Discriminating Normal from "Abnormal" Pregnancy Cases Using an Automated FHR Evaluation Method

Jiří Spilka[1], George Georgoulas[2], Petros Karvelis[2], Václav Chudáček[1], Chrysostomos D. Stylios[2], and Lenka Lhotská[1]

[1] Department of Cybernetics, Czech Technical University, Prague, Czech
[2] Laboratory of Knowledge and Intelligent Computing, Technological Educational Institute of Epirus, Department of Computer Engineering Arta, Greece

Abstract. Electronic fetal monitoring has become the gold standard for fetal assessment both during pregnancy as well as during delivery. Even though electronic fetal monitoring has been introduced to clinical practice more than forty years ago, there is still controversy in its usefulness especially due to the high inter- and intra-observer variability. Therefore the need for a more reliable and consistent interpretation has prompted the research community to investigate and propose various automated methodologies. In this work we propose the use of an automated method for the evaluation of fetal heart rate, the main monitored signal, which is based on a data set, whose labels/annotations are determined using a mixture model of clinical annotations. The successful results of the method suggest that it could be integrated into an assistive technology during delivery.

Keywords: Electronic fetal monitoring, Fetal Heart Rate, Random Forests, Classification.

1 Introduction

Fetal heart rate (FHR) monitoring has become an indispensable part of fetal assessment during pregnancy and, more importantly, during the delivery. It most commonly refers to the monitoring of fetal heart rate and uterine contractions (UC). These two signals comprise what is also known as the Cardiotocogram (CTG). CTG monitoring provides obstetricians with insight into fetal well-being acting as the main source of information for the fetus which is obviously not amenable to direct observation.

Since its introduction, the goal of fetal monitoring is to detect potential adverse outcomes and provide information about fetal well-being. However, the initial enthusiasm was followed by skepticism since the CTG was accused for the increased rate of cesarean sections [1] while high intra- and inter-observer variability was also reported [2],[3]. Despite the skepticism, CTG remains the most prevalent method for intrapartum fetal surveillance [4], often supported by ST-analysis, which nonetheless does not diminish the need for a correct interpretation of CTGs.

A. Likas, K. Blekas, and D. Kalles (Eds.): SETN 2014, LNAI 8445, pp. 521–531, 2014.

International Federation of Gynecology and Obstetrics (FIGO) guidelines [5] introduced in 1986 serve as the basis for CTG interpretation although several national updates have also been released – see e.g. [6] for references. The guidelines were meant to assure the lowest number of asphyxiated neonates as possible while avoiding false alarms (which leads to unnecessary cesarean sections). An additional goal of the guidelines is to lower the high inter and intra-observer variability.

In an attempt to reach a more objective interpretation of the CTG, computerized systems appeared, some of them being as old as the FIGO guideline themselves. Beginning with the work of [7] the automated analysis of CTG was based upon clinical guidelines [8]. Additionally, beyond the morphological features used in the guidelines, new features were introduced. These were primarily based on research in adult heart rate variability [9]. Therefore, time domain [10],[11], frequency domain [12], time-frequency [13], and nonlinear descriptors/features [14] were proposed over the past years and combined with various machine learning paradigms such as Support Vector Machines (SVMs) [15] and artificial neural networks (ANNs) [16],[17] to name just a few.

In this work we use for the first time a Random Forest (RF) classifier along with a sophisticated model for the definition of classes based on the latent class analysis (LCA). The results are promising indicating that this kind of modelling is probably more suitable for building a decision support system compared to systems that rely on information coming from the pH. Such a decision system would be closer to clinical reality than a system based solely on pH.

The rest of this paper is structured as follows: Section 2 provides the necessary background for FHR preprocessing and feature extraction as well as a short description of the RF classifier. In section 3 the data set along with the LCA are presented in brief, followed by a description of the experimental procedure and the respective results. Finally section 4 summarizes the findings offering also some hints for future work.

