

# PARENTAL IMPRINTS ON BIRTH WEIGHT: A DATA-DRIVEN MODEL FOR NEONATAL PREDICTION IN LOW RESOURCE PRENATAL CARE

Rajeshwari Mistri<sup>1</sup>, Harsh Joshi<sup>1</sup>, Nachiket Kapure<sup>1</sup>, Parul Kumari<sup>1</sup>,  
Manasi Mali<sup>1</sup>, Seema Purohit<sup>1</sup>, Neha Sharma<sup>2</sup>, Mrityunjay Panday<sup>3</sup>,  
Chittaranjan S. Yajnik<sup>4</sup>

<sup>1</sup>B.K. Birla College of Arts, Science and Commerce, Kalyan

<sup>2</sup>Tata Consultancy Services, India

<sup>3</sup>Cognizant Technologies, India

<sup>4</sup>Diabetes Unit, KEM Hospital and Research Centre, India

{rajeshwarimistri11, joshiharsh0506, kapnachi1904,  
parulkumari2307, malimanasi2002, nvsharma1975, mrityunjay.0113,  
csyajnik}@gmail.com, seema.purohit@bkbck.edu.in

April 9, 2025

## Abstract

Accurate fetal birth weight prediction is a cornerstone of prenatal care, yet traditional methods often rely on imaging technologies that remain inaccessible in resource-limited settings. This study presents a novel machine learning-based framework that circumvents these conventional dependencies, using a diverse set of physiological, environmental, and parental factors to refine birth weight estimation. A multi-stage feature selection pipeline filters the dataset into an optimized subset, demonstrating previously underexplored yet clinically relevant predictors of fetal growth. By integrating advanced regression architectures and ensemble learning strategies, the model captures non-linear relationships often overlooked by traditional approaches, offering a predictive solution that is both interpretable and scalable. Beyond predictive accuracy, this study addresses a question: whether birth weight can be reliably estimated without conventional diagnostic tools. The findings challenge entrenched methodologies by introducing an alternative pathway that enhances accessibility without compromising clinical utility. While limitations exist, the study lays the foundation for a new era in

prenatal analytics, one where data-driven inference competes with, and potentially re-defines, established medical assessments. By bridging computational intelligence with obstetric science, this research establishes a framework for equitable, technology-driven advancements in maternal-fetal healthcare.

**Keywords:** Obstetric Modeling, Longitudinal Analysis, Paragenetics, BART Modeling, Gestoplac, SHAP Analysis

# 1 Introduction

Birth weight(BW) is an important factor of neonatal health, with both low and high birth weights associated with a range of adverse outcomes, including increased risks of infant mortality, developmental delays, metabolic disorders, and long-term chronic diseases [1]. Accurately predicting fetal BW during pregnancy is essential for effective prenatal care, enabling healthcare providers to identify at-risk pregnancies, implement timely interventions, and optimize resource allocation [2]. Traditional methods for estimating BW rely heavily on ultrasound-derived biometric parameters, which, while effective, require specialized equipment and expertise often unavailable in resource-constrained settings [3]. This limitation underscores the need for alternative predictive models that can use readily available physiological, demographic, and environmental data to improve prenatal care and risk stratification.

Building on previous research that studied the use of maternal physiological characteristics for BW prediction, this study expands the scope by incorporating paternal features, an approach that recognizes the significant role fathers' attributes play in influencing fetal development [4]. Unlike earlier studies, which were limited by smaller datasets and a focus solely on maternal factors, this research utilizes a more extensive dataset and a broader set of predictors. By combining both maternal and paternal characteristics, the aim is to develop a more accurate predictive framework. Furthermore, this study refines the methodology by working with advanced data imputation strategies and feature selection process to ensure that only the most relevant prenatal attributes are used, thereby enhancing model interpretability and predictive power [5].

