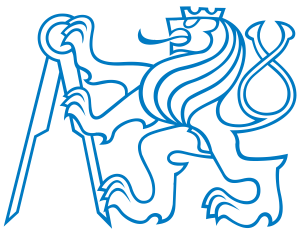
Complex approach to fetal heart rate analysis: A hierarchical classification model



Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics

Jiří Spilka

Doctoral thesis

Supervisor: Doc. Ing. Lenka Lhotská, CSc.

Study Programme No. P2612-Electrotechnics and Informatics
Branch No. 3902V035-Artificial Intelligence and Biocybernetics

Prague, September 2013
revised: December 2013

Abstract

The electronic fetal monitoring (EFM) is used for fetal behaviour surveillance via measurement of fetal heart rate. In the course of labour, fetus can suffer by severe hypoxic insults that might lead to possible adverse long term consequences. The main goal of EFM is to provide indirect information about fetal well-being and help obstetricians to indicate timely intervention to prevent adverse consequences.

Nowadays, the EFM is an integral part of every day obstetrics practice. The EFM most commonly refers to cardiotocography (CTG) that is a measurement of fetal heart rate and uterine contractions. The reliability, validity, and efficiency of CTG have neither been confirmed nor disproved. Also all the computerized systems have yet to prove its efficiency. The reasons for stagnating or even non-existent technical progress are: *i)* use of very small and ad-hoc created databases, *ii)* neglect of the high intra/inter observer variability of clinical evaluation of CTG, *iii)* unclear definition of pathological labour outcome (usually the outcome is imprecisely defined by a pH value), *iv)* strict technical approach disconnected from the clinical reality.

Therefore, in this thesis, we introduce the first open access database of intrapartum CTG. The database that enables other researchers to developed and test new algorithms for CTG analysis and classification. We show that it is possible to overcome the high inter observer variability using a model of clinical annotation. We also show that there is a group of pathological CTG records on which the clinicians have good agreement. Finally, we develop a novel approach for CTG evaluation using a hierarchical model. The model considers different outcome measures as a mixture of individual components. Thus is able to overcome discrepancies between biochemical markers (pH, base excess, base deficit), Apgar score, and clinical evaluation of CTG. The developed model is able to answer the difficult question, whom to trust when you are given multiple noisy and imprecise information about labour outcome.

Keywords Cardiotocography, Fetal heart rate, Time series analysis, Feature selection, Mixture models, Latent class analysis, Latent class regression, Classification

Acknowledgements

This work would not have been possible without my supervisor Lenka Lhotská, who always supported me to pursue my own ideas and directions and found a solution for all my problems and difficulties during the years, and my colleague Václav Chudáček, who introduced me into the field of CTG processing and analysis and with whom I had the pleasure to collaborate closely and pursue the goal of automatic CTG processing and analysis. I appreciate his enthusiasm and value his friendship.

My special thanks go to my co-authors: Michal Huptych for discussions on experiments and statistics, Miroslav Burša for his extraordinary programming skills and maintenance of the obstetrics medical database, Lukáš Zach for development of the CTGAnnotator software, and Chrysostomos Stylios, George Geourgoulas, and Petros Karvelis for the collaboration on classification experiments.

I gratefully acknowledge my colleagues: Jakub Kužílek for discussions on various topics on signal processing, Martin Macaš for discussions on machine learning, Karla Štěpánová for valuable comments to the thesis, Michal Vavrečka for being there and making the work place cheerful, and Martin Hanuliak from whom I had the pleasure to learn a lot in many fields, mainly in ECG signal processing.

I am also very grateful for co-operation with medical experts: Petr Koucký for providing the initial impulse for the CTG processing, Petr Janků and Lukáš Hruban for the stimulating discussions, patience, and positive attitude towards the automatic analysis of fetal heart rate.

I would like to thank to all medical experts that participated on the annotation of CTG signals: P. Janků, L. Hruban, A. Hudec, V. Korečko , M. Kacerovský, M. Koucký, M. Procházka, J. Segeťa, and O. Šimetka.

At last but not at least I would like to express my gratitude to my family, without them this work would not have been possible, literally. I would like to thank to my lovely wife, Ajka, for her understanding, patience, never ending optimism, and for keeping me happy, to my parents who always supported me, and to Jana and Jamie for their suggestions and comments on English writing.

Research work in this thesis was partly supported by the research projects: ČVUT Grant SGS10/279/OHK3/3T/13, and by the research programs No. NT11124-6/2010 Cardiotocography evaluation by means of artificial intelligence of the Ministry of Health Care.

Contents

Abbreviations and symbols

Abbreviations

ACC	Accuracy
ACOG	American College of Obstetricians and Gynaecologists
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
AUC	Area Under receiver operating Characteristic
BDecf	Base Deficit
BE	Base Excess
BIC	Bayes Information Criterion
BPM	Beats Per Minute
CS	Caesarean Section
CTG	Cardiotocogram
CTU-UHB	Czech Technical University – University Hospital Brno
CV	Cross-Validation
DFA	Detrend Fluctuation Analysis
ECG	Electrocardiogram
EFM	Electronic Fetal Monitoring
EM	Expectation Maximization
DECG	Fetal Electrocardiogram
FHR	Fetal Heart Rate
FHRV	Fetal Heart Rate Variability
FIGO	International Federation of Gynaecology and Obstetrics
FN	False Negative
FP	False Positive
FS	Feature Selection

HF	High Frequency
LCA	Latent Class Analysis
LCM	Latent Class Model
LCR	Latent Class Regression
LF	Low Frequency
LTV	Long Term Variability
MF	Middle Frequency
MV	Majority Voting
NICU	Neonatal Intensive Care Unit
NICHD	National Institute of Child Health and Human Development
PA	Proportion of Agreement
PCA	Principal Component Analysis
PPV	Positive Predictive Value (= PR)
PR	Precision (= PPV)
RCOG	Royal College of Obstetricians and Gynaecologists
SE	Sensitivity
SMOTE	Synthetic Minority Over-sampling Technique
SP	Specificity
STAN	ST ANalysis
STV	Short Time Variability
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UC	Uterine Contractions
VAG	Vaginal Delivery
VLF	Very Low Frequency

Symbols

π	a mixing parameter of a mixture model
α_{ck}^{j}	probabilities for j-th annotator, a latent class c, and a predicted class k
$\mathbf{B}$	a matrix of basis vectors

C	number of classes
$\hat{\rho}$	sample correlation coefficient
$\mathcal{D}$	an input dataset
d	number of features
$\delta(a, b)$	an indicator function that equals 1 when $a = b$ and 0 otherwise
f	a classifier mapping features to class labels
J	number of annotators
κ	the kappa coefficient
l	number of selected features or principal components
λ	an eigen value
m	embedding dimension or number of components of a mixture model
$\mathbf{A}$	$C \times C$ confusion matrix with entries $[\mathbf{A}]_{ck} = \alpha_{ck}$
μ_i	an estimate of a latent class for i-th observation
μ_{ic}	an estimate of a latent class for i-th observation and class c
N	number of examples/observations
$\mathbf{p}$	vector of prevalences
p_o	overall proportion of agreement
p_s	proportion of agreement with respect to different classes
q	number of folds of cross validation
$\mathcal{S}_{acc}$	accuracy based score
$\mathcal{S}_{sp}$	spammer score
T	time series of RR intervals
$\boldsymbol{\theta}$	vector of model parameters
$\mathbf{w}$	coefficients of logistic regression
$\mathbf{x}_i$	d-dimensional feature vector
y_i	a class label for i-th instance
y_i^j	evaluation of j-th annotator for i-th observation
z	fetal heart rate signal

Chapter 1

Introduction

"A journey of a thousand miles began with a single step."
saying of Lao Tzu tr. L. Giles 51, 1904

Being born is one of the most crucial events in our life. After intrauterine growth and development a baby is going to establish itself as an independent individual. During labour, a fetus can repeatedly suffer from oxygen insufficiency, which is normal but for fetuses with weakened defence mechanism a metabolic acidosis could be developed. The metabolic acidosis can lead to neuro-development disability, cerebral palsy, neonatal encephalopathy, or death resulting from excessively long oxygen insufficiency. To handle the labour stress a fetus is equipped with a defence mechanism. Good understanding of how an individual fetus reacts to the stress of labour helps to indicate timely intervention when the fetal defence has been activated but before the risk of long-term consequences increases.

The fetal heart rate (FHR) reflects changes in fetal behaviour. In the past a fetal stethoscope was used to intermittently monitor FHR and its changes. However, the stethoscope could not detect subtle changes in FHR and continuous monitoring was also impracticable. Introduction of electronic fetal monitoring (EFM) overcame these disadvantages and offered continuous fetal surveillance during pregnancy and, more importantly, during delivery. The EFM most commonly refers to cardiotocography (CTG) that is a measurement of FHR and uterine contractions (UC). Since its introduction the CTG has served as the main information channel providing obstetricians with insight into fetal well-being.

The introduction of CTG in late 1960's was accompanied by great expectations. Initially, the CTG was intended for high risk pregnancies but it has become commonly used even for normal pregnancies. However promising the technology at the time, it has been surrounded by great controversies from the very beginnings. The rationale of CTG is that it should prevent adverse labour outcomes by enabling clinicians to timely intervene in labour. For this rationale to be true three conditions must hold (Haggerty, 1999; Paneth et al., 1993). The CTG is *i) reliable*: substantial inter-observer agreement exist as to the identity and meaning of CTG patterns, *ii) valid*: one or more CTG patterns are statistically significant to an adverse labour outcome, and *iii) efficient*: an intervention based on a CTG pattern could prevent an adverse labour outcome. The great research effort was devoted to the reliability of CTG by introduction of various guidelines (ACOG, 2009; FIGO, 1986; Macones et al., 2008) to the validity by examining different patterns in connection to labour outcomes (Hamilton et al., 2012; Parer et al., 2006; Westgate et al., 2007) and to the efficiency (Alfirevic et al., 2006).

The attempts of computerized CTG were aimed on improving the reliability and efficiency. Beginning with work of (Dawes et al., 1981) the automatic analysis of CTG was aligned with clinical guidelines, which has become fundamental for almost every work on automatic CTG analysis. In addition to the morphological features used in the guidelines, new features were introduced for FHR analysis in order to reveal a possible new information hidden to the clinical guidelines. These were mostly based on the research in the adult heart rate variability (Task-Force, 1996) and consisted mainly of frequency, joint time-frequency, and nonlinear features. The morphological features were included into automatic systems for CTG analysis, among the best known are Omniview SisPorto® (de Campos

et al., 2008) developed at University of Porto, INFANT® (Greene and Keith, 2002; Keith and Greene, 1994) developed by K2 Medical Systems™, UK, and the PeriCALM™ (Elliott et al., 2010; Parer and Hamilton, 2010), developed by LMS Medical systems, Canada and PeriGen, USA. However, all the systems are mainly used at the institution or region they were developed and, more importantly, none of the system has proven its efficiency in a clinical trial.

The possible reasons that hindered the development of automatic CTG analysis are: *i)* unclear relationship between FHR patterns and labour outcome as measured from fetal blood by pH or base deficit (BDecf) after delivery (Parer et al., 2006; Westgate et al., 2007), *ii)* high variability in CTG interpretation (Blackwell et al., 2011; Vayssiere et al., 2009), *iii)* use of small and proprietary CTG databases in many studies, and/or *iv)* disconnection between strictly technical papers and clinical practice.

The only measurable improvement of the EFM was introduction of the ST-analysis method (Rosén and Luzietti, 1994) (Neoventa Medical, Sweden), which is based on analysis of fetal electrocardiogram. The ST-analysis improved the labour outcomes slightly (Amer-Wåhlin and Maršál, 2011; Norén et al., 2003) but its use is not always possible or feasible since it requires invasive measurement. Moreover, the ST-analysis is not an alternative to the CTG but rather a support. In order to use the ST-analysis correct interpretation of CTG is still required.

Nowadays, the CTG remains the most prevalent method for intrapartum fetal surveillance (Bernardes et al., 1997; Chen et al., 2011) often supported by the ST-analysis. Since the rationale of CTG was neither confirmed nor disproved, Sartwelle (2012) proposed to abandon the CTG monitoring completely. Nevertheless, with the widespread use of technology in all areas of clinical practice, it is unlikely that the CTG will be abandoned and a solution to improve the CTG is still desired. Steer (2008) concluded that the weakness of CTG lies in a generally poor standard of interpretation and the contribution of the human factor, demonstrated by high intra- and inter-observer variability. Either more education and training on CTG interpretation should be performed (Doria et al., 2007; Westerhuis et al., 2007a) or one should use a more cost-effective solution by developing a decision support system serving as a source of additional information (Bernardes and Ayres-De-Campos, 2010; Hasley, 2011; Steer, 2008). A development of methods that could enable to create such a decision system are the main aim of this work.

1.1 Motivation and goals of thesis

The motivation of this work is to overcome the possible reasons that hindered a development of automatic CTG interpretation as described above. The main goals are to propose a methodology of CTG evaluation, to design a new classification paradigm, and to develop a novel classification system that would support the clinicians with assessment of CTG. The methodology will be based on the design and implementation of a model for evaluation of CTG. The main goals of the thesis can be summarized as follows:

1. **To perform a critical analysis of used databases and algorithms.** The comprehensive overview of databases that were used for CTG processing and/or classification. The critical analysis would also give overview of different approaches (algorithms) with respect to their classification performance.

2. **To create and describe a new open access database of CTG records.** The CTG database that would be used for design and verification of a model for classification and that would be open to other researches.

3. **To design a model for clinical evaluation of CTG.** The model that is able to account for the high inter-observer variability in clinical decision and that would estimate the hidden (unknown) truth of CTG evaluation from multiple clinical annotations. Until now there has not been such a model developed.

4. **To design, implement, and verify a classification of FHR features using pH.** The FHR features would reflect the complex behaviour of fetus and would be suitable to discriminate normal and abnormal fetuses. Further, to produce a classifier of FHR features where a pH value is used to discriminate between normal and abnormal labour outcome.

5. **To design and develop a classification system** that is able to account for uncertainty with labour outcome definition. A system that would consider the discrepancy between objective evaluation using biochemical markers and subjective evaluation using Apgar score and clinical assessment of CTG. The results of the system would be a classifier of FHR features and estimated labour outcome from multiple, possibly noisy and imprecise sources. The system would provide accurate information about fetal well-being.

1.2 Structure of the thesis

In Chapter 2 we introduce the fetal physiology and CTG from an obstetrician's perspective and present the surveillance methods used for fetal monitoring. We thoroughly describe the assessment of labour and neonate outcome. In Chapter 3 we present the state of the art of automatic evaluation of fetal heart rate and in Chapter 4 we introduce the new open access CTG database. We describe the FHR preprocessing and analysis using a comprehensive set of features originating from different domains in Chapter 5. We provide a through analysis of clinical evaluation in Chapter 6 and we show that the inter observer variability in clinical decision could be lowered using a latent class model. We perform a classification of FHR features using the pH value as a discriminator between normal and abnormal CTG records in Chapter 7 and in Chapter 8 we present a novel hierarchical model for FHR evaluation. We prove that the model is able to account for discrepancy in biochemical markers and clinical evaluation and provide the best classification results.

1.3 List of publications

This thesis is based on the following papers:

Journal papers

J. Spilka, V. Chudáček, M. Koucký, L. Lhotská, M. Huptych, P. Janků, G. Georgoulas, and C. Stylios. Using nonlinear features for fetal heart rate classification. *Biomedical Signal Processing and Control*, 7(4):350–357, 2012.

J. Spilka, V. Chudáček, P. Janků, L. Hruban, M. Burša, M. Huptych, L. Zach, A. Hudec, M. Kacerovský, M. Koucký, L. Lhotska, M. Procházka, V. Korečko, J. Segeťa, and O. Šimetka. First step to automated obstetrics alarm system: Analysis of annotations derived from expert-obstetricians. *Methods of Information in Medicine*, Manuscript submitted for publication, 2013a.

V. Chudáček, J. Spilka, P. Janků, M. Koucký, L. Lhotská, and M. Huptych. Automatic evaluation of intrapartum fetal heart rate recordings: A comprehensive analysis of useful features. *Physiological Measurement*, 32:1347–1360, 2011.

V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych, and L. Lhotská. Open access intrapartum CTG database. *BMC Pregnancy and Childbirth*, Manuscript submitted for publication, 2013.

Conference papers

J. Spilka, V. Chudáček, M. Koucký, and L. Lhotská. Assessment of Non-Linear Features for Intrapartal Fetal Heart Rate Classification. In *Proceedings of 9th International Conference on Information Technology and Applications in Biomedicine*, 2009.

J. Spilka, G. Georgoulas, P. Karvelis, V. P. Oikonomou, V. Chudáček, C. Stylios, L. Lhotská, and P. Janků. Automatic evaluation of FHR recordings from CTU-UHB CTG database. In M. Burša, S. Khuri, and M. Renda, editors, *Information Technology in Bio- and Medical Informatics*, Lecture Notes in Computer Science, pages 57–66. Springer Berlin Heidelberg, 2013b.

V. Chudáček, J. Spilka, M. Huptych, G. Georgoulas, L. Lhotská, M. Stylios, C. Koucký, and P. Janků. Linear and Non-Linear Features for Intrapartum Cardiotocography Evaluation. In *Computers in Cardiology*, volume 35, 2010.

V. P. Oikonomou, J. Spilka, C. Stylios, and L. Lhotská. An adaptive method for the recovery of missing samples from FHR time series. In *26th International Symposium on Computer-Based Medical Systems (CBMS)*, 2013.

L. Zach, V. Chudáček, M. Huptych, J. Spilka, M. Burša, and L. Lhotská. CTG Annotator–Novel Tool for Better Insight into Expert-obstetrician Decision Making Processes. In *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*, pages 1280–1282. Springer, 2013.

Chapter 2

Obstetrics preliminaries

This chapter contents is largely based on (Spilka, 2009, 2011) with modifications of state of the art knowledge. The summarized medical information is based on general medical textbooks (Guyton and Hall, 2005; Čech et al., 2006) and other general texts such as (Sundström et al., 2000).

Labour is a very stressful period for fetus as well as for mother. Fetus is affected by mother's behaviour and condition. The way fetus reacts to its changing environment gives an important information about its status. For instance, a change in fetal heart rate can be caused by nervous system that is activated by receptors reacting to the change of internal environment.

One of the major fetus's tasks is to handle reoccurring hypoxic events that could lead to severe consequences for further child development. Fetus has its own physiological protective mechanism able to sustain repetitive hypoxic episodes. However, if the fetus is not able to adequately response or to recover from hypoxic stage, the hypoxia could be developed into the next stage of oxygen deficiency called asphyxia that could lead to cerebral palsy, neonatal encephalopathy, or to death. Hypoxia, with prevalence lying in the region of 0.6% (Heintz et al., 2008) to 3.5% (Strachan et al., 2000), is considered still to be the third most common cause of newborn death (d'Aloja et al., 2009).

Several studies elaborated more on the cause of neonatal encephalopathy and cerebral palsy. Pierrat et al. (2005) examined 90 neonates with moderate or severe newborn encephalopathy with prevalence 1.64/1000. Birth asphyxia prevalence was 0.86/1000 per term live birth. The main cause of newborn encephalopathy was birth asphyxia, diagnosed in 52% cases. From these cases, asphyxia was caused intrapartum in 56% of cases, antepartum in 13%, ante-intrapartum in 10%, and post-partum in 2%. In 19% of cases, no underlying cause was identified during the neonatal course. Locatelli et al. (2010) investigated risk factors (described below) related to neonatal encephalopathy (prevalence 0.88/1000) and compared these factors with a control group. In neonatal encephalopathy group the risk factors were present antepartum in 74%, and intrapartum in 68% while in the control group the occurrence was lower, 18% antepartum and 19% intrapartum.

It is apparent that adverse delivery outcomes are not necessarily connected to intrapartum events but can be linked to antepartum period. In the both periods there are risk factors which occurrence significantly contribute to neonatal encephalopathy (Locatelli et al., 2010) and cerebral palsy (Evans et al., 2001). The antepartum risk factors are: obesity, diabetes, thyroid dysfunction, previous caesarean delivery, pre-eclampsia, fetal growth restriction, abnormal amniotic fluid volume, and abnormal FHR tracing before labour. The intrapartum risk factors are: bleeding during labour, epidural analgesia, intrauterine infection, meconium-stained liquor, post-term delivery, induced labour, and caesarean section. In presence of some risk factors, electronic fetal monitoring is necessary for fetal surveillance. On the other hand, for low risk pregnancies the use of electronic fetal monitoring does not offer significant contribution to fetal outcomes (Alfirevic et al., 2006).

This chapter is organized as follows: first, we outline the basics of fetal physiology and fetus response to different stages of oxygen deficiency. Next, we describe an interaction between mother and fetus during gestation with emphasis on the antepartum and intrapartum period. Finally, we introduce methods for the fetal hypoxia diagnostics with focus on electronic fetal monitoring.

2.1 Fetal physiology

Fetal development lasts about 40 weeks. Complex systems, such as circulatory, respiratory, nervous, gastrointestinal, etc. are being developed during that time. In this work we discuss in detail only the circulatory system; the others are mentioned only to give overall insight into fetal behaviour.

Fetal heart begins beating approximately at 4^{th} week of pregnancy with frequency about 65 beats per minute (BPM). This frequency increases during a gestation up to 140 bpm before delivery. The main function of fetal heart is to pump oxygenated blood from placenta to the organs and, in turn, to carry carbon dioxide back to placenta, where an exchange between mother and fetus is maintained. The exchange is not limited to gases only but is performed for all substances such as nutrition and fetus' waste products.

Fetal circulation The oxygenated blood from mother's aorta is distributed to the uterine arteries and further to the spiral arteries that deliver blood to placenta. Here, in the thin capillaries membranes, the exchange of gases and substrates is performed. The fetus respiration system is non-functional and placenta works as the fetal lungs. Therefore, blood flows bypass lungs by ductus arteriosus. The same situation applies for liver, only with the difference, that liver are partially functioning and blood is not completely bypassed by ductus venous. The whole organization of the fetal circulation is illustrated in Figure 2.1. The oxygenated blood from placenta enters the right atrium and continues directly to the left atrium throughout foramen ovale. From there it is pumped into left ventricle and then to aorta and further back to placenta via umbilical arteries. The de oxygenated blood returning from the upper part of the body enters the right atrium and is pumped into the right ventricle. Then, after ventricle contraction, blood is pumped through ductus arterious into the descending aorta.

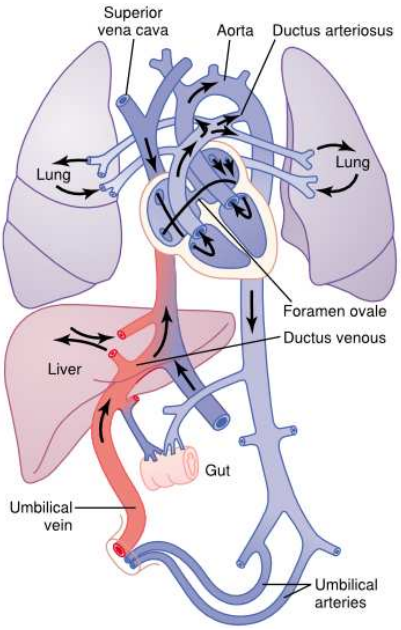
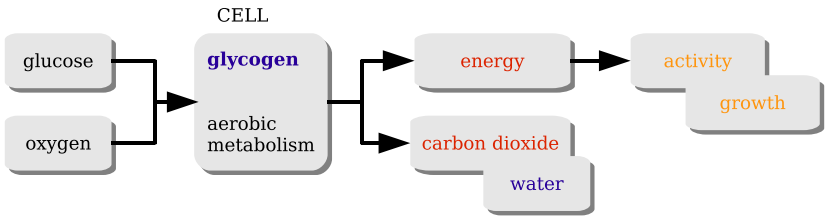


Figure 2.1: Organization of the fetal circulation. The difference between fetal and neonatal circulation lies in so called "blood short-cuts". It involves ductus venous, ductus arterious and foramen ovale. If these are not closed at the first breath, there is a serious risk for a new born development (Guyton and Hall, 2005).

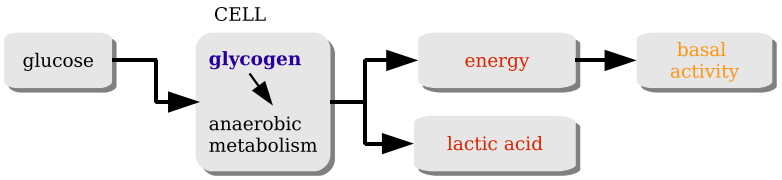
2.1.1 Energy metabolism

Placenta maintains an exchange of oxygen and carbon dioxide between mother and fetus. This exchange can only be performed due to different partial pressures of gases. In placenta, oxygen

is bound to haemoglobin and released in the capillaries in the fetal circulatory system. There the carbon dioxide replaces the oxygen and is carried back to placenta. Depending on oxygen availability we distinguish aerobic and anaerobic metabolism. These are illustrated in Figure 2.2. The aerobic metabolism utilises glycogen (or fatty acids), oxygen, adenosine diphosphate, and phosphate (P) in order to create adenosine triphosphate which serves as energy source. The waste product are carbon dioxide and water. The anaerobic metabolism is also used for glycolysis but with the difference that the oxygen is not available and cannot be used.



(a) aerobic metabolism



(b) anaerobic metabolism

Figure 2.2: Energy metabolism. The aerobic metabolism is oxygen dependent. In cases of oxygen insufficiency, the so called anaerobic metabolism produces enough energy to cover basal activity (modified from (Sundström et al., 2000)).

The waste product of anaerobic metabolism is lactic acid. The anaerobic metabolism only provides energy for basal (vital) activity and, as a consequence, fetus growth is restricted. Therefore, the anaerobic metabolism should not last for hours. If the supply of oxygen is not re-established, hypoxanemia, hypoxia, and sequentially asphyxia are developed. These terms express different stages of decreased oxygen saturation of the fetal artery blood. Asphyxia is the last and worst stage that might occur. Before describing the individual stages, it is necessary to explain autonomic nervous system and its reaction to oxygen deficiency. This system adapts fetal heart rate to changing environment and regulates blood distribution. It consists of humoral and neural (parasympathetic and sympathetic) systems that function antagonistically. Parasympathetic system reacts rapidly on abrupt changes, whereas the sympathoadrenal system works at more fundamental level prevailing during stage of fetal hypoxia (Amer-Wåhlin, 2003). Parasympathetic activation causes reduction in fetal heart rate called bradycardia, while sympathetic activation leads to surge of stress hormones from the adrenals and FHR may increase up to tachycardia. It is worth to mention that transition between sympathetic and parasympathetic system is not linear, i.e. changing constantly in time, but rather shows non-linear behaviour (Goldberger et al., 2002). In Figure 2.3 is illustrated how the nervous systems reflect a change in blood gases.

Hypoxanemia Hypoxanemia is an initial stage of oxygen deficiency. The oxygen is depleted in the arterial blood at the periphery. Central organs and peripheral tissues are intact and enough oxygen is provided to maintain aerobic metabolism. The fetal response is activated by chemoreceptors located in major vessels. It involves several safety precautions. First, the more efficient uptake of oxygen is performed by increased blood flow or increased number of erythrocytes. Second, the fetal movements are reduced and also growth is restricted in order to save the oxygen. The fetus can sustain hypoxanemia for days and weeks. However, in presence of fetal hypoxanemia before labour, fetus has less ability to handle labour stress because of restriction of energetic reserves.

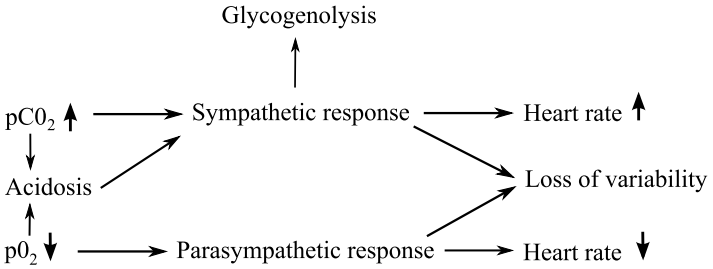


Figure 2.3: Relationship between blood gases and heart function (Amer-Wåhlin, 2003).

Hypoxia Hypoxia represents second stage of oxygen deficiency when the peripheral tissues are affected. Blood flow is redistributed in favour of central organs guaranteeing aerobic metabolism. On the contrary, anaerobic metabolism is utilised at peripheral tissues. The prime reaction to hypoxia is surge of stress hormones (adrenalin, noradrenalin) and sympathetic activation. Without any damage to fetus, hypoxia can last several hours.

Asphyxia Asphyxia is the most critical stage. The oxygen is depleted and high priority organs utilise anaerobic metabolism. The energy is created from glucose stored in liver and myocardium. Brain has very low glucose level, therefore glucose is supplied by liver. The fetal response to asphyxia involves release of stress hormones and activation of sympathetic nervous system. The fetus attempts to maintain function of central organs as long as possible. The final stage of asphyxia is the collapse of system with brain and heart failure. Asphyxia that lasts only few minutes might cause irrecoverable damage.

2.2 Fetal surveillance methods

The reliable assessment and diagnosis of changes in fetus condition is of major importance. The fetus hypoxia activate defence mechanism and anaerobic metabolism is utilised at the peripheral tissues. Using diagnostic tools these can be detected and evaluated. The diagnosis can be roughly split into two groups: fetal blood measurement (fetal blood sampling, pulse oximetry) and electronic fetal monitoring (cardiotocogram, fetal electrocardiogram).

2.2.1 Cardiotocogram

The fetal heart rate reflects changes in fetal behaviour and condition. Cardiotocogram (CTG) involves monitoring of fetal heart rate and uterine pressure. It offers valuable insight into fetal condition and serves intrapartum as well as peripartum (the admission CTG) when it might diagnose potential fetal compromise. The electronic fetal monitoring was introduced in 1960s and is a successor of auscultation method where the FHR was monitored periodically by stethoscope.

Cardiotocogram recording

We distinguish two types of CTG monitoring based on different stages of labour. Before rupture of membranes the external ultrasound probe and transducer are used to acquire FHR and uterine pressures, respectively. After the rupture of membranes an electrode could be attached at fetus scalp and FHR is computed directly from ECG's R-R intervals. The uterine pressures are obtained using internal electrode placed in vagina. The record is called intrauterine pressure. The external and internal monitoring is shown in Figure 2.4.

External monitoring has certain limitations in comparison to internal. In external monitoring the ultrasound Doppler principle is utilized to detect fetal heart pulsations. Therefore the ultrasound

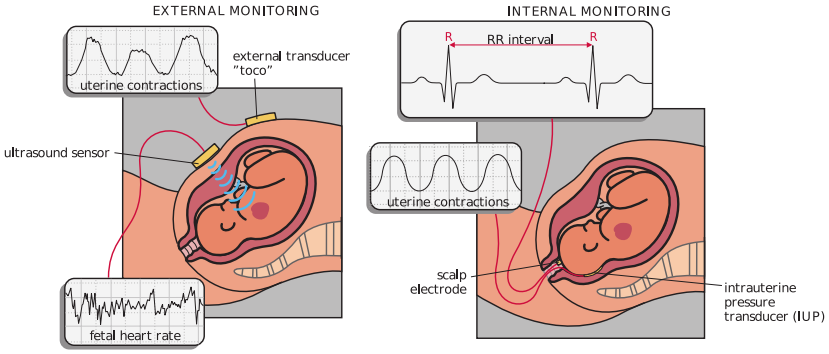


Figure 2.4: Recording of fetal heart rate and uterine activity (Sundström et al., 2000).

probe must be located precisely at the position of fetal heart and any movement either mother's or fetus' may cause distortions. The great advantage of external monitoring lies in easy application and non-invasibility. Internal monitoring, the direct electrocardiogram measurement (DECG), is invasive and can be used only when the fetal position is normal, i.e. head first presentation, and after fetal membranes' rupture. The electrode is screwed to fetal scalp without any damage to fetus and complete electrocardiogram is acquired. Then the fetal heart rate is computed as difference of successive beats. The pressure transducer is placed in vagina and intrauterine pressure is recorded. The internal monitoring has a higher signal to noise ratio than the external one and, in addition, DECG and its morphological changes can be examined.

Changes in fetal heart rate The changes of fetal heart rate may either occur during oxygen insufficiency or could be caused by aspects, such as mother behaviour or external influences. The FHR changes and its causes are as follows:

- Normal changes – the FHR is different during quiet and active sleep (REM[1]). There are rapid shifts in autonomic nervous system resulting in accelerations and increased heart variability during active sleep.
- Changes in placental blood flow – mainly due to cord compression. When the cord is compressed, the blood is pushed into fetus. The heart must pump more blood and the heart rate increases. The increase in blood volume results to increase in blood pressure. Hence, sensitive baro-receptors are activated and cause decrease in fetal heart rate. When the compressed cord is released, the FHR returns to normal.
- Adaptation to oxygen insufficiency – when oxygen content decrease, chemo-receptors are activated and stimulate sympathetic and parasympathetic nervous system. The changes in fetal heart rate depends on the stages of hypoxia. In case of acute hypoxanemia, immediate fall in FHR occurs while gradually developing hypoxia causes increase in FHR.
- External stimuli – due to the contraction there is an increase of head pressure that may cause deceleration. Also pressure on eye bulb might induce bradycardia.
- Increase in mother's temperature – in case of mother fever, the fetal metabolism increase which leads to higher oxygen consumption and may result in fetal tachycardia.
- The effect of drugs – the fetus could be affected by various drugs and the ability to handle labour stress may decrease, e.g. mother over-stimulation with xytocin results in increased uterine activity and fetus is affected by more intensive contraction.
- Fetal activity – the state of fetus (active and quiet) affects the frequency spectrum of FHR.

[1] rapid eyes movement

Assessment of fetal heart rate changes

The following patterns and features are usually assessed in CTG records: baseline rate, variability, acceleration, and deceleration. These patterns and their properties are strictly defined in guidelines for fetal monitoring (FIGO, 1986; NICE, 2007) and according to their occurrence the appropriate reaction is suggested. The normal CTG record is presented in Figure 2.5. It shows accelerations and normal heart variability that are markers of fetal well-being.

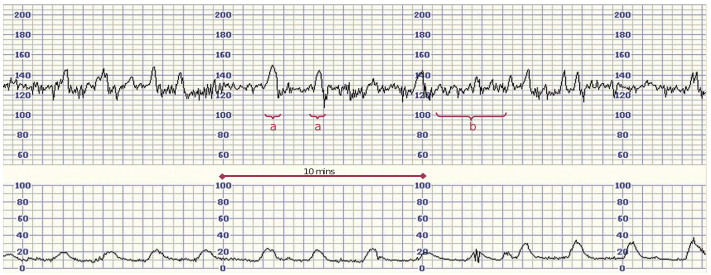


Figure 2.5: Normal reactive trace. (a) Accelerations; (b) normal variability (Hinshaw and Ullal, 2007).

Baseline heart rate Baseline fetal heart rate is determined over time period of 5 or 10 minutes when acceleration and deceleration are absent. Normal baseline rate is in range of 110 – 160 bpm. The decrease of heart rate below 110 bpm is called bradycardia and the increase of heart rate up to 150 bpm is called tachycardia.

Variability FHR variability is defined as amplitude oscillations around baseline heart rate. Normal values are between 5–25 bpm. Example of normal variability is shown in Figure 2.5. The so called saltatory pattern is an increase in variability of more than 25 bpm. Complete loss of variability for more than 40 minutes is the most abnormal sign and fetus may no longer fine-tune its circulation. The FHR could also have sinusoidal pattern with smooth, undulating sine-wave. In case of sinusoidal pattern, immediate intervention is required.

Accelerations Acceleration is a transient increase in the heart rate of more than 15 bpm lasting 15 seconds or more. This is associated with fetal movements or stimulation, and indicates fetal well-being, see Figure 2.5.

Decelerations Deceleration is characterized as a transient decrease of FHR below the baseline level of more than 15 bpm lasting at least 10 seconds. The decelerations are linked to uterine activity and distinguished as uniform or variable. Uniform deceleration has the same pattern and shape from one deceleration to another, whereas the variable decelerations might vary from one contraction to another; for illustration see Figure 2.6. Uniform decelerations can be further divided into early and late depending on time of occurrence. Early deceleration represents transient decrease in FHR when the drop in FHR matches the onset of contraction. On the contrary, late decelerations are characterized as those with different onset of the contraction and deceleration. Note that only late decelerations are connected with hypoxia. The variable decelerations have different shape from one deceleration to another. As for uniform deceleration, the variable can be also split into two groups: uncomplicated and complicated. Uncomplicated deceleration is defined as deceleration lasting less than 60 seconds; below this time fetus is able to sustain it.

CTG guidelines

In order to standardize CTG interpretation and classification the guidelines were introduced by the International Federation of Gynaecology and Obstetrics (FIGO, 1986), see Table 2.1. They were

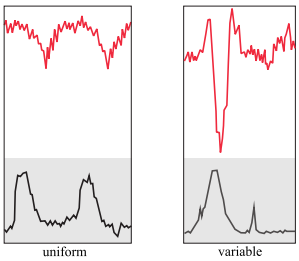


Figure 2.6: Uniform (rounded pattern, shape is similar) and variable (rapid loss of beats, pattern may vary) decelerations (Sundström et al., 2000).

further adapted by the national obstetrics organizations and resulted into modified guidelines (NICE, 2007; RCOG, 2001). In 2008 the guidelines were reviewed by the diverse group of investigators from the three organizations (National Institute of Child Health and Human Development, American College of Obstetricians and Gynaecologists, and Society for Maternal-Fetal Medicine) and resulted into the guidelines described in (Macones et al., 2008) and (ACOG, 2009). Despite the efforts made and variety of guidelines introduced, the interpretation of CTG still remains subjective with high inter and intra observer variability documented back in 1982 (Beaulieu et al., 1982; Lotgering et al., 1982) as well as in the recent studies (Blackwell et al., 2011; Vayssiere et al., 2009). Note that the FIGO guidelines still remain the only international consensus on interpretation of CTG. (de Campos and Bernardes, 2010) showed comparison between various guidelines and concluded that for the normal patterns the guidelines are consistent but for the suspicious and pathological patterns they are in wide disagreements. They also stated that guidelines are difficult to interpret and proposed a simplified view on guidelines.

Table 2.1: FIGO guidelines. Adapted from (de Campos and Bernardes, 2010).

NORMAL PATTERN	SUSPICIOUS PATTERN	PATHOLOGICAL PATTERN
– Baseline heart rate between 110 and 150 bpm – Amplitude of heart rate variability between 5 and 25 bpm	– Baseline heart rate between 150 and 170 bpm or between 100 and 110 bpm – Amplitude of variability between 5 and 10 bpm for more than 40 minutes – Increased variability above 25 bpm – Variable decelerations	– Baseline heart rate below 100 or above 170 bpm – Persistence of heart rate variability of less than 5 bpm for more than 40 minutes – Severe variable decelerations or severe repetitive early decelerations. – Prolonged decelerations – Late decelerations: the most ominous trace is a steady baseline without baseline variability and with small decelerations after each contraction – A sinusoidal pattern

2.2.2 Fetal electrocardiogram analysis

ST analysis of fetal electrocardiogram was successfully introduced into clinical practise by Neoventa Medical, Moelndal, Sweden. This technique is commonly referred to as STAN® (ST ANalysis). Contrary to CTG, the complete ECG curve is used to examine and evaluate morphological changes. The ST analysis is not intended to be used autonomously but only as addition to standard CTG. It serves as source of additional information validating or invalidating hypothesis of fetal condition and behaviour observed on CTG. The analysis of ST segment is well established in detecting and monitoring of myocardial insufficiency in adults cardiology and the development of ST analysis of fetal ECG has been based on this experience and knowledge. The fetal brain and heart are equally sensitive to changes in oxygen content; therefore, myocardial function serves as indirect measurement of brain condition.

The ECG signal is acquired by internal electrodes screwed into the fetal scalp without any damage to fetus. The continuous ECG is displayed and important markers of ECG are automatically computed. These markers involve changes in T wave amplitude and ST segment. For illustration of important ECG waves and intervals see Figure 2.7.

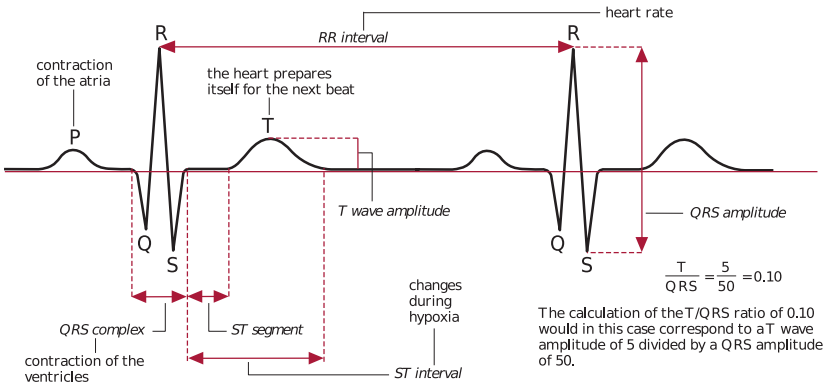


Figure 2.7: The representation of ECG curve and its important features (Sundström et al., 2000).

The T wave amplitude is used for computation of T/QRS ratio. This is performed periodically on ensemble average of several consequent beats. An increase in T wave reflects to fetus hypoxia and the degree of rise corresponds to degree of hypoxia. The second important feature of ECG is ST segment where its changes are examined. The biphasic ST is defined as a downward-leaning ST segment. We distinguish different degrees of biphasic ST segment starting at Grade 1 and continuing to Grade 2 and 3. With progression of disturbance in myocardial function, there is a shift in degree from Grade 1 to Grade 2 or even to the worst Grade 3. The morphologies of particular biphasic degrees are shown in Figure 2.8.

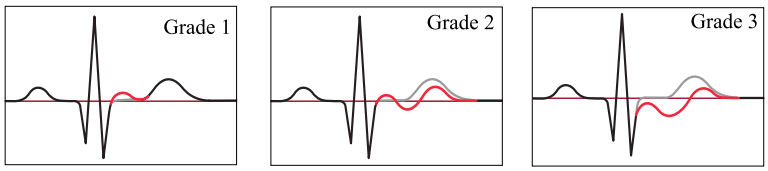


Figure 2.8: The morphology of biphasic ST. In Grade 1 the ST segment is above isoelectric line whereas in Grade 3 is completely below (Sundström et al., 2000).

As mentioned above, the ST analysis should be performed after occurrence of suspicious patterns on CTG. The sole assessment of ST segment could lead to misleading results and rise in the labour intervention (Sundström et al., 2000) and potential adverse outcomes for fetus as well as for mother. As for the CTG, the interpretation of ST segment was standardized and guidelines were created in order to avoid subjective assessment of ST changes. In this guidelines we distinguish three types of events: episodic T/QRS rise, baseline T/QRS rise, and biphasic ST. The T/QRS rise is considered as episodic when the T/QRS rises and returns to the baseline in time period no longer than 10 minutes. The degree of change indicates the fetal stress and corresponds to short lasting hypoxia. The T/QRS increased of more than 0.10, in connection with abnormal CTG, is considered as significant and registered as an ST event. Baseline T/QRS rise is similar to episodic rise with exception that increase of T/QRS has duration longer than 10 minutes. The baseline T/QRS rise of 0.05 with CTG classified as abnormal, is consider as significant and indicates persistent stress and zero opportunity to recover. The last event assessed is the biphasic ST with different degrees where the degree corresponds to the degree of abnormality. The grade 2 and 3 are generally considered as abnormal. The above mentioned events are connected with CTG interpretation in guidelines that are called STAN simplified guidelines (Sundström et al., 2000).

2.2.3 Other methods

Auscultation

Auscultation in general refers to listening for sounds produced within the body. Fetal heart tones can also be monitored during pregnancy by auscultation with a specialized stethoscope. The auscultation gives the clinician short term discrete evaluation of the fetal heart rate function.

Continuous phonocardiography

Recently automated, microphone based, auscultation techniques try to tackle the biggest disadvantage of the method by introducing continuous monitoring (Jiménez-González and James, 2009, 2010). Continuous phonocardiography has been nearly non-existent in Europe since the beginning of the 1990's. It is used more in the developing countries nevertheless there are signs that such concept could be useful for telemonitoring applications as developed recently within the EU funded ENIAC-MAS project.

Fetal blood sample

Fetal blood sampling (FBS) is used in conjunction with EFM and serves as an accurate tool for measurement of metabolic acidosis. In case of non-reassuring patterns on CTG or DECG the FBS might be performed in order to acquire precise value of pH. The small sample is obtained from fetal scalp capillary. The fetal blood sampling requires expertise and is time-consuming. It may also cause complications (Cunningham, 2005) but it is generally considered to be safe (Ojala et al., 2006).

Pulse oximetry

This method uses light reflection from blood where light is differently reflected or inhibited depending on the oxygen saturation ($\mathrm{FSp0_2}$) in fetal blood. The electrode emitting and receiving light is placed against fetal scalp and continuous $\mathrm{FSp0_2}$ is acquired. However, as it was noted by (Cunningham, 2005; Steer, 2008), low oxygen saturation has poor specificity for acidosis. Therefore, application of pulse oximetry made no significant contribution to any measures of fetal outcome.

2.3 Assessment of labour and neonate outcome

In the previous section we introduced electronic fetal monitoring as the methodology to identify fetal distress and oxygen insufficiency. When child is born, we need to assess its status in order to acquire additional information whether to what extent baby suffered. The commonly used methods for assessment are Apgar score, cord acid-base analysis, and the occurrence of neonatal complications.

2.3.1 Apgar score

This method was devised by Virginia Apgar in 1953. It was not initially intended to asses neonates that suffered from asphyxia. However, this methodology was established in clinical settings and is widely used. The Apgar score includes five parameters that are examined at the neonate's age of 1, 5, and 10 minutes. These parameters are heart rate, breathing, skin colour, muscular tone, and excitability. Each parameter is given score in range of 0 - 2 points and then all parameters are summed up giving the score at particular child's age. The maximum score that could be achieved is 10 points. Note that the assessment of child is subjective and a high observer variability was reported (O'Donnell et al., 2006).

There is a high correlation between low Apgar score at 5 minutes and neonates that suffered from asphyxia during labour (Manganaro et al., 1994). However, there are also many reasons for low Apgar score that are not related to asphyxia, such as immaturity, labour trauma, drugs, infection, the

activation of reflexes through manipulation of the upper airways, meconium aspiration, or carbon dioxide narcosis (Sundström et al., 2000).

The Apgar score below or equal to 7 at 5 minutes is considered as an indicator of metabolic acidosis (Doria et al., 2007; Westerhuis et al., 2007b) and cerebral palsy (MacLennan, 1999).

2.3.2 Acid-base analysis

Analysis of blood gases from umbilical cord blood can indicate to what extent a baby suffered by hypoxia during a labour. When a child is born, the cord is immediately doubly clamped and samples are taken from artery and vein. From these samples the values of blood gases are calculated, offering information on labour outcome. Many papers have studied the blood gases and their relation to adverse outcomes of labour; a great review is provided in (Armstrong and Stenson, 2007). There are wide controversies regarding biochemical markers (pH, base excess, and base deficit) obtained from blood gas analyses. The most prominent being discussion on exact relation to possible further complications for child development.

The pH is determined by presence of respiratory and metabolic acids and computed as logarithm of hydrogen ion activity. Because of the logarithm the relation of cumulative exposure to hypoxia and value of pH is nonlinear, e.g. the change of pH from 7 to 6.9 is almost twice as much as the change from 7.3 to 7.2. The other biochemical markers, base excess (BE) and base deficit (BDecf), are more linear.

There are many other factors that influence value of pH, the best example is elective caesarean section without labour (Riley and Johnson, 1993) where the pH values are similar to adults. Another aspect is the connection of CTG patterns to pH values. A performed caesarean section, because of suspicious/pathological CTG trace, prevents a baby to get into real asphyxia and the suspicious/pathological trace is not reflected by low pH value. Also the pH is only weakly correlated to clinical annotation (Schiermeier et al., 2008b; Spilka et al., 2013a; Valentin et al., 1993).

Abnormal labour outcome There is no general agreement, which biochemical marker (pH, BE, BDecf) is the best for identifying abnormal labour outcome, nor there is agreement on which threshold value should be used. Different studies were performed, each focused on different biochemical markers and slightly different outcome measures, e.g cerebral palsy (MacLennan, 1999), neonatal encephalopathy, perinatal mortality, 5-minute Apgar scores, and neonatal unit admission (Yeh et al., 2012), Apgar less than 7 at 5 minutes, NICU admission (Victory et al., 2004). We note here that another aspect, not discussed here, is to relate the adverse outcome directly to intrapartum period. The comparison on different biochemical markers is not straightforward and below we present only short overview.

The median pH values is 7.22 (interquartile range 7.17 – 7.27) (Yeh et al., 2012) with similar values, 7.24 ± 0.07 reported earlier by (Victory et al., 2004). Thorp et al. (1989) stated that pH should be preferred to other biochemical markers. Georgieva et al. (2013b) concluded that pH is the most robust marker to potential adverse outcomes even though its relation to adverse outcome is weak (Georgieva et al., 2013b; Yeh et al., 2012).

From the studies on cerebral palsy in neonates pH and BDecf are recommended as preferred measures (Pierrat et al., 2005) even though (Low, 2005) provided the contrary. Additionally intrapartum events and cerebral palsy are very rarely related by the intrapartum hypoxia only (Schifrin, 2004) and the real outcome of the delivery can be seen only in several years-long follow up (Ingemarsson et al., 1997). The base deficit was established by (Siggaard-Andersen and Huch, 1995) and Rosén et al. (2007) stated that it is the only usable measure for assessment of metabolic hypoxia and that base excess is erroneously used in many papers as well as in the clinical practice (Rosén et al., 2007). Below we present a short overview of thresholds used for biochemical markers.

- pH < 7.00 together with BDecf≥ 12 (MacLennan, 1999) was found to be related to significant increase of possibility of cerebral palsy. The pH < 7.00 was also recommend as a value that

defines pathological acidemia (Goldaber et al., 1991).

- pH ≤ 7.05 or pH < 7.05 is used as a threshold by many authors (Amer-Wåhlin and Maršál, 2011; Costa et al., 2009; Siira et al., 2007). Even though this value is not used unanimously it is generally accepted as the threshold between pathological and not-pathological delivery outcomes. Combination with BDecf was used e.g. in (Keith et al., 1995; Westerhuis et al., 2007b).

- pH ≤ 7.10 or pH < 7.10 (Cahill et al., 2012; Fulcher et al., 2012; Georgoulas et al., 2006; Yeh et al., 2012) – this value is supported by recent works on the large Oxford database as well as used heuristically in this thesis as a sign of severe problems with the delivery.

- pH ≤ 7.15 (Chung et al., 1995; Tommaso et al., 2013) – this value is based on the standard deviation (or 25th percentile). For simplicity rounded to 7.15 but if adhered strictly the threshold should be 7.17 (Victory et al., 2004; Yeh et al., 2012).

In general pH is more robust (Georgieva et al., 2013b) but is affected more by respiratory asphyxia, BDecf is more about metabolic asphyxia. Biochemical measures are very dependent on the measuring procedure – pH is in general considered to be more robust than the BDecf; since for the BDecf the pCO_2 has to be used, which could be measured erroneously (Kro et al., 2010).

Chapter 3

Automatic analysis of FHR – state of the art

A lot of attempts have been made to tackle the unresolved problem of reliable automatic analysis of CTG signal but, unfortunately, none of them were successful enough to be able to meet demands and expectation of clinicians. The automatic classification of fetus behaviour and condition is still challenge for many researches. In this chapter we briefly introduce solutions that were developed and used for automatic assessment of CTG records. It is important to mention that none of the complete systems we are going to describe is widely applied in clinical settings and obstetricians still rely on visual assessment of CTG tracings. Each system is merely used in the place or in the country where it was developed and a solution that would improve CTG interpretation still awaits (Bernardes and Ayres-De-Campos, 2010).

3.1 Clinical point of view

From the clinical point of view, there are still efforts to link antepartum and intrapartum events to adverse fetal outcomes, either regarding risk factors (Locatelli et al., 2010; Wayenberg, 2005) or connecting to FHR (Parer et al., 2006; Westgate et al., 2007). The later paper described a link between hypoxia and decelerations and presented possible new features like variability between decelerations and overshoot. The different types of variable decelerations were investigated by (Hamilton et al., 2012) and only the most serious decelerations (amplitude more than 60 BPM and length of 60 seconds and more) were found significant to labour outcome.

There is still lack of international consensus on clinical guidelines and FIGO guidelines from 1986 are still in use – even it is well known that they are in some cases inappropriate. There are international alternatives (ACOG, 2009; Macones et al., 2008; NICE, 2007; RCOG, 2001) but there is lack of agreement at many key concepts (de Campos et al., 2010). There are efforts to simplify guidelines or create ones (Parer and Ikeda, 2007; Parer et al., 2009), which are claimed superior to the traditional guidelines (Coletta et al., 2012; Tommaso et al., 2013) but these efforts are not generally acknowledged (Miller and Miller, 2012).

3.2 Overview of CTG databases

Most works use very small ad-hoc acquired datasets, differently sampled with various parameters used as outcome measures. We aimed to bring as detailed overview of databases as possible but the exhaustive description of databases was infeasible, therefore, several inclusion criteria were applied. First, if a CTG database was used in multiple works, we included the paper where the database was described in the most detail, e.g. we preferred paper of (Jezewski et al., 2010) rather than of (Czabanski et al., 2012). If the description was the same, we included the most recent paper. Second, only those

works that used intrapartum CTG signals were considered, e.g. we did not included the work of (Ocak, 2013) since he worked with Cardiotocography Data Set (UCI[1]). Third, we preferred journal papers and works that attempted to show results with regards to objective annotation (pH, base deficit, etc.).

We do not provide exhaustive description of used databases in text since we believe that the overview in the tables is self-explanatory. Due to the space limitation the overview had to be split into two tables, Table 3.1 and 3.2. In Table 3.1 we present used databases regarding the CTG signals and clinical parameters. The number of cases varies from study to study, the lowest being around 50 cases, and the highest being 7568 cases. In Table 3.2 we present the overview of databases from classification point view, it is apparent that in each paper different criteria for classes division were used, thus, making any comparison of results between different studies virtually impossible.

