Journal of computer science and software applications
https://www.mfacademia.org/
ISSN:2377-0430
Vol. 4, No. 4, 2024

Cardiotocography Graph Classification Using Deep Learning and Ensemble Methods

Emily Johnson

University of California, Berkeley

johnson980@berkeley.edu

Abstract:

Fetal monitoring, essential for obstetric assessment and prenatal health, enables timely detection of abnormalities, reducing birth defects and mortality. Cardiotocography (CTG) is an effective method that records fetal heart rate changes and their relationship to movements and uterine contractions, allowing for the assessment of fetal condition. This paper presents a novel CTG-based classification method using deep learning and ensemble classification, utilizing eight selected characteristics from a cloud-based CTG analysis system.Key contributions include the collection and feature extraction of fetal heart rate data, deep learning model training, creation of a dataset with four reshuffled fetal data features, and ensemble classification. An empirical study using five-fold minus one cross-validation on a dataset of 21,570 records shows the proposed method achieves an average accuracy of 96.83%, compared to 94.88% from traditional methods.

Keywords:

cardiotocography;deep learning; feature extraction ; ensemble classifier; cardiotocography classification.

1. Introduction:

As an important obstetric assessment measure of prenatal fetal health, fetal monitoring can detect fetal abnormalities and reduce birth defect rates and mortality in a timely manner. Cardiotocography (CTG) is a true record of the relation between the changes of fetal heart rate and fetal movement(contraction in particular) ,analysis of CTG has become an effective and common practice for fetal monitoring to help understand the status of the fetus in the womb in an indirect manner. Manual CTG analysis is based on empirical knowledge.,its success relies heavily on the obstetricians' experience and the conclusion could be obscure in some cases.This is mainly due to the fact that their experiences rely on the prior cases they have previously encountered, which could be obscure and inaccurately presented. This may easily lead to divergence in their conclusions. This is a common case in the areas such as biomedical analysis and pattern analysis. [1] Automatic analysis of fetal data (in particular heart rate data) has been boosted by not only the ever-developing monitoring instruments but also the computing science and technologies. A number of fetal heart systolic analysis system have been successfully developed and deployed in obstetrics. However, only SonicAid developed in Oxford University has been widely accepted whilst the others failed because of their lack of specificity and poor in mainland China, companies such as Sunray and Mindray have developed their fetal analysis systems which target domestic markets and better satisfy the local requirements.

Table 1 summarizes the related work in CTG-based fetal health analysis. It is worth noting that most of them are based on the dataset collected by SisPorto, which is a part of the public UCI dataset. It contains only 2128 data items of 21 features. Various machine learning approaches such as artificial neuron networks and decision trees have been applied and achieved accuracy of 90% or higher. However, most of these prior works can only identify abnormal pathological cases from normal one, whilst modern obstetrical practice expects finer classification which imposes a more demanding task for such existing approaches.

Table 1. Research situation at home and abroad

researcher	method	data set	accuracy
N Chamidah [4] ,etc	1.SVM 2.K-Means+SVM	UCI	1.76.72% 2.90.64%
Z Comert[5] , etc	1.ANN 2.ELM	Not clear(UCI/CTU-UHB)	1.91.84% 2.93.42%
D Gavrilis[6] , etc	Grammatical Evolution	UCI	92.50%
N Krupa[7] , etc	EMD,SVM	Random selection of 90 data records of 20 minutes from fifteen volunteers, divided into normal and risk categories. The training set and test set ratio is 2: 1.	86%
S Das[8] , etc	ANN	UCI	92.14%
B Johnson[9] , etc	Four layers of neural networks, 200 neurons per layer, a scaled conjugate gradient back propagation method, and a threshold of 1.915.	UCI	The absolute accuracy is 84.26%, the positive error is 10.18%, and the negative error is 5.56%.
H Ocak[10] , etc	ANFIS	Including 1,831 CTG records ， 1655 were classified as normal, the remaining 176 were classified as pathological.	Normal 97.2%, Pathologic 96.6％
H Sahin[11] , etc	ANN SVM SL RBF C4.5 CART RF k-NN	Including 1831 instances with 21 attributes, the results are only divided into normal and pathological two categories	98.63% 98.96% 98.74% 98.53% 99.13% 98.91% 99.18% 98.42%
M Jezewski[12] , etc	Clustering and fuzzy if-then rules	Data from the system MONAKO file includes 685 signals, the results are divided into normal and abnormal two categories	>79%
A Pereira[13] , etc	Rapid excavation classification model	2126 data, 23 attributes, the results are divided into three types of NSP	>80%
P Tomas[14] , etc	Random forest, using seven features of AC, DS, DP, ASTV, MSTV, ALTV and Mean	UCI, the results are divided into three types of NSP	93.70%
Zhou Hongbiao[15] , etc	Genetic Algorithm and BP Neural Network	UCI	Normal 98.24%, Suspicious 82.67%, Pathologic 95.65%

