Data Analysis and Optimal Prediction Model Search based on CTG Data

Siyuan Jiang [1, *, †], Tengyang Lian [2, †] and Yuwen Wei [3, †]

[1] Pharmaceutical engineering, University of Shanghai for Science and Technology, Shanghai, China

[2] Fujian Medical University, Fuzhou, China

[3] College of Future Technology, Xinjiang University, Urumqi, China

* Corresponding author: weiyuwen@stu.xju.edu.cn

†These authors contributed equally.

Abstract. This paper presents a data analysis and optimal prediction model search of fetal health based on cardiotocography (CTG) data. The objective of this study is to develop an accurate and efficient method to predict fetal health outcomes using CTG data. This paper first analyzes the dataset to identify potential predictors of fetal health and investigate their relationship with fetal distress. The result found a strong relationship between some of the variables in the dataset and fetal health. This paper then uses machine learning algorithms to build and compare several prediction models, including elastic net regression and lasso model. The result of this paper shows that a random forest model performs best in terms of AUC. The model can accurately predict fetal health outcomes with an AUC of 0.925. The finding of this paper suggests that CTG data analysis combined with machine learning algorithms can provide a useful tool for prenatal monitoring and management.

Keywords: CTG, fetal health, elastic net regression, Lasso model.

1. Introduction

Fetal health monitoring during pregnancy is a critical aspect of obstetric care, with the aim of preventing adverse pregnancy outcomes such as fetal distress, preterm labor, and intrauterine growth restriction. Fetal monitoring involves the recording of fetal heart rate (FHR) and uterine contractions, typically using external tocodynamometers and ultrasound transducers. Cardiotocography (CTG) is a widely used method for monitoring fetal health, providing information about fetal heart rate and uterine contractions. However, interpretation of CTG data is complex, and the accuracy of the traditional CTG interpretation system has been questioned.

This paper proposes to develop and evaluate prediction models for fetal health outcomes using two different methods: lasso model and elastic net regression. These methods have been widely used in data analysis and have shown promising results in various applications. By applying these methods to the fetal health classification dataset from Kaggle, the paper aims to develop prediction models that can accurately predict the fetal health status based on CTG data. The fetal health classification dataset from Kaggle consists of 21 features extracted from CTG recordings of 2126 fetuses. The features include baseline FHR, accelerations, decelerations, and various other indicators of fetal health. The dataset is labeled with three classes: Normal, Suspect, and Pathological. The dataset is imbalanced, with the Normal class having the majority of the samples.

To improve the accuracy of fetal health prediction based on CTG data, researchers have explored the use of machine learning techniques. Among these techniques, the elastic net regression model has been shown to be effective in predicting fetal health outcomes based on CTG data [1]. This model combines the strengths of ridge and lasso regression, allowing for the selection of relevant variables while also avoiding overfitting.

Other machine learning models have also been applied to fetal health prediction. For example, a study by Oweis et al. applied the random forest algorithm to CTG data and achieved high accuracy

in predicting fetal acidosis [2]. Another study by Tharwat et al. used a support vector machine to predict fetal distress based on CTG data and achieved an accuracy of 88.4% [3].

To further improve the accuracy of fetal health prediction, researchers have also explored the use of deep learning techniques. For example, Wang et al. applied a convolutional neural network to CTG data and achieved an accuracy of 92.4% in predicting fetal distress [4]. Another study by Ren et al. used a deep belief network to predict fetal hypoxia based on CTG data and achieved an accuracy of 91.2% [5].

In addition to machine learning models, researchers have also explored the use of feature selection techniques to identify the most relevant CTG features for fetal health prediction. A study by Shoushtarian et al. used a correlation-based feature selection approach to identify the most important CTG features for predicting fetal distress [6]. Another study by Zhang et al. used a genetic algorithm-based feature selection approach to identify the most informative CTG features for predicting fetal acidosis [7].