2 The Automatic FHR Analysis Method

2.1 FHR Preprocessing

The FHR could be contaminated by large amount of artefacts, especially when it is recorded using ultrasound probe. An example of FHR with artefacts can be seen in Fig. 1. Therefore preprocessing aims at removing these artefacts before proceeding to the feature extraction stage. Our preprocessing methodology employed a simple artefact rejection scheme: let $x(i)$ be a FHR signal in beats per minute (bpms), where N is the number of samples and $i = 1,2, \ldots, N$, whenever $x(i) > 50$ or $x(i) < 210$ we interpolated $x(i)$ using cubic Hermite spline interpolation implemented in MATLAB version 7.14.0.739.

2.2 Feature Extraction

The FIGO guidelines' morphological features were the first features used to describe the FHR and further used as inputs to classification schemes. Later, in order to examine FHR in more detail, other features originating from different domains were introduced. These were essentially based on adult heart rate variability analysis and mostly included frequency and nonlinear methods. Since all features used in this work are described in our previous works [14], [15] for the sake of brevity we present the features in Table 1 and provide necessary information to be able to repeat the analysis. We refer the interested reader to the referenced works in Table 1 for a more detailed description of the used features. In total we worked with 21 features such that different parameter settings yielded a total number of 49 features.

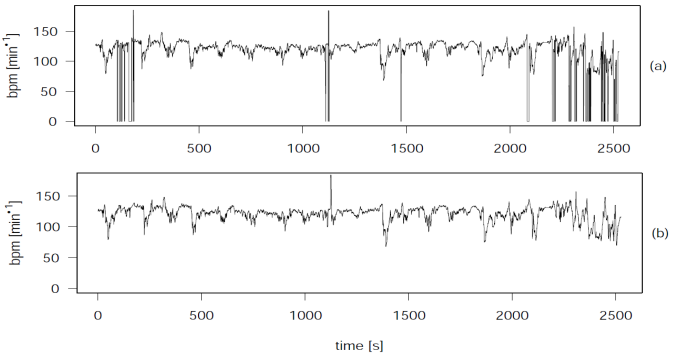

Fig. 1. Artefacts rejection. (a) Raw signal with artefacts, (b) signal after artefacts removal.

2.3 Random Forest Classification

Most classification tasks after the feature extraction stage include a feature selection stage [28] in order to alleviate the often encountered curse of dimensionality. Feature selection methods are usually divided into filter, wrapper, and embedded methods [29]. Decision Trees (DT) classifiers are of the last variety having the feature selection part inherently encoded during their construction process.

RFs are a learning paradigm that as it is implied by its name are comprised by a set of DTs that act together in order to reach a classification decision. RFs were introduced by Breiman [31] and since then they have been employed in many classification as well as regression tasks [32]-[34]. RFs are very competitive compared to other state of the art classification algorithms, such as boosting. Unlike boosting RFs provide fast training.

Each member of the ensemble of trees operates on a bootstrapped sample of the training data. Moreover at each node of a tree random feature selection is performed. More specifically, a subset S with M features from the original set of n features is selected and then the best feature among M is selected to split the node. With this mechanism there is no need for explicitly excluding a set of features before the classification process.

Table 1. Features involved in this study

Feature set	Features	parameters
FIGO-based [5]	baseline	mean, standard deviation
	number of accel. and decel, Δ_{total}	
Statistical	STV, STV-HAA[18], STV-YEH[19], Sonicaid[20], SDNN [9],Δ_{total}[10], LTI-HAA[18]	
Frequency	Energy03[9]	LF, MF, HF, LF/HF
	Energy04[21]	VLF, LF, MF, HF, LF/(MF+HF)
Fractal dim.	FD_Variance, FD_BoxCount, FD_Higuchi[22], DFA[23], FD_Sevcik[24]	
Entropy	ApEn[25], SampEn[26]	M= 2, r = 0.15,0.2
Complexity	Lempel Ziv Complexity (LZC) [27]	
Other	Poincaré	SD1, SD2

3 Experimental Analysis

For the experimental evaluation of the proposed approach, we employed a newly released CTG database [35] and a multiple trial resampling method. The database, the evaluation procedure and the results are presented in the rest of this section.

