Automated Fetal Health Classification from Cardiotocography Using Logistic Regression, Random Forest, and Lasso

Sarvar Khamidov (sk10583), Sofia Jenssen (sgj3544)
MLPH
Spring 2025

Introduction

Child and maternal mortality are critical global health challenges which are Sustainable Development Goals (SDGs) of the United Nation. SDG aims to reduce under 5 mortality to below 25 per 1000 live births and ending preventable newborn deaths by 2030. Despite the progress, an estimated 5.2 million children under five and nearly 300,000 mothers still die annually from largely preventable causes, with over 90 percent of these deaths occurring in low-resource settings. Early identification of fetal disorder is essential to guide timely clinical interventions that can prevent stillbirths and neonatal complications, thereby helping to meet SDG targets and improve survival outcomes for mothers and infants.

Cardiotocography (CTG) is a cost effective technology widely used to monitor fetal heart and rate and uterine activity during labor. However, CTG reading and interpretation is highly dependent on operator and prone to observer variability which can delay or misdirect critical decision making. Automated fetal monitoring and classification systems help to standardize CTG analysis by extracting quantitative features, like quantitative and qualitative features and applying ML algorithms to detect fetal health anomalies.

In this study, we implement and compare three supervised machine learning approaches such as multinomial logistic regression with 10 fold cross validation, random forest, and lasso penalized logistic regression. Our results show that the random forest model outperforms both baseline logistic regression and the sparse lasso model achieving the highest overall accuracy, Cohen's k and balanced accuracy for the critical Pathological class. Even though logistic regression and lasso offer interpretability and simplicity, random forest was able to capture nonlinear interaction which made it best suited for this problem.

Related Work

There has been much previous work done to automate the fetal health classification focused on using cardiotocography data. However, many of them didn't use any explicit regularization or feature selection mechanisms. For example Ayres-de Campos et al. (2001) developed the SisPorto 2.0 software to extract a comprehensive set of CTG features and applied simple neural networks and support-vector machines, achieving promising accuracy on first-stage labor recordings. However, because all 21 metrics were entered without penalization, models often overfit to noisy or correlated inputs, and no external validation was reported (Ayres-de Campos et al., 2001; Chudáček et al., 2014).

More recent studies have implemented advanced to ensemble and deep-learning approaches but still face critical drawbacks. Afridi, Khan, and Ahmed (2019) compared logistic regression, random forest, and SMOTE-augmented ensembles, reporting high overall accuracy but persistently low recall for the "suspect" class, precisely the group needing early clinical intervention. Islam, Rahman, and Kabir (2022) evaluated similar classifiers but only on internal cross-validation folds, leaving generalizability to new hospitals untested. Meanwhile, Salini, Nisha, and Padma (2024) combined SMOTE with LightGBM to push accuracy above 99 %, yet produced a "black-box" model with little insight into which CTG patterns drive predictions, making it difficult for clinical trust and adoption.

These gaps create the need for a hybrid strategy that balances interpretability and predictive power. Lasso logistic regression can yield sparse, clinically meaningful feature subsets, while random forests capture nonlinear interactions, and standard logistic regression provides a transparent baseline.

Methods

We frame the fetal health classification as a supervised, multi-class prediction problem where the goal is to assign each cardiotocogram data to one of three categories: Normal, Suspect or Pathological. The response variable Y takes those values and predictor variables include 21 continuous features derived from CTG signal analysis.

To evaluate unbiased models performance we split the dataset into training and testing. We preserve the three-class prevalence in both training and test sets to avoid misleading performance estimates on under-represented “Suspect” or “Pathological” groups.

We fit a multinomial logistic regression model with 10-fold cross-validation to tune model complexity. An L2 penalty helps against overfitting when decision boundaries are roughly linear in the feature space. This model provides a performance baseline and interpretable coefficients.

We also trained a random forest model to capture nonlinear interactions and complex dependencies. We chose an ensemble of 500 classification trees, each grown on a bootstrap sample of the training set. That large number of trees guarantees stable class probability estimates. We also performed a feature importance score from this model for future variable selection.