As healthcare systems increasingly adopt digital and data-driven approaches, the potential to use maternal and paternal physiological data for predictive analytics is developed as a promising research domain. This study contributes to this growing field by developing an interpretable, clinically relevant machine learning(ML) model capable of accurately estimating BW without reliance on imaging-based diagnostics. By using periodically collected data, the proposed framework offers a scalable solution adaptable to diverse healthcare settings, particularly those with limited access to ultrasound technology [6, 7].

## 1.1 OBJECTIVE

The objective of this research is to develop an accurate, interpretable ML model for BW prediction using prenatal data without advanced diagnostics. Specifically, the study aims to:

- Identify the maternal and paternal key features from a high-dimensional dataset to enhance predictive performance.
- Evaluate various ML models to capture complex relationships in the data in low-resource settings.
- Validate the clinical relevance of ML models in predicting birth outcomes, replacing traditional imaging methods.

To achieve the research objectives, this study is structured into several key sections. Section II provides a comprehensive literature review, examining existing approaches to BW prediction and identifying gaps in current research. Section III details the methodology, including data preprocessing techniques, imputation methods, feature selection strategies, and the ML algorithms employed. Following this, the study in Section IV presents the results, highlighting the performance of various models and identifying the most influential predictors. A discussion of the findings contextualizing them within the broader field of prenatal care and neonatal health is followed in Section V. Finally, Section VI concludes the paper by summarizing key insights, acknowledging limitations, and proposing future research directions to enhance the accuracy and clinical applicability of BW prediction models.

## 2 Literature Review

Recent advancements in ML techniques have significantly enhanced the predictive accuracy of fetal BW models, particularly when relying on non-invasive data. For instance, a study introduced a hybrid Long Short-Term Memory (LSTM) framework that demonstrated a 6% improvement in accuracy over conventional estimation methods [2]. By analyzing a dataset encompassing more than 5,700 pregnancies, this model underscores the viability of utilizing existing clinical data while illuminating the challenges that arise from the dependence on single-center data, which may limit the generalizability of the findings [2].

Handling missing data is critical in longitudinal prenatal datasets. Previous research has evaluated various imputation techniques, with Multiple Imputation by Chained Equations (MICE) demonstrated superior performance in handling missing values compared to K-Nearest Neighbors (KNN) imputation, particularly in preserving temporal consistency within longitudinal datasets, as evidenced by the Pune Maternal Nutrition Study (PMNS) dataset (Varma et al., 2024) [8]. The phase 1 study of this birth cohort further validated MICE’s effectiveness, reducing imputation error by 23% over conventional methods. Then in phase 2 study by N. Kapure et al., 2025 [9], the research uses advanced imputation, supervised feature selection, and ensemble-based regression models, with Gradient Boosting excelling in capturing complex maternal-fetal interactions. Identified key predictors, including gestational age, placental weight, and maternal anthropometric features, enhance neonatal risk assessment, reinforcing ML’s role in maternal and neonatal care [9].

In 2019, a study by Lu et al. demonstrated the effectiveness of ensemble learning methods combined with a genetic algorithm for fetal weight estimation from physiological data. While this approach achieved promising results, limitations in accuracy at extreme fetal weight values were noted [3]. Furthermore, regression studies remain underexplored, as most current

research focuses on classification-based outcomes such as low birth weight (LBW) [10]. The precise analysis of maternal nutritional biomarkers has revealed specific mechanisms influencing fetal growth patterns. A study conducted in Tarragona involving 729 pregnant women has found that mid-pregnancy plasma concentrations of vitamin B12 and folate were significant predictors of LBW, achieving a classification accuracy of 96.19% [1]. While these approaches achieve high predictive accuracy for classification tasks, such as identifying growth-restricted or macrosomic fetuses, they often fail to address the continuous nature of BW prediction. However, these biomarkers may not be universally accessible or relevant for populations in low-resource settings like rural India, where plasma biomarker data is scarce. Hence, while valuable, this study’s findings may not directly translate to Indian contexts. Our study focuses on regression; we rely on features more feasible to measure in Indian rural healthcare setups. India accounts for one-sixth of the global population, yet existing models often fail to address its unique demographic and socio-economic challenges, such as limited access to advanced diagnostic tools [11]. This research also addresses this gap by focusing on a rural Indian population and employing features that are both practical and culturally relevant. A 2020 study by Wahab et al. examined the impact of maternal early-pregnancy dietary glycemic load on fetal development in 3,471 pregnancies. The findings suggest that a higher glycemic load is associated with increased fetal abdominal circumference and estimated fetal weight in late pregnancy [12].