While machine learning and feature selection techniques have shown promise in improving fetal health prediction, there are still challenges that need to be addressed. One challenge is the imbalance of fetal health outcomes in the dataset, with normal outcomes being more common than abnormal outcomes. To address this issue, researchers have explored the use of data augmentation techniques such as over-sampling and under-sampling. A study by Gutiérrez-Ríos et al. applied oversampling techniques to balance the dataset and achieved improved performance in fetal health prediction [8].

Another challenge is the lack of interpretability of machine learning models, which can make it difficult for clinicians to understand the reasoning behind a prediction. To address this challenge, researchers have explored the use of explainable artificial intelligence (XAI) techniques. A study by Hu et al. used a decision tree-based XAI approach to provide clinicians with explanations for fetal distress prediction based on CTG data [9].

This paper aims to develop and evaluate prediction models for fetal health outcomes using lasso model and elastic net regression methods and to compare their performance. By analyzing the fetal health classification dataset from Kaggle, the paper aims to provide insights into the development of accurate prediction models for fetal health monitoring. The results of this study have the potential to improve prenatal care and contribute to better fetal health outcomes [10].

2. Method

This dataset consists of 2126 Cardiotocography measurements that were classified by three expert obstetricians into three categories: Normal, Suspect, and Pathological. The features were extracted from the exams and used for analysis [11]. Table 1 shows all the variables and some information of the dataset.

Table 1. Data information

Variable name	Symbol	Variable name	symbol
baseline value	X1	histogram_width	X12
accelerations	X2	histogram_min	X13
fetal_movement	X3	histogram_max	X14
uterine_contractions	X4	histogram_number_of_peaks	X15
light_decelerations	X5	histogram_number_of_zeroes	X16
severe_decelerations	X6	histogram_mode	X17
prolongued_decelerations	X7	histogram_mean	X18
abnormal_short_term_variability	X8	histogram_median	X19
mean _of_short_term_variability	X9	histogram_variance	X20
percentage_of_time_with_abnormal_long_term_variability	X10	histogram_tendency	X21
mean_ of_long_term_variability	X11	fetal_health	X22

To deal with the above variables, this paper further utilized correlation analysis, LASSO regression and Plastic Net Regression upon this dataset. The results are shown in the following section.

3. Results and Discussion

3.1. Preliminary Correlation Processing

First, this paper makes a preliminary analysis of the correlation between all variables in the data set. And the correlation graph is obtained, as Figure 1 shows.