Lastly, we fit multinomial logistic regression with L1 penalty using 10 fold cross validation to choose alpha. This approach improves the interpretability by shrinking many coefficients to zero. We assume that only a small number of variables from available carry the majority of the signal.

All three models were evaluated using overall accuracy, Cohen’s k, per-class balanced accuracy, confusion matrices and multi class ROC-AUC. We also looked at Gini Coefficients for each model to quantify separability.

Data and Experiment Setup

We used the Fetal Health Classification Dataset, published in the UCI Machine Learning Repository available via Kaggle. The dataset consists of 2,126 observations, each a CTG session with a pregnant person. To preprocess the data, we factored the classification categories Normal, suspect, or pathological. There were no missing values in the dataset. The data from each CTG was classified by a doctor that labeled each observation into one of the three fetal health categories (Normal, Suspect, or Pathological), which was our outcome variable. The majority of cases were normal (77.9%), 13.9% were suspect, and 9.3% were pathological cases (Figure 1). An 80/20 stratified split was employed and all numeric predictors were standardized (z-scores) using centering and scaling to ensure equal contribution across features.

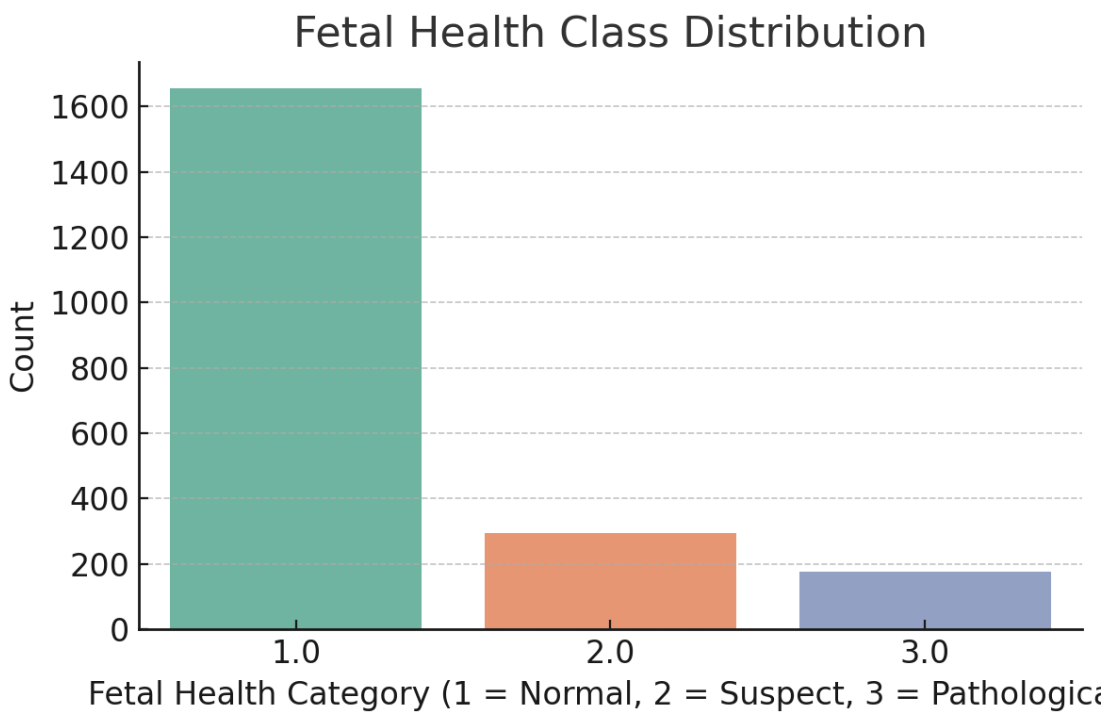

Figure 1. Fetal Health Class Distribution bar chart