Findings gained further validation through dual-energy X-ray absorptiometry measurements in 400 women, which quantified a 12% increase in newborn fat mass corresponding to elevated maternal glycemic loads [13]. The precision of these measurements, coupled with the consistent correlation patterns across different sample sizes and measurement techniques, provides compelling evidence for incorporating nutritional biomarkers into BW prediction frameworks [13]. These insights offer a foundation for developing targeted prenatal nutrition protocols, though validation across broader demographic groups remains essential for establishing universal guidelines [12]. The analysis of paternal characteristics through multiple studies identified that paternal BMI and other anthropometric measurements can influence neonatal BW [6]. These measurements, evaluated across thousands of cases, demonstrated consistent correlations with infant BW when integrated with existing maternal health parameters in predictive assessments [6]. The study by Wu et al. established updated BW references based on gestational age using the Generalized Additive Model for Location, Scale and Shape (GAMLSS) model, providing a more precise framework for identifying growth abnormalities [7, 14]. Additionally, placenta weighing less than 400g at term are strongly linked to reduced fetal BW, as placental size directly influences nutrient and oxygen transfer to the fetus. Early detection of a small placenta can help predict LBW and guide timely interventions to improve neonatal outcomes [15].

Extensive research has been conducted on fetal BW prediction using ultrasound-derived parameters for fetal weight estimation. Studies employing the Hadlock and INTERGROWTH-21st methods demonstrated the utility of biometric parameters such as head circumference (HC), abdominal circumference (AC), femur length (FL), and biparietal diameter (BPD) in predicting small-for-gestational-age (SGA) and large-for-gestational-age (LGA) neonates [16, 17]. Moreover, the reliance on expansive feature sets encompassing the above given biometric parameters introduces complexities, particularly in low-resource environ-

ments. Similarly, a competing-risks model combining second-trimester ultrasound biometry with maternal demographics effectively stratified risks for small-for-gestational-age outcomes [18]. Ensemble learning models are optimized with genetic algorithms, have further refined these predictions, particularly in settings with access to high-quality imaging technologies, albeit with limitations in predicting weight extremes, when supplemented with ultrasound data and historical clinical records [3]. Despite these advancements, reliance on ultrasound measurements poses constraints, particularly in resource-limited settings. This study circumvents these limitations by exclusively employing non-ultrasound features, instead harnessing maternal physiological and anatomical data for expanding the scope of BW prediction research for broadening its applicability in resource-constrained environments, non-clinical applications.

In addition to ML approaches, the exploration of paragenetics and environmental determinants of BW has profound implications for predictive modeling. Research stemming from the Tohoku Medical Megabank Project, which analyzed a substantial cohort of 22,711 neonates and their parents, revealed distinct patterns of influence associated with genetic versus environmental factors [19]. The findings indicate that lifestyle choices and maternal behaviors, such as diet and physical activity, predominantly impact term BW's, whereas paragenetics predispositions have a more certain effect in cases of preterm births [19]. This differentiation highlights the necessity for integrative models that incorporate paragenetics, environmental, and socio-behavioral variables to enhance predictive accuracy for BW [19]. Furthermore, challenges related to clinical implementation were identified, including the influence of maternal characteristics such as obesity and height on estimation accuracy [10], which underscores the critical need for tailored prediction strategies that account for individual patient profiles and the complex nature of fetal growth determinants [20].

