Prediction of Delivery Mode from Fetal Heart Rate and Electronic Medical Records Using Machine Learning

Xue Kang[1], Rongdan Zeng[1], Hao Yi[1], Chuan Wang[1], Mujun Liu[2], Zheng Zheng[3], Yaosheng Lu[1,4], Huijin Wang[1] and Jieyun Bai[1,4]

[1]College of Information Science and Technology, Jinan University, Guangzhou, China
[1]College of Science & Engineering, Jinan University, Guangzhou, China
[3]Department of Obstetrics, Preterm Birth Prevention and Treatment Research Unit, Guangzhou Women Children's Medical Center, Guangzhou Medical University, Guangzhou, China
[4]Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Information Technology, Jinan University, Guangzhou 510632, China

Abstract

Background: Cardiotocography (CTG) is the most common method for monitoring the fetus during the early stages of labor, making effective decisions based on CTG data is a challenge. The American College of Obstetrics and Gynecology suggests that obstetricians in the proposal of delivery mode should take into account the maternity's prenatal electronic medical information and risk factors such as pregnancy complications.

Methods: This paper is conducted in two parts to identify the delivery mode: First, traditional fetal heart rate (FHR) features and maternal electronic medical record (EMR) features are extracted from the available FHR and prenatal medical records respectively. Then, several predi-ction models (Logistic Regression (LR), Gaussian Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), AdaBoost, Gradient Boost, Bagging Classifier, Extremely Randomized Trees Classifier (Extra Tree Classifier), Voting Classifier, Stacking Classifier) are trained to classify delivery mode for patients: vaginal or Cesarean Ssection (CS). An associa-tion based feature selection algorithm is employed on the features sets to remove the unnecessary features to improve the performance of classifiers. The data sets were composed of 784 signals collected between 2017 and 2020.

Results: The prediction models based on FHR features and EMR features take into account accuracy, sensitivity, specificity, and area under the curve receiver operating characteristic (AUC) in the outcomes. AdaBoost has the highest specificity (58.80%), KNN has the best accuracy (62.72%) and sensitivity (0.8131), and bagging has the highest AUC (58.03%). As predicted, the AUC of most classifiers improve after feature selection.

Conclusion: According to the obtained results, machine learning is proved with high value in predicting CS, and is a useful tool for reducing the false-positive rate and unnecessary operative interventions by employing FHR data and EMR information.

1. Introduction

According to United Nations International Children's Emergency Fund (UNICEF), 130 million babies are born every year, one million of which will be intrapartum stillbirths and more than 3.5 million of which will die from perinatal complications[1]. In many cases, caesarean section (CS) can save the lives of both the mother and the newborn. However, CS is a surgical procedure, which may lead to surgical wound infection and other complications arising from anesthesia surgical procedures[2].

Cardiotocography (CTG) is the most common method for monitoring the fetus during the early stages of labor[3]. At present, the assessment of CTG traces is usually carried out through visual inspection. However, visual inspection succumbs to qualitative interpretation of the operators, which leads to high inter- and intra-observer variability. Overinterpretation of CTG is common. 40%-60% of infants are born without any evidence to support pathological findings such as hypoxia and metabolic acidosis[4]. It is also the direct cause of unnecessary CS.

This paper aims to detect pathological cases using Fetal Heart Rate (FHR) and Machine Learning (ML), an objective measure of fetal status, which will provide obstetricians and midwives with an additional level of interpretation of fetal status and help decide whether surgical intervention is required. The experimental results show that the methods are highly predictive.

2. Methods and Materials

As shown in Figure 1, the methodology used in this paper includes feature extraction of FHR and maternal

Electronic Medical Record (EMR), and the extracted features are sorted by importance, screen the important features to a classifier with suitable functions to classify delivery mode.