The 21 physiological features capture various aspects of fetal heart rate (FHR) and uterine activity, including variability measures, contraction indicators, and statistical properties of the FHR signal distribution (histogram descriptors). Descriptive statistics showed that the baseline fetal heart rate has a mean of 133.16 bpm, ranging from 106 to 160. Short term variability had a mean of 1.33 and SD of 0.88 and long-term variability shows more spread with a mean of 8.19 and a maximum of 50.7. Histogram-based descriptors like histogram mean, width, and spread reflected the FHR. Histogram width and variance show wide variability across cases while the histogram mean has a value of 134.6 bpm, close to the FHR mean. Descriptives also showed that the dataset contains features that are very zero-inflated, skewed but may be informative when non-zero. For example, they might be event-based like Accelerations, uterine contractions and light decelerations, which all have a mean of 0.00 and standard deviation of 0.00, indicating that they were less common during the CTG. This was particularly true for severe and prolonged decelerations, of which the vast majority of observations in the dataset had values of zero.

We then conducted a correlation matrix to identify collinearity in features (correlation higher than 0.85). Histogram features like median, mode, minimum, maximum, number of peaks and number of zeroes exhibited clusters of strong positive correlations. These variables were removed before model fitting to avoid inflating model variance and enhancing the interpretability of our results.

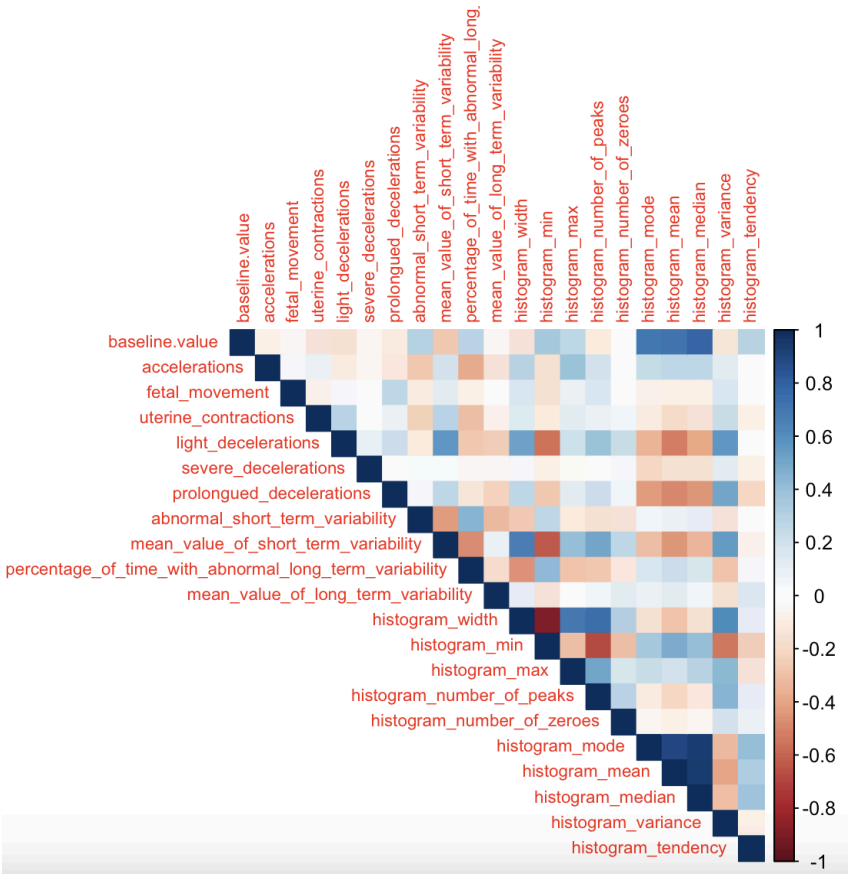

Figure 2. Correlation matrix of 21 original features.

