Fetal Health Classification Using Cardiotocography Data

Vincent Tian (vinnyt@mit.edu)
Jennifer Chen (jennchen@mit.edu)
Tessie Xie (cc906@mit.edu)
Kevin Sheng (ks2237@mit.edu)

Fall 2023

Contents

1 Introduction

Child and maternal mortality are urgent global concerns. The UN targets ending preventable deaths in newborns and under-5 children by 2030. Cardiotocograms (CTGs) provide vital fetal health data using ultrasound pulses, allowing timely interventions. This cost-effective technology significantly reduces child and maternal mortality, especially in resource-constrained areas. The project aim is to classify fetal health to prevent child and maternal mortality.

2 Data Source

The dataset, sourced from Larxel on Kaggle, comprises 22 distinct features and encompasses a total of 2,126 measurements extracted from Cardiotocogram examinations. These features encapsulate critical parameters, including fetal health rate, fetal movements, uterine contractions, and various aggregate summary metrics. Three expert obstetricians meticulously classified the records into three distinct categories: *Normal*, *Suspect* and *Pathological*. The dependent variable, denoted as 'fetal_health,' encompasses multiple classes, each assigned a numeric value, with *Normal* labeled as 1, *Suspect* as 2, and *Pathological* as 3. Notably, the distribution of these labels is imbalanced 4, with *Normal* accounting for 78%, *Suspect* for 14%, and *Pathological* for 8% of the dataset.

- Link to Data Set

It is important to highlight the presence of columns labeled 'histogram' within the dataset. These columns encapsulate summary statistics derived from histogram charts of heart rate distribution for each infant included in the dataset.

3 Exploratory Data Analysis

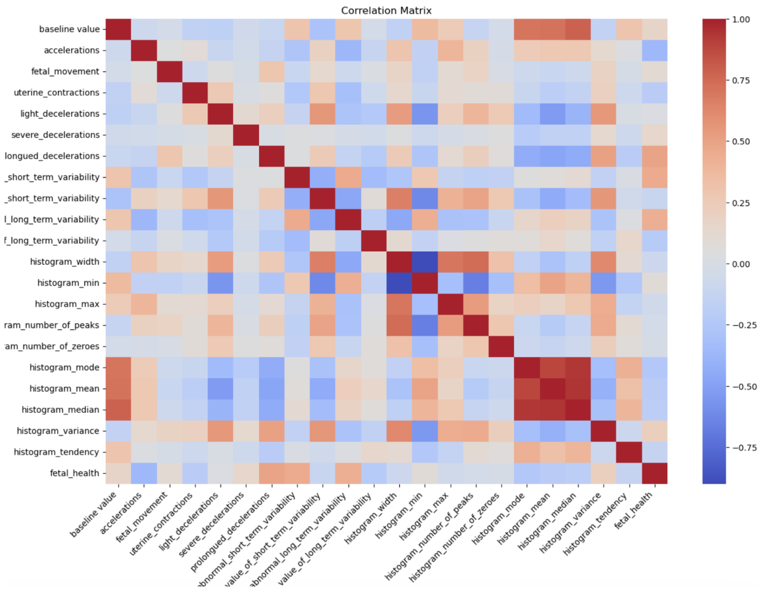

Figure 1: Correlation between features

In the context of the histogram-related features, a notable high correlation is observed among 'mode,' 'mean,' and 'median.' To mitigate the effects of collinearity in our models, we have opted to retain only the 'median' as a representative feature. Additionally, it's worth mentioning a significant negative correlation between 'min' and 'width.' Given the intuitiveness and relevance of 'min' in our analysis, we have decided to retain it while omitting 'width' from consideration.

Furthermore, an important consideration is the substantial class label imbalance within the dataset. To address this, we have employed a stratified sampling technique for the purpose of train/test data splitting.

4 Data Pre-Processing

The dataset exhibits a complete absence of missing values, and we have meticulously removed 13 duplicate rows. To maintain consistency in the distribution of class labels between the training and testing sets, we have employed stratified sampling. Specifically, 80% of the data is allocated for training, with the remaining 20% reserved for testing. This process results in the creation of two distinct train/test datasets:

- original dataset: stratified sampling for train test split
- 1. use Synthetic Minority Over-Sampling Technique (SMOTE) to over-sample the dataset
  2. use stratified sampling to do train test split

5 Methods

5.1 Evaluation Metrics

To reduce instances where the model overestimates fetal health, it's crucial to establish a relevant evaluation metric for our model assessments. Given an idealized scenario with ample resources, we aim to minimize the false negative rate (FNR), which measures the ratio of falsely predicting susceptible or pathological infants as *Normal*. For simplicity, we have assigned equal weights to the false negative rates for both *Suspect* and *Pathological* categories.