2. Our Contribution

This paper presents a novel CTG-based classification method based on deep learning and ensemble classification above 8 fetal heart monitoring characteristics which were selected from the CTG chart of the fetal heart rhythm information analysis system based on the cloud service. Its main contribution includes:

(1) Identification and Acquisition of CTG-based features.There are 21 features considered in the public UCI dataset. However, the dataset itself is too small and the criteria and acquisition of these features are not unveiled. Given a much larger dataset, more features could be obtained to meet the modern clinical practice’s requirements. In this paper, we use a much larger CTG dataset of 21570 which have been collected from a cloud-based fetal analysis system used by multiple domestic hospitals.

(2) A deep-learning based approach to reshuffle the dataset. Most traditional classification approaches have difficulties in analyzing fetal data in a real-time manner. Furthermore, latest researches in various domains suggest that they are outperformed by deep learning approaches which are capable of unveiling inherent features for better classification.

(3) Ensemble classification of the reshuffled CTG dataset. Traditional classifiers such as GradientBoosting, ExtraTrees, RandomForest, MLP (Multi-Layer Perceptron) and DecisionTree[17] have been easily ensemble in our work and the results are far less vulnerable than those given by individual classifiers.

3. Data and methods

3.1 Data

The raw fetal dataset used in this study is the fetal heart rate monitoring chart collected from the cloud-based fetal analysis system used in the obstetrics and gynecology in hospital such as The First Affiliated Hospital of Jinan University. Table 2 summarizes the dataset contains 215 preliminary features of 21570 data items, these 215 features were then reshuffled into four by an autoencoder [18].

Table 2. Number of data collection in each health category

Category	numbers
Normal	20818
suspicious	725
abnormal	28

3.2 Methods

Inspired by prior works in fetal analysis [19], we present in this paper a novel fetal analysis approach based on deep learning and ensemble classification. This approach contains five major steps: data acquisition, feature extraction and selection, deep learning model construction, model training, and data recognition and classification.

3.2.1 Collection of fetal heart rate data

The raw CTG data such as fetal heart rates and uterine pressure were collected via ultrasonic and pressure sensors from daily obstetric practice in hospitals. Once a pregnant woman completes her check, her CTG is then analyzed by an obstetrician who is to classify it into either normal, suspicious, or abnormal. These classification result is denoted as 0, 1 and 2 respectively and saved in a record alongside the heart rate data. We have collected totally 21570 records for our analysis.

3.2.2 Feature extraction and selection

Various features could be extract from the raw CTGs, among them we select eight features based on both domain expertise and results of prior works. They are listed below in:

Table 3. Eight features and their meanings

name of the feature	meaning of the feature
FHRBaseline	Baseline value
variability	Baseline variance value
decTimes	Deceleration times
accTimes	Acceleration times
criteriaNSTFHR	Baseline baseline values for fetal heart rate
criteriaNSTVar	Baseline variation standard value
criteriaNSTDec	Deceleration standard value
criteriaNSTOAcc	Acceleration standard value

Each record in the data set contains eight features and one diagnostic results. Analysis of the fetal heart rate curve can be obtained by the above characteristics. Each record within our dataset contains an assessment result and eight features which are obtained via analysis of the raw CTG data. The first four features were obtained by first calculating the baseline values of the fetal heart rate and then identifying the acceleration, deceleration and mutation signals based on the baseline values. [20] The assessment was made by a computer-aided system which is based on a set of criteria given by SOGC (THE SOCIETY OF OBSTETRICANS AND GYNAECISIS OF CANADA), as shown in Table 4[21].