3.1 Database

The CTU-UHB database [35] consists of 552 records and it is a subset of 9164 intrapartum CTG records that were acquired between years 2009 and 2012 at the obstetrics ward of the University Hospital in Brno, Czech Republic. The CTG signals were carefully selected with clinical as well as technical considerations in mind. The records selected were as close as possible to the birth and in the last 90 minutes of labor there is at least 40 minutes of usable signal. Additionally, since the CTG signal is difficult to evaluate in the second (active) stage of labor, we have included only those records which had second (active) stage's length at most 30 minutes. The CTGs were recorded using STAN and Avalon devices. The acquisition was done either by scalp electrode (FECG 102 records), ultrasound probe (412 records), or combination of both (35 records). For three records the information was not available. All recordings were sampled at 4Hz. The majority of babies were delivered vaginally (506) and the rest using cesarean section (46). A more detailed description of the CTU-UHB is provided in [35].

3.2 Latent Class Analysis

In this work we used clinical annotations from 9 clinicians. All clinicians are currently working in delivery practice with experience ranging from 10 to 33 years (with a median value of 15 years). Clinicians evaluated the CTG recordings into three classes: normal, suspicious, and pathological (FIGO classes). Since there is a large inter-observer variability in evaluation the simple majority voting among clinicians cannot be used. Therefore we employed a more powerful approach - the latent class analysis (LCA) [36]. The LCA is used to estimate the true (unknown) evaluation of CTG and to infer weights of individual clinicians' evaluation. The LCA and its advantages over majority voting were described in [37]. For other examples on LCA in machine learning see, e.g. [38] and [39]. The clinical evaluations were considered as coming from mixture of multinomial distribution with unknown parameters and unknown mixing proportions. The Expectation Maximization (EM) algorithm [40] was used to estimate the unknown parameter and proportions. The EM algorithm was restarted several times with different starting values to avoid local maximum. The limit of log-likelihood convergence was set to 10e−3. The resulting class for individual examples was determined by the largest posterior probability.

The application of LCA leads to different labeling compared to the majority voting annotation as it is summarized in the following cross (Table 2), which corresponds to the data set described in Section 3.1. For the calculation of this table four cases from the original 552 CTGs were removed because the majority voting was inconclusive. This is a situation which is unlikely to occur with the LCA.

Table 2. Cross table (contingency table) of the annotations resulting from applying the majority voting (MV) and the (LCA)

	Normal by LCA	Suspicious by LCA	Pathological by LCA
Normal by MV	168	50	0
Suspicious by MV	7	185	66
Pathological by MV	0	3	69

In this preliminary study,we merged the suspicious and pathological class (according to the LCA) for simplicity reasons into a "super class" of abnormal cases. Thus we are interested in those records that deviate from normality.

3.3 Evaluation Procedure

For evaluating our approach we employed 5 trials of 2 fold cross validation (5x2 CV) [41], [42]. In other words we divided the available data into two sets and we used one for training the random forest classifier and the other one (testing) for estimating its performance and then we reversed their roles (the training became the testing set). The whole procedure was repeated 5 times with reshuffling taking place between each one of the five different trials.

Since RFs are not a parameter free algorithm, some parameter tuning needs to take place which however should be decoupled from the performance estimation process [42],[43]. Therefore during each training phase we performed a grid search increasing the number of trees from 100 to 1000 with a step increment of 50 and the number of features from 1 to 10 with a step increment of 1 (Breiman suggested a value for the feature equal to $\lfloor \log_2 n+1 \rfloor$ where n is the number of features, while also other suggested values can be found in the literatures ($\sqrt{n}$ or even as low as 1 [44]), so we tried a search in the vicinity of these suggested values). To perform this grid search each time the training set was divided again into a training and testing set (2/3 of the original training set comprise the new training) a performance metric was evaluated (see paragraph bellow) and the procedure was repeated five times (not to be confused with the 5 repetitions of the 5x2 CV procedure) after reshuffling the cases. The averaged performance metric over these five trials was used for selecting the "best" set of parameters. Using this set of parameters a new random forest was trained using the original training set and its performance was tested using the test set.