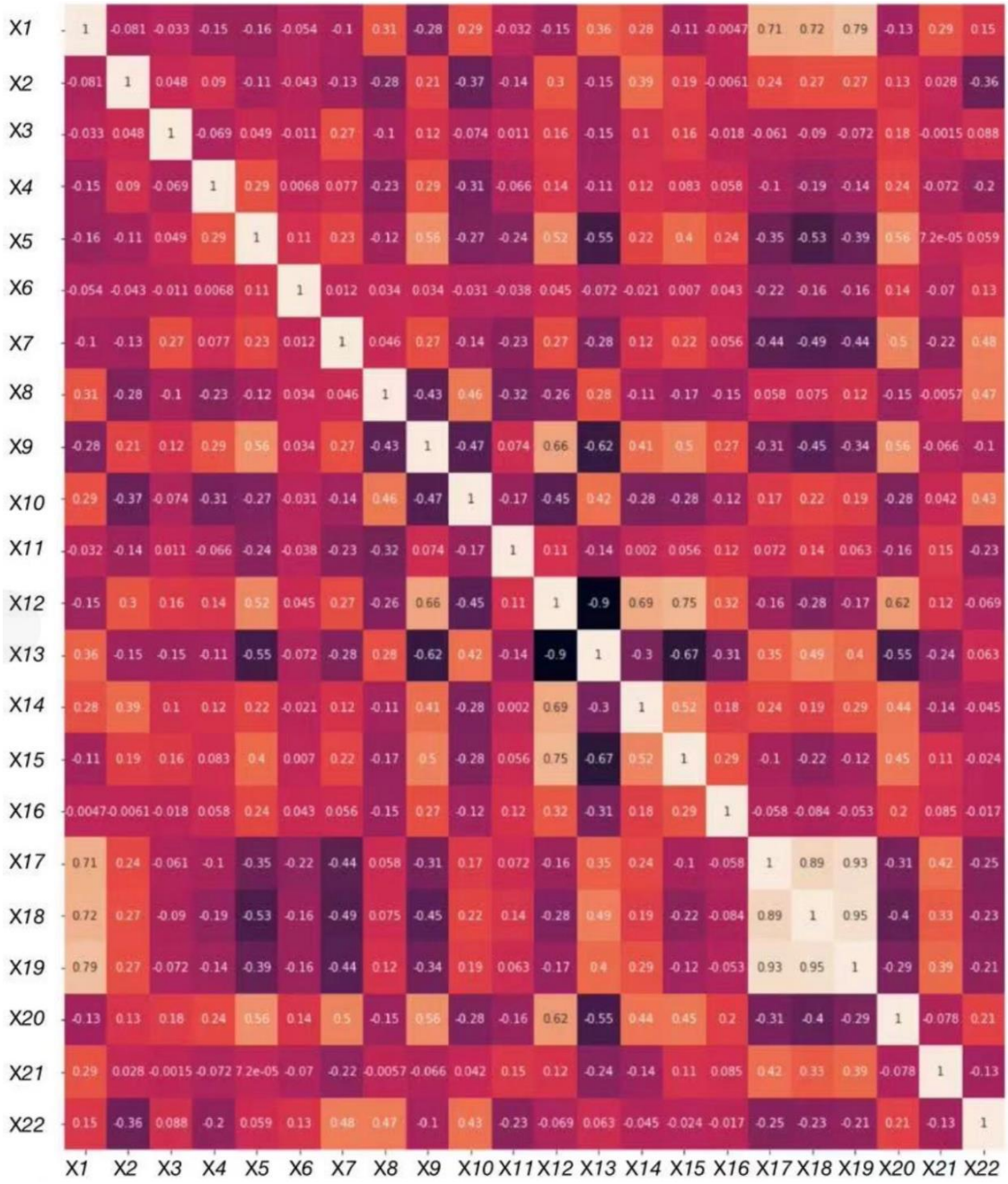

Figure 1. Correlation between variables

This paper is more concerned about the relationship between different columns and fetal health. So, the paper plotted out Table 2. The columns "Prolongued Decelerations", "Abnormal Short-Term Variability", and "Percentage of Time with Abnormal Long-Term Variability" have been shown to be strongly correlated with fetal health.

Table 2. Correlations between fetal health and 22 columns

Variables	X22	Variables	X22
X1	0.148	X12	-0.069
X2	-0.364	X13	0.063
X3	0.088	X14	-0.045
X4	-0.204	X15	-0.023
X5	0.059	X16	-0.017
X6	0.131	X17	-0.250
X7	0.485	X18	-0.227
X8	0.471	X19	-0.205
X9	-0.103	X20	0.207
X10	0.426	X21	-0.132
X11	-0.227	X22	1

By examining Fig. 2, despite not all pairs of features are displayed, we can gain several insights into the relationships between different features and fetal health. "Abnormal Short-Term Variability" appears to be correlated with fetal health, as the mean of the "Normal" class is shifted towards lower values compared to the other classes. "Percentage of Time with Abnormal Long-Term Variability" has high kurtosis, while the "Suspect" and "Pathological" classes have low kurtosis. Conversely, the "Number of Accelerations Per Second" feature exhibits low kurtosis for the "Normal" class, but high kurtosis for the other classes. "Histogram Mean" and "Histogram Mode" are highly correlated with each other and with the "Pathological" class, as their distribution means are shifted towards lower values. The relationships between the features are largely linear, although there may be some slight non-linearity present, between "Histogram Mean" and "Histogram Mode", as well as between "Abnormal Long Term Variability Time" and "Abnormal Short-Term Variability". This paper obviously showed that "Abnormal Short-Term Variability" and "Abnormal Long Term Variability Time" are very negative correlated with our class normal, and the opposite with "Pathological" and "Suspect". This same pattern holds true for the next negatively correlated feature.