3.3 Automatic FHR evaluation – the origins

A very first attempt for automatic CTG analysis was to follow the clinical guidelines used for CTG assessment (FIGO, 1986). The morphological (FIGO) features have become fundamental for almost all works that have attempted to classify fetus status. Beginning with work of (Dawes et al., 1981) the guidelines were essential for any automatic evaluation. Dawes et al. (1982b) proposed an algorithm for baseline estimation and extraction of accelerations and decelerations. Mantel et al. (1990a,b) also developed an iterating procedure for complete FIGO features extraction. The extraction of morphological features were improved by (Bernardes et al., 1991) and resulted into development of automatic system, SisPorto (de Campos et al., 2008), for CTG analysis which is briefly described below. The first automatic classification system were described by (Nielsen et al., 1988) and further by (Chung et al., 1995; Keith et al., 1995). From the early works of (Dawes et al., 1981) the research efforts in the field of CTG extended into various areas and focused in detail on particular components. The description of the research of CTG in the past years follows.

3.4 Features for FHR

FIGO features There exist many approaches to estimate a baseline of fetal heart rate, which is the key concept in the analysis of CTG based on clinical guidelines. Initially (Dawes et al., 1982a) proposed an algorithm for baseline estimation and extraction of accelerations and decelerations. Among the most commonly used approaches are stable segments (de Campos and Bernardes, 2004; de Campos et al., 2004), filtering approach (Pardey et al., 2002; Taylor et al., 2000), fetal heart rate density approach (Georgieva et al., 2011; Jimenez et al., 2002), or others (Kupka et al., 2006; Mantel et al., 1990a). The complete extraction of FIGO features (baseline, accelerations, and decelerations) was first proposed by (Dawes et al., 1982b) and further in (Mantel et al., 1990a,b).

Short term variability (STV) STV is uninterpretable by naked eye and thus remains one of the few automatically detected and assessed features FHR in clinical practice. Comparison of the most used STV indexes is presented in (Cesarelli et al., 2009). STV is in general used mainly for antepartum evaluation. Evaluation of STV for intrapartum period with negative outcome was done by (Schiermeier et al., 2008a).

Frequency features The frequency features are commonly used for FHR analysis though the delineation of different spectral bands is not well-studied as in the adult HRV where interpretation of different bands was stated in (Task-Force, 1996). Frequency features were examined in (Sibony et al., 1994; Signorini et al., 2003) and further in (Siira et al., 2007). The recent paper (Laar et al., 2008) gives a short overview of papers, which analysed spectrum to FHR either antepartum or intrapartum.

[1] http://archive.ics.uci.edu/ml/datasets/Cardiotocography

The association of frequency features to the uterine contractions was investigated by (Warrick and Hamilton, 2012).

Nonlinear features Use of non-linear methods for FHR analysis has also its roots in adults HRV research where these methods have proven their usefulness. The measure of fractal dimension of reconstructed attractor was performed by (Chaffin et al., 1991; Felgueiras et al., 1998; Kikuchi et al., 2006). Felgueiras et al. (1998) also examined waveform fractal dimension. A slightly different approach was applied by (Gough, 1993) who measured the length of FHR at different scales and thus estimated fractal dimension. Probably the most successful non-linear methods for FHR analysis are approximate entropy (ApEn) and sample entropy (SampEn). They are widely used for examination of non-linear systems and also proved their applicability in FHR analysis. Let us mention only few studies that employed ApEn or SampEn (Georgoulas et al., 2006; Gonçalves et al., 2006a; Lake et al., 2002; Pincus and Viscarello, 1992). Other methods for non-linear analysis are detrend fluctuation analysis applied by (Echeverria et al., 2004) and Lempel Ziv complexity used by (Ferrario et al., 2005). The different estimation of fractal dimension were reviewed by (Hopkins et al., 2006) and, more comprehensively, in our work (Spilka et al., 2012). Recently a multi-fractal analysis was employed by (Doret et al., 2011) and multi-scale analysis by (Helgason et al., 2011).

3.5 Classification methods

As well as the abundance of features used for FHR description a lot of methods were used for the classification task. The methods were primarily based on the preference of researches the most used were Artificial Neural Networks and Support Vector Machines. The Artificial Neural Networks were employed in many works, e.g. in (Georgieva et al., 2013b; Jezewski et al., 2010; Keith et al., 1995; Maeda et al., 1998; Magenes et al., 2000). The exhaustive work of CTG analysis was performed by Georgoulas et al. For CTG classification they used Hidden Markov Models (Georgoulas et al., 2004), Support Vector Machines (Georgoulas et al., 2005), and a hybrid approach utilizing grammatical evolution (Georgoulas et al., 2007). They compared the classification performance of respective methods to conventional methods, such as k-nn (k-nearest neighbours), qdc (quadratic discriminant classifier), and ldc (linear discriminant classifier). The support vector machines were also used in (Czabanski et al., 2010; Warrick et al., 2010) and in our work (Spilka et al., 2012).

In Table 3.3 we provide comprehensive comparison of classification results. The table is aimed to be used in conjunction with Tables 3.1 and 3.2 where we present the used databases in more detail. Table 3.3 present subset of works used in Tables 3.1 and 3.2. The same selection criteria applied with additional one that we included only those work that presented any classification results. We believe that Table 3.3 is self-explanatory and no additional comments are required.

We aimed to visualize the relationship between results reported in the literature and data size that was used in particular work. The complete results are presented in Figure 3.1 where we plot accuracy (ACC), area under ROC (AUC), sensitivity (SE), specificity (SP), and positive predictive value (PPV) as a function of data size. Note that the for the x-axis the logarithm of the data size is plotted. Because of different metrics used the relationship is unclear. For the works that reported SE and SP we computed geometric mean, where $G-mean = \sqrt{SE \cdot SP}$. A better approach would be to use harmonic mean of SE and PPV (precision), the so called F-measure (He and Garcia, 2009) though very few studies reported the PPV. The relationship is shown in Figure 3.2. With increasing data size the $G-mean$ decreases. For the regression line estimated for all papers this decrease is about 4.5% per 100 examples. Note that the x-axis is not in logarithm scale since we excluded the largest study of (Elliott et al., 2010) where the sensitivity and specificity was not reported.

Table 3.1: Overview of databases used in various works – CTG signal and clinical point of view. Legend: "N/A" – information not available, "–" – used for column (FHR sig. only) and express that authors used the whole signal without specifying the length. The works are ordered by publication date. Parameters: type of acquisition (ultrasound Doppler (US), direct fetal electrocardiogram measurement (FECG)); timing of recording antepartum (ante.) or intrapartum (inte.) phase; stage of labour (I. or II.); length of FHR signal (FHR sig.); time to actual delivery; use of uterine contractions (UC), description of inclusion criteria; description of clinical data; evaluation type: objective (obj.), subjective (subj.), or combination of both (comb.); number of total cases.

Reference	acquisition	timing	labour stage	FHR sig. [min.]	time to delivery [min]	UC used	incl. criteria	clinical info.	evaluation type	# total cases
(Nielsen et al., 1988)	N/A	intra.	I.	30	N/A	yes	no	no	obj.	50
(Chung et al., 1995)	FECG	intra.	N/A	N/A	N/A	yes	yes	yes	obj.	73
(Keith et al., 1995)	N/A	intra.	N/A	> 120	until del.	yes	no	yes	comb.	50
(Bernardes et al., 1998)	US, FECG	ante., intra.	I.,II.	–	until del.	yes	no	yes	obj.	85
(Maeda et al., 1998)	N/A	intra.	N/A	50	N/A	no	no	no	subj.	49
(Lee and Dorffner, 1999)	FECG	intra.	N/A	–	N/A	yes	no	no	subj.	53
(Chung et al., 2001)	US	ante., intra.	I.,II.	N/A	120	no	no	yes	comb.	76
(Strachan et al., 2001)	FECG	intra.	I.,II.	> 30	until del.	yes	no	yes	obj.	679
(Siira et al., 2005)	FECG	intra.	I.,II.	60	95% bellow 9	yes	yes	yes	obj.	334
(Cao et al., 2006)	US, FECG	intra.	N/A	30	N/A	yes	no	no	subj.	148
(Salamalekis et al., 2006)	US	intra.	I.,II.	N/A	until del.	no	yes	yes	comb.	74
(Georgoulas et al., 2006)	FECG	intra.	I.,II.	20-60	until del.	no	no	no	obj.	80
(Gonçalves et al., 2006a)	US, FECG	intra.	I.,II.	32-60	until del.	no	yes	yes	obj.	68
(Costa et al., 2009)	FECG	intra.	I.,II.	–	until del.	yes	yes	yes	obj.	148
(Elliott et al., 2010)	N/A	intra.	I.,II.	> 180	until del.	yes	yes	yes	subj.	2192
(Warrick et al., 2010)	US, FECG	intra.	I.,II.	> 180	until del.	yes	yes	no	obj.	213
(Jezewski et al., 2010)	US	ante., intra.	N/A	–	N/A	yes	yes	yes	obj.	749[a]
(Helgason et al., 2011)	FECG	intra.	I.,II.	> 30	until del.	yes	no	no	comb.	47
(Chudáček et al., 2011)	US, FECG	intra.	I.,II.	20	until del.	no	yes	yes	comb.	552
(Spilka et al., 2012)	US, FECG	intra.	I.,II.	20	until del.	no	yes	no	obj.	217
(Georgieva et al., 2013b)	N/A	intra.	I.,II.	–	until del.	no	yes	yes	obj.	7568

[a] 749 recordings, 103 woman

Table 3.2: Overview of databases used in various works – description of criteria used for division into different categories. The works are ordered by publication date. Legend: NR – non-reassuring (NR), neonatal encephalopathy (NE).

Reference	Classes (categories)	division criteria for classes	# classes	# cases in classes	# total cases
(Nielsen et al., 1988)	normal ; pathological	apgar 1 min. < 7 or pH < 7.15 or BE < -10	2	34 ; 16	50
(Chung et al., 1995)	normal ; abnormal	pH < 7.15	2	65 ; 8	73
(Keith et al., 1995)	normal ; poor outcome	pH, BDecf, Apgar	2	38 ; 12	50
(Bernardes et al., 1998)	norm. ; susp. ; pathol.	pH, Apgar, neonatology	3	56 ; 22 ; 7	85
(Maeda et al., 1998)	norm. ; susp. ; pathol.	manual clinical rules	3	12 ; 18 ; 19	49
(Lee and Dorffner, 1999)	normal CTG ; decels.	1 clinician	2	N/A	53
(Chung et al., 2001)	normal ; presumed distress; acidemic	norm. FHR ; abnorm. and pH > 7.15 ; abnorm. and pH < 7.15 and BE < -8	3	36 ; 26 ; 14	76
(Strachan et al., 2001)	normal ; abnormal	pH ≤ 7.15 and BDecf ≥ 8	2	608 ; 71	679
(Siira et al., 2005)	normal ; acidemic	pH < 7.05	2	319 ; 15	334
(Cao et al., 2006)	reassuring ; NR	2 clinicians	2	102 ; 44	148
(Salamalekis et al., 2006)	normal ; NR [NR and pH > 7.20 ; NR and pH < 7.20]	FIGO, pH < 7.20	2	32 ; 42	74
(Georgoulas et al., 2006)	normal ; at risk	pH > 7.20 ; pH < 7.10	2	20 ; 60	80
(Gonçalves et al., 2006a)	normal ; mildly acidemic ; modest-severe acid.	pH ≥ 7.20 ; pH > 7.10 and pH < 7.20 ; pH ≤ 7.10	3	48 ; 10 ; 10	68
(Costa et al., 2009)	Omniview-SisPorto 3.5 alerts	pH < 7.05	2	7 ; 141	148
(Elliott et al., 2010)	normal ; abnormal	BDecf > 12 and NE	2	60 ; 2132	2192
(Warrick et al., 2010)	normal ; pathological	BDecf < 8 ; BDecf ≥ 12	2	187 ; 26	213
(Jezewski et al., 2010)	normal ; abnormal	apgar N/A min. < 7 or birth weight < 10^{th} perc. or pH < 7.20	2	28% abnorm.	749[a]
(Helgason et al., 2011)	FIGO-TN ; FIGO-FP ; FIGO-TP	norm. FHR and pH ≥ 7.30 ; abnorm. and pH ≥ 7.30 ; abnorm. and pH ≤ 7.05 [b]	3	15 ; 17 ; 15	47
(Chudáček et al., 2011)	norm. ; susp. ; pathol.	3 clinicians	3	139 ; 306 ; 107	552
(Spilka et al., 2012)	normal ; pathological	pH < 7.15	2	123 ; 94	217
(Georgieva et al., 2013b)	normal ; adverse	pH < 7.1 and neonatology	2	N/A	7568

[a] 749 recordings, 103 woman

[b] 1 clinician eval. using FIGO guidelines

Table 3.3: Overview of works that presented classification results. Due to space limitation not all details and results are shown. For the overview of database that were used for classification refer to Table 3.1 and 3.2. The works are ordered by publication date. Legend: CI – confidence intervals, CV – cross validation, NN – neural networks, SVM – support vector machine, LDA – linear discriminant analysis, εDA – ε-insensitive learning with deterministic annealing, SE – sensitivity, SP - specificity, AUC – area under ROC, ACC – accuracy, PPV – positive predictive value, G-mean – geometric mean, F-meas. – F-measure

Reference	# classes	data size	error estimation	classification	SE [%]	SP [%]	AUC	other
(Nielsen et al., 1988)	2	50	CI	discr. fun.				ACC: 86%
(Chung et al., 1995)	2	73		manual rules	88	75		ACC: 77%
(Bernardes et al., 1998)	3	85	CI	FIGO[a]	100	80		
(Maeda et al., 1998)	3	49	hold-out not stratified	NN				ACC: 86%
(Cao et al., 2006)	2	148		log. reg.	47	95	0.85	PPV 44%
(Georgoulas et al., 2006)	2	80	CV stratified	SVM	70	85	0.75	G-mean 77
(Gonçalves et al., 2006a)	3	68	leave one out	LDA	70	74		
(Salamalekis et al., 2006)	3	74		ANOVA	67	72	0.66	PPV 36%
(Costa et al., 2009)	2	148	CI	Omniview-3.5	57	97		PPV 50%
(Elliott et al., 2010)	2	2192		PeriCalm			0.53 (0.83)[b]	
(Czabanski et al., 2010)	2	685 (191)[c]	see[d]	εDA				ACC: 96%
(Jezewski et al., 2010)	2	749 (103)[c]	see[d]	NN	67	68		ACC: 67%; PPV 62%
(Warrick et al., 2010)	2	213	multiple 10-fold CV	SVM	70	75		
(Helgason et al., 2011)	3	47		adaptive complexity params.	100	60		
(Spilka et al., 2012)	2	217	10-fold CV	SVM	73	76	0.78	PPV 70%, F-meas. 72%
(Georgieva et al., 2013b)	2	376 (7568)[e]	CV; hold out	NN	61	68	0.64	

[a] baseline estimated by clinicians, complete FIGO estimated by computer
[b] red grade and above ; results in brackets blue grade and above
[c] presented as: recordings (woman)
[d] 50 rand. division to training, test, and validation set
[e] 376 records used for learning/testing NN, rest used for NN evaluation using EveREst plot

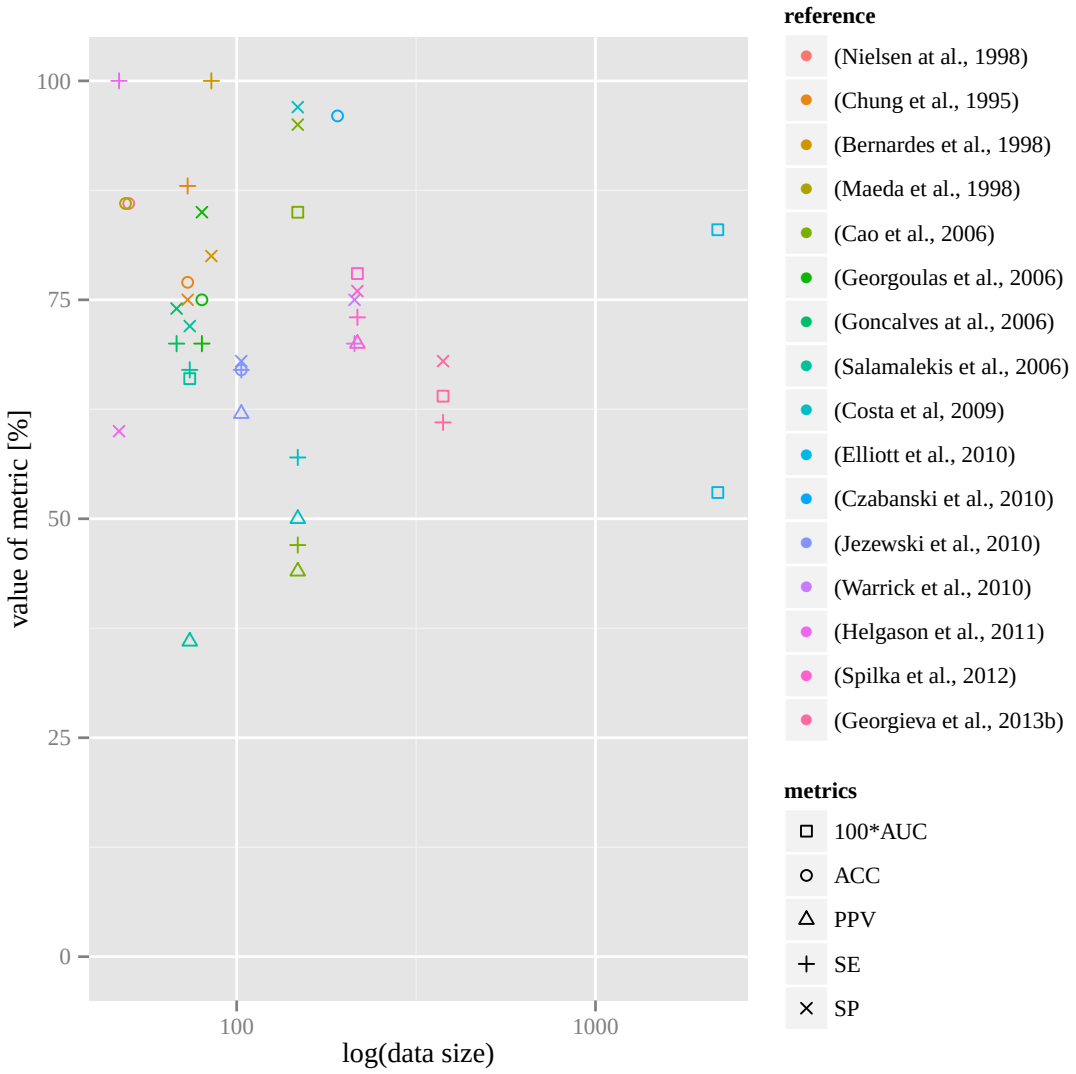


Figure 3.1: Relationship between different performance metrics and logarithm of data size – log(data size). Legend: 100*AUC – area under the ROC multiplied by 100% for visualization, ACC – accuracy, PPV – positive predictive value (= precision), SE – sensitivity, SP – specificity.

3.6 Fetal monitoring systems

A several systems have been developed and some of them resulted into commercial applications; the complete systems used for fetal assessment mostly employ an expert system. A brief description of each system follows. Based on the work of (Dawes et al., 1982b) a system for antenatal analysis was created – System 8000 (Dawes et al., 1991). This system was further improved (Dawes et al., 1996) and is nowadays commercially available, known as *sonicadFetalCare*. It uses FIGO-like features with additional parameter of short term variability. For antepartum monitoring there is also a *2CTG2* system (Magenes et al., 2007) that is result of works (Magenes et al., 2000, 2003; Signorini et al., 2003). *NST-Expert* (Non-Stress Test) (Alonso-Betanzos et al., 1995) is a non-invasive method used for fetal assessment. The main part of this software is an expert system that is capable of proposing a diagnose and treatment. Moreover, it might also estimate the potential problems of neonate. *CAFE* (Computer Aided Fetal Evaluation) (Guijarro-Berdiñas and Alonso-Betanzos, 2002) is successor of NST-Expert. *SisPorto* system has been developed by Bernardes at al. at University of Porto, Portugal, since 1990. It consists of an expert system which evaluates individual features described according to guidelines for CTG assessment. Today, SisPorto has matured to its 3-rd version and is known as Omniview-SisPorto ® 3.5. (de Campos et al., 2008). The *K2 Medical Systems* (Greene and Keith, 2002; Keith and

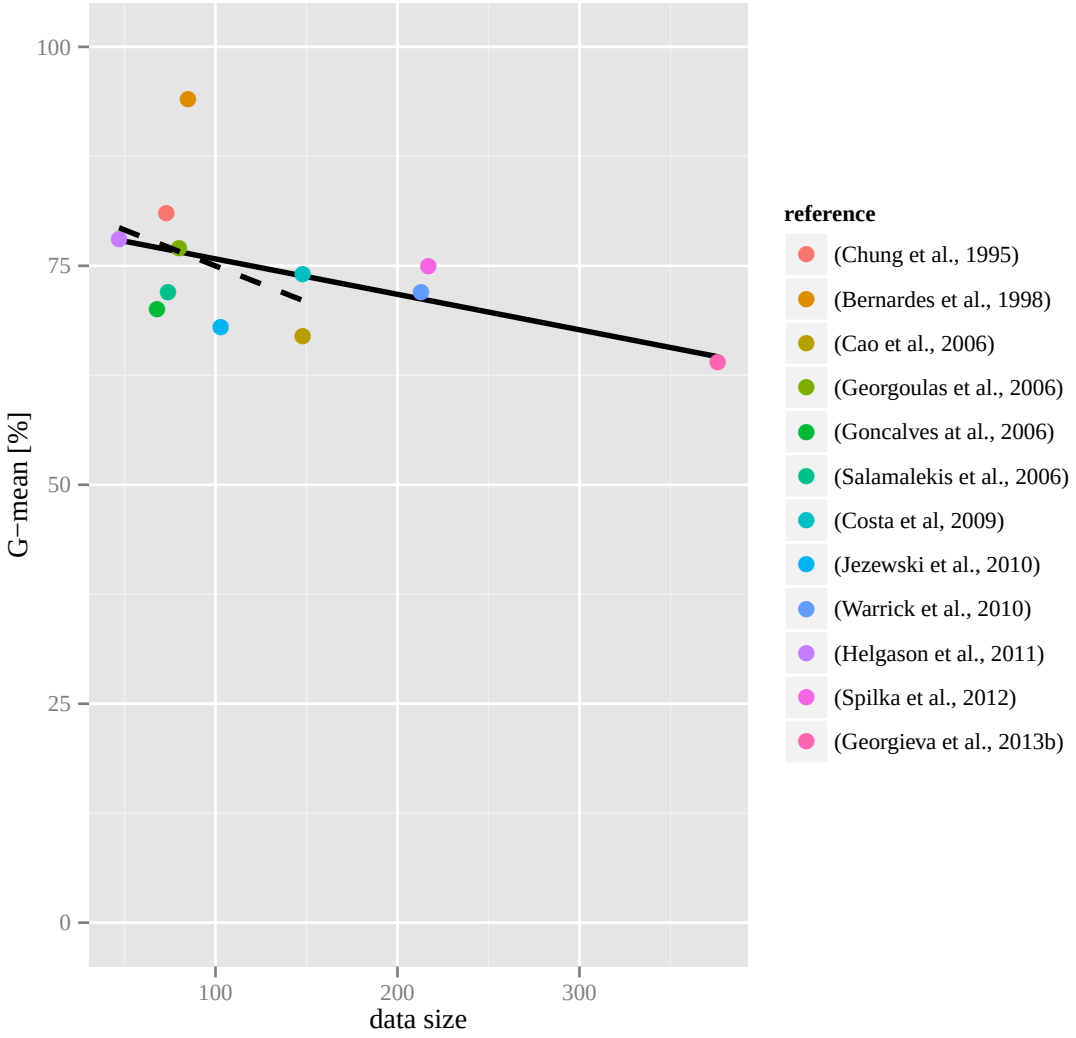


Figure 3.2: Performance of classification ($G-mean$) as a function of data size. Trend was estimated using linear least squares. Legend: solid black line – trend using all papers present in legend, dashed black line – trend estimated for papers with data size < 200.

Greene, 1994) has been developed by K2 Medical Systems™, Plymouth, UK. It is a distributed system consisting of central PC and local units that are situated at the patient's bed and gathering information, such as CTG and results of blood sample analysis. Nowadays the system is known as INFANT®. The alarm is evoked in case of abnormalities. The new NIHCD guidelines are employed in the monitoring system PeriCALM™ (Elliott et al., 2010; Parer and Hamilton, 2010) , developed by LMS Medical systems, Montreal, Canada and PeriGen, Princeton, USA. The comprehensive overview of central monitoring system was provided by (Nunes et al., 2013).

3.7 Other techniques and alternatives to CTG

Beside the CTG analysis there is a research effort in fetal electrocardiogram using abdominal electrodes (Clifford et al., 2011; Piéri et al., 2001), direct fetal electrocardiogram (ST-analysis) (Rosén et al., 2004), phonocardiography (Kovács et al., 2011), or magnetocardiography (Kariniemi et al., 1974; Kiefer-Schmidt et al., 2012). The magnetocardiography has not been established in the clinical practice yet and the phonocardiography is only used in antenatal period for screening potential fetal complications. On the other hand, the fetal electrocardiogram is used widely and has become common

in obstetrics. The so called ST analysis (STAN) of fetal electrocardiogram was employed in the clinical practice recently. Many randomized control trials were published and, most probably, will be published in the future on the topic whether STAN improves fetal outcomes in comparison to sole use of CTG. The majority of studies proved that addition of STAN indeed lead to better fetal outcomes (Amer-Wåhlin and Maršál, 2011; Amer-Wåhlin et al., 2001; Norén et al., 2003, 2006) but there were also few studies disproving this (Ojala et al., 2006; Westerhuis et al., 2007a). A promising research has been announced by Neoventa Medical AB; they, together with a company Nanexa AB (nanotechnology), plan to develop a nanomaterial sensor that will measure lactate intrapartum, see press release (Nanotechnology opens up new possibilities in perinatal care, http://nanexa.com/). It will therefore offer a possibility to monitor oxygen insufficiency and indicate fetal complications.

Chapter 4

Experimental data (collection and structure)

One of the main obstacles for improvements in the CTG analysis and classification is the lack of any publicly available database. Based on the critical analysis presented in Chapter 3 we decided to systematically design and develop a CTG database satisfying the following requirements: open access, reasonable size, systematic selection, and complete clinical information.

The CTU-UHB[1] database consists of two parts, CTG recordings and clinical data. In total 552 records were carefully selected from 9164 intrapartum recordings, which were acquired between 27th April 2010 and 6th August 2012 at the obstetrics ward of the University Hospital in Brno, Czech Republic and stored in electronic form in the OB TraceVue® system. The resulting signals for the database were selected with clinical as well as technical considerations in mind. Detailed description of the database and reasoning behind the selection of the parameters is presented.

When reviewing literature on automatic CTG processing, two things are striking. First, there is a large disconnection between approaches and goals in the clinical and technical papers. While the clinical papers are mostly looking for applicable solutions to their problems (lack of agreement, sometimes critically misclassified recordings), the technical papers often use CTG data as just an another input to the carefully tuned classifiers. Most works use very small ad-hoc acquired datasets, differently sampled with various parameters used as outcome measures, though we have to concede that our previous works (Chudáček et al., 2011; Spilka et al., 2012) were done in the exact same manner. It is hard to believe that it is more than 30 years since computer processing of CTG began (Dawes et al., 1981) and since then, no common database of CTG records is available. There is no way how to compare/improve/disregard among different results that hinder any progress towards the ultimate goal of a usable and working automated classification of the CTG recordings.

In this chapter we present a novel open-access CTU-UHB database consisting of CTG records and clinical information. The CTU-UHB database was designed and developed based on a new methodology that is proposed for the future development of similar databases that can serve both for extraction of medical knowledge, routine classification, and development and testing of new algorithms. The criteria for the selection of records for the database are discussed from both a clinical and technical point of view. We also present a detailed description of the main clinical and technical parameters, which, in our opinion, are important for understanding and should be taken into account when using the database. This chapter is largely based on the paper (Chudáček et al., 2013).

4.1 Ethics statement

The CTG recordings and clinical data were matched by anonymized unique identifier generated at the side of hospital information system. The timings of CTG records were matched to stages of labour

[1] Czech Technical University – University Hospital Brno

(first and second stage) and were made relative to time of the birth, thus also de-identified. This study was approved by the Institutional Review Board of University Hospital Brno; all women signed informed consent.

4.2 Data collection

The data were collected between 27th of April 2010 and 6th of August 2012 at the obstetrics ward of the University Hospital in Brno, Czech Republic. The data consisted of two main components, the first were intrapartum CTG recordings and the second were clinical data.

The CTGs were recorded using STAN S31 (Neoventa Medical, Mölndal, Sweden) and Avalon FM40 and FM50 (Philips Healthcare, Andover, MA). All CTG signals were stored in an electronic form in the OB TraceVue® system (Philips) in a proprietary format and converted into text format using proprietary software provided by Philips. Each CTG record contains time information and signal of fetal heart rate and uterine contractions sampled at 4 Hz. When a signal was recorded using internal scalp electrode it also contained T/QRS ratio and information about biphasic T-wave. From 9164 intrapartum recordings the final database of 552 carefully selected CTGs was created keeping in consideration clinical as well as technical point of view; the details about recordings selection are provided further.

The clinical data were stored in the hospital information system (AMIS) in the relational database. Complete clinical information regarding to delivery and fetal/maternal information were obtained. The clinical data included: delivery descriptors (presentation of fetus, type of delivery and length of first and second stage), neonatal outcome (seizures, intubation, etc.), fetal and neonatal descriptors (sex, gestational week, weight, etc.), and information about mother and possible risk factors. For the final CTU-UHB database clinical data were exported from relational database and converted into physionet text format (Goldberger et al., 2000).

4.3 Data selection and criteria considered

The selection procedure of records was based on clinical and CTG signal parameters and performed in steps over-viewed in Figure 4.1.

4.3.1 Clinical criteria

In the following paragraphs we describe criteria and their reasoning that were used for exclusion of portion of recordings. Then we will shortly discuss criteria that were included in the final database but no restrictions on creation of the final database were based upon them.

Clinical selection criteria The following parameters were taken into account for selection of recordings for the final database. References in this section refer to a description of particular parameter.

- Women's Age – although the women's high age plays significant role in the probability of congenital diseases, for the intrapartum period no significance was found (Callaway et al., 2005). Low age (< 18 years) could have an adverse effect and was therefore excluded (Berglund et al., 2010).
- Week of gestation – maturity of the fetus plays significant role in the shape and behaviour of the FHR antepartum as well as intrapartum (Park et al., 2001). Therefore the selection was limited to mature fetuses: $week\_of\_gestation \geq 37$ according to last menses counting, which was in majority cases confirmed by ultrasound measurement during antepartum check-ups.
- Known fetal diseases – fetuses with known congenital defects or known intrauterine growth restriction (IUGR) that could influence the FHR and/or outcome of the delivery were excluded

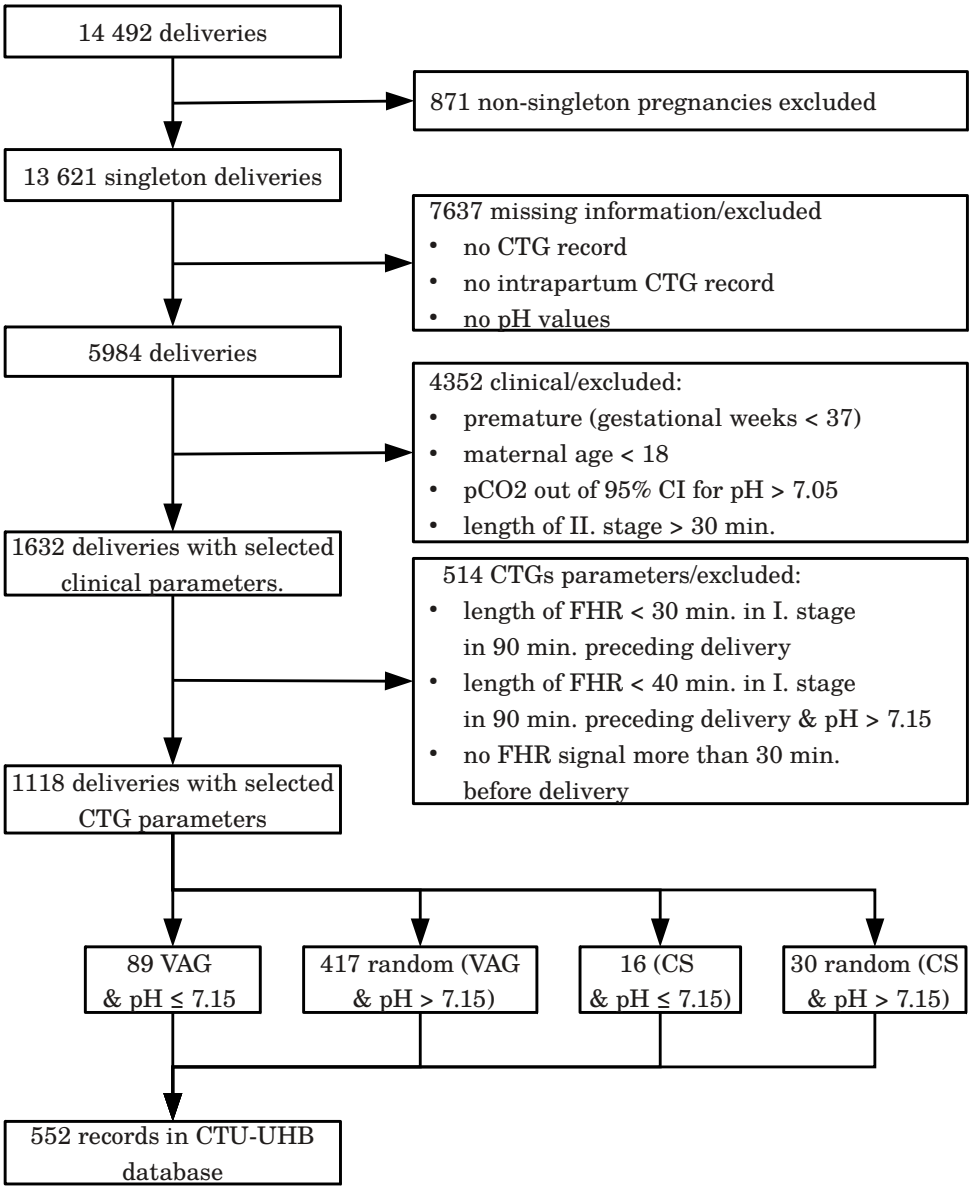


Figure 4.1: Selection of recordings for the final database.

from the database. Additionally, post-natally detected defects were consulted and two cases with transposed large veins were left in the set, since these two particular changes should not have influenced the FHR.

- Type of gravidity – only singleton, uncomplicated pregnancies were included.
- Type of delivery – the majority of the database consists of vaginal deliveries. Nevertheless to increase the number of cases with pathological outcome in the database, 16 CS recordings with pH ≤ 7.15 were included and consequently control group consisting of 30 CS with normal outcomes was also included to enable separate evaluation if necessary.

Additional clinical criteria provided Together with criteria used for selection, following criteria were considered and are available together with the CTG data:

- Sex of the fetus – both sexes were included even though the sex of fetus significantly influences the outcome according to (Bernardes et al., 2009).

- Parity – although the first delivery can be "more difficult" in general clinical sense (Singh et al., 2008) it is the same from the point of view of the FHR recording.
- Risk factors – to be able to describe and identify the a priori high-risk pregnancies we have included risk factors that could have influenced the state of the baby before entering the labour. For full review of the parameters and further references we recommend paper of (Badawi et al., 1998). The final risk-factors included in the database were gestational diabetes, preeclampsia, maternal fever (>37.5 °C), hypertension and meconium stained fluid.
- Drugs – especially those administered during delivery were considered only with regard to their influence on FHR. Opiates may influence the FHR directly but are rarely used in the Czech Republic during delivery and were not used in any of the cases included in the database. Therefore, we do not provide information about drugs administration in the database. Note that e.g. oxytocin used for enhancement of the uterine activity influences the FHR in majority indirectly, via increase of uterine activity, and thus can be assessed from the CTG alone.
- Other criteria – complementary information in order to offer insight why e.g. operative delivery was chosen. These include: induced delivery, type of presentation (occipital/breech), no progress of labour, dystocia cephalokorporal (in-coordinate uterine activity), dystocia cephalopelvic.

4.3.2 Labour outcome measures

Since our main intention was to prepare database that could be used for comparison of different automated approaches we have selected only those recordings that included umbilical artery pH. We added all additional outcome measures that were available for the recording in the hospital information system. Some of these measures are often misused and we will discuss their disadvantages below. The measures include:

Outcome measure selection criteria To enable objective classification the pH measure was considered as essential for the evaluation of the database.

- Umbilical artery pH (pH) – is the most commonly used outcome measure, sign of respiratory hypoxia. Records with missing pH were excluded. Following suggestion by (Rosén et al., 2007) records, which had values of pCO_2 outside 95th percentile (Kro et al., 2010) were excluded except those with pH ≤ 7.05, which even according to (Kro et al., 2010) should be approached with care.

Additional outcome measures provided Even though the is pH is the most commonly used measure, it is worth to include additional measures such as following:

- Base excess (BE) – is often used in the clinical setting as a sign for metabolic hypoxia, but is often false positive (Rosén et al., 2007).
- Base deficit in extracellular fluid (BDecf) – is according to (Rosén et al., 2007) better measure of metabolic hypoxia than BE. Still pH remains more robust measure and according to last study of remains the most informative (Georgieva et al., 2013b).
- Neonatalogy – complete neonatological reports were acquired for all the cases in pre-prepared database. No severe cases of neonatal morbidity were found, no hypoxic ischemic encephalopathy, no seizures (for details on neonatal morbidity see (McIntyre et al., 2012)).
- Subjective evaluation of the outcome of the delivery based on Apgar's score (Apgar), where five categories are used to assess the newborn child in 1st, 5th and 10th minute (Finster and Wood, 2005).

The complete database was used for inter-intra observer variability study described in Chapter 6.

4.3.3 Signal criteria

When the data were filtered according to the clinical information, we have applied the following criteria on CTG records:

- Signal length – we have decided to include preceding 90 minutes before delivery, where delivery time is represented by the time when the first objective evaluation of labour was acquired.
  - I. stage – the length of the 1st stage was limited to maximum of 60 minutes in order to keep recordings easily comparable. The minimal length was dependent on the pH of the recording in question – to include as much abnormal recordings as possible. Thus minimal length of the I. stage of 30 minutes was required for recording with pH ≤ 7.15 and 40 minutes for others. The distance to birth was not allowed to be further than 30 minutes.
  - II. stage – based on our previous experience with analysis of II. stage of labour (active pushing phase), we limited the II. stage to 30 minutes at maximum. This also limits possibility of adverse events occurring in the II. stage, which could disconnect CTG recording in the I. stage with objective evaluation of the delivery.

  Given the restriction above the signals are 30(40)–90 minutes long depending on the length of the II. stage and also available signal in the I. stage. No signal ends earlier than 30 minutes before delivery.
- Missing signal – amount of missing signal was, except of the II. stage, kept to possible minimum. Nevertheless the trade-off between having full-signal and having recordings with abnormal outcomes had to be made. No more than 50% of signal was allowed to be missing in the I. stage.
- Noise and artifacts – these are a problem especially for the recordings acquired by the ultrasound probe. Certainly in some recordings maternal heart rate is intermittently present. But even though it can pose a challenge for user of the database it also reflects the clinical reality.
- Type of measurement device – the database is composed as a mixture of recordings acquired by ultrasound Doppler probe, direct scalp measurement or combination of both – again reflecting the clinical reality at the obstetrics ward of UHB.

4.4 Results

4.4.1 Description of the Database

Records for the CTU-UHB database were selected based on clinical and technical criteria described above. Table 4.1 provides overview of patient and labour outcome measure statistics and Table 4.2 presents main parameters regarding the CTG signals. The CTG signals were transformed from proprietary Philips format to open Physionet format (Goldberger et al., 2000), all data were anonymized at the hospital and de-identified (relative time) at the CTU side. An example of one CTG record is shown in Figure 4.2.

CTG database – vaginal deliveries

The main part of the CTG database consists of 506 intrapartum recordings delivered vaginally. It means the deliveries got always to the II. stage of labour (fully dilated cervix, periodical contractions), even though not all deliveries had active pushing period. Some were delivered operatively by means of forceps or vacuum extraction (VEX). The main outcome measures are presented in Table 4.1. Please note that column "Comment", which gives additional information either with regard to number of potential outliers or points out interesting features of the database such as number of pathological cases based on certain parameters or quality of the recording in each window.

Table 4.1: Patient and labour outcome statistics for the whole CTG-UHB cardiotocography database.

506 – Vaginal (44 – operative); 46 – Caesarean Section US = 412; DECG = 102; US-DECG = 35; N/A = 3				
	Mean	**Min**	**Max**	**Comment**
Maternal age (years)	29.8	18	46	over 36y: 40.
Parity	0.43	0	7	
Gravidity	1.43	1	11	
Gestational age (weeks)	40	37	43	over 42 weeks: 2
pH	7.23	6.85	7.47	pat.: 48; abnormal.: 64
BE	-6.36	-26.8	-0.2	pat.: 39; abnormal: 121
BDecf (mmol/l)	4.60	-3.40	26.11	pat.: 25; abnormal.: 68
Apgar 1min	8.26	1	10	AS1 < 3: 18
Apgar 5min	9.06	4	10	AS5 < 7: 50
Neonate's weight (g)	3408	1970	4750	small: 17; large: 44
Neonate's sex (F/M)		259 / 293		

Table 4.2: CTG signal statistics. W1 – 30 minute window beginning 60 minutes before end of the 1st stage of labour, W2 – 30 minute window before the end of 1st stage of labour

506 – Vaginal (44 – operative); 46 – Caesarean Section US = 412; DECG = 102; US-DECG = 35; N/A = 3				
	Mean	**Min**	**Max**	**Comment**
Length of I. stage (min)	225	45	648	
Length of II. stage (min)	11.87	0	30	
Dist. SignalEnd to Birth (min)	2.70	0	29	over 10 min: 9
Noisy data W1 (%)	12.38	0	74	
Missing data W1 (%)	3.59	0	87	
Overall W1 (%)	15.98	0	89	over 50%: 18
Noisy data W2 (%)	13.42	0	49	
Missing data W2 (%)	0	0	0	
Overall W2 (%)	13.14	0	49	over 25%: 98
Noisy data II.stage (%)	22.62	0	91	
Missing data II. stage (%)	8.47	0	100	
Overall II. stage (%)	31.26	0	100	over 50%: 97

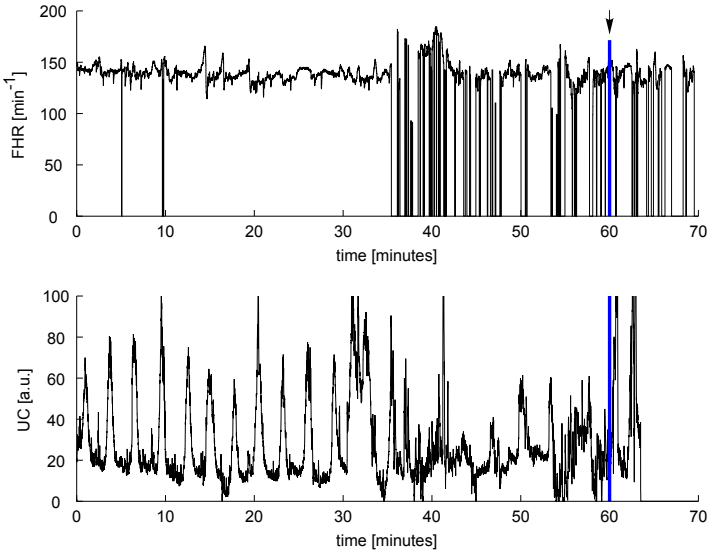


Figure 4.2: Record of fetal heart rate and uterine contractions. An example record from the database. The end of I. stage of labour is marked with blue line and arrow.

CTG database – deliveries by Caesarean Section

The database was selected to have the majority of intrapartum recordings with vaginal delivery. Nevertheless due to low number of cases with severely-abnormal outcomes, we have decided to add all recordings delivered by Caesarean Section (CS) with abnormal outcomes that conformed with the requirements mentioned above. Additional 30 CS recordings with normal outcome were randomly selected and added as control-group. This control should enable the user of the database to evaluate CS recordings separately, if necessary.

4.5 Conclusion

The CTU-UHB database is the first open-access database for research on intrapartum CTG signal processing and analysis. The database design was based on our proposed methodology that could be used for any future development of a database of similar nature. A database that could be used for data mining of knowledge, routine classification, or testing.

In the following paragraphs we will highlight the subjects, that could if unobserved, lead to problems with the use of the database.

The CTU-UHB users should be aware that there is a possible noise in the clinical data, since some information had to be mined from free text. Even though the whole data was carefully checked it is possible that some noise is still present. However, this noise should not significantly disrupt any results obtained. Also we note that, due to the selection process, the database is biased from normal population but this bias is evident in all other studies and, more importantly, if we would keep the database in the original form, the potential users would be forced to select the data themselves – resulting in different selection criteria and making, again, any comparison across studies infeasible.

From Table 3.2 it is evident that each study used different outcome measures, or their combinations. Again, this makes any comparison across studies infeasible. There are two main sources of evaluation: objective e.g. by umbilical artery pH, which is a prominent example, and subjective evaluation by experts according to their knowledge and/or guidelines used. The former is described in more detail in Section 2.3 and the latter is described and analysed in Chapter 6.

Among undocumented parameters in the database, which could influence the shape and/or different properties of FHR one could count e.g. smoking (Oncken et al., 2002), which can increase the heart rate or epidural analgesia (Cleary-Goldman et al., 2005; Hill et al., 2003) responsible for intermittent

fetal bradycardia due maternal intermittent hypotension. Some risk factors can influence the look of the FHR such as diabetes melitus, where FHR looks more immature (Tincello et al., 2001). Also technical parameters can influence the FHR itself – such as the size of a autocorrelation window for deriving FHR from ultrasound (Roj et al., 2008), or the derived parameters such as power spectral density of FHR, which can be affected by the type of interpolation (Cesarelli et al., 2011).

This question is usually limited by the availability of the data. Really long signals (spanning from the check-in to delivery) enable us to create an individualized approach to each fetus with regard to its starting point (Rosén et al., 2007). We have much more information to analyse, which can be positive (Graatsma et al., 2009) or confusing based on the point of view (Sisco et al., 2009). Short signals e.g. 70-min long (Schiermeier et al., 2008a) enable us to try to find direct relation between the features measured and the outcome.

Another question is how to treat the II. stage of labour. General opinion on the second stage is that it is different from the I. stage – in the shape of the signal. It is also very often noisy and it differs even in the clinical treatment where obstetricians are much more likely to apply operative delivery in unclear traces (Sheiner et al., 2001).

The CTU-UHB database is the first open-access database for research on intrapartum CTG signal processing and analysis. It is available at the physionet[2] (Goldberger et al., 2000). The database is reasonably large and allows researches to test and developed algorithms/methods for CTG analysis and classification. Using CTU-UHB database different approaches can be easily compared with one another in the objective fashion. Intuitively, the use of common database can stimulate research in CTG signal processing and classification.

[2] At the time of publication of this thesis the database was under review in the journal of BMC Pregnancy and Childbirth. For the purpose of review a portion of database was available at: http://bio.felk.cvut.cz/users/spilkaj/CTU_UHB_database. After the review the database will be available at http://physionet.org/.

Chapter 5

Signal processing and analysis

One of the most important aspects of signal processing is the quality of input data. Fetal heart rate can be distorted by variety of reasons (e.g. fetal or maternal movements, misplaced electrode etc.), leading to a corrupted or missing signal. Even though the missing intrapartum FHR is common (0-40% missing for ultrasound measurement (US) and 0-10% for the direct measurement (DECG) (Bakker et al., 2004)), there are no guidelines stating when a signal is unusable either for visual inspection or for automatic analysis. The usual empirical value given by clinicians is 50%. Even though the external monitoring using US has a lower signal to noise ratio than that recorded using DECG there is no clinical difference between these two approaches. However, it matters for automatic analysis as was shown in (Gonçalves et al., 2006b) and later in the similar paper (Gonçalves et al., 2013). In this chapter we describe the three preprocessing steps: artefacts rejection, interpolation, and detrend. The preprocessed signals were further described by features that originated from different fields. We divided the features based on the previous works into several groups: morphological, time-domain, frequency-domain, and nonlinear. The features described contain almost the complete set of features used for FHR analysis.

The traditional approach to CTG analysis is to study morphological changes of signal, i.e. baseline, variability, accelerations, and decelerations, which are used by obstetricians. These morphological features are defined by guidelines (FIGO, 1986) and are usually estimated visually. Another type of features are those that are either difficult to visually estimate or can not be estimated by the naked eye at all. This category includes short/long term variability and also variety of other features (frequency, entropy, complexity, and fractal dimension).

Chapter at a glance. First, we present the preprocessing steps and then we describe linear, frequency, and nonliner methods for fetal heart rate analysis. The chapter brings a comprehensive review of almost all features that were used for FHR analysis.

5.1 Signal preprocessing

Preprocessing is the main part in every signal processing task and is always the first step to be made. Values of extracted features and further classification are highly dependent on the preprocessing quality. For instance preprocessing steps could distorts the deterministic nature of the data and add some stochastic components making the use of nonlinear methods unsuitable. The ideal signals for analysis would be those measured directly in the heart. This is, however, not possible and signals are measured either externally using Doppler ultrasound or internally by a scalp electrode. As mentioned above, signals recorded externally have lower signal to noise ratio than those recorded internally but even internal records are contaminant with noise and artefacts. In our case the preprocessing consisted of the following steps: artefacts rejection, detrend, and interpolation.

Artefacts rejection The FHR signal contains a lot of artefacts caused by mother and fetal movements or displacements of the transducer. In general the amount of data being removed as artefacts or missing

values is in the range between 0% – 40% of all data. The algorithm suggested by (Bernardes et al., 1991) was used for artefact rejection. Any successive five beats with a difference lower than 10 bpm among them are considered as a stable segment. Then, whenever the difference between adjacent beats is higher than 25 bpm, the sample is substituted by linear interpolation between the previous beat and the new stable segment. Thus, all abrupt changes in FHR are removed and replaced. The result of artefacts rejection is presented in Figure 5.1b. Notice that the artefacts occur mostly at the end of labour.

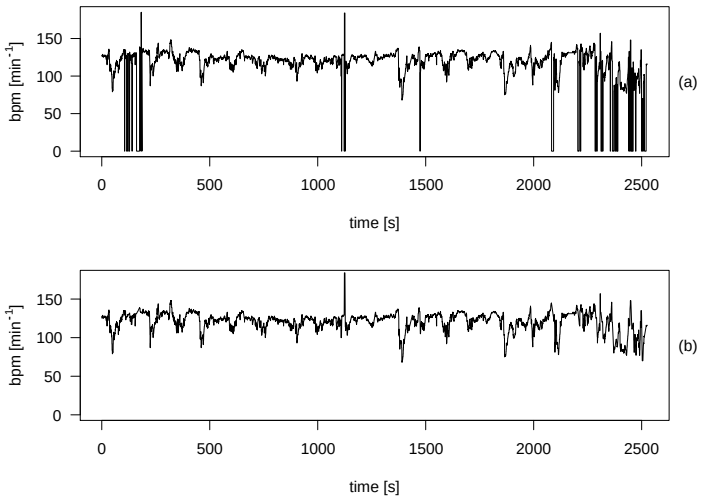


Figure 5.1: Rejection of artefacts. (a) the raw signal with artefacts, (b) signal after artefacts rejection.

Interpolation and gap treatment We used cubic Hermite spline interpolation (Kahaner et al., 1989), implemented in MATLAB®, to replace the missing data. We did not compute across a gap (Sprott, 2003) when the length of the missing data was 15 seconds or more – the value obtained based on our experiments. The spline interpolation also introduces nonlinearity, however, the amount of nonlinearity should be approximately the same for normal and abnormal FHR. Another approach, which was recently introduced, is to replace missing data using an adaptive method (Oikonomou et al., 2013) based on two steps: i) reconstruction step to obtain estimate of missing data using empirical dictionary, ii) construction of the dictionary using updated values from the first step. These two steps are repeated until convergence. This method show promising results (Oikonomou et al., 2013) and is currently verified at the CTU-UHB database.

Detrend Physiological time series are generally considered as nonstationary, i.e. statistical properties of physiological signal (mean, variance, and correlation structure) vary during time. We describe stationarity and nonstationarity in Section 5.3. For the frequency and nonlinear methods that require a signal to be stationary we carefully detrend signal using estimated baseline. The baseline estimation is described in more detail in Section 5.2.3.

5.2 Linear time series analysis

We examine oscillation in intervals between consecutive beats and also variations in difference of adjacent beats. For data analysis we use statistical methods in the time domain such as first and second order statistics (Task-Force, 1996). Another approach is to examine frequency spectrum by Fourier transform. A signal is decomposed to its single frequencies where each frequency is represented either by amplitude or power. Let $z(i)$ be a FHR signal for $i = 1, 2, \ldots, N$ where N is a length of FHR. The

$z(i)$ is expressed in beats per minute (BPM). Another, corresponding expression of $z(i)$ used in this work are known as RR series with time increments $T(i)$ (in seconds).

5.2.1 Time domain

The methods described in this section are mainly based on (Task-Force, 1996) and (Magenes et al., 2000) if not referenced. The time domain features representing the variation between consecutive R-R intervals are as follows:

- The mean heart rate: $\bar{T} = \frac{1}{N}\sum_{i=1}^{N} T(i)$ [ms]
- Standard deviation of the FHR: $\mathrm{SDNN} = \left(\frac{1}{N-1}\sum_{i=1}^{N}(T(i)-\bar{T})^2\right)^{1/2}$ [ms]

Short term variability (STV) There are two principal ways how to estimate the short term variability depending on signal acquisition technique. For the DECG the beat-to-beat variability approach is used, on the other hand, when CTG is acquired using Doppler ultrasound technique there is no real beat-to-beat variability because of intrinsic smoothing due to correlation based technique. Instead epoch-to-epoch variation is used when the FHR is averaged over short period of time. Mantel et al. (1990a) suggests 2.5 seconds for averaging while in the Sonicaid 8000 system the period of 3.75 seconds is used (Pardey et al., 2002). The STV is estimated for signal of length 60 sec.; for longer signals the 60 sec. estimations are averaged.