While significant advancements have been made in predicting fetal BW, current models often rely on limited feature sets and may not be directly applicable to diverse populations, especially those with unique demographic and socio-economic contexts like India. This research demonstrates the potential of achieving comparable results using a more targeted feature set, integrating a broader range of paternal factors beyond anthropometric data and exploring simple, scalable predictors are essential for improving model performance and applicability in resource-constrained settings [7]. Study also examines a specific, underrepresented Indian population. By using a feature set encompassing body size, dietary intake, blood pressure, biochemistry, delivery outcomes, paternal features, obstetric history, and age of elder child, this research demonstrates a scalable and context-specific solution for improving prenatal care [19]. By identifying key predictors and their interactions with fetal growth patterns, this informs the targeted interventions and improves pregnancy outcomes in diverse and complex population.

As studies based on regression remain scarce, with few investigations; in contrast, the study not only shifts the focus to regression-based prediction but also employs a significantly smaller and meticulously curated observations of neonatal-maternal pair, prioritizing features with the strongest predictive capabilities while ensuring model simplicity and applicability across diverse settings [2]. Unlike most existing research that relies on ultrasound-derived parameters this approach utilizes maternal physiological and anatomical features, addressing the challenge of BW prediction without imaging data, while emphasizing explainable AI to

enhance model transparency and interpretability [5].

### 3 Methodology

The methodology adopts a similar structured approach encompassing data collection, pre-processing, imputation, feature selection, ML model training, and evaluation, as outlined in PMNS Study Phase 2 [9], with modifications to incorporate the expanded dataset. This previous study focused on maternal health indicators to predict BW outcomes. The current phase 3 continues this work further by incorporating both paternal and maternal factors, substantially increasing the feature set. Although the data preprocessing and imputation approaches are still similar to Phase 2, Phase 3 adds complexity because of the larger and more diverse dataset, enabling a broader analysis.

#### 3.1 Data Source

The dataset utilized in this study originates from the Pune Maternal Nutrition Study (PMNS), an extensive birth cohort initiated in 1993 by the Diabetes Unit of KEM Hospital, Pune, in collaboration with researchers from Southampton, UK [21]. The PMNS, conducted across six villages near Pune, provides a comprehensive dataset focused on maternal and fetal health. The strength of the PMNS dataset lies in its long-term follow-up design, capturing serial measurements from pregnancy through neonatal and adolescent stages. The structured data collection includes both clinical and biochemical assessments, with biological samples such as blood, urine, and placental tissues archived for future investigations. This enables a comprehensive exploration of maternal and fetal health determinants, as well as their potential transgenerational impacts [8].

Data collection methodologies included standardized protocols for anthropometric measurements, blood sample analyses, dietary assessments using validated food frequency questionnaires, and socio-economic evaluations through structured interviews. Clinical evaluations such as gestational assessments, and metabolic profiling further enrich the dataset, providing an in-depth understanding of pregnancy outcomes. This meticulously curated dataset serves as a critical resource for investigating maternal, fetal, and neonatal health trajectories, offering valuable insights into developmental origins of health and disease across diverse population groups.

The PMNS dataset provides critical insights into:

- Determinants of fetal growth specific to Indian populations, allowing for a better understanding of regional and genetic influences on birth outcomes.
- The long-term evolution of cardiometabolic risks, including diabetes, within the framework of the Developmental Origins of Health and Disease (DOHaD), offering valuable data for preventive healthcare strategies.
- Transgenerational influences on health outcomes by incorporating data from maternal, paternal, and neonatal sources, supported by the inclusion of a third generation.

- The development of a preconceptional micronutrient intervention informed by maternal nutritional and metabolic data, aiding in the formulation of targeted maternal health interventions.