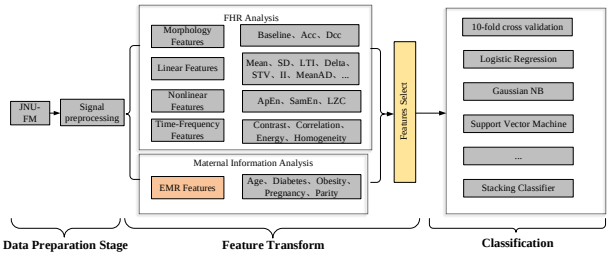

Figure 1. Procedure of Delivery Mode Prediction

2.1. Dataset

A total of 43,388 CTG records were collected at NanFang Hospital of Southern Medical University from January 2012 to November 2020. This study has been approved by the Medical Ethics Committee of NanFang Hospital of Southern Medical University (NFEC-2019-024). We select fetal monitoring data according to the following rules: singleton pregnancy, the rate of missed FHR signal missing per 10 minutes is less than 10%, the delivery time is between 36-42 gestational weeks and the CTG recording time exceeds 60 minutes. Finally, 784 data have been used in this paper.

2.2. Pre-processing

The artifact sampling points of the FHR signal are defined as invalid data points, which are numerically expressed as outliers outside 50-220bpm, or the value deviation of two adjacent points exceeds 25bpm. Referring to the research of Bernardes J[5], invalid data points are detected and eliminated through a 5-min sliding window, and linear interpolation is performed between two adjacent stable FHR segments, each FHR value is replaced by the average of 5 values center around it to obtain the pre-processed FHR signal.

2.3. Features Extraction

The International Federation of Gynecology and Obstetrics (FJGO) has established several guidelines for the interpretation of CTG signals[6], of which the most important features of FHR include baseline, acceleration, deceleration, Short-Term Variability (STV), Long-Term Variability (LTV)[7], root mean square, mean, standard deviation, mean absolute deviation, median absolute deviation, skewness, kurtosis[8], delta and interval index[9]. These features can be interpreted as visual cues for obstetricians to monitor the fetus.

Nonlinear features, including Sample Entropy (SampEn), Approximate Entropy (ApEn)[10], and Lempel-Ziv complexity (LZC), are obtained by nonlinear changes to the pre-processed FHR signal, which can measure the irregularity of the FHR, and it is a useful indicator for the detection of fetal hypoxia and metabolic acidosis.

A study has shown that fetal autonomic nervous system activities can be observed in different frequency domains[11]. In this paper, fractal texture features, including contrast, correlation, energy and homogeneity of time-frequency images, are extracted from four frequency domains: very low frequency (0-0.03Hz), low frequency (0.03-0.15Hz), medium frequency (0.15-0.50Hz) and high frequency (0.50-1Hz).

A study has shown that maternal age, parity, gravidity and some pre-existing diseases such as gestational obesity and gestational diabetes increase the possibility of CS[12], so the corresponding EMR features are extracted in this paper, including maternal age, gravity, parity, gestational diabetes, gestational obesity.

At the end of the feature extraction stage, 43 features and parameters are extracted from the morphological domain, linear domain, nonlinear domain, time-frequency domain, statistical features and EMR respectively.

2.4. Feature Importance Analysis

From the perspective of interpretability, in order to measure the value of feature variables in building a decision model, the Shapley value is used to estimate the relative importance of each feature. In machine learning training tasks[13]. The essence of Shapley value is to calculate the average marginal benefit of a specific feature by looking for possible feature subset permutations and combinations, representing the contribution to the model prediction.

The python SHAP tool is used in this study. The Shapley value decomposes each individual's deviation from the delivery mode into the contribution of each category feature, generates the Shapley value of the category feature matrix, and calculates the overall feature importance for this sample by averaging the individual Shapley values for each sample.

