 2003.

[172] N. Tasnim, G. Mahmud, and S. Akram. Predictive accuracy of intrapartum cardiotocography in terms of fetal acid base status at birth. *Journal of the College of Physicians and Surgeons Pakistan*, 19(10):632–5, 2009.

[173] S. B. Thacker, D. F. Stroup, and H. B. Peterson. Efficacy and safety of intrapartum electronic fetal monitoring: An update. *Obstet. Gynecol.*, 86(4):613–20, 1995.

[174] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer. Testing for nonlinearity in time series: The method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1):77–94, 1992.

[175] M. E. Tipping. Personal website of Mike Tipping. www.miketipping.com/sparsebayes.htm. Accessed 20/09/15.

[176] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–44, 2001.

[177] P. Tomas, J. Krohova, P. Dohnalek, and P. Gajdos. Classification of cardiotocography records by random forest. In *36th International Conference on Telecommunications and Signal Processing (TSP)*, pages 620–3, Rome, IT, 2013. IEEE.

[178] I. Tsoulos, G. Georgoulas, D. Gavrilis, C. Stylios, J. Bernandes, and P. Groumpos. Introducing grammatical evolution in fetal heart rate analysis and classification. In *3rd International IEEE Conference on Intellignet Systems*, pages 560–5, London, 2006.

[179] Dimitris G. Tzikas, Liyang Wei, Aristidis Likas, Yongyi Yang, and P. Galatsanos. A tutorial on relevance vector machines for regression and classification with applications. Online.

[180] University College London Clinical Trial Centre. INFANT Trial news - UCL Clinical trial centre. www.ucl.ac.uk/cctu/researchareas/womenshealth/infant. Accessed 01/09/15.

[181] M. van der Hout-van der Jagt, S. Oei, and P. Bovendeerd. Simulation of reflex late decelerations in labor with a mathematical model. *Early Hum. Dev.*, 89(1):7–19, 2013.

[182] M.B. van der Hout-van der Jagt, S.G. Oei, and P.H.M Bovendeerd. A mathematical model for simulation of early decelerations in the cardiotocogram during labor. *Med. Eng. Phys.*, 34:579–89, 2012.

[183] H. P. van Geijn. Developments in CTG analysis. *Baillieres Clin. Obstet. Gynaecol.*, 10(2):185–209, 1996.

[184] J. O. E. H. Van Laar, M. M. Porath, C. H. L. Peters, and S. G. Oei. Spectral analysis of fetal heart rate variability for fetal surveillance: Review of the literature. *Acta Obstetrica et Gynaecologica Scandinavica*, 87(3):300–6, 2008.

[185] E.G. Vázquez, A.Y. Escolano, P.G. Riaño, and J.P. Junquera. Repeated measures multiple comparison procedures applied to model selection in neural networks. In *Bio-Inspired Applications of Connectionism*, pages 88–95. Springer, 2001.

[186] A. C. Vidaeff and S Ramin. Fetal pulse oximetry: 8 vital questions. *OBG Management*, 16(3):28–44, 2004.

[187] P. Warrick, E. Hamilton, and M. Macieszczak. Neural network based detection of fetal heart rate patterns. In *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, volume 4, pages 2400–2405, July 2005.

[188] P. A. Warrick and E. F. Hamilton. Subspace detection of the impulse response function from intrapartum uterine pressure and fetal heart rate variability. In *Computing in Cardiology Conference (CinC)*, pages 85–8, Zaragoza, 2013. IEEE.

[189] P. A. Warrick, E. F. Hamilton, D. Precup, and R. E. Kearney. Detecting the temporal extent of the impulse response function from intra-partum cardiotocography for normal and hypoxic fetuses. In *Engineering in Medicine and Biology Society, 30th Annual International Conference of the IEEE*, volume 56, pages 1587–1597, 2008.

[190] P. A. Warrick, E. F. Hamilton, D. Precup, and R. E. Kearney. Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography. *IEEE Transactions on Biomedical Engineering*, 57(4):771–9, 2010.