- Standard beat-to-beat variability $\mathrm{STV} = \frac{1}{N}\sum_{i=1}^{N-1}|T(i+1)-T(i)|$ [ms]
- De Haan (de Haan et al., 1971): $\text{STV-HAA} = \mathrm{IQR}(\arctan\left(\frac{T(i)}{T(i-1)}\right))$ [a.u.], where IQR is inter-quartile range with $i = 1,\ldots,N-1$.
- Yeh (Yeh et al., 1973): $\text{STV-YEH} = \sqrt{\sum_{i=1}^{N-1}\frac{(D(i)-\bar{T})^2}{N-2}}$ [ms], where $D(i) = 1000\cdot\frac{T(i)-T(i+1)}{T(i)+T(i+1)}$.
- Sonicaid 8000 STV (Pardey et al., 2002): $\text{Sonicaid} = \frac{1}{M}\sum_{t=1}^{M} R_t$ [ms], where M is number of minutes of FHR and R_t is difference between adjacent epochs in the particular minute: $R_t = \frac{1}{H-1}\sum_{j=1}^{H-1}|\bar{s}_j - \bar{s}_{j+1}|$, where H is number of subintervals in 60 sec ($H = 60/K$), K is number of samples in 3.75 seconds, $K = f_s \cdot 3.75$, and $\bar{s}_j$ is average value in 3.75 seconds for a subinterval $j = \{1,2,\ldots,H\}$.

Long term variability (LTV) Long term variability is computed over 60 seconds and there is no need of averaging the signal in 60 seconds as for the STV. For longer FHR than 60 sec. estimations of LTV are averaged over each 60 sec.

- The Delta value: $\Delta = \frac{1}{M}\sum_{i=1}^{M}\left[\max_{i\in M}(T(i)) - \min_{i\in M}(T(i))\right]$ [ms], where M is the number of minutes of a signal.
- Total value of the Delta: $\Delta_{total} = \max_{i\in[1,N]}(T(i)) - \min_{i\in[1,N]}(T(i))$ [ms].
- Long term irregularity: $\mathrm{LTI-HAA} = \mathrm{IQR}\sqrt{T^2(i)+T^2(i-1)}$ [a.u.], where IQR is inter-quartile range with $i = 1,\ldots,N-1$ (de Haan et al., 1971).

Note that the total value of Delta, Δ_{total}, corresponds to long term variability defined in the FIGO guidelines (FIGO, 1986).

5.2.2 Frequency domain

Signal decomposition into frequency components is a fundamental analysis technique. With this approach we lose the notion of time and only frequency components of signal are provided. The power as a function of frequency constitutes to what is known as power spectral density (PSD). The PSD could be estimated by various methods. One of them is Fourier transformation which considers signal as a composition of cosine waves with different amplitudes, phases, and frequencies.

We estimated the power spectral density (PSD) using fast Fourier transform (FFT). The PSD is usually divided into non-overlapping energy bands. These bands represent underlying physiological activity of either mother or fetus. The division of power spectrum into individual bands is not such straightforward as for adult heart rate variability and exact bands for fetal monitoring still remain unknown (Laar et al., 2008). Slightly different spectral bands were examined and described by (Sibony et al., 1994) and (Signorini et al., 2003). The former approach divides spectra into four bands: very low frequency VLF : 0 – 0.03 Hz, low frequency LF: 0.03 – 0.15 Hz that reflects sympathetic activity, mild frequency MF: 0.15 – 0.5 Hz, which is associated with fetal movement and maternal breathing, high frequency HF: 0.5 – 1 Hz that represents fetal breathing[1], and LF/(MF + HF) ratio that corresponds with balance of two autonomous systems. Other frequency bands were proposed by Sibony et.al. They partitioned spectra similarly as Signorini et.al., with the modification that number of bands was reduced into three and boundaries of bands changed : very low frequency VLF: 0 – 0.05 Hz, low frequency LF: 0.05 – 0.15 Hz, high frequency HF: 0.15 – 0.5 Hz, and LF/HF ratio. Implementation was provided by (Kaplan and Staffin, 1998).

We shall note here that power spectral density of fetal heart rate has power law scaling relationship. The energy as a function of frequency decreases in power low fashion $1/f^{\beta}$. The spectral index β is estimated as a slope of line fitted to the spectrum estimate, see Figure 5.2 for illustration. The β equals 0 for white noise, 1 for pink noise, and 2 for fractional Brownian motion (Eke et al., 2002). Note that spectral analysis performed on the whole record obscures detailed information about autonomic modulation of RR intervals (Furlan et al., 1990)

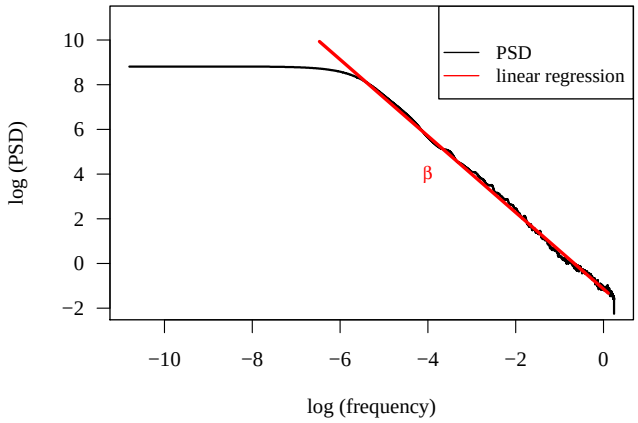


Figure 5.2: Spectrum of fetal heart rate. Estimated spectral index β as a slope of linear regression.

5.2.3 Morphological features

The following group of descriptive features is based on guidelines for CTG evaluation (FIGO, 1986). These features and patterns are used by clinicans for CTG assessment and were previously described in Section 2.2.1. The set of features is defined as follows:

- baseline – the mean level of fetal heart rate where acceleration and deceleration are absent

[1] Note that fetal lungs are non-functional and only movements are performed

- number of accelerations
- number of decelerations

Baseline is the most fundamental morphological feature. The improper baseline estimation destroys subsequent analysis of accelerations and decelerations. The developed algorithm for baseline estimation was based on kernel density estimate of FHR probability density function. It was inspired by (Georgieva et al., 2011) and compared to other algorithms (Jimenez et al., 2002; Pardey et al., 2002; Taylor et al., 2000) in the work of (Zach, 2013). The modified algorithm was further used in (Abry et al., 2013). The algorithm works on consecutive windows with 5 minutes overlap. In each window a probability density function is estimated using the kernel density of certain width h. The h is estimated for each window and depends on median and variance. The estimated baseline in the window corresponds to maximum of the density and is employed as an anchor point of baseline in this window. The anchor points from individual windows are then used for estimating the baseline for the whole signal. In order to determine the stability of signal two measures were introduced (Georgieva et al., 2011). The so called signal stability index (SSI) together with minimum expected value (MEV). The SSI corresponds to maximum of the density function and MEV to minimum value of this distribution. The lower SSI and MEV the less stable signal is. Clearly, when SSI and MEV are low the baseline is not estimated and accelerations and decelerations are not considered. The details on baseline estimation can be found in (Georgieva et al., 2011; Zach, 2013). An example of estimated baseline is shown in Figure 5.3.

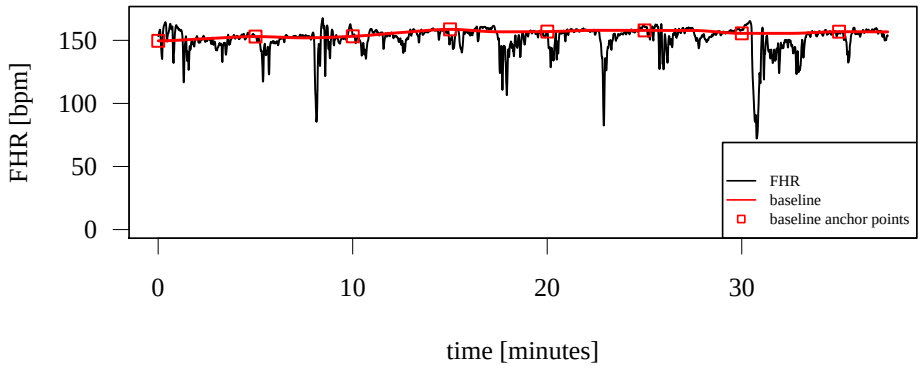


Figure 5.3: Estimated baseline of fetal heart rate.

5.3 Nonlinear time series analysis

The nonlinear approach may reveal relevant clinical information of FHR hidden to conventional time series analysis. Goldberger et al. (1985) observed that a human heart beat fluctuates on different time scales and is self-similar (self-affine), see Figure 5.4. Despite that there remains ongoing controversy over whether a normal heart rate is chaotic or not (Glass, 2009), tools used for examination of chaotic time series could also be useful for FHR analysis. There exist several approaches for nonlinear time series analysis; in this work fractal dimension, entropy, and complexity measures were utilized. When analysing FHR by nonlinear methods we have to be aware of at least two major pitfalls. First, FHR contains stochastic components induced by motion artefacts and measurement process. These distortions could severely damage the nature of FHR; therefore we used a surrogate data test to establish nonlinearity of FHR. Second, a certain data length is necessary to reliably estimate values of nonlinear methods. The required data length for each method is discussed in the corresponding sections below.

Stationarity and nonstationarity Time series are considered stationary when the statistical measures, i.e. mean, variance, and correlation structure, are the same irrespective of time. On the other

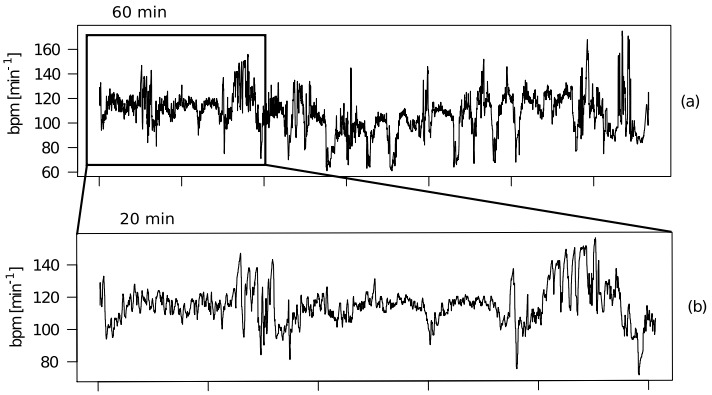


Figure 5.4: Self-affinity of fetal heart rate. Fluctuations of FHR at different time scales that are statistically self-similar (self-affine).

hand, the nonstationary time series do not possess this property and statistical measures fluctuate over time. According to the dichotomous model (Eke et al., 2002), signals are seen as realization of one of two temporal processes: fractional Brownian motion (fBm) and fractional Gaussian noise (fGn). The fBm signal is nonstationary with stationary increments. Physiological signals are generally consider as fBm, e.g. see Figure 5.4, where statistical properties of FHR varies over time. The fGn is considered as stationary. Since FHR is generally accepted as to be fBm methods able to overcome long-term statistical fluctuation should be applied or the trend making FHR nonstationary could be removed.

State space reconstruction There are two approaches to estimate dimension of a signal either by direct measurement of the waveform or by operating in reconstructed state space. The former approach considers a signal in $\mathbb{R}^2$ as a geometric object and directly uses it without any further transform. On the other hand the state space is reconstructed from coordinates representing the variables needed to specify the state of a dynamical system.

As time evolves, a system moves from one state to another creating a trajectory, which provides a geometrical interpretation of system dynamics. The trajectories that never intersect and touch each other are called strange attractors and are typical for chaotic systems. Packard et al. (1980) showed that it is possible to reconstruct state space from scalar time series and this reconstructed space is diffeomorphically[2] equivalent to the original state space. The state space can be reconstructed using Taken's embedding theorem (Takens, 1981). It states that it is possible to reconstruct state space from signal $z(t)$ delayed by time τ as long as the embedding dimension m is larger than $2d + 1$, where d is a box counting dimension, $z(t) \to z_{m,\tau}(t) = [z(t), z(t+\tau), \ldots, z(t+(m-1)\cdot\tau)]$. Different choice of τ and m leads to different reconstruction. Optimal embedding parameters cannot be established in general but are connected to specific application. The mutual information approach (Fraser and Swinney, 1986) is usually used to search the time delay and Cao's method (Cao, 1997) for examination of the embedding dimension. For other methods see any literature about nonlinear time series analysis, e.g. (Kantz and Schreiber, 2004). The state space reconstruction from FHR for normal and pathological fetus is shown in Figure 5.5; see Figure caption for details.

5.3.1 Fractal dimension

Box counting dimension

The box-counting dimension is based on evaluation of signal's capacity by covering it by N boxes of the side length ϵ. A minimal number of boxes needed to cover whole signal is counted and then the side length of boxes is decreased. Thus, repetitively descreasing size of boxes and counting them

[2] A diffeomorphism is a map between manifolds which is differentiable and has a differentiable inverse

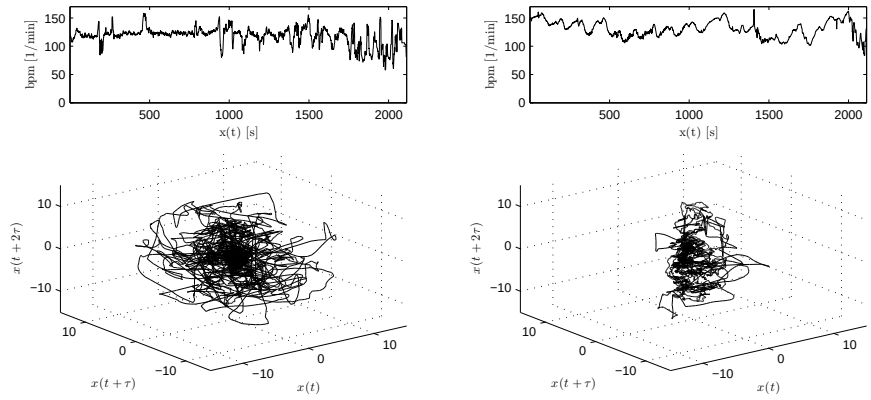


Figure 5.5: Fetal heart rate and state space reconstruction for normal and pathological fetus. The upper left and right signals represent fetal heart rate. The corresponding state space is shown below. (a) FHR and state space for normal fetus. The optimal delay time was $\tau = 2.5\ s$. (b) FHR and state space for pathological fetus. The optimal delay time was $\tau = 5\ s$. The low complexity of FHR for fetus with developed acidemia, (b), is clearly visible in both time and state space. In the state space the delayed coordinates of FHR span less area thus showing reduced variability.

we are able to estimate box counting dimension D_B as a slope of a linear regression fit to pairs on a log-log plot of $N(\epsilon)$ versus $1/\epsilon$

$$\begin{aligned} N(\epsilon) &= (1/\epsilon)^{D_B}, \\ D_B &= \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)}. \end{aligned}$$

Higuchi's dimension

The Higuchi method (Higuchi, 1988) calculates fractal dimension from estimated length of curve, i.e. fetal hear rate in our case. From the time series $z(1), z(2), \ldots, z(N)$ of length N a new time series Z_ϵ^s is constructed such that

$$\{Z_\epsilon^s\} = \{z(s), z(s+\epsilon), z(s+2\epsilon), \ldots, z(s+\lfloor (N-s)/\epsilon \rfloor \epsilon)\}, \quad s = 1, 2, \ldots, \epsilon,$$

where $\lfloor a \rfloor$ denotes the floor function that gives largest integer lower or equal to a, s defines the initial time, and ϵ the time interval. The ϵ represents time displacement and number of new created subsets is equal to ϵ. For example, for $\epsilon = 3$ and $N = 100$ we create following sequences

$$\{Z_3^1\} = \{z(1), z(4), z(7), \ldots, z(97), z(100)\}, \tag{5.1}$$
$$\{Z_3^2\} = \{z(2), z(5), z(8), \ldots, z(98)\}, \tag{5.2}$$
$$\{Z_3^3\} = \{z(3), z(6), z(9), \ldots, z(99)\}. \tag{5.3}$$

The length of curve Z_ϵ^s is defined as follows

$$L_s(\epsilon) = \left\{ \left(\sum_{i=1}^{\lfloor \frac{N-s}{\epsilon} \rfloor} |Z(s+i\epsilon) - Z(s+(i-1)\epsilon)| \right) \frac{N-1}{\lfloor \frac{N-s}{\epsilon} \rfloor \epsilon} \right\} \Big/ \epsilon,$$

where $(N-1)/\lfloor (N-s)/\epsilon \rfloor \epsilon$ represents the normalization factor for the curve length of subset time series. Then the length of curve for time interval ϵ, $\langle L(\epsilon) \rangle$, is defined as the average value over ϵ sets of $L_s(\epsilon)$

$$\langle L(\epsilon) \rangle = \frac{1}{\epsilon} \sum_{s=1}^{\epsilon} L_s(\epsilon).$$

The computed curve length $\langle L(\epsilon) \rangle$ for different ϵ is related to the fractal dimension D by exponential formula

$$\langle L(\epsilon) \rangle \propto \epsilon^{-D}.$$

The fractal dimension is estimated as a slope of fitted regression to log-log plot of $\langle L(\epsilon) \rangle$ versus ϵ. Note that Higuchi's method estimates the Hurst exponent H that is related to the fractal dimension $D = E + 1 - H$, where E stands for Euclidean dimension which is equal to one for time series.

Next, we estimated the two scaling regions as were described by (Higuchi, 1988). He suggested two scaling regions on the log-log plot of some measurement function, e.g. number of boxes, versus size of region, e.g size of box. These two regions are illustrated in Figure 5.6. Higuchi named the time where the curve bends as critical time τ_c. This time separates short D_s and long D_l scale waveform fractal dimension for $\leq \tau_c$ and $> \tau_c$, respectively. The region of the short scale reflects the short time variability while the longer scale represents the long time irregularity. To standardize estimated dimension we determined the τ_c for all methods. The τ_c was approximately same for all methods, $\tau_c \approx 3$ s. In addition, in order to estimate both regions by one parameter, we also fitted the log-log plot with a second order polynomial which coefficients (first order p_1 and second order p_2 polynomial coefficient) correspond to the both STV and LTV.

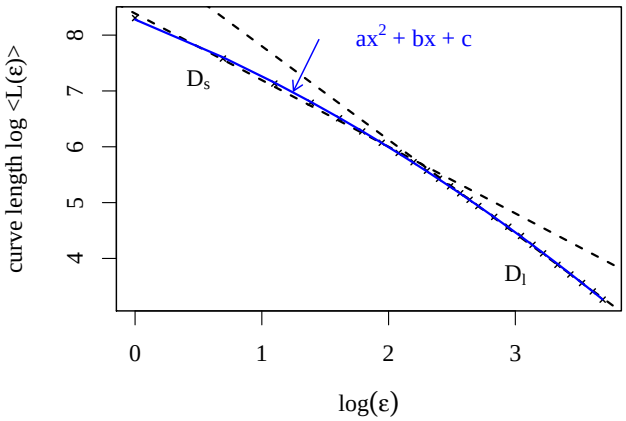


Figure 5.6: Short, D_s, and long, D_l, scale of waveform fractal dimension estimated by Higuchi's method. The curve breaks at $\log(\epsilon) = 2.5$. This equals to critical time $\tau_c = e^{\epsilon}/f_s \approx 3$ s, where sampling frequency is $f_s = 4$ Hz. The estimated second order polynomial is shown in blue colour.

Dimension of Variance

The variance technique of fractal dimension calculation is based on properties of fractional Brownian motion (fBm). It is a very useful approach because it is robust to noise. Let $z(t)$ be a signal continuous in time t and Δt a time increment. The variance σ^2 is then related to the Δt according to the power law (Kinsner, 1994)

$$Var\{\Delta z(t_n, \Delta t)\} = \langle z^2(t_n, \Delta t) \rangle \propto |\Delta t|^{2H},$$

where $\Delta z(t_n, \Delta t) = z(t_n + \Delta t) - z(t_n)$ and H is the Hurst exponent computed from a log-log plot using

$$H = \lim_{\Delta t \to 0} \frac{1}{2} \frac{\log Var\{\Delta z(t_n, \Delta t)\}}{\log(\Delta t)}.$$

Finally, the variance dimension is defined as: $D_\sigma = E + 1 - H$, where E is the Euclidean dimension which equals to one for time series. The variance dimension is robust to noise.

5.3.2 Detrend Fluctuations Analysis

The detrend fluctuation analysis (DFA) was proposed by (Peng et al., 1995) and probes the signal at different time scales. The result of the DFA is the fractal scaling exponent α. The whole process of estimating α is as follows. First, the time series $z(1), z(2), \ldots, z(N)$ is integrated giving

$$Y(i) = \sum_{i=1}^{N} [z(i) - \bar{z}],$$

where $Y(i)$ is the sum of i-th sample (cumulative sum) and $\bar{z}$ is the averaged value of the entire signal. Then $Y(i)$ is divided into windows Y_ϵ of equal length ϵ. For each window Y_ϵ a least square line $Y_{l\epsilon}$ representing the trend in the window is estimated. This line is subtracted from a summed $Y(i)$ (in the window Y_ϵ) in order to reduce possible non-stationarity. The formula for computation of fluctuations $F(\epsilon)$ in a window is following

$$F(\epsilon) = \sqrt{\frac{1}{\epsilon} \sum_{j=1}^{\epsilon} [Y_\epsilon(j) - Y_{l\epsilon}(j)]^2}.$$

This procedure is repeated for all time scale (different sizes of window ϵ). Then the $F(\epsilon)$ is plotted on log-log graph against all size of window ϵ. Typically, the relationship between $F(\epsilon)$ and ϵ is exponential $F(\epsilon) \sim \epsilon^\alpha$. This indicates the presence of self-similarity, i.e. for small windows size ϵ the fluctuations are similar to those for large ϵ.

The resulting scaling exponent α gives us information about origin of time series. For instance, $\alpha = 0$ indicates random process (white noise), $1/f$ pink noise has $\alpha = 1$, and $\alpha = 1.5$ indicates Brownian noise. Note the relation between α and spectral index $\beta = 2\alpha - 1$. Also note the relationship to the Hurst exponent $H = \alpha - 1$ (Eke et al., 2002).

Peng et al. (1995) suggested the minimal data length to be $N = 8200$ samples. For shorter time series, Govindan et al. (2007) provides a method to estimate DFA with help of generated phase randomized surrogates.

5.3.3 Entropy

Entropy describes behaviour of a system in terms of randomness and quantifies information about the underlying dynamics. Entropy is simply a fancy word for the "disorder". A stochastic, irregular, and less predictable signal has higher entropy than a completely deterministic. In other words, entropy is a measure of the amount of energy in a system that is unable to do work (Eckmann and Ruelle, 1985).

Approximate entropy

The approximate Entropy (ApEn) is able to distinguish low-dimensional deterministic system, chaotic system, stochastic, and mixed systems (Pincus, 1995). It has its roots in the work of (Grassberger and Procaccia, 1983) and (Eckmann and Ruelle, 1985). A time series z of length N is divided into a set of m-length vectors $u_m(i)$. Then the number of vectors $u_m(i)$ and $u_m(j)$, close to each other, in an Euclidean sense, $d[u_m(i), u_m(j)] \leq r$, is expressed by the number $n_i^m(r)$. This number is used to

calculate the probability of vectors being close according to $C_i^m(r) = n_i^m/(N-m+1)$. Let us define a function $\Phi^m(r) = 1/(N-m+1)[\sum_{i=1}^{N-m+1} \ln C_i^m(r)]$. Consequently the ApEn can be defined as

$$ApEn(m,r) = \lim_{N\to\infty}[\Phi^m(r) - \Phi^{m+1}(r)].$$

Sample entropy

A slightly modified estimation of approximate entropy was proposed by (Richman and Moorman, 2000) and resulted in what is known as sample entropy (SampEn). This estimation overcame the shortcomings of the ApEn mainly because the self-matches are excluded. Secondly, conditional probabilities are not estimated by a template-wise approach. SampEn requires only that one template finds a match of length $m+1$. The calculation of SampEn is as follows

$$SampEn(m,r) = \lim_{N\to\infty} -\ln \frac{C^{m+1}(r)}{C^m(r)}. \tag{5.4}$$

In the following Figure 5.7, we present simulated time series and the procedure for calculating sample entropy. We define the template length $m = 2$ and r as a positive value (usually $r = (0.1-0.2)\cdot SD$, where SD stands for standard deviation). The samples similar to the first sample $u[1]$ are marked by filled circle, to the second sample $u[2]$ by filled square, and to the third sample by filled triangle. Then we count occurrence of two-patterns and three-patterns. These are as follows: three two-patterns $(u[1],u[2];u[9],u[10];u[24],u[25])$ and two three-patterns $(u[1],u[2],u[3];u[9],u[10],u[11])$. Since we do not count self-matches, they are reduced to two and one, respectively. This is repeated for all two-patterns and three-patterns in sequence and then computed using formula (5.4).

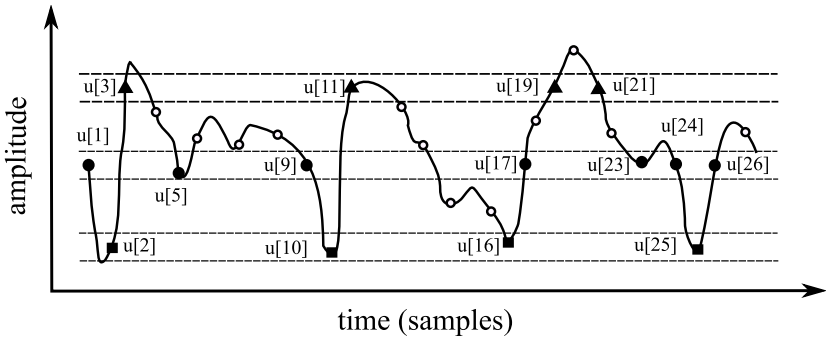


Figure 5.7: Simulated time series and its sample entropy estimation, for details see text. Modified from (Costa et al., 2005).

Data length Pincus (1995) showed that ApEn is broadly applicable for data series of length $N > 100$. Nevertheless, this was suggested for wide spectrum of applications. In our case, a meaningful data length for ApEn is $N \geq 1000$.

Parameters settings Parameters used for ApEn and SampEn estimation: tolerance $r = \{0.15; 0.2\}\cdot SD$ and the embedding dimension $m = \{2,3\}$ (Liu et al., 2011; Pincus and Viscarello, 1992).

Implementation The ApEn was implemented by (Kaplan and Staffin, 1998). Implementation of SampEn can be found at physionet web page (Goldberger et al., 2000). Note that for long time series a fast computations of entropies are available (Manis, 2008; Pan et al., 2011).

5.3.4 Lempel Ziv Complexity

The Lempel Ziv Complexity (LZC) (Lempel and Ziv, 1976) is widely used in data compression. It is based on information theory approach. The LZC estimates reoccurring patterns contained in the time series irrespective of time. A periodic signal has the same reoccurring patterns and low complexity while in random signal individual patterns are rarely repeated and signal complexity is high. To be more precise, (Lempel and Ziv, 1976) defined complexity as "a measure on the extent to which the given sequence resembles a random one".

To the time series $z(1), z(2), \ldots, z(N)$ the encoding procedure is applied in order to form sequences S of strings. For the binary encoding this sequence contains only {0,1}. The increase in signal value $z(i+1) > z(i)$ is encoded by 1 and decrease $z(i+1) \leq z(i)$ by 0. To indicate that substring of S starts at position i and ends at position j we write $S(i,j)$. The vocabulary of the sequence $v(S)$ contains all substring of S, e.g. for S = 101, $v(S) = \{1, 0, 10, 01, 101\}$. Let S and Q denotes two strings and SQ their concatenation. When the length of sequence is not specified a operator π is used to remove last string from concatenated SQ. The operator π comes as a postfix $SQ\pi$.

The whole procedure of computation complexity $c(N)$ is following: At the start the complexity $c(N)$ is set to 1, $S = s_1$, $Q = s_2$, $SQ = s_1, s_2$, $SQ\pi = s_1$, and the vocabulary $v(SQ\pi)$ is empty. For generalization purpose, let us assume that we moved in sequence to sample r. The $v(SQ\pi)$ is not empty and strings S and Q are the following $S = s_1, s_2, \ldots, s_r$, $Q = s_{r+1}$. If $Q \in SQ\pi$ then Q contains the substring of S and do not provide new information, therefore, the S remain unchanged and a new character s_{r+2} is add to Q. Again we check if $Q \in SQ\pi$ and if Q is not substring of $SQ\pi$ we increase $c(N)$ by one and concatenate S and Q, otherwise we continue in adding the new characters to Q until the end of the sequence is reached.

At the end the number of different strings is equal to $c(N)$. By convention, when the sequence reaches its last element, the $c(N)$ is increased by 1. It is apparent that $c(N)$ is dependent on the length of original sequence N. We use the normalization form to avoid this dependence on the number of data points (Lempel and Ziv, 1976). The normalized $c(N)$ is defined as

$$c(N) = \frac{c(N) \log_2 N}{N}.$$

Note that another coding scheme can be used in order to encode signal. The above described binary encoding can be extended to a ternary and even more quantizing encoding. However, as (Kaspar and Schuster, 1987) pointed out, the higher encoding should not be used in order to minimize the dependence of results on quantification criteria and normalization procedures. The required data length for binary encoded data is 1000 samples (Ferrario et al., 2004).

5.3.5 Poincaré plot

The Poincaré plot, also known as a return map, is useful for visualization and analysis of FHR series. The FHR signal is embedded in dimension m = 2 with time delay τ = 1; on the x-axis is plotted $z(i)$ with respect to $z(i+1)$ on the y-axis. The Poincaré plot is analysed using fitted ellipses. Two measures are estimated SD_1 as the standard deviation of points perpendicular to the line $y = x$ and SD_2 as the standard deviation of points along the $y = x$ line (Brennan et al., 2001).

5.4 Table of all features

In Table 5.1 we present overview of all extracted features and their settings parameters. In total we worked with 21 features; different settings parameters yielded 49 features.

Feature groups The features were divided based on their "origin". The division into different groups followed our previous work (Spilka et al., 2012) and is also consistent with work of others,

Table 5.1: Table of all extracted features and their parameters.

Feature set	Features	parameters
FIGO-based	baseline	mean, standard deviation
	number of accel. and decel., Δ_{total}	
Statistical	STV, STV-HAA, STV-YEH, Sonicaid, SDNN, Δ, LTI-HAA	
Frequency	energy03	LF, MF, HF, LF/HF
	energy04	VLF, LF, MF, HF, LF/(MF+HF)
Fractal dim.	FD_Variance, FD_BoxCount, FD_Higuchi, DFA, FD_Sevcik	D, D_s, D_l, p_1, p_2
Entropy	ApEn, SampEn	$r = \{0.15, 0.2\}$, $m = 2$
Complexity	LZC	
other	Poincaré	SD_1, SD_2

e.g. (Georgoulas et al., 2006; Magenes et al., 2000). In the *FIGO-group* we included the morphological features (baseline, number of accelerations and decelerations) and long term variability termed Δ_{total}. The short term variability features are not included in the FIGO group since these are not considered in the FIGO guidelines because they can not be estimated visually (FIGO, 1986). From the remaining features we created two groups: *HRV-based* (statistical and frequency features inspired by adult HRV analysis), and *nonlinear* (fractal dimension, entropy, complexity, and Poincaré plot).

5.5 Surrogate data test

So far we assumed that fetal heart rate is nonlinear driven by deterministic chaos. In order to verify this hypothesis we used surrogate data test. In this test we formulated a null hypothesis e.g. that data are generated by gaussian linear stochastic process. Then, if this hypothesis is rejected on some significance level, we can conclude that data do not origin from such process and nonlinear methods may reveal important information about underlying system dynamics. All nonlinear methods were used as a discriminator between original data and its surrogates; the null hypothesis for rejected for all methods on significance level $p < 0.05$.

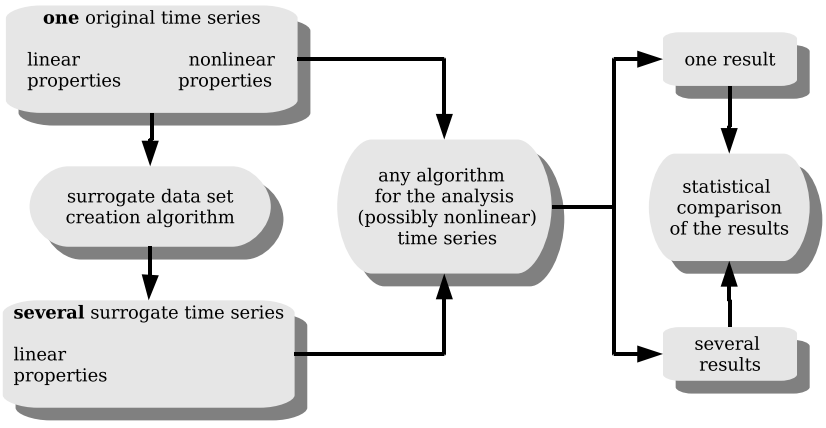


Figure 5.8: Scheme of surrogate date test for the case of the null hypothesis of a linear process. Modified from (Galka, 2000).

There are many available null hypothesis against which we can test our time series. For example, null hypothesis could be that data are independent, identically distributed random variables of unspeci-

fied mean μ and variance σ^2. In our work, we employed general hypothesis that data are produced by gaussian linear stochastic process ($AR(p)$ process). During data generation we required that surrogate and original data have the same power spectrum and probability density function. There exists broad area of methods one can use for data generation. The main idea behind creation of surrogate data is following: $i)$ apply Fourier transform to original time series, $ii)$ replace the phases by random numbers ranging from $(-\pi, \pi)$, $iii)$ apply inverse Fourier transform to the Fourier coefficients. In our work, we used iteratively refined surrogates proposed by (Schreiber and Schmitz, 1996). For more information see referenced paper or book of (Kantz and Schreiber, 2004). The level of significance is commonly set to be $p \leq 0.05$, therefore we need at least 19 or 39 surrogate data for one- and two-sided test, respectively. The whole scheme of surrogate data test is presented in Figure 5.8.

Chapter 6

Analysis of clinical evaluation

Interpretation of CTG recordings is an integral part of every day clinical practice though, since the introduction of CTG, it has been a subject of many controversies. Not only because of difficulty to interpret individual patterns of CTG but because of CTG utility in general (Sartwelle, 2012). Also, and more importantly, the high intra and inter-observer variability still persists (Beaulieu et al., 1982; Vayssiere et al., 2009).

In this chapter we propose a new approach for the annotation of CTG records and implement a new software for collecting these annotations – the CTGAnnotator. We offer a detailed insight into clinical evaluation of CTG. We analysed CTG evaluation obtained from nine clinicians where each clinician evaluated 634 CTG records. Our study performed on unique, open access, CTU-UHB database is the largest study ever performed when both the number of clinicians and number of records are considered. We provided comprehensive analysis of observer agreement and, in contrast to other works, we did not restrict the analysis to simple quantitative measures such as proportion of agreement and kappa coefficient but we also used simple visualizations in order to provide a clear picture of clinicians agreement/disagreement. This chapter is in part based on paper (Spilka et al., 2013a).

We proposed, implemented, and tested a novel approach for analysis of the clinical evaluation of CTG – the latent class model (LCM) of CTG evaluation. We use this model to estimate the hidden true class of CTG evaluation. This method provide superior results to the majority voting and, in addition, it enables us to resolve the ongoing controversy on how many classes should be used for CTG evaluation. The LCM offers deeper insight into clinical decision making and provides weights of individual clinicians. We show results of sensitivity and specificity for biochemical markers and Apgar score. Our study is the first study that provides sensitivity and specificity regarding the clinical evaluation based on FIGO guidelines. Finally, we statistically compare the clinical evaluation to the fetal heart rate features extracted in the same time window.

6.1 Clinical evaluation

Since the very introduction of the CTG into clinical practice its merit was widely disputed. The method was introduced without proper clinical trials (MacDonald et al., 1985) and its evaluation suffers from large inter-observer disagreement (Beaulieu et al., 1982; Bernardes et al., 1997; Lotgering et al., 1982; Vayssiere et al., 2009) among others. Even though guidelines (e.g. the most prominent FIGO guidelines (FIGO, 1986)) were introduced to tackle the heterogeneity of the CTG evaluation, high inter- and intra-observer variability is reported frequently even today (Blackwell et al., 2011). According to (de Campos et al., 2010), guidelines are in general too complex, with many parameters that are hardly possible to assess precisely in the clinical environment.

Large body of literature exists where clinicians try to look for alternative approaches to the current evaluation of the CTG according to FIGO (and from FIGO derived) guidelines. Since year 2000 the ST-analysis (STAN) (Rosén and Lindecrantz, 1989; Rosén et al., 2004) has spread worldwide. Although the most studies show that ST-analysis is performing better than the CTG alone (Amer-Wåhlin and

Maršál, 2011), it is important to keep in mind that the necessary first step to correctly interpret the ST ratio in ST-analysis is to correctly evaluate the CTG itself. Incorrect use of STAN combined with poor CTG interpretation can have disastrous effects (Westerhuis et al., 2007a).

Secondly, tweaks to the FIGO guidelines were proposed extensively (ACOG, 2009; Macones et al., 2008; NICE, 2007; RCOG, 2001) but no general agreement on the guidelines exists (de Campos et al., 2010). Some more complex guidelines were proposed by (Parer and Hamilton, 2010) and even though they claim superiority over classical guidelines in the inter-observer agreement (Coletta et al., 2012) clinicians remain conservative. None of the major guidelines changes were studied in larger group exceeding couple of interested hospitals.

CTG Annotator

We developed a software, the *CTGAnnotator* (Zach et al., 2013), which has been used to obtain annotation of the CTG recordings from nine obstetricians working on delivery wards of six Obstetrics and Gynaecology Clinics of all the medical schools in the Czech Republic. All clinicians have been currently practising delivery ward doctors with median experience of 15 years (minimum 10, maximum 33).

Annotation has been acquired using stand-alone platform independent application. The application has adopted the most commonly used display layout of CTG machines (in European format – 1 min./cm and 30 bpm/cm), and therefore poses no difficulty for obstetricians to adjust. CTG annotator GUI is shown in Figure 6.1.

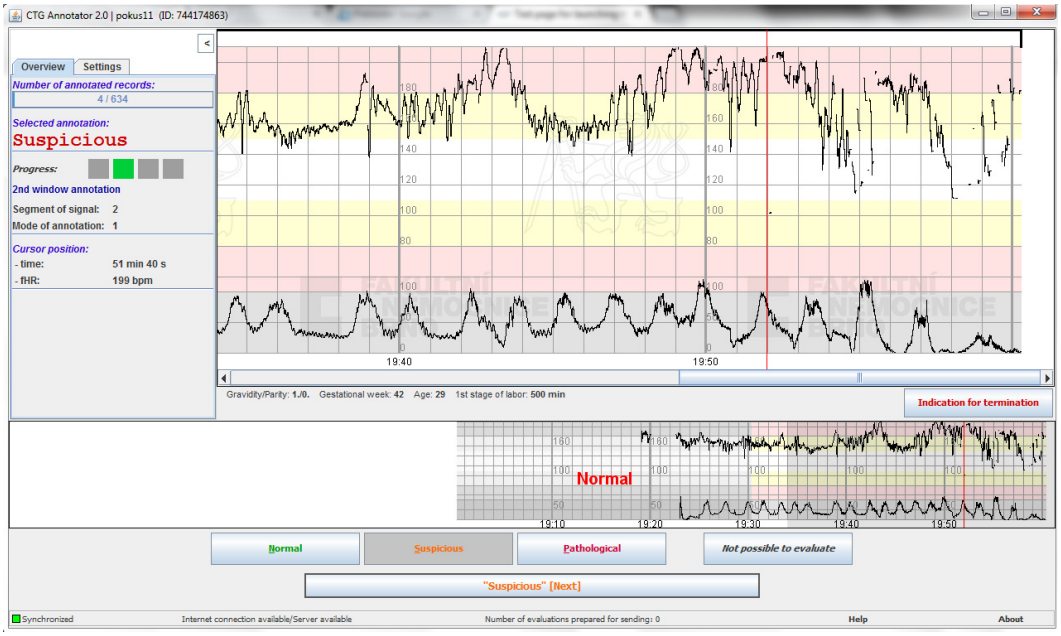


Figure 6.1: Cardiotocographic recording (CTG) and an example screen of the CTGAnnotator software that was used for CTG annotation.

The initial and final run of the application needs an internet connection, otherwise it is able to run in an off-line mode. The application is able, when connected to server, synchronize the data after each evaluated recording.

6.1.1 Annotation methodology

Simple introduction to the application was provided individually to each expert at their workplace. The introduction included running through a test mode of the application for the expert to get acquainted with the application interface. The test was run on specially selected CTG recordings that were not used

later for final set evaluation. Even though we expect that all experts adhered to the FIGO guidelines criteria (as required for the clinical decision making by the official Czech Obstetrics body) we did not provide any special training nor did we encouraged it. Our goal was to get as close to the *real* clinical evaluation as possible outside the delivery ward.

Based on the data structure in our database each CTG recording was presented for annotation in four steps, see also Figure 6.2:

1. 30-minutes long window with beginning of the CTG signal at maximum one hour before the end of the first stage of labour (Step/Window 1).

2. 30-minutes long window with beginning of the CTG signal at maximum 30 minutes before the end of the first stage of labour (Step/Window 2).

3. Full second stage of labour signal which was presented for evaluation only when more than 5 minutes of CTG signal was available (Step/Window 3).

4. Evaluation of labour outcome – prediction of umbilical artery biochemical parameters after delivery (in general training pH value was suggested) (Step 4).

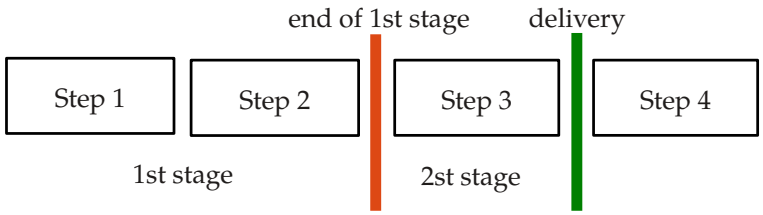


Figure 6.2: Annotation work-flow. Two 30 minutes windows were evaluated in the first stage of labour and one window in the seconds stage (if the CTG signal was longer then 5 min.). The labour outcome was evaluated in the 4-th step.

In steps 1 to 3 the experts evaluated CTG recordings as normal, suspicious, pathological, or uninterpretable according to their daily practice. Clinicians were provided with general clinical information about mother's age, gestational age, gravidity, parity, and total length of the first stage. The clinicians were made aware à priori of the general outlines of the database as described in the previous section and the way the records were presented to them.

In the 4-th step the clinicians stated their prediction of the delivery outcome as measured on umbilical cord artery – divided into four tiers as no hypoxia (normal), mild hypoxia (abnormal), severe hypoxia (pathological), or undecidable. During introduction these classes were described also in terms of arterial pH values (severe hypoxia: $pH \leq 7.05$, mild hypoxia $7.05 > pH \leq 7.15$, no hypoxia $pH > 7.15$). For this last step additional clinical data, presence of risk factors, were provided. Clinicians' evaluation of FHR and time needed for decision as well as changes in their decisions were recorded for each step of annotation.

Number of hidden features embedded in the CTG Annotator enabled us to acquire the annotation in the best controlled way possible. As a precaution and to limit the speedy clinicians that might have had lost concentration, CTG Annotator application closed automatically when a number of records annotated in a row exceeded 150. It was enabled to reopen again the following day. The order of records was pre-set randomly, prior to the experiment, and was exactly the same for all clinicians. In order to establish intra-observer agreement some of the records were randomly selected and presented more than once for annotation. Additionally to acquire intra-observer agreement with respect to particular class some of the data were selected for repeated annotation based on the previous annotation by the expert – this was done automatically by the application. The additional occurrences of the same

record (repeated ones) were ordered automatically to appear for the next time at the largest distance from the first occurrence as possible. With repeated records each clinician evaluated 634 recordings.

We have examined the percentage of recordings to normal, suspicious, pathological, and non-interpretable CTGs categories as evaluated by each clinician. We have also examined time needed to arrive at decision for each record also with respect to category chosen by an expert. Additionally we have collected predictions of the experts on the adverse outcome of the newborn and compared them to objective value of the pH or BDecf, based solely on the CTG recording.

6.2 Observer agreement measures

The assessment of agreement between observers is not an easy task. Among statisticians there is no agreement how the observer agreement should be measured. The kappa coefficient and its derivatives has been used to measure agreement in the past but it has been shown that the kappa is influenced by prevalence and base rate and is not suitable for comparison across different studies (populations). There is no single measure of agreement that could outperform the others; hence, the general advice is to use more measures until the proper will be available. For details refer to the great overview of statistical methods of rater agreement (Uebersax, 2010). In our work we used the proportion of agreement (PA) and for the sake of completeness also Fleiss kappa coefficient (Fleiss et al., 2004). We computed overall PA as well as PA with respect to different categories. In addition, we also aimed to visualize the inter/intra-observer agreement in a simple way in order to offer a clear and simple picture of agreement rather than to use a quantitative measure.

Proportion of agreement The proportion of agreement is simply probability that clinicians agree on evaluation. The generalized formula for proportion of agreement holds for multiple annotators with multiple classes. We follow the description used in (Uebersax, 2010). Let N is the number of annotated observations $i = 1, \ldots, N$ and C is the number of classes $c = 1, \ldots, C$. The number of annotations performed on observation i is defined as n_i and number of times observation i is annotated using class c is defined as n_{ci}. For example when $C = 2$ the i-th observation can be annotated as $1, 1, 2, 2, 2$. Then $n_{1i} = 2, n_{2i} = 3$, and $n_i = 5$. The summation across different c leads to a total number of annotations on i-th observation

$$n_i = \sum_{c=1}^{C} n_{ci}. \tag{6.1}$$

Intuitively, the n_i equals number of annotators (if each annotator annotates each observation only once). The number of all possible annotator-annotator pairs on class c for observation i is given by $n_{ci}(n_{ci} - 1)$. The number of agreements on class c across all observations is

$$S(c) = \sum_{i=1}^{N} n_{ci}(n_{ci} - 1). \tag{6.2}$$

Further, given n_{ci} (the number of annotations on class c for i-th observation) the number of possible annotator-annotator pairs in agreement can be computed as: $n_{ci}(n_i - 1)$. Note that the summation of $n_{ci}(n_i - 1)$ across all classes $(\sum_{c=1}^{C} n_{ci})(n_i - 1)$ is equal to $n_i(n_i - 1)$. Next, the sum across all observations is termed as the total number of *possible* agreements on class c

$$S_{poss}(c) = \sum_{i=1}^{N} n_{ci}(n_i - 1).$$

Finally the proportion of agreement specific to particular class c is equal to the total number of agreements on c divided by the total number of possible agreements on c

$$p_s(c) = \frac{S(c)}{S_{poss}(c)}.$$

The overall proportion of agreement, irrespective of category c, is computed in the similar way. The summation of (6.2) across all categories

$$O = \sum_{c=1}^{C} S(c)$$

and the total number of possible agreements

$$O_{poss} = \sum_{c=1}^{C} S_{poss}(c).$$

The overall proportion of agreement is equal to

$$p_o = \frac{O}{O_{poss}}.$$

Confidence intervals The confidence intervals for the proportion of agreement were estimated using bias-corrected and accelerated bootstrap method (Efron, 1994, 2003).

Kappa coefficient The kappa coefficient (Cohen, 1960) is widely used measure of observer agreement though there is a great disagreement on its appropriateness. The kappa coefficient is considered as chance corrected agreement, i.e. in its computation it corrects the agreement expected by chance. Let p_{obs} be observed probability of agreement and $\hat{p_e}$ be probability of agreement obtained by a chance (simply guessing the right class). The kappa coefficient is defined as

$$\kappa = \frac{p_{obs} - \hat{p_e}}{1 - \hat{p_e}}$$

with standard errors

$$SE(\kappa) = \sqrt{\frac{p_{obs}(p_{obs} - \hat{p_e})}{N_p(1 - \hat{p_e})^2}},$$

where N_p is number of pairs of ratings. There is large quantity of papers dealing with appropriateness/inappropriateness of kappa coefficient. The most serious disadvantages are: i) dependence on observed marginal proportions making comparison across different population infeasible, ii) lack of natural extension for multiple rates and multinomial classes (Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990), for details refer to great discussion on kappa coefficient (Uebersax, 2010).

6.3 Majority voting

The majority voting is the simplest voting mechanism to aggregate evaluation from multiple clinicians. Let y_i^j be evaluation of i-th observation $i = 1, \ldots, N$ for j-th annotator, $j = 1, \ldots .J$. The probability that i-th observation is assigned to the c-th class is

$$\mu_{ic} = (1/J) \sum_{j=1}^{J} \delta(y_i^j, c),$$

where $\delta(y_i^j, c)$ is indicator function that equals 1 when $y_i^j = c$ and 0 otherwise. The majority voting, or more precisely plurality voting, is simply choosing a class c for maximum of μ_{ic}. In the case of ties a flip of fair coin is performed.

6.3.1 Problems with majority voting

For its simplicity the majority voting is usually preferred. However, there are some limitations when using majority voting of clinicians, the summary of drawbacks is listed below. The summary does not only highlights the disadvantages of majority voting but touch the problems with clinical evaluation in general.

1. There is high inter and intra-observer variability in clinical evaluation, which has been widely reported, see (Blackwell et al., 2011; Blix et al., 2003; Lotgering et al., 1982; Vayssiere et al., 2009) among others.

2. Each clinicians has different expertise not only based on length of his/her career (experienced vs. inexperienced) but also influenced by labour management at working place. For example, a clinician who is called only to the most serious cases could loose, to some extent, knowledge on normal cases.

3. Clinicians could loose concentration/motivation or be simply distracted/inattentive during annotation.

4. A wrong class could be entered by an accident.

5. The annotation was performed in an artificial settings (different from reality/practice).

6.3.2 Condorcet's jury theorem

The different schemes of voting were thoroughly studied in social sciences. For the completeness we describe the famous Condorcet's jury theorem (1786), details can be found in (Boland, 1989), which states: if voters are right with probability $p > 1/2$, then majority vote is likely to be right than wrong and the probability of being right tends to 1 when number of voters goes to infinity. This theorem holds for dichotomous voting (e.g. normal/pathological). When there is c categories for voting the so called Condorcet's paradox applies, however (List and Goodin, 2001) provides the contrary. For details we refer the interested readers to the referenced article. It is intuitive that with increasing number of voters the likelihood of correct decision increase. Below we performed an experiment with clinical evaluation where we examined the stability of majority voting with respect to number of clinicians.

6.3.3 Stability of majority voting

The examination of stability of majority voting can be motivated in the following way. Let consider that we have majority votes of J clinicians. We would like to know if the created majority was obtained simply by a chance or if the majority is stable and possible variability in clinicians was cancelled out by using high number of clinicians. We summarize the definition of stability in Proposition 1.

Proposition 1 *We consider a majority vote of J clinicians stable if a majority voting of $J+1$ clinicians is not different (measured by proportion of agreement).*

In the proposition the term "different", our criterion, is not rigours thus offering space for possible misinterpretation. This vague term should be replaced by proper evaluation, i.e. statistical testing. However the created majority votes are not independent hence making any statistical comparison impossible.

We performed a simple experiment. We computed majority votes (MV) for all combinations of clinicians $\binom{J}{k}$, where $k = 3, \ldots, J-1$. Then we compared this majority with majority vote of all clinicians, $J = 9$. The procedure is shown in Algorithm 1.

Algorithm 1: Procedure for comparing different majority votes.

Input: $K = \{3, \ldots, 8\}$ number of clinicians, $\mathbf{Y}$ clinical evaluation of size $N \times J$, mv_J majority vote of all J clinicians, mv_b majority vote of combination b of clinicians

Result: pa - proportion of agreement

```
begin
    for j ∈ K do
        comb ← (J choose j) – all combinations of j clinicians from J
        for b ∈ comb – for all combinations do
            Y_b = Y(:, b) – get evaluation for selected combination of clinicians
            mv_b ← majorityVoting(Y_b)
            pa(j, b) ← proportionOfAgreement(mv_b, mv_J) – compare majority voting
        end
    end
end
```

6.4 Latent class analysis of clinical evaluation

6.4.1 A model of fetal heart rate evaluation

We described several outcome measures in Section 2.3 that are used to evaluate fetal well-being either during delivery (clinical evaluation of CTG) or after baby is born (biochemical markers and Apgar score). These measures could be divided into two subgroups based on their nature: subjective (clinical evaluation of CTG and Apgar score) and objective (pH, BE, and BDecf). Regarding the both groups there exist wide controversies and none of the measure is superior to the others. Because of their nature they can not be used interchangeably but are not complementary either. In the first group the pH value is the most common measure. However; it was shown that intrapartum metabolic acidosis only slightly corresponds to adverse fetal outcomes (Yeh et al., 2012). In the second group the subjective evaluation is too subjective and large intra- and inter-observer variability were reported in several studies, e.g. (Bernardes et al., 1997; Blix et al., 2003; Vayssiere et al., 2009) among others. While the difficulties with objective evaluation could not be diminished but only account for, the subjective measure (CTG evaluation) and its high intra/inter observer variability could be reduced.