Table 4. Classification according to Basic characteristics of fetal heart rate of electronic fetal monitoring in Canadian standards SOGC guideline, 2007)

parameter	normal NST	atypical NST	abnormal NST
FHRBaseline	•110-160bpm	•100-110bpm •>160bpm •Baseline rises	•Heart rate too slow <100bpm •Heart rate too fast >160bpm，>30min •Baseline uncertain
variability	•6-25bpm(moderate variation)	•≤5bpm(No variation or minimal variation），40-80min	•≤5bpm，≥80min •≥25bpm，>10min •Sine type
decTimes	•No deceleration or occasional variation	•Variable deceleration last 30-60s	•Variable deceleration last>60s •Late deceleration
accTimes	•≥2 times，≥15bpm，last >15s，<40min	•≤2 times，≥15bpm，last >15s，40-80min	•≤2 times，≥15bpm，last >15s，>80min
accTimes（<32weeks）	•≥2 times，≥10bpm，last >10s，<40min	•≤2 times，≥10bpm，last >10s，40-80min	•≤2 times，≥10bpm，last >10s，>80min

The other four criteria (in criteriaNSTFHR, criteriaNSTVar, criteriaNSTDec, and criteriaNSTOAcc) are obtained from the SOGC guideline in Table 4, with up to three options for each cell in the table. The parameters have three criteria, so they can be represented by numbers 0-8. Each number corresponds to one of the options in the cell, 0-2 indicates the normal NST, 3-5 represents the atypical NST, and 6-8 represents the abnormal NST. The system is responsible for providing the obstetricians with both these above parameters and the classification results based on Table 4, so that they can make their judgment based on their prior experience.

3.2.3 Construction of the Deep Learning Model

Among the eight eigenvalues, features criteriaNSTFHR, criteriaNSTVar, criteriaNSTDec and criteriaNSTOAcc are found correlated to the other four features in FHRBaseline, variability, decTimes and accTimes, this suggests redundancy among these eight features[22]. An autoencoder [23] is therefore employed to reshuffle these features, which results in a modified dataset with four features.

Autoencoder is a network model capable of transforming high-dimensional data into low-dimensional one in an unsupervised learning manner. It has been applied to unsupervised feature learning in deep learning, particularly in areas such as image classification [24]. In the autoencoder, the hidden layer is regarded as an encoder and decoder, the input data through the hidden layer of coding and decoding, to reach the output layer to ensure that the results of the output is as much as possible consistent with the input data. In other words, the hidden layer is to ensure that the output data is equal to the input data. One of the benefits of this is that the hidden layer can capture the characteristics of the input data, leaving its features unchanged, while greatly reducing the dimension of the data [25].

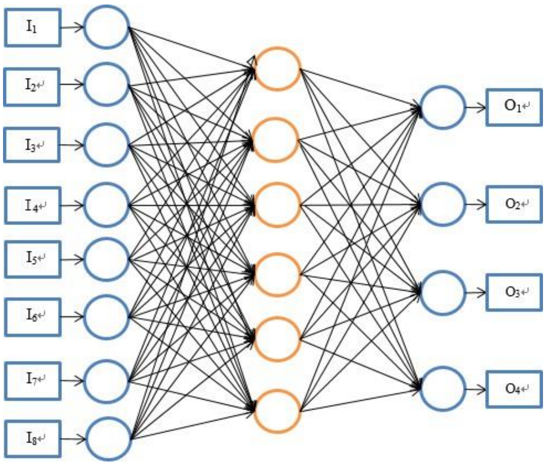

Fig.1 The specific structure of the network used in this experiment

The autoencoder consists of three layers in our case, where the input fetal heart rate data of each layer is the output data of the previous layer. Assuming that the system of three layers is S (S1, S2, S3), the input and output are I and O respectively, the system can be expressed as I => S1, S1 => S2, S2 => S3, S3 => O. If there are no changes from I to O, the whole system does not have any information loss of damage, which means that each layer of information has not changed but replaced by another expression. This is the main idea of automatic encoders: automatic learning, step by step progressive. If we use our own data as input, and constantly adjust the training so that the output and input remain the same, you can get the data of multiple levels of features.

