



Comparing Operative Delivery Risk Using Explainable Artificial Intelligence (XAI) and Maternal and Fetal Clinical Information

C. Ramani Vijay Chelvi¹ and B. Shunmugapriya ²●.

^{1,2}Department of Computer Science and Engineering, National Engineering College, Kovilpatti, India.

● These authors contributed equally to this work

DOI: <https://www.doi.org/10.29194/NJES.29030542>

Received: March 21, 2026 Revised: May 15, 2026 Accepted: May 23, 2026 Published: Sep. 20, 2026

Abstract

Predicting the risk of operative delivery, especially Cesarean section, is important for improving outcomes for both mothers and newborns. In this study, a machine-learning-based framework is developed using explainable artificial intelligence techniques across three clinical datasets: fetal cardiotocography, maternal health parameters, and pregnancy outcome data.

Three classification models—Logistic Regression, Decision Tree, and Random Forest—were evaluated to identify the most effective approach. Among these, the Random Forest model achieved the best performance, with an F1 Score of 0.88 and an Area Under the Curve of 0.99 on the fetal health dataset.

To enhance interpretability in a clinical context, a dual-layer explanation strategy was adopted. SHapley Additive exPlanations was used to analyze global feature importance, while Local Interpretable Model-agnostic Explanations was applied to explain individual predictions. The results indicate that abnormal short-term variability is a key factor influencing operative delivery risk. The explanations generated by both methods were consistent with established clinical understanding, making the model both accurate and interpretable.

Keywords: *Machine Learning, Explainable Artificial Intelligence, Random Forest Classifier, Ensemble Learning, Predictive Analytics*

Corresponding author: Provide the corresponding author information and publisher here. E-mail address: ramanivijic8@gmail.com

1. Introduction

The choice of delivery method is a key decision in modern obstetrics, especially when Cesarean section or operative delivery is involved. While the procedure is vital for instances where either the fetus is under distress or the pregnant woman is under critical circumstances, the rising number of Cesarean sections performed across the world today has raised considerable medical and health concerns. Unnecessary surgeries can increase risk, recovery time, and healthcare costs. At the same time, poor risk assessment may lead to serious complications.

Fetal condition is commonly monitored using cardiotocography (CTG), which is the observation of fetal cardiac rhythm and contractions of the uterus. In fact, cardiotocography (CTG) is considered a normal clinical practice, but its interpretation can vary depending on the clinician's experience. Therefore, there is a dramatic variance in results among various observers and even among observations by the same observer on various occasions. In contrast, maternal condition, as measured by blood pressure, blood sugar, age, and Body Mass Index, is usually evaluated independently. Recent advances in machine learning, or artificial intelligence, have introduced a fresh approach to enhance obstetric clinical decision-support systems. Unlike regular statistical models with linear

relationships, machine learning models can efficiently interpret complex, nonlinear patterns in health phenomena in the medical domain. In previous research, Machine learning models have been shown to successfully predict adverse outcomes in pregnancy, as well as identify early indicators for fetal distress based on cardiotocography (CTG) signals and maternal clinical indicators. Nonetheless, despite the success in performance results, there remain two key shortcomings in existing research work. Second, existing research tends to explore a single dataset or quantify a small set of factors, with unclear value to overall fetal systematic comparison of classical models, such as Logistic Regression, with more advanced non-linear methods, including Decision Trees and ensemble techniques like Random Forest, is still needed under a unified prediction goal. This comparative analysis is crucial to measure the real benefits of complex models in practical obstetric applications.

Another limitation is that the lack of interpretability in highly accurate models is a hindrance to the usage of Machine learning models in a medical context. It is known that ensemble models, as well as deep learning models, have received criticism for being "black-box" models. The models are very accurate in their predictions, but they cannot explain the reasons behind these predictions. In a medical emergency, there is a need for explanations

to be provided for models in critical areas like healthcare for establishing accountability in these models.

Explainable Artificial Intelligence emerged as a potential solution to this problem by allowing the global model properties or specific predictions to be understood. Contrary to models being accurate while opaque, the goal of Explainable Artificial Intelligence algorithms is to provide explanations of this specific accuracy or decision of the model, rather than being solely accuracy-driven. They provide population-level feature importance as well as patient-level explanations through Explainable Artificial Intelligence. To address these challenges, this study presents a machine learning-based approach that can predict the risk of operative delivery. This study involves the evaluation of diverse machine learning approaches for predicting operative risk based on three diverse data sources that incorporate fetal cardiotocography, the mother's health vital signs, and the overall outcome of the pregnancy.

This study aims to balance predictive performance with interpretability and the ability for its results to be interpreted by providing Shapley Additive Explanations and Local Interpretable Model-Agnostic Explanations for the diverse approaches explored in the research work.

The main contributions of this study are summarized as follows:

- (1) A unified comparative analysis of three obstetric-related datasets, including fetal cardiotocography, maternal clinical vital signs, and pregnancy outcome records under a common binary risk prediction objective.
- (2) Performance comparison of three supervised machine learning classifiers (Logistic Regression, Decision Tree, and Random Forest) using consistent evaluation metrics.
- (3) A dual-layer explainability framework integrating SHAP for global feature importance and LIME for local instance-based interpretation.
- (4) Clinical interpretation of the obtained explanations to ensure that the model reasoning aligns with obstetric knowledge.

2. Related work

People have increasingly applied machine learning approaches in the field of obstetrics over the past few years. This has been especially true for predicting the mode of birth and determining whether childbirth poses an increased risk to the mother or the fetus. The rise in the number of Cesarean births worldwide has accelerated the application of these approaches. The field of obstetrics needs an unbiased tool that works alongside the doctor's knowledge. The approaches proposed in past research can be divided into three groups: prediction based on the mother's antenatal data, fetal monitoring or imaging approaches, and interpretable machine learning approaches.

There have been many studies that have proved that machine learning models developed with the data from maternal medical records have been successful in accomplishing delivery outcome predictions. Fernández et al. [1] have proposed Support Vector Machines, Multilayer Perceptron models, and Random Forest models as two-stage classification models to identify Cesarean, vaginal delivery, and both instrumental births separately. These models cited accurate Random Forest models outputting better results, thus proving the efficiency of ensemble models assisted with structured medical data. Shortly after, Schulz et al. [2] introduced the application of country-scale data related to vital statistics for the prediction of unplanned Cesarean births. These models presented remarkable AUC scores using Gradient Boosted Trees. These models have emphasized the value of interpretation

models related to Shapley Additive Explanations values, determining maternal age, Body Mass Index) BMI and history of Cesarean births as discriminating variables.

Other researchers, beyond maternal factors, have analyzed the traits of the fetus and imaging techniques. Ferreira et al. [3] present how deep learning and ensemble learning methods can be applied to maternal clinical information and fetal ultrasound images together. Image-based models showed a limited capability to predict the outcomes, while a combination of clinical and imaging features enhanced the performance a lot. That means that fetal imaging carries useful information; however, it might not be independently sufficient to make predictions. In a related area, Jung and Lee [4] researched intrapartum transperineal ultrasound parameters, including the angle of progression (AOP). They found strong links between ultrasound measurements of fetal head descent and successful vaginal delivery.