2.5. Classifier and Evaluation Metrics

In this paper, we consider several simple and powerful classifiers, such as Logistic Regression (LR), Gaussian Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF)[14], Decision Tree(DT), eXtreme Gradient Boosting (XGB-oost), Light Gradient Boosting Machine (LightGBM)[15], AdaBoost, Gradient Boost, Bagging Classifier, Extremely Randomized Trees Classifier (ExtraTree Classifier)[16],

Voting Classifier, Stacking Classifier. Before using the above methods for classification, we divide the dataset into training set, validation set, and test set by 8:1:1, all algorithms use 10-fold cross-validation.

In this study, "positive" represents "CS" and "negative" represents "vaginal delivery", and a confusion matrix consisting of four prognostic indicators, True Positive (TP), False Positive (TP), True Negative (TN) and False Negative (FN), is used to evaluate Classifier performance. The confusion matrix ensures several statistical performance metrics such as Accuracy, Sensitivity, Specificity, Quality Index (QI) and Area Under the receiver operating characteristic Curve (AUC) to measure the efficiency of the classifier. The performance indicators are as follows:

- Accuracy = the sum of TP and TN over the total sample;
- Sensitivity = the amount of TP over the sum of TP and FN;
- Specificity = the amount of TN over the sum of TN and FP;
- AUC = Area under the receiver operating characteristic curve;
- Quality index (QI) = The square root of the product of Sensitivity and Specificity;

3. Result and Discussion

In this paper, we only focus on the last 30 minutes of the first stage of labour for distinguishing between vaginal deliveries and CS samples, we obtained 233 CS and 551 vaginal delivery signals.

Based on all the features, various ML algorithms are used to predict the delivery mode, and the results are shown in Table 1. It is evident from the experimental results that the Bagging Classifier outperforms all other classifiers with accuracy of 59.18%, sensitivity of 61.16%, specificity of 57.74%, and AUC of 58.03%. The second classifier is SVM, which has QI value of 57.16%. Among all classifiers, KNN has the highest sensitivity of 81.31%, but the specificity is only 18.88%, which means that KNN is accurate in predicting CS, but the accuracy of predicting vaginal delivery is low. As an integrated learner, XGBoost also has a high accuracy rate for CS prediction, and its sensitivity is 67.15%. The Voting Classifier has the highest AUC of 58.42%, indicating that the model has high reliability.

In order to measure the value of feature variables in the construction of decision-making models, this study uses Shapley value for feature importance calculation. The importance ranking of the Shapley value of the Voting Classifier (top 20) is shown in Figure 2. In order to establish an interpretable streamlined delivery mode prediction model, we rank the importance of the Shapley value of the Voting Classifier, and calculate the classification effects of the top 5, 10, 15, 20, 25, 30, 35, 40 features respectively. It can be seen from Table 2 that the top 20 features have the best classification performance.

Table 1. Classification performance metrics for all features after pre-processing.

	accuracy	sensitivity	specificity	QI	AUC
LR	0.5689	0.5789	0.5451	0.5618	0.5847
Gaussian NB	0.5239	0.5154	0.5622	0.5383	0.5581
SVM	0.5702	0.5681	0.5751	0.5716	0.5729
KNN	0.6276	0.8131	0.1888	0.3918	0.5302
RF	0.5370	0.5426	0.5236	0.5330	0.5274
DT	0.5561	0.5898	0.4761	0.5310	0.5372
XGB	0.5982	0.6715	0.4249	0.5342	0.5493
LGB	05281	0.5390	0.5021	0.5203	0.5260
AdaBoost	0.4732	0.4247	0.5880	0.4997	0.5115
Gradient Boost	0.5395	0.5517	0.5107	0.5308	0.5347
Bagging	0.5918	0.6116	0.5451	0.5774	0.5803
Extra Tree	0.5077	0.4374	0.6738	0.5429	0.5483
Voting	0.5625	0.5572	0.5751	0.5661	0.5842
Stacking	0.5497	0.5227	0.6137	0.5664	0.5772