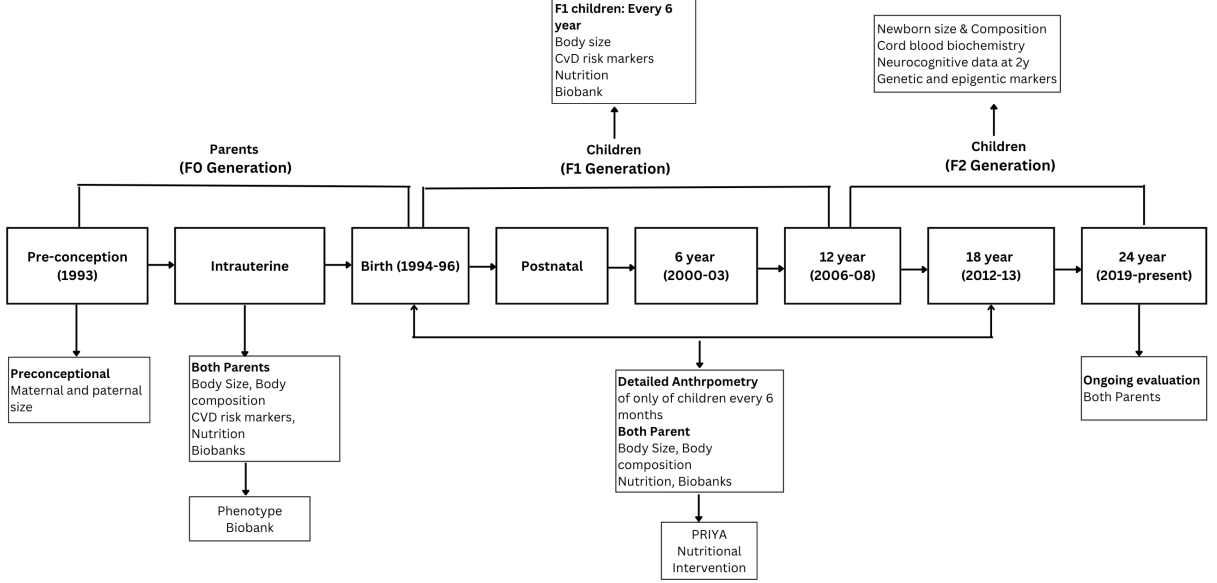


Figure 1: Data Source [9]

Phase 2 of the study focused solely on maternal characteristics, analyzing key health indicators such as anthropometric measurements, biochemical markers, and obstetric history to predict BW outcomes [9]. While this phase provided valuable insights, the absence of paternal accounts limits the scope of predictive modeling. Phase 3 expands upon this prior research by incorporating 5,979 features across 800 pregnancies, integrating both maternal and paternal data to enhance predictive modeling of BW and related health outcomes. This phase introduces paternal anthropometric, metabolic, and genetic factors, broadening the analytical scope and ensuring a more comprehensive representation of pregnancy and birth-related determinants. Additionally, the dataset includes detailed dietary intake records, biochemical markers, lifestyle behaviors, and genetic predisposition indicators, allowing for a multidimensional analysis of maternal and fetal health outcomes.

The dataset is securely stored on dedicated servers with restricted access, ensuring data integrity and confidentiality. Advanced data management protocols, including systematic validation checks and imputation techniques, are employed to address missing values and enhance data quality. For the current analysis, this extensive dataset is utilized to predict both maternal and paternal influences on BW, particularly in low-resource settings where early risk identification is crucial for improving neonatal health outcomes.

## 3.2 Data Exploration and Preprocessing

Exploratory data analysis (EDA) is conducted to assess the dataset’s structure, including column types (categorical or numerical) and the extent of missing data. In total, the dataset contained 4,729,389 data points, with 47.37% missing values. The high proportion of missing data required a thorough examination and informed the selection of preprocessing and imputation strategies to preserve the integrity of the subsequent analysis. To ensure that predictions of birth weight are exclusively based on prenatal data, a series of filtering steps is employed. Initially, columns representing data collected after childbirth are removed, reducing the dataset from 5,979 to 1,122 columns. Further refinement is applied to exclude postnatal maternal data, leaving 886 columns focused on prenatal factors. To maintain data quality and balance feature diversity, researchers retained only features with at least 60% non-missing data, resulting in 867 columns. Finally, only numerical columns, which are suitable for regression models are chosen, reducing the dataset to 852 columns. These filtering steps effectively reduces the amount of missing data from 47.37% to 4.37%, providing a more manageable and informative dataset for the analysis.