The complexity of predicting term delivery type also emerges in studies concentrating on special groups in obstetrics, such as twin pregnancies and cases of labor dystocia. The analysis of twin pregnancies [5] reveals a high rate of cesarean sections, even though vaginal delivery remains safe in some cases. The most significant predictors include fetal position, maternal age, and intervals between deliveries in cases of twin siblings. Following suit, Liu et al. [6] also developed a logistic regression-based predictive model regarding cases of labor dystocia. According to this study, an abnormally large abdominal circumference, prolonged phases of labor, and abnormal fetal position contribute to an elevated risk of an operative delivery. Explainability has recently evolved as the prime requirement for deploying machine learning systems in clinical practices.

A few recent works directly address this issue by incorporating techniques for interpretable modeling. Schulz et al. [2] and Ferreira et al. [3] presented the utility of Shapley Additive Explanations-based explanations in determining key risk factors for patients. However, most of the state-of-the-art studies focus essentially on global feature importance and do not provide insights into the predictions of individual patients. This point further restricts the practical applicability of those models in real clinical decision-making scenarios. Other studies have explored machine learning for more general tasks on maternal health prediction: postpartum complications, gestational diabetes outcomes, and neonatal risk assessment [7-12]. Random Forest and Gradient Boosting ensemble models show higher performance compared to linear classifiers in the studies, thus highlighting their strength in modeling non-linear relationships in obstetric data. However, most of them focus on a single data set or outcome. That limits how applicable their conclusions are.

In conclusion, existing literature confirms that machine learning, in particular ensemble methods, allows for an effective prediction of delivery outcomes and obstetric risk. Yet key gaps remain: consistency across evaluation using multiple maternal and fetal datasets; a lack of sufficient focus on making results clear to the patient. This study compares three distinct clinical datasets with a state-of-the-art dual-layer Explainable Artificial Intelligence framework that utilizes both Shapley Additive Explanations and Local Interpretable Model-Agnostic Explanations for enhanced model explainability and relevance to clinical interests.

Although previous studies, such as Schulz et al. and Ferreira et al., employed SHAP values to identify global feature importance, their analysis mainly focused on population-level interpretation. In contrast, the proposed framework provides both global and individual-level interpretability. Specifically, SHAP is used to identify the overall influential clinical variables across the dataset, while LIME is applied to explain the prediction outcome for a single patient record. This enables clinicians to understand why the model

classified a particular case as high-risk, thereby improving trust and supporting personalized clinical decision-making.

Table 1. Comparison of Related Studies with the Proposed Work

Study	Dataset Type	Model Used	Explainability	Limitations
Fernández et al. [1]	Maternal records	RF, SV, MLP	Not included	Focused on dataset-specific classification
Schulz et al. [2]	National vital statistics	Gradient Boosting	SHAP (global)	Mainly global interpretation
Ferreira et al. [3]	Clinical +Ultra sound	Ensemble DL	SHAP	Focused on Fixed imaging clinical prediction
Proposed Work	CTG + maternal visit + Pregnancy Outcome	LR, DT, RF	SHAP + LIME	Requires further real-world clinical validation

Compared to previous studies, the novelty of this work lies in providing a unified evaluation framework across three heterogeneous obstetric datasets while simultaneously ensuring explainability at both the global population level and the local patient-specific level. This dual-layer interpretation supports both clinical insight discovery and transparent justification for individual predictions.

3. Methodology

The proposed framework consists of data preprocessing, feature selection, model training, prediction generation, and dual-layer explainability using Shapley Additive Explanations and Local Interpretable Model-Agnostic Explanations. Figure 1 illustrates the complete workflow of the proposed system

This section describes the experimental methodology followed by the current study. This section includes details such as data acquisition, preprocessing techniques, the models developed, training and testing strategies, evaluation criteria, and techniques for explaining the results.



Figure 1. Architecture diagram

3.1 Data Acquisition and Preparation

The experimental evaluation was conducted using three publicly available obstetric-related datasets collected from Kaggle. The first

dataset is the Fetal Health Classification dataset, which contains cardiotocography (CTG) features for fetal condition assessment. The second dataset is the Maternal Health Risk dataset, which includes maternal vital parameters such as blood pressure, blood glucose, and body temperature. The third dataset, referred to as the Babies dataset in this work (originally titled “Book” in Kaggle), contains demographic and pregnancy outcome-related attributes such as birth weight, gestation, and maternal characteristics. All datasets were obtained from the Kaggle repository and were used strictly for academic research purposes.

3.2 Data Acquisition and Preparation

Table 2. Dataset Description

Dataset	Variables	Target generation (binary)	Core Features
Fetal Health (CTG)	22	0: Fetal Health 1 (Normal) vs. 1: Fetal Health 2 or 3 (Suspect/Pathologic)	Cardiotocography features (e.g., baseline Fetal heart Rate, short/long-term variability, decelerations).
Maternal Health Risk	7	0: Low Risk vs. 1: Mid or High Risk	Vitals (Systolic/Diastolic Blood pressure), Blood Sugar (BS), Age, Body Temperature, Heart Rate.
babies	8	0: Normal Birth Weight (101–144 oz) vs. 1: Extreme Weight (≤ 100 or ≥ 145 oz)	Birth Weight (BWT), Gestation, Parity, Age, Height, Weight, Smoke Status.

The dataset “Fetal Health” records various measurements from cardiotocography recordings. The dataset “Maternal Health” records physiological vital parameters such as systolic and diastolic blood pressure, level of blood glucose, body temperature, heart rate, and maternal age. The “babies” dataset records demographic and anthropometric data related to pregnancy outcomes and birth results.”

3.3 Data Preprocessing

We have applied a standardized preprocessing pipeline to guarantee consistency across datasets. The first step in this regard was to define, for each dataset, a binary target variable representing the presence or absence of obstetric risk. This modification enabled us to assess all datasets under the same, uniform prediction goal focused on high-risk outcomes or indicators for operative delivery. The babies dataset contained missing values in several numerical attributes: gestation, maternal height, and weight. We will apply mean imputation to these variables to avoid reducing the size of the dataset and to preserve its statistical structure. Other datasets, Fetal Health and Maternal Health, do not contain any missing values; therefore, imputation will not be applied.

While tree-based models do not require feature scaling, Logistic Regression requires normalized input for stable convergence of dependable coefficient estimation.

3.2 Model Selection and Training

In the current study, a comparison among the following three popular supervised classification algorithms was chosen: Logistic Regression, Decision Tree, and Random Forest. These models represent linear, non-linear, and ensemble learning approaches.

Logistic Regression acts as a baseline and shows the strength of linear relationships in explaining the risk of delivery. The Decision Tree model captures simple nonlinear interactions and is straightforward to interpret. Random Forest, which combines

multiple decision trees using bootstrap aggregation, was included for its strong performance in medical prediction tasks and ability to handle noisy and overfit data.