The specific network structure shown in Fig.1, the first and input layer contains eight nodes, the second and hidden layer is of six nodes, and the third and output layer is composed of four nodes. In brief, feature extraction of the raw fetal heart rate dataset is carried out in a dimensional reduction manner, which results in a new dataset of four reshuffled features.

3.2.4 Training

Training of this autoencoder with the fetal dataset is composed of two steps, namely pre-training and structural reshuffling. They are carried out in a left-to-right unsupervised learning manner and right-to-left supervised learning manner respectively.

(1) Pre-training: First, the autoencoder is trained with the raw fetal dataset without diagnosis results in order to capture the features for each of the three layers. The parameters of all the nodes at all layers are determined during this process.

(2) Structural reshuffling: After pre-training, we use the network parameters for fine-tuning, reverse

generating the bottom of the data through the parameters learned from the previous step, while correcting the network layer forward parameters. The previous step is similar to the random initialization process, so that the input value is more approximated to the optimal value, the overall learning effect is superior to the ordinary neural network and that's why the deep learning is better than ordinary neural network.

Once the training is over, this autoencoder becomes capable of reshuffling the input raw fetal dataset for output.

3.2.5 Classification of the Reshuffled Data

Various classifiers have been considered for classification of the reshuffled fetal dataset. They have their pros and cons in terms of accuracy, robustness and others. Ensemble classifier such as Boosting is chosen to ensemble various classifiers of lower accuracy in order to achieve a highly accurate one.

Gradient Boosting Machine proposed by Friedman is a representative algorithm for boosting integration categorization, which takes the tree of all the previously generated trees and uses the gradient descent method to generate the tree in the direction of minimizing the objective function. This makes it need for a large number of iterations when the amount of data is large and complex, leading to a very large time requires of classification. This classifier used in this article is eXtreme Gradient Boosting (XGBoost) [26], a popular C++ implementation of the Gradient Boosting Machine with improved accuracy and high efficiency due to its use of multi-threading. XGBoost is packaged as a Python library for easy access.

Various classifiers can be ensembled in XGBoost. Considering the nature of the reshuffled fetal dataset and the applicability of various classifiers, we chose five of them, which are Gradient Boosting, Extra Trees, Random Forest, MLP and Decision Tree, as shown in Figure 2.

Compared with the traditional GBDT (Gradient Boosted Decision Trees), the larger difference of XGBoost is the definition of the objective function. XGBoost uses Taylor series to make an approximate equivalent. The final objective function depends only on the first derivative and the second derivative on the error function of each data point. At the same time, XGBoost adds a regular term to the cost function. The regular term contains the number of leaf nodes of the tree, and the sum of the squares of the L2 modules of the scores output on each leaf node. From the perspective of Bias-variance tradeoff, the regular term reduces the variance of the model and controls the complexity of the model. This makes the learning model more simple, to prevent over-fitting, which is a characteristic that XGBoost better than a traditional GBDT. [27]