The custom false negative formula is given by 2:

$$\text{Custom FN} = \frac{1}{2}\left(\frac{sn}{sn+ss+sp}\right) + \frac{1}{2}\left(\frac{pn+ps}{pn+ps+pp}\right)$$

In addition, the following were also used for further comparison:

- Accuracy
- Suspect FN
- Pathological FN

5.2 Modelling

After conducting data pre-processing, each model will undergo training through a 5-fold cross-validation procedure applied to both the stratified and SMOTE sampling datasets. The criterion for model selection will be the Custom FN metric, and the chosen model will be utilized for predictions on the test set.

In this study, a diverse set of models will be employed, encompassing Logistic Regression, K-Nearest Neighbors (K-NN), Classification and Regression Trees (CART), Evolutionary Learning of Globally Optimal Trees (EVTree), Random Forest (RF), XGBoost, and Neural Networks. The expectation is that XGBoost will emerge as the top-performing model, while comparatively simpler models like Logistic Regression or K-NN are anticipated to yield inferior results. An ideal scenario would involve models such as CART or EVTree achieving superior outcomes, as they offer greater interpretability, aligning with the study's goal of interpretable model performance.

Additionally, an ensemble model comprising various "weak" models, including Logistic Regression, K-Nearest Neighbors (K-NN), Classification and Regression Trees (CART), and Neural Networks with a single hidden layer, will be employed. The underlying concept is to leverage their collective predictions and implement a hard voting mechanism to mitigate errors present in each individual model. In cases where a tie occurs among the predictions of these four models, a conservative approach will be adopted, selecting the least favorable prediction as the final output. For instance, if *Normal* and *Suspect* are in a tie, the prediction *Suspect* will be assigned.

6 Results

Based on our preliminary analysis, the utilization of both stratified and SMOTE sampling techniques across a range of models has demonstrated promising outcomes. Notably, each model that incorporated

SMOTE sampling exhibited significant enhancements when compared to those relying solely on stratified sampling, demonstrating an impressive average 100% increase in our Custom FN metric. Although certain models may have experienced a decrease in overall accuracy with the adoption of SMOTE, it is important to underscore that our project places a higher emphasis on mitigating false negatives rather than solely pursuing maximal accuracy.

In the case of the stratified sampling dataset, XGBoost emerged as the top-performing model, consistently achieving the highest scores across metrics such as Accuracy, Suspect FN, Pathological FN, and Custom FN. The ensemble model secured the second-best performance, while Logistic Regression consistently exhibited the least favorable results across the majority of metrics.

The dynamics shifted when considering the SMOTE dataset, as the ensemble model exhibited a remarkable 48% improvement in Custom FN over the XGBoost model. Surprisingly, the K-Nearest Neighbors (K-NN) model also delivered impressive results, surpassing XGBoost by 25% in Custom FN. Interestingly, despite Logistic Regression's lower overall Custom FN score, it excelled in identifying suspect patients, securing the second position just behind K-NN and the ensemble model. On the other hand, Neural Networks underperformed, possibly due to the limitations of a relatively small dataset. Moreover, Random Forests demonstrated proficiency in detecting Pathological FN cases but faced challenges in identifying Suspect FN cases.

Table 1: Test results across models for stratified and SMOTE sampling.

	Stratified Sampling				SMOTE			
Method	Accuracy	Suspect FN	Pathological FN	Custom FN	Accuracy	Suspect FN	Pathological FN	Custom FN
Logistic Regression	0.903	0.224	0.229	0.226	0.884	0.043	0.100	0.071
K-NN	0.905	0.276	0.086	0.181	0.877	0.034	0.086	0.060
CART	0.920	0.259	0.057	0.158	0.868	0.103	0.057	0.089
EVTree	0.922	0.310	0.171	0.241	0.879	0.121	0.057	0.089
RF	0.922	0.293	0.086	0.189	0.931	0.155	0.057	0.106
XGBoost	0.936	0.138	0.057	0.098	0.943	0.103	0.057	0.080
NN - H1	0.910	0.190	0.200	0.195	0.896	0.138	0.171	0.155
NN - H2	0.901	0.190	0.143	0.166	0.908	0.121	0.086	0.103
Ensemble	0.929	0.155	0.057	0.106	0.863	0.034	0.057	0.046

Our initial hypothesis, suggesting that the ensemble model composed of weak learners would outperform all other models, was substantiated by our findings. On the SMOTE testing set, the ensemble model demonstrated superior performance in terms of Suspect FN and Pathological FN, ultimately leading to the highest Custom FN score. However, it's noteworthy that the accuracy of the ensemble model was unexpectedly the lowest among the models. A more detailed analysis can shed light on this discrepancy, particularly when considering the confusion matrix.