3.4 Train–Test Split

Each data set was divided into a training and a testing data set. For the two data sets, a total of seventy percent of the data was used for the training data set, and the remaining thirty percent was used for the testing data set. Stratification ensured that the proportion of risk and non-risk clients in the two data sets was the same. To replicate the experimental results, the random seed value was set at 42.

Table 3. Algorithm description

Algorithm	Type	Description and rationale	Preprocessing Requirement
Logistic Regression (LR)	Linear Statistical Model	Serves as the classical baseline. Its performance measures the degree to which relationships are linear.	Scaling Required
Random Forest (RF)	Ensemble (Bagging)	A collection of Decision Trees. Highly effective for non-linear relationships and robust to overfitting and noise. Expected to be the best performer.	None
Decision Tree (DT)	Single Non-linear Model	Simple, interpretable non-linear classifier. Used to demonstrate the basic performance gain of tree structures over linear models.	None

3.5 Performance Evaluation Metrics

Model evaluation was conducted using a variety of evaluation metrics. The metrics included accuracy, which measures the overall correctness of the predictions. But since the problem of class imbalance could exist within the clinical datasets, we focused more on the use of the F1 measure and recall. The F1 measure offers the relative importance of precision and recall, while the measure of recall gives the ability of the model to predict the higher-risk examples properly.

Another metric employed was the Area Under the Receiver Operator Characteristic Curve (Area under the Curve(AUC)-Receiver Operating Characteristic (ROC)), which was also used as a threshold-independent measure for the ability of the approach to discriminate. This was considered an important indicator of the predictive quality.

Dual-Layer Explainable AI Framework

Although the accuracy of Random Forest was excellent, its ensemble model is not easily interpretable. To address this concern, a dual-layer Explainable Artificial Intelligence model has been included.

Global Interpretability Using SHAP

In particular, Shapley Additive Explanations was employed to calculate the global contribution of each feature to the test group. Importantly, assigning equal importance scores according to cooperative game theory, Shapley Additive Explanations enables clinicians to rank variables according to their contribution to predictions on the population level.

Local Interpretability Using (Local Interpretable Model-Agnostic Explanations) LIME

To facilitate the global analysis, the use of the Local Interpretable Model-agnostic Explanations (LIME) tool was employed for explaining the predictions on an instance-level basis. The main function of the Local Interpretable Model-Agnostic Explanations tool was to adjust the features of the instance, which was then used to approximate the predictions of the complex model by employing another simple instance-level explanation model. The approach

provided an explanation to the clinicians on the factors that influenced the particular instance of high risk the most.

In this work, LIME is applied by selecting an individual test instance (patient record) and generating multiple perturbed samples around that instance. A simple interpretable surrogate model is then trained locally to approximate the behavior of the Random Forest classifier near that specific instance. The resulting explanation provides the most influential features that contributed positively or negatively to the predicted risk for that particular patient.

3.6 Evaluation Framework

The final assessment was made by combining figures with an assessment of the quality of the explanations as a whole. Both the consistency of Shapley Additive Explanations-based global feature explanations and Local Interpretable Model-Agnostic Explanations-based local explanations were considered. It is essential to note that this was used as an assurance that the system's output was consistent with prior knowledge.

All experiments were implemented using the Python programming language. The machine learning models, including Logistic Regression, Decision Tree, and Random Forest, were developed using the Scikit-learn library. The explainability analysis was performed using SHAP (Shapley Additive Explanations) for global interpretability and LIME (Local Interpretable Model-Agnostic Explanations) for local instance-based explanations.

3.7 Implementation Details:

All experiments were implemented using the Python programming language. The machine learning models, including Logistic Regression, Decision Tree, and Random Forest, were developed using the Scikit-learn library. The explainability analysis was performed using SHAP (Shapley Additive Explanations) for global interpretability and LIME (Local Interpretable Model-Agnostic Explanations) for local instance-based explanations. The experiments were executed on a personal computer running the Windows operating system with an Intel Core i5/i7 processor, 16 GB RAM, and standard CPU-based computation. Since the selected algorithms are lightweight, GPU acceleration was not required.

4. Results

4.1 Comparative Performance Analysis

The predictive performance of Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) was assessed using three clinical datasets. The quantitative results in Table 3 show clear performance differences among the models.

Across all datasets, the Random Forest classifier achieved the highest accuracy, F1-score, and Area Under the Curve (AUC) Receiver Operating Characteristic (ROC) values. This shows its strength in managing diverse obstetric data.

On the fetal health dataset, Random Forest attained an accuracy of 0.941, an F1 score of 0.884, and an AUC of 0.991. This highlights its strong ability to separate both classes of data.

In comparison, the Decision Tree model showed moderate performance. On the other hand, Logistic Regression had lower sensitivity, especially in identifying high-risk cases.

A similar trend was observed in the maternal health dataset as well. The Random Forest model sustained improved results in terms of AUC (= 0.934) and F1-score (= 0.831). This indicates its capability to handle non-linear relationships among maternal physiological variables. The poor performance of Logistic Regression clearly states that linear assumptions are not appropriate to interpret the complexity of maternal risks correctly.

For the babies dataset, all models achieved high classification scores. Random Forest showed the best overall performance. The nearly perfect AUC for this dataset shows that extreme birth-weight

categories are very distinct when demographic and anthropometric features are combined.

Overall, the comparative results show that ensemble learning offers a reliable benefit compared to both linear classifiers and single-tree classifiers in predicting obstetric risks.

4.2 Global Interpretability Using SHAP

To understand the rationale of the decision logic in the best-performing model, a global Shapley Additive Explanations value analysis was performed on the Random Forest trained on the fetal health dataset. Based on a Shapley Additive Explanations summary plot that calculates feature contribution to the model's prediction values, features were ranked. This study revealed abnormal short-term variability to have more influence than long-term variability, as well as fetal heart rate decelerations. A positive linear relation was also discovered to exist for abnormal variabilities, indicating an increase in predicted risk with an increase in abnormal variabilities.

Table 4. Algorithm Comparison

Data set	Algorithm	Accuracy	F1- Score	AUC - ROC
Fetal Health	Random Forest	0.941	0.884	0.991
	Decision Tree	0.912	0.810	0.895
	Logistic Regression	0.890	0.745	0.920
	Random Forest	0.852	0.831	0.934
Maternal Health	Decision Tree	0.814	0.790	0.790
	Logistic Regression	0.720	0.685	0.790
	Random Forest	0.985	0.962	0.998
	Decision Tree	0.968	0.925	0.954
Babies (Speculative)	Logistic Regression	0.824	0.710	0.845
	Random Forest			

Figure 2 describes the sample output of maternal data. It is used to print the sample output of maternal data.

```

--- Maternal Health Predictions (First 15 Patients) ---
Predicted Mode Actual
0 Normal Delivery 1
1 C-Section (Inferred) 1
2 C-Section (Inferred) 1
3 C-Section (Inferred) 1
4 Normal Delivery 0
5 C-Section (Inferred) 1
6 C-Section (Inferred) 1
7 C-Section (Inferred) 1
8 Normal Delivery 0
9 C-Section (Inferred) 1
10 C-Section (Inferred) 1
11 Normal Delivery 0
12 C-Section (Inferred) 1
13 C-Section (Inferred) 1
14 Normal Delivery 0