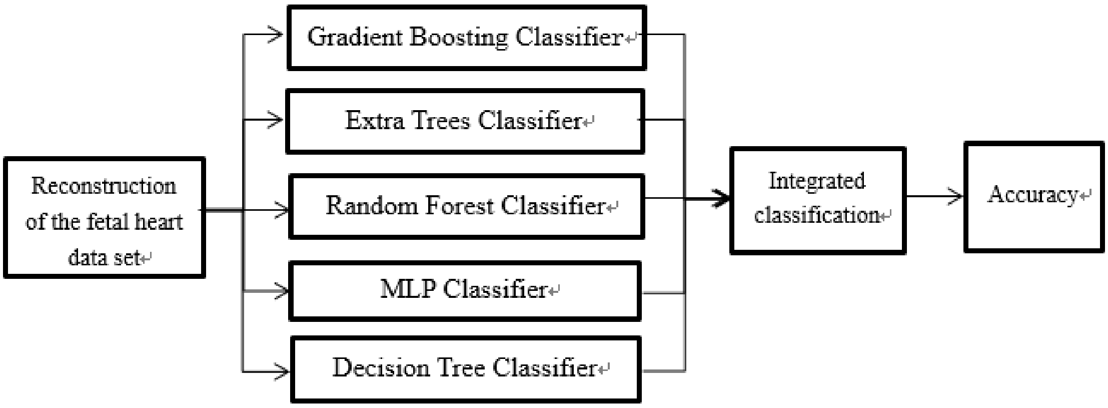

Fig.2 Framework of the ensemble classifier in scratch

4. Experimental Results and Analysis

4.1 Setup

Experiment are carried out in a five-fold minus one cross validation manner[28], which means that the dataset was evenly divided into five sections and, in each experiment, four sections were used for training purpose and the last one was used for testing. This was repeated for five times.divides all the training data into five parts and completes five experiments. The performances of the five chosen classifiers with the raw and reshuffled fetal datasets were investigated and compared against each other, before they are compared against that of the XGBoost ensemble classifier. The parameter settings of these five classifiers are given as below (default values were used if not specified).

(1) Gradient Boosting: the number of trees is 60.

(2) Extra Trees: the number of trees is 80, the maximum depth of the tree is 12, the minimum number of leaves is 2, and max_features is set to 0.5.

(3) Random Forest: the number of trees is 80, the maximum depth is 12, the minimum number of leaves is 2, max_features is set to 0.2.

(4) MLP: the use of single-layer hidden layer structure, the number of hidden layer nodes is 5.

(5) Decision Tree: the use of ID3 decision tree, the evaluation criteria using the Gini coefficient.

4.2 Experimental results

The numbers of not match distribution between the results of experimental classification and the results of the diagnosis of doctors is as shown in Table 5 below, the health class classification accuracy rate is 96.88%, the suspicious class accuracy is 96.55% and the abnormal class accuracy rate is 71.43% by calculating.

Table 5. Error distribution from the ensemble classifier

XGBoost classification Manual diagnosis	normal	suspicious	abnormal
normal		324	326
suspicious	9		16
abnormal	4	4	

4.3 Case Study

Table 6 presents some of the cases that XGBoost has mis-classified, where each of the six cases illustrates a type of mis-classification (normal mis-classified as suspicious, etc).

Table 6. Examples of incorrect judgment

NO	FHRBaseline	variability	decTimes	accTimes	criteriaNSTFHR	criteriaNSTVar	criteriaNSTDec	criteriaNSTO Acc	Artificial result	Experimental result
18054	126	7	1	14	0	7	3	0	0	0
6709	113	9	2	12	0	7	3	0	0	2
18331	209	11	6	1	3	7	6	3	1	0
6292	122	10	0	2	0	7	0	0	1	2
6312	119	6	0	2	0	7	0	0	2	0
3067	167	12	4	14	3	7	6	0	2	1

A closer look at the fetal data of NO.18054 and NO.6709 shows that their criterialNSTDec are 3, which is considered abnormal by obstetricians. The suspicious case of NO.18331 was mis-classifed as normal because the variability is still considered as acceptable by obstetricians. In the cases of NO.6292 and No.6312, the manual diagnoses were made by the obstetricians after further checks of the pregnant women, which showed that all the other features were rated normal. In the case of NO.3067, the FHRBaseline of 167 was considered suspicious, whilst both the variability of 12 and the criteriaNSTDec of 6 were rated normal. Additional expertise of the obstetricians was needed for a manual judgment, which is beyond the capability of the ensemble classifier.