We follow the works of (Dawid and Skene, 1979; Raykar et al., 2010; Smyth et al., 1995) and use similar notation to the most recent one (Raykar and Yu, 2012). To keep the description general we refer clinicians as annotators and assigned clinical evaluation as a label. For simplicity we begin the description with a simple model using random variables. Let us define the discreate random variables Ψ representing fetal status (fetal well-being), S as a pattern of CTG (providing, to some extent, information about fetal status), and Y^j as a label assigned by an annotator j, $j \in J$. We are interested in the conditional probability of fetal status given the clinical annotation $p(fetal\_status|clinical\_annotation) = p(\Psi|Y^j)$. This leads to a casual model (6.3), where fetal status is reflected by a CTG pattern which in turn is mapped to an annotation

$$\Psi \longrightarrow S \longrightarrow Y^j. \tag{6.3}$$

The conditional probability $p(\Psi|Y^j)$ could be expressed as

$$p(\Psi|Y^j) = p(\Psi|S, Y^j)p(S|Y^j). \tag{6.4}$$

Since Ψ is conditionally independent of Y^j given S the above equation can be rewritten as

$$p(\Psi|Y^j) = p(\Psi|S)p(S|Y^j).$$

We reformulate the casual model by assuming that Ψ is reflected by an unknown (unobservable) true category Y. We can imagine this as mapping of S into a category Y. This category could correspond to a different classification scheme (e.g. FIGO, NICHD). A clinical annotation Y^j is an estimate of this category Y based on observation of CTG pattern S

$$\Psi \longrightarrow S \longrightarrow Y \longrightarrow Y^j. \tag{6.5}$$

Without any loss of information it is convenient to omit S because it is reflected by Y and does not provide any other information. In fact Y could be viewed as quantization of S. This could seem as a strong assumption but without it the clinical annotation would be meaningless. The casual model is simplified to

$$\Psi \longrightarrow Y \longrightarrow Y^j.$$

To rephrase above relationship in words. The fetal status corresponds to unknown ground truth (a category), which is based on CTG pattern. Then a clinical annotation is only estimate of this ground truth, hence fetal well-being. This leads to the following equation for posterior probability

$$p(fetal\_status|clinical\_annotation) = p(\Psi|Y^j) = p(\Psi|Y)p(Y|Y^j).$$

However, in real world scenario the posterior probability $p(\Psi|Y)$ is not feasible to estimate or could not be estimated at all. In the work of (Smyth et al., 1995) it was subjectively estimated by scientist but this approach is inappropriate in our setting because it would be too subjective. Because we are not able to estimate posterior $p(\Psi|Y)$, we are rather interested in the posterior probability

$$p(Y|Y^j) = \frac{p(Y)p(Y^j|Y)}{p(Y^j)}$$

There are two catches however. First, Y is unknown (unobservable) truth value and, second, the Y^j suffers by high inter- and intra-observer variability. To overcome these difficulties (Dawid and Skene, 1979) proposed method to estimate Y from annotations Y^j, which will be introduced below. We first present a general mixture model and Expectation Maximization (EM) algorithm to find a maximum of likelihood function. Then we introduce the special cases of the mixture model: the binomial and multinomial models that are used to estimate Y from Y^j.

6.4.2 Finite mixture models

The finite mixture models are used for unsupervised learning (but are not limited to) when class labels are not available or when we do not restrict ourselves to particular class labels. Mixture models are well-studied statistical inference technique (McLachlan and Peel, 2000) that are able to represent arbitrarily complex probability mass functions. Finite mixture models has fixed number of parameters and a standard method to estimate these parameters is expectation maximization (EM) algorithm (Dempster et al., 1977) or, alternatively, one can use Markov Chain Monte Carlo (MCMC) technique (Diebolt and Robert, 1994).

Let $\mathbf{X} = [X_1, \ldots, X_d]$ be a d-dimensional random discreate variable, where $\mathbf{x}_i \in \mathcal{X}$ is d-dimensional vector and N is a number of observations. Further let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ be an input data set. It is assumed that the data is from a mixture of initially specified M components in some unknown proportions $\pi_1, \ldots, \pi_M$ (mixing probabilities). That is, each data point is a realization of the mixture probability mass function

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_m p(\mathbf{x}_i|\boldsymbol{\theta}_m), \tag{6.6}$$

where $\boldsymbol{\theta}$ corresponds to unknown mixing proportion π_m and the elements of $\boldsymbol{\theta}_m$. The parameters to be estimated are $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M, \pi_1, \ldots . \pi_{M-1}\}$, which are needed to specify the mixture. For the $\boldsymbol{\pi}$ the following must hold

$$\pi_m \geq 0, \quad m = 1, \ldots, M, \quad \text{and} \sum_{m=1}^{M} \pi_m = 1.$$

To estimate parameters of a distribution the maximum likelihood is commonly used (but other methods could be used as well). The likelihood of $p(\mathcal{D}|\boldsymbol{\theta})$ gives probability of $\mathcal{D}$ under the parameters $\boldsymbol{\theta}$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}).$$

Taking the logarithm of above equation and plugging it into the mixture probability mass function from equation (6.6) we compute the log-likelihood of mixture model as

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \log \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \sum_{m=1}^{M} \pi_m p(\mathbf{x}_i|\boldsymbol{\theta}_m).$$

The maximum likelihood estimate can not be found analytically and EM algorithm is powerful iterative technique for maximizing the likelihood

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \{\log p(\mathcal{D}|\boldsymbol{\theta})\}.$$

The Expectation Maximization algorithm

The expectation maximization algorithm (Dempster et al., 1977) is an iterative procedure to maximize the likelihood when some variables (parameters) are unobserved. In other words, the EM algorithm aims to find $\boldsymbol{\theta}$ that maximize the $\log p(\mathcal{D}|\boldsymbol{\theta})$ given observed data $\mathcal{D}$. The algorithm starts with initial guess of parameters and than repeats two steps: i) expectation (E-step) and ii) maximization (M-step) until some convergence criteria is met or until a predefined number of iteration is reached.

The EM is based on the interpretation of incomplete data. We consider $\mathcal{D}$ as being incomplete because of missing labels $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$, where N is number of observations and $\mathbf{y}_i = \{y_1, \ldots, y_M\}$ is a binary vector of labels. A label y_m is defined to be one if a sample $\mathbf{x}_m$ was produced by the m-th component. If we would know the labels $\mathcal{Y}$ the complete log likelihood would be

$$\log p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{m=1}^{M} y_m \log \left[\pi_m p(\mathbf{x}_i|\boldsymbol{\theta}_m)\right].$$

Let $t = 0, 1, \ldots, T$ be iterations of the EM algorithm, we introduce a Q-function such as

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^t) = E\left[\log p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})|\mathcal{X}, \hat{\boldsymbol{\theta}}^t\right]. \tag{6.7}$$

This equation is the central part of expectation maximization algorithm; interpretation is as follows (Duda et al., 2000): the parameter $\hat{\boldsymbol{\theta}}^t$ is current (best) estimate for the complete data, the $\boldsymbol{\theta}$ is a candidate improved estimate, $\boldsymbol{\theta} \in \Theta$. The interpretation of equation (6.7) is that given a candidate $\boldsymbol{\theta}$ the right hand side computes the likelihood of data including the unknown labels $\mathcal{Y}$ marginalized with respect to the current best distribution described by $\hat{\boldsymbol{\theta}}^t$. The EM algorithm selects the best candidate from $\boldsymbol{\theta}$ that maximizes $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^t)$. The new chosen $\boldsymbol{\theta}$ is denoted $\boldsymbol{\theta}^{t+1}$. The overall algorithm is presented in Algorithm 2.

Algorithm 2: Expectation maximization algorithm

set: stopping condition ε, max iterations T, $t = 0$
begin
  i) initialize $\boldsymbol{\theta}^0$ to an initial guess (or random values)
  ii) **for** $t \in T$ **do**

    E-step compute $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^t)$
    M-step $Q(\boldsymbol{\theta}^{t+1}) = \arg\max_\theta Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^t)$

    if $Q(\boldsymbol{\theta}^{t+1}, \hat{\boldsymbol{\theta}}^t) - Q(\hat{\boldsymbol{\theta}}^t, \boldsymbol{\theta}^{t-1}) > \varepsilon$ **then**
      break
    end

  end
  iii) return $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{t+1}$
end

6.4.3 Binomial and multinomial mixture models

In this section we introduce a principal approach to estimate ground truth from multiple annotations. We consider the clinical annotations as a mixture model of binomial and multinomial distribution. We introduced general finite mixture model above. Here we describe the model for binary classification (binomial distribution) and for multiclass classification (multinomial distribution).

Let N is the number of instances and $y_i \in \mathcal{Y}$ is the unobservable ground truth for i-th instance, $i = 1, 2, \ldots, N$, $y_i^j \in \mathcal{Y}$ is an annotation assigned by an annotator j, where $j \in J$. For the binary case the $\mathcal{Y} = \{0, 1\}$ and for the multiclass case the $\mathcal{Y} = \{1, 2, \ldots, C\}$, where C is number of categories. For simplicity we begin description with binary classification and then extend it to multi-class classification. We provide only a short introduction for the more details we refer interested reader to the original work of (Dawid and Skene, 1979).

Binary classification

Let $\mathcal{Y} = \{0, 1\}$ be a binary class. In the case the true class y_i equals to 1, the parameter α is referred to as sensitivity

$$\alpha^j = \Pr[y_i^j = 1 | y_i = 1] = \frac{\Pr[y_i^j = 1, y_i = 1]}{\Pr[y_i = 1]}.$$

Conversely, if the true class equals to 0 the parameter β is known as specificity

$$\beta^j = \Pr[y_i^j = 0 | y_i = 0] = \frac{\Pr[y_i^j = 0, y_i = 0]}{\Pr[y_i = 0]}.$$

The assumption for α and β is that they are independent on the observed data. The assumption, which is violated in practise since some instances are more difficult than the others and each annotator posses a different expertise. The approach dealing with dependence on observed data was described in (Yan et al., 2010).

Unlike (Dawid and Skene, 1979) we formulate the model in simplified way, that is every annotator provides one label for each instance. With this simplification we completely rule out the possible violation of conditional independence between two labels assigned by one annotator for one instance. In other words, in the scenario when an annotator labels an instance for a second time, there is a probability that he/she remembers the decision made earlier. Even thought this probability is small

it would be different for each instance. By simplifying the model we eliminated it. Note that in the original paper of (Dawid and Skene, 1979) the α and β are denoted by a common variable π, where the π_{00} is sensitivity and π_{11} is specificity.

Let $p = \Pr[y_i = 1]$ be prevalence of class 1. Given observations $\mathcal{D} = \{y_i^1, \dots, y_i^J\}_{i=1}^N$ and parameters as $\boldsymbol{\theta} = \{\alpha^1, \dots, \alpha^j, \beta^1, \dots \beta^j, p\}$ the goal is to estimate unknown ground truth y_i and also the sensitivity α^j and specificity β^j for each annotator j. The y_i is treated as latent (hidden) variable and EM algorithm could be used to estimate it. With assumption that observed data are independent the likelihood can be formulated as

$$\Pr[\mathcal{D}|\boldsymbol{\theta}] = \prod_{i=1}^{N} \Pr[y_i^1, \dots, y_i^J|\boldsymbol{\theta}]$$

Under assumption that labels y_i^j are conditionally independent given α^j, β^j, and y_i, the likelihood written for two classes is

$$\Pr[\mathcal{D}|\boldsymbol{\theta}] = \prod_{i=1}^{N} \left[\sum_{c=0}^{1} \Pr[y_i^1, \dots, y_i^J|y_i = c, \boldsymbol{\theta}] \cdot \Pr[y_i = c|\boldsymbol{\theta}] \right].$$

We assume that $y_i^1, \dots y_i^J$ are independent, i.e. all annotators make their evaluation independently. Taking logarithm of the log likelihood yields

$$\log \Pr[\mathcal{D}|\boldsymbol{\theta}] = \sum_{i=1}^{N} \log \left[\sum_{c=0}^{1} \prod_{j=1}^{J} \Pr[y_i^j|y_i = c, \boldsymbol{\theta}] \cdot \Pr[y_i = c|\boldsymbol{\theta}] \right].$$

Further we rewrite the log likelihood as

$$\log \Pr[\mathcal{D}|\boldsymbol{\theta}] = \sum_{i=1}^{N} \log[pa_i + (1-p)b_i], \tag{6.8}$$

where

$$\begin{aligned} a_i &= \prod_{j=1}^{J} \Pr[y_i^j|y_i = 1, \alpha^j] = \prod_{j=1}^{J} [\alpha^j]^{y_i^j} [1-\alpha]^{1-y_i^j}, \\ b_i &= \prod_{j=1}^{J} \Pr[y_i^j|y_i = 0, \beta^j] = \prod_{j=1}^{J} [\beta^j]^{1-y_i^j} [1-\beta]^{y_i^j}. \end{aligned} \tag{6.9}$$

The maximum likelihood is found by maximizing the log likelihood function

$$\hat{\theta}_{\text{ML}} = \{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, p\} = \arg\max_{\theta} \{\log \Pr[\mathcal{D}|\boldsymbol{\theta}]\}.$$

Two methods are commonly used for the maximization of log likelihood with latent variables: EM and Markov Chain Monte Carlo simulation. The choice of the method is merely based on the researches preference. Also it is possible to derive analytical expression as was shown in (Pepe and Janes, 2007) but this approach has several flaws, which were summarized in the response letter (Formann and Böhning, 2008). In our work we decided to follow the line of Dawid and Skene and use the EM algorithm.

Estimation using EM algorithm The hidden variables to be estimated are sensitivity $\boldsymbol{\alpha}$, specificity $\boldsymbol{\beta}$, prevalence p, and true (unknown/hidden) label y_i. If we would known the hidden labels $\boldsymbol{y} = [y_i, \dots, y_N]$ the complete likelihood would be computed as

$$\log \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{\theta}] = \sum_{i=1}^{N} y_i \log pa_i + (1-y_i) \log(1-p)b_i.$$

Since we do not know the y_i, we replace it by its estimate μ_i. First we initialize the $\mu_i, p, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ and then we repeat the E and M step until convergence.

E-step. In this step we compute the conditional expectation of y_i given the observations from annotators $\mathcal{D}$ under the current estimates of parameters $\boldsymbol{\theta}$:

$$\mathbb{E}\{\log \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{\theta}]\} = \sum_{i=1}^{N} \mu_i \log p a_i + (1-\mu_i)\log(1-p)b_i, \tag{6.10}$$

where the expectation is with respect to $\Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{\theta}]$ and $\mu_i = \Pr[y_i = 1|y_i^1, \ldots, y_i^J, \boldsymbol{\theta}]$. By Bayes theorem the μ_i is computed as

$$\mu_i \propto \Pr[y_i^1, \ldots, y_i^J|y_i = 1, \boldsymbol{\theta}] \cdot \Pr[y_i = 1|\boldsymbol{\theta}] = \frac{a_i p}{a_i p + b_i(1-p)}.$$

M-step. The current estimate of μ_i is used to maximize the conditional expectation. By taking derivative of (6.10) and equating it to zero we compute the new estimates of α^j, β^j, and p

$$\alpha^j = \frac{\sum_{i=1}^N \mu_i y_i^j}{\sum_{i=1}^N \mu_i}, \quad \beta^j = \frac{\sum_{i=1}^N (1-\mu_i)(1-y_i^j)}{\sum_{i=1}^N (1-\mu_i)}, \quad p = (1/N)\sum_{i=1}^N \mu_i.$$

The EM algorithm is only guaranteed to converge to local maximum; therefore, it is usually restarted several times with different set of starting values. Other possible solution is to use the majority voting for initialization: $\mu_i = (1/J)\sum_{j=1}^J y_i^j$ as it was proposed in (Dawid and Skene, 1979).

Multi-class classification

The extension from binary classification to multi-class classification is straightforward. Let us define $c \in C$ as a category to which y_i^j could be assigned and $\boldsymbol{\alpha}_c^j = (\alpha_{c1}^j, \alpha_{c2}^j, \ldots, \alpha_{ck}^j)$ as a multinomial parameter describing a probability that annotator j assign a class $k \in C$ to an instance given the true class is c

$$\alpha_{ck}^j = \Pr[y_i^j = k|y_i = c], \quad \sum_{k=1}^{C} \alpha_{ck}^j = 1.$$

For the multi-class classification the sensitivity and specificity are usually computed with one-versus-all approach, where one class is taken as positive and the other classes as negative; this is repeated for all classes c. For binary case, $C = 2$, α_{00}^j and α_{11}^j refer to sensitivity and specificity, respectively. In order to keep the link to the original paper of (Dawid and Skene, 1979) we note here that they denoted the α_{ck}^j by $\pi_{ck}^{(j)}$.

Let $\delta(y_i^j, c)$ be a statement that equals 1 when $y_i^j = c$ and 0 otherwise and $p_c = \Pr[y_i = c]$ be a prevalence of category c. Furthermore, let denote observations $\mathcal{D} = \{y_i^1, \ldots, y_i^J\}_{i=1}^N$ and parameters as $\boldsymbol{\theta} = \{\alpha_{ck}^j, p_c\}$. Dawid and Skene (1979) treat the unknown truth y_i as latent (hidden) variable and uses the EM algorithm to estimate it. Similarly to (Raykar and Yu, 2012) the likelihood is proportional to

$$\Pr[\mathcal{D}|\boldsymbol{\theta}] = \prod_{i=1}^{N}\left[\sum_{c=1}^{C} \Pr[y_i = c] \prod_{j=1}^{J} \Pr[y_i^j|y_i = c]\right] = \prod_{i=1}^{N}\left[\sum_{c=1}^{C} p_c \prod_{j=1}^{J}\prod_{k=1}^{C} (\alpha_{ck}^j)^{\delta(y_i^j,k)}\right]. \tag{6.11}$$

If we know the missing labels $\boldsymbol{y}$ the log likelihood can be written as

$$\log \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{\theta}] = \sum_{i=1}^{N}\sum_{c=1}^{C} \delta(y_i, c) \log\left[p_c \prod_{j=1}^{J}\prod_{k=1}^{C} (\alpha_{ck}^j)^{\delta(y_i^j,k)}\right].$$

Estimation using the EM algorithm As in the binary case we use expectation maximization algorithm to estimate the latent parameters. In the **E-step** the conditional expectation is computed as

$$\mathbb{E}\{\log \Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{\theta}]\} = \sum_{i=1}^{N}\sum_{c=1}^{C} \mu_{ic} \log\left[p_c \prod_{j=1}^{J}\prod_{k=1}^{C} (\alpha_{ck}^{j})^{\delta(y_i^j,k)} \right], \tag{6.12}$$

where $\mu_{ic} = \Pr[y_i = c|y_i^1, \ldots, y_i^J, \boldsymbol{\theta}]$ is estimated probability of ground truth given the y_i^j and $\boldsymbol{\theta}$ and can be computed as

$$\mu_{ic} \propto p_c \prod_{j=1}^{J}\prod_{k=1}^{C} (\alpha_{ck}^{j})^{\delta(y_i^j,k)}.$$

In the **M-step** we use the current estimates to maximize the conditional expectation. Taking gradient of (6.12) and equating it to zero, the parameter α_{ck}^{j} is updated using the following equation

$$\alpha_{ck}^{j} = \frac{\sum_{i=1}^{N} \mu_{ic}\delta(y_i^j, k)}{\sum_{i=1}^{N} \mu_{ic}}.$$

The E and M step are repeated until convergence; the μ_{ic} was initialized using majority voting.

Latent class analysis with different number of classes

The latent model is powerful not only for estimating the latent class from multiple, possibly noisy, annotations but could be also used to infer the number of classes the annotators are actually using. We discussed the FIGO guidelines in Section 2.2.1. Here we briefly discuss their alternatives. The FIGO guidelines (FIGO, 1986) were the first recognized international guidelines. Since then many alternatives were introduced (ACOG, 2009; Macones et al., 2008; RCOG, 2001) all employing 3-tier classification system based on FIGO. The comparison of guidelines and their statements was performed by (de Campos et al., 2010) with conclusion that the guidelines are, in general, too complex and hard to follow in clinical environment. Attributing to high inter/intra observer variability. To better interpret the CTG patterns and to lower the variability alternatives to 3-tier were devised. Schifrin (2004) advocates a simple 4-tier system while (Parer and Ikeda, 2007; Parer et al., 2009) propose a 5-tier system claiming its superiority over the classical guidelines (Parer and Hamilton, 2010). Tommaso et al. (2013) showed that the NIHCD guidelines had better sensitivity and specificity over 5-tier system but, in general, the performance of 5-tier was better. Further, Coletta et al. (2012) claimed that there is a better sensitivity of 5-tier system though (Miller and Miller, 2012) provided the contrary.

The latent class model for CTG evaluation was properly introduced in the section above. Using this model we can gain insight into the discrepancy of guidelines and disagreement on number of categories, which should be used for evaluation. In the equation (6.11) we computed the likelihood of parameters $\boldsymbol{\theta}$ given the labels in set $\mathcal{D}$. In this equation we supposed that number of classes is fixed. However, the guidelines are not precise nor they are strictly followed by clinicians leaving and open space for different evaluation. Our goal is to examine if choosing different number of classes offers better description of clinical evaluation in terms of model fit. The extension of equation (6.11) to encompass different number of classes is straightforward. We replace C by a number R representing different number of classes

$$\Pr[\mathcal{D}|\boldsymbol{\theta}] = \prod_{i=1}^{N}\left[\sum_{r=1}^{R} p_r \prod_{j=1}^{J}\prod_{k=1}^{C} (\alpha_{rk}^{j})^{\delta(y_i^j,k)} \right],$$

where

$$\alpha_{rk}^{j} = \Pr[y_i^j = k|y_i = r], \quad \sum_{k=1}^{C} \alpha_{rk}^{j} = 1.$$

The α_{rk}^j are probabilities, which represent class-conditional probability that a label in r-th class corresponds to label in k-th class given by j-th annotator. In our experiments, we varied R to be $R = \{2, 3, \ldots, 8\}$, creating models $Mr_2, Mr_3, \ldots, Mr_8$.

Number of estimated parameters The number of estimated parameters Λ increase rapidly with increasing R, J, and C and is computed as

$$\Lambda = R - 1 + \sum_{j=1}^{J} (C - 1) + (R - 1). \tag{6.13}$$

For instance when $C = 2, J = 2$, and $R = 2$ the parameters to be estimated are as follows $\boldsymbol{\theta} = \{\alpha^1, \alpha^2, \beta^1, \beta^2, p\}$. If the Λ exceeds number of observations the model will be unidentified. The model will be also unidentified if the probabilities α_{rk}^j will be sparse.

Rank of annotators

The latent class model (LCM) is not only useful to estimate the latent class but it can be also used to evaluate the agreement between observers or, more precisely, from the LCM model we can infer the contribution of individual clinicians to the latent class estimate.

The ranking/scoring of individual annotators was thoroughly investigated in (Raykar and Yu, 2012). Here we provide a brief summary with addition of our devised score, which is simplest, better interpretable, and therefore more suitable for our approach.

Recall that for the *binary classification* the parameters α^j and β^j refer to sensitivity and specificity, respectively. To represent the α^j and β^j in one number, (Raykar and Yu, 2012) proposed a score to detect spammers (bad annotators) in a crowd-sourcing task

$$\mathcal{S}_{sp}^j = (\alpha^j + \beta^j - 1)^2.$$

A spammer assign labels randomly and therefore sensitivity and specificity is almost equal. The score for spammer tends to zero while ideal annotators have $\mathcal{S}_{sp}^j = 1$.

For the *multiclass classification* the situation is similar but instead of sensitivity and specificity we work with parameter α_{ck}^j where the estimated latent class was c and annotator j assigned class k. Again, note that when $C = 2$ the α_{00}^j and α_{11}^j refer to sensitivity and specificity, respectively. Let $\mathbf{A}^j$ be a $C \times C$ confusion matrix with entries $[\mathbf{A}^j]_{ck} = \alpha_{ck}^j$. In the case of spammer the rows in $\mathbf{A}^j$ would be equal one another. For example when $C = 3$ the $\mathbf{A}^j$ for spammer could be

$$\mathbf{A}^j = \begin{pmatrix} \alpha_{11} & \ldots & \alpha_{13} \\ \vdots & \cdots & \vdots \\ \alpha_{31} & \ldots & \alpha_{33} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \end{pmatrix}. \tag{6.14}$$

The score is defined as Frobenius norm of the confusion matrix to the closest rank one approximation. For each row of $\mathbf{A}^j$ the mean of this row is subtracted. Then the summation of the squares of all entries yields the final score. Equivalently

$$\mathcal{S}_{sp}^j = \left\| \mathbf{A}^j - \frac{1}{C} \mathbf{e}\mathbf{e}^T \mathbf{A}^j \right\|_F^2 = \left\| \left(\mathbf{I} - \frac{1}{C} \mathbf{e}\mathbf{e}^T \right) \mathbf{A}^j \right\|_F^2,$$

where $\mathbf{e}$ is a column vector of ones and $\mathbf{I}$ is identity matrix with ones on diagonal. A spammer is when $\mathcal{S}_{sp}^j$ tends to zero, a good annotator tends to $\mathcal{S}_{sp}^j = C - 1$. The normalization of this score to C classes yields

$$\mathcal{S}_{sp}^j = \frac{1}{C-1} \left\| \left(\mathbf{I} - \frac{1}{C} \mathbf{e}\mathbf{e}^T \right) \mathbf{A}^j \right\|_F^2.$$

Even though this score is effective for penalizing random annotations it does not cover situations when an annotator assigns only one class (or uses one class prevalently). For example consider the following confusion matrices for good annotator $j = g$ and bad $j = b$

$$\mathbf{A}^g = \begin{pmatrix} 0.9 & 0.1 & 0 \\ 0.1 & 0.9 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix} \quad \mathbf{A}^b = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.05 & 0.95 \\ 0 & 0.05 & 0.95 \end{pmatrix}.$$

The good annotator $\mathbf{A}^g$ has performed well on the classes $C = \{1, 2\}$ and badly on the third class, where the probability of correct decision was 0.5. The bad annotator $\mathbf{A}^b$ evaluated prevalently the third class. Evidently the performance was good on the third class but poor on the first and even poorer on the second class. Nevertheless the score $\mathcal{S}_{sp}$ is similar for both $\mathcal{S}^g_{sp} = 0.49$ and $\mathcal{S}^b_{sp} = 0.45$.

Accuracy based scoring In order to differentiate good and bad annotators, which are not spammers (random evaluation) but use dominantly one class, we proposed a different scoring function. This scoring is more intuitive and in a sense corresponds to the classification accuracy. Again, consider matrix $\mathbf{A}^j$ the diagonal elements represent probabilities of correct classifications with respect to latent class $c = k$ and off-diagonal elements represent probabilities of misclassification $c \neq k$. The proposed score is defined as

$$\mathcal{S}^j_{acc} = \frac{1}{C}\left(2 \cdot \mathrm{trace}(\mathbf{A}^j) - \sum_{c=1}^{C}\sum_{k=1}^{K} \mathbf{A}^j_{ck}\right).$$

The score simply equal to summation of diagonal elements with subtraction of summation of off-diagonal elements. The score for very bad annotators is $\mathcal{S}^j_{acc} = -1$ and for the good annotators is $\mathcal{S}^j_{acc} = 1$. For the example above the scores are $\mathcal{S}^g_{acc} = 0$ and $\mathcal{S}^b_{acc} = 0.53$. For the case of spammer, confusion matrix in (6.14), the score is even lower $\mathcal{S}_{acc} = -0.33$. Clearly, in these cases the proposed score is able to differentiate the various annotators better than the $\mathcal{S}_{sp}$.

Ranking for different number of classes The proposed accuracy based score has a limitation when we consider that the latent variable has different number of classes than the annotators actually used. We described this approach in the previous section. Let $\mathbf{A}^j_{rk}$ be a matrix with entries $[\mathbf{A}^j]_{rk} = \alpha_{rk}$, where $R \neq K$. Then the $\mathbf{A}^j_{rk}$ is not square and the accuracy based score S^j_{acc} can not be computed and only the $\mathcal{S}^j_{sp}$ can be used.

6.4.4 Model selection and fit

With the latent class analysis we can use miscellaneous techniques to evaluate model fit and to determine, which model is more appropriate for different R. Increasing R from r_{min} to r_{max} (in our case $r_{min} = 2$ and $r_{max} = 8$) will increase model fit but also with possibility to over-fit and need of estimating additional model parameters, see equation (6.13). A trade-off between better model fit and number of parameters to be estimated is usually sought and tackled by penalizing the log likelihood by a function of parameters $\boldsymbol{\theta}$ that are needed to be estimated. Criterion for selecting number of classes r can be formulated as

$$\hat{r} = \arg\min_r \left\{\mathcal{C}(\hat{\boldsymbol{\theta}}(r)), r = r_{min}, \ldots, r_{max}\right\}, \tag{6.15}$$

where $\mathcal{C}(\hat{\boldsymbol{\theta}}(r))$ is a model selection criterion and $\hat{\boldsymbol{\theta}}(r)$ is an estimate of parameters of r classes. The two most common measures are the Akaike information criterion (AIC) (Akaike, 1973) and Bayes information criterion (BIC) (Schwarz, 1978). We can express both of these criteria in the common form

$$\mathcal{C}(\hat{\boldsymbol{\theta}}(r)) = -\ln \Pr[\mathcal{D}|\hat{\boldsymbol{\theta}}(r)] + \mathcal{P}(r),$$

where $\mathcal{P}(r)$ is a function that penalizes higher number of classes r and $\Pr[\mathcal{D}|\hat{\boldsymbol{\theta}}(r)]$ is likelihood of θ given $\mathcal{D}$. Let L is the likelihood, N is the number of examples, and ϑ is the number of estimated parameters, the AIC and BIC are defined as

$$AIC(r) = -2\ln L + 2\vartheta,$$
$$BIC(r) = -2\ln L + \vartheta \ln N.$$

The better model the lower BIC and/or AIC. Usually the AIC over estimates the number of r while BIC underestimates it, the compromise between them is often sought. Though, for the latent class models, the BIC is usually preferred because of its simplicity (Lin and Dayton, 1997).

Why AIC/BIC are preferred over cross-validation? In this stage, when choosing the best model, we are not interested in prediction capabilities of the model. Rather we aim to select the model, which has the best descriptive properties therefore we use AIC and BIC.

6.4.5 Statistical measures

In this chapter we use sensitivity, specificity, and precision (positive predictive value) to assess clinical evaluation of CTG with respect to different markers (pH, BE, BDecf, and Apgar score). Let us consider the confusion matrix in Table 6.1. TN (true negative) express number of correctly classified negative examples, TP (true positive) is number of correctly classified positive examples, FN (false negative) is number of incorrectly classified negative examples, and FP (false positive) is the number of incorrectly classified positive examples.

Table 6.1: Confusion matrix. p/n – actual positive/negative, p′/n′ – predicted positive/negative.

	p′	n′
p	TP	FP
n	FN	TN

Sensitivity is accuracy on positive examples $SE = TP/(TP + FN)$, specificity is accuracy on negative examples $SP = TN/(FP + TN)$, and precision (PR) is proportion of predicted positive results that are true positive $PR = TP/(TP + FP)$. For more information on statistical measures on confusion matrices c.f. Section 7.4.3 or see any general textbook.

6.5 Statistical analysis of features with respect to clinical evaluation

In Chapter 5 we described set of features extracted from fetal heart rate record. The features were extracted for the whole FHR (Step 1 plus Step 2). We evaluated the clinical evaluation without FHR so far, here we are interested in the relationship between extracted FHR features and corresponding clinical evaluation. The assessment is performed using statistical testing.

To be able to select appropriate statistical tests all features were tested for normal distribution using Lilliefors test. The test operates with null hypothesis that sample comes from normal distribution. Only features with normal distribution in all three classes (normal, suspicious, pathological) were considered to have normal distribution in general. For the normal distributed features we used Analysis of variance (ANOVA) and for not normally distributed we used non-parametric Kruskal-Wallis test, which makes no distributional assumptions. We tested the null hypothesis that features comes from the same distribution against alternative hypothesis that they don't. The null hypothesis was rejected when $p < 0.01$.

Features were divided into groups based on their origin: FIGO-based, HRV-based, and nonlinear. Since we obtained a large amount of features we filtered them in each group using correlation. Only

one representative feature was retained and other correlated were removed when the correlation was $\hat{\rho} > |0.9|$.

6.6 Results

6.6.1 Proportion of agreement and inter/intra observer variability

To keep the results as clear as possible we focused on presentation of the results acquired from the second step in the first stage of labour and the prediction of the outcome based on the full CTG recording (Steps 2 and 4 as described above). The detailed results for individual clinicians are shown only for Step 2 since for Step 4 the results are similar.

Recall that for Step 2 the classes were: normal, suspicious, pathological, and uninterpretable and for step 4 they were: no hypoxia, mild hypoxia, severe hypoxia, and uninterpretable.

In total, 552 unique records were presented to nine clinical experts for annotations and together with almost 20 % of records that were presented repeatedly with random repetitions and class-dependent repetitions amounted to 634 recordings to annotate. All clinicians were randomly assigned by a number, which can be used to connect the respective results across all the figures in this section.

Percentage of assigned classes. We evaluated the median percentage of clinical annotation. The results are present in first column of Table 6.2. Note that percentages do not add to 100% since we computed median across all clinicians. We can see that evaluation in step 4 had higher percentage of normal records than evaluation in the Step 2. The details for percentages of recordings assigned to particular classes by individual experts are shown in Figure 6.3. Even from this figure we can see that proportion of normal, suspicious, pathological, and uninterpretable evaluation from each clinician differed substantially, experts 3 and 8 being the most defensive ones, the expert 6 on the other hand the most confident one. The Figure 6.3 shows results for Step 2 only (detailed results for Step 4 are not presented).

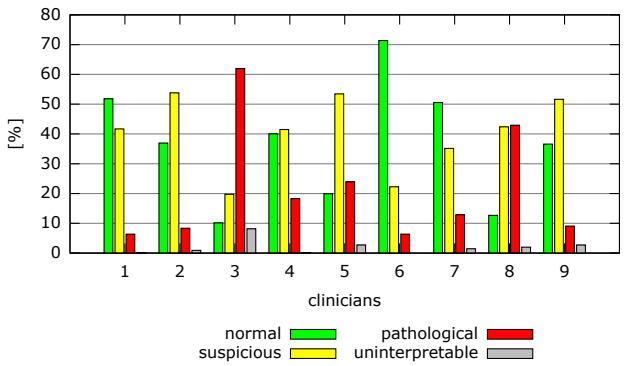


Figure 6.3: Percentage of normal, suspicious, pathological, and uninterpretable evaluation based on evaluation of Step 2 by 9 expert-clinicians.

Average time of evaluation The average length of annotation, i.e. time elapsed between evaluation of the first and last record, was 38 days. The average time distance of records that were presented for the second time (for intra-observer agreement) was five days; hence it is unlikely that clinicians would remember the repeated CTGs. The median time needed for decision for all types of evaluation with 25th and 75th percentiles are shown in second column in Table 6.2. Details for individual clinicians for Step 2 are shown in Figure 6.4. For all clinicians but 3 and 8 the pathological evaluation took more time than normal and suspicious. All clinicians had significantly different times for each type of evaluation on significance level $p < 0.01$ (Kruskall-Wallis test was used).

Table 6.2: Assessment of clinical evaluation. Percentage of class evaluation, average time per evaluation, intra-observer agreement, and proportion of agreement,

	class	median percentage of evaluation [%]	average time (median (25-75 perc.)) [sec.]	intra-observer agreement [%]	proportion of agreement (95% CI) [%]
Step 2	overall	—	4.1 (2.6 – 8.5)	71.2	48 (47 – 50)
	normal	37.0	2.9 (2.1 – 4.2)	80.7	57 (54 – 60)
	suspicious	41.7	3.8 (2.7 – 6.3)	65.7	46 (48 – 48)
	pathological	12.9	7.0 (3.9 – 11.9)	50	41 (36 – 46)
	uninterpretable	1.5	6.0 (4.4 – 7.5)	0	15 (10 – 21)
Step 4	overall	—	4.8 (2.5 – 8.4)	71.6	50 (48 – 52)
	no hypoxia	67.8	6.8 (3.6 – 12.1)	86.3	65 (63 – 68)
	mild hypoxia	23.7	11.5 (5.2 – 24.6)	58.6	32 (30 – 34)
	severe hypoxia	7.3	8.6 (6.3 – 9.7)	53.3	29 (25 – 33)
	undecidable	1.8	6.1 (5.1 – 8.2)	0	20 (16 – 24)

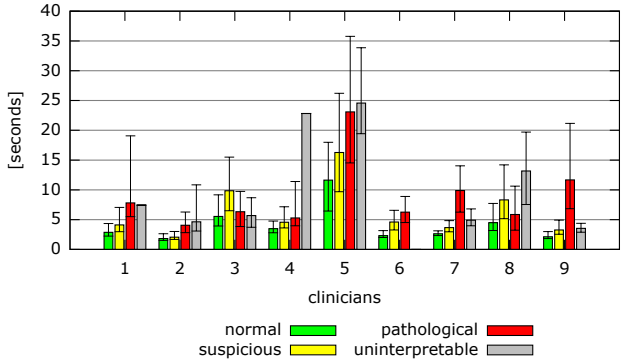


Figure 6.4: Average times for evaluation of the recording as normal, suspicious, pathological, and uninterpretable. Values presented as median with 25th and 75th percentiles.

Intra-observer agreement. The average intra-observer proportion of agreement is shown in third column of Table 6.2. Again, detailed results for individual clinicians on Step 2 are presented in Figure 6.5. We should stress out that from Figure 6.3, we already know that experts 3 and 8 evaluated the data defensively – they had disproportionate amount of pathological evaluations – thus it can partly explain the large agreement of experts 3 and 8 on pathological class in comparison to most of the others in Figure 6.5. Similar results were obtained for the labour evaluation.

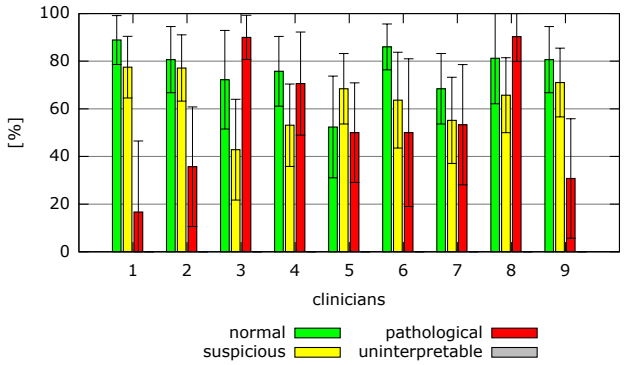


Figure 6.5: Intra-observer proportion of agreement with 95% confidence intervals in respect to assessed CTG category.

Proportion of agreement. The overall proportion of agreement (PA) is present in the last column of Table 6.2. Evidently, the PA values are low and there are two reasons for that. First, by the common sense, the more experts asked for an opinion the less agreement we expect; second, two clinicians could be considered as outliers since they evaluated CTG defensively, as was shown in Figure 6.3 and also evidenced in the figures hereinafter. In order to examine inter-observer agreement in more detail we computed PA of all clinicians and majority voting ($PA_1, \ldots, PA_9$) with results: PA_j = {75.9, 78.6, 45.8, 77.2, 46.4, 78, 76.6, 35.1, 73}. We sorted the PA_j in descending way and evaluated inter-observer agreement iteratively. In each iteration we added a clinician and evaluated PA. Thus obtaining PA for groups of two clinicians {2, 6}: three clinicians {2, 6, 4}, four {2, 6, 4, 7} and so forth. The results are shown in Figure 6.6. It could be seen that for normal evaluation the PA decreased almost linearly.

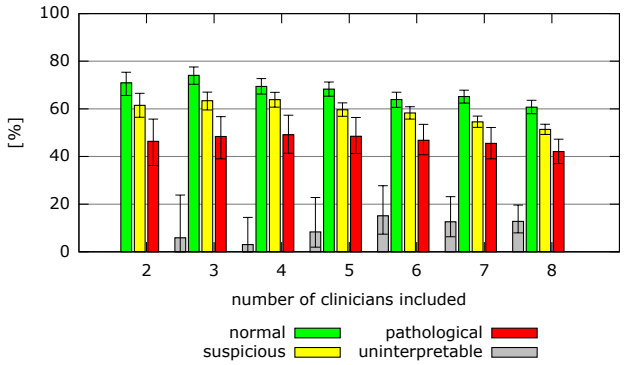
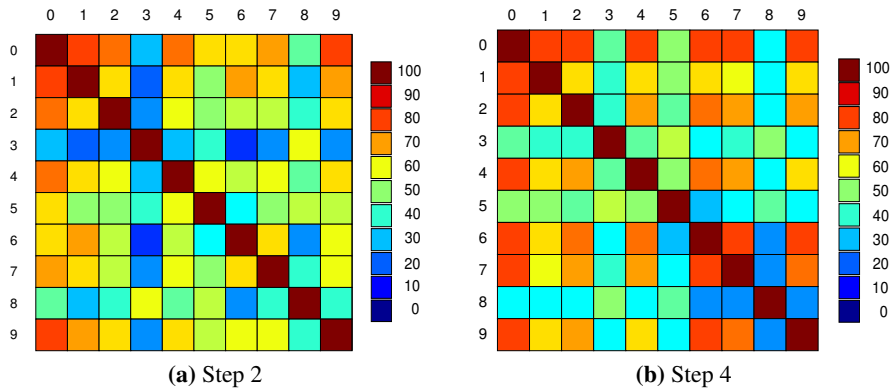


Figure 6.6: Inter-observer agreement with 95% confidence intervals for groups of 2, 3, 4, 5, 6, 7, 8, and 9 clinicians.

Inter-observer agreement. The more details of inter-observer PA between individual clinicians and majority voting are shown in the matrix in Figure 6.7. The matrix is symmetric along its diagonal. The majority voting outcome based on evaluation of all clinicians is marked by 0 and shown in the first row and the first column. The rest of the rows and columns represent individual clinicians as in the previous figures. For example the first clinician (marked with 1) agrees with the majority voting (marked with 0) on 80 % and with the clinician (marked with 2) on 65 %.



(a) Step 2 **(b)** Step 4

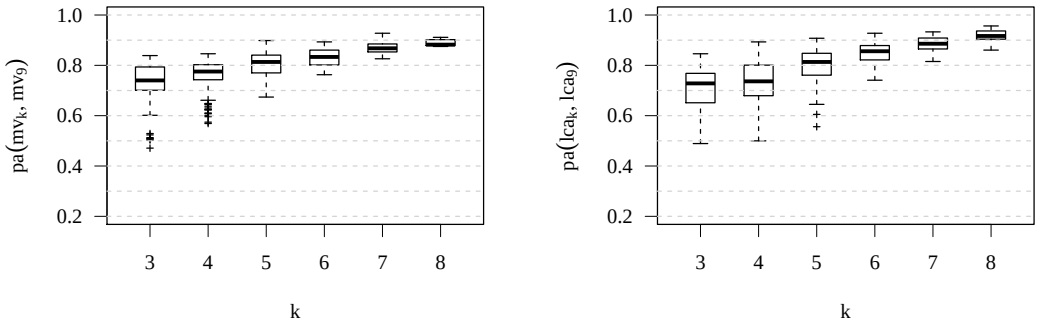
Figure 6.7: Matrix of inter-observer proportion of agreement. (Majority voting = 0, clinicians = {1,2, . . . , 9}). On the right of the figure, bars represent levels of proportion of agreement.

For the sake of completeness we also computed Fleiss kappa coefficient, however we warn against its improper use for comparing different populations. The overall kappa was: 0.255 with 95% CI (0.253–0.258) being the fair agreement.

Clinical evaluation from hospital records For 262 CTGs records we were able to obtain CTG evaluation directly from hospital records of the delivery ward of the UHB[1]. The evaluation – usually of the last segment of the first stage and second stage – was given on the printed-out CTG recording. The proportion of agreement between evaluation of the first stage from the clinical documentation and the evaluation using the majority voting of nine experts was 59% CI (52–64 %). With respect to different categories the PAs were: 61 % with 95% CI:(53–70) for normal, 63% (56–69) for suspicious, and 32% (18–47) for pathological.

6.6.2 Stability of majority voting

We analysed stability of majority voting and latent class model. The stability was defined in Proposition 1. In Figure 6.8 we present overall results (irrespective the classes) of majority voting (MV) and latent class model (LCM) stability. The stability of LCM is better for higher number of clinicians while for the lower number the MV perform better and has also lower variance. For $k \leq 5$ the MV should be preferred and for $k > 5$ the LCM should be favoured. The same conclusion holds when the overall evaluation is split into the individual classes as it is shown in Figures 6.9 for MV and 6.10 for LCM. Recall that the comparison is not rigours but statistical testing could be hardly employed as discussed in Section 6.3.3.



(a) majority voting **(b)** latent class estimate

Figure 6.8: Stability of majority voting and latent class model for all classes. Legend: mv_k and mv_9 is majority voting for k and 9 clinicians, respectively; lca_k and lca_9 is latent class analysis of k and 9 clinicians, respectively; pa(a,b) is proportion of agreement between a and b.

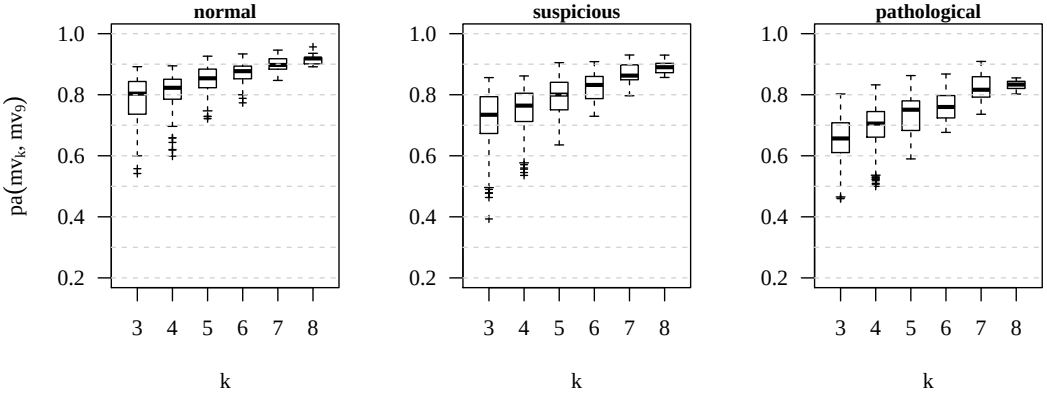


Figure 6.9: Stability of majority voting (MV) for normal, suspicious, and pathological evaluation. Legend: mv_k and mv_9 is majority voting for k and 9 clinicians, respectively; pa(a,b) is proportion of agreement between a and b.

[1] University hospital in Brno

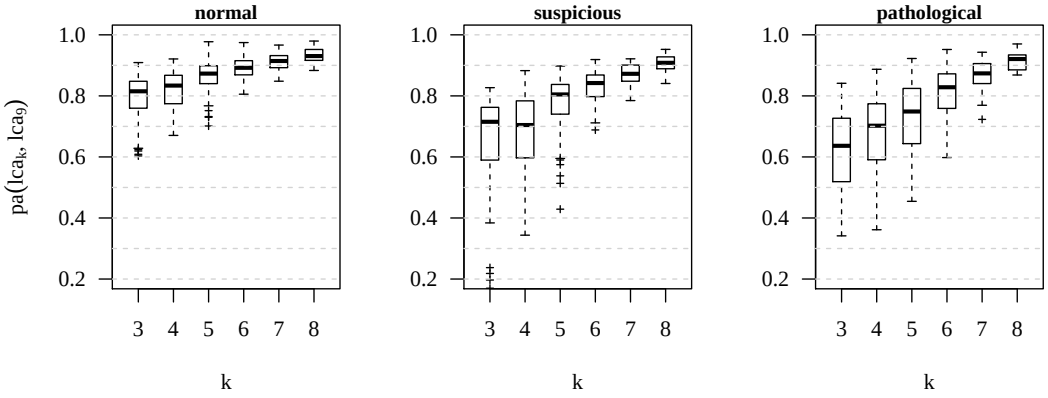


Figure 6.10: Stability of latent class model (LCM) for normal, suspicious, and pathological evaluation. Legend: lca_k and lca_9 is latent class analysis of k and 9 clinicians, respectively; pa(a,b) is proportion of agreement between a and b.

6.6.3 Latent class analysis

Different number of classes We analysed the clinical evaluation using the latent class model for different number of classes, thus creating 7 models: $M_{r2}, M_{r3}, M_{r4}, M_{r5}, M_{r6}, M_{r7}, M_{r8}$. The model fit statistics are show in Table 6.3. The progression of AIC and BIC for increasing number of r is shown in Figure 6.11. Clinicians should evaluate the CTG using three FIGO classes (normal, suspicious, and pathological) but from Figure 6.11 we can conclude that the best fit is for model M_{r4}. From the model M_{r3} to M_{r4} the both measures AIC and BIC climb down. The BIC starts rising from M_{r4} to M_{r5} while the AIC only slightly decreases, hence the best fitted model is M_{r4}.

Table 6.3: Fit statistics for different number of classes (df – degrees of freedom, AIC – Akaike information criterion, BIC – Bayes information criterion).

model	df	AIC	BIC
M_{r2}	515	7316	7476
M_{r3}	496	6842	7083
M_{r4}	477	6677	7000
M_{r5}	458	6656	7062
M_{r6}	439	6666	7154
M_{r7}	420	6669	7239
M_{r8}	401	6688	7340

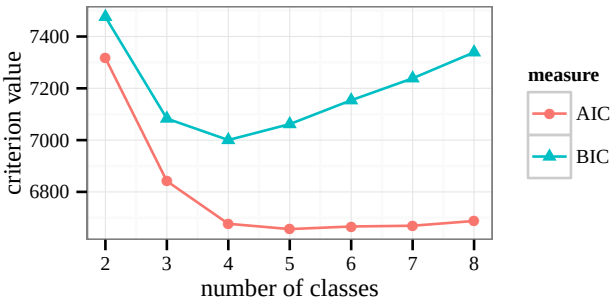


Figure 6.11: Progression of AIC and BIC for different number of classes ($r = \{2, 3 \ldots, 8\}$).

In order to have better insight into models M_{r3} and M_{r4}. We evaluated conditional clinicians response probability with respect to latent class (outcome). Model M_{r3} is presented in Figure 6.12

and M_{r4} in Figure 6.13. The clinicians are marked with numbers $(1, 2, \ldots, 9)$ and their responses are separated with respect to estimated latent class.

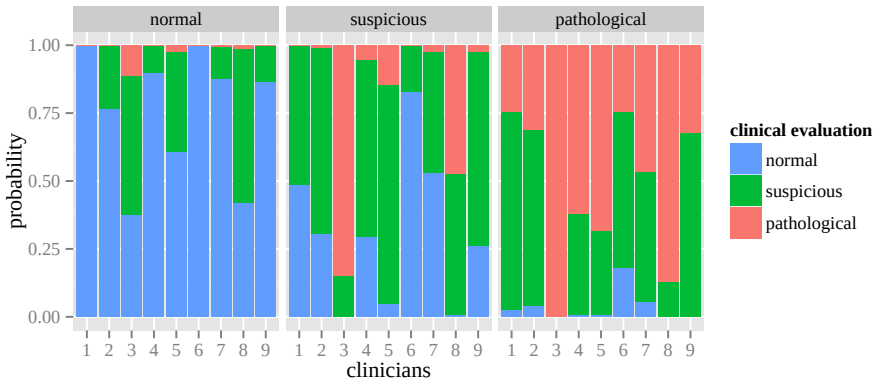


Figure 6.12: Conditional clinicians response probability with respect to latent class (outcome). Model M_{r3}. Estimated latent classes were as follows: normal, suspicious, and pathological (shown in grey headings). Class population shares: $P(normal)$ = 0.30, $P(suspicious)$ = 0.45, $P(pathological)$ = 0.25.

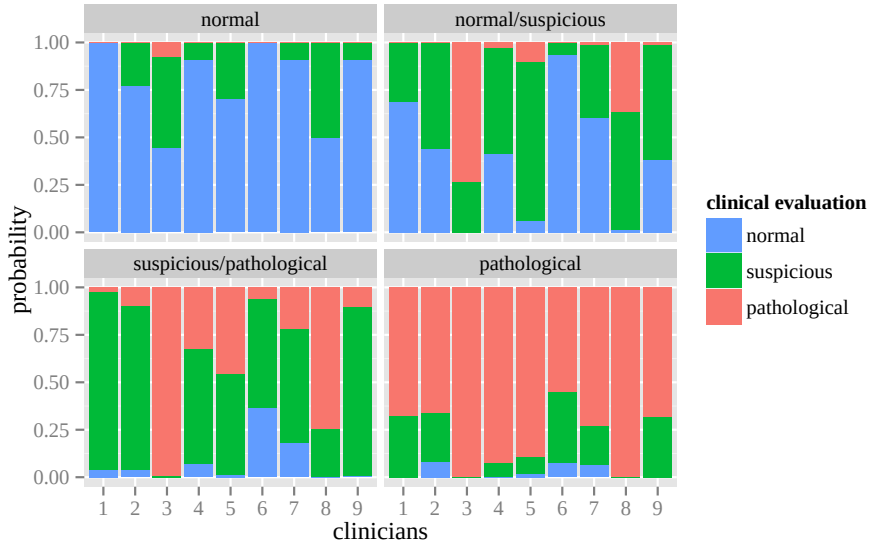


Figure 6.13: Conditional clinicians response probability with respect to latent class (outcome). Model M_{r4}. Estimated latent classes were named as follows: normal, normal/suspicious, suspicious/pathological, and pathological (shown in grey headings). Class population shares: $P(normal)$ = 0.25, $P(normal/suspicious)$ = 0.38, $P(suspicious/pathological)$ = 0.29, $P(pathological)$ = 0.08.

For the M_{r3} the latent class can be separated into normal, suspicious, and pathological based on the majority of clinician's evaluation. For the M_{r4} the situation is more complicated. The assignment of classes is rather intuitive and could be determined with help of knowing proportions of classes shown in Figure 6.3. For the first class, that we ex-post assigned as normal, we can see that majority of clinicians evaluation was normal. For the second class, ex-post assigned to normal/suspicious we can observe discrepancy in clinical evaluation, prevalent normal evaluation (clinicians 1,6,7), prevalent suspicious (clinicians 2, 4, 5, 8, 9), and prevalent pathological (clinician 3). Keeping in mind proportions of evaluation, Figure 6.3, we know that clinician 6 mostly evaluated CTG as normal, while clinicians 3 mostly evaluated CTG as pathological. In this case we can neglect them for decision on class assignment and use other clinicians. Hence we assigned it as normal/suspicious. Again, we would like to note that the assignment of final class is based on the intuition rather than rigours classification.

Comparing models M_{r3} and M_{r4} we can conclude that pathological class for M_{r4} is better separated for M_{r4} that it is for M_{r3}. When we further increased r, model M_{r5}, the pathological class remained almost identical, however the classes normal – suspicious were slightly better separated (figure not shown). Nevertheless, the separation comes at a expense of estimating more model parameters and hence increasing BIC.

Rank of clinicians The rank of individual clinicians contributing to the latent class estimate was determined using $\mathcal{S}_{acc}$ score. These scores are presented in Table 6.4 and the progression of the score during the iterations of the EM algorithm is presented in Figure 6.14.

Table 6.4: Score of individual clinicians for model M_{r3}.

clinician	$\mathcal{S}_{acc}$	rank
1	0.182	7
2	0.184	6
3	0.008	8
4	0.456	1
5	0.396	2
6	−0.045	9
7	0.198	4
8	0.195	5
9	0.274	3

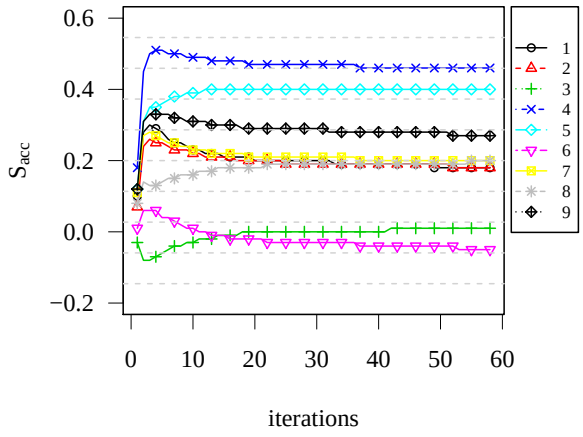


Figure 6.14: Scoring of individual clinicians with increasing number of iterations for model M_{r3}. Clinicians are represented by numbers.