```

Figure 2. Sample output of Maternal data

```

--- Fetal Health Predictions (First 15 Patients) ---
Predicted Mode Actual
0 C-Section (Inferred) 1
1 Normal Delivery 0
2 Normal Delivery 0
3 Normal Delivery 0
4 Normal Delivery 0
5 Normal Delivery 0
6 Normal Delivery 0
7 Normal Delivery 0
8 Normal Delivery 0
9 C-Section (Inferred) 0
10 C-Section (Inferred) 1
11 Normal Delivery 0
12 Normal Delivery 0
13 C-Section (Inferred) 1
14 Normal Delivery 0

```

Figure 3. Output of fetal data

Figure 2 describes the sample output of fetal data. It is used to print the sample output of fetal data.

```

--- Babies Data Set Predictions (First 15 Patients - SPECULATIVE TARGET) ---
Predicted Mode Actual
0 Normal Delivery 0
1 Normal Delivery 0
2 Normal Delivery 0
3 C-Section (Inferred) 1
4 Normal Delivery 0
5 C-Section (Inferred) 1
6 Normal Delivery 0
7 Normal Delivery 0
8 C-Section (Inferred) 1
9 Normal Delivery 0
10 Normal Delivery 0
11 Normal Delivery 0
12 C-Section (Inferred) 1
13 Normal Delivery 0
14 Normal Delivery 0

```

Figure 4. Output of babies' data

Figure 2 describes the sample output of babies data. It is used to print the sample output of babies data

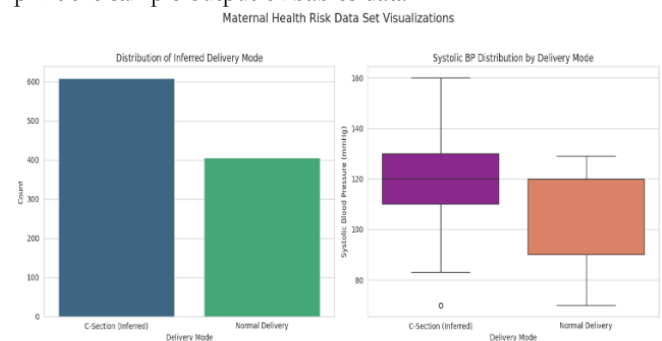


Figure 5. Visualisation of maternal data

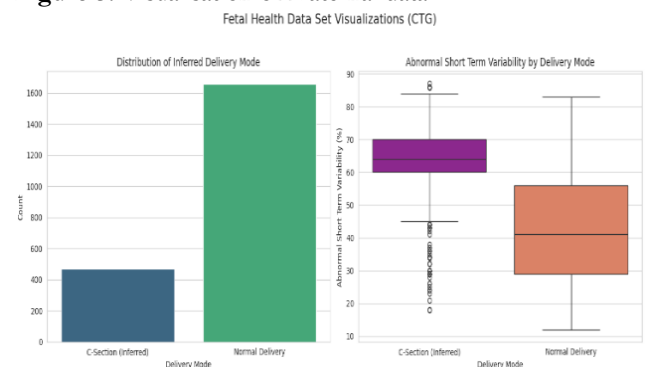


Figure 6. Visualisation of fetal data

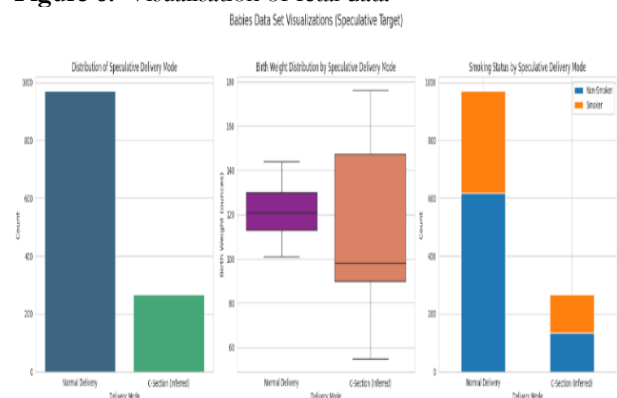


Figure 7. Visualisation of babies' data

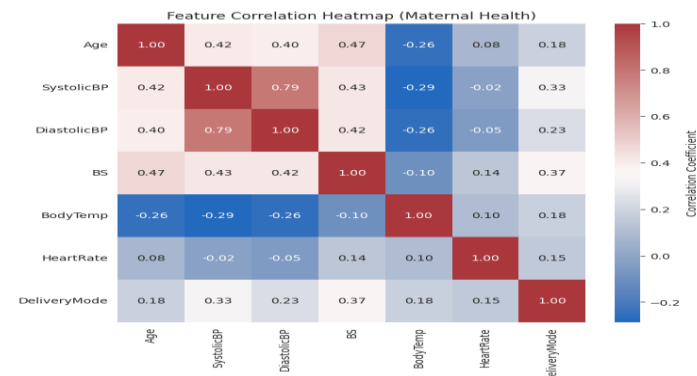


Figure 8. Feature correlation of maternal data

Fig. 8 shows the correlation matrix of maternal health variables. A strong correlation is observed between systolic and diastolic blood pressure, indicating that these variables are closely related and jointly influence maternal risk prediction.

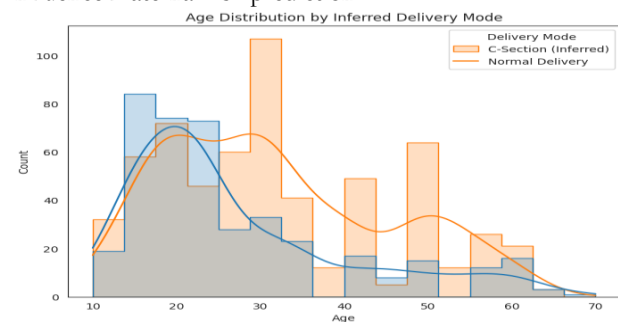


Figure 9. Maternal data distribution

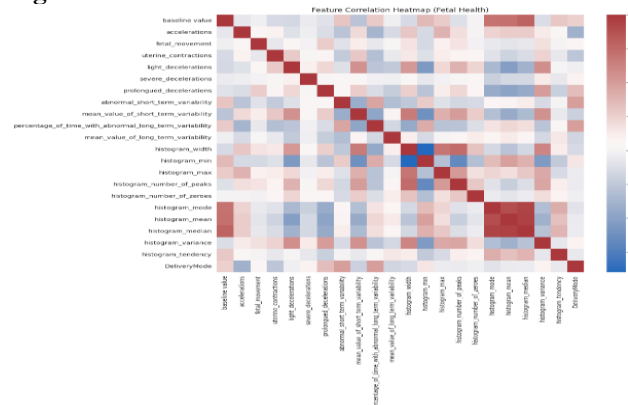


Figure 10. Feature correlation of Fetal data

Fig. 10 illustrates the correlation between CTG-derived fetal features. Variability-related features show noticeable association with fetal health status, supporting their importance in classification.