Further comparisons are made between various classifiers, and the results are shown in Figure 3. The classification accuracy of Gradient Boosting, Extra Trees, Random Forest, MLP and Decision Tree applied to the raw dataset respectively was 95.30%, 94.11%, 95.96%, 96.14% and 93.47% respectively, that of XGBoost applied to the raw dataset directly is 94.48%; and that of the combination of autoencoder and XGBoost is 96.83%.

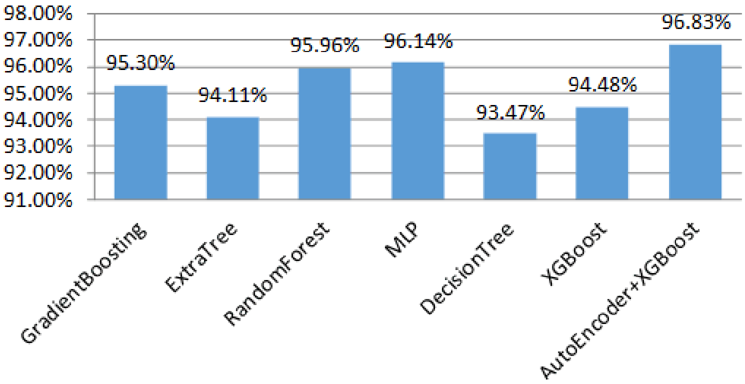

Fig.3 Classification accuracy of initial data sets by different classifiers

4.4 Discussion

The above results suggest that:

(1) The proposed approach provides a far higher accuracy for “normal” cases than that for “suspicious” and “abnormal” cases. This is likely to be caused by these two reasons. First, the data distribution in the sample dataset is neven, the proportion of normal cases is very high, compared to that of the other cases, which results in a biased classifier after training. In addition, domain expertise was heavily involved in manual classification of the fetal dataset during which the final decisions were made after re-investigation for all the cases be initially labeled “abnormal”. However, these outliers have been neglected by the classifier(s) for the sake of overall performance.

(2) It is worth noting that the propose autoencoder plus ensemble classifier approach achieves better results than classifiers such as decision tree, ExtraTree and simple XGBoost, and slightly better than the other classifiers in Random Forrest and Neuron Network. It is believed that this is relevant to both the uneven data distribution and the simple data structure of the raw CTG dataset.

5. Conclusion

A 21570-strong CTG dataset collected from the obstetrics and gynecology in multiple hospitals such as the First Affiliated Hospital of Jinan University has been used to evaluate a novel CTG analysis approach which is based on deep learning and ensemble classifier. Results from XGBoost suggest that the it achieves an accuracy of 96.83% where the dataset is reshuffled by the autoencoder, compared to that of 95.30%, 94.11%, 95.96%, 96.14%, 93.74%, and 94.48% provided by Gradient Boosting, Extra Trees, Random Forest, MLP, Decision Tree, and XGBoost (on raw CTG dataset). This suggests that the proposed hybrid approach outperforms the others in CTG analysis.

References

[1] Langellė, R., Devoux, T.: Training Multilayer Perceptrons Layer Using an ObjectiveFunctions for Internal Representations. J. Neural Networks 58(8), 83–98 (1996)

[2] Pardey J,Mouldenb M,Redman C W G. A computer system for the numerical analysis of nonstress tests [J]. American journal of obstetrics and gynecology,2002,186(5):1095-1103.

[3] ZHU Cai-hong.Calculation of fetal heart rate baseline variation or disappearance with maternal delivery and neonatal outcome [J]. National Medical Journal, 2014,18: 74-77.

[4] Chamidah N, Wasito I. Fetal state classification from cardiotocography based on feature extraction using hybrid K-Means and support vector machine[C]// International Conference on Advanced Computer Science and Information Systems. IEEE, 2015:37-41.

[5] Comert Z, Kocamaz A F, Gungor S. Cardiotocography signals with artificial neural network and extreme learning machine[C]// Signal Processing and Communication Application Conference. 2016:1493-1496.

[6] Gavrilis D, Tsoulos I G. Classification of fetal heart rate using grammatical evolution[C]// Signal Processing Systems Design and Implementation, 2005. IEEE Workshop on. IEEE Xplore, 2005:425-429.