The "best" set of parameters was selected using two different criteria, which were derived from the elements of the confusion matrix (Table 3). Following the standard practice in the medical field we labeled the abnormal cases as positive and the normal cases as negative:

Table 3. Confusion matrix for a typical dichotonomous (2-class) problem

		predicted class	
		Positive	Negative
True class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

a) Balanced Error Rate (BER) [45]:

$$\mathrm{BER} = \left(\mathrm{FP}/\left(\mathrm{FN}+\mathrm{TP}\right)+\left(\mathrm{FN}/\left(\mathrm{FP}+\mathrm{TN}\right)\right)\right)/2$$

b) Geometric mean (G-Mean) [46]:

$$\mathrm{G-Mean} = \sqrt{\mathrm{TPrate}\cdot\mathrm{TNrate}}$$

Where:

True Positive Rate (TPrate) also known as Sensitivity or Recall:

$$\mathrm{TPrate} = \mathrm{TP}/\left(\mathrm{TP}+\mathrm{FN}\right)$$

True Negative Rate (TNrate) also known as Specificity:

$$\mathrm{TNrate} = \mathrm{TN}/\left(\mathrm{TN}+\mathrm{FP}\right)$$

The aforementioned criteria were selected instead of the more common accuracy measure due to the slight imbalanced of the data set, since these criteria are not affected by the distribution of cases into the different classes.

3.4 Results

Tables 4 and 5 summarize the results for the two different performance metrics that were used during the random forest tuning process and Fig. 2 includes the respective specificity and sensitivity values in a box plot format, which reveals that under this setting the two approaches are very similar.

Table 4. Aggregated Confusion matrix using BER for tuning

		predicted class	
		abnormal	normal
True class	abnormal	1365	520
	normal	189	686

Table 5. Aggregated Confusion matrix using g-mean for tuning

		predicted class	
		abnormal	normal
True class	abnormal	1359	526
	normal	187	688

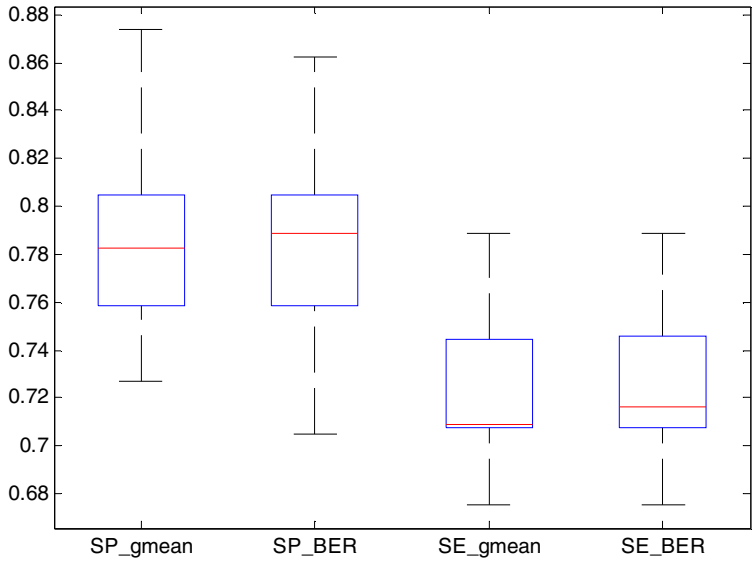

Fig. 2. Boxplot of the specificities and the sensitivities for the 2 different criteria used during tuning

4 Conclusions

This research work integrates a method for combining experts' evaluation of CTG recording with an automatic approach that attempts to reproduce their decision. The automated method uses a number of diverse features, coming from different domains, along with an advanced ensemble method, the RF paradigm. Our preliminary results indicate that the "latent" labeling approach creates different annotations compared to simple majority voting and that the resulting classification problem can be tackled by an automated method, even though the performance should be further improved before it can be adopted into clinical practice.

Moreover the sensitivity (~72%) and specificity (~78%) values achieved are higher (even though there is no one-to-one correspondence) than those achieved using the pH value for labeling [47] indicating that the proposed data model (features-LCA labeling) may provide more consistent approach than the one relying on the pH level.

In future work we will continue testing of the proposed three class setting approach (normal, suspicious, and pathological) and especially with in an ordinal classification setting. This way we will exploit the natural ranking of categories which in most cases leads to higher classification performance [48] compared to a scheme that does not take into account the natural ordering of the classes.