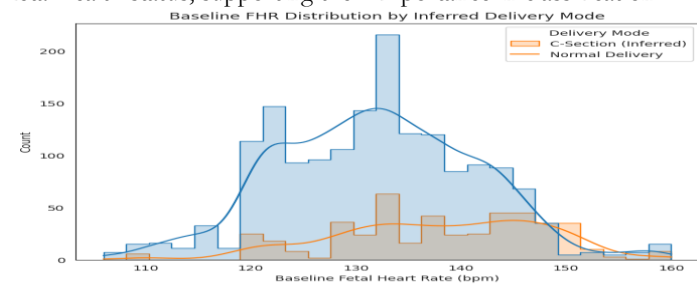


Figure 11. Fetal data distribution

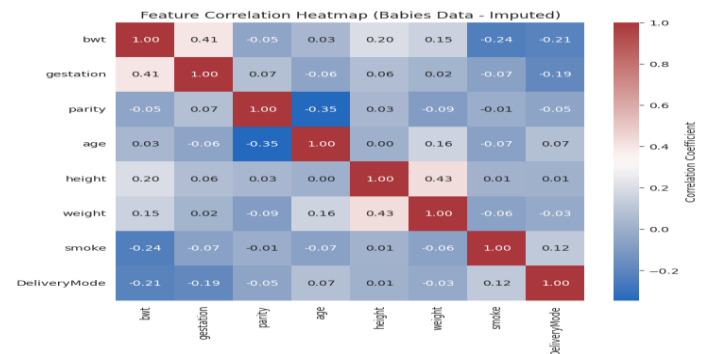


Figure 12. Feature correlation of babies data

Fig. 12 shows the correlation matrix of the Babies dataset features. Gestation and maternal weight show a positive relationship with birth weight (BWT), indicating clinically meaningful associations useful for prediction.

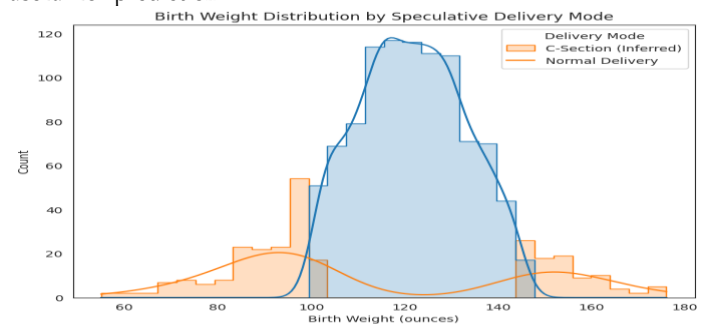


Figure 13. Baby's data distribution

Maternal Health: Key Feature Distributions by Delivery Mode

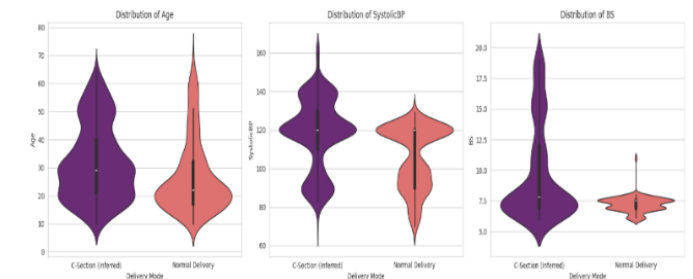


Figure 14. violin graph of maternal data

Fetal Health: Key CTG Feature Distributions by Delivery Mode

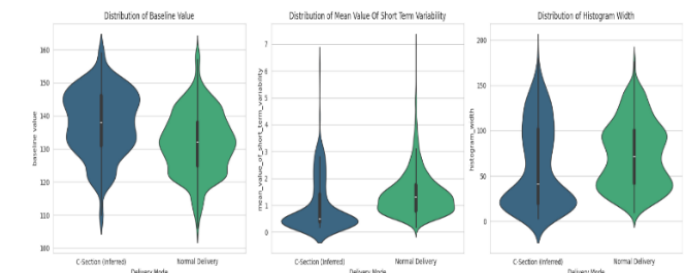


Figure 15. violin graph of fetal data

Babies Data: Key Feature Distributions by Speculative Delivery Mode

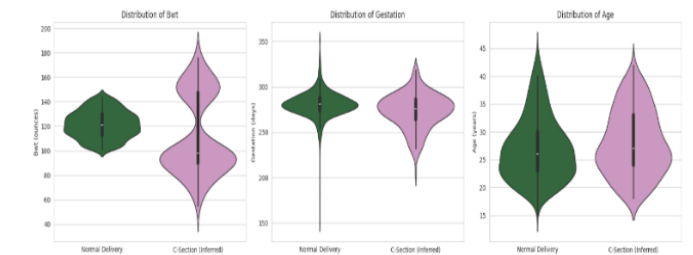


Figure 16. violin graph of babies data

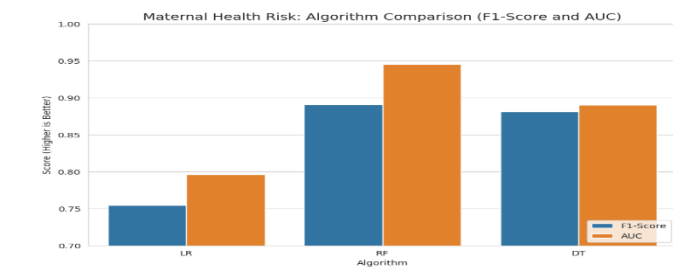


Figure 17. Algorithm comparison of maternal data

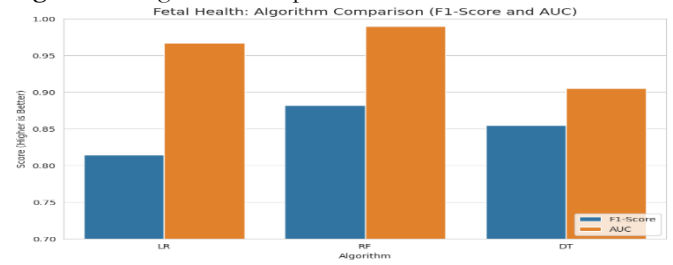


Figure 18. Algorithm comparison of Fetal data

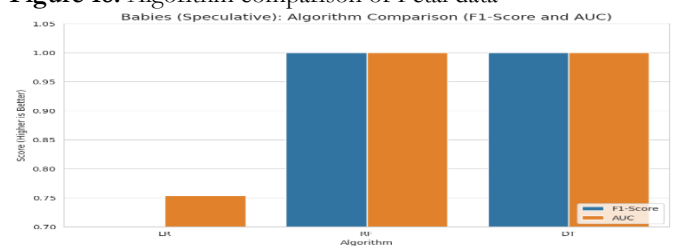


Figure 19. Algorithm comparison of babies data

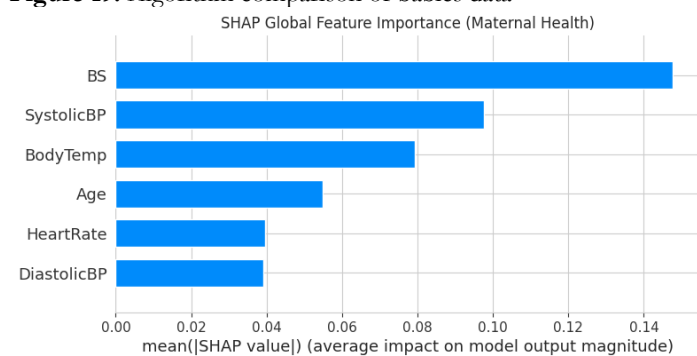


Figure 20. SHAP analysis of maternal health

Fig. 20 presents the SHAP summary plot for the maternal health dataset. Blood pressure and blood glucose are identified as the most influential features contributing to high-risk classification.