[7] Krupa N, Mohd Ali M A, Zahedi E, et al. Antepartum fetal heart rate feature extraction and classification using empirical mode decomposition and support vector machine[J]. BioMedical Engineering OnLine, 2011, 10(1):6.

[8] Das S, Roy K, Saha C K. Fuzzy Membership Estimation Using ANN: A Case Study in CTG Analysis[C]// Advances in Intelligent Systems and Computing. 2014:221-228.

[9] Johnson B, Bennett A, Kwak M, et al. Automated evaluation of fetal cardiotocograms using neural network[C]// IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2012:408-413.

[10] Ocak H, Ertunc H M. Prediction of fetal state from the cardiotocogram recordings using adaptive neuro-fuzzy inference systems[J]. Neural Computing and Applications, 2013, 23(6):1583-1589.

[11] Sahin H, Subasi A. Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques[J]. Applied Soft Computing, 2015, 33(C):231-238.

[12] Jezewski M, Leski J. Cardiotocographic Signals Classification Based on Clustering and Fuzzy If-Then Rules[M]// 5th European Conference of the International Federation for Medical and Biological Engineering. Springer Berlin Heidelberg, 2011:121-124.

[13] Pereira A, Salgado F, Reis L P, et al. Classification model for cardiotocographies[C]// Information Systems and Technologies. IEEE, 2016:1-6.

[14] Tomas P, Krohova J, Dohnalek P, et al. Classification of cardiotocography records by random forest[M]. 2013.

[15][J]. Journal of Chongqing Medical University, 2011, 36 (7): 846-849. [J]. Journal of Chongqing Medical University, 2011, 36 (7): 846-849. [J]. Journal of Chongqing Medical University, 2011, 36 (7): 846-849.

[16] UC Irvine.UCI Machine Learning Repository[DB/OL].http://www.uci.edu/.

[17] Scholz M, Klinkenberg R. An ensemble classifier for drifting concepts[J]. Proceedings of the Second International Workshop on Knowledge Discovery in Data Streams, 2005, 6(11):53--64.

[18] Krizhevsky A, Hinton G E. Using very deep autoencoders for content-based image retrieval[C]// Esann 2011, European Symposium on Artificial Neural Networks, Bruges, Belgium, April 27-29, 2011, Proceedings. DBLP, 2011.

[19] Peterek T, Smondrk M, Krohova J, et al. Comparison of classification techniques for actual fetal state determination[J]. Wit Transactions on the Built Environment, 2014, 145:187-194.

[20] Inbarani H H, Banu P K N, Azar A T. Feature selection using swarm-based relative reduct technique for fetal heart rate[J]. Neural Computing and Applications, 2014, 25(3):793-806.

[21] Yao Qiang, Liu Xinghui. Interpretation of Foreign Guide to Fetal Heart Electronic Monitoring [J]. Chinese Journal of Clinical Pharmacology, 2010, 6 (1): 11-13.

[22] Friedman J H. Greedy function approximation: A gradient boosting machine.[J]. Annals of Statistics, 2001, 29(5):1189-1232.

[23] Vishnubhotla S, Fernandez R, Ramabhadran B. An autoencoder neural-network based low-dimensionality approach to excitation modeling for HMM-based text-to-speech[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, Usa. DBLP, 2010:4614-4617.

[24] ZHANG Kai-xu, ZHOU Chang-le.Study on unsupervised learning of Chinese vocabulary based on automatic encoder [J] .Chinese Journal of Information, 2013, 27 (5): 1-7.

[25] Wang J, He H, Prokhorov D V. A Folded Neural Network Autoencoder for Dimensionality Reduction[J]. Procedia Computer Science, 2012, 13(6):120-127.

[26] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[J]. 2016:785-794

[27] Chen T, He T, Benesty M. xgboost: Extreme Gradient Boosting[J]. 2016.

[28] Attanayake, A.M.C.H., Jayasundara, D.D.M. and Peiris, T.S.G.. An Application of 5-fold Cross Validation on a Binary Logistic Regression Model. Advances and Applications in Statistics, 2016,49(6): 443-451.