Detection of Fetal State using Cardiotocogram Features

Nehal Linganur
University of Massachusetts, Amherst
Amherst, United States
nlinganur@umass.edu

Matthew Hawley
University of Massachusetts, Amherst
Amherst, United States
mhawley@umass.edu

***Abstract*—Fetal mortality is an ever-growing health hazard that can be prevented by using certain exams and procedures. While performing Cardiotocography, specialists can identify certain features that can influence fetal heart rates and uterine contractions during labor in order to prevent complications. In this project, we aim to train 3 different classification models - Logistic Regression, Random Forest, and Support Vector Classifier - to predict the classification of the fetal state depending on the Cardiotocogram (CTG) features provided in a data set. We also intend to identify which features contribute to each class the most by looking at the coefficients, and which of the 3 models is the best one. The data set consists of 21 feature variables and 3 class labels - ”Normal”, ”Suspect”, and ”Pathological” - and was tested using the methods mentioned above. After training the models and conducted statistical analyses, we concluded that the model with the highest performance was the Random Forest model since it produced the highest F1-scores. We also found that there was a statistically significant difference between each pair of classes for each of the top features identified by the Random Forest Classifier.**

***Index Terms*—Cardiotocography, CTG, Logistic Regression, Random Forest, Support Vector Classifier, F1-scores**

I. Introduction

A. Background

Fetal heart rate is an essential aspect of labor and childbirth when considering the safety and well-being of both the mother and child. In order to ensure that there are no complications Cardiotocograms, more commonly represented as CTGs, are used. These devices measure many vital features, such as fetal heart rate, fetal movement, and uterine contractions. Fetal heart rate (FHR) is influenced by many factors, and these factors give rise to distinct FHR patterns, which are categorized as baseline, variability, accelerations, and decelerations. Baseline patterns refer to the average FHR, excluding accelerations and decelerations. Variability refers to fluctuations of the FHR around the baseline. Accelerations occur when there is an increase in the FHR above the baseline, which indicates a healthy response of the fetus to an external stimuli. Decelerations are a decrease in the FHR below the baseline, which could be a result of delayed responses to stimuli.

Uterine contractions, which are resulted from the coordinated activity of uterine smooth muscles, also play an important role during labor. Oxytocin influences the frequency and intensity of uterine contractions. Intense uterine contractions can limit blood flow, which can cause decelerations in the FHR, indicating an abnormal situation or complication.

B. Motivation

Now that we understand the importance of measuring FHR and uterine contractions, we can discuss some of the common problems that arise in the context of fetal mortality. Fetal mortality still remains to be one of the most traumatic complications to occur during the journey of pregnancy. The loss of a fetus before birth not only devastates a family emotionally, but it also has major impacts on overall maternal health and societies. Maternal health can be compromised by engaging in activities such as substance abuse, isolation, neglect, and self harm. Similarly, fetal mortality can have negative impacts on societies, as high rates of fetal mortality can indicate health care challenges, insufficient infrastructure, and inadequate training of healthcare professionals.

In order to prevent these concerns and reduce fetal mortality, it is important to be able to identify certain features that can affect FHR and uterine contractions. By using data obtained from CTGs, we can predict the classification of the fetal state at a much earlier stage so precautions can be taken to prevent fetal mortality.

II. Methodology

A. Hypothesis

For this report, we wanted to perform testing on the best model’s top 3 features. Before training the models and performing the hypothesis tests, we wanted to make sure that we could identify whether there was a difference in the mean value of the percentage of time with the top 3 features in the model. We also wanted to identify whether in each pair of the classes, if there would be a difference in the mean value of the percentage of time with the top 3 features (i.e. ”Normal” vs ”Suspect”, ”Normal” vs ”Pathological”, and ”Suspect” vs ”Pathological”). Our null hypothesis was that there wouldn’t be a difference in the mean values with or without the pairs. Our alternative hypothesis was that there would be a significant difference in the mean values of the top 3 features. In order to ensure that the variances among the classes for each feature was not equal, we used Pingouin’s homoskedacity method to verify that they were indeed not equal to each other.

B. Dataset

The dataset we used for this project was obtained from the archives of the UC Irvine Machine Learning Repository. This dataset consists of measurements of the FHR and uterine contraction features on CTGs classified by expert obstetricians. The dataset has a total of 21 feature variables and 3 class labels, and our first step was to split all the variables into feature variables and class labels. We also obtained the proportion of the labels to see if the dataset was imbalanced and our results indicated that the prevalence of ”Normal” cases was much higher than the other classes, suggesting that our dataset was indeed imbalanced. This is the data we used in our classification models to predict the classification of the fetal state.