7 Interpretation

7.1 Confusion Matrix

In the analysis of model performance, it is noteworthy to focus on the 'Suspect' and 'Pathological' categories within the confusion matrix. The Logistic Regression model exhibited a higher rate of misclassification in the *Suspect* category, erroneously classifying an additional infant as *Normal* compared to the Ensemble model. Conversely, in the *Pathological* category, the Ensemble model demonstrated superior performance by making only two misclassifications, assigning *Suspect* labels to *Pathological* cases. In contrast, the Logistic Regression model displayed a less robust performance, misclassifying two *Pathological* infants as *Normal*.

As reiterated, our paramount objective is to minimize false negatives as they carry substantial clinical significance in the context of fetal health classification. In this regard, the performance of the Ensemble model warrants special attention. Notably, the Ensemble model succeeded in significantly reducing the false negatives for *Pathological* cases, with only two instances where *Pathological* babies were predicted as *Suspect*. In such cases, it is anticipated that vigilant parents would engage in further medical assessments to confirm the true health status of their infants, potentially averting more severe conditions.

However, it remains a pertinent concern that two *Suspect* babies were incorrectly predicted as *Normal* by the Ensemble model. Such misclassifications may lead parents to assume that everything is in order with their infants, potentially delaying necessary medical interventions if further conditions were to manifest. This underscores the delicate balance between minimizing false negatives and avoiding misclassification, emphasizing the intricate nature of fetal health classification.

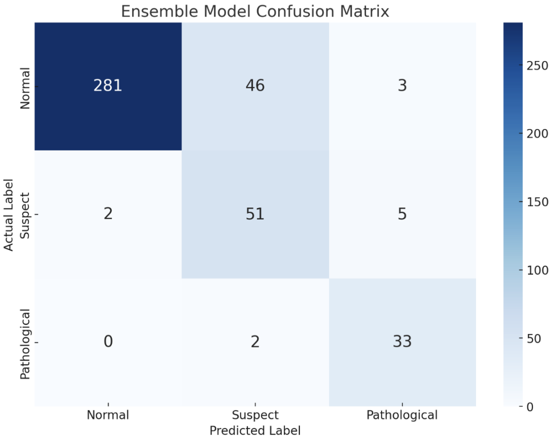

Figure 2: Ensemble model on test set

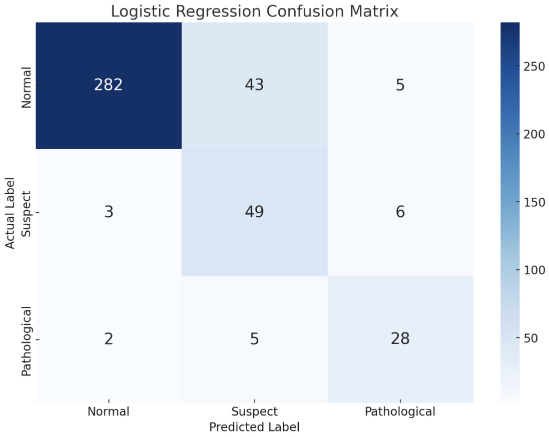

Figure 3: Logistic Regression model on test set

7.2 SHAP

For further interpretation, we can refer to the SHAP summary plot 5. Here are some key insights.

Accelerations and Mean Value of Short Term Variability: These features exert substantial influence in distinguishing *Normal* and *Suspect* fetal health classifications, yet their impact on *Pathological* cases is notably limited.

Abnormal Short Term Variability: Pathological infants exhibit a notably greater susceptibility to influencing factors compared to their *Normal* or *Suspect* counterparts.

8 Conclusion

In conclusion, our project on ”Fetal Health Classification Using Cardiotocography Data” has successfully demonstrated the substantial potential of machine learning in improving fetal health assessment. The application of diverse algorithms, particularly the effective use of XGBoost and ensemble models, has shown promise in accurately categorizing fetal health states from CTG data. The innovative approach to addressing data imbalance, notably through the SMOTE technique, significantly enhanced the predictive accuracy while minimizing false negatives.

9 Next Steps

The next crucial steps for this project involve acknowledging the real-world constraints of hospital budgets. Our current models have overlooked the potential misclassification of *Normal* babies as either *Suspect* or *Pathological*, which could result in unnecessary costs for hospitals. To address this issue, it is imperative to incorporate cost considerations into our evaluation metrics.