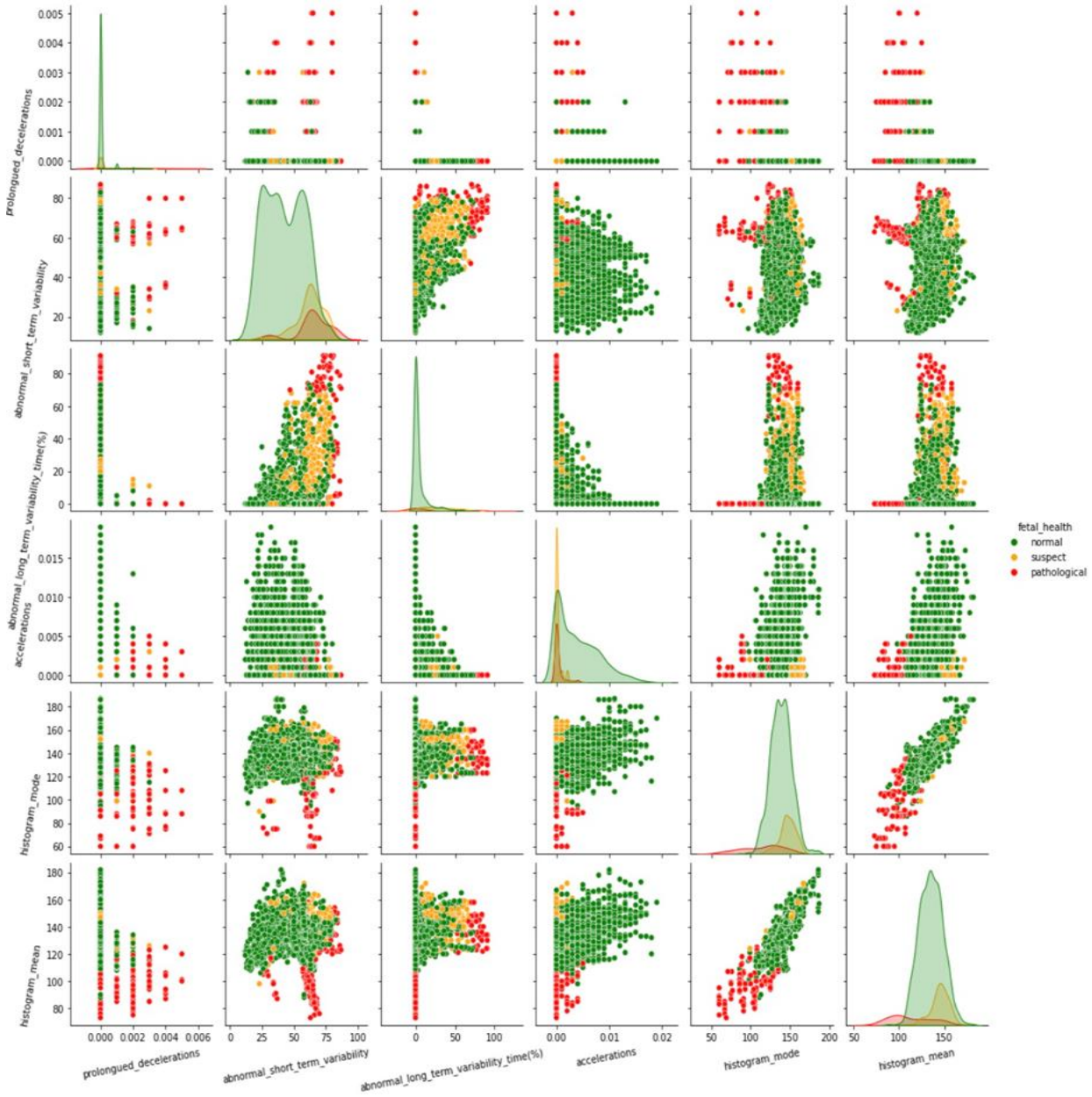

Figure 2. Dot plot of correlation of columns

3.2. LASSO Regression and Plastic Net Regression

This research focus on the relationship between the different values and the output. Because the number of the kinds of the values is 21, LASSO Regression can be useful in both modeling and training. LASSO method is also called least absolute shrinkage and selection operator. It performs both variable selection and regularization to enhance the prediction accuracy, interpretability of the resulting statistical model, and it also helps to avoid overfitting. LASSO method can distinguish the significant values out of others in an efficient way. For a set of casual data, some of the values is associated with the output, while others are not able to affect the output. LASSO Regression will test whether a value will affect the output. Through this process, the P value representing the significance will be showed, indicating which of the values can be used in modeling.