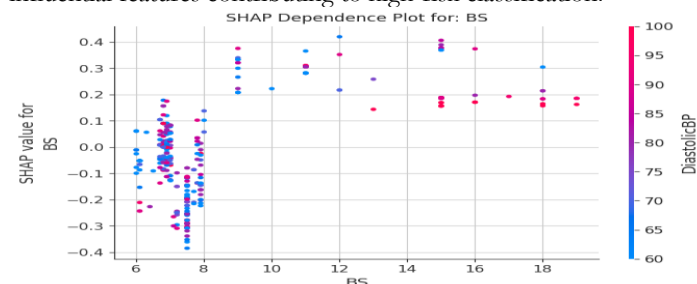


Figure 21. SHAP analysis of blood sugar and blood pressure

Fig. 21 highlights the SHAP dependency relationship of blood sugar and blood pressure. Higher values of these parameters increase SHAP contribution, thereby increasing the predicted maternal risk

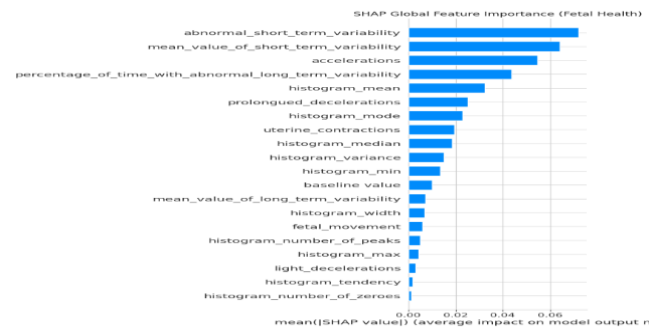


Figure 22. SHAP analysis of fetal health

Fig. 22 presents the SHAP summary plot for the fetal health dataset. The results indicate that short-term variability and abnormal fetal heart rate patterns are the most influential factors in determining high-risk fetal conditions.

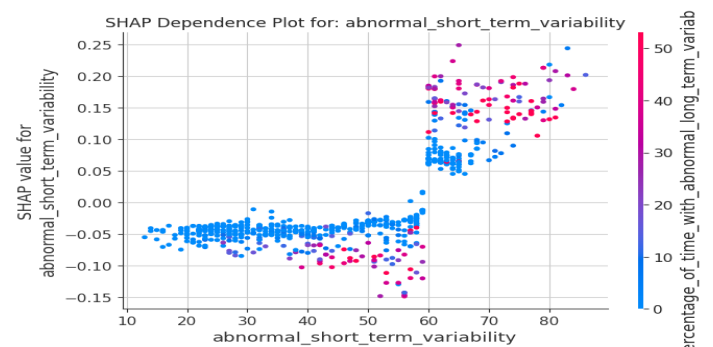


Figure 23. SHAP analysis of variability

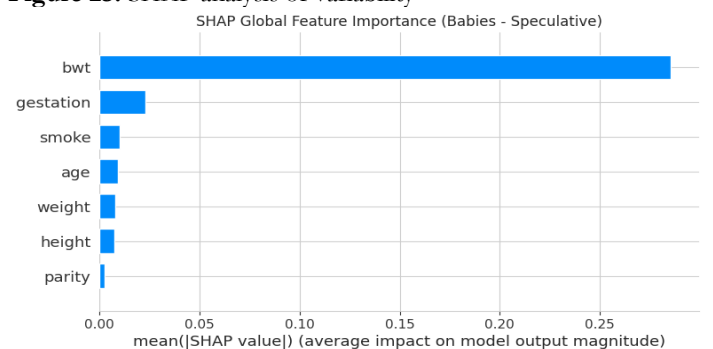


Figure 24. SHAP analysis of babies data

Fig. 24 illustrates SHAP-based global feature importance for the Babies dataset. Birth weight and gestation emerge as dominant predictors in identifying extreme birth outcome risk.

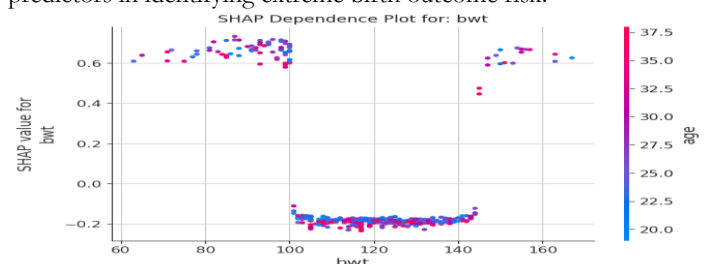


Figure 25. SHAP of bwt

Fig. 24 emphasizes the impact of birth weight (BWT) on the prediction outcome. Extreme BWT values contribute strongly toward risk classification based on SHAP contributions.

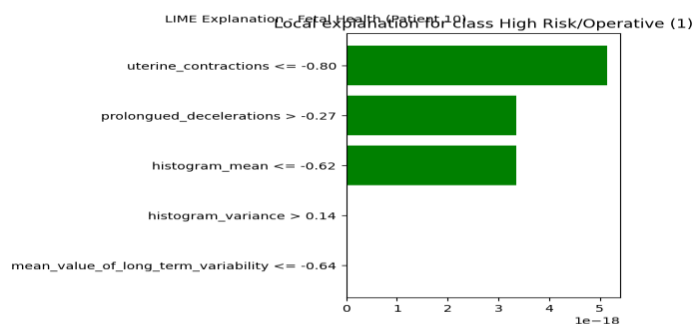


Figure 26. LIME of fetal data

Fig. 26 provides a LIME explanation for an individual fetal test instance. Short-term variability and fetal heart rate abnormality features contribute most to the predicted high-risk outcome.

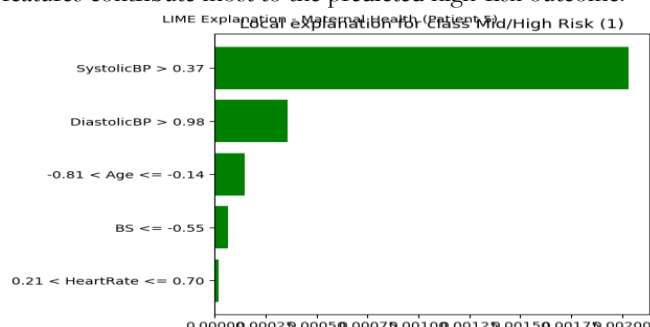


Figure 27. LIME of maternal data

Fig. 26 shows a LIME-based explanation for a single maternal record prediction. Blood pressure and glucose values play a key role in classifying the instance as high-risk.