However, the decision to implement these models ultimately rests with the hospitals. It is essential to recognize that if the current ensemble model is put into practice, there may be instances where babies in more critical conditions are mistakenly classified as either *Normal* or *Suspect*. In such cases, careful attention should be paid to interpreting these classifications, considering the relatively low probability of misclassification. This thoughtful approach is vital to ensure the successful and responsible deployment of the model in a clinical setting.

References

[1] Andrew Mvd, *Fetal Health Classification Dataset*, Kaggle, [Online; accessed 11-2023], Available at: `https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification`.

[2] Wikipedia, *Ensemble Learning*, [Online; accessed 11-2023], Available at: `https://en.wikipedia.org/wiki/Ensemble_learning`.

[3] Sophia Yang, *Multiclass Logistic Regression from Scratch*, Towards Data Science, [Online; accessed 11-2023], Available at: `https://towardsdatascience.com/multiclass-logistic-regression-from-scratch-9cc0007da372`.

[4] Jason Brownlee, *SMOTE Oversampling for Imbalanced Classification*, Machine Learning Mastery, [Online; accessed 11-2023], Available at: `https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/`.

[5] imbalanced-learn Documentation, *SMOTE: Synthetic Minority Over-sampling Technique*, [Online; accessed 11-2023], Available at: `https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html`.

Appendix

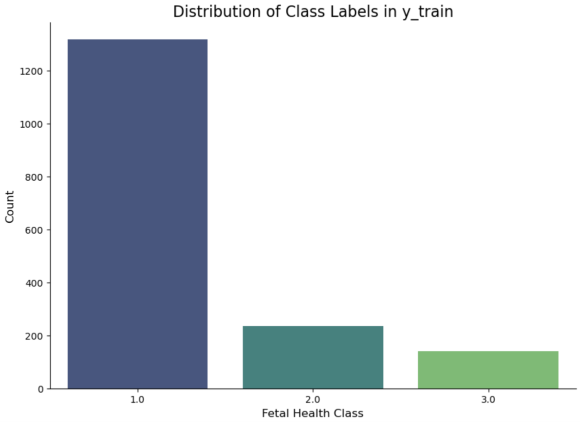

Figure 4: Distribution of class labels

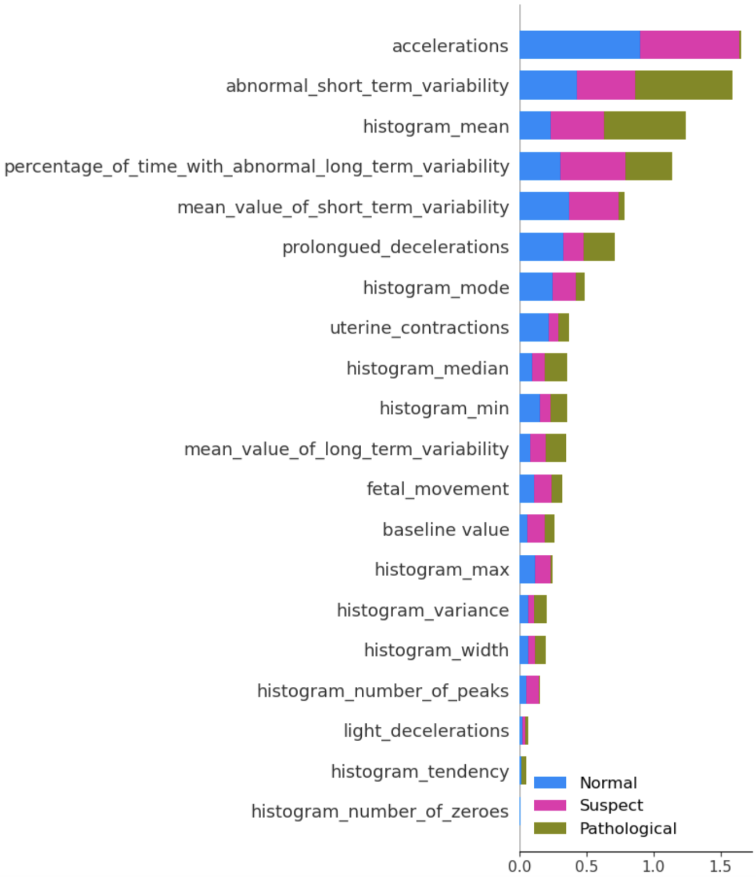

Figure 5: SHAP summary plot from XGBoost

Table 2: Confusion Matrix

Actual	Predicted		
	Normal	Suspect	Pathological
Normal	nn	ns	np
Suspect	sn	ss	sp
Pathological	pn	ps	pp