The problem of LASSO Regression is that the model may delete some less significant value, leading to less fit of the model. In order to overcomes the limitations of the LASSO method, another kind of regression model is better. Elastic Net Regression is helpful when the significance of the data is still unknown and the relationship is undecided. This method is mixed with LASSO Regression and Ridge Regression, using the alpha value to balance the proportion of two regression. The alpha

value in this research is 0.5, meaning that the percentage of LASSO Regression and Ridge Regression is equal.

ROC curve is used to evaluate both the models. Receiver Operating Characteristic (ROC) curve is a plot that displays the performance of a binary classifier system at different discrimination thresholds. The plot shows the trade-off between the true positive rate and the false positive rate for different threshold values. It is widely used in analyzing whether the model is good for predicting in machine learning. Area Under the Curve, also known as AUC value, is between 0 and 1. The closer it is to 1, the more reliable of the model is. Specificity stands for false positives are found, while sensitive means true positives are found, as Figure 3 and 4 showing.

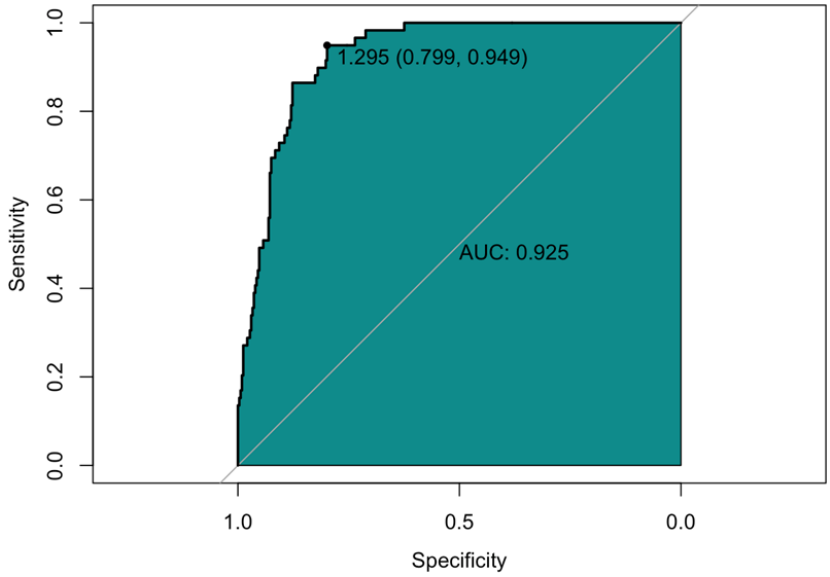

Figure 3. AUC value (Lasso Model)