Results

This section covers the performance of the three supervised learning models Logistic Regression with 10-fold Cross-Validation, Random Forest, and Lasso Logistic Regression in classifying fetal health into Normal, Suspect or Pathological categories using the features derived from CTG signals. All models were trained on 80% of the dataset using a stratified split and evaluated on the remaining 20% test set. Models were compared based on classification accuracy Cohen's kappa, balanced accuracy.

The logistic regression model was implemented using a multinomial framework with 10-fold cross-validation to select the optimal regularization parameter. Three decay values were tested: 0, 0.0001, and 0.1. The best performance was observed at decay = 0.1, which introduces a modest penalty on the magnitude of model coefficients, effectively regularizing the

model without excessively shrinking useful predictors. With this configuration, the logistic regression model achieved an accuracy of 90.6% and a Cohen's kappa of 0.747, indicating substantial classification agreement. The confusion matrix revealed strong performance on Normal cases (312/330, or 94.3% sensitivity), and good performance for Pathological cases (29/35, or 82.9% sensitivity), but relatively weaker precision for Suspect cases (44/59 correctly classified, with a positive predictive value of 68.8%). The balanced accuracy for each class was 88.6% (Normal), 84.6% (Suspect), and 91.0% (Pathological). This model demonstrated solid baseline performance and maintained interpretability.

The Random Forest model was trained using 500 decision trees, with 3 variables randomly sampled at each split (mtry = 3), following the standard method of using the square root of the total number of predictors. This configuration was selected to balance performance with computational efficiency. The number of trees was not further tuned because, in practice, model accuracy tends to stabilize well before 500 trees, while additional trees only marginally improve results but increase training time. The out-of-bag (OOB) error, a built-in estimate of test error in Random Forests, was reported at 5.29%, indicating a high-performing model with low generalization error. Evaluation on the held-out test set revealed an accuracy of 94.6%, and a Cohen's kappa of 0.851, reflecting excellent agreement with ground truth labels. The confusion matrix showed that 97% of Normal cases, 81.4% of Suspect cases, and 94.3% of Pathological cases were correctly identified. Importantly, the balanced accuracy for the Pathological class was 96.9%, surpassing the clinical benchmark of 85% and suggesting that the model was particularly effective at identifying the most high-risk fetal health cases without sacrificing performance in the majority class. The Random Forest's feature importance analysis also aligned well with known physiological risk markers.

To interpret how the Random Forest model made its predictions, we examined variable importance using the mean decrease in Gini impurity. The most influential feature was abnormal short-term variability (importance = 0.134), which reflects the proportion of time the fetal heart rate displayed irregular beat-to-beat fluctuations. This is a well-established marker of fetal distress, often linked to hypoxia. The second most important feature was the mean value of short-term variability (importance = 0.112), reinforcing the role of high-frequency variability in identifying at-risk fetuses. Histogram mean ranked third (importance = 0.089), capturing the central tendency of the heart rate distribution and helping distinguish between normal and elevated baselines.

Other key features included prolonged decelerations (importance = 0.076), which reflect sustained drops in fetal heart rate and are associated with severe distress, and percentage of time with abnormal long-term variability (importance = 0.071), a measure of slow baseline shifts. Baseline fetal heart rate (importance = 0.068), histogram width (importance = 0.063), and histogram variance (importance = 0.061) also contributed meaningfully, capturing the overall spread and rhythm instability of the fetal heart signal. Overall, the model prioritized features that improved classification performance.