Feature	Description
LB	FHR baseline (bpm)
AC	number of accelerations/sec
FM	number of fetal movements/sec
UC	number of uterine contractions/sec
DL	number of light decelerations/sec
DS	number of severe decelerations/sec
DP	number of prolonged decelerations/sec
ASTV	percentage of time with abnormal short term variability
MSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
MLTV	mean value of long term variability
Width	width of FHR histogram
Min	minimum of FHR histogram
Max	maximum of FHR histogram
NMax	number of histogram peaks
NZeros	number of histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency

TABLE I
Table 1. UC Irvince dataset used

III. Classification Models

A. Logistic Regression

Before we trained the Logistic Regression model, we first split the dataset into a 70% training set and a 30% testing set. We then made sure to scale the training data to make sure that all the features were at the same scale in order to prevent certain features from dominating over others. All the training features had a mean of 0 and a standard deviation of 1.

The hyperparameters used for our Regression model included a ”C” value and the GridSearchCV function. A ”C” value is used to regularize the model to prevent overfitting, and a smaller ”C” value relates to stronger regularization. The value of ”C” we obtained was 0.219. The GridSearchCV function is used to find the optimal ”C” value, and we performed a 5-fold cross validation to find this value.

We then trained our Logistic Regression model for the training and testing set and obtained very similar precision, recall, f1-scores, and average scores for all 3 classes. Since the results were similar across all classes for both sets of data, we can assume that the model wasn’t overfitted, despite it performing slightly worse for the ”Suspect” class compared to the other two classes.

	Precision	Recall	F1-Score
Normal	0.92	0.96	0.94
Suspect	0.68	0.59	0.63
Pathological	0.85	0.75	0.80

TABLE II
Confusion Matrix for Testing set

B. Random Forest Classifier

The hyperparameters used for our Random Forest Classifier model include the n_estimators and the ”class_weights = balanced” parameters. N_estimators tells us the number of trees used in our forest, and the number of trees should be a relatively higher value in order to improve the performance of the model. We used an n_estimators value of 100. The ”class_weights = balanced” parameter deals with an imbalanced dataset and gives equal priority to all classes, despite the prevalence of imbalances.

For our testing set, the Random Forest model performed well since the scores for all 3 classes were high. However for the training set, we observed that all the scores were a perfect score of 1.0. This was somewhat suspicious, since the model may not have generalized the data properly, which is an indication that the model may be overfitted.

	Precision	Recall	F1-Score
Normal	0.96	0.87	0.96
Suspect	0.84	0.81	0.82
Pathological	0.91	0.91	0.91

TABLE III
Confusion Matrix for Testing set

C. Support Vector Classifier

For SVC models, the ”C” hyperparameter is similar to that of the Logistic Regression model in the sense that it prevents overfitting. A larger ”C” value allows for individual data points to be included in the model, which could potentially be a risk for overfitting. A smaller ”C” value allows for more training data to be misclassified, which is preferred for an imbalanced data set like the one we were using. Another hyperparameter used in this model is the ”gamma” parameter. This parameter works similarly to the ”C” parameter - it must be smaller to prevent overfitting.

For our testing set, the SVC model performs very well since the scores across the 3 classes are relatively high. For our training set, the scores were also high, which suggests good performance. However, we saw that the performance across all 3 classes is not consistent in the training and testing set - the scores for the ”Suspect” class in the testing set are lower than in the training one. This could be a result of some sort of overfitting.

	Precision	Recall	F1-Score
Normal	0.95	0.95	0.95
Suspect	0.76	0.75	0.75
Pathological	0.84	0.87	0.85

TABLE IV
CONFUSION MATRIX FOR TESTING SET

IV. ANALYSES

A. Best Model?

In order to decide which classification model worked the best, we decided to compare the F1-scores between all the models using the testing set and see which one gave the highest values. The reason for choosing F1-scores and not precision or recall is because the F1-scores display a weighted average of precision and recall.