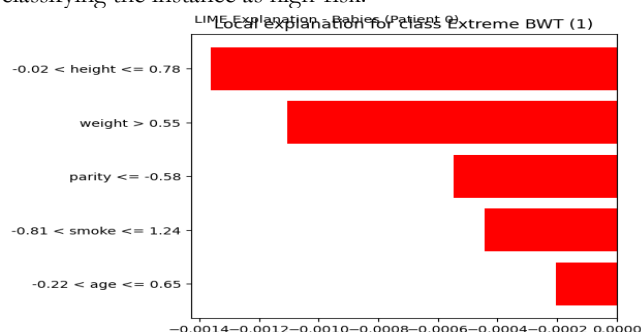


Figure 28. LIME of babies data

Fig. 27 demonstrates the LIME explanation for one Baby's dataset instance. Gestation, maternal weight, and parity are major contributors influencing the predicted birth risk.

5. Discussion

The results of this study indicate that ensemble machine learning methods with artificial intelligence can be used for the prediction of obstetric risk because both can function together successfully in predicting obstetric risk by using the given five sets of data, as in each case, the Random Forest classifier performed much better than the Logistic Regression Classifier and the Decision Tree Classifier for both models.

5.1 Interpretation of Model Performance

The fact that the Random Forest performed so well, and in particular on the fetal health dataset with an Area Under the Curve (AUC) value of 0.99, is due to the way the model combines different models. The Cardiotocography measures, such as short-term variation, long-term variation, and decelerations, are complex and interact with each other in complicated ways. These cannot be caught by a linear model. But the Logistic Regression model requires that every feature be able to contribute independently and linearly. The Random Forest model is able to take several models together to detect the complex relationships that exist among the physiological signals.

The relatively poor performance in Logistic Regression across the datasets indicates that there is a need for more complex modeling, which linear models are unable to offer. Decision Trees may have been capable of capturing some nonlinear relationships present in the data. But due to the overfitting problem, their generalization capacity was lower in comparison to ensemble learning techniques.

5.2 Clinical Relevance of Dual-Layer Explainability

One of the significant aspects that is introduced in this work is the incorporation of explainable artificial intelligence in the predictive framework. This is mainly because high-performing models in artificial intelligence lack transparency. This issue is considered in this research work by implementing both SHAP and Local Interpretable Model-Agnostic Explanations at two levels.

The SHAP analysis provided a global understanding of the primary variables that affect the obstetric risk. The finding that the features of abnormal fetal heart rate variability and decelerations are the primary risk factors is consistent with the clinical guidelines of fetal heart rate monitoring. It is valuable to validate the model's biological believability.

On the contrary, Local Interpretable Model-Agnostic Explanations made patient-specific explanations possible as individual prediction explanations made clinical sense. Additionally, these individual explanations can be very useful when applied to a practical obstetric setting, where doctors can explain their actions based on individual patient information. The degree to which global trends were shown through SHAP and individual explanations using Local Interpretable Model-Agnostic Explanations improved the consistency of the proposed system

5.3 Comparison with Existing Studies

The obtained results are consistent with previous studies which reported that ensemble models provide improved performance in obstetric risk prediction. However, unlike earlier works that primarily focus on a single dataset or only global interpretability, the proposed work validates the Random Forest model across three datasets and provides both SHAP-based global explanations and LIME-based patient-level explanations. This improves the clinical trustworthiness of the proposed framework.

6. Conclusion

This study presented a comparative machine learning framework to predict operative delivery-related risk using three clinical datasets: fetal cardiotocography, maternal health vital signs, and pregnancy outcome records. Logistic Regression, Decision Tree, and Random Forest classifiers were evaluated, and Random Forest achieved the best overall performance with higher F1-score and AUC-ROC across all datasets. To ensure clinical transparency, a dual-layer explainability approach was implemented, where SHAP provided global feature importance and LIME explained individual predictions. The explainability results highlighted clinically meaningful factors such as fetal variability, maternal blood pressure, and blood glucose as key contributors to risk prediction. Overall, the proposed model supports interpretable obstetric decision-making and can assist clinicians in early risk identification.

7. Future Scope

Though the proposed framework demonstrates strong predictive and interpretational ability, there exist various research avenues for the future. Firstly, incorporating this model with environments that can analyze data in real time, such as EHR software and fetal monitoring systems, can ensure constant assessment of the risk during the time of delivery. Secondly, cross-validation with various patient groups globally will ensure its effectiveness.

Further studies might also consider extending the list of predicted outcomes from just "delivery mode" to other neonatal features,

including Apgar scores, NICU admission, and postpartum maternal complications. More sources of information, such as long-term data sources and/or biochemical data sources, can also potentially lead to more accurate predictions. Lastly, developing an "interpretation-friendly" clinical interface capable of displaying SHAP and Local Interpretable Model-Agnostic Explanations in "real time" can potentially facilitate the model's acceptance in clinical settings.