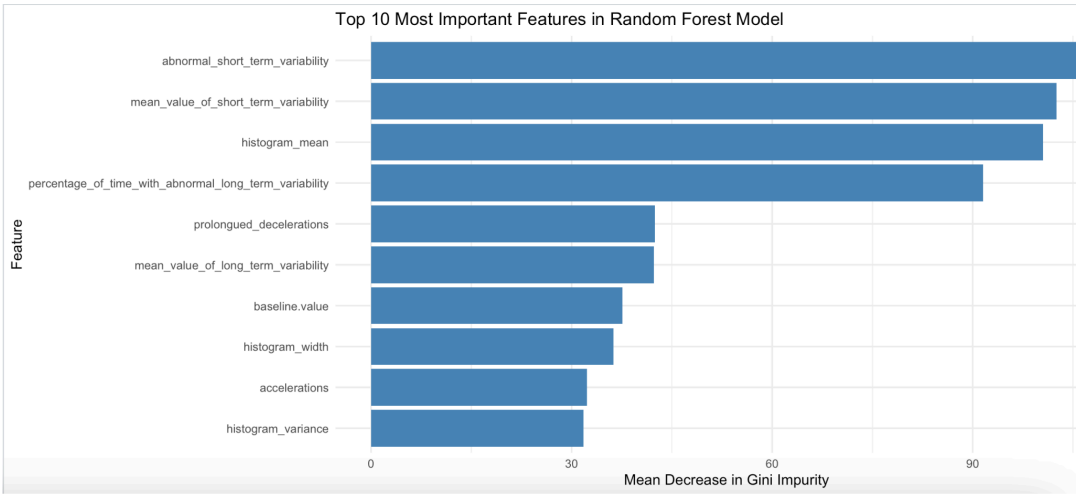

Figure 3. Important features in the Random Forest Model according to mean decrease in Gini Impurity.

The Lasso regression model was trained using multinomial logistic regression with L1 regularization (alpha = 1), which induces sparsity by shrinking some coefficients to zero. The optimal λ (lambda) was selected via 10-fold cross-validation, with the value minimizing classification error being λ = 0.00214 (Figure 4). This regularization strength was chosen to balance model complexity with generalization ability. On the test set, the model achieved an accuracy of 89.9% and a Cohen's kappa of 0.728, slightly lower than both logistic regression and Random Forest. The confusion matrix showed that 311 out of 330 Normal cases were correctly classified (94.0% sensitivity), but only 27 of 35 Pathological cases were detected (77.1% sensitivity). While Lasso achieved a balanced accuracy of 87.8% for Pathological cases—just above the minimum threshold for clinical utility—it was notably less sensitive than Random Forest in detecting these high-risk outcomes.

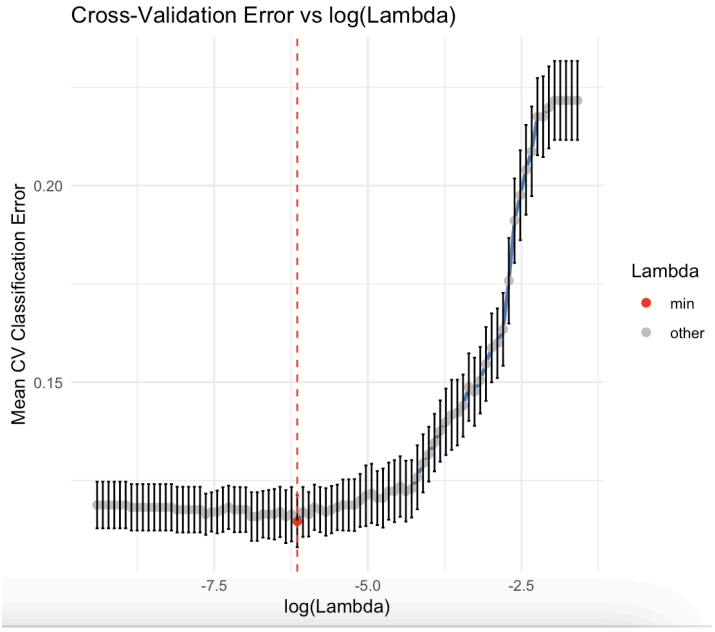

Figure 4. Cross-Validation Error vs Lambda

To better understand how the Lasso logistic regression made the predictions, we examined the features that had non-zero coefficients since Lasso performed automatic feature selection for each health classification (Figure 5). For Normal cases, it positively weighted accelerations, which are brief increases in fetal heart rate typically associated with healthy autonomic responsiveness. In contrast, it assigned negative coefficients to prolonged decelerations and abnormal short-term variability, both of which are warning signs of fetal distress. This aligns with clinical expectations: the absence of decelerations and normal variability are reassuring indicators. For Suspect cases, the model identified histogram mean as a key predictor. A higher histogram mean (indicating higher FHR mean) may reflect mild distress or compensatory response. For Pathological cases, the only retained feature was abnormal short-term variability, which received a positive coefficient.