[191] P. A. Warrick, R. E. Kearney, D. Precup, and E. F. Hamilton. Linear models of intrapartum uterine pressure-fetal heart rate interaction for the normal and hypoxic fetus. In *Engineering in Medicine and Biology Society,*, volume 1 of *28th Annual International Conference of the IEEE*, pages 6434–6437. EMBS '06, 2006.

[192] P. A. Warrick, R. E. Kearney, D. Precup, and E. F. Hamilton. System-identification noise suppression for intra-partum cardiotocography to discriminate normal and hypoxic fetuses. In *Computers in Cardiology*, pages 937–940. CinC, September 2006.

[193] P. A. Warrick, R. E. Kearney, D. Precup, and E. F. Hamilton. Low-order parametric system identification for intrapartum uterine pressure-fetal heart rate interaction. In *Engineering in Medicine and Biology Society, 29th Annual International Conference of the IEEE*, pages 5043–5046. EMBS, 2007.

[194] P. A. Warrick, R. E. Kearney, D. Precup, and E. F. Hamilton. Time progression of a parametric impulse response function estimate from intra-partum cardiotocography for normal and hypoxic fetuses. In *Computers in Cardiology*, 2007.

[195] P.A. Warrick, E.F. Hamilton, and R.E. Precup, D. Kearney. Identification of the dynamic relationship between intrapartum uterine pressure and fetal heart rate

for normal and hyposic fetuses. *IEEE Transactions on Biomedical Engineering*, 56(6):1587–97, 2009.

[196] S. Weiner. Independent validation of a fetal heart rate pattern recognition software. *Am. J. Obstet. Gynecol.*, 208(1):S316–7, 2013.

[197] M. E. M. H. Westerhuis, A. Kwee, A. A. van Ginkel, A. P. Drogtrop, W. J. A. Gyelaers, and G. H. Visser. Limitations of ST analysis in clinical practice: Three cases of intrapartum metabolic acidosis. *Br. J. Obstet. Gynaecol.*, 114(10):1194–1201, 2007.

[198] J. A. Westgate, L. Bennet, and A. J. Gunn. Fetal heart rate variability changes during brief repeated umbilical cord occlusion in near term fetal sheep. *Br. J. Obstet. Gynaecol.*, 106(7):664–71, 1999.

[199] J. A. Westgate, B. Wibbens, L. Bennet, G. Wassink, J. T. Parer, and A. J. Gunn. The intrapartum deceleration in center stage: a physiologic approach to the interpretation of fetal heart rate changes in labor. *Am. J. Obstet. Gynecol.*, 197(3):236.e1–11, 2007.

[200] Jennifer Westgate, Maureen Harris, John S.H. Curnow, and Keith R. Greene. Plymouth randomized trial of cardiotocogram only versus ST waveform plus cardiotocogram for intrapartum monitoring in 2400 cases. *American Journal of Obstetrics and Gynecology*, 169(5):1151–60, 1993.

[201] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[202] L. Xu. *Computerised analysis of fetal heart rate.* PhD thesis, University of Oxford, 2014.

[203] L. Xu, A. Georgieva, C. W. Redman, and S. J. Payne. Feature selection for computerized fetal heart rate analysis using genetic algorithms. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 445–8, 2013.

[204] L. Xu, C. W. G. Redman, S. J. Payne, and A. Georgieva. Feature selection using genetic algorithms for fetal heart rate analysis. *Physiol. Meas.*, 35(7):1357–71, 2014.

[205] P. Yeh, K. Emary, and L. Impey. The relationship between umbilical cord arterial ph and serious adverse neonatal outcome: Analysis of 51519 consecutive validated samples. *Br. J. Obstet. Gynaecol.*, 119(7):824–31, 2012.

[206] E. Yilmaz and Kilikçier. Determination of fetal state from cardiotocogram using LS-SVM with particle swarm optmization and binary decision tree. *Computational and Mathematical Methods in Medicine*, 2013.