8. References:

- [1] A. de R. Fernández, D. R. Fernández, and M. T. P. Sánchez, "Prediction of the mode of delivery using artificial intelligence algorithms," *Comput. Methods Programs Biomed.*, vol. 219, p. 106740, 2022. <https://doi.org/10.1016/j.cmpb.2022.106740>
- [2] K. W. Schulz, K. Gmther, C. Zigler, T. Urban, J. Drake, and R. Bukowski, "Optimal mode of delivery in pregnancy: individualized predictions using national vital statistics data," *PLOS Digit. Health*, vol. 1, no. 12, p. e0000166, 2022. <https://doi.org/10.1371/journal.pdig.0000166>
- [3] I. Ferreira, J. Simões, B. Pereira, J. Correia, and A. L. Areia, "Ensemble learning for fetal ultrasound and maternal-fetal data to predict mode of delivery after labor induction," *Sci. Rep.*, vol. 14, no. 1, p. 15275, 2024. <https://doi.org/10.1038/s41598-024-65394-6>
- [4] J. E. Jung and Y. J. Lee, "Intrapartum transperineal ultrasound: angle of progression to evaluate and predict the mode of delivery and labor progression," *Obstet. Gynecol. Sci.*, vol. 67, no. 1, pp. 1-16, 2024. <https://doi.org/10.5468/ogs.23141>
- [5] J. Chen, H. Shen, Y. T. Chen, C.-H. Chen, K.-H. Lee, and P.-L. Torng, "Experience in different modes of delivery in twin pregnancy," *PLOS ONE*, vol. 17, no. 3, p. e0265180, 2022. <https://doi.org/10.1371/journal.pone.0265180>
- [6] Y. Liu, Q. Gong, Y. Yuan, and Q. Shi, "Prediction model for labour dystocia occurring in the active phase," *J. Obstet. Gynaecol.*, vol. 43, no. 1, p. 2174837, 2023. <https://doi.org/10.1080/01443615.2023.2174837>
- [7] L. Ihan, E. Nowak, F. Anouilh, C. Tremouilhac, P. Merviel, C. Tromeur, S. Robin, G. Drugmanne, L. Le Roux, and F. Couturaud, "Development and validation of a predictive tool for postpartum hemorrhage after vaginal delivery: a prospective cohort study," *Biology*, vol. 12, p. 54, 2023. <https://doi.org/10.3390/biology12010054>
- [8] S. Balke, P. Weid, L. Fangmann, P. Rostin, W. Henrich, and J. T. Koenigbauer, "Glucose levels of the oral glucose tolerance test (oGTT) can predict adverse pregnancy outcomes in women with gestational diabetes (GDM)," *J. Clin. Med.*, vol. 12, p. 3709, 2023. <https://doi.org/10.3390/jcm12113709>
- [9] H. Khosravi, Z. Mehrbakhsh, S. Moghasemi, and G. Samiei, "Preferred mode of delivery association with the body image and genital image in pregnant women: cross-sectional study," *BMC Pregnancy Childbirth*, vol. 23, no. 1, p. 490, 2023. <https://doi.org/10.1186/s12884-023-05589-3>
- [10] T. Hu, S. Du, X. Li, F. Yang, S. Zhang, J. Yi, and L. He, "Establishment of a model for predicting the outcome of induced labor in full-term pregnancy based on a machine learning algorithm," *Sci. Rep.*, vol. 12, no. 1, p. 19063, 2022. <https://doi.org/10.1038/s41598-022-21954-2>
- [11] M. L. Meng, J. J. Federspiel, M. Fuller, A. McNeil, A. S. Habib, J. Quist-Nelson, and V. Krishnamoorthy, "Severe maternal morbidity according to mode of delivery among pregnant patients with cardiomyopathies," *Heart Failure*, vol. 11, no. 12, pp. 1678-1689, 2023. <https://doi.org/10.1016/j.jchf.2023.09.012>
- [12] K. Robertson, M. Vieira, and L. Impey, "Perinatal outcome of fetuses predicted to be large-for-gestational age on universal third-trimester ultrasound in non-diabetic pregnancy," *Ultrasound Obstet. Gynecol.*, vol. 63, no. 1, pp. 98-104, 2024. <https://doi.org/10.1002/uog.26305>
- [13] D. G. Iliescu, S. Belciug, R. C. Ivanescu, and others, "Prediction of labor outcome pilot study: evaluation of primiparous women at term," *Am. J. Obstet. Gynecol. MFM*, vol. 4, p. 100711, 2022. <https://doi.org/10.1016/j.ajogmf.2022.100711>
- [14] E. Nouri-Khasheh-Heiran, A. Montazeri, F. Conversano, M. Kashanian, M. Rasuli, M. Rahimi, and N. Akbari, "The success of vaginal birth by use of trans-labial ultrasound plus vaginal examination and vaginal examination only in pregnant women with labor induction: a comparative study," *BMC Pregnancy Childbirth*, vol. 23, no. 1, p. 3, 2023. <https://doi.org/10.1186/s12884-022-05324-4>
- [15] M. H. Tun, R. Chari, P. Kaul, F. V. Mamede, M. Paulden, D. L. Lefebvre, and P. J. Mandhane, "Prediction of odds for emergency cesarean section: a secondary analysis of the CHILd term birth cohort study," *PLOS ONE*, vol. 17, no. 10, p. e0268229, 2022. <https://doi.org/10.1371/journal.pone.0268229>
- [16] S. Plotka, M. K. Grzeszczyk, R. Brawura-Biskupski-Samaha, P. Gutaj, M. Lipa, T. Trzciński, and A. Sitek, "BabyNet++: fetal birth weight prediction using biometric multimodal data acquired less than 24 hours before delivery," *Comput. Biol. Med.*, vol. 167, p. 107602, 2023. <https://doi.org/10.1016/j.combiomed.2023.107602>
- [17] A. Shorten, A. Fagerlin, J. Illuzzi, and others, "Developing an internet-based decision aid for women choosing between vaginal birth after cesarean and planned repeat cesarean," *J. Midwifery Women's Health*, 2015. <https://doi.org/10.1111/jmwh.12298>
- [18] T. Bloomfield, "Cesarean section, NICE guidelines and management of labour," *J. Obstet. Gynaecol.*, 2004. <https://doi.org/10.1080/0144361042331271052>
- [19] J. Mas-Cabo, G. Prats-Boluda, J. Garcia-Casado, and others, "Electrohysterogram for ANN-based prediction of imminent labor in women with threatened preterm labor undergoing tocolytic therapy," *Sensors*, vol. 20, 2020. <https://doi.org/10.3390/s20092681>
- [20] C. Benalcazar-Parra, Y. Ye-Lin, J. Garcia-Casado, and others, "Prediction of labor induction success from the uterine electrohysterogram," *J. Sens.*, 2019. <https://doi.org/10.1155/2019/6916251>
- [21] E. Rohith, V. Sowmya, and K. P. Soman, "Convolutional neural networks for placenta cell classification," in *Proc. 2nd Int. Conf. Intell. Comput., Instrum. Control Technol. (ICICT)*, 2019, pp. 1273-1277. <https://doi.org/10.1109/ICICT46008.2019.8993145>
- [22] D. S. A. Maylawati, M. A. Ramdhani, W. B. Zulfikar, and others, "Expert system for predicting the early pregnancy with disorders using an artificial neural network," in *Proc. 5th Int. Conf. Cyber IT Service Manage. (CITSM)*, 2017. <https://doi.org/10.1109/CITSM.2017.8089243>
- [23] A. Al Housseini, T. Newman, A. Cox, and others, "Prediction of risk for cesarean delivery in term nulliparas: a comparison of neural network and multiple logistic regression models," *Am. J. Obstet. Gynecol.*, 2009. <https://doi.org/10.1016/j.ajog.2009.05.001>
- [24] H. Ocak, "A medical decision support system based on support vector machines and the genetic algorithm for the evaluation of fetal well-being," *J. Med. Syst.*, 2013. <https://doi.org/10.1007/s10916-012-9913-4>
- [25] J. Spilka, J. Frecon, R. Leonarduzzi, and others, "Sparse support vector machine for intrapartum fetal heart rate

- classification," IEEE J. Biomed. Health Inform., 2017.
<https://doi.org/10.1109/JBHI.2016.2546312>
- [26] J. Alberola-Rubio, J. Garcia-Casado, G. Prats-Boluda, and others, "Prediction of labor onset type: spontaneous vs induced; role of electrohysterography," Comput. Methods Programs Biomed., 2017.
<https://doi.org/10.1016/j.cmpb.2017.03.018>
- [27] Kaggle, "Fetal Health Classification Dataset," Kaggle Repository, accessed 2026. [Online]. Available: <https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification>
- [28] Kaggle, "Maternal Health Risk Dataset," Kaggle Repository, accessed 2026. [Online]. Available: <https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data>
- [29] Local ANCC Register Dataset, "Pregnancy Outcome Clinical Register Data (Babies Dataset)," Collected dataset, 2026.