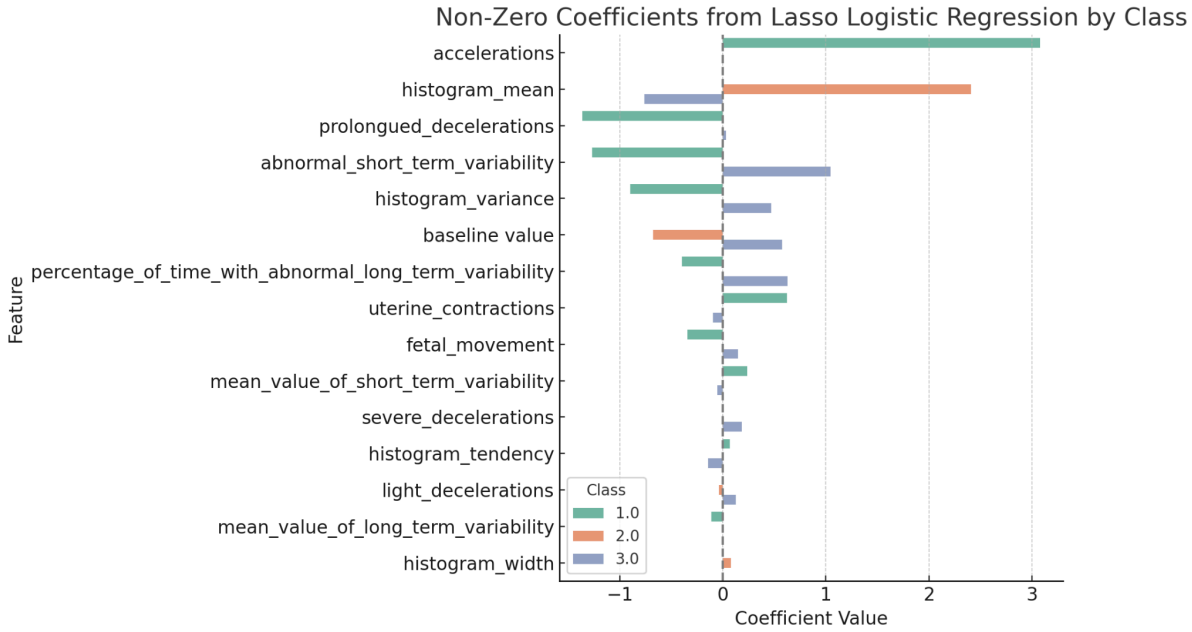

Figure 5. Chart showing the non-zero coefficients selected by the Lasso logistic regression model grouped by fetal health class. The direction and magnitude of each coefficient reflects how strongly the feature influences classification for the outcome.

Model Comparison

When comparing the three models, there is a revealed tradeoff between predictive accuracy and model interpretability. Random Forest consistently outperformed the other models across all evaluation metrics, achieving the highest overall accuracy (94.6%), strongest agreement beyond chance (Cohen's kappa = 0.851), and best balanced accuracy in the Pathological class (96.9%). Its ability to capture nonlinear relationships and complex interactions among variables made if most effective in handling the patterns present in CTG data. In contrast, Logistic Regression was less complex and offered a simpler, more interpretable model while still achieving robust accuracy (90.6%) and good class-level

performance, especially for Normal and Pathological cases. Lasso Regression, while slightly lower in overall performance (89.9% accuracy), provided the most compact and interpretable model by reducing the feature set to only the most essential variables. Ultimately, the Random Forest model is best suited for medical applications requiring high sensitivity and balanced class performance.