Acknowledgments. This research work was supported by the joint research project "Intelligent System for Automatic CardioTocoGraphic Data Analysis and Evaluation using State of the Art Computational Intelligence Techniques" by the programme "Greece-Czech Joint Research and Technology projects 2011-2013" of the General Secretariat for Research & Technology, Greek Ministry of Education and Religious Affairs, co-financed by Greece, National Strategic Reference Framework (NSRF) and the European Union, European Regional Development Fund.

References

1. Alfirevic, Z., Devane, D., Gyte, G.M.: Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database Syst. Rev. 3 (2006)
2. Bernardes, J., Costa-Pereira, A., Ayres-de-Campos, D., van Geijn, H.P., Pereira-Leite, L.: Evaluation of interobserver agreement of cardiotocograms. Int. J. Gynaecol. Obstet. 57(1), 33–37 (1997)
3. Blix, E., Sviggum, O., Koss, K.S., Oian, P.: 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 (2003)
4. Chen, H.Y., Chauhan, S.P., Ananth, C.V., Vintzileos, A.M., Abuhamad, A.Z.: 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.e10 (2011)
5. FIGO, Guidelines for the Use of Fetal Monitoring. Int. J. Gynaecol. Obstet. 25, 159–167 (1986)

6. 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 (2009)
7. Dawes, G.S., Visser, G.H.A., Goodman, J.D.S., Redman, C.W.G.: Numerical analysis of the human fetal heart rate: the quality of ultrasound records. Am. J. Obstet. Gynecol. 141(1), 43–52 (1981)
8. de Campos, D.A., Ugwumadu, A., Banfield, P., Lynch, P., Amin, P., Horwell, D., Costa, A., Santos, C., Bernardes, J., Rosen, K.: A randomised clinical trial of intrapartum fetal monitoring with computer analysis and alerts versus previously available monitoring. BMC Pregnancy Childbirth 10(71) (2010)
9. 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(1996)
10. Magenes, G., Signorini, M.G., Arduini, D.: Classification of cardiotocographic records by neural networks. In: Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000, vol. 3, pp. 637–641 (2000)
11. Goncalves, H., Rocha, A.P., de Campos, D.A., Bernardes, J.: Linear and nonlinearfetal heart rate analysis of normal and acidemic fetuses in the minutes precedingdelivery. Med. Biol. Eng. Comput. 44(10), 847–855 (2006)
12. Van Laar, J.O.E.H., Porath, M.M., Peters, C.H.L., Oei, S.G.: Spectral analysis of fetal heartrate variability for fetal surveillance: Review of the literature. Acta Obstetricia et Gynecologica Scandinavica 87(3), 300–306 (2008)
13. Georgoulas, G., Stylios, C.D., Groumpos, P.P.: Feature Extraction and Classiffication of Fetal Heart Rate Using Wavelet Analysis and Support Vector Machines. International Journal on Artifficial Intelligence Tools 15, 411–432 (2005)
14. Spilka, J., Chudáček, V., Koucký, M., Lhotská, L., Huptych, M., Janků, P., Georgoulas, G., Stylios, C.: Using nonlinear features for fetal heart rate classification. Biomedical Signal Processing and Control 7(4), 350–357 (2012)
15. Georgoulas, G., Stylios, C.D., Groumpos, P.P.: 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 (2006)
16. Czabanski, R., Jezewski, M., Wrobel, J., Jezewski, J., Horoba, K.: 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 (2010)
17. Georgieva, A., Payne, S.J., Moulden, M., Redman, C.W.G.: Artifficial neural networks applied to fetal monitoring in labour. Neural Computing and Applications 22(1), 85–93 (2013)
18. De Haan, J., Van Bemmel, J.H., Versteeg, B., Veth, A.F.L., Stolte, L.A.M., Janssens, J., Eskes, T.K.A.B.: Quantitative evaluation of fetal heart rate patterns. I. Processing methods. European Journal of Obstetrics and Gynecology and Reproductive Biology 1(3), 95–102 (1971)
19. Yeh, S.Y., Forsythe, A., Hon, E.H.: Quantification of fetal heart beat-to-beat interval differences. Obstet. Gynecol. 41(3), 355–363 (1973)
20. Pardey, J., Moulden, J., Redman, C.W.G.: A computer system for the numerical analysis of nonstress tests. Am. J. Obstet. Gynecol. 186(5), 1095–1103 (2002)
21. Signorini, M.G., Magenes, G., Cerutti, S., Arduini, D.: Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. IEEE Trans. Biomed. Eng. 50(3), 365–374 (2003)