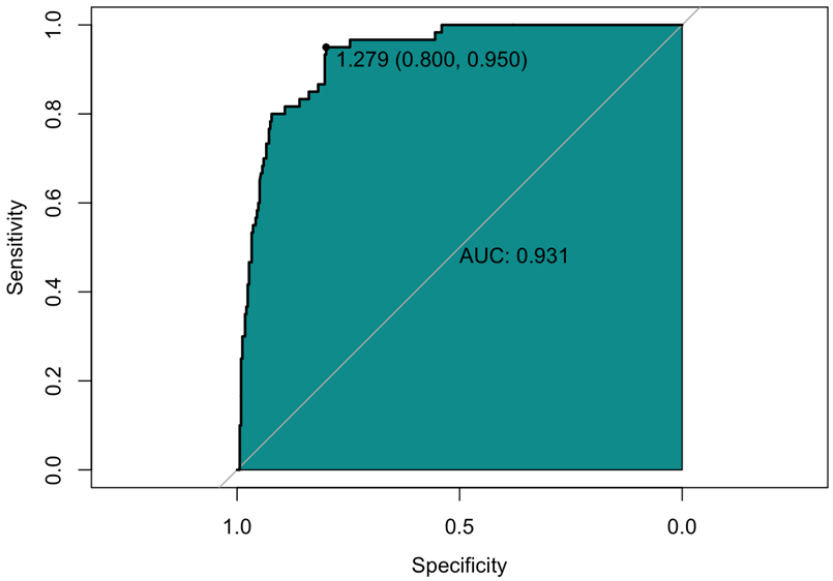

Figure 4. AUC value (Elastic Net Regression)

4. Conclusion

Elastic Net Regression was used to develop a predictively model for fetal health through CTG data. The final model with the AUC value of 0.931, indicates a high-level accuracy in predicting fetal health. The Elastic Net Regression technique integrates the strengths of both Lasso and Ridge regression methods, enabling feature selection and regularization simultaneously. In the analysis, the Elastic Net Regression method was used to minimize the risk of overfitting.

The three most important variables in the model were baseline value (mean fetal heart rate during the recording), accelerations (increases in fetal heart rate), and the number of prolonged decelerations (decreases in fetal heart rate lasting longer than 3 minutes). These variables provide valuable information about fetal health during cardiotocography exams. These variables have shown to be important predictors of fetal health outcomes in previous studies The results are consistent with these findings, demonstrating the importance of these variables in predicting fetal health outcomes.

In conclusion, this study highlights the best predictive model through CTG data, and also shows the potential of Elastic Net Regression in model building. The variables selected in the model are consistent with previous studies, indicating the importance of these variables in predicting fetal health outcomes. Further studies could explore the use of other machine learning methods and the incorporation of additional variables to improve the accuracy of prediction models for fetal health outcomes.

References

[1] Ayres-de-Campos D, Bernardes J. Fetal Monitoring: The Old and the New. Best Pract Res Clin Obstet Gynae Col, 2016, 30: 3 - 12.

[2] Oweis R J, Haddad F S, Kanaan M. Fetal Acidosis Prediction Using Random Forest Algorithm. Int J Med Inform, 2019.

[3] Tharwat A, Gaber T, Hassanien A E. A Hybrid Model for Fetal Distress Classification. Comput Biol Med, 2017, 81: 67 - 75.

[4] Wang Y, Yao Y, Lu X. Fetal Distress Detection by Convolutional Neural Network. IEEE Access, 2019, 7: 6607 - 36614.

[5] Ren H, Zhao T, Wang L, et al. Fetal Hypoxia Detection by Deep Belief Network with Intrapartum Cardiotocography. J Matern Fetal Neonatal Med, 2020, 33 (4): 558 - 565.

[6] Shoushtarian M, Hajizadeh S, Khalilzadeh MA, et al. Correlation-Based Feature Selection for Fetal Distress Classification. Comput Biol Med, 2014, 45: 1 - 8.

[7] Zhang X, Zhou Y, Lin C, Zhao Y. Genetic Algorithm Based Feature Selection for Fetal Distress Classification. Comput Biol Med, 2015, 59: 42 - 51.

[8] Gutiérrez-Ríos R, González-Lorenzo M, Alcalá-Fdez J, et al. Fetal Health Classification by Means of Data Augmentation and Random Forests. PLoS One, 2019, 14 (4): e0215154.

[9] Hu J, Li Y, Yan J, Li M, He X. Explainable Artificial Intelligence for Fetal Distress Prediction. BMC Med Inform Decis Mak, 2020, 20 (1): 192.

[10] Ilea D E, Siriteanu T, Kovacs L. Fetal health classification using support vector machines and neural networks. 2012 IEEE 10th International Symposium on Intelligent Systems and Informatics, 2012.

[11] Ayres de Campos et al. SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med, 2000, 5: 311 - 318.