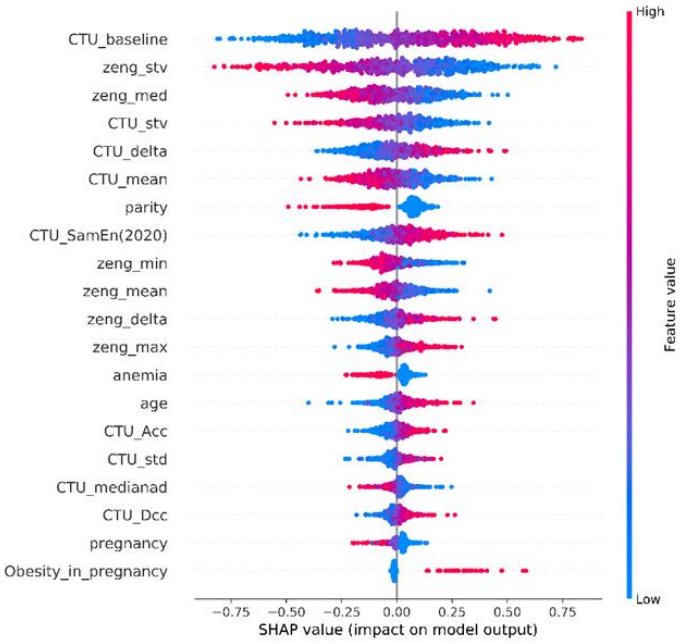

Figure 2. Shapley value importance ranking of Voting Classifier (top 20)

Finally, the top 20 important features are used to train the model. The prediction results of various algorithms based on the selected features for the mode of delivery are shown in Table 3. It can be seen from the table that better results can be obtained by training the model after proper feature selection., where the QI value of the Bagging Classifier is improved to 59.30% and the QI value of the Voting Classifier is improved to 58.20%. The AUC of the Stacking Classifier is improved to 59.10%.

Table 2. Classification performance metrics of different importance features in Voting Classifier

	accuracy	sensitivity	specificity	QI	AUC
Top5 features	0.5064	0.4519	0.6352	0.5358	0.5637
Top10 features	0.5497	0.4991	0.6695	0.5781	0.5986
Top15 features	0.5587	0.5263	0.6352	0.5782	0.6070
Top20 features	0.5599	0.5227	0.6481	0.5820	0.5909
Top25 features	0.5561	0.5263	0.6266	0.5743	0.5813
Top30 features	0.5574	0.5481	0.5794	0.5635	0.5886
Top35 features	0.5561	0.5517	0.5665	0.5591	0.5873
Top40 features	0.5612	0.5554	0.5751	0.5651	0.5841

Table 3. Classification performance metrics for top 20 features using Shapley value importance

	accuracy	sensitivity	specificity	QI	AUC
LR	0.5663	0.5681	0.5622	0.5651	0.5989
Gaussian NB	0.5115	0.4864	0.5708	0.5269,	0.5378
SVM	0.5051	0.4719	0.5837	0.5248	0.5460
KNN	0.6237	0.8185	0.1631	0.3654	0.5210
RF	0.5153	0.4592	0.6481	0.5455	0.5426
DT	0.5612	0.6189	0.4249	0.5128	0.5213
XGB	0.5574	0.5935	0.4724	0.5293	0.5471
LGB	0.4401	0.3848	0.5708,	0.4686	0.4686
AdaBoost	0.4184	0.2995	0.6996	0.4577	0.4987
Gradient Boost	0.5969	0.6969	0.3605	0.5012	0.5270
Bagging	0.5829	0.5717	0.6094	0.5903	0.5855
Extra Tree	0.5051	0.4428	0.6524	0.5375	0.5550
Voting	0.5599	0.5227	0.6481	0.5820	0.5909
Stacking	0.5625	0.5499	0.5923	0.5707	0.5951

4. Conclusion and Future Work

According to the results obtained in this study, ML can be incorporated into the clinic as an aid for physicians in decision-making about the delivery mode. In fact, the ensemble approach is the best in this case. This study only used the characteristics of FHR. We can add relevant features of Uterine Contraction (UC) signals, explore possible relationships between UC and FHR signals, and add them to the feature set in the future work.