22. Higuchi, T.: Approach to an irregular time series on the basis of the fractal theory. Phys. D 31(2), 277–283 (1988)
23. Peng, C.K., Havlin, S., Stanley, H.E., Goldberger, A.L.: Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5(1), 82–87 (1995)
24. Sevcik, C.: A Procedure to Estimate the Fractal Dimension of Waveforms. Complexity International 5 (1998)
25. Pincus, S.: Approximate entropy (ApEn) as a complexity measure. Chaos 5(1), 110–117 (1995)
26. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate entropy and sample entropy. Am. J. Physiol. Heart Circ. Physiol. 278(6), H2039–H2049 (2000)
27. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Transactions on Information Theory IT-22(1), 75–81 (1976)
28. Theodoridis, S., Koutroumbas, K.: Pattern recognition, 4th edn. Academic Press (2009)
29. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature extraction: foundations and applications. STUDFUZZ, vol. 207. Springer (2006)
30. Liu, H., Motoda, H.: Computational methods of feature selection. Chapman & Hall/CRC (2007)
31. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
32. Athanasiou, L., Karvelis, P., Tsakanikas, V., Naka, K., Michalis, L., Bourantas, C., Fotiadis, D.: A novel semi-automated atherosclerotic plaque characterization method using grayscale intravascular ultrasound images: Comparison with Virtual Histology. IEEE Transactions on Information Technology in Biomedicine 16(3), 391–400 (2012)
33. Liaw, A., Wiener, M.: Classification and Regression by random Forest. R News 2(3), 18–22 (2002)
34. Díaz-Uriarte, R., De Andres, S.A.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(1), 3 (2006)
35. Chudáček, V., Spilka, J., Burša, M., Janků, P., Hruban, L., Huptych, M., Lhotská, L.: Open access intrapartum CTG database. BMC Pregnancy and Childbirth 14 (2014)
36. Lazarsfeld, P.F.: The Logical and Mathematical Foundations of Latent Structure Analysis. In: Samuel, A., Stouffer (eds.) Measurement and Prediction, pp. 362–412. John Wiley & Sons, New York (1950)
37. Spilka, J., Chudáček, V., Janků, P., Hruban, L., Burša, M., Huptych, M., Zach, L., Lhotská, L.: Analysis of obstetricians' decision making on CTG recordings. Journal of Biomedical Informatics (2014) (manuscript submitted for publication)
38. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statistics 28, 20–28 (1979)
39. Raykar, V.C., Yu, A.: Eliminating Spammers and Ranking Annotators for Crowd sourced Labeling Tasks. Journal of Machine Learning Research 13, 491–518 (2012)
40. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1–38 (1977)
41. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10(7), 1895–1923 (1998)
42. Japkowicz, N., Shah, M.: Evaluating learning algorithms: A classification perspective. Cambridge University Press (2011)
43. Salzberg, S.L.: On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and knowledge discovery 1(3), 317–328 (1997)

44. Hastie, T.J., Tibshirani, R.J., Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction. Springer (2009)
45. Xuewen, C., Wasikowski, M.: Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 124–132. ACM (2008)
46. Kubat, M., Stan, M.: Addressing the curse of imbalanced training sets: one-sided selection. In: ICML, vol. 97, pp. 179–186 (1997)
47. Spilka, J., Georgoulas, G., Karvelis, P., Oikonomou, V.P., Chudáček, V., Stylios, C.D., Lhotská, L., Janků, P.: Automatic evaluation of FHR recordings from CTU-UHB CTG database. In: Bursa, M., Khuri, S., Renda, M.E. (eds.) ITBAM 2013. LNCS, vol. 8060, pp. 47–61. Springer, Heidelberg (2013)
48. Frank, E., Hall, M.: A simple approach to ordinal classification. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 145–156. Springer, Heidelberg (2001)