Looking at the tables we obtained, we concluded that the Random Forest Model gave the highest F1-scores across all 3 classes compared to the other two models, hence this is the best model. Even though we observed that the training set for the Random Forest Model showed a perfect score of 1 and we suspected some overfitting, it is considered normal for there to be some higher accuracy and better performance in training sets compared to testing sets, and it is acceptable that the difference is not that significant.

According to the model feature importances, the top 3 features in our dataset that influence the classes are ALTV, ASTV, and Mean.

B. Hypothesis Tests

As we mentioned before, each of the 3 features were found to have unequal variances among the different class labels with p values < 0.05. Hence, we decided to use the Welch_ANOVA and pairwise_gameshowell methods from Pingouin.

Since we had 3 features, we came up with 3 hypothesis questions as well as hypothesis questions for each pair of classes. This gave us a total of 12 hypothesis questions:

1. Is there a difference in the mean value of the percentage of time with abnormal long term variability (ALTV) between classes?

2. For each pair of classes: is there a difference mean value of the percentage of time with abnormal long term variability (ALTV) between the two classes?

a) H0: Mean(Normal) = Mean(Suspect), HA: Mean(Normal) $\neq$ Mean(Suspect)

b) H0: Mean(Normal) = Mean(Pathological), HA: Mean(Normal) $\neq$ Mean(Pathological)

c) H0: Mean(Suspect) = Mean(Pathological), HA: Mean(Suspect) $\neq$ Mean(Pathological)

3. Is there a difference in the mean value of the percentage of time with abnormal short term variability (ASTV) between classes?

4. For each pair of classes: is there a difference mean value of the percentage of time with abnormal long term variability (ALTV) between the two classes?

d) H0: Mean(Normal) = Mean(Suspect), HA: Mean(Normal) $\neq$ Mean(Suspect)

e) H0: Mean(Normal) = Mean(Pathological), HA: Mean(Normal) $\neq$ Mean(Pathological)

f) H0: Mean(Suspect) = Mean(Pathological), HA: Mean(Suspect) $\neq$ Mean(Pathological)

5. Is there a difference in the mean value of the percentage of time with histogram mean (Mean) between classes?

6. For each pair of classes: is there a difference mean value of the percentage of time with abnormal long term variability (ALTV) between the two classes?

10: H0: Mean(Normal) = Mean(Suspect), HA: Mean(Normal) $\neq$ Mean(Suspect)

11: H0: Mean(Normal) = Mean(Pathological), HA: Mean(Normal) $\neq$ Mean(Pathological)

12: H0: Mean(Suspect) = Mean(Pathological), HA: Mean(Suspect) $\neq$ Mean(Pathological)

C. ANOVA and Games_Howell PostHoc Test

For each hypothesis test involving a single feature between classes, we performed an ANOVA test and obtained the p-value. For each pair of classes, we used the Games_Howell PostHoc Test, since the variances for each feature are unequal to each other. The p-values for each hypothesis test is given below, and using this value we were able to either reject or accept the null hypothesis at 95% confidence.

1) 1.631148e-130 = reject the null and accept the alternate
2a) 0.000000e+00 = reject the null and accept the alternate
2b) 2.518810e-10 = reject the null and accept the alternate
2c) 7.449421e-02 = reject the alternate and accept the null
3) 4.472368e-130 = reject the null and accept the alternate
4a) 0.000000 = reject the null and accept the alternate
4b) 0.000000 = reject the null and accept the alternate
4c) 0.102143 = reject the alternate and accept the null
5) 1.145660e-114 = reject the null and accept the alternate
6a) 0.0 = reject the null and accept the alternate
6b) 0.0 = reject the null and accept the alternate
6c) 0.0 = reject the null and accept the alternate

V. CONCLUSION

We tested 3 different classification models and found that the Random Forest Classifier with n_estimates = 100 to be the best one. Based on the model's highest weighted features, we did some tests to see if there was a statistical difference in the data between classes. We found that there was a statistically significant difference between each pair of classes for each ALTV, ASTV, and Mean, excepting ALTV for the classes ”Suspect” and ”Pathological”, as well as ASTV for the classes ”Suspect” and ”Pathological” (hence the null hypothesis was accepted and the alternate hypothesis was rejected due to a p-value being greater than 0.05). These features were not found to have a statistical difference in the means between them.

REFERENCES

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

[2] View of predicting fetal health using Cardiotocograms: A machine learning approach. (n.d.). Tensorgate Journals.

[3] UCI machine learning repository. (n.d.). UCI Machine Learning Repository.