## 3.3 Statistical Analysis and Data Summary

Statistical analysis is utilized to summarize discrete (ordinal and nominal) and continuous attributes. Descriptive statistics like mean, standard deviation, skewness, kurtosis, and range are employed to define the distribution of the data. It supports the determination of patterns as well as anomalies, so that there can be an informed decision in choosing a model and imputation procedures. For example, determining features that were normally distributed directed the application of appropriate statistical and ML methodologies, ensuring that the models developed on these data have an informed understanding of the structure of the dataset.

## 3.4 Data Imputation

Building on these insights, data imputation is performed to address the remaining 4.37% of missing values in the filtered dataset, effective imputation strategies are critical for maintaining data consistency. Discrete variables are imputed using k-nearest neighbors (KNN), well-suited to categorical data [22]. For continuous variables, a combination of techniques are utilized, including Multiple Imputation by Chained Equations (MICE) and KNN [23]. These advanced methods capture complex patterns in the data, providing more reliable imputations than simpler techniques such as mean or median, which are deliberately avoided due to the non-normal distribution of many continuous features.

## 3.5 Feature Selection

Subsequently, to identify the most relevant features for predicting BW, 25 different feature selection techniques are implemented. This comprehensive approach includes Filter methods such as Improved Normalized Information-based Feature Selection (INMIFS), mutual information (MI), and Pearson correlation. These techniques assessed the relevance of



individual features based on their statistical dependencies, providing insights into feature importance from a univariate perspective [4]. Wrapper methods are also utilized, including Recursive Feature Elimination (RFE), forward selection, and Boruta to iteratively select features that maximize the performance of ML models. These techniques are particularly suited for capturing complex interactions among features that filter methods might overlook, as they evaluate subsets of features in conjunction with model training [24].

Table 1: Feature Selection Techniques categorized by method type.

Category	Techniques
Filter Methods	Pearson Correlation Spearman Correlation Kendall Correlation Mutual Information (MI-based) ANOVA Select K Best Relief EasyFS INMIFS MXM
Wrapper Methods	Forward Selection Recursive Feature Elimination (RFE) Boruta
Embedded Methods	Lasso Regressor Ridge Regressor Elastic Net SHAP Tree-based Gradient Boosting BART
Model-Agnostic Methods	Permutation Feature Importance (PFI) Regression-based Method MARS (Multivariate Adaptive Regression Splines) PLS (Partial Least Squares)
Novel/Hybrid Approaches	Hopular

The use of Boruta further ensures that only statistically significant features are retained, while RFE and forward selection provides systematic approaches to refine feature subsets. Embedded methods, such as regularization techniques like Lasso and Ridge, along with decision tree-based approaches, helps to identify important features during the model training process itself. These methods integrates feature selection into the learning algorithm, based on the intrinsic properties of the models to rank features, making them efficient and aligned with the predictive goal [25]. The combination of these techniques ensures the capture of both linear and non-linear relationships in the data. Depending on the specific method applied, features between 1-20 instances are selected without feature engineering or creation,

performed to maintain data integrity (Table 1).

### 3.6 Regression Models, Evaluation Criteria and Residual Analysis

The study utilizes 6 regression models to predict fetal birth weight: BART, Linear Regression (with L1 and L2 regularization), Gradient Boosting Regressor (GBR), CatBoost, LightGBM, and NGBoost. These models are selected for their ability to capture linear and nonlinear relationships among the features. Model evaluation is conducted using two primary metrics,  $R^2$  and RMSE, to ensure a robust assessment of predictive accuracy. While residual analysis incorporates visual tools, including error scatter plots, QQ-plots, bar plots, and Kernel Density Estimation (KDE) plots, to examine error distributions and alignment. Finally, the results are presented to subject matter expert to validate the study’s findings. This expert review ensured that the methodology and conclusions are grounded in biomedical expertise and could guide future research.

## 4 Results

### 4.1 Data Preprocessing

The dataset includes 305 continuous variables and 547 discrete variables, providing a diverse feature set for analysis. Normality tests (skewness and kurtosis) indicate that only 5 features follow a normal distribution, while 300 are non-normally distributed (see figure 2).