Interestingly, the Suspect class consistently showed the lowest predictive precision across all models. This trend may reflect ambiguity in the Suspect label itself—clinically and statistically, it likely represents a heterogeneous group with overlapping features from both Normal and Pathological categories. As a result, even high-performing models like Random Forest struggled to achieve precision as high as it was in other classes (PPV = 84.2%) for this class. Misclassifications in this middle category were more frequent and should be interpreted in light of its inherent diagnostic ambiguity. In addition, the models also revealed interesting trends when looking at feature importance analysis. Abnormal short-term variability and deceleration patterns were consistently prioritized in Random Forest and Lasso models, showing a pattern in what should be considered when determining high-risk pregnancies.

Results: Model Comparison

Model	Accuracy	Kappa	Normal (BA)	Suspect (BA)	Pathological (BA)
Logistic Regression (CV)	0.906	0.75	0.886	0.846	0.91
Random Forest	0.946	0.85	0.921	0.895	0.969
Lasso Regression (CV)	0.899	0.73	0.879	0.846	0.878

Discussion

Three models showed distinct strength and trade offs. Random forest delivered the best overall performance, with 94,6 % accuracy, k = 0,85 and a 96.9 % balanced accuracy for Pathological cases. This confirms that nonlinear interactions among CTG features carry essential predictive power. Logistic regression (90.6 % accuracy, k = 0.75) demonstrated that a simple linear decision boundary already captures most of the class structure and Lasso

regression (89.9 % accuracy, k = 0.73) revealed that fewer than ten predictors is enough for almost high performance.

Some of the limitations include relance a single center CTG dataset. Performance may drop on data from different hospitals or devices without calibration. We also assumed equal misclassification costs across classes, whereas in practice mislabeling a pathological case may carry far higher clinical consequences than a false-positive normal call.

Future work should validate these classifiers on external cohorts and explore cost-sensitive learning frameworks that weight pathological detections more heavily. Integrating maternal and clinical covariates (e.g., gestational age or comorbidities) could further improve the classification.

Code Availability

https://github.com/khsarvar/MLPH_Final

Reference

Ayres-de Campos, D., Bernardes, J., Garrido, A., Sa, J., & Pereira-Leite, L. (2001). SisPorto 2.0: A program for automated analysis of cardiotocograms. *Journal of Maternal-Fetal and Neonatal Medicine*, *9*(5), 311–318.

Chudáček, V., Spilka, J., Chládek, J., Huptych, M., & Burša, M. (2014). CTU-UHB intrapartum cardiotocography database: A comprehensive open-access resource. *Physiological Measurement*, *35*(5), 887–900.

Afridi, S., Khan, A., & Ahmed, H. (2019). Cardiotocography data classification using ensemble methods. *Biomedical Signal Processing and Control*, *49*, 48–55.

Islam, M., Rahman, M., & Kabir, M. (2022). Comparative evaluation of machine learning classifiers for fetal health prediction using CTG signals. *Health Informatics Journal*, *28*(3), Article 14604582221084554.

Salini, K., Nisha, R., & Padma, K. (2024). Enhanced CTG classification using SMOTE and LightGBM. *IEEE Access*, *12*, 34567–34575.

Teammates Contributions

Both teammates contributed to the project equally across all phases of the project, from initial design through analysis, interpretation, and report preparation.

Sarvar Khamidov:

- Conducted the related-work survey, drafted the Methods section, and justified modeling choices and simplifying assumptions.
- Co-developed the study objectives, selected the CTG features for analysis, and defined the classification framework.
- Implemented and tuned the random forest classifier (ntree=500, mtry=3), computed OOB error, and generated the variable-importance plots.

Sofia Jenssen:

- Co-developed the study objectives, selected the CTG features for analysis, and defined the classification framework.
- Performed the stratified train test split, implemented and tuned the multinomial logistic regression and Lasso models, and generated performance metrics.
- Co-drafted the Results and Data and experiment sections, formatted tables and figures, and performed critical revisions to ensure clarity and coherence.