From the point of view of latent class model, there are two distinct clinicians 4 and 5. Then follow clinician 9 and further group of clinicians 1, 2, 7, and 8 with similar scores. From Figure 6.14 we can see that after the 10th iteration the scores of clinicians remained more or less stable. We note here that the score do not correlate with the clinicians' experience nor it correlates with the work place. Using the score $\mathcal{S}_{sp}$ different results were obtained (table and figure not shown), however, as we pointed out above, this score has lower discriminability than the $\mathcal{S}_{acc}$ score.

6.6.4 Sensitivity and specificity of clinical evaluation

We evaluated the sensitivity and specificity with respect to different aggregation of clinical evaluation: i) majority voting (MV) and ii) latent class model (LCM) of clinical evaluation. We present only figures regarding the MV with respect to pH value, the other results are presented in tables. In order to compute sensitivity and specificity we have to shrink the three class evaluation into two class. When a

clinician assess a CTG recording as pathological he/she thinks that there is a serious problem with the fetus. On the other hand, when a clinician assess it as suspicious it is likely more of a hunch and there might be something wrong but it is most likely not that serious. We decided to compute the sensitivity and specificity as pathological evaluation vs (normal + suspicious evaluation) even though the assignment of suspicious to normal is not entirely correct.

Distribution of pH values regarding to majority voting of the clinical evaluation (Step 2) is shown in Figure 6.15 and the distribution of pH with respect to labour evaluation (Step 4) in Figure 6.16. On the both figures the pathological evaluations are scattered across the whole pH range. The interesting records are those having low pH and normal evaluation and vice versa (e.g. in Step 2 there are normal evaluations for pH 6.98 and 7.00). In accordance to Table 6.2 it could be seen that for Step 4 there is larger proportion of normal evaluation than suspicious from Step 4.

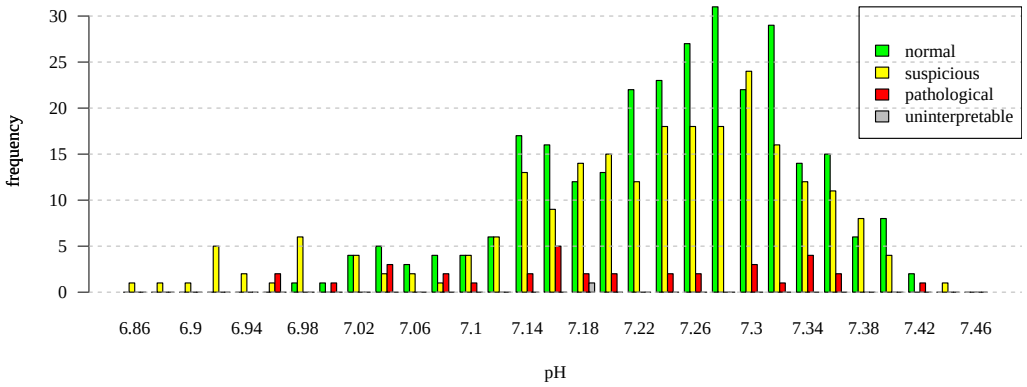


Figure 6.15: Majority voting vs. umbilical artery pH (Step 2).

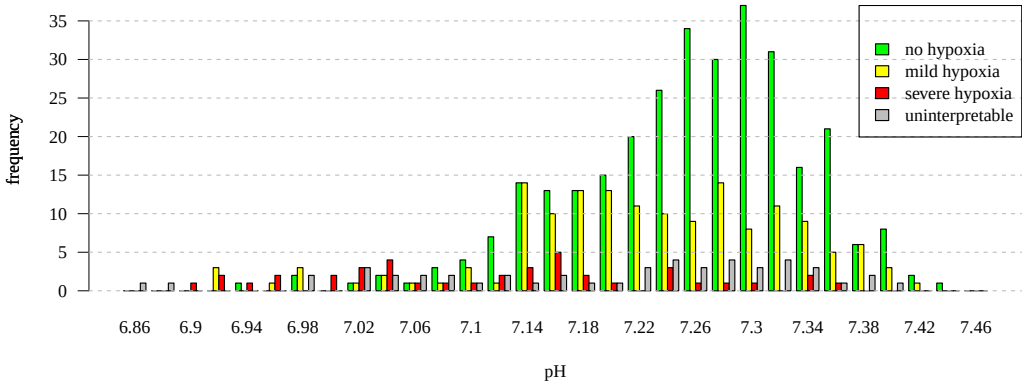


Figure 6.16: Majority voting vs. umbilical artery pH (Step 4).

Table 6.5 presents evaluation obtained from hospital records (262 CTGs) regarding to pH, base excess (BE), base deficit (BDecf), and Apgar score at 5 minutes. For all markers the sensitivity is lower than specificity. The best sensitivity and specificity was achieved for BDecf ≥ 12.

The results of sensitivity and specificity of different markers with respect to majority voting for Step 2 and Step 4 are presented in Table 6.6. These results are inferior to the results obtained from the clinical evaluation from hospital records (Table 6.5). However, the hospital records represent only subset of recordings that were not selected randomly. Very similar results were obtained when we recomputed sensitivity and specificity for the same subset of records (table not shown). Based on the comparison, the selected hospital evaluation are not representative, hence not used further.

In Table 6.7 we present result of latent class model (LCM), M_{r3}, of clinical evaluation of Step 2 and Step 4. In comparison to majority voting the LCM has better sensitivity but at a price of lower specificity. We note that the LCM model is preferred since the majority voting is less stable as was shown above.

Table 6.5: Sensitivity (SE), specificity (SP), and precision (PR) of clinical evaluation obtained from hospital records (eval. hosp. recs.). Timing of evaluation corresponds to Step 2. Results are presented as pathological vs. (suspicious + normal).

	objective	SE (95% CI)	SP (95% CI)	PR (95% CI)
eval. hosp. recs.	pH ≤ 7.05	41 (35–47)	94 (91–97)	32 (26–38)
	pH ≤ 7.10	42 (36–48)	95 (92–98)	45 (39–51)
	BE ≤ -12	44 (38–50)	94 (91–97)	32 (26–38)
	BDecf ≥ 12	60 (54–66)	94 (91–97)	27 (22–32)
	Apgar < 7	0 (0– 0)	91 (88–94)	0 (0– 0)

Table 6.6: Sensitivity (SE), specificity (SP), and precision (PR) of majority voting of clinical evaluation for Step 2 and 4. Results for 533 records because for nine records the majority vote was uninterpretable. Results presented as pathological evaluation vs. (suspicious + normal).

	objective	SE (95% CI)	SP (95% CI)	PR (95% CI)
Step 2	pH ≤ 7.05	27 (23–31)	89 (86–92)	18 (15–21)
	pH ≤ 7.10	26 (22–30)	90 (87–93)	24 (20–28)
	BE ≤ -12	34 (30–38)	89 (86–92)	17 (14–20)
	BDecf ≥ 12	39 (35–43)	89 (86–92)	11 (8–14)
	Apgar < 7	11 (8–14)	88 (85–91)	3 (2– 4)
Step 4	pH ≤ 7.05	25 (21–29)	95 (93–97)	30 (26–34)
	pH ≤ 7.10	19 (16–22)	95 (93–97)	33 (29–37)
	BE ≤ -12	23 (19–27)	95 (93–97)	21 (18–24)
	BDecf ≥ 12	24 (20–28)	94 (92–96)	12 (9–15)
	Apgar < 7	21 (18–24)	94 (92–96)	12 (9–15)

Table 6.7: Sensitivity (SE), specificity (SP), and precision (PR) of latent class model, M_{r3}, of clinical evaluation for Step 2 and 4. Results for 552 records are presented as pathological evaluation vs. (suspicious + normal).

	objective	SE (95% CI)	SP (95% CI)	PR (95% CI)
Step 2	pH ≤ 7.05	50 (46–54)	78 (75–81)	16 (13–19)
	pH ≤ 7.10	43 (39–47)	78 (75–81)	19 (16–22)
	BE ≤ -12	56 (52–60)	77 (73–81)	13 (10–16)
	BDecf ≥ 12	67 (63–71)	77 (73–81)	9 (7–11)
	Apgar < 7	26 (22–30)	75 (71–79)	4 (2– 6)
Step 4	pH ≤ 7.05	39 (35–43)	93 (91–95)	31 (27–35)
	pH ≤ 7.10	31 (27–35)	93 (91–95)	35 (31–39)
	BE ≤ -12	38 (34–42)	92 (90–94)	22 (19–25)
	BDecf ≥ 12	33 (29–37)	91 (89–93)	11 (8–14)
	Apgar < 7	21 (18–24)	91 (89–93)	7 (5– 9)

In Table 6.8 we present results of sensitivity and specificity for the latent class model M_{r4} for pathological evaluation, see the conditional probability response probability in Figure 6.13. For the pathological class all clinicians more or less agree on the evaluation. However this agreement is not highlighted in better sensitivity and specificity for biochemical markers and Apgar score.

Progression of sensitivity and specificity with respect to pH and BDecf

In Figures 6.17 and 6.18 we present sensitivity and specificity with varying pH and BDecf levels. The results are shown only for MV (similar progression obtained using LCM). The sensitivity and specificity were computed for each step of pH and BDecf, starting from pH = 6.95 (BDecf = 0) up to pH = 7.35 (BDecf = 20) with pH step of 0.025 (BDecf = 1). In Figure 6.17 we can see that specificity

Table 6.8: Sensitivity (SE), specificity (SP), and precision (PR) of latent class model, M_{r4}, of clinical evaluation for Step 2. Results are presented as pathological evaluation vs. (suspicious + normal).

	objective	SE (95% CI)	SP (95% CI)	PR (95% CI)
Step 2	pH ≤ 7.05	23 (19–27)	93 (91–95)	21 (18–24)
	pH ≤ 7.10	23 (19–27)	93 (91–95)	30 (26–34)
	BE ≤ -12	28 (24–32)	93 (91–95)	19 (16–22)
	BDecf ≥ 12	28 (24–32)	92 (90–94)	11 (8–14)
	Apgar < 7	11 (8–14)	92 (90–94)	4 (2– 6)

remains almost constant for the whole range of pH. This almost constant specificity is because of large proportion of normal and suspicious evaluation. The sensitivity is at maximum for pH = 7. From this point it has started to decrease.

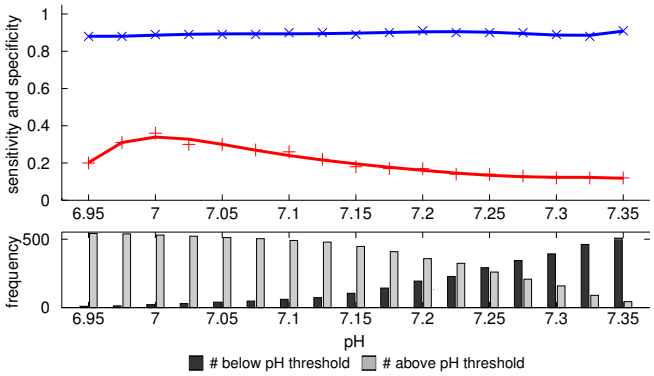


Figure 6.17: Sensitivity (+) and specificity (x) with respect to different pH values. Evaluation of the Step 2. Results presented as pathological vs (suspicious + normal). The proportions of the two classes are shown in the bottom graph.

For the BDecf, see Figure 6.18, the behaviour of specificity is similar to the previous Figure 6.17. The sensitivity is at maximum for BDecf = 13. The obtained maximum value more or less corresponds to pH and BDecf cut-off points described in Section 2.3.2.

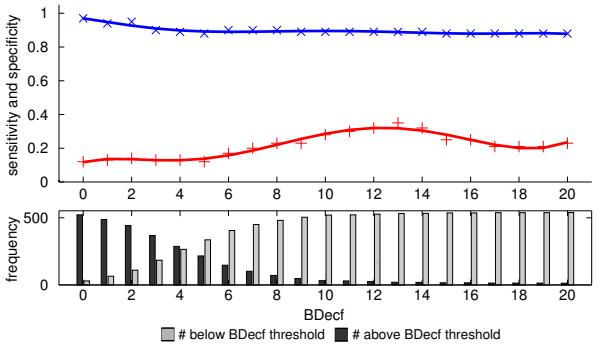


Figure 6.18: Sensitivity (+) and specificity (x) with respect to different BDecf values. Evaluation of the Step 2. Results presented as pathological vs (suspicious + normal). The proportions of the two classes are shown in the bottom graph.

6.6.5 Statistical analysis – FHR features vs. clinical evaluation

We analysed the relationship of extracted features with respect to clinical evaluation. First we removed inter-correlated features in the predefined groups: FIGO-based, HRV-based, and nonlinear. We removed those features having correlation $\mathcal{R} > |0.9|$. Only one representative feature was kept from

the correlated group. The normal distribution in each class (normal, suspicious, and pathological) was tested using the Lilliefors test. Most of the features were found not normally distributed mainly because of pathological class. Next, based on the results of Lilliefors test, we used either ANOVA test or Kruskall-Wallis test. We performed tests for each clinicians and also for majority voting and latent class model M_{r3}. In Figure 6.19 we present probability density for feature baselineMean and evaluation from 4-th clinician. It could be seen that from normal to pathological evaluation the baselineMean increases. Even though the probability density overlaps, the confidence intervals for average values do not overlap (not show in the figure). Hence the null hypothesis was rejected for this clinician.

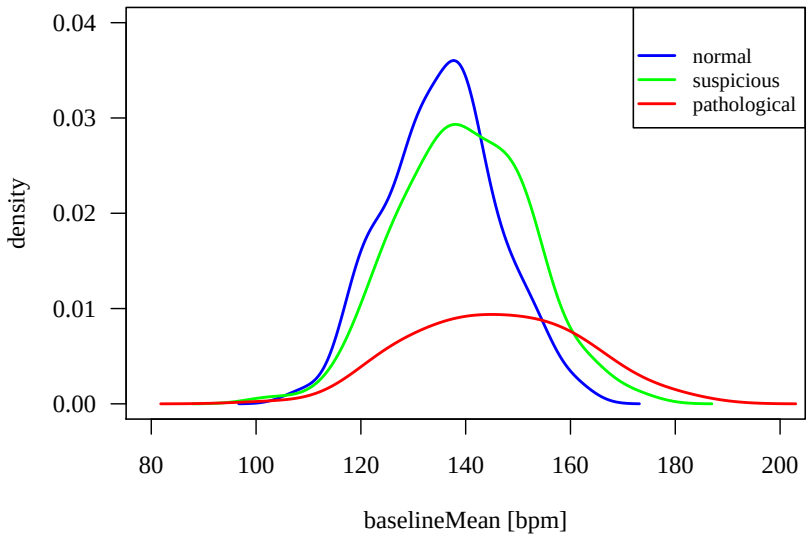


Figure 6.19: Probability density function of feature baselineMean for evaluation from clinician 4.

In Table 6.9 we present only those features for which the null hypothesis was rejected for at least eight clinicians and for both MV and LCM.

Table 6.9: Significant features to clinical evaluation. Features organized by their origin: FIGO-based, HRV-based and Nonlinear.

Feature set	Significant features
FIGO-based	baselineMean, number of decelerations, Δ_{total}
HRV-based	energy04_VLF, energy04_LF, LF/(MF + HF), LF/HF, LTI-HAA
Nonlinear	ApEn(2,0.15), FD_BoxCountP1, FD_HiguchiDl, Poincaré_SD1, Poincaré_SD2

We would like to highlight here an important conclusion that the listing of significant features corresponds to the (Chudáček et al., 2011, Table 2) where features were extracted on different database and with different clinicians but one used for evaluation.

6.7 Discussion and conclusion

In this chapter we described and implemented a new methodology for annotation of CTG records. The clinical annotations were obtained using developed CTGAnnotator software. We gathered clinical evaluation from nine practising clinicians where each clinician evaluated 634 records (approximately 691 hours of CTG records). We provided a comprehensive analysis of clinical evaluation of CTG using common statistical measures as well as using a novel approach of latent class analysis. In terms of number of clinicians and number of evaluated records our study is the largest study that has been performed providing broader insight into clinical evaluation and its variability.

Proportion of agreement and inter/intra observer variability. We showed that there is large intra and inter observer variability, which support the results of previous studies (Blackwell et al., 2011; Vayssiere et al., 2009). Obstetricians using FIGO guidelines or its derivatives struggle with the consistency of their assessment of CTG recordings. The low average intra-observer agreement of 71.5% was mainly because of agreement on pathological class (50%). For the normal class the intra-observer agreement was 80.7%. The large inter-observer agreement is clearly apparent even from the proportion of evaluation (normal, suspicious, and pathological) in Figure 6.3. The inter/intra observer variability is large irrespective of clinicians experience or work place. The detailed view on inter observer agreement offered inter-observer matrix in Figure 6.7, where two clinicians (3,8) are distinct (defensive) in their decision for the CTG evaluation in Step 2 and three clinicians (3,5,8) are distinct for the labour outcome evaluation in Step 4. The overall proportion of agreement (PA) was 48 % (95% CI: 47–50%). There are two contributing factors behind the low PA: i) two clinicians (3,8) evaluated CTG distinctly and ii) the more clinicians asked for evaluation the more heterogeneous the evaluation is expected. The PA with increasing number of clinicians decreased almost linearly as it was shown in Figure 6.6. The direct comparison of inter/intra observer variability to the state of the art publications is rather difficult. In other works the size of the population ranges from 3 (Devane and Lalor, 2005) to 845 (Blix et al., 2003) with 30 to 50 recordings being the most common size e.g. (Bernardes et al., 1997; Keith et al., 1995; Vayssiere et al., 2009). Regarding the number of annotators, again, wide range and professional background of the experts can be found from 3 experts of (Ojala et al., 2008), 28 midwives in (Devane and Lalor, 2005) to 116 ob/gyn residents in (Lidegaard et al., 1992). Generally, the outcome of the studies is presented using the kappa coefficient, of which the use is inappropriate since it can not be compared across different populations. Also the limited space of journal articles does not allow authors to present a comprehensive set of figures offering a clear and simple picture of observer agreement as was shown in this chapter.

Latent class analysis We described a novel approach for the FHR evaluation – the latent class model (LCM). With the LCM model with varying number of classes we contributed to the controversy how many classes should be used for CTG evaluation. We showed that the model has the best fit for 4-tier classification. The difference between 3 and 4 classes lied in better separation of pathological records from the other ones. In other words, there is a clear pathological group for which there is good agreement among clinicians; for the other classes the evaluation is more diverse and splitting these classes to more and more finer classes did not contribute to better model fit and lower clinicians variability.

For the latent class model we assessed the contribution of each clinicians using proposed scoring function. For the M_{r3} model the clinicians were ranked in the following order (score $\mathcal{S}_{acc}$ presented in brackets): 4 (0.456), 5 (0.396), 9 (0.274), 7 (0.198), 8 (0.195), 2 (0.184), 1 (0.182), 3 (0.008), and 6 (-0.045).

Stability of majority voting We proved that even with a high number of clinicians the consensus (simple majority voting) can not be reached. The unreached consensus is because of large inter/intra observer variability. We employed a novel approach using the latent class model, which achieved better consensus than the majority voting. A large improvement was obtained for the pathological evaluation.

Sensitivity and specificity of clinical evaluation The sensitivity (SE) of clinical evaluation to pH, BE, BDecf, and Apgar score at 5 min. was low. In Figure 6.17 it was shown that the pathological evaluation was scattered across the whole pH range. The highest SE for majority voting was achieved for BDecf $\geq$ 12. The results were: SE: 39% (95% CI 35 – 43), specificity (SP): 89 (95% CI 86 – 92). The LCM model, M_{r3} improved SE (67%, 95% CI 63 – 71) but at the price of lower SP (77%, 95% CI 73 – 81). The precision (PR) was slightly better for the majority voting than for LCM model. In general the SE was higher for the Step 2 than for the labour outcome (Step 4) since at this step clinicians rather tend to assign normal class. To the best of our knowledge there is no other work

that presents results of SE and SP to clinical evaluation according to FIGO guidelines. Most of the works evaluates only single patterns: baseline, accelerations, decelerations (Cahill et al., 2012; Donker et al., 1993; Valensise et al., 1997). The different guidelines were compared in (Tommaso et al., 2013) though they used a different pH threshold value ($pH \leq 7.15$) and did not use FIGO guidelines. They concluded that the best SE and SP were achieved for NIHCD guidelines, SE 67%, SP 92%, and Parer & Ikeda system (Parer et al., 2009), SE 55% and SP 67%. The FIGO guidelines were compared only using a computer system (Schiermeier et al., 2008b) with SE 85% and SP 22%. The results are in contrast to ours, highlighting the fact that clinicians do not strictly follow the guidelines and rather use their experience in evaluation.

Statistical analysis – FHR features vs. clinical evaluation We showed that clinical evaluation is statistically significant to clinical features: baselineMean and number of decelerations. The statistical significance was not proved for a number of acceleration and for all of the short term variability indices. Intuitively, the short term variability could not be assessed visually; hence it is unlikely to be significant. On the contrary, the quantity of significant non-linear features suggest that the "intuition" based part of the decision process is rather large and there is a contributing factor of pattern-like memory acquired during working experience. The important finding is the correspondence of statistically significant features with our previous work (Chudáček et al., 2011) that used a different database and different clinicians but one.

Chapter 7

Classification using the pH

This chapter follows the most used "traditional" approach of fetal heart rate classification where a pH value is used to discriminate between normal and abnormal recordings. This approach has several advantages and disadvantages. We discuss them in context and detail at the end of this chapter. Briefly speaking, the "traditional" approach is simple and easy reproducible at the first glance. However, in almost every work a different pH level is used and almost every research work is performed on different, usually small (50-100 recordings), database. Another disadvantage of the "traditional" is the relationship between CTG and pH, which is not fully understood and, finally, it is not natural that there would be a simple separating point between the normal and abnormal (pathological) group. For instance, when choosing pH threshold to be pH ≤ 7.10, the probability the two fetuses, one having pH = 7.09, the other having pH = 7.11, belong to two different groups is small.

In general the main idea of this chapter is not novel many of the building blocks were used before. We followed the traditional approach in order to gain an insight into the new experimental database (CTU-UHB database) and provide results regarding the pH classification. Our study is the largest study that presents results of sensitivity and specificity of FHR classification. This study also evaluates behaviour of almost a complete set of features used for FHR analysis on the largest CTG database. The results are evaluated in terms of ability to discriminate normal (pH > 7.10) and abnormal (pH ≤ 7.10) fetuses. The level of pH threshold was thoroughly discussed in Section 2.3. The results are presented in the context of the three published papers: first, analysis of features and their link to the degree of hypoxia – estimated by pH value (Spilka et al., 2012, 2013b). second; investigation of useful features suitable for mimicking obstetricians evaluation (Chudáček et al., 2011).

7.1 Correlation of features

The correlation of features could provide the first insight into a relationship between them. Even though the correlation assess linear dependence it is useful method to discover redundancy in the feature set. There are many ways to compute correlation coefficient each one suitable for different task. We choose Pearson correlation coefficient for its general applicability. Let consider two features x_1 and x_2 with N examples written as $x_1(i)$ and $x_2(i)$ where $i = 1, 2, \ldots, N$. The population correlation coefficient is defined as

$$\rho = \frac{cov(x_1, x_2)}{\sqrt{var(x_1)var(x_2)}},$$

where cov stands for the covariance and var for the variance. The estimate of correlation coefficient is given by

$$\hat{\rho} = \frac{\sum_{i=1}^{N}(x_1(i) - \bar{x}_1)(x_2(i) - \bar{x}_2)}{\sqrt{\left(\sum_{i=1}^{N}(x_1(i) - \bar{x}_1)^2\right)\left(\sum_{i=1}^{N}(x_2(i) - \bar{x}_2)^2\right)}},$$

where $\bar{x}_1$ and $\bar{x}_2$ are mean values of x_1 and x_2, respectively. The coefficient ranges from -1 to 1. For coefficient values of $[-1, 0)$ the two variables have negative correlation and for values of $(0, 1]$ they have positive correlation. When $\hat{\rho} = 0$ the x_1 and x_2 are uncorrelated. The two variables have the so called complete positive correlation if $\hat{\rho} = 1$. The following hypothesis are used to test statistical significance of correlation

$H_0 : \hat{\rho}(i) = 0$, there is no correlation between the features.

$H_1 : \hat{\rho}(i) \neq 0$, there is a correlation between the features.

To perform t-test we compute t value as

$$t = \hat{\rho}\sqrt{\frac{N-2}{1-\hat{\rho}^2}}.$$

Then using t we find p-value for $N - 2$ degrees of freedom. The p-value is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data. We used the significance level $p = 0.01$. Next, we created a correlation matrix that describes relationship of all measured features by correlation coefficient. We also included the pH values and observed, which features are best correlated to the pH.

7.2 Statistical analysis of features with respect to pH

We perform similar testing as in Chapter 6 (Section 6.5) though, here we use pH value instead of clinical evaluation. The methodology of statistical testing is the same. All features were tested for normal distribution with Lilliefors test. For the normally distributed features we used Analysis of variance (ANOVA) and for not normally distributed we used non-parametric Kruskal-Wallis test. We tested the null hypothesis that features comes from the same distribution against alternative hypothesis that they do not. The null hypothesis was rejected when $p < 0.01$.

7.3 Feature selection

Feature selection (FS) reduces the input dimensionality, because in real world applications we tend to extract more features than necessary in an effort to include all possible information. However, sometimes some of the extracted features can be correlated, hence redundant information is likely to be included or sometimes some features are irrelevant to the application at hand and may negatively affect the performance of the classifier. The term performance refers to the discriminative capability of a classifier.

Let N is number of examples and $\mathbf{x}_i$ is a feature vector, $\mathbf{x}_i \in \mathbb{R}^d$. In feature selection a search problem of finding a subset of l features from a given set of d features, $l < d$ has to be solved in order to optimize a specific evaluation measure, i.e the performance of a classifier. There are number of approaches that try to tackle this problem, which could roughly be divided into three categories: filters, wrappers, and embedded methods (Guyon et al., 2006). The filter approach ranks features based on a performance evaluation metric calculated directly from the data, the wrapper approach employs a predictive model and uses its output to determine the quality of the selected features, and the embedded approach integrates the selection of features in model building. In our work we used the simplest filter approach since we did not want to be restricted to a particular classifier. We incorporated several filter techniques and then used feature meta-selection. A feature was selected when a majority of methods selected this feature. Note that because of filter approach we restricted each method to select at most ten best features. We employed a simple measure to assess the performance of feature selection methods. In each fold of cross-validation the individual methods created candidate sets of selected features $\{S_i\}_{i=1}^{E}$, where E is a number of FS methods used. Then a meta-selection was applied

to create the resulting set S_{ms}. In order to assess the feature selection methods we evaluated the intersection $S_{ms} \bigcap S_i$. The cardinality of intersection provides information how a particular method was used in the feature selection.

We worked with features separated based on their "origin" into following groups: FIGO-based (features based on FIGO-guidelines), HRV-based (features inspired by adult HRV analysis), Nonlinear, and complete that contained all of the features. The details on the division can be found in Section 5.4. The purpose of features division based on their "origin" was to prove that within each group there are features with information value and, hence, their computations were performed correctly.

Below we present the five feature selection methods that were used for selecting the most appropriate features.

Information gain

The information gain evaluates an attribute by measuring the amount of information gain with respect to a class. The mutual information termed $\mathrm{InfoGain}$, is computed using entropy H:

$$\mathrm{InfoGain}(Class, Attribute) = H(Class) - H(Class|Attribute)$$

Correlation based feature selection

Correlation based feature selection (CFS) uses a heuristic evaluation function to rank feature subsets. This algorithm chooses features that are in strong relationship with a class while having low intercorrelation (Hall, 1998). It is similar to already used correlation coefficient but with difference that subsets of features are used.

Maximum relevance and minimum redundancy

The maximum relevance and minimum redundancy (mRMR) (Peng et al., 2005) algorithm finds a feature set S with l features from the set of all features U, where $S \subseteq U$. It attempts to maximize relevance (features' dependence on target class) while minimizing redundancy (excluding features with same information value). The relevance could be determined by correlation coefficient, as described above, or using mutual information, also described above using entropy. The mutual information between two features x_1 and x_2 is defined as

$$I(x_1, x_2) = \sum_{x_1} \sum_{x_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)},$$

where $p(x_1, x_2)$ is the joint probability distribution and $p(x_1)$ and $p(x_2)$ are the marginal probability distribution functions. A maximally relevant feature $\mathbf{x}_i$ has the largest mutual information with a class c, $I(\mathbf{x}_i, c)$. The global measure of relevance (dependence) $D(S)$ with respect to c is defined as

$$D(S) = \frac{1}{|S|} \sum_{\mathbf{x}_i \in S} I(\mathbf{x}_i, c).$$

The global measure of features redundancy in S is defined by

$$R(S) = \frac{1}{|S|^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \in S} I(\mathbf{x}_i, \mathbf{x}_j).$$

The maximum relevance minimum redundancy algorithm combines D and R to optimize the both of them simulatneously. To find a good subset of features a maximum is sought

$$S^* = \underset{S \subseteq U}{\mathrm{argmax}}(D(S) - R(S)).$$

Relief

Relief (RELevance In Estimating Features) (Kira and Rendell, 1992) is a popular feature selection algorithm based on a weight vector over all features, which is updated according to the sample points presented (the higher the weight the better the feature). The scoring function for binary class can be expressed as

$$R_r(\mathbf{x}_i) = \frac{1}{2}\sum_{j=1}^{N}\left(x_i(j) - nearmiss(\mathbf{x}_i))^2 - (x_i(j) - nearhit(\mathbf{x}_i)\right)^2),$$

where N is the number of examples and $nearmiss(\mathbf{x}_i)$ and $nearhit(\mathbf{x}_i)$ denote the nearest point to $x_i(j)$ from $\mathbf{x}_i$ that belongs to the other and the same class, respectively.

Fisher score

The Fisher score (Duda et al., 2000) is widely used feature selection technique. It is related to principal component analysis (PCA), which will be described in detail in Chapter 8. The PCA finds projection of d dimensional space into l dimensional by minimizing the quadratic mean square error. The Fisher score selects features such that in the space spanned by l features, the distance between data points from different classes is maximal while the distance from points within class is minimal. The PCA uses a combination of features while the Fisher selects (ranks) features individually. The Fisher score is computed as follows

$$F(\mathbf{x}_i) = \frac{\sum_{c=1}^{C} n_c(\bar{\mathbf{x}}_{i,c} - \bar{\mathbf{x}}_i)^2}{\sum_{c=1}^{C} n_c\sigma_{i,c}^2}$$

where n_c is number of examples for class c, $\bar{\mathbf{x}}_i$ is the mean of feature $\mathbf{x}_i$, $\bar{\mathbf{x}}_{i,c}$ and $\sigma_{i,c}$ are the mean and variance of feature $\mathbf{x}_i$ for class c, respectively.

Implementation The Information gain, Correlation based feature selection, and Relief were implemented in WEKA (Witten and Frank, 2005) and wrapped in Matlab code (Zhao et al., 2010). The Fisher score was implemented in (Zhao et al., 2010) and the Maximum relevance minimum redundancy in (Peng et al., 2005) and used from (Zhao et al., 2010).

7.4 Classification

Recall that by a classification we consider a classical learning task of finding a function that maps feature space $\mathcal{X}$ to class labels $\mathcal{Y}$ as follows $f : \mathcal{X} \to \mathcal{Y}$. The f should generalize well on unseen data. Below we a present technique for balancing the dataset and description of three classifiers: Naive Bayes, Support Vector Machine, and C4.5.

7.4.1 Imbalanced data

The data we are using are strongly imbalanced. The abnormal (pathological) class is heavily under-sampled in comparison to the normal one. This creates an extra challenge to the already difficult task of fetus well-being diagnosis. The class imbalance is a fundamental problem arising when pattern recognition methods are dealing with real life problems and many approaches have been proposed to overcome this situation (Chawla et al., 2004; He and Garcia, 2009). There are two dominating approaches for handling the class imbalance: i) sampling methods that under-sample the majority class or oversample the minority class, ii) cost sensitive learning when a penalty of misclassification of minority class is higher than misclassification of majority class. We used the sampling method Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) because of its general

applicability and based on our previous experience (Spilka et al., 2010, 2013b). For other methods refer to (He and Garcia, 2009) and references therein.

Synthetic Minority Oversampling Technique (SMOTE) is popular technique to compensate for imbalance in data. It operates on the minority class creating artificial data. SMOTE is based on real data belonging to the minority class and it operates in the feature space rather than the data space (Chawla et al., 2002). The algorithm for each instance (in feature space) of the minority class introduces a synthetic example along any/all of the lines joining that particular instance with its k nearest neighbours that belong to the minority class. Usually after the SMOTE the training set has approximately equal numbers of examples in each class.

7.4.2 Classifiers

The features sets were used to train the following classifiers: Naive Bayes, Support Vector Machine (SVM), and C4.5 decision tree. For more information about classifiers, see e.g. (Duda et al., 2000). The Naive Bayes and SVM were implemented in Pattern Recognition Toolbox (PRTools)[1], version 4.1.10 and C4.5 (J48) was implemented in WEKA (Witten and Frank, 2005) and used from (Zhao et al., 2010).

Naive Bayes

Naive Bayes classifier is based on the Bayes theorem where posterior probability is computed as

$$p(Y|X_1,\ldots,X_d) = \frac{p(Y)p(X_1,\ldots,X_d|Y)}{p(X_1,\ldots,X_d)},$$

where Y is the class variable and $X_1,\ldots,X_d$ are features. The Naive Bayes uses this theorem but with strong (naive) assumption that features are conditionally independent given the class. Therefore, we can estimate posterior probability as

$$p(Y|X_1,\ldots,X_d) = \frac{p(Y)}{p(X_1,\ldots,X_d)}\prod_{i=1}^{d} p(X_i|Y).$$

Then, to minimize error classification, we chose the decision rule with maximum a posterior probability referred as MAP

$$Y_{MAP} = \operatorname*{argmax}_y p(Y=y)\prod_{i=1}^{d} p(X_i = x|Y=y).$$

Decision tree – C4.5

The C4.5 algorithm was proposed by (Quinlan, 1992) and is used to generate a decision tree. Generally, decision tree divides data into subgroups where it is desired that one class prevails in each subgroup. The C4.5 employs the concept of information entropy for choosing an attribute. First, we create root of tree using the attribute with highest information gain (difference in entropy). Then we make decision and move to the sub-lists of the tree until every example is covered. The decision tree is prone to over-fitting, hence it has to be pruned in order to ensure the generalization capabilities. The C4.5 utilises error based pruning. The algorithm goes backwards and removes branches that do not help towards the goal by replacing them with leaf nodes.

[1] `http://www.prtools.org/`

Support Vector Machine

The main purpose of Support Vector Machine (SVM) is to minimize the structural risk, i.e. the risk of error prediction on unseen data. Given N observations and input set, $\mathcal{D} = \{\mathbf{x}_i, y_{i=1}\}^N$, where $\mathbf{x}_i$ are features and $y_i = \{-1, 1\}$ the resulting class. The SVM algorithm searches a hyperplane $\mathbf{w}$, which maximize the distance (margin) between the hyperplane and instances closest to it (Vapnik, 1995). These instances are called support vectors. The goal is to find a vector $\mathbf{w}$ and a constant b such that they satisfy the constraints

$$\mathbf{w}^T\mathbf{x}_i + b \geq +1, \quad y_i = 1, \tag{7.1}$$

$$\mathbf{w}^T\mathbf{x}_i + b \leq -1, \quad y_i = -1, \tag{7.2}$$

and the vector has the smallest norm, i.e. solution minimize $||\mathbf{w}||^2$. The problem can be solved either in primal space (space of parameters $\mathbf{w}$ and b) or in dual space (space of Lagrange multipliers). The purpose is to find α_i that are solution of dual optimization task

$$\alpha_i = \underset{\alpha_i}{\operatorname{argmax}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j,$$

under the constraints $\sum_{i=1}^{N} \alpha_i y_i = 0$ and $\alpha_i \geq 0$. Points that satisfy condition $\alpha_i > 0$ are called support vectors and determine the hyperplane. Classification of an instance $\mathbf{x}$ is then obtained as the $sign$ of following function

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b\right).$$

The features can be mapped into another space using a kernel function $\varphi()$. For nonseparble case the so called slack variables ξ_i can be introduced that allow the margin constraints to be violated, for details on different choice of kernel functions and slack variables, cf. (Vapnik, 1995). In our experiments we used the radial basis function kernel $\gamma = 1/2\sigma^2$.

7.4.3 Performance evaluation

The classification task is to generalize well on unseen/independent data. A classifier is learned on training/learning data and then tested on data that has not been used for learning (unseen test data). There exist many measures to assess performance of a classifier and a lot of techniques to create training and test data in order to estimate generalization ability of a classifier on test (unseen) data. For overview of measures suitable for imbalanced data refer to (He and Garcia, 2009) and for overview on error estimation techniques refer to (Dougherty et al., 2010). In this section we briefly introduce the most common measures (sensitivity, specificity, precision, and F-measure) and the most common methods for error estimation (hold-out sample, cross-validation, and leave one out cross validation).

Statistical measures

The most common form to represent performance is by a confusion matrix shown in Table 7.1. In the confusion matrix, TN (true negative) expresses number of correctly classified negative examples, TP (true positive) is number of correctly classified positive examples, FN (false negative) is number of incorrectly classified negative examples, and FP (false positive) is the number of incorrectly classified positive examples.

The overall classification *accuracy* can be computed as $a = (TP + TN)/TP + FP + TN + FP$. This could be further divided into accuracy observed separately on positive examples (sensitivity) $SE =$

Table 7.1: Confusion matrix. p/n – actual positive/negative, p′/n′ – predicted positive/negative.

	p′	n′
p	TP	FP
n	FN	TN

$TP/(TP + FN)$ and accuracy observed only on negative examples (specificity) $SP = TN/(FP + TN)$. When dealing with imbalanced dataset, the negative class has usually larger proportion than the positive class and the sensitivity is not good measure to assess the performance on the negative class. The number of TN is usually high thus masking the number of FP. For example when negative class amounts for 95 examples and positive class for 5 examples. The results of classification in a form of confusion matrix can be: $TP = 5, TN = 90, FN = 0, FP = 5$. Then the sensitivity and specificity would be 100% and 95%, respectively. Nevertheless, the positive classification has only 50% accuracy. To better assess the FP a precision (positive predictive value) could be used, where $PR = TP/(TP + FP)$.

In order to combine sensitivity (also referred as recall (RE)) and precision to one number a harmonic mean can be used. The harmonic mean is referred to as F-measure

$$F_\beta = \frac{(1 + \beta^2)(PR \cdot RE)}{\beta^2 \cdot PR + RE},$$

where the parameter β is usually set to one. The F-measure penalizes the low numbers of PR or RE and, therefore, better assess the performance on positive and negative class.

Other approach for classifiers comparison is to use receiver operation characteristic (ROC) or simply ROC curve. The ROC curve is shown in Figure 7.1. In this graphical plot the sensitivity is plotted as a function of (1 - specificity).

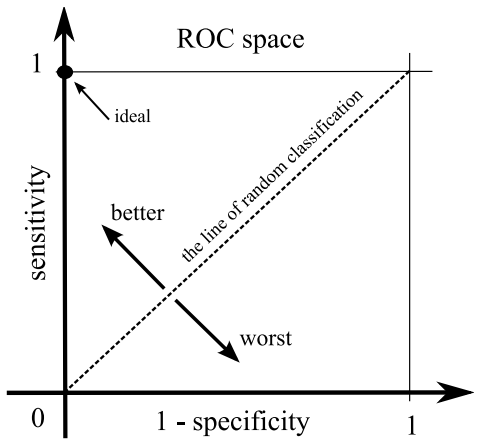


Figure 7.1: The receiver operation characteristic. Sensitivity is plotted as function of (1 - specificity). Ideal classifier is marked in the upper left corner with coordinates (0,1)

The line of random classification (random guess) is a straight line at a 45° diagonal. Successful classifier is placed above this line and tends to upper left corner with coordinates (0,1) That is, all positive examples are classified correctly and no negative example is misclassified. Another useful value for comparison is area under the ROC curve (AUC). The AUC expresses probability that classifier rank randomly chosen positive instance higher than randomly chosen negative instance.

The similar plot offers the so called PR curves where recall (sensitivity) is plot as a function of precision (positive predictive value) (Davis and Goadrich, 2006).

Error estimation

A classifier should be general enough to perform well on unseen data. e.g. when a classifier is implemented in real application and is used to classify a new data. The generalization ability of a classifier could be estimated using various techniques (Dougherty et al., 2010) where the test ("unseen") data are used to estimate classification performance (error).

The computed classification error is an estimate of the true (generally unknown) error. There are two related properties of the estimated error: bias and variance. The bias is a difference between estimated value and true (unknown) error value. The low bias means that we accurately estimate the true error. The variance express how the estimated error changes under different training sets. Intuitively bias and variance are tight together and we are seeking a technique, which estimates the error on unseen data with low bias and low variance.

The largest bias and lowest variance provides the hold-out sample method where data set is split into training and test set. This method has a disadvantage that a portion of data is not used for learning. The cross-validation (CV) method overcomes this problem. The CV has determined number of folds, e.g 10-fold CV. In each step, the data set is divided into training and test data. Then a classifier is learned on training set and performance evaluated on test data. This procedure is repeated in each fold of CV with differently divided data sets. The CV is commonly used and recommended technique to estimate classification performance (Kohavi, 1995). The bias and variance for CV is different for different number of folds. Kohavi (1995) recommended to use 10-fold CV for its general applicability. When number of folds is equal number of examples the CV is called leave-one-out CV where in each fold on example is left as test set and the rest is used for learning. The leave-one-out CV is unbiased but can have high variance since the training data are very similar to each other.

Statistical testing

The difference between individual classifiers trained using different feature sets should be statistically confirmed. Although there is no unified framework the use of McNemar's test is recommended (Salzberg, 1997). However, when dealing with a relatively small dataset there is not enough data to acceptably minimize both errors: (i) when estimating classification performance, (ii) in statistical testing. Apparently, the better way is to minimize the former error and refrain from statistical testing. In addition, a statistical comparison is usually needed when introducing a new classifier where the new classifier is compared against other classifiers on various datasets – a scenario not pursued in this work.

7.5 Proposed experimental methodology

We worked with the CTU-UHB database described in Chapter 4. Based on the literature review we considered recordings as abnormal when $\mathrm{pH} \leq 7.10$ (61 records), and normal when $\mathrm{pH} > 7.10$ (491 records), for more details on chosen pH threshold refer to Section 2.3.2.

The features described in Chapter 5 were extracted on the last 60 minutes of the first stage of labour. The extracted features were represented by feature set and served as an input to the procedure depicted in Figure 7.2. The procedure consists of $50 \times q$-fold cross validation (CV) where data were 50 times randomly split for q-fold CV and results for each run were aggregated. Since the SMOTE technique involves random sampling we executed the training branch (SMOTE → feature selection → classifier learning) 5 times. For the sake of simplicity this inner loop is not shown in the procedure in Figure 7.2. The prediction of a classifier, in the inner loop, were grouped using majority voting.

In our experiments we first sought for such q when bias and variance were minimal. We found the best number of folds to be $q = 4$ and used it for all experiments. Note that in this initial experiment we did not use the SMOTE for balancing the training data and also used only the Naive Bayes classifier. The 50×4-fold CV was used for all feature groups (FIGO-based, HRV-based, nonlinear, complete set). The feature set was 50 times split for 4-fold CV. For the training set and the SMOTE was applied in

order to approximately balance the number of normal and abnormal examples (9 nearest neighbours were used to create new samples along 6 randomly chosen lines). Then the features were selected using five different methods. The feature meta-selection consisted of choosing those features that were selected by a three methods, at least. The selected features were used to train Naive Bayes, SVM, and C4.5 classifiers. Then, performance of classification was estimated on the test set.

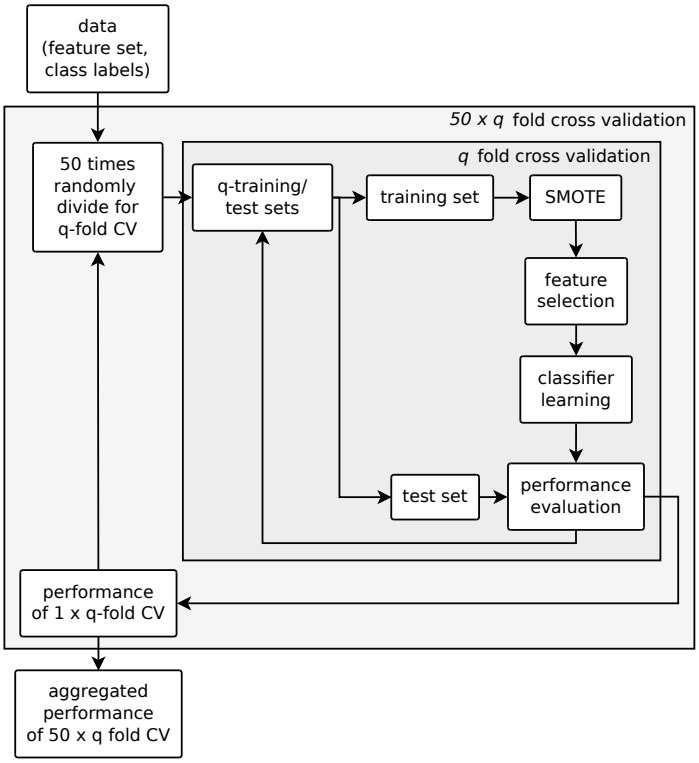


Figure 7.2: Experimental methodology for classification, where q is the number of cross-validation folds.

7.6 Results

7.6.1 Feature correlation

First, we examined inter-correlation between features and with features and pH values. The results form the correlation matrix where rows and columns are equal to number of selected features plus pH. To be able to better distinguish data we picture the correlation matrix as an image where colour scales and shape of ellipses are used to symbolize the values of correlation coefficient $\hat{\rho}$. In Figure 7.3 we present correlation matrix for representative features and pH.

The features originating from different domains: (morphological, time, frequency, and nonlinear) were presented in Chapter 5. In general, we can conclude that features are more or less correlated in the domain they operate. There is interesting correlation between Sonicaid and LTI-HAA, the former represents short term variability while the later long term irregularity. Also these features have good correlation to energy in spectral bands. The ApEn(2,0.2) and ApEn(2,0.15) are in strong correlation, as expected. They also correlate with SampEn(2,0.2) and SampEn(2,0.15), not show in the figure. The interesting features are baselineMean and LZC. They have strong inter-correlation but with other features the correlation (except the correlation of LZC and STV-HAA) is much lower.

In general the correlation of features and pH was low. The highest positive correlation $\hat{\rho} = 0.18$ was found for ApEn(2,0.15) and lowest negative correlation $\hat{\rho} = -0.29$ for energy03_LF. These correlations were statistically significant $p < 0.01$.

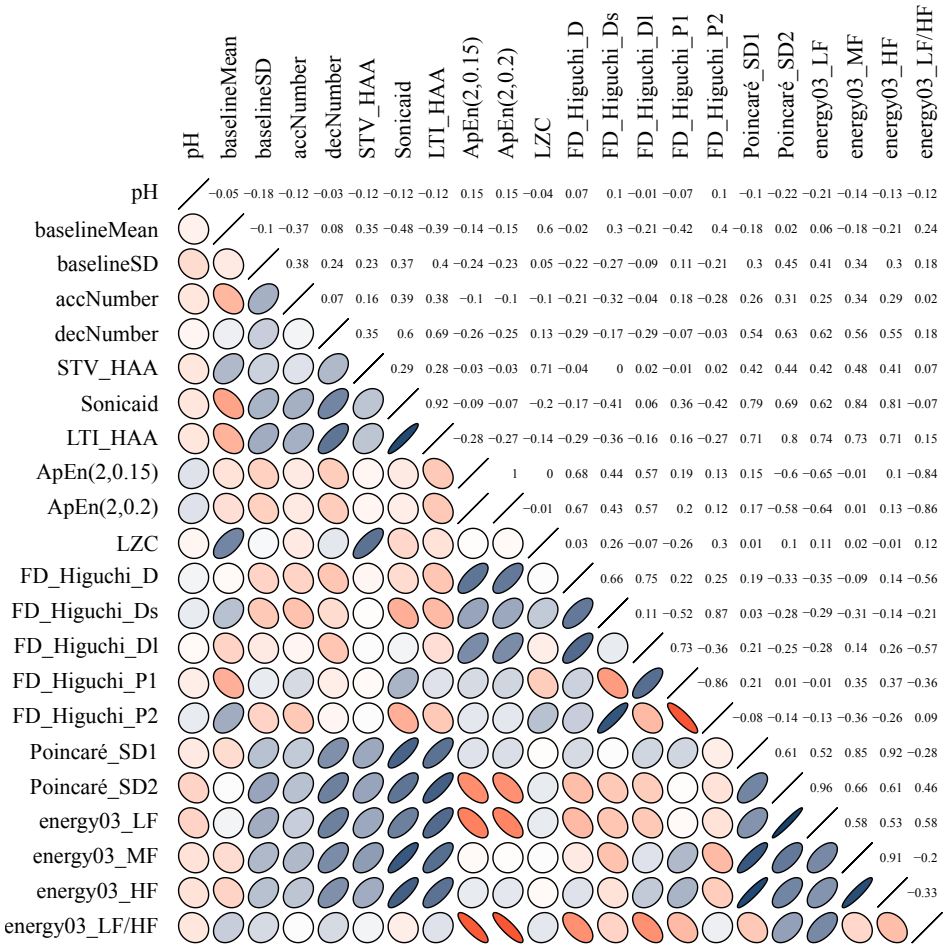


Figure 7.3: Correlation matrix (map) of representative features. The blue colour marks positive correlation while the red colour marks negative correlation. The scale of colour and shape of ellipses represent the value of correlation. The darker the colour and thinner the ellipse the more correlated features.

7.6.2 Statistical analysis of features

In Table 7.2 we present only those features for which the null hypothesis was rejected on significance level $p < 0.01$. To keep the table brief we removed those features that had inter-correlation $\hat{\rho} > |0.9|$ and kept only one representative from such correlated group of features. For example ApEn(m, r) and SampEn(m, r) were highly correlated, therefore we included only ApEn(2,0.15) in the table.

Table 7.2: Significant features to pH ≤ 7.10. Features organized by their origin: FIGO-based, HRV-based, and Nonlinear

Feature set	Significant features
FIGO-based	Δ_{total}
HRV-based	STV-HAA, Sonicaid, energy04_VLF, energy04_LF, energy04_MF, energy04_HF, energy03_LF/HF
Nonlinear	ApEn(2,0.15), Poincaré_SD1, Poincaré_SD2

The number of significant features was high (in total 18 but only 11 representative features are presented in Table 7.2). It is interesting that, from FIGO-based features, only long term variability, Δ_{total}. was significant to the pH. The highest number of significant features was from frequency

domain. Even-though the frequency features are correlated the correlation did not reach the $\hat{\rho} > |0.9|$ threshold. In Figure 7.4 we provide density of energy03_LF regarding to pH. Because of lower proportion of abnormal cases with pH ≤ 7.10 the density is smaller. In Figure 7.5 we sub-sampled the normal class to have equal proportion as abnormal class. The confidence intervals for the normal class were estimated using bias-corrected and accelerated bootstrap method (Efron, 1994, 2003). The normal and abnormal class is "separated" but the true normal class density, Figure 7.4, have to be considered.

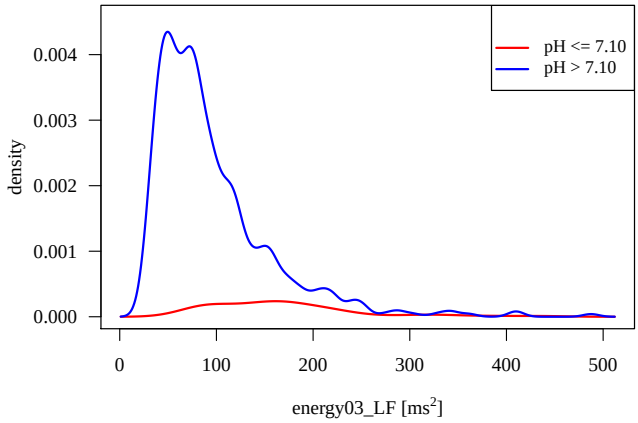


Figure 7.4: Density function of energy03_LF and pH.

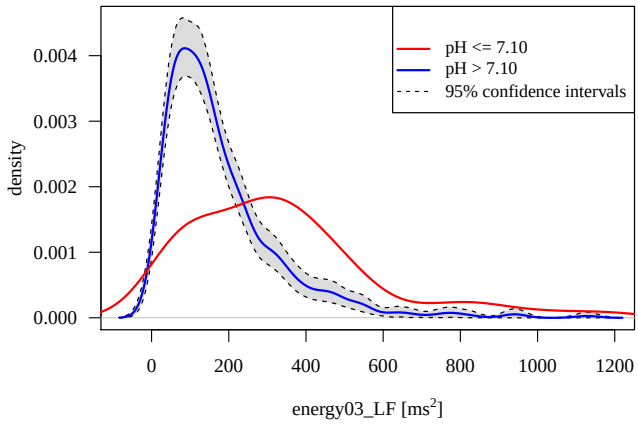


Figure 7.5: Density function normal class (pH > 7.10) and abnormal class (pH $\leq$ 7.10) with equal proportions for feature energy03_LF and pH.

7.6.3 Feature selection/classification

The complete feature set was used to determine the bias and variance of classification performance (sensitivity, specificity, and F-measure) for different number of folds of cross validation (CV), ($q \in \{2, \ldots, 20\}$). In this experiment the normal and abnormal class in training set were not balanced using SMOTE (in order not to be dependent on the SMOTE) and only Naive Bayes classifier was used for learning. The results for sensitivity and specificity are present in Figure 7.6. The variance of estimated error increased with increasing q since very few examples remained in the test set. For example, when $q = 20$, the test set consist only of 57 or 58 examples. Therefore the variance, especially on abnormal class, is very large.