References

[1] J. B. Warren, W. E. Lambert, R. Fu, J. M. Anderson, A. B. J. R. Edelman, and R. i. Neonatology, "Global neonatal and perinatal mortality: a review and case study for the Loreto Province of Peru," vol. 2, pp. 103-113, 2012.

[2] S. M. J. M. c. r. Koroukian and review, "Relative risk of postpartum complications in the Ohio Medicaid population: vaginal versus cesarean delivery," vol. 61, no. 2, pp. 203-224, 2004.

[3] A. J. B. A. I. J. o. O. Ugwumadu and Gynaecology, "Are we (mis) guided by current guidelines on intrapartum fetal heart rate monitoring? Case for a more physiological approach to interpretation," vol. 121, no. 9, pp. 1063-1070, 2014.

[4] J. Spilka, G. Georgoulas, P. Karvelis, V. Chudáček, C. D. Stylios, and L. Lhotská, "Discriminating normal from "abnormal" pregnancy cases using an automated fhr evaluation method," in Hellenic Conference on Artificial Intelligence, 2014, pp. 521-531: Springer.

[5] D. Ayres-de-Campos, J. Bernardes, A. Garrido, J. Marques-de-Sa, and L. J. J. o. M.-F. M. Pereira-Leite, "SisPorto 2.0: a program for automated analysis of cardiotocograms," vol. 9, no. 5, pp. 311-318, 2000.

[6] M. Cesarelli, M. Romano, P. Bifulco, F. Fedele, M. J. C. i. b. Bracale, and medicine, "An algorithm for the recovery of fetal heart rate series from CTG data," vol. 37, no. 5, pp. 663-669, 2007.

[7] I. Nunes, D. J. B. P. Ayres-de-Campos, R. C. Obstetrics, and Gynaecology, "Computer analysis of foetal monitoring signals," vol. 30, pp. 68-78, 2016.

[8] Z. Cömert and A. F. Kocamaz, "Using wavelet transform for cardiotocography signals classification," in 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017, pp. 1-4: Ieee.

[9] D. Arduini, G. Rizzo, G. Piana, A. Bonalumi, P. Brambilla, and C. J. J. o. M.-F. I. Romanini, "COMPUTERIZED ANALYSIS OF FETAL HEART-RATE. 1. DESCRIPTION OF THE SYSTEM (2CTG)," vol. 3, no. 3, pp. 159-163, 1993.

[10] J. S. Richman, J. R. J. A. J. o. P.-H. Moorman, and C. Physiology, "Physiological time-series analysis using approximate entropy and sample entropy," vol. 278, no. 6, pp. H2039-H2049, 2000.

[11] H. Gonçalves, A. P. Rocha, D. Ayres-de-Campos, J. J. M. Bernardes, B. Engineering, and Computing, "Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery," vol. 44, no. 10, pp. 847-855, 2006.

[12] I. Mylonas and K. J. D. Ä. I. Friese, "Indications for and risks of elective cesarean section," vol. 112, no. 29-30, p. 489, 2015.

[13] L. S. J. C. i. g. t. Shapley, "A value for n-person games," vol. 69, 1997.

[14] N. Ramakrishnan et al., "The Top Ten Algorithms in Data Mining," ed: Taylor & Francis Group, 2009.

[15] G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," vol. 30, 2017.

[16] P. Geurts, D. Ernst, and L. J. M. l. Wehenkel, "Extremely randomized trees," vol. 63, no. 1, pp. 3-42, 2006.

Address for correspondence:

Jieyun Bai
College of Information Science and Technology, Jinan University, Guangzhou 510632, China
bai_jieyun@126.com