Figure 7.7 provides good example of bias and variance of the estimated F-measure. When $q = 2$ and $q = 3$ there is a bias in the estimated error. From the 4-fold CV to 10-fold the median of F-measure remains almost the same while variance increases. Therefore, for further experiments, we choose the

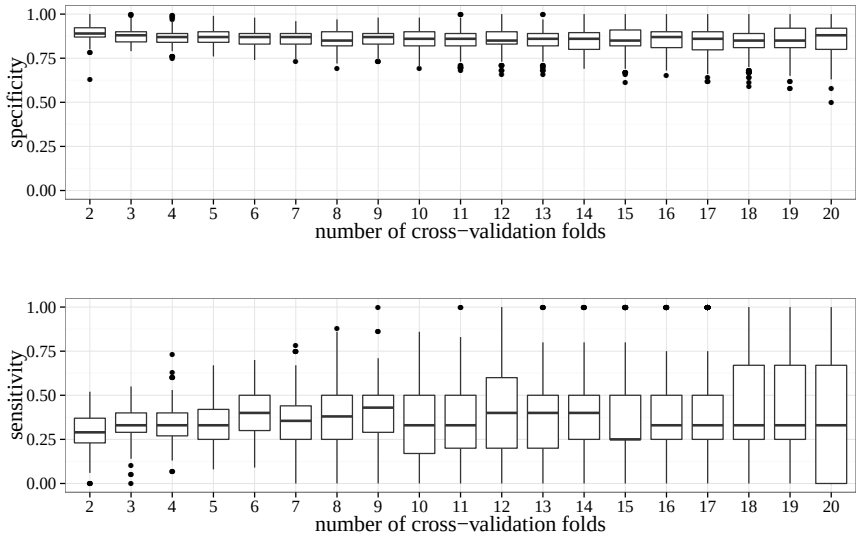


Figure 7.6: Progression of sensitivity and specificity for q-fold cross validation ($q \in \{2, \ldots, 20\}$). Naive Bayes learner with original dataset.

4-fold CV since it provides the lowest bias and variance, c.f. Figure 7.6. Note that the chosen number of folds ($q = 4$) is, again, only an estimate of the best number of folds since the true error is unknown.

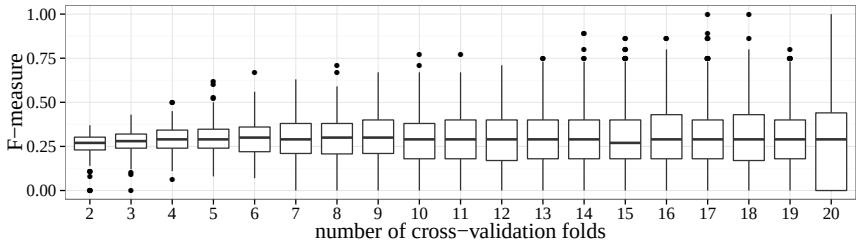


Figure 7.7: Progression of F-measure for q-fold cross validation ($q \in \{2, \ldots, 20\}$). Naive Bayes learner with original dataset.

The features were divided into four groups based on their origin. For each group we performed the 50×4-fold cross validation using procedure depicted in Figure 7.2. The same procedure was done using the whole feature set (termed Complete set hereinafter). In order to evaluate the feature selection methods, we assessed them in each fold of cross-validation. On average the CFS method selected 6 features that were further used for the classifiers learning. Other methods: mRMR, Fisher score, Relief, and InfoGain selected 4,5,6,7 features, respectively. The numbers of selected features are presented for the Complete set only since, for the other groups, the results were similar. In Table 7.3 we present features selected; only those features selected more than 50% of times are included. In the individual groups the features are order based on the number of times they were selected except for FIGO-based group. In this group there is only five features and all were selected in each fold of CV. The feature selection from all available features showed dominance of frequency features, from which very low frequency (VLF) and low frequency (LF) were used in almost every fold of cross-validation.

The distributions of selected features from the Complete set are present in Figure 7.8. The discrimination between normal and abnormal examples is most apparent for the following features: energy04_VLF, energy03_LF, and Poincaré_SD2. The distribution of energy03_LF was already shown in Figure 7.5, here we use another representation using boxplot.

Each feature set was used to train Naive Bayes, SVM, and C4.5 classifiers. The results, shown in Table 7.4, were aggregated from each fold of 50×4 fold CV and median, 25th, and 75th percentiles were estimated. The best results were achieved using the SVM and Naive Bayes with features selected

Table 7.3: Selected features organized by their origin: FIGO-based, HRV-based, Nonlinear, and Complete set.

Feature set	Selected features
FIGO-based	baselineMean, baselineSD, accNumber, decNumber, Δ_{total}
HRV-based	energy04_VLF, energy04_LF, energy03_LF, STV-HAA, energy03_MF, energy04_MF, SDNN
Nonlinear	Poincaré_SD2, SampEn(2,0.15), ApEn(2,0.15), Poincaré_SD1, FD_Sevcik
Complete set (all features)	energy04_VLF, energy03_LF, accNumber, Poincaré_SD2, decNumber, energy04_LF, STV-HAA

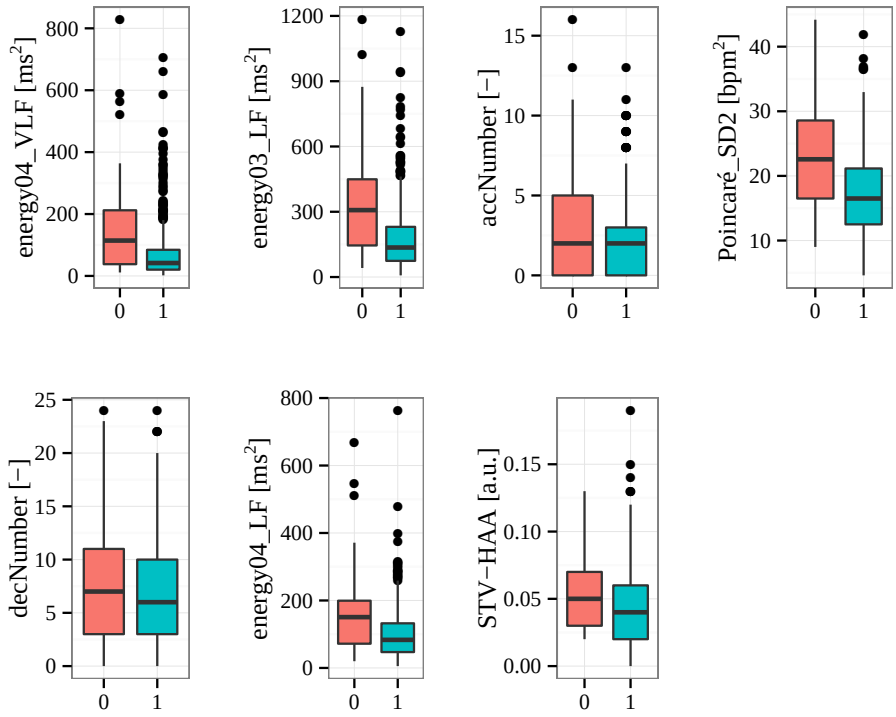


Figure 7.8: Distributions of selected features; 0 – normal class, 1 – abnormal class.

from all available features (Complete set).

7.7 Discussion and conclusion

In this chapter we analysed an almost complete set of features used for FHR analysis. The features originated from different domains (in total 49 more or less distinct features). In contrast to previous works we evaluated the behaviour of features on a reasonably large database. In general, the analysis, feature selection, and classification performed in this chapter are not novel. Many of the buildings blocks were used before though employed on ad-hoc created and small databases. Therefore we followed the classical approach in order to gain an insight into the new experimental database (CTU-UHB database) and provide results regarding the pH classification.

Correlation of features Intuitively, the features were correlated in the domain in which they operate. The only exceptions are LZC and baselineMean, they are strongly correlated but with other features they are not correlated (with the exception of correlation of LZC and STV-HAA). The correlation between

Table 7.4: Classification results for selected features from different groups: FIGO-based, HRV-based; Nonlinear, and Complete set. The results are averaged across all folds of CV (50×4 folds CV) and presented using median (25th – 75th) percentiles. (SE – sensitivity, SP – specificity, PR – precision, F – F-measure).

Feature set	[%]	Naive Bayes	SVM	C4.5 Tree
FIGO-based	SE	38 (31–44)	53 (47–60)	20 (13–27)
	SP	77 (74–80)	68 (66–72)	90 (86–94)
	PR	17 (14–20)	17 (15–20)	19 (13–27)
	F	23 (19–27)	26 (23–29)	19 (13–26)
HRV-based	SE	53 (47–63)	53 (44–60)	50 (40–60)
	SP	74 (71–77)	76 (72–79)	78 (73–82)
	PR	21 (18–24)	21 (18–24)	22 (19–25)
	F	30 (27–34)	29 (25–34)	31 (26–35)
Nonlinear	SE	53 (47–63)	53 (47–63)	31 (20–40)
	SP	67 (63–72)	76 (74–80)	83 (79–85)
	PR	17 (15–20)	22 (19–26)	18 (14–22)
	F	26 (23–29)	32 (27–36)	23 (16–28)
Complete set	SE	60 (53–67)	53 (47–60)	33 (27–47)
	SP	75 (72–77)	78 (75–80)	84 (80–87)
	PR	23 (20–25)	23 (20–26)	21 (17–26)
	F	33 (29–36)	33 (28–37)	26 (21–32)

the features and biochemical marker (pH) was low. The highest positive correlation of $\hat{\rho} = 0.18$ had ApEn(2,0.15) and the lowest negative correlation had energy03_LF, $\hat{\rho} = -0.29$. These correlations were significantly different from zero. Even though the correlation is small, similar results were reported in (Fulcher et al., 2012). They found the correlation coefficient of $\hat{\rho} = -0.28$ for second order coefficient of variation $(\sigma/\mu)^2$ and $\hat{\rho} = -0.28$ for median absolute deviation, $median(|\mathbf{x} - \text{median}(\mathbf{x})|)$.

Statistical evaluation of features We analysed which features are statistically significant to the normal and abnormal recordings determined by pH value. From the FIGO-based features, only long term variability Δ_{total} was significant. The frequency features were the most prevalent type of features significant to pH. About half of the significant features corresponded to the analysis presented in the previous Chapter 6 where we examined features significance with respect to clinical evaluation, c.f. Table 6.9.

Results of classification Our study is the largest study, which shows the results of sensitivity and specificity with respect to pH. We compared a full spectrum of features for fetal heart rate analysis. Three different classifiers were used to discriminate between normal and abnormal recordings for three types of features groups (FIGO-based, HRV-based, and nonlinear). We showed that in each group there are features able to discriminate between normal and abnormal recordings. The worst classification results were obtained for the FIGO-based features and the best results for a combination of all groups (the Complete feature set) with sensitivity (SE) 60% (25th and 75th percentiles: 53–67), specificity (SP) 75% (72–77), precision (PR), 23% (20–25), and F-measure 33% (29–36). The most useful features for classification were frequency features; the very low (energy04_VLF) and low (energy03_LF) were selected in almost every fold of 50×4 CV. These features correspond to activation (dominance) of sympathetic system (part of autonomous nervous systems), which is responsible for an increase in fetal heart rate. The additional features were number of accelerations and decelerations, Poincaré_SD2, and STV-HAA.

Comparison to other works The comparison to other works is difficult if not impossible. We have provided a comprehensive overview of classification performance presented in other works (Chapter 3,

Table 3.3). Most of the works used very small and ad-hoc created databases the typical size is 50 – 100 records. In addition, in every work a seemingly similar, yet different criteria was used to define normal/abnormal (pathological) recordings as shown in Table 3.2. Despite the difficult comparison our results compare favourably to studies using "large" databases ($N \approx 100$). Our results are slightly inferior to (Costa et al., 2009); they achieved SE 57%, SP 97%, and PR 50% but they used a four times smaller sample size. The work of (Georgieva et al., 2013b) is the most appropriate for comparison because of the same pH threshold used. They reported better sensitivity of 61% but lower specificity of 68%, the precision was not used in their work. Achieved precision in our work was low, about 23% for the Naive Bayes classifier. There are two contributing factors. First, the precision in other works is also low, 50% (Costa et al., 2009), 36% (Salamalekis et al., 2006) with the exception of 70% (Spilka et al., 2012). Also the clinical evaluation of CTG has low precision (approximately 20 – 25%) as documented in Chapter 6 (Table 6.6) or 44% as reported in (Cao et al., 2006). Other factor for low precision is the utilization of SMOTE technique where a new minority abnormal samples were created. These new samples possibly overlapped into the normal group causing the false positive results. We performed a simple verification using uniformed sub-sampling of the normal group. We used 500×4 CV where in every repetition (1,2, ..., 500) the normal group was randomly sub-sampled to have equal proportion as the abnormal group. The approach improved the precision but yielded very poor generalization on the test set, a problem well known (He and Garcia, 2009).

Comparison to our previous work Our previous work (Spilka et al., 2012) on presumably large database (217 records) showed promising results, though, the verification of this approach on a larger database has not confirmed the results and provided inferior ones. The reason for worse performance on the current database is unclear. One possibility is different pH threshold. In this work we followed the recent studies (Georgieva et al., 2013b; Yeh et al., 2012) and used pH ≤ 7.10, in contrast to threshold pH ≤ 7.15 in (Spilka et al., 2012). We verified this hypothesis by setting pH level to pH ≤ 7.15. The results were essentially similar to those obtained using the pH ≤ 7.10 (higher sensitivity with similar specificity and precision, detailed results are not present). However, the difference was insignificant. Thus the different choice of pH threshold did not affect the results significantly. We believe the main reason behind the worse performance is inappropriate sampling of the previous database when we focused on acquiring as many abnormal records as possible and aimed to keep normal and abnormal groups balanced. It seems that we neglected the sampling of the normal groups and included too few normal examples. It is tempting to include the previous database (Spilka et al., 2012) and use it, for instance, as the test set. But previously we worked with much shorter signals, whose length is insufficient for some features we are using now, e.g. Detrend Fluctuation Analysis. Also, and more importantly, we did not have information about begin/end of first/second stage of labour. Many works, including our previous papers, do not differentiate between first and second stage of labour. Although in the second stage the shape of CTG is different and the signal has much lower quality (Dupuis and Simon, 2008). Another problem is incomplete clinical information of the previous database (e.g. neonatology, BE, BDecf, etc.) making the post-analysis infeasible.

Issues with the pH (the target class) Regarding the classification results there is an important questions that need to be carefully considered. How to interpret the classification results? The answer is not straightforward since there are distinct approaches to processing and classification of intrapartum CTG. The first one is a more technical approach that uses the objective evaluation (pH, BE, and/or BDecf) of the data. Another approach is subjective evaluation (Apgar score, clinicians assessment of CTG). Yet another approach is to combine the both (objective and subjective) together.

The approaches utilizing objective evaluation (pH) suffers from, at least, two major drawbacks. The relation of hypoxia to the fetal cord arterial pH after delivery is widely discussed in several papers (Cao et al., 2006). The predominant conclusion is that only an overall examination of the baby at about four years of age can bring a confident enough conclusion on the occurrence of effective asphyxia during the delivery. In addition, in many cases where timely interventions based on suspicious/pathological

FHR signal is made, the arterial pH values of the instrumentally delivered baby will be above the pathological threshold.

A second approach is to use subjective evaluation (Apgar score, CTG assessment by clinicians) and use it for the classification process to try to adopt clinicians behaviour. Nevertheless, this approach has several drawbacks as well. First, the inter and even the intra observer variabilities are substantial, as document in Chapter 6. Second, clinicians categorize the signals usually according to FIGO-based guidelines into the three classes (normal, suspicious, and pathological). Large subset of signals are evaluated as suspicious, but suspicious class does not exist after delivery, there is usually normal or (possibly) asphyxiated baby (about whom, there will likely not be any decisive proof for at least the next several months). Third, the sensitivity and precision of clinical evaluation to pH values are low, as was shown in the previous Chapter 6. Hence, simple joining of subjective (clinical evaluation) and objective (pH) might not be appropriate. For instance, whom to trust when pH and clinical evaluation are distinct, e.g. when clinicians are saying that FHR is pathological but the pH is clearly normal? The answer to this question is subject of the next chapter.

Chapter 8

A hierarchical model for FHR evaluation

The main advantage of the fetal heart rate monitoring is a continuous surveillance of fetal well-being. Despite the research efforts the automatic evaluation of fetal status is still not used widely in clinical practice. One of the possible causes is improper and imprecise evaluation of labour outcome by individual markers and their relation to FHR.

In this chapter we propose and implement a novel hierarchical model for FHR evaluation. The model is organized in a hierarchical structure and is made of a combination of biochemical markers, Apgar score, and latent class model of clinical evaluation. This novel model is able to overcome the discrepancies in biochemical markers and also suppress the inter-observer variability of clinical evaluation. The model simultaneously produces the latent class (hidden truth) of labour outcome and a classifier of FHR features. The results of the model are in all measures superior to those achieved using pH value.

This chapter is organized as follows: first, we introduce the unsupervised learning as a tool to discover the underlying unknown structure of a data. Second, we define the most difficult examples for the classical scenario when the pH is used as a discriminator. We show that there are examples constantly misclassified irrespective of classification technique used. Third, we thoroughly describe the disadvantages of different markers used for evaluation of labour outcome. Next, we propose a novel hierarchical model able to suppress the disadvantages of these markers. Last we show the classification performance of the hierarchical model and thoroughly describe the difference between the model and the scenario when a pH is used solely as global (singular) marker for labour outcome.

8.1 Unsupervised learning

8.1.1 Feature extraction/ dimensionality reduction

In the previous Chapter 7 we described a feature selection as a tool for reduction of feature space dimensionality. The performed feature selection required class labels. In the scenario when we either do not have the class labels or we do not want to be restricted to particular choice of classes, we can use unsupervised technique for lowering dimensionality of feature space. One of the most common methods is the principal component analysis (PCA).

Principal component analysis

The principal component analysis (PCA) (Bishop, 1995) approximate data with linear model (system of linear equations) and transform the data into a new coordinate system with lower dimension. Let N is number of observations and d is number of features (typically $d \ll N$), the aim is to lower dimensionality so that each observation can be represented using l variables $1 \leq l < d$. In other words, we want to linearly transform the data into another coordinate system with dimensionality l. The PCA minimize quadratic mean square error and maximize the variance of projected vectors. Below we present PCA by minimizing the mean square error (Hyvärinen et al., 2001). Let $\mathbf{x}_v$ is a column vector

of d features $\mathbf{x}_v = (x_1, \ldots, x_d)^T$ and $\mathbf{b}_1, \ldots, \mathbf{b}_l$ are basis vectors spanning the l-dimensional subspace. Each vector $\mathbf{x}_v$ represents a single observation with d features. The vectors representing individual observations could be arranged in a data matrix $\mathbf{X}$ of size $d \times N$, the basis vectors could be arranged into matrix $\mathbf{B}$ of size $l \times d$. Without loss of generality we assume rows of $\mathbf{X}$ to have zero mean. Further, we assume for basis vectors $\mathbf{b}_1, \ldots, \mathbf{b}_l : \mathbf{b}_i^T \mathbf{b}_j = \delta_{ij}$. The PCA can be viewed as linear transformation

$$\mathbf{y} = \mathbf{B}\mathbf{x}_v$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_l \end{pmatrix} = \begin{pmatrix} b_{11} & \ldots & b_{1d} \\ \vdots & \cdots & \vdots \\ b_{l1} & \ldots & b_{ld} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix},$$

where $\mathbf{y}$ are projected data and $y_i = \sum_{k=1}^{d} b_{ik} x_k = \mathbf{b}_i \mathbf{x}_v$ is i-th projected data. The projection of $\mathbf{x}_v$ on the subspace spanned by basis vectors is

$$\hat{\mathbf{x}}_v = \sum_{i=1}^{l} (\mathbf{b}_i^T \mathbf{x}_v) \mathbf{b}_i.$$

The mean square error to be minimized (with respect to $\mathbf{b}_1, \ldots, \mathbf{b}_l$)

$$\varepsilon^2 = E\left\{ \|\mathbf{x}_v - \hat{\mathbf{x}}_v\|^2 \right\} = E\left\{ \|\mathbf{x}_v - \sum_{i=1}^{l} (\mathbf{b}_i^T \mathbf{x}_v) \mathbf{b}_i\|^2 \right\}.$$

This can be further expressed as

$$\varepsilon^2 = trace(\mathbf{C}_x) - \sum_{j=1}^{l} \mathbf{b}_j^T \mathbf{C}_x \mathbf{b}_j,$$

where $\mathbf{C}_x$ is a covariance matrix. The minimum of ε^2 under orthonormality condition on $\mathbf{b}_i$ is given by any orthonormal basis spanned by the l first eigenvectors.

Choosing the number of principal components The order of l principal components is given by vector of eigenvalues $\boldsymbol{\lambda}$. The physical unit of λ_i is power of i-th component, i.e. its variance. Hence, by sorting the eigenvectors by eigenvalues give us the order (importance) of principal components. Since λ_i represents variance we can choose l in the way that variance of chosen components with respect to overall variance is above some predefined threshold, u:

$$\frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq u.$$

The typical value of u is 0.95 when 95% of variance of the original data is retained. Note that the value of mean square error is equal to $\varepsilon^2 = \sum_{i=l+1}^{d} \lambda_i$, i.e. the sum of eigen values (components) that were not used.

8.1.2 Gaussian mixture model

In Chapter 6 we described a finite mixture model for random discrete variable, eq. (6.6). Here we assume the random variable to be continuous. A general, convenient, practice is to assume the variables are multivariate normal (Gaussian) distribution. The normal density is a convenient choice because of its complete theory, analytical tractability, and natural occurrence. Let X be a continuous random variable. We define the probability density function $f(x)$ as a derivative of a distribution function $F(x)$, that is $f(x) = F'(x)$. Further we assume that $f(\mathbf{x}_i|\boldsymbol{\theta})$ is multivariate probability density given as

$$f(\mathbf{x}_i|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}|\sum|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_i-\mu)^T \sum^{-1}(\mathbf{x}_i-\mu)},$$

where $\mathbf{x}_i$ represents a d-dimensional feature vector, $\boldsymbol{\theta} = \{\mu, \sum\}$, μ is a mean value and $\sum$ is covariance matrix. For the full rank covariance matrix $\sum$ a number of distinct elements has to be estimated $\frac{1}{2}d(d+1)$. As for the general mixture model the log likelihood function is given as

$$\log f(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \sum_{m=1}^{M} \pi_m f(\mathbf{x}_i|\boldsymbol{\theta}_m).$$

where π_m is mixing parameter, the prior probability, for m-th component, $m = 1, \ldots, M$ and $\boldsymbol{\theta}_m$ are parameters for m-th component. For finding a maximum of the log likelihood function we take its derivative with respect to the parameters $\boldsymbol{\theta}$ and equal it to zero

$$\frac{\partial \log f(\mathcal{D}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

Model selection and fit

In Chapter 6 (Section 6.4.4) we described the two most common measures for a model selection: the Akaike information criterion (AIC) (Akaike, 1973) and Bayes information criterion (BIC) (Schwarz, 1978). For the convenience we provide their computation below. For the details on these measures refer to Section 6.4.4. The AIC and BIC are computed as

$$AIC(r) = -2 \ln L + 2\vartheta,$$
$$BIC(r) = -2 \ln L + \vartheta \ln N,$$

where r is number of classes, L is likelihood, N is number of examples, and ϑ is number of estimated parameters. The better model the lower BIC and/or AIC.

8.1.3 Clustering of FHR using Gaussian mixture model

Quantization of fetal behaviour using FHR

The fetal heart rate is the main information channel of the fetal behaviour, which is very complex and during pregnancy undergoes a great changes that are reflected by fetal heart rate (Van Leeuwen et al., 2003, 2013). In this section we aim to infer the underlying structure of FHR and to quantize the FHR features into m-finite states. The clinical "quantization" is performed using the guidelines into three classes (ACOG, 2009; FIGO, 1986; Macones et al., 2008), four classes (Schifrin, 2004), or five classes (Parer and Ikeda, 2007; Parer et al., 2009). Without any restriction to particular guidelines we examine the fetal behaviour using the Gaussian mixture model (GMM) of fetal heart rate features. The GMM is unsupervised technique and this procedure is commonly know as clustering.

The clustering is used to infer underlying unknown structure and to describe/explain the data by a model. Let $\mathbf{x}_i \in \mathcal{X}$ is a d-dimensional feature vector for $i = 1, \ldots, N$, where N is number of examples. Our goal is to describe $\mathcal{X} \in \mathbb{R}^d$ by a model that helps us to better explain the data. In Section 6.4.2 we described a general finite mixture model and above we focused on the special case of mixture model being the Gaussian mixture model. We use this model to represent the feature space $\mathcal{X}$ by mixture of Gaussians. The GMM is described by following parameters (for $m = 1, \ldots, M$ components): mixing proportions α_m, mean values μ_m and covariance matrices Σ_m. The feature space is high-dimensional where individual features are highly correlated or some are even identical. In order to remove correlated features and, more importantly, to reduce the d dimension to lower dimension l, we used the principal component analysis. Using the PCA we linearly transformed the features into a new coordinate system with lower dimension.

Recall that in Section 6.6.3 we sought the appropriate number of classes for clinical evaluation. Here, we search for the number of components, $\hat{m}$ that best represents the feature space $\mathcal{X}$ so that AIC and BIC are minimal: $\hat{m}$ = argmin$_m\{AIC; BIC\}$. We use the procedure shown in Algorithm 3. We note here that finding minimum of AIC and BIC is often a compromise. AIC tends to overestimate $\hat{m}$ while BIC underestimates it.

Algorithm 3: Procedure for finding best model fit for $m = 2, \ldots, M$

Input: l principal components (PCs), maximum number of iterations I = 100, maximum number of components M = 8

Result: $\hat{m}$ optimal number of components

begin
  for $m \in M$ **do**
    for $i \in I$ **do**
      i) $gm \leftarrow$ EM(PCs,m) (run EM algorithm with random initialization)
      ii) $AIC(i, m) \leftarrow gm.AIC$
      iii) $BIC(i, m) \leftarrow gm.BIC$
    end
  end
  v) return $\hat{m} \leftarrow \underset{m}{\operatorname{argmin}}\{\text{median}(AIC); \text{median}(BIC)\}$
end

Clusters with respect to different classes

The best model with the lowest AIC and/or BIC is compared against to different markers (classes) of labour outcome (pH, BE, BDecf, Apgar score) or directly to the clinical evaluation of fetal heart rate. The comparison helps us to link the patterns of FHR, represented by $\hat{m}$ components, directly to classes assessing fetal well-being.

The model was estimated for $\hat{m}$ components using the l principal components. The resulting class for individual examples were assigned using the posterior probability where a class $c \in \{1, \ldots, C\}$ was determined by the largest posterior probability

$$\hat{y} = \underset{c}{\operatorname{argmax}}\, p(c|\mathbf{x}, \boldsymbol{\theta}_c).$$

The given components (classes) were unordered and their order did not correspond to the ordering of the classes based on, e.g. pH (normal, suspicious, and pathological). Knowing the dataset we ordered the components based on their prior probability (the mixing parameter α_m). The ordering corresponded to the distribution of pH in individual classes. The different target classes were determined as follows (order pathological; suspicious; normal) pH: $\{\text{pH} \leq 7.10; \text{pH} > 7.10 \wedge \text{pH} \leq 7.15; \text{pH} > 7.15\}$.

8.2 In search of the most difficult examples

In the previous Chapter 7 we presented results on classification using the target class determined by pH level. The achieved results compared favourably with other works (Costa et al., 2009; Georgieva et al., 2013b) and also with the results of clinical evaluation presented in Chapter 6 (Tables 6.6 and 6.7). Nevertheless, the precision of the classification was very low and sensitivity was inferior to specificity.

The analysis of the results in each fold of 50×4-fold CV revealed that there were examples constantly misclassified. In this section we analyse classification results in more detail and with addition of two independent techniques: unsupervised classification using GMM and latent class model (LCM) of clinical evaluation. We aim to find distinct records that are difficult with respect to classification and/or clinical evaluation. We use the pH as a target class and, in contrast to the previous Chapter 7, we utilize additional suspicious class.

The misclassified examples are called difficult hereinafter. A record is considered as difficult if is simultaneously misclassified in: i) unsupervised learning scenario (clustering of fetal heart rate using GMM), ii) supervised learning scenario (classification using Naive Bayes, SVM, and C4.5), and iii) clinical evaluation of CTG. There are another criteria, which could be considered for definition of difficult examples but, for simplicity, we do not pursue them in this work. A record could be also considered as difficult if: lies near a separating boundary in the supervised/unsupervised learning scenario, the biochemical markers (pH, BE, BDecf) are largely different, i.e. each of the biochemical marker points to a different class, or clinicians do not agree on evaluation (either the votes for different classes are equal or there is a weak plurality).

The difficult records represent the most serious errors in the classification, i.e. errors spanning across one class (misclassified normal to pathological and vice versa).

8.2.1 Unsupervised learning

We used GMM model with $\hat{m}$ components. We estimated the model for the l principal components that accounted for 95% variance in data in order to include as much information as possible into the model but also keep the dimensionality low and reduce redundant information. The learning of GMM was restarted 20 times and a model with the best likelihood was chosen. After the model was estimated, the classes were assigned using posterior probability as we described above for the clustering of FHR (Section 8.1.3).

8.2.2 Supervised learning

In the previous Chapter 7 we used the procedure depicted in Figure 7.2 for classification. In this section we employ this approach but we replace the 50×4 cross-validation (CV) by leave-one-out cross-validation (LOOCV) method. That is in each fold of CV the whole dataset but one example is used for training and the remaining one for testing. In this scenario a classifier is supplied with maximum of information. We used the three classifiers (Naive Bayes, SVM, and C4.5) and combined their predictions by majority voting. It is possible that results could be biased to the used methods (feature selection and classification). Therefore, we verified results with more straightforward technique by employing instance based learning (the Adaboost algorithm).

Adaboost is powerful technique to discover difficult examples. The Adaboost (Freund and Schapire, 1996) is an iterative procedure where in each iteration the so called weak learner is used to learn a new rule. This new rule is created in the way that the most difficult examples (misclassified) from the previous iteration are weighted more than the correctly classified examples. At the end of the learning the weights of individual examples could be used to assess the difficulty. As the weak learner we used the simple decision stump. The basic steps of the algorithm are summarized in Algorithm 4.

8.2.3 Clinical evaluation

We proposed and described the latent class model (LCM) for clinical evaluation and its advantages over the simple majority voting in Chapter 6. Here we compare the model predictions (outcome) to the classes determined by pH levels (normal, suspicious, and pathological). The classes are assigned based on the maximum of posterior probability given by the LCM, similarly as for the unsupervised learning.

Algorithm 4: Adaboost algorithm

Input: a data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, where N is number of examples, $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$
T – the maximum number of weak learners to be included in the ensemble
h – a weak learner
begin
i) initialize the weight vector $\mathbf{v}_1 = (1/N, 1/N, \ldots, 1/N)$
ii) **for** $t = 1, \ldots, T$ **do**
train weak learner h_t sampling $\mathcal{D}$ according to $\mathbf{v}_t$
$\epsilon_t = \sum_i v_t(i)[h_t(\mathbf{x}_i) \neq y_i]$, where $h_t(\mathbf{x}_i)$ is the weak classifier
$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$
$\mathbf{v}_{t+1} = \frac{\mathbf{v}_t}{Z_t} \times \begin{cases} \epsilon^{-\alpha_t}, & \text{if } h_t(\mathbf{x}_i) = y_i \\ \epsilon^{\alpha_t}, & \text{if } h_t(\mathbf{x}_i) \neq y_i \end{cases}$, Z_t is normalizing constant
end
iii) classify any new instance $\mathbf{x}$ using $H(\mathbf{x}) = sign\left(\sum_{t=1}^T \alpha_t h_t(\mathbf{x})\right)$
end

8.3 Building a hierarchical model for FHR evaluation

In this section we introduce a hierarchical model for fetal heart rate evaluation, which main idea is to better model the imprecise definition of adverse labour outcome.

The main purpose of fetal heart rate monitoring is to prevent a baby from adverse short term and long term (years) sequels though the long term progress of a baby is difficult to monitor for years and it is even more difficult to link a baby's possible complications to the intrapartum period. Therefore, in virtually all works, the adverse labour outcomes are replaced by more readily obtained indicators. The most used are pH, BE, BDecf, and Apgar score. Despite their common usage, they have many flaws, which could be divided roughly into common problems and individual issues related to a particular marker. Below we briefly introduce these from both angles. The proper description is beyond the scope of this work; more information can be found in (Armstrong and Stenson, 2007; Malin et al., 2010) among others.

The common issues with biochemical markers Briefly, the common problems for biochemical markers are: improper measurement, swapped samples from vein and artery, unclear relationship between FHR and a marker (e.g. pH), and miscellaneous thresholds to define a pathology. The biochemical measures are very dependant on the measuring procedure, for example the BDecf is dependant on a correct measurement of the pCO_2. The main difficulty lies in the exact relation of biochemical markers to fetal heart rate, which is not fully understood. Time between the recording and actual delivery plays a crucial role. The best example to understand the connection is timely Caesarean section (CS) due to suspicious CTG – the CTG is suspicious/pathological but an intervention prevented baby to get into real asphyxia that would be reflected in the pH value. In addition (Yeh et al., 2012) proves on 51519 cases that pH is weakly associated to adverse outcomes. The published meta-analysis (Malin et al., 2010) showed significant relationship between low pH and neonatal mortality, hypoxic ischaemic encephalopathy, intraventricular haemorrhage or periventricular leucomalacia, or cerebral palsy.

Disadvantages of individual markers Despite the common problems, there are individual problems related to each marker. The pH does not change linearly even though changes in the fetus are almost linear (Ross, 2011), additionally the value deteriorates "automatically" with time (Lynn and Beeby, 2007). The BE is claimed as obsolete (Rosén et al., 2007) providing false positives but it is still used

widely (Roemer, 2007). The recent study (Georgieva et al., 2013a) showed that pH is better than BDecf for predicting seizures and other cerebral problems. The Apgar score, although very simple and old, has still its merit (Finster and Wood, 2005). However, it is subjective measure with high inter-observer variability (O'Donnell et al., 2006).

8.3.1 The hierarchical model and its components

Why all models are imprecise The models used for analysis and classification of FHR are just approximation of real situation. There are two major sources of imprecisions: i) the FHR contains noise and artefacts caused by the measurement techniques and ii) the measurement of labour outcome is obscured by imprecise fetal well-being classification. In general the common practice is to use some marker (pH, BE, BDecf) and choose a threshold/s as a separating boundary for normal/suspicious-/pathological or any other type of classes. Many different thresholds have been used in the past as documented in Tables 3.1 and 3.2. The abundance of these resulted into the situation in which any comparison across different works is impossible. Even more, to use a single value as a separation boundary between normal/suspicious/pathological is simple but imprecise. The strict boundary is necessary for being able to, at least, quantify newborn and evaluate labour in general but, on the other hand, it attributes for mixing or separating different types of fetuses, especially those lying near to boundary, e.g. fetuses with pH 7.09 and 7.11 for pH threshold pH = 7.10. This is clearly present in the histograms of pH provided by (Yeh et al., 2012) or by EveREst plot (visualisation of population percentiles) by (Georgieva et al., 2013a).

The inability to select a proper, singular, marker that should be used, lead to their combination performed in many works, as documented in Tables 3.1 and 3.2. However, the markers are combined using strict logical conjunction/disjunction (and/or) but it is unlikely that markers are equally good. A solution is a scoring system that weights individual markers and their values, such a system was proposed for predicting neonatal morbidity (Portman et al., 1990). In this system the Apgar score at 5 min., BDecf, and patterns of FHR were assigned a score, which was combined to predict neonatal morbidity. The higher the score the more probable the neonatal morbidity was.

In this section we propose and describe a simple yet effective way how to combine the markers in a hierarchical structure and infer the weights of individual markers from available data. The resulting class is modelled as latent class model where biochemical markers, Apgar score, and clinical evaluation of CTG are hierarchically structured and used for latent class estimation. Initially our model was designed without knowing the scoring system of (Portman et al., 1990) despite that the models have common ground. Though, in contrast to the scoring system, our model captures the uncertainty in biochemical markers and clinical evaluation using a hierarchical structure. The main idea and purpose of the model is to better model the imprecise classification of labour outcome.

The hierarchical model is composed of majority voting of biochemical markers, Apgar score at 5 min., and latent class model of clinical evaluation of CTG. The model is presented in Figure 8.1. In order to simplify the model and improve its interpret-ability we categorized biochemical markers and Apgar score into three, commonly used, categories: normal, suspicious, and pathological. Below we present details for individual components of the model.

Biochemical markers There is a strong linear relationship between biochemical markers (pH, BE, and BDecf) as shown in Figure 8.2. The BDecf has negative correlation to pH and BE. If we consider the absolute value of correlation the pH is less correlated to BDecf than to BE.

We used the following thresholds to define pathological, suspicious, and normal categories. The pH was considered as pathological pH ≤ 7.10 (Georgieva et al., 2013b; Yeh et al., 2012), suspicious with pH of one standard deviation from median value pH $> 7.10 \wedge$ pH ≤ 7.15 (Victory et al., 2004), and normal with pH > 7.20 (Bernardes et al., 1998; Maharaj, 2008). For the BE the following thresholds were used: pathological BE ≤ 12, suspicious with BE of one standard deviation from median value BE $> -12 \wedge$ BE ≤ -8 (Roemer, 2007; Victory et al., 2004) and normal with BE > -8. The categories for BDecf were inverse to BE: pathological BDecf ≥ 12 (MacLennan, 1999), suspicious

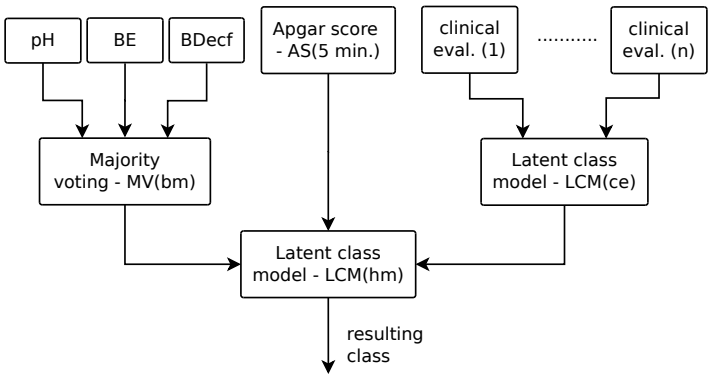


Figure 8.1: Hierarchical model of biochemical markers, Apgar score, and clinical evaluation.

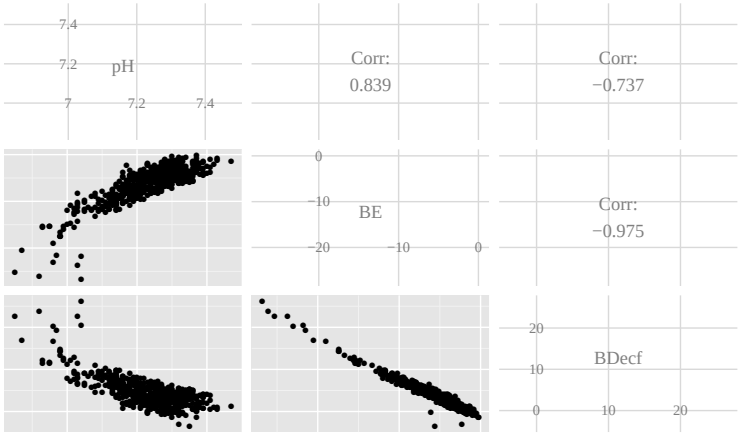


Figure 8.2: Relationship between pH, BE, and BDecf. Correlations between the biochemical markers are shown in the upper triangle.

BDecf $\geq 8 \ \wedge$ BDecf < 12 (Roemer, 2007; Victory et al., 2004) and normal BDecf < 8. In order to group biochemical markers together we used majority voting on the categories. The relationship between pH, BE, and BDecf with class determined by the majority voting is present in Figure 8.3.

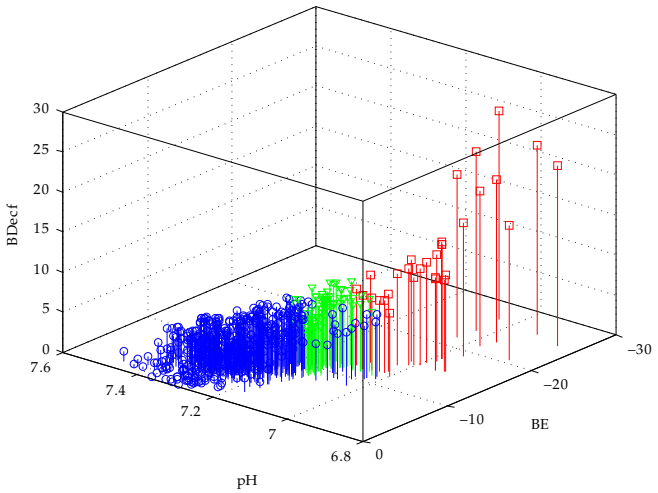


Figure 8.3: Scatter plot of pH, BE, and BDecf where the majority vote of biochemical markers determines the resulting class. Legend: normal (blue ○), suspicious (green △), pathological (red □). The figure is simplified so that negative BDecf (only for normal class) are set to zero.

Apgar score at 5 min. The categorization of Apgar score (AS) was more complicated than for the biochemical markers. The definition of pathological Apgar score is clear, AS < 7 (MacLennan, 1999). On the other hand, the boundary for suspicious/normal is unclear. Based on the distribution of the Apgar score we chose suspicious Apgar score as AS > 7 $\wedge$ AS < 9, and normal AS $\geq$ 9.

Clinical evaluation The categories of clinical evaluation were based on FIGO guidelines. Nine clinicians provided annotation of CTG in the last 60 minutes of the first stage of labour. Since there is high inter observer variability we used the latent class model to estimate a latent class. This model was implemented and thoroughly described in Chapter 6.

8.4 Classification using the hierarchical model

The hierarchical model was used to estimate a latent class (the resulting class in the LCM_{HM} model, see Figure 8.1). Then the estimated latent class was used for learning a classifier in the classical learning scenario when class assignment is known prior to learning. Despite that this approach has been shown as inferior to the latent class regression (Raykar et al., 2010) we used it in order to gain a knowledge on possible classification. We employed exactly the same classification procedure as we described in Chapter 7, see Figure 7.2. To follow Chapter 7 we joined the normal and suspicious classes. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be a dataset, where N is number of observations, $\mathbf{x}_i \in \mathcal{X}$ is d-dimensional feature vector, and $y_i \in \{0, 1\}$ is estimated latent class (0 – normal, 1 – abnormal). The 50×4 cross- validation was utilized. In each fold we balanced the normal and abnormal cases using SMOTE and then selected the best features that were used for classification. We did not divide the features into several groups but used the whole feature set. The results on each fold were aggregated and final results were computed. For more information on the classification refer to the previous Chapter 7.

8.5 Latent class regression using the hierarchical model

In Section 6.4.3 we showed that it is possible to estimate class labels from multiple noisy and imprecise annotations. We treated the unknown (hidden) class as latent variable and used the latent class analysis for its estimation. As we have said above the estimated latent class can be further used for learning a classifier. Here, we present a simple yet effective extension to the latent class analysis by employing the so called covariates (covariate can be considered as another explaining variable), in which not only a hidden class is estimated but a classifier is learned as well (McLachlan and Peel, 2000). The extension on the latent class analysis is straightforward.

8.5.1 Latent class regression

The goal of a latent class regression (LCR) is the same as for the classical learning scenario; find a function $f : \mathcal{X} \to \mathcal{Y}$ which generalizes well on unseen data. In the typical scenario the labels y_i are known before learning a classifier. In our case however, it is infeasible/impossible to obtain the class labels for training. Instead we have a variety of noisy labels $y_i^1, y_i^2 \ldots, y_i^J$ from multiple sources J, which are imprecise and prone to errors. The latent class regression simultaneously estimate the unknown class labels y_i and a classifier f as well. The structure of model is similar to the model shown in Figure 8.1, however in the latent class regression model the feature are employed as it is shown in Figure 8.4.

Classification model The method could be used for any classifier that produce soft probabilistic estimates, for ease of exposition and to keep a link to the paper (Raykar et al., 2010) we use logistic regression. The logistic regression introduce a non-linearity to a linear classifier. Consider a linear discriminant function $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, where $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$. For the *binary classification*, $\mathcal{Y} = \{0, 1\}$, the

classifier is written in a form $\hat{y} = 1$ if $\mathbf{w}^T\mathbf{x} > \gamma$ and 0 otherwise, where threshold γ defines decision boundary. The probability of positive class is modelled using logistic sigmoid

$$\Pr[y = 1|\mathbf{x}, \mathbf{w}] = \sigma(\mathbf{w}^T\mathbf{x}),$$

where the logistic sigmoid is defined as

$$\sigma(u) = 1/(1 + e^{-u}).$$

For the *multiclass classifier*, $\mathcal{Y} = \{1, \ldots, C\}$, the probability of class is computed as

$$\Pr[y = c|\mathbf{x}, \mathbf{w}] = \frac{e^{\mathbf{w}^T\mathbf{x}}}{1 + e^{\mathbf{w}^T\mathbf{x}}} \quad c < C$$
$$\Pr[y = C|\mathbf{x}, \mathbf{w}] = \frac{1}{1 + e^{\mathbf{w}^T\mathbf{x}}},$$

the resulting class is given to that with maximum probability $\hat{y} = \underset{c}{\operatorname{argmax}}(\Pr[y = c|\mathbf{x}, \mathbf{w}])$.

Binary classification

The proposed model is very similar to the model we already described in Section 6.4. The likelihood function of the parameters $\boldsymbol{\theta} = \{\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ given the observations $\mathcal{D}$ is defined as

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} \Pr[y_i^1, \ldots, y_i^J|\mathbf{x}_i, \boldsymbol{\theta}].$$

The derivation is the same as described above. Recall the likelihood equation, (6.8), where we used prevalence of a class p. Here we use output of a classifier and computed probability $p_i = \sigma(\mathbf{w}^T\mathbf{x})$ instead. The log likelihood is computed as

$$\log \Pr[\mathcal{D}|\boldsymbol{\theta}] = \sum_{i=1}^{N} \log[p_i a_i + (1 - p_i)b_i], \tag{8.1}$$

where a_i and b_i are defined by equations (6.9). The maximum-likelihood estimator is found by maximizing the log likelihood

$$\hat{\theta}_{\mathrm{ML}} = \{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{w}}\} = \underset{\theta}{\operatorname{argmax}}\{\log \Pr[\mathcal{D}|\boldsymbol{\theta}]\}.$$

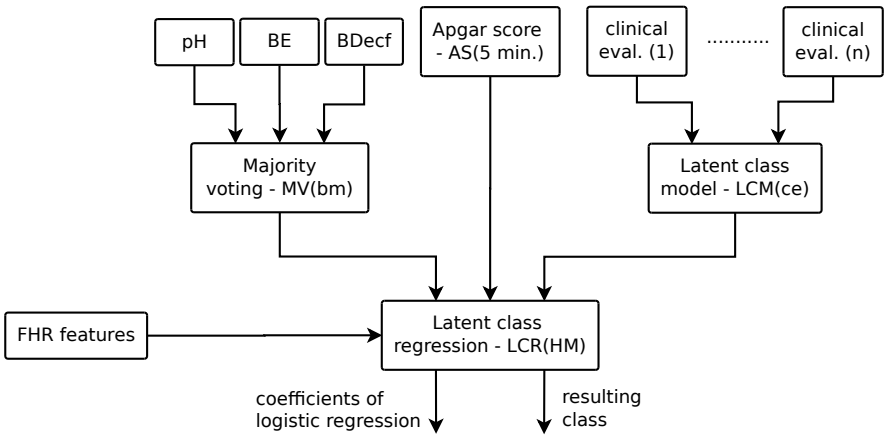


Figure 8.4: Hierarchical model using the latent class regression of biochemical markers, Apgar score, clinical evaluation and FHR features.

Estimation using EM algorithm We use expectation maximization algorithm to estimate model parameters and hidden variables. We described the EM algorithm in Section 6.4.2, here we offer brief summary. The EM algorithm is iterative procedure that first uses an initial estimate of hidden data, and than repeat two steps. First, the expectation step (E step) where the initial values are used to estimate the maximum likelihood for the interested variables. Second, the maximization step (M step) where new estimates of hidden variables are computed. The E and M step are repeated until convergence. The details of derivation were presented in Section 6.4, here we present only final equations.

E-step: We compute the conditional expectation

$$\mathbb{E}\{\log \Pr[\mathcal{D}, \mathbf{y}|\boldsymbol{\theta}]\} = \sum_{i=1}^{N} \mu_i \log p_i a_i + (1 - \mu_i) \log(1 - p_i) b_i,$$

where μ_i is defined as

$$\mu_i = \frac{a_i p_i}{a_i p_i + b_i(1 - p_i)}.$$

M-step. The current estimate of μ_i is used to maximize the conditional expectation

$$\alpha^j = \frac{\sum_{i=1}^{N} \mu_i y_i^j}{\sum_{i=1}^{N} \mu_i} \quad \beta^j = \frac{\sum_{i=1}^{N} (1 - \mu_i)(1 - y_i^j)}{\sum_{i=1}^{N} (1 - \mu_i)}.$$

The classifier weights $\mathbf{w}$ need to be updated for every iterations of EM algorithm. Since sigmoid function is nonlinear and we don't have the closed form solution we have to use the gradient ascent based optimization method. The Newton-Raphson update is given by

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \mathbf{H}^{-1} \mathbf{g},$$

where $\mathbf{g}$ is a gradient vector, $\mathbf{H}$ is a Hessian matrix, and η is a length of step. The gradient is

$$\mathbf{g}(w) = \sum_{i=1}^{N} \left[\mu_i - \sigma(\mathbf{w}^T \mathbf{x}_i)\right] \mathbf{x}_i$$

and the Hessian matrix is defined as

$$\mathbf{H}(w) = -\sum_{i=1}^{N} \left[\sigma(\mathbf{w}^T \mathbf{x}_i)\right] \left[1 - \sigma(\mathbf{w}^T \mathbf{x}_i)\right] \mathbf{x}_i \mathbf{x}_i^T.$$

Two steps of EM are repeated until convergence, i.e. when difference between iterations is smaller than a predefined threshold $Q(\boldsymbol{\theta}^{t+1}, \hat{\boldsymbol{\theta}}^t) - Q(\hat{\boldsymbol{\theta}}^t, \boldsymbol{\theta}^{t-1}) > \varepsilon$, see Algorithm 2.

Multi-class classification

The extension to the multiple classes is straightforward. Instead of α^j and β^j representing the sensitivity and specificity, we use a common parameter α_{ck}^j, representing accuracy of classification k to a class c for an element/annotator j. For binary case, α_{00}^j equals sensitivity and α_{11}^j equals specificity. The likelihood function of the parameters $\boldsymbol{\theta} = \{\mathbf{w}, \alpha_{ck}^j\}$ given the observations $\mathcal{D}$:

$$\Pr[\mathcal{D}|\boldsymbol{\theta}] = \prod_{i=1}^{N} \left[\sum_{y_i=1}^{C} \prod_{j=1}^{J} \Pr[y_i^j | y_i = c, \mathbf{w}] \cdot \Pr[y_i = c | \mathbf{x}_i, \mathbf{w}] \right].$$

The derivation is similar as to the approach described in Section 6.4. Recall equation (6.12), which were used to compute the conditional dependence in the **E-step**. In this equation we replace the prevalence p_c with output of multinomial logistic regression, p_{ic}.

$$\mathbb{E}\{\log \Pr[\mathcal{D}, \mathbf{y}|\boldsymbol{\theta}]\} = \sum_{i=1}^{N}\sum_{c=1}^{C} \mu_{ic} \log\left[p_{ic} \prod_{j=1}^{J}\prod_{k=1}^{C} (\alpha_{ck}^{j})^{\delta(y_i^j,k)} \right],$$

where $\mu_{ic} = \Pr[y_i = c|y_i^j, \boldsymbol{\theta}]$ is estimated probability of unknown ground truth

$$\mu_{ic} = p_{ic} \prod_{j=1}^{J}\prod_{k=1}^{C} (\alpha_{ck}^{j})^{\delta(y_i^j,k)}.$$

In the **M-step** we use the current estimates to maximize the conditional expectation; the parameter α_{ck}^{j} is updated using the following equation

$$\alpha_{ck}^{j} = \frac{\sum_{i=1}^{N} \mu_{ic}\delta(y_i^j, k)}{\sum_{i=1}^{N} \mu_{ic}}.$$

Similarly to the binary classification we update weights $\mathbf{w}$ using the Newton-Raphson update, which is given as

$$\mathbf{w}^{t+1} = \mathbf{w}^{t} - \eta \mathbf{H}^{-1}\mathbf{g},$$

where $\mathbf{g}$ is a gradient vector, $\mathbf{H}$ is a Hessian matrix, and η is a length of step.

8.6 Proposed experimental methodology

In this chapter we use the hierarchical model for CTG evaluation. We utilize two experimental methodologies. First, the classical learning scenario when the resulting class is estimated using the latent class analysis of hierarchical model (LCM_{HM}) and used for learning a classifier. The classification approach was re-used from Chapter 7 and briefly summarized in Section 8.4. The second approach use the latent class regression (LCR) when not only the class labels are estimated but a classifier is learned as well (the LCR_{HM} model). The methodology for this approach is described in detail below. Both approaches works with the CTU-UHB database (Chapter 4) using the same FHR features extracted on the last 60 minutes of the first stage of labour.

The whole procedure for using LCR is presented in Figure 8.5. The input is data set (feature set and class labels from multiple sources). The procedure consisted of $50\times q$-fold cross validation (CV) where data were 50 times randomly split for q-fold CV and results for each run were aggregated. We used the same number of folds $q = 4$ as in Chapter 7. The training set consisted of d features and, because of high dimensionality, we used the principal component analysis to reduce the dimensionality to l principal components, where $l < d$. The number of l was determined in each fold of CV. The same transformation was applied to the test dataset, where the test set was transformed using the matrix of basis vectors $\mathbf{B}$ obtained from the training set. Then the LCR was used to estimate the latent class and coefficients of logistic regression on the training set. The model parameters learned using the LCR were used to estimate the labels for the test set: the parameters $[\mathbf{A}]_{ck} = \alpha_{ck}^{j}$ and prevalence for individual classes $\mathbf{p}$. Then the learned classifier $\mathbf{w}$ on training data set was tested on the test set. The results on each fold of 50×4-fold CV were aggregated.

8.7 Performance evaluation

In this chapter we used the same metrics for performance evaluation as we used in Chapter 7. The statistical measures such as sensitivity, specificity, precision, and F-measure were described in Section 7.4.3. These measures can be used for binary classification only. For more than two classes the classification performance is computed as one class versus group of the other classes – the so called one-vs-all approach.

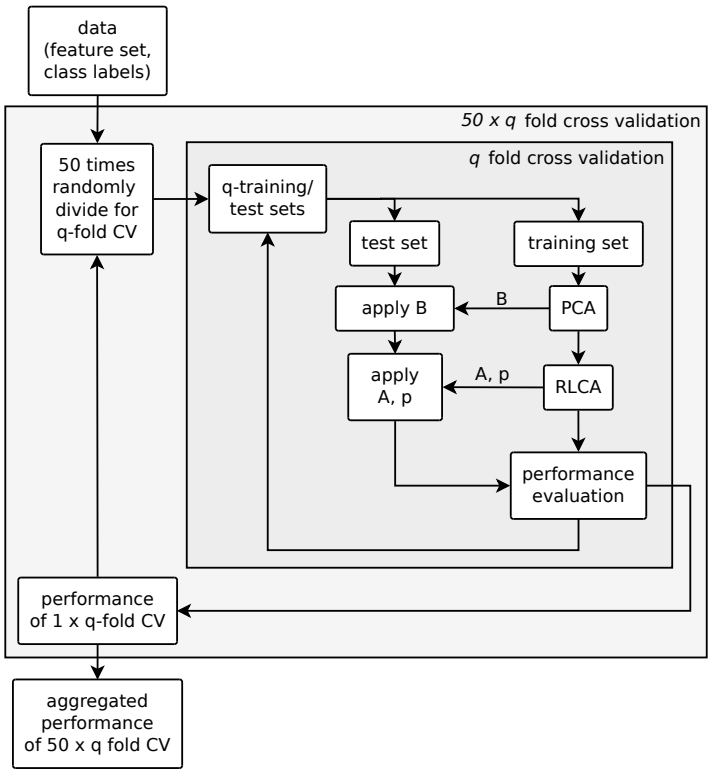


Figure 8.5: Experimental methodology for the latent class regression. The notations are not printed in mathematical symbols, they are as follows: q = 4 is the number of CV folds, B = $\mathbf{B}$ is matrix of basis vectors, A = $[\mathbf{A}]_{ck}$ is matrix of parameters, and p = $\mathbf{p}$ is prevalence of estimated classes.

Confusion matrix For the analysis of difficult examples and classification performance we used visualisation using a confusion matrix, see Table 8.1. In this matrix the correct classification is present on the diagonal line, TP_{ck}, where $c = k, c = \{1, 2, 3\}$ and $k = \{1, 2, 3\}$. Misclassifications are present by off diagonal elements. The classes are: 1 – normal, 2 – suspicious, and 3 – pathological. The most serious misclassifications are those the most distant from the diagonal: FP_{13} and FN_{31}.

Table 8.1: Structure of confusion matrix used in this chapter. (n/s/p – actual values, n′/s′/p′ – predicted values, n – normal, s – suspicious, p – pathological). The classes are 1 – normal, 2 – suspicious, and 3 – pathological.

	n′	s′	p′
n	TP_{11}	FP_{12}	FP_{13}
s	FN_{21}	TP_{22}	FP_{23}
p	FN_{31}	FN_{32}	TP_{33}

The overall classification accuracy for the three classes can be computed as ACC = $(\mathrm{TP}_{11} + \mathrm{TP}_{22} + \mathrm{TP}_{33})/N$, where N is number of instances. When we evaluate sensitivity and specificity on three classes we shrink three classes into two (one versus all approach). We assess pefromance on pathological versus joined normal and suspicious classes. For such created two classes the same metrics could be used as described in Section 7.4.3.

The similar confusion matrix is used for visualisation of the probability of an estimated latent class to a particular component $[\mathbf{A}]_{ck} = \alpha_{ck}$, where α_{ck} is probability of the estimated latent class c when the predicted class was k. The estimated latent classes are marked as $n/s/p$ (normal/suspicious/pathological) and the predicted classes as $n'/s'/p'$ (normal/suspicious/pathological). The matrix has structure as follows

$$\mathbf{A}^{\mathbf{j}} = \begin{pmatrix} \alpha_{nn'} & \alpha_{ns'} & \alpha_{np'} \\ \alpha_{sn'} & \alpha_{ss'} & \alpha_{sp'} \\ \alpha_{pn'} & \alpha_{ps'} & \alpha_{pp'}. \end{pmatrix}$$

8.8 Results

8.8.1 Clustering of FHR using GMM

We used the Gaussian mixture model (GMM) to model the fetal heart rate represented by features in a space $\mathcal{X} \in \mathbb{R}^d$. The feature space was high dimensional and most probably contained redundant information, therefore we used the principal component analysis to transform it into a lower dimension l, where $l < d$. Then we aimed to find a model that best described the FHR patterns, hence the behaviour of fetus. The model with $\hat{m}$ components minimizing AIC and BIC was sought.

Feature extraction/ dimensionality reduction

Choosing the dimension l is a trade off between number of components retained and information (variance) lost. Generally the l is sought to contain 95% variance of data. In Figure 8.6 we show a cumulative sum of sorted eigenvalues ($\lambda_1 \geq \lambda_2, \ldots \geq \lambda_d$,). The biggest λ_i corresponds to greatest variance. From Figure 8.6 it can be seen that the great proportion of variance is contained in the firsts principal components. Each added principal component comes at expense of higher dimensionality and need of estimating bigger covariance matrices, where ($l[l+1]/2$) is the number of estimated elements in Σ_k for l features). Therefore we sought l as a trade off between low l with as much variance retained as possible. If we rather prefer low l we can chose $l = 7$ when λ_i "stops decrease" while, on the other hand, the variance is preferred we chose $l = 13$ to contain 95% of variance. In our experiments we use range of $l = \{7, \ldots, 13\}$, we comment on different performance below.

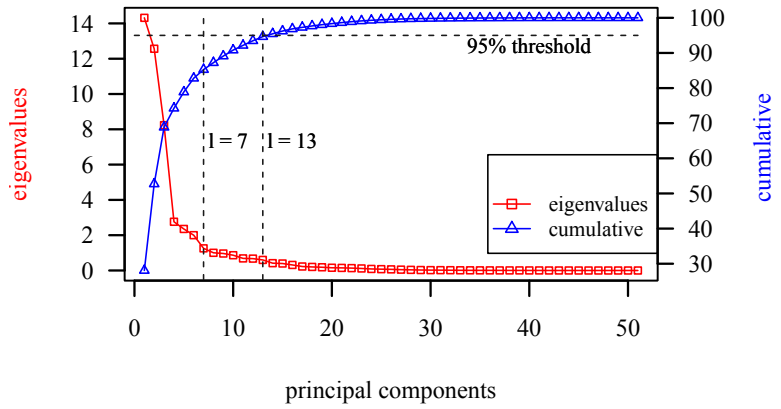


Figure 8.6: Values of eigen values and their cumulative sum in percentage of their variance.

Quantization of fetal behaviour using FHR

We aimed to quantize the fetal behaviour using fetal heart into m-finite states. The optimal number of components $\hat{m}$ was found as value of m that minimizes the AIC and BIC. In practice however, the balance between AIC and BIC has to be sought (AIC overestimates $\hat{m}$ while BIC underestimates it). The results for $m = \{2, \ldots, 8\}$ are shown using box-plots in Figure 8.7. Clearly, the AIC decreased for increasing m and BIC increased. Nevertheless, from $m = 2$ to $m = 3$ BIC climbed up slightly while the AIC sharply decreased. Apparently the good choice was to chose $\hat{m} = 3$ when there was the small change in BIC but the large in AIC. The chosen $\hat{m}$ was the same irrespective of number of principal components $l = \{7, \ldots, 13\}$. However, intuitively, when choosing lower number of principal components, $l = 7$, the model had better fit than for $l = 13$ because of lower dimensionality.

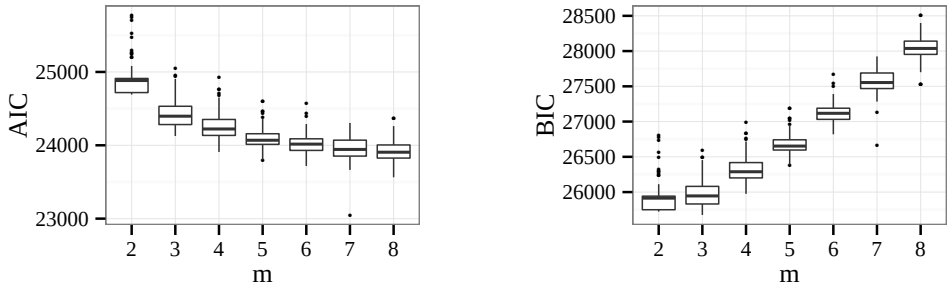


Figure 8.7: Number of components m with respect to AIC and BIC (for 13 principal components).

8.8.2 The most difficult examples

We searched for the difficult examples using various techniques. First, the unsupervised learning (GMM model from the previous section) with three components, $\hat{m} = 3$. We estimated the model for the $l = 13$ principal components to include as much information as possible into the model. The learning of GMM was restarted 20 times and model with the best likelihood was chosen. Second, we employed supervised learning (Naive Bayes, SVM, and C4.5) classifiers with the LOOCV method. The prediction of classifiers were combined using majority voting. Third, we re-used the latent class model from Chapter 6 (Section 6.4) to obtain an estimate (unknown "ground truth") of clinical evaluation. The model worked with three classes, $c = 3$. We were interested in the most serious errors in the classification, i.e. those errors spanning across one class. The results of classification are presented using the confusion matrices in Table 8.2.

Table 8.2: Confusion matrices for unsupervised learning, supervised learning, and LCM of clinical evaluation, The difficult false negatives, false positives, and easy true positives are presented in the lower left, upper right, and lower right corner, respectively. (n/s/p – actual values, n′/s′/p′ – predicted values, n – normal, s – suspicious, p – pathological).

(a) unsupervised learning

	n′	s′	p′
n	174	161	104
s	21	17	14
p	24	10	27

(b) supervised learning

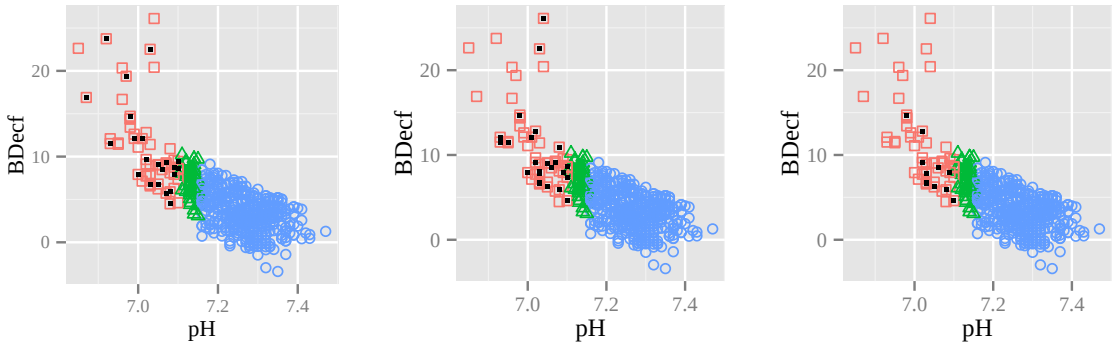
	n′	s′	p′
n	336	29	74
s	35	3	14
p	25	6	30

(c) LCM of clinical eval.

	n′	s′	p′
n	149	191	99
s	15	26	11
p	11	24	26

The number of "easy records" (true positives) TP_{33} is similar for all methods (unsupervised, supervised, and clinical evaluation). The number of false positives FP_{13} varies for all methods, while the number of false negatives FN_{31} is similar for unsupervised and supervised technique and is the lowest for the LCM of clinical evaluation (LCM_{ce}). In order to examine FN_{31} in more detail we plot their relationship to pH and BDecf. The similar plot to Figure 8.3 only the BE is not included since it is highly correlated to BDecf. The relation can be seen in Figure 8.8. The misclassified FN_{31} are more concentrated at the boundary between pathological and suspicious. This holds especially for the LCM_{ce}. On the one hand there is higher density while, on the other hand, the examples next to boundary are similar to the examples lying at the other side of the boundary.

The misclassified examples in the supervised scenario could be biased towards the used feature selection and classification methods. We verified the results by employing instance based learning (the Adaboost algorithm), which is more "natural" way to assess the difficulty of examples for classification. We used the LOOCV with $t_{max} = 100$ iterations (= 100 weak learners). From each q-th fold of LOOCV we obtained the weights $\mathbf{v}^q_{t_{max}}$. Then these weights were averaged across all folds $\bar{\mathbf{v}} = \frac{1}{N-1} \sum_{q=1}^{N-1} \mathbf{v}^q_{t_{max}}$ giving an estimate of examples difficulty. The distribution of weights is depicted in Figure 8.9. The threshold for choosing difficult examples is marked by an arrow ($\downarrow$).



(a) unsupervised learning **(b)** supervised learning **(c)** LCM of clinical evaluation

Figure 8.8: The relationship between pH and BDecf with marked false negatives FN_{31} for different techniques. Classes: normal (blue ○), suspicious (green △), pathological (red □), false negative (black ▪).

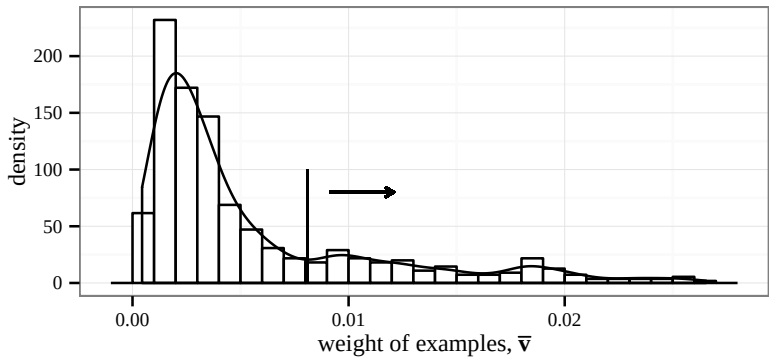


Figure 8.9: Adaboost weights of difficult examples. The threshold for difficult is marked by an arrow (| →).

The application of selected threshold identified 77 examples as difficult. We verified their correspondence to the supervised learning scenario. In the TP_{33} group the adaboost identified 22 (from 30) same examples. In the FN_{31} group the number of same examples was 24 (from 25) and for the FP_{13} was 12 (from 74). The adaboost identified almost the same set of easy/difficult examples as supervised learning for TP_{33} and FN_{31} while for the FP_{13} group agreed with supervised learning only on 12 examples.

Combination of the methods We examined the intersection of the records obtained from the three methods: unsupervised, supervised, and LCM of clinical evaluation (LCM_{ce}) for the true positives TP_{33}, false positives FP_{13}, and false negatives FN_{31}. Let U, S, C be sets that contain TP_{33}, FP_{13}, FN_{31} records for the supervised, unsupervised scenario, and clinical evaluation, respectively. In Table 8.3 we present pair-wise intersection of methods, where e.g. $|U \bigcap S|$ denotes the cardinality of intersection (number of the same examples) for unsupervised and supervised scenarios.

The cardinality of intersection shows that there are 14 records that are distinct and easily classified, TP_{33}, for all three methods. The number of FP_{13} is high irrespective the method used while, interestingly, the intersection of FP_{13} for all three methods is small. This imply that using different techniques the number of false positives could be reduced. Also the cardinality of intersection for FN_{31} is low mainly because of small number of FN_{31} for the LCM_{ce}. This suggest that by using the LCM_{ce} we can lower the number of false negatives.

We searched for similar patterns in the TP_{33}, FP_{13}, and FN_{31} groups. For example if among the TP_{33} were vaginal deliveries only or any other dominating factor, which could affect the results. In

Table 8.3: Number of records that are the same with respect to different approaches. FN_{31} and FP_{13} represent the most difficult false negatives and false positives, respectively. On contrary, the TP_{33} marks the most easiest records, i.e. records identified as true positive.

	FN_{31}	FP_{13}	TP_{33}
$\lvert U \cap S \rvert$	10	42	19
$\lvert U \cap C \rvert$	5	52	20
$\lvert S \cap C \rvert$	11	33	17
$\lvert U \cap S \cap C \rvert$	5	26	14

the TP_{33} and FP_{13} there were no common patterns. The FN_{31} group contained only babies delivered vaginally and FHR recorded only by Doppler ultrasound. Nevertheless these factor should not attribute to misclassification. The more important finding is that for the FN_{31} group the BDecf and BE are not pathological but one example. Average values of the biochemical markers for FN_{31} were: pH = 7.06, BE = -10.9, and BDecf = 8.4. The distribution of the FN_{31} with respect to pH and BDecf is shown in Figure 8.10. Clearly the FN_{31} examples are close to suspicious class but one example.

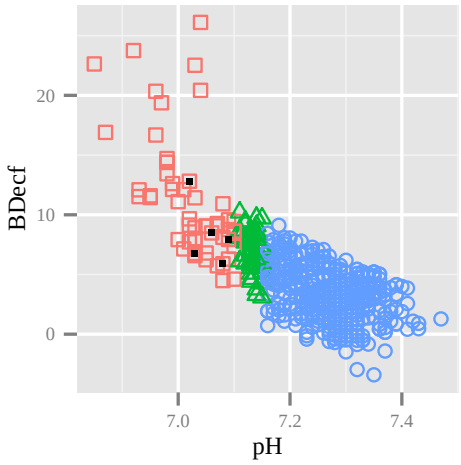


Figure 8.10: The relationship between pH and BDecf with marked false negatives FN_{31} for the intersection of unsupervised, supervised, and LCM ($\lvert U \cap S \cap C \rvert$). Classes: normal (blue ○), suspicious (green △), pathological (red □), false negative (black ■).

8.8.3 The hierarchical model – latent class analysis and regression

We have designed a hierarchical model in order to improve the imprecise classification of labour outcome when we must account for uncertainty related to the individual markers.

In this section we describe the learning of hierarchical model (HM) and weights of individual components. First, we estimated the HM without regression (LCM_{HM}) and used the estimated latent class (the resulting class in Figure 8.1) for learning a classifier. The latent class was estimated using approach described in Chapter 6 and the classification was performed as described in Chapter 7. Second, we estimated the latent class of hierarchical model and coefficients of logistic regression using the latent class regression LCR_{HM}. We used three latent classes (normal, suspicious, pathological) for LCM_{HM} and LCR_{HM}.

Latent class analysis using the hierarchical model

The HM was learned using the EM algorithm (see Algorithm 2) that was iterated until convergence, i.e. while the log-likelihood was decreasing, expressed using the Q-function, eq. (6.7). The rate of convergence was determined by setting the $\varepsilon = 10^{-3}$. The $\mathcal{S}_{acc}$ score as a function of iterations is shown in Figure 8.11.

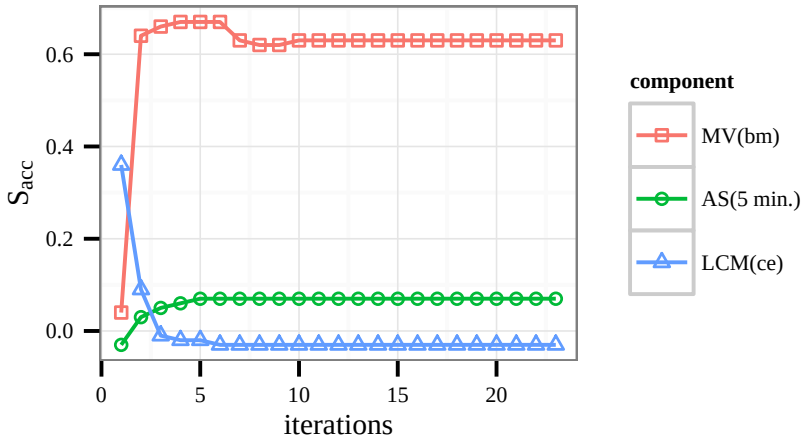


Figure 8.11: Progression of $\mathcal{S}_{acc}$ score for the majority vote of pH, BE, and BDecf (MV(bm)), Apgar score at 5 min. (AS(5 min.)), and latent class analysis of clinical evaluation LCM(ce).

The initial settings for latent class μ_{ic} was determined by majority voting. Then for the growing number of iterations, the MV_{bm} increased rapidly while the LCM_{ce} decreased. For the higher number of iterations the $\text{AS}_{5\text{ min.}}$ and LCM_{ce} are inferior to the μ_{ic}. The $\text{AS}_{5\text{ min.}}$ because of poor results on pathological class and LCM_{ce} because of poor distinction between normal and suspicious classes. Details for the final model can be observed in the following confusion matrices

$$\mathbf{A}^{bm} = \begin{pmatrix} 0.91 & 0.09 & 0 \\ 0.01 & 0.72 & 0.27 \\ 0.19 & 0.00 & 0.81 \end{pmatrix} \quad \mathbf{A}^{AS} = \begin{pmatrix} 0.87 & 0.11 & 0.02 \\ 0.00 & 0.74 & 0.26 \\ 0.53 & 0.47 & 0 \end{pmatrix} \quad \mathbf{A}^{LCM} = \begin{pmatrix} 0.34 & 0.44 & 0.22 \\ 0.33 & 0.44 & 0.23 \\ 0 & 0.32 & 0.68 \end{pmatrix}.$$

Recall that $[\mathbf{A}]_{ck} = \alpha_{ck}$, where α_{ck} is probability of the estimated latent class c when the predicted class was k. In these matrices the correspondence of estimated latent class to a particular component is on the diagonal when $c = k$ while on the off-diagonal elements are present "misclassification's", $c \neq k$. It can be seen that latent class corresponds to the MV_{bm} for all classes. For the $\text{AS}_{5\text{ min.}}$ the results are inferior to the estimated pathological (p) class ($c = p, k = p'$), when $\alpha_{pp'} = 0$ (the lower right corner of $\mathbf{A}^{AS}$). In contrast to the $\text{AS}_{5\text{ min.}}$ the LCM_{ce} has a good performance on the pathological class while poor on the normal and suspicious, where the LCM_{ce} could not distinguish between normal and suspicious. For the sake of clarity the confusion matrices are presented in Figure 8.12.

Clustering with respect to different elements (components) In order to examine the individual elements of the model we compared them to the unsupervised GMM model. Such comparison helps us to understand, which elements of the model corresponds best to the FHR features respective their clusters. The biochemical markers compared were: pH, BE, and BDecf. We do not included into comparison individual clinicians but we used their majority voting (MV_{ce}) instead. In addition we included the three main components: majority voting of biochemical markers (MV_{bm}), Apgar score at 5 min. ($\text{AS}_{5\text{ min.}}$), and latent class model of clinical evaluation (LCM_{ce}) and also the resulting hierarchical model (LCM_{HM}). The results are present in Table 8.4.

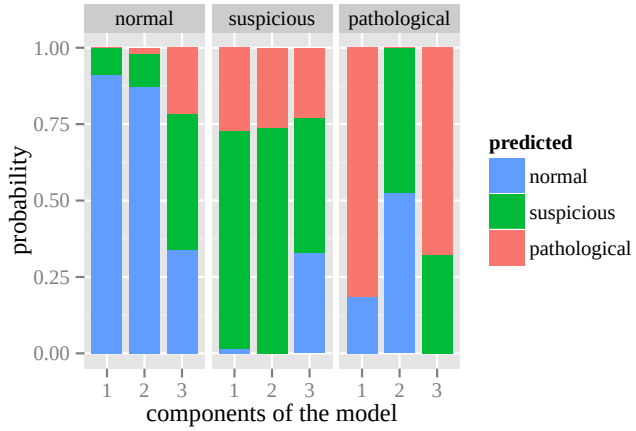


Figure 8.12: Probability of estimated latent class μ_{ic} by individual components α_{ck}. Components: 1 = MV_{bm}, 2 = $AS_{5\ min.}$, 3 = LCM_{ce}. The μ_{ic} is present in the upper grey horizontal bar, the classes given by the components are marked as predicted. For example when μ_{ic} was normal the MV_{bm}, $AS_{5\ min.}$, and LCM_{ce} predicted normal with probabilities 0.91, 0.87, and 0.34, respectively.

Table 8.4: Clustering results with respect to different elements (components). (ACC – accuracy, SE – sensitivity, SP – specificity, PR – precision, F – F-measure). The accuracy assess all classes. The SE, SP, PR, and F assess the pathological class versus joined normal and suspicious class.

element (component)	overall ACC [%]	pathological vs. (normal + suspicious) SE [%]	SP [%]	PR [%]	F [%]
pH	39	44	76	19	26
BE	38	53	76	12	20
BDecf	39	56	75	7	13
MV_{ce}	41	71	80	32	45
MV_{bm}	40	54	76	14	23
$AS_{5\ min.}$	39	32	74	4	7
LCM_{ce}	41	58	84	54	56
LCM_{HM}	41	61	76	14	22

Note that the proportion of of normal/suspicious/pathological classes are different for each element and therefore the results on pathological vs. (normal + suspicious) are slightly misleading. If we consider the F-measure, the LCM_{ce} has the best performance with F = 56%. Then follow the MV_{ce} with F = 45%. Thus, the clustering of FHR features best corresponds to the clinical evaluation (LCM_{ce}, MV_{ce}). This results were expected since the clinicians actually used the FHR patterns for evaluation. Considering the both, the overall results and results on pathological class, the LCM_{HM} is in the middle of the individual elements. It has better sensitivity than biochemical markers (pH, BE, BDecf, or MV_{bm}) but lower precision than pH. In other words the number of false negatives was decreased while the number of false positives was slightly increased. The LCM_{HM} model best corresponds to the MV_{bm} component, which coincide with Figure 8.12 and confusion matrices presented for the individual components above. We have to keep in mind that clustering is unsupervised technique and might not well relate to the actual fetal status/well-being. On the other hand, it helps to understand the data and mapping between the FHR features and different labour outcome measures.

The relation to difficult examples The ability of hierarchical model to explain pH false negatives examples (showed in Figure 8.8a) is depicted in Figure 8.13, where the LCM_{HM} is used to determine the resulting class.

The boundary between different classes (normal/suspicious/pathological) is not clear and individual classes overlaps, compare to Figure 8.8a. About half of the false negative examples (pH) is not truly

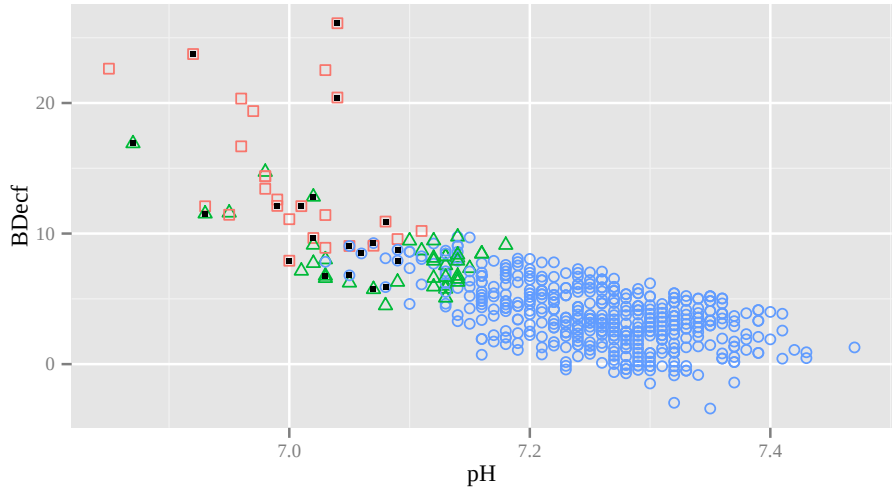


Figure 8.13: The false negatives FN_{31} from unsupervised learning when compared to pH classes, i.e. those examples there were clustered into normal class but pH was pathological. The underlying class is determined by hierarchical model: normal (blue ○), suspicious (green △), pathological (red □), false negatives (black ■).

false negative when the LCM_{HM} is considered (the normal ○ and suspicious ▽ examples marked as false negatives ■). To better explain the LCM_{HM} model and to understand how the imprecise definition of labour outcome is treated, consider the the left out-most false negative example (resulting class is suspicious). The biochemical markers (pH = 6.87, BE = -20.5, and BDecf = 16.9) and Apgar score ($AS_{5\ min.}$ = 4) imply the labour outcome being really pathological. However, the clinicians in majority considered the FHR as normal. There is a clear disconnection between outcome measures and FHR. The LCM_{HM} assigned the suspicious class to this example because, in the final model, the Apgar score is not sensitive to pathological cases $\alpha_{pp'} = 0$, hence not contributing to this class and the final decision on this case was strongly determined by MV_{bm} and LCM_{ce} and rather inclined to the clinical evaluation.

Classification using estimated class from the hierarchical model We performed the same experiment as in Chapter 7 with the estimated latent class from the LCM_{HM}. We employed the 50×4-fold cross validation using the procedure depicted in Figure 7.2. In each fold of CV we used the feature meta-selection to select the best features for classification. The features selected more than 50% of folds are present in Table 8.5. The selected features were similar to the features selected using pH (Table 7.3). The exceptions were energy03_LF_HF and ApEn(2,0.15) that were not selected for pH but were found useful for the class estimated by the hierarchical model. Again, the frequency features dominated over the other types of features.

Table 8.5: Selected features sorted based on the importance.

Feature set	Selected features
Complete set (all features)	energy04_VLF, energy03_LF, Poincaré_SD2, accNumber, energy03_LF_HF, STV-HAA, decNumber, ApEn(2,0.15), energy04_LF

The selected features were used to train the Naive Bayes, SVM, and C4.5 classifiers. The results, shown in Table 8.6, were aggregated from each fold of 50×4 fold CV and median, 25th, and 75th percentiles were estimated.

The best results were achieved using the Naive Bayes. In contrast to results in Chapter 7 (Table 7.4), the results were better for both, sensitivity and specificity, confirming that number of false negatives

Table 8.6: Classification results for resulting class estimated by the hierarchical model LCM_{HM}. The results are averaged across all folds of CV (50×4 folds CV) and presented using median and (25th – 75th) percentiles. The results were evaluated as pathological vs. (normal + suspicious). (SE – sensitivity, SP – specificity, PR – precision, F – F-measure).

Feature set	[%]	Naive Bayes	SVM	C4.5 Tree
Complete set	SE	63 (56–75)	44 (33–63)	25 (13–38)
	SP	78 (76–81)	84 (82–87)	92 (89–94)
	PR	16 (14–19)	15 (12–19)	14 (8–21)
	F	25 (22–29)	23 (18–29)	17 (10–25)

FN_{31} was reduced. On the other hand, the precision decreased (because of higher FP_{13}) and hence the F-measure decreased.

Latent class regression using the hierarchical model

In the previous section we presented results when we estimated the latent class first and then learned a classifier. In this section we simultaneously estimated the latent class and learnt a classifier as well using the latent class regression. First we describe the model that was estimated on the whole dataset. Second, we present the performance of the model and prediction error.

The properties of LCR The latent class regression model (LCR_{HM}) was estimated iteratively using the EM algorithm (Section 8.5.1), where the limit of log-likelihood convergence was set to $\varepsilon = 10^{-3}$. The whole database was used for learning the LCR model. The number of principal components for learning was set to $l = 13$ (95% variance of data retained). Initial setting of latent class was determined using majority voting. In Figure 8.14 we present the progression of accuracy based scoring $\mathcal{S}_{acc}$ of the main model components. Because of the initial setting based on majority voting the first iteration of the LCR_{HM} corresponds to LCM_{HM}, c.f. Figure 8.12. However, from the second iteration the models differ. The accuracy score $\mathcal{S}_{acc}$ of individual components computed over all classes changes slightly with increasing iterations.

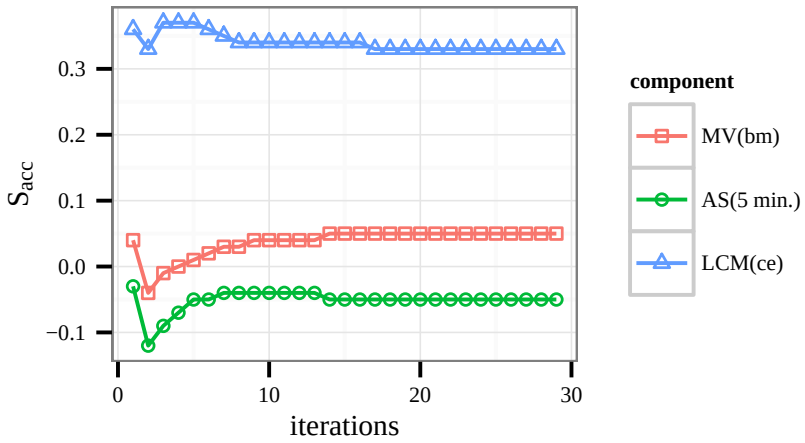


Figure 8.14: Progression of $\mathcal{S}_{acc}$ score for the main components of hierarchical model: majority vote of pH, BE, and BDecf (MV(bm)), latent class analysis of clinical evaluation (LCA(ce)), and Apgar score at 5 min. (AS(5 min.)).

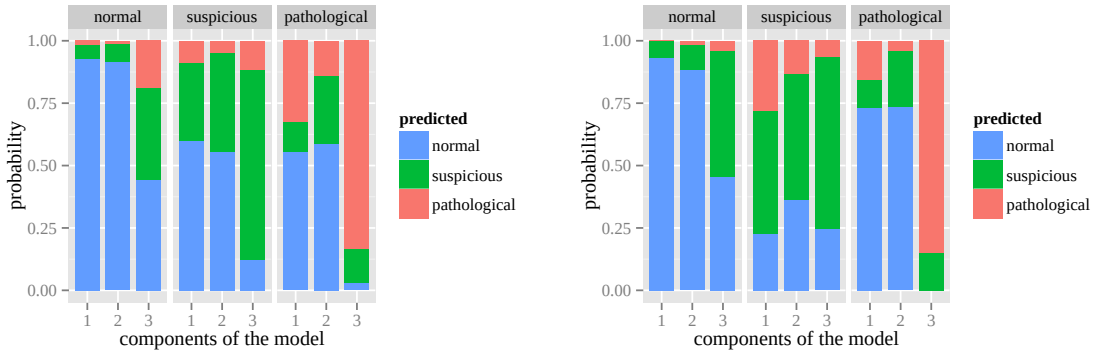
Even though the $\mathcal{S}_{acc}$ for the first iteration t_1 is similar to the $\mathcal{S}_{acc}$ for the last iteration t_{28} the parameters (weights) for individual components have changed. This change can be observed in the

confusion matrices of $\mathbf{A}_{[ck]}$ showed for the first iteration $\mathbf{A}(t_1)$ and the last iteration $\mathbf{A}(t_{28})$. The abbreviations are as follows: bm = MV_{bm}, as = $\text{AS}_{5\text{ min.}}$, and ce = LCM_{ce}.

$$\mathbf{A}(t_1)^{bm} = \begin{pmatrix} 0.93 & 0.06 & 0.1 \\ 0.60 & 0.31 & 0.09 \\ 0.56 & 0.12 & 0.32 \end{pmatrix} \quad \mathbf{A}(t_1)^{as} = \begin{pmatrix} 0.91 & 0.08 & 0.01 \\ 0.55 & 0.40 & 0.05 \\ 0.59 & 0.27 & 0.14 \end{pmatrix} \quad \mathbf{A}(t_1)^{ce} = \begin{pmatrix} 0.45 & 0.36 & 0.19 \\ 0.12 & 0.76 & 0.12 \\ 0.03 & 0.13 & 0.84 \end{pmatrix}$$

$$\mathbf{A}(t_{28})^{bm} = \begin{pmatrix} 0.93 & 0.07 & 0 \\ 0.23 & 0.49 & 0.28 \\ 0.73 & 0.12 & 0.15 \end{pmatrix} \quad \mathbf{A}(t_{28})^{as} = \begin{pmatrix} 0.88 & 0.10 & 0.02 \\ 0.37 & 0.50 & 0.13 \\ 0.74 & 0.22 & 0.04 \end{pmatrix} \quad \mathbf{A}(t_{28})^{ce} = \begin{pmatrix} 0.46 & 0.51 & 0.04 \\ 0.25 & 0.69 & 0.07 \\ 0 & 0.15 & 0.85 \end{pmatrix}$$

The confusion matrices are also present in Figure 8.15. It can be seen that mainly the suspicious and pathological latent classes were changed during the learning. The normal class corresponds to the MV_{bm} and $\text{AS}_{5\text{ min.}}$ while the pathological mainly corresponds to the LCM_{ce}. The created LCR_{HM} model is similar to the model without regression (LCM_{HM}) for normal class but for the suspicious class and mainly for the pathological these model do differ significantly. For the LCM_{HM} the MV_{bm} contributed more than for LCR_{HM}. The prevalence of estimated latent classes were as follows: $P(normal)$ = 0.65, $P(suspicious)$ = 0.11, and $P(pathological)$ = 0.24.



(a) first iteration (t_1) **(b)** last iteration (t_{28})

Figure 8.15: Probability of estimated latent class by individual components α_{ck} for the first iteration and the last iteration of EM algorithm. The latent class is present in the upper grey horizontal bar the classes given by the components are marked as predicted. (1 = MV_{bm}, 2 = $\text{AS}_{5\text{ min.}}$, 3 = LCM_{ce}).

The classification performance of LCR We used the 50×4-fold cross-validation procedure presented in Figure 8.5. The number of principal components l retained for the LCR was determined in each fold of CV. The threshold was set to u = 0.95 so that 95% variance was retained. The average number of l across all folds of CV was l = 13. The transformed feature set was used for learning the model parameters $\boldsymbol{\theta} = \{\alpha_{ck}^j, \mathbf{w}\}$. The classification accuracy was estimated on the test set and aggregated from all folds of CV and median with 25th – 75th percentiles were computed.

In Table 8.7 we present the results of latent class regression for the pathological class versus joined normal and suspicious class. In Figure 8.16 we show ROC and PR curves. The threshold γ determined the resulting class $\hat{y}$ = 1 if $\mathbf{w}^T\mathbf{x} > \gamma$ and $\hat{y}$ = 0 otherwise (1 – pathological, 0 – suspicious + normal). In Figure 8.16 it can be seen how the value of γ affects the performance of the classification. The best value in terms of the F-measure is γ = 0.5. Imagine that we desire a better sensitivity, then the red point would move up to the point were sensitivity is approximately 80% and specificity is 75%. In the PR curve the point would climb down to 80% of sensitivity and 50% of precision leading to decrease in F-measure from 65% to 62%.

Table 8.7: Classification results for latent class regression using the hierarchical model. The results are averaged across all folds of CV (50×4 folds CV) and presented as median with (25th – 75th) percentiles. The results were evaluated as pathological vs. (normal + suspicious). (SE – sensitivity, SP – specificity, PR – precision, F – F-measure)

type	[%]	median (25th-75th)
pathological vs. (normal + suspicious)	SE	66 (59–71)
	SP	89 (86–92)
	PR	66 (59–73)
	F	65 (60–70)

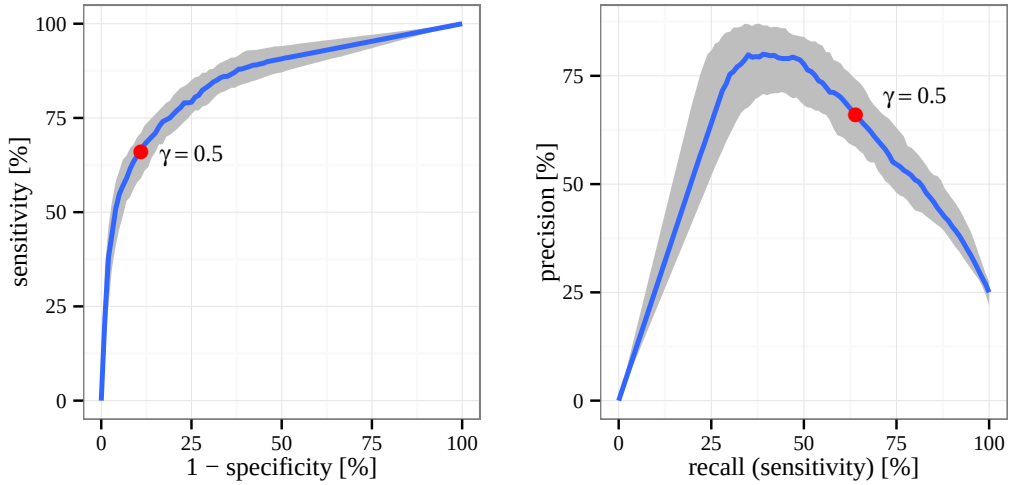


Figure 8.16: Receiver operation characteristic (left) and precision-recall curve (right). The median is presented in blue with 25th and 75th percentiles in grey. The threshold γ between two classes is marked with the red point (•).

The presented results were superior, in all measures, to the Naive Bayes classifier learned using the LCM_{HM}. The sensitivity, specificity, and precision were greatly improved. These results were expected since the classification probabilities p_{ic} were, in part, used to estimate the latent class μ_{ic}. In order to clarify the LCR_{HM} prediction capabilities we evaluated the model performance using all elements of the model. These results are presented in Table 8.8 for pathological vs. (normal + suspicious) class and in Table 8.9 for overall performance on all classes. In the two class scenario the performance of LCR_{HM} is similar to LCM_{ce} because the probability of estimated pathological latent class $\alpha_{pp'}$ was the highest for the LCM_{ce} component. However, the accuracy of LCM_{ce} is much lower than accuracy of LCR_{HM}. Note that results in Table 8.8 are slightly misleading because of different proportion of normal/suspicious/pathological cases for each element.

The connection between LCM_{HM} and LCR_{HM} Intuitively the both models LCM_{HM} and LCR_{HM} are different. For the latter model the features were used as covariates and helped to model the latent class. It is not surprising that pathological class for LCR_{HM} yielded to LCM_{ce} since this model of clinical evaluation had the best performance for unsupervised learning, see Table 8.4 where the LCM_{ce} corresponded the best to the FHR clusters. In Table 8.10 we present confusion matrix between LCM_{HM} and LCR_{HM}. The both models are similar for the normal and suspicious class. The included covariates altered mostly the pathological class.

The hierarchical model and its relation to pH There are two aspect that have to be considered. First the LCR_{HM} produced the class labels μ_{ic} and also coefficients of logistic regression $\mathbf{w}$. The

Table 8.8: Classification results for latent class regression with respect to different elements (components). The results are averaged across all folds of CV (50×4 folds CV) and presented using median (25th – 75th) percentiles. The results were evaluated as pathological vs. (normal + suspicious). (SE – sensitivity, SP – specificity, PR – precision, F – F-measure).

element	SE [%]	SP [%]	PR [%]	F [%]
pH	36 (28–44)	77 (74–80)	16 (12–21)	22 (17–27)
BE	43 (30–50)	77 (74–80)	11 (9– 6)	18 (14–24)
BDecf	43 (25–50)	77 (74–79)	7 (4–10)	13 (7–17)
MV_{ce}	77 (69–83)	83 (80–85)	38 (32–42)	50 (44–55)
MV_{bm}	43 (30–50)	77 (74–80)	11 (9–16)	18 (14–24)
$AS_{5\ min.}$	31 (18–40)	76 (73–79)	4 (3– 6)	7 (5–11)
LCM_{ce}	67 (60–72)	89 (86–92)	67 (60–73)	66 (62–70)

Table 8.9: Classification results for LCR_{HM} with respect to different elements (components). The overall measure that asses all classes (ACC – accuracy).

element	ACC [%]
pH	62 (58–65)
BE	64 (61–68)
BDecf	57 (54–60)
MV_{ce}	50 (48–53)
MV_{bm}	61 (57–64)
$AS_{5\ min.}$	58 (55–61)
LCM_{ce}	49 (46–52)
LCR_{HM}	71 (69–75)

Table 8.10: The confusion matrix of estimated latent classes for LCM_{HM} and LCR_{HM} models. $n_r/s_r/p_r$ normal/suspicious/pathological from the LCR_{HM} and $n_m/s_m/p_m$ normal/suspicious/pathological from the LCM_{HM}.

	n_r	s_r	p_r
n_m	353	20	106
s_m	5	26	9
p_m	0	13	20

ability of logistic regression to learn the μ_{ic} was tested using the 50×4 CV and shown in Table 8.8. Below we present the estimated latent class μ_{ic} with respect to the classes determined by pH, see Table 8.11. We can observe that the pathological pH class (p) is largely in the suspicious (s′) (27 records) or pathological (p′) (26 records) class given by the LCR_{HM}.

Table 8.11: Confusion matrix for the μ_{ic} estimated by LCR_{HM} to classes determined by pH. (n′/s′/p′ determined by LCR_{HM}, n/s/p determined by pH).

	n′	s′	p′
n	335	8	96
s	15	24	13
p	8	27	26

In order to highlight the strength of the LCR_{HM} to model the uncertainty of evaluation of labour and to give an example of discrepancy of pH to the main components we show the details of the last row of Table 8.11, i.e. we analyse the pathological examples determined by pH. Following our notation we define the false negatives $FN_{pn'}$ (pH pathological, LCR_{HM} normal) and $FN_{ps'}$ (pH pathological, LCR_{HM} suspicious) and true positives $TP_{pp'}$. In Table 8.12 we present detailed results for each category for the main components of the model.

Table 8.12: Confusion matrices for the false negatives $FN_{pn'}$ (pH pathological, LCR_{HM} normal) and $FN_{ps'}$ (pH pathological, LCR_{HM} suspicious) and true positives $TP_{pp'}$. The false negatives and true positives were obtained by comparing the LCR_{HM} to pathological class determined by pH, see the last row of Table 8.11. (n′/s′/p′ – normal/suspicious/pathological for individual components of the model).

(a) $FN_{pn'}$ (8 recs.)

	n′	s′	p′
MV_{bm}	1	7	0
$AS_{5\ min.}$	6	2	0
LCM_{ce}	5	3	0

(b) $FN_{ps'}$ (27 recs.)

	n′	s′	p′
MV_{bm}	0	10	17
$AS_{5\ min.}$	10	13	4
LCM_{ce}	6	20	1

(c) $TP_{pp'}$ (26 recs.)

	n′	s′	p′
MV_{bm}	0	5	21
$AS_{5\ min.}$	10	12	4
LCM_{ce}	0	1	25

For the $FN_{pn'}$ there were 8 records determined by pH as pathological but the LCR_{HM} resulted to normal class. In Table 8.12a we can see that the MV_{bm} was not pathological for any case, either BDecf or BE was normal/suspicious. The presence of normal and suspicious only holds also for the $AS_{5\ min.}$ and LCM_{ce}. For the $FN_{ps'}$ there were 27 record determined as pathological by pH. In Table 8.12b we can see the higher occurrence of pathological for MV_{bm} but low number of pathological for $AS_{5\ min.}$ and LCM_{ce} highlighting the discrepancy of evaluation. The true positive $TP_{pp'}$ are present in Table 8.12c. The number of pathological given by MV_{bm} and LCM_{ce} is prevalent in contrast to the $AS_{5\ min.}$, which proportion remained almost the same as for the $FN_{ps'}$.

Comparison of the model to the supervised learning (pH based) The classification using the pH is the most common and followed approach. In this chapter we showed that this approach is imprecise; however, for the sake of completeness, we present the results of the LCR_{HM} regarding the pH classification below and compare the results to the supervised learning of Naive Bayes. The procedure for classification was the same as describe above only for the test set the pH labels were used. Note that, in contrast to Chapter 7 (Table 7.4), the Naive Bayes was learned with the three classes (normal/suspicious/pathological). The confusion matrices estimated on the test set are present in Table 8.13 and the results of classification are present in Table 8.14.

Table 8.13: Comparison of classification to pH for the latent class regression and supervised learning using Naive Bayes. The confusion matrices are computed for the test set on 50×4 fold CV. n/s/p – actual values, n′/s′/p′ – predicted values, n – normal, s – suspicious, p – pathological

(a) LCR_{HM}

	n′	s′	p′
n	78.7	6.7	24.4
s	8.3	0.9	3.8
p	8.4	1.3	5.6

(b) Naive Bayes

	n′	s′	p′
n	71.2	13.2	25.4
s	7	1.3	4.7
p	5.7	1.8	7.7

Table 8.14: Results of the LCR_{HM} and Naive Bayes (NB) to classes determined by pH. The results are presented as pathological vs. (normal + suspicious).)

	SE [%]	SP [%]	PR [%]	F [%]
LCR_{HM}	36 (28–44)	77 (74–80)	16 (12–21)	22 (17–27)
NB	50 (48–57)	79 (70–81)	23 (16–29)	30 (24–38)

The results of LCR_{HM} are worse in all measures than the results of Naive Bayes though this was expected since the LCR_{HM} was learned with the hierarchical model and tested on pH labels.

8.9 Discussion and conclusion

In this chapter we presented a novel hierarchical model for fetal heart rate evaluation. We showed that the model is able to overcome the uncertainty in the biochemical markers (pH, BE, and BDecf). The model does not weigh individual markers equally and, in addition, minimizes the inter-observer variability of clinical evaluation of CTG using the latent class model.

The most difficult examples When the pH is considered as a singular discriminant between classes there are a number of records that are constantly misclassified in the classical learning scenario. The same misclassified examples across one class (normal to pathological and vice versa) for unsupervised learning, supervised learning, and clinical evaluation were considered as difficult. The lowest number of difficult (false negative) records was for the clinical evaluation. There was no common characteristic of the difficult examples from the clinical point of view nor was there a technical factor that would be significant. In addition, the difficult false negatives were not pathological when a BE or BDecf was considered. From the supervised point of view (Naive Bayes, SVM, C4.5, and Adaboost) the difficult false negatives and easy true positives were almost identical while the false positives were different in majority. Probably, the high ratio of false positive results is caused by class imbalance where the minority class were oversampled using SMOTE. The number of false positives could be reduced by addition of pathological examples or by using another technique (an alternative to the SMOTE).

Clustering of fetal heart rate We described the fetal heart rate with comprehensive set of features and used Gaussian mixture model (GMM) to model the feature space. The best model was found to have three components, $\hat{m} = 3$. The GMM is an unsupervised technique and helps to discover the underlying structure of the data. The created model best corresponded to the latent class model of clinical evaluation, the LCM_{ce}.

Latent class regression using the hierarchical model The classification using the latent regression analysis yielded the best results. The features were modelled with logistic regression (LCR_{HM} model) and the latent class was simultaneously estimated with coefficients of logistic regression. For the pathological latent class the LCR_{HM} most inclined to the LCM_{ce} model. This behaviour was expected since also the GMM model best corresponded to the LCM_{ce}. On the other hand, the normal and suspicious latent class were better modelled using MV_{bm} and $\text{AS}_{5\text{ min.}}$.

The results of sensitivity 66% (59–71) specificity 89% (86–92), precision 66% (59–73), and F-measure 65% (60–70) were superior to the results when a latent class was estimated first (LCM_{HM} model) and, more importantly, the results were superior to those achieved when a pH was used as the singular, fool proof, marker of labour outcome. The comparison to other works is impossible since the concept of the model is novel and has not been used in any other work so far. More importantly, in other works a much smaller and ad-hoc created databases were used. Most of the works use a database size of lower than 100 records. It is unlikely that database of this size would reflect the inter-individual differences in very complex fetal behaviour and mechanism of their defence to the labour stress. Even the database presented in our work might be insufficient.

The hierarchical model and its use The hierarchical model presented in this chapter is simple and intuitive. The model could be easily used in any other work. Only the probabilities of the model $\mathbf{A}(t_{28})^{bm}$, $\mathbf{A}(t_{28})^{as}$, and $\mathbf{A}(t_{28})^{ce}$ and prevalence of classes $\mathbf{p} = \{P(normal), P(suspicious), P(pathological)\}$ are required. The limitation of the model is the need of multiple clinical evaluations of CTG though the annotation of CTG by clinicians is not that time consuming (approximately 4 seconds for 30 min. of CTG). The hierarchical model provided encouraging results for automatic classification of CTG records and overcome the variability in different markers used for labour outcome evaluation. We showed that it is possible to automatically analyse the CTG and provide information, which could be used as a support for clinical decision making.

Chapter 9

Conclusion and discussion

In this work we described and implemented a novel classification system for fetal well-being evaluation. The system includes preprocessing of fetal heart rate and its analysis using a comprehensive set of features. Further, the system combines different sources of information in order to properly evaluate fetal well-being during labour. The developed system considers biochemical markers, Apgar score, and clinical evaluation of CTG as imprecise sources of information and is able to overcome discrepancies between them. This work clearly met the goals set in the Introduction (Chapter 1). We summarize our objectives in Section 9.1 and in the next Section 9.2 we detail the thesis contributions to the state of the art of CTG field.

9.1 Accomplishment of the objectives

In this section we summarize the achieved goals of the thesis stated in Section 1.1 (Chapter 1). The objectives of this thesis were as follows:

1. **We performed a critical analysis of used databases and algorithms.** In the review (Chapter 3) we showed that in most of the works a small, ad-hoc created, databases are used. Even more, in almost every work different criteria are applied for division into target classes. We showed that classification performance decrease with increasing data size.

2. **We introduced the first open-access database** for research on intrapartum CTG signal processing and analysis in Chapter 4. The database is reasonably large and allows researches to develop algorithms/methods for CTG analysis and classification. Using the CTU-UHB database different approaches can be easily compared with one another in the objective fashion. We firmly believe that this unique database will stimulate the research in CTG processing and classification.

3. **We proposed a novel model for clinical evaluation of CTG in Chapter 6**. The model better accounts for high inter-observer variability and is able to estimate the unknown truth from multiple noisy clinical annotations. The model also allows us to analyse different number of classes than clinicians commonly use and provide more stable results than majority voting.

4. **We classified the FHR features using pH** that was used as a discriminator between two types of FHR records (normal and abnormal) in Chapter 7. We showed that frequency based features are good descriptors of abnormal fetal heart rate. We performed the classification using different techniques and provided unique results on the largest database.

5. **We designed and developed a novel hierarchical model** for fetal heart rate evaluation in Chapter 8. The model defines the labour outcome as a mixture of biochemical markers (pH, BE, BDecf), Apgar score, and latent class model of clinical evaluation of CTG. The model is able to overcome the discrepancy between the individual components and model the imprecise

definition of labour outcome. Moreover, the model provides accurate information about fetal well-being and relates the FHR patterns to adverse labour outcome using the logistic regression.

9.2 Scientific contributions

In addition to the presented achievements the thesis contributed scientifically to the CTG analysis, evaluation, and classification as follows:

Analysis of clinical evaluation We performed the largest study on clinical evaluation (largest in terms when both, number of clinicians and number of evaluated records are considered). We gathered clinical evaluation from nine practising clinicians, where each clinician evaluated 634 records (approximately 691 hours of CTG records). We showed that there is a large inter and intra observer variability and, more importantly, even with a large number of clinicians, the consensus (simple majority voting) can not be reached. This problem we refer to as stability of majority voting and, in Chapter 6, we showed how to improve the clinicians' consensus (stability of consensus) using a latent class model.

Latent class analysis of clinical evaluation We contributed to the unresolved controversy of how many classes should be used for CTG evaluation in Chapter 6. We showed that clinicians evaluate the CTG into 4 classes, despite the fact that they should use 3 classes based on FIGO guidelines. The difference between 3-tier and 4-tier classes lies in better separation of pathological records from the other ones. In other words, there is a clear pathological group for which there is a good agreement among clinicians; for the other classes the evaluation is more diverse and splitting these classes to more and more smaller classes would not contribute in lowering clinicians variability.

Link between FHR features and clinical evaluation In Chapter 6 we showed that clinical evaluation was in accordance to clinical features such as decelerations and baseline. Intuitively, these features must perform the best. The clinical evaluation was not significant to the short term variability features, which are impossible to assess visually. On the other hand the quantity of frequency and nonlinear found significant suggests that the 'intuition' based part of the decision process is rather large. The general approach to the FHR/CTG assessment is indeed based on the official FIGO guidelines. But the guidelines contain crisp and clear thresholds and rules which are difficult to precisely adhere to in a clinical setting.

Clustering of fetal behaviour via FHR We showed that fetal behaviour represented by fetal heart rate could be the best quantized into three categories (Chapter 8). However, the three groups (clusters) of FHR features, represented by a mixture of multivariate Gaussian model, only vaguely corresponded to the biochemical markers. The correspondence was better for clinical evaluation. This supports the conclusion that the link between FHR and biochemical markers is imprecise.

Methodology of database development and annotation We designed and developed the database based on a new methodology that is also proposed for the development of similar databases that can be used for extraction of medical knowledge, classification, and design and testing of new methods and algorithms. We also developed a new software (the CTGAnnotator) to obtain clinical annotations to the existing database. The CTGAnnotator and process of annotation used the methodology designed for annotation of clinical data.

9.3 Future work

When directing further research many works use a similar sentence that could be formulated as follows: "we obtained promising results for future research, which should be confirmed on a larger database". Even though we worked with one of the largest database in the field and provided unique results we must use the same words. More data is needed to confirm our encouraging results. In particular, the estimated trend (see Figure 3.2) between data size and quality of results would be of great interest. Another direction would be to further study the difficult (misclassified) records from a clinical point of view and try to discover an underlying relationship.

In the design of the novel hierarchical model we aimed to use simple techniques in order not to encapsulate the model into complicated structure and provide a clear picture of its interpretation capabilities. For the hierarchical model we used the latent class model when we categorized the pH into three classes. Possibly the latent trait model that would use the original continuous pH might offer additional information. Another improvement of the model might be gained by using latent class model of multiple Apgar score evaluations, i.e. to obtain several estimates of Apgar score from practitioners. On the input side of the model, there could be benefit of using other features to describe the complex behaviour of fetus. Especially the multi-fractal features seems to be promising lately. For the feature extraction another technique could be used instead of principal component analysis. A technique that would be able to also map a nonlinear relationship between the individual features and for the classification part a more powerful classifier instead of logistic regression could be employed. A better classifier or combination of classifiers might provide even better results than those achieved.

Bibliography

P. Abry, S. Roux, V. Chudáček, P. Borgnat, P. Gonçalves, and M. Doret. Hurst Exponent and IntraPartum Fetal Heart Rate: Impact of Decelerations. In *26th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6, 2013.

ACOG. American College of Obstetricians and Gynecologists Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. *Obstet Gynecol*, 114(1): 192–202, Jul 2009.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akadémiai Kiado.

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 Syst Rev*, 3(3):CD006066, 2006.

A. Alonso-Betanzos, B. Guijarro-Berdiñas, V. Moret-Bonillo, and S. López-Gonźalez. The NST-EXPERT project: the need to evolve. *Artif Intell Med*, 7(4):297–313, Aug 1995.

I. Amer-Wåhlin. *Fetal ECG waveform analysis for intrapartum monitoring*. PhD thesis, Dept. Obstetrics and Gynecology, Lund University Hospital, 2003.

I. Amer-Wåhlin and K. Maršál. ST analysis of fetal electrocardiography in labor. *Seminars in Fetal and Neonatal Medicine*, 16(1):29–35, 2011.

I. Amer-Wåhlin, C. Hellsten, H. Norén, H. Hagberg, A. Herbst, I. Kjellmer, H. Lilja, C. Lindoff, M. Månsson, L. Mårtensson, P. Olofsson, A. Sundström, and K. Maršál. Cardiotocography only versus cardiotocography plus ST analysis of fetal electrocardiogram for intrapartum fetal monitoring: a Swedish randomised controlled trial. *Lancet*, 358(9281):534–538, Aug 2001.

L. Armstrong and B. J. Stenson. Use of umbilical cord blood gas analysis in the assessment of the newborn. *Arch Dis Child Fetal Neonatal Ed*, 92(6):F430–F434, Nov 2007.

N. Badawi, J. J. Kurinczuk, J. M. Keogh, L. M. Alessandri, F. O'Sullivan, P. R. Burton, P. J. Pemberton, and F. J. Stanley. Intrapartum risk factors for newborn encephalopathy: the Western Australian case-control study. *BMJ*, 317(7172):1554–1558, Dec 1998.

P. C. A. M. Bakker, G. J. Colenbrander, A. A. Verstraeten, and H. P. V. Geijn. The quality of intrapartum fetal heart rate monitoring. *Eur J Obstet Gynecol Reprod Biol*, 116(1):22–27, Sep 2004.

M. Beaulieu, J. Fabia, and B. Leduc. The reproducibility of intrapartum cardiotocogram assessments. *Canadian Medical Association Journal*, 127(3):214–216, 1982.

S. Berglund, C. Grunewald, H. Pettersson, and S. Cnattingius. Risk factors for asphyxia associated with substandard care during labor. *Acta Obstet Gynecol Scand*, 89(1):39–48, 2010.

J. Bernardes and D. Ayres-De-Campos. The persistent challenge of foetal heart rate monitoring. *Current Opinion in Obstetrics and Gynecology*, 22(2):104–109, 2010.

J. Bernardes, C. Moura, J. P. de Sa, and L. P. Leite. The Porto system for automated cardiotocographic signal analysis. *J Perinat Med*, 19(1-2):61–65, 1991.

J. Bernardes, A. Costa-Pereira, D. A. de Campos, H. P. van Geijn, and L. Pereira-Leite. Evaluation of interobserver agreement of cardiotocograms. *Int J Gynaecol Obstet*, 57(1):33–37, Apr 1997.

J. Bernardes, D. A. de Campos, A. Costa-Pereira, L. Pereira-Leite, and A. Garrido. Objective computerized fetal heart rate analysis. *Int J Gynaecol Obstet*, 62(2):141–147, Aug 1998.

J. Bernardes, H. Gonçalves, D. Ayres-De-Campos, and A. Rocha. Sex differences in linear and complex fetal heart rate dynamics of normal and acidemic fetuses in the minutes preceding delivery. *Journal of Perinatal Medicine*, 37(2):168–176, 2009.

C. Bishop. *Neural networks for pattern recognition*. Clarendon Press, Oxford, 1995.

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):378.e1–378.e5, Oct 2011.

E. Blix, O. Sviggum, K. S. Koss, and P. Oian. Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts. *BJOG*, 110(1): 1–5, Jan 2003.

P. J. Boland. Majority systems and the Condorcet jury theorem. *The Statistician*, 38 (3):181–189, 1989.

M. Brennan, M. Palaniswami, and P. Kamen. Do existing measures of Poincaré plot geometry reflect nonlinear features of heart rate variability? *IEEE Trans Biomed Eng*, 48(11):1342–1347, Nov 2001.

A. G. Cahill, K. A. Roehl, A. O. Odibo, and G. A. Macones. Association and prediction of neonatal acidemia. *Am J Obstet Gynecol*, 207(3):206.e1–206.e8, Sep 2012.

L. K. Callaway, K. Lust, and H. D. McIntyre. Pregnancy outcomes in women of very advanced maternal age. *Aust N Z J Obstet Gynaecol*, 45(1):12–16, Feb 2005.

H. Cao, D. E. Lake, I. Ferguson, J. E., C. A. Chisholm, M. P. Griffin, and J. R. Moorman. Toward quantitative fetal heart rate monitoring. *IEEE Trans Biomed Eng*, 53(1):111–118, 2006. ISSN 0018-9294.

L. Cao. Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D*, 110:43–50, 1997.

M. Cesarelli, M. Romano, and P. Bifulco. Comparison of short term variability indexes in cardiotocographic foetal monitoring. *Comput Biol Med*, 39(2):106–118, Feb 2009.

M. Cesarelli, M. Romano, M. Ruffo, P. Bifulco, G. Pasquariello, and A. Fratini. PSD modifications of FHRV due to interpolation and CTG storage rate. *Biomedical Signal Processing and Control*, 6(3):225–230, 2011.

D. G. Chaffin, C. C. Goldberg, and K. L. Reed. The dimension of chaos in the fetal heart rate. *Am J Obstet Gynecol*, 165(5 Pt 1):1425–1429, Nov 1991.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

H.-Y. Chen, S. P. Chauhan, C. V. Ananth, A. M. Vintzileos, and A. Z. Abuhamad. Electronic fetal heart rate monitoring and its relationship to neonatal and infant mortality in the United States. *Am J Obstet Gynecol*, 204(6):491.e1–491.10, Jun 2011.

V. Chudáček, J. Spilka, P. Janků, M. Koucký, L. Lhotská, and M. Huptych. Automatic evaluation of intrapartum fetal heart rate recordings: A comprehensive analysis of useful features. *Physiological Measurement*, 32: 1347–1360, 2011.

V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych, and L. Lhotská. Open access intrapartum CTG database. *BMC Pregnancy and Childbirth*, Manuscript submitted for publication[1], 2013.

[1]For the purpose of review the submitted article is available at: `http://bio.felk.cvut.cz/~spilkaj/2013_Chudacek_BMC_Preg_database_submitted.pdf`

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):109–116, May 2001.

T. K. Chung, M. P. Mohajer, Z. J. Yang, A. M. Chang, and D. S. Sahota. The prediction of fetal acidosis at birth by computerised analysis of intrapartum cardiotocography. *Br J Obstet Gynaecol*, 102(6):454–460, Jun 1995.

D. V. Cicchetti and A. R. Feinstein. High agreement but low kappa: II. Resolving the paradoxes. *Journal of clinical epidemiology*, 43(6):551–558, 1990.

J. Cleary-Goldman, M. Negron, J. Scott, R. A. Downing, W. Camann, L. Simpson, and P. Flood. Prophylactic ephedrine and combined spinal epidural: maternal blood pressure and fetal heart rate patterns. *Obstet Gynecol*, 106(3):466–472, Sep 2005.

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.e1–47.e5, Jul 2011.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46, 1960.

J. Coletta, E. Murphy, Z. Rubeo, and C. Gyamfi-Bannerman. The 5-tier system of assessing fetal heart rate tracings is superior to the 3-tier system in identifying fetal acidemia. *Am J Obstet Gynecol*, 206(3): 226.e1–226.e5, Mar 2012.

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):464.e1–464.e6, Nov 2009.

M. Costa, A. L. Goldberger, and C.-K. Peng. Multiscale entropy analysis of biological signals. *Phys Rev E Stat Nonlin Soft Matter Phys*, 71(2 Pt 1):021906, Feb 2005.

F. Cunningham. *Williams Obsterics*, chapter 18. Intrapartum assesment, pages –. Mc Graw Hill, 2005.

R. Czabanski, M. Jezewski, J. Wrobel, J. Jezewski, and K. Horoba. Predicting the risk of low-fetal birth weight from cardiotocographic signals using ANBLIR system with deterministic annealing and epsilon-insensitive learning. *IEEE Trans Inf Technol Biomed*, 14(4):1062–1074, Jul 2010.

R. Czabanski, J. Jezewski, A. Matonia, and M. Jezewski. Computerized analysis of fetal heart rate signals as the predictor of neonatal acidemia. *Expert Systems with Applications*, 39(15):11846–11860, 2012.

E. d'Aloja, M. Müller, F. Paribello, R. Demontis, and A. Faa. Neonatal asphyxia and forensic medicine. *J Matern Fetal Neonatal Med*, 22 Suppl 3:54–56, 2009.

J. Davis and M. Goadrich. The Relationship Between Precision-Recall and ROC Curves. In *In ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM Press, 2006.

G. S. Dawes, G. H. Visser, J. D. Goodman, and C. W. Redman. Numerical analysis of the human fetal heart rate: the quality of ultrasound records. *Am J Obstet Gynecol*, 141(1):43–52, Sep 1981.

G. S. Dawes, C. R. Houghton, and C. W. Redman. Baseline in human fetal heart-rate records. *Br J Obstet Gynaecol*, 89(4):270–275, Apr 1982a.

G. S. Dawes, C. R. Houghton, C. W. Redman, and G. H. Visser. Pattern of the normal human fetal heart rate. *Br J Obstet Gynaecol*, 89(4):276–284, Apr 1982b.

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

G. S. Dawes, M. Moulden, and C. W. Redman. Improvements in computerized fetal heart rate analysis antepartum. *J Perinat Med*, 24(1):25–36, 1996.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28:20–28, 1979.

D. A. de Campos and J. Bernardes. Comparison of fetal heart rate baseline estimation by SisPorto 2.01 and a consensus of clinicians. *Eur J Obstet Gynecol Reprod Biol*, 117(2):174–178, Dec 2004.

D. A. de Campos and J. Bernardes. Twenty-five years after the FIGO guidelines for the use of fetal monitoring: Time for a simplified approach? *International Journal of Gynecology & Obstetrics*, 110(1):1 – 6, 2010. ISSN 0020-7292.

D. A. de Campos, J. Bernardes, K. Marsal, C. Nickelsen, L. Makarainen, P. Banfield, P. Xavier, and I. Campos. Can the reproducibility of fetal heart rate baseline estimation be improved? *Eur J Obstet Gynecol Reprod Biol*, 112(1):49–54, Jan 2004.

D. A. 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. *Journal of Perinatal Medicine*, 36(3):260–264, 2008.

D. A. de Campos, A. Ugwumadu, P. Banfield, P. Lynch, P. Amin, D. Horwell, A. Costa, C. Santos, J. Bernardes, and K. Rosén. A randomised clinical trial of intrapartum fetal monitoring with computer analysis and alerts versus previously available monitoring. *BMC Pregnancy Childbirth*, 10:71, 2010.

J. de Haan, J. van Bemmel, B. Versteeg, A. Veth, L. Stolte, J. Janssens, and T. Eskes. Quantitative evaluation of fetal heart rate patterns. I. Processing methods. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 1(3):95–102, 1971. cited By (since 1996) 13.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

D. Devane and J. Lalor. Midwives' visual interpretation of intrapartum cardiotocographs: intra- and inter-observer agreement. *J Adv Nurs*, 52(2):133–141, Oct 2005.

J. Diebolt and C. P. Robert. Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375, 1994. ISSN 00359246.

D. K. Donker, H. P. van Geijn, and A. Hasman. Interobserver variation in the assessment of fetal heart rate recordings. *Eur J Obstet Gynecol Reprod Biol*, 52(1):21–28, Nov 1993.

M. Doret, H. Helgason, P. Abry, P. Gonçalvès, C. Gharib, and P. Gaucherand. Multifractal analysis of fetal heart rate variability in fetuses with and without severe acidosis during labor. *American Journal of Perinatology*, 28(4):259–266, 2011. ISSN 07351631.

V. Doria, A. T. Papageorghiou, A. Gustafsson, A. Ugwumadu, K. Farrer, and S. Arulkumaran. Review of the first 1502 cases of ECG-ST waveform analysis during labour in a teaching hospital. *BJOG*, 114(10):1202–1207, Oct 2007.

E. R. Dougherty, C. Sima, B. Hanczar, U. M. Braga-Neto, et al. Performance of error estimators for classification. *Current Bioinformatics*, 5(1):53–67, 2010.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000. ISBN 0471056693.

O. Dupuis and A. Simon. [Fetal monitoring during the active second stage of labor]. *J Gynecol Obstet Biol Reprod (Paris)*, 37 Suppl 1:S93–100, Feb 2008.

J. C. Echeverria, B. R. Hayes-Gill, J. A. Crowe, M. S. Woolfson, and G. D. H. Croaker. Detrended fluctuation analysis: a suitable method for studying fetal heart rate variability? *Physiol Meas*, 25(3):763–774, Jun 2004.

J. P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617–656, 1985.

B. Efron. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426): 463–475, 1994.

B. Efron. Second thoughts on the bootstrap. *Statistical Science*, 18:135–140, 2003.

A. Eke, P. Herman, L. Kocsis, and L. R. Kozak. Fractal characterization of complexity in temporal physiological signals. *Physiol Meas*, 23(1):R1–38, Feb 2002.

C. Elliott, P. Warrick, E. Graham, and E. Hamilton. Graded classification of fetal heart rate tracings: association with neonatal metabolic acidosis and neurologic morbidity. *American Journal of Obstetrics and Gynecology*, 202(3):258.e1–258.e8, 2010.

K. Evans, A. S. Rigby, P. Hamilton, N. Titchiner, and D. M. Hall. The relationships between neonatal encephalopathy and cerebral palsy: a cohort study. *J Obstet Gynaecol*, 21(2):114–120, Mar 2001.

A. R. Feinstein and D. V. Cicchetti. High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549, 1990.

C. S. Felgueiras, J. P. de Sá, J. Bernardes, and S. Gama. Classification of foetal heart rate sequences based on fractal features. *Med Biol Eng Comput*, 36(2):197–201, Mar 1998.

M. Ferrario, M. G. Signorini, and S. Cerutti. Complexity analysis of 24 hours heart rate variability time series. *Conf Proc IEEE Eng Med Biol Soc*, 6:3956–3959, 2004.

M. Ferrario, M. G. Signorini, and G. Magenes. Complexity analysis of the fetal heart rate for the identification of pathology in fetuses. In *Proc. Computers in Cardiology*, pages 989–992, 2005.

FIGO. Guidelines for the Use of Fetal Monitoring. *International Journal of Gynecology & Obstetrics*, 25: 159–167, 1986.

M. Finster and M. Wood. The Apgar score has survived the test of time. *Anesthesiology*, 102(4):855–857, Apr 2005.

J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2004.

A. K. Formann and D. Böhning. Re: Insights into latent class analysis of diagnostic test performance. *Biostatistics*, 9(4):777–778, Oct 2008.

A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2):1134–1140, 1986.

Y. Freund and R. E. Schapire. Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

B. Fulcher, A. Georgieva, C. Redman, and N. Jones. Highly comparative fetal heart rate analysis. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 3135 –3138, 28 2012-sept. 1 2012.

R. Furlan, S. Guzzetti, W. Crivellaro, S. Dassi, M. Tinelli, G. Baselli, S. Cerutti, F. Lombardi, M. Pagani, and A. Malliani. Continuous 24-hour assessment of the neural regulation of systemic arterial pressure and RR variabilities in ambulant subjects. *Circulation*, 81(2):537–547, Feb 1990.

A. Galka. *Topics in nonlinear time series analysis, with implications for EEG analysis, in: Advanced Series in Nonlinear Dynamics*. World Scientific, Singapore., 2000.

A. Georgieva, S. J. Payne, M. Moulden, and C. W. G. Redman. Computerized fetal heart rate analysis in labor: detection of intervals with un-assignable baseline. *Physiol Meas*, 32(10):1549–1560, Oct 2011.

A. Georgieva, M. Moulden, and C. W. Redman. Umbilical cord gases in relation to the neonatal condition: the EveREst plot. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 168(2):155 – 160, 2013a. ISSN 0301-2115.

A. Georgieva, S. J. Payne, M. Moulden, and C. W. G. Redman. Artificial neural networks applied to fetal monitoring in labour. *Neural Computing and Applications*, 22(1):85–93, 2013b.

G. Georgoulas, C. Stylios, G. Nokas, and 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.

G. Georgoulas, C. D. Stylios, and P. P. Groumpos. Feature Extraction and Classification of Fetal Heart Rate Using Wavelet Analysis and Support Vector Machines. *International Journal on Artificial Intelligence Tools*, 15:411–432, 2005.

G. Georgoulas, C. D. Stylios, and P. P. Groumpos. Predicting the risk of metabolic acidosis for newborns based on fetal heart rate signal classification using support vector machines. *IEEE Trans Biomed Eng*, 53(5): 875–884, May 2006.

G. Georgoulas, D. Gavrilis, I. G. Tsoulos, C. D. Stylios, J. Bernardes, and P. P. Groumpos. Novel approach for fetal heart rate classification introducing grammatical evolution. *Biomedical Signal Processing and Control*, 2:69–79, 2007.

L. Glass. Introduction to controversial topics in nonlinear science: is the normal heart rate chaotic? *Chaos*, 19 (2):028501, Jun 2009.

K. G. Goldaber, L. Gilstrap, 3rd, K. J. Leveno, J. S. Dax, and D. D. McIntire. Pathologic fetal acidemia. *Obstet Gynecol*, 78(6):1103–1107, Dec 1991.

A. Goldberger, V. Bhargava, B. West, and A. Mandell. On the mechanism of cardiac electrical stability. The fractal hypothesis. *Biophysical Journal*, 48(3):525–528, 1985. ISSN 00063495.

A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–E220, Jun 2000.

A. L. Goldberger, L. A. N. Amaral, J. M. Hausdorff, P. C. Ivanov, C.-K. Peng, and H. E. Stanley. Fractal dynamics in physiology: alterations with disease and aging. *Proc Natl Acad Sci U S A*, 99 Suppl 1:2466–2472, Feb 2002.

H. Gonçalves, A. P. Rocha, D. A. de Campos, and J. Bernardes. Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery. *Med Biol Eng Comput*, 44(10):847–855, Oct 2006a.

H. Gonçalves, A. P. Rocha, D. A. de Campos, and J. Bernardes. Internal versus external intrapartum foetal heart rate monitoring: the effect on linear and nonlinear parameters. *Physiol Meas*, 27(3):307–319, Mar 2006b.

H. Gonçalves, A. Costa, D. A. de Campos, C. Costa-Santos, A. P. Rocha, and J. Bernardes. Comparison of real beat-to-beat signals with commercially available 4 Hz sampling on the evaluation of foetal heart rate variability. *Med Biol Eng Comput*, 51(6):665–676, Jun 2013.

N. A. Gough. Fractal analysis of foetal heart rate variability. *Physiol Meas*, 14(3):309–315, Aug 1993.

R. Govindan, J. Wilson, H. Preißl, H. Eswaran, J. Campbell, and C. Lowery. Detrended fluctuation analysis of short datasets: an application to fetal cardiac data. *Physica D: Nonlinear Phenomena*, 226(1):23–31, 2007.

E. M. Graatsma, B. C. Jacod, L. A. J. van Egmond, E. J. H. Mulder, and G. H. A. Visser. Fetal electrocardiography: feasibility of long-term fetal heart rate recordings. *BJOG*, 116(2):334–7; discussion 337–8, Jan 2009.

P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9:189–208, 1983.

K. Greene and R. Keith. K2 Medical System. [online], June 2002. Available: `http://ww.k2ms.com/` [Accessed 2013-06-20].

B. Guijarro-Berdiñas and A. Alonso-Betanzos. Empirical evaluation of a hybrid intelligent monitoring system using different measures of effectiveness. *Artif Intell Med*, 24(1):71–96, Jan 2002.

I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2006.

A. C. Guyton and J. E. Hall. *Textbook of medical physiology*. Saunders Book Company, 11 edition edition, 2005.

L. A. Haggerty. Continuous electronic fetal monitoring: contradictions between practice and research. *J Obstet Gynecol Neonatal Nurs*, 28(4):409–416, 1999.

M. A. Hall. Correlation-based Feature Selection for Machine Learning. Technical report, The University of Waikato, 1998.

E. Hamilton, P. Warrick, and D. O'Keeffe. Variable decelerations: do size and shape matter? *J Matern Fetal Neonatal Med*, 25(6):648–653, Jun 2012.

S. K. Hasley. Decision support and patient safety: the time has come. *Am J Obstet Gynecol*, 204(6):461–465, Jun 2011.

H. He and E. A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, Sept. 2009.

E. Heintz, T.-H. Brodtkorb, N. Nelson, and L.-A. Levin. The long-term cost-effectiveness of fetal monitoring during labour: a comparison of cardiotocography complemented with ST analysis versus cardiotocography alone. *BJOG*, 115(13):1676–1687, Dec 2008.

H. Helgason, P. Abry, P. Goncalves, C. Gharib, P. Gaucherand, and M. Doret. Adaptive Multiscale Complexity Analysis of Fetal Heart Rate. *IEEE Trans Biomed Eng*, 58:2186–2193, Mar 2011.

T. Higuchi. Approach to an irregular time series on the basis of the fractal theory. *Phys. D*, 31(2):277–283, 1988. ISSN 0167-2789.

J. B. Hill, J. M. Alexander, S. K. Sharma, D. D. McIntire, and K. J. Leveno. A comparison of the effects of epidural and meperidine analgesia during labor on fetal heart rate. *Obstet Gynecol*, 102(2):333–337, Aug 2003.

K. Hinshaw and A. Ullal. Peripartum and intrapartum assessment of the fetus. *Anaesthesia & Intensive Care Medicine*, 8(8):331–336, 2007.

P. Hopkins, N. Outram, N. Lofgren, E. C. Ifeachor, and K. G. Rosén. A Comparative Study of Fetal Heart Rate Variability Analysis Techniques. In *Proc. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS '06*, pages 1784–1787, 2006.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

I. Ingemarsson, A. Herbst, and K. Thorngren-Jerneck. Long term outcome after umbilical artery acidaemia at term birth: influence of gender and duration of fetal heart rate abnormalities. *Br J Obstet Gynaecol*, 104(10): 1123–1127, Oct 1997.

M. Jezewski, R. Czabański, J. Wróbel, and K. Horoba. Analysis of extracted cardiotocographic signal features to improve automated prediction of fetal outcome. *Biocybernetics and Biomedical Engineering*, 30(4):29–47, 2010.

L. Jimenez, R. Gonzalez, M. Gaitan, S. Carrasco, and C. Vargas. Computerized algorithm for baseline estimation of fetal heart rate. In *Proc. Computers in Cardiology*, pages 477–480, 2002.

A. Jiménez-González and C. J. James. Extracting sources from noisy abdominal phonograms: a single-channel blind source separation method. *Med Biol Eng Comput*, 47(6):655–664, Jun 2009.

A. Jiménez-González and C. J. James. Time-structure based reconstruction of physiological independent sources extracted from noisy abdominal phonograms. *IEEE Trans Biomed Eng*, 57(9):2322–2330, Sep 2010.

D. K. Kahaner, C. Moler, and S. Nash. *Numerical Methods and Software*. Prentice-Hall, 1989.

H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge Univesity Press, UK, 2. edition, 2004.

D. Kaplan and P. Staffin. Software for Heart Rate Variability. [online], April 1998. Available: `http://www.macalester.edu/~kaplan/hrv/doc` [Accessed 2013-07-16].

V. Kariniemi, J. Ahopelto, P. Karp, and T. Katila. The fetal magnetocardiogram. *Journal of Perinatal Medicine*, 2(3):214–216, 1974.

F. Kaspar and H. Schuster. Easily calculable measure of the complexity of spatiotemporal patterns. *Physical Review A*, 36:842–848, 1987.

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.

R. D. Keith, S. Beckley, J. M. Garibaldi, J. A. Westgate, E. C. Ifeachor, and K. R. Greene. A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. *Br J Obstet Gynaecol*, 102(9):688–700, Sep 1995.

I. Kiefer-Schmidt, M. Lim, A. Wacker-Gussmann, E. Ortiz, H. Abele, K. O. Kagan, R. Kaulitz, D. Wallwiener, and H. Preissl. Fetal magnetocardiography (fMCG): moving forward in the establishment of clinical reference data by advanced biomagnetic instrumentation and analysis. *J Perinat Med*, 40(3):277–286, Apr 2012.

A. Kikuchi, T. Shimizu, A. Hayashi, T. Horikoshi, N. Unno, S. Kozuma, and Y. Taketani. Nonlinear analyses of heart rate variability in normal and growth-restricted fetuses. *Early Hum Dev*, 82(4):217–226, Apr 2006.

W. Kinsner. Batch and real-time computation of a fractal dimension based on variance of a time series. Technical report, Department of Electrical & Computer Engineering, University of Manitoba, Winnipeg, Canada, 1994.

K. Kira and L. A. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256. Morgan Kaufmann Publishers Inc., 1992.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8.

F. Kovács, C. Horváth, A. T. Balogh, and G. Hosszú. Fetal phonocardiography–past and future possibilities. *Comput Methods Programs Biomed*, 104(1):19–25, Oct 2011.

G. A. B. Kro, B. M. Yli, S. Rasmussen, H. Norè n, I. Amer-Wåhlin, O. D. Saugstad, B. Stray-Pedersen, and K. G. Rosén. A new tool for the validation of umbilical cord acid-base data. *BJOG*, 117(12):1544–1552, Nov 2010.

T. Kupka, J. Wrobel, J. Jezewski, A. Gacek, and M. Jezewski. Evaluation of Fetal Heart Rate Baseline Estimation Method Using Testing Signals Based on a Statistical Model. In *Proc. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS '06*, pages 3728–3731, Aug. 2006.

J. V. 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 Obstet Gynecol Scand*, 87(3):300–306, 2008.

D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman. Sample entropy analysis of neonatal heart rate variability. *Am J Physiol Regul Integr Comp Physiol*, 283(3):R789–R797, Sep 2002.

C. Lee, A. amd Ulbricht and G. Dorffner. Application of artificial neural networks for detection of abnormal fetal heart rate pattern: a comparison with conventional algorithms. *Journal of Obstetrics & Gynecology*, 19 (5):482–485, 1999.

A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, IT-22 (1):75–81, 1976.

O. Lidegaard, L. M. Bøttcher, and T. Weber. Description, evaluation and clinical decision making according to various fetal heart rate patterns. Inter-observer and regional variability. *Acta Obstet Gynecol Scand*, 71(1): 48–53, Jan 1992.

T. H. Lin and C. M. Dayton. Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics*, 22(3):249–264, 1997.

C. List and R. E. Goodin. Epistemic democracy: generalizing the Condorcet jury theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001.

C. Liu, C. Liu, P. Shao, L. Li, X. Sun, X. Wang, and F. Liu. Comparison of different threshold values r for approximate entropy: application to investigate the heart rate variability between heart failure and healthy control groups. *Physiological Measurement*, 32(2):167, 2011.

A. Locatelli, M. Incerti, G. Paterlini, V. Doria, S. Consonni, C. Provero, and A. Ghidini. Antepartum and intrapartum risk factors for neonatal encephalopathy at term. *Am J Perinatol*, 27(8):649–654, Sep 2010.

F. K. Lotgering, H. C. Wallenburg, and H. J. Schouten. Interobserver and intraobserver variation in the assessment of antepartum cardiotocograms. *Am J Obstet Gynecol*, 144(6):701–705, Nov 1982.

J. A. Low. The current crisis in obstetrics. *J Obstet Gynaecol Can*, 27(11):1031–1037, Nov 2005.

A. Lynn and P. Beeby. Cord and placenta arterial gas analysis: the accuracy of delayed sampling. *Arch Dis Child Fetal Neonatal Ed*, 92(4):F281–F285, Jul 2007.

D. MacDonald, A. Grant, M. Sheridan-Pereira, P. Boylan, and I. Chalmers. The Dublin randomized controlled trial of intrapartum fetal heart rate monitoring. *Am J Obstet Gynecol*, 152(5):524–539, Jul 1985.

A. MacLennan. A template for defining a causal relation between acute intrapartum events and cerebral palsy: international consensus statement. *BMJ*, 319(7216):1054–1059, Oct 1999.

G. A. Macones, G. D. V. 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. *J Obstet Gynecol Neonatal Nurs*, 37(5):510–515, 2008.

Maeda, Utsu, Makio, Serizawa, Noguchi, Hamada, Mariko, and Matsumoto. Neural Network Computer Analysis of Fetal Heart Rate. *J Matern Fetal Investig*, 8(4):163–171, Dec 1998.

G. Magenes, M. G. Signorini, and D. Arduini. Classification of cardiotocographic records by neural networks. In *Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN 2000*, volume 3, pages 637–641, 2000.

G. Magenes, M. G. Signorini, M. Ferrario, L. Pedrinazzi, and D. Arduini. Improving the fetal cardiotocographic monitoring by advanced signal processing. In *Proc. 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 3, pages 2295–2298 Vol.3, 2003.

G. Magenes, M. Signorini, M. Ferrario, and F. Lunghi. 2CTG2: A new system for the antepartum analysis of fetal heart rate. In R. Magjarevic, T. Jarm, P. Kramar, and A. Zupanic, editors, *11th Mediterranean Conference on Medical and Biomedical Engineering and Computing 2007*, volume 16 of *IFMBE Proceedings*, pages 781–784. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-73044-6.

D. Maharaj. Intrapartum Fetal Resuscitation: A Review. *The Internet Journal of Gynecology and Obstetrics*, 9 (2), 2008.

G. L. Malin, R. K. Morris, and K. S. Khan. Strength of association between umbilical cord pH and perinatal and long term outcomes: systematic review and meta-analysis. *BMJ*, 340:c1471, 2010.

R. Manganaro, C. Mami, and M. Gemelli. The validity of the Apgar scores in the assessment of asphyxia at birth. *Eur J Obstet Gynecol Reprod Biol*, 54(2):99–102, Apr 1994.

G. Manis. Fast computation of approximate entropy. *Computer methods and programs in biomedicine*, 91(1): 48–54, 2008.

R. Mantel, H. P. van Geijn, F. J. Caron, J. M. Swartjes, E. E. van Woerden, and H. W. Jongsma. Computer analysis of antepartum fetal heart rate: 1. Baseline determination. *Int J Biomed Comput*, 25(4):261–272, May 1990a.

R. Mantel, H. P. van Geijn, F. J. Caron, J. M. Swartjes, E. E. van Woerden, and H. W. Jongsma. Computer analysis of antepartum fetal heart rate: 2. Detection of accelerations and decelerations. *Int J Biomed Comput*, 25(4):273–286, May 1990b.

S. McIntyre, D. Taitz, J. Keogh, S. Goldsmith, N. Badawi, and E. Blair. A systematic review of risk factors for cerebral palsy in children born at term in developed countries. *Developmental Medicine & Child Neurology*, (in press), Nov 2012.

G. McLachlan and D. Peel. *Finite Mixture Models*. New York, John Wiley & Sons., 2000.

D. A. Miller and L. A. Miller. Three-tier versus five-tier fetal heart rate classification systems. *Am J Obstet Gynecol*, 207(6):e8–9; author reply e9, Dec 2012.

NICE. National Collaborating Centre for Women's and Children's Health, commissioned by the National Institute for Health and Clinical Excellence. Intrapartum care. London: RCOG Press, 2007.

P. V. Nielsen, B. Stigsby, C. Nickelsen, and J. Nim. Computer assessment of the intrapartum cardiotocogram. II. The value of compared with visual assessment. *Acta Obstet Gynecol Scand*, 67(5):461–464, 1988.

H. Norén, I. Amer-Wåhlin, H. Hagberg, A. Herbst, I. Kjellmer, K. Maršál, P. Olofsson, and K. G. Rosén. Fetal electrocardiography in labor and neonatal outcome: data from the Swedish randomized controlled trial on intrapartum fetal monitoring. *Am J Obstet Gynecol*, 188(1):183–192, Jan 2003.

H. Norén, S. Blad, A. Carlsson, A. Flisberg, A. Gustavsson, H. Lilja, M. Wennergren, and H. Hagberg. STAN in clinical practice–the outcome of 2 years of regular use in the city of Gothenburg. *Am J Obstet Gynecol*, 195 (1):7–15, Jul 2006.

I. Nunes, D. Ayres-de Campos, C. Figueiredo, and J. Bernardes. An overview of central fetal monitoring systems in labour. *J Perinat Med*, 41(1):93–99, Jan 2013.

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):9913, Apr 2013.

C. P. F. O'Donnell, C. O. F. Kamlin, P. G. Davis, J. B. Carlin, and C. J. Morley. Interobserver variability of the 5-minute Apgar score. *J Pediatr*, 149(4):486–489, Oct 2006.

V. P. Oikonomou, J. Spilka, C. Stylios, and L. Lhotská. An adaptive method for the recovery of missing samples from FHR time series. In *26th International Symposium on Computer-Based Medical Systems (CBMS)*, 2013.

K. Ojala, M. Vääräsmäki, K. Mäkikallio, M. Valkama, and A. Tekay. A comparison of intrapartum automated fetal electrocardiography and conventional cardiotocography–a randomised controlled study. *BJOG*, 113(4): 419–423, Apr 2006.

K. Ojala, K. Mäkikallio, M. Haapsamo, H. Ijäs, and A. Tekay. Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies. *Acta Obstet Gynecol Scand*, 87(5): 536–540, 2008.

C. Oncken, H. Kranzler, P. O'Malley, P. Gendreau, and W. A. Campbell. The effect of cigarette smoking on fetal heart rate characteristics. *Obstet Gynecol*, 99(5 Pt 1):751–755, May 2002.

N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a Time Series. *Physical Review Letters*, 45(9):712+, 1980.

Y.-H. Pan, Y.-H. Wang, S.-F. Liang, and K.-T. Lee. Fast computation of sample entropy and approximate entropy in biomedicine. *Computer methods and programs in biomedicine*, 104(3):382–396, 2011.

N. Paneth, M. Bommarito, and J. Stricker. Electronic fetal monitoring and later outcome. *Clin Invest Med*, 16 (2):159–165, Apr 1993.

J. Pardey, M. Moulden, and C. W. G. Redman. A computer system for the numerical analysis of nonstress tests. *Am J Obstet Gynecol*, 186(5):1095–1103, May 2002.

J. Parer and E. Hamilton. Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation. *American Journal of Obstetrics and Gynecology*, 203(5):451.e1–451.e7, 2010.

J. T. Parer and T. Ikeda. A framework for standardized management of intrapartum fetal heart rate patterns. *Am J Obstet Gynecol*, 197(1):26.e1–26.e6, Jul 2007.

J. T. Parer, T. King, S. Flanders, M. Fox, and S. J. Kilpatrick. Fetal acidemia and electronic fetal heart rate patterns: is there evidence of an association? *J Matern Fetal Neonatal Med*, 19(5):289–294, May 2006.

J. T. Parer, T. Ikeda, and T. L. King. The 2008 National Institute of Child Health and Human Development report on fetal heart rate monitoring. *Obstet Gynecol*, 114(1):136–138, Jul 2009.

M. I. Park, J. H. Hwang, K. J. Cha, Y. S. Park, and S. K. Koh. Computerized analysis of fetal heart rate parameters by gestational age. *Int J Gynaecol Obstet*, 74(2):157–164, Aug 2001.

C. K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos*, 5(1):82–87, 1995.

H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (8):1226–1238, 2005.

M. S. Pepe and H. Janes. Insights into latent class analysis of diagnostic test performance. *Biostatistics*, 8(2): 474–484, Apr 2007.

J. Piéri, J. Crowe, B. Hayes-Gill, C. Spencer, K. Bhogal, and D. James. Compact long-term recorder for the transabdominal foetal and maternal electrocardiogram. *Medical and Biological Engineering and Computing*, 39(1):118–125, 2001.

V. Pierrat, N. Haouari, A. Liska, D. Thomas, D. Subtil, P. Truffert, and G. d'Etudes en Epidémiologie Périnatale. Prevalence, causes, and outcome at 2 years of age of newborn encephalopathy: population based study. *Arch Dis Child Fetal Neonatal Ed*, 90(3):F257–F261, May 2005.

S. Pincus. Approximate entropy (ApEn) as a complexity measure. *Chaos*, 5 (1):110–117, 1995.

S. M. Pincus and R. R. Viscarello. Approximate entropy: a regularity measure for fetal heart rate analysis. *Obstet Gynecol*, 79(2):249–255, Feb 1992.

R. J. Portman, B. S. Carter, M. S. Gaylord, M. G. Murphy, R. E. Thieme, and G. B. Merenstein. Predicting neonatal morbidity after perinatal asphyxia: a scoring system. *Am J Obstet Gynecol*, 162(1):174–182, Jan 1990.

J. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, 1992.

V. C. Raykar and S. Yu. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *Journal of Machine Learning Research*, 13:491–518, 2012.

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *The Journal of Machine Learning Research*, 11:1297–1322, 2010.

RCOG. Royal College of Obstetricians and Gynaecologists. The use of electronic fetal monitoring. Evidence-based clinical guidelines. RCOG Press, London, 2001.

J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol*, 278(6):H2039–H2049, Jun 2000.

R. J. Riley and J. W. Johnson. Collecting and analyzing cord blood gases. *Clin Obstet Gynecol*, 36(1):13–23, Mar 1993.

V. M. Roemer. How to determine and use base excess (BE) in perinatal medicine. *Z Geburtshilfe Neonatol*, 211 (6):224–229, Dec 2007.

D. Roj, T. Fuchs, T. Przybyla, M. Jezewski, A. Matonia, and A. Gacek. The influence of window size of autocorrelation function on fetal heart rate variability measurement using the Doppler ultrasound signal. *Journal of Medical Informatics and Technologies*, 12:111–116, 2008.

K. G. Rosén and K. Lindecrantz. STAN–the Gothenburg model for fetal surveillance during labour by ST analysis of the fetal electrocardiogram. *Clin Phys Physiol Meas*, 10 Suppl B:51–56, 1989.

K. G. Rosén and R. Luzietti. The fetal electrocardiogram: ST waveform analysis during labour. *J Perinat Med*, 22(6):501–512, 1994.

K. G. Rosén, I. Amer-Wåhlin, R. Luzietti, and H. Norén. Fetal ECG waveform analysis. *Best Pract Res Clin Obstet Gynaecol*, 18(3):485–514, Jun 2004.

K. G. Rosén, S. Blad, D. Larsson, H. Norén, and N. Outram. Assessment of the fetal bioprofile during labor by fetal ECG analysis. *Expert Review of Obstetrics & Gynecology*, 2(5):609–620, Sept. 2007. ISSN 1747-4108.

M. G. Ross. Labor and fetal heart rate decelerations: relation to fetal metabolic acidosis. *Clin Obstet Gynecol*, 54(1):74–82, Mar 2011.

E. Salamalekis, E. Hintipas, I. Salloum, G. Vasios, C. Loghis, N. Vitoratos, C. Chrelias, and G. Creatsas. Computerized analysis of fetal heart rate variability using the matching pursuit technique as an indicator of fetal hypoxia during labor. *J Matern Fetal Neonatal Med*, 19(3):165–169, Mar 2006.

S. Salzberg. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Min. Knowl. Discov.*, 1:317–328, 1997.

T. P. Sartwelle. Electronic fetal monitoring: a bridge too far. *J Leg Med*, 33(3):313–379, Jul 2012.

S. Schiermeier, H. Hatzmann, and J. Reinhard. The value of Doppler cardiotocogram computer analysis system 70 minutes before delivery. *Z Geburtshilfe Neonatol*, 212(5):189–193, Oct 2008a.

S. Schiermeier, S. P. 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. *BJOG*, 115(12):1557–1563, Nov 2008b.

B. S. Schifrin. The CTG and the timing and mechanism of fetal neurological injuries. *Best Pract Res Clin Obstet Gynaecol*, 18(3):437–456, Jun 2004.

T. Schreiber and A. Schmitz. Improved Surrogate Data for Nonlinearity Tests. *Phys. Rev. Lett.*, 77 (4):635–638, 1996.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978. ISSN 00905364.

E. Sheiner, A. Hadar, M. Hallak, M. Katz, M. Mazor, and I. Shoham-Vardi. Clinical significance of fetal heart rate tracings during the second stage of labor. *Obstet Gynecol*, 97(5 Pt 1):747–752, May 2001.

O. Sibony, J. Fouillot, M. Benaudia, A. Benhalla, P. Blot, and C. Sureau. Spectral analysis: a method for quantitating fetal heart rate variability. In H. van Geijn and F. Copray, editors, *A Critical Appraisal of Fetal Surveillance*, pages 325–332. New Your, Elsevier Science, 1994.

O. Siggaard-Andersen and R. Huch. The oxygen status of fetal blood. *Acta Anaesthesiol Scand Suppl*, 107: 129–135, 1995.

M. G. Signorini, G. Magenes, S. Cerutti, and D. Arduini. Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. *IEEE Trans Biomed Eng*, 50(3):365–374, Mar 2003.

S. Siira, T. Ojala, E. Ekholm, T. Vahlberg, S. Blad, and K. G. Rosén. Change in heart rate variability in relation to a significant ST-event associates with newborn metabolic acidosis. *BJOG*, 114(7):819–823, Jul 2007.

S. M. Siira, T. H. Ojala, T. J. Vahlberg, J. O. Jalonen, I. A. Välimäki, K. G. Rosén, and E. M. Ekholm. Marked fetal acidosis and specific changes in power spectrum analysis of fetal heart rate variability recorded during the last hour of labour. *BJOG*, 112(4):418–423, Apr 2005.

T. Singh, S. Sankaran, B. Thilaganathan, and A. Bhide. The prediction of intra-partum fetal compromise in prolonged pregnancy. *J Obstet Gynaecol*, 28(8):779–782, Nov 2008.

K. M. Sisco, A. G. Cahill, D. M. Stamilio, and G. A. Macones. Is continuous monitoring the answer to incidentally observed fetal heart rate decelerations? *J Matern Fetal Neonatal Med*, 22(5):405–409, May 2009.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, pages 1085–1092, 1995.

J. Spilka. Fetal Electrocardiogram Analysis. Master's thesis, Czech Technical University in Prague, 2009.

J. Spilka. Analysis of intrapartum fetal heart rate. Unpublished Technical report, Czech Technical University in Prague, August 2011.

J. Spilka, V. Chudáček, J. Kužílek, L. Lhotská, and M. Hanuliak. Detection of Inferior Myocardial Infarction: A Comparison of Various Decision Systems and Learning Algorithms. In *Computers in Cardiology*, volume 37, 2010.

J. Spilka, V. Chudáček, M. Koucký, L. Lhotská, M. Huptych, P. Janků, G. Georgoulas, and C. Stylios. Using nonlinear features for fetal heart rate classification. *Biomedical Signal Processing and Control*, 7(4):350–357, 2012.

J. Spilka, V. Chudáček, P. Janků, L. Hruban, M. Burša, M. Huptych, L. Zach, A. Hudec, M. Kacerovský, M. Koucký, L. Lhotská, M. Procházka, V. Korečko, J. Seget'a, and O. Šimetka. First step to automated obstetrics alarm system: Analysis of annotations derived from expert-obstetricians. *Methods of Information in Medicine*, Manuscript submitted for publication[2], 2013a.

J. Spilka, G. Georgoulas, P. Karvelis, V. P. Oikonomou, V. Chudáček, C. Stylios, L. Lhotská, and P. Janků. Automatic evaluation of FHR recordings from CTU-UHB CTG database. In M. Burša, S. Khuri, and M. Renda, editors, *Information Technology in Bio- and Medical Informatics*, Lecture Notes in Computer Science, pages 47–61. Springer Berlin Heidelberg, 2013b.

J. C. Sprott. *Chaos and Time-Series Analysis*. Oxford University Press, 2003. ISBN 0198508409.

P. J. Steer. Has electronic fetal heart rate monitoring made a difference. *Semin Fetal Neonatal Med*, 13(1):2–7, Feb 2008.

B. K. Strachan, W. J. van Wijngaarden, D. Sahota, A. Chang, and D. K. James. Cardiotocography only versus cardiotocography plus PR-interval analysis in intrapartum surveillance: a randomised, multicentre trial. FECG Study Group. *Lancet*, 355(9202):456–459, Feb 2000.

B. K. Strachan, D. S. Sahota, W. J. van Wijngaarden, D. K. James, and A. M. Chang. Computerised analysis of the fetal heart rate and relation to acidaemia at delivery. *BJOG*, 108(8):848–852, Aug 2001.

A. Sundström, D. Rosén, and K. Rosén. Fetal Surveillance - textbook. [online] Gothenburg, Sweden: Neoventa Medical AB, January 2000.

F. Takens. Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence*, 4:366–381, 1981.

Task-Force. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Eur Heart J*, 17(3):354–381, Mar 1996.

G. M. Taylor, G. J. Mires, E. W. Abel, S. Tsantis, T. Farrell, P. F. Chien, and Y. Liu. The development and validation of an algorithm for real-time computerised fetal heart rate monitoring in labour. *BJOG*, 107(9): 1130–1137, Sep 2000.

J. A. Thorp, J. E. Sampson, V. M. Parisi, and R. K. Creasy. Routine umbilical cord blood gas determinations? *Am J Obstet Gynecol*, 161(3):600–605, Sep 1989.

D. Tincello, S. White, and S. Walkinshaw. Computerised analysis of fetal heart rate recordings in maternal type I diabetes mellitus. *BJOG*, 108(8):853–857, Aug 2001.

M. D. Tommaso, V. Seravalli, A. Cordisco, G. Consorti, F. Mecacci, and F. Rizzello. Comparison of five classification systems for interpreting electronic fetal monitoring in predicting neonatal status at birth. *J Matern Fetal Neonatal Med*, 26(5):487–490, Mar 2013.

J. Uebersax. Latent Structure Analysis. [online], April 2010. Available: `http://www.john-uebersax.com/stat/` [Accessed 2013-06-20].

[2]For the purpose of review the submitted article is available at: `http://bio.felk.cvut.cz/~spilkaj/2013_Spilka_MIM_observer_var_submitted.pdf`

H. Valensise, D. Arduini, F. Giannini, R. Conforti, F. Giacomello, and C. Romanini. Role of antepartum computerized fetal heart rate analysis in the prediction of fetal distress during labor. *Eur J Obstet Gynecol Reprod Biol*, 73(2):129–134, Jun 1997.

L. Valentin, G. Ekman, P. E. Isberg, S. Polberger, and K. Maršál. Clinical evaluation of the fetus and neonate. Relation between intra-partum cardiotocography, Apgar score, cord blood acid-base status and neonatal morbidity. *Arch Gynecol Obstet*, 253(2):103–115, 1993.

P. Van Leeuwen, D. Geue, S. Lange, W. Hatzmann, and D. Grönemeyer. Changes in the frequency power spectrum of fetal heart rate in the course of pregnancy. *Prenat Diagn*, 23(11):909–916, Nov 2003.

P. Van Leeuwen, D. Cysarz, F. Edelhäuser, and D. Grönemeyer. Heart rate variability in the individual fetus. *Autonomic Neuroscience: Basic and Clinical*, 2013. doi: 10.1016/j.autneu.2013.01.005. Article in Press.

V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0387945598.

C. Vayssiere, V. Tsatsaris, O. Pirrello, C. Cristini, C. Arnaud, and F. Goffinet. Inter-observer agreement in clinical decision-making for abnormal cardiotocogram (CTG) during labour: a comparison between CTG and CTG plus STAN. *BJOG*, 116(8):1081–7; discussion 1087–8, Jul 2009.

E. Čech, Z. Hájek, K. Maršál, and B. Srp. *Porodnictví*. Grada Publishing, 2006.

R. Victory, D. Penava, O. D. Silva, R. Natale, and B. Richardson. Umbilical cord pH and base excess values in relation to adverse outcome events for infants delivering at term. *Am J Obstet Gynecol*, 191(6):2021–2028, Dec 2004.

P. Warrick, E. Hamilton, D. Precup, and R. Kearney. Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography. *IEEE Transactions on Biomedical Engineering*, 57(4):771–779, 2010.

P. A. Warrick and E. F. Hamilton. Fetal heart-rate variability response to uterine contractions during labour and delivery. In *Computing in Cardiology (CinC), 2012*, pages 417–420. IEEE, 2012.

J.-L. Wayenberg. Threshold of metabolic acidosis associated with neonatal encephalopathy in the term newborn. *J Matern Fetal Neonatal Med*, 18(6):381–385, Dec 2005.

M. Westerhuis, A. Kwee, A. A. van Ginkel, A. P. Drogtrop, W. J. A. Gyselaers, and G. H. A. Visser. Limitations of ST analysis in clinical practice: three cases of intrapartum metabolic acidosis. *BJOG*, 114(10):1194–1201, Oct 2007a.

M. Westerhuis, K. G. M. Moons, E. van Beek, S. M. Bijvoet, A. P. Drogtrop, H. P. van Geijn, J. M. M. van Lith, B. W. J. Mol, J. G. Nijhuis, S. G. Oei, M. M. Porath, R. J. P. Rijnders, N. W. E. Schuitemaker, I. van der Tweel, G. H. A. Visser, C. Willekes, and A. Kwee. A randomised clinical trial on cardiotocography plus fetal blood sampling versus cardiotocography plus ST-analysis of the fetal electrocardiogram (STAN) for intrapartum monitoring. *BMC Pregnancy Childbirth*, 7:13, 2007b.

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.e1–236.11, Sep 2007.

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.

Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, J. Dy, and P. Malvern. Modeling annotator expertise: Learning when everybody knows a bit of something. In *International Conference on Artificial Intelligence and Statistics*, volume 9, pages 932–939, 2010.

P. Yeh, K. Emary, and L. Impey. The relationship between umbilical cord arterial pH and serious adverse neonatal outcome: analysis of 51,519 consecutive validated samples. *BJOG*, 119(7):824–831, Jun 2012.

S. Y. Yeh, A. Forsythe, and E. H. Hon. Quantification of fetal heart beat-to-beat interval differences. *Obstet Gynecol*, 41(3):355–363, Mar 1973.

L. Zach. Automatická analýza kardiotokografického záznamu plodu. Master's thesis, Czech Technical University in Prague, 2013.

L. Zach, V. Chudáček, M. Huptych, J. Spilka, M. Burša, and L. Lhotská. CTG Annotator–Novel Tool for Better Insight into Expert-obstetrician Decision Making Processes. In *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*, pages 1280–1282. Springer, 2013.

Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu. Advancing Feature Selection Research – ASU Feature Selection Repository. Technical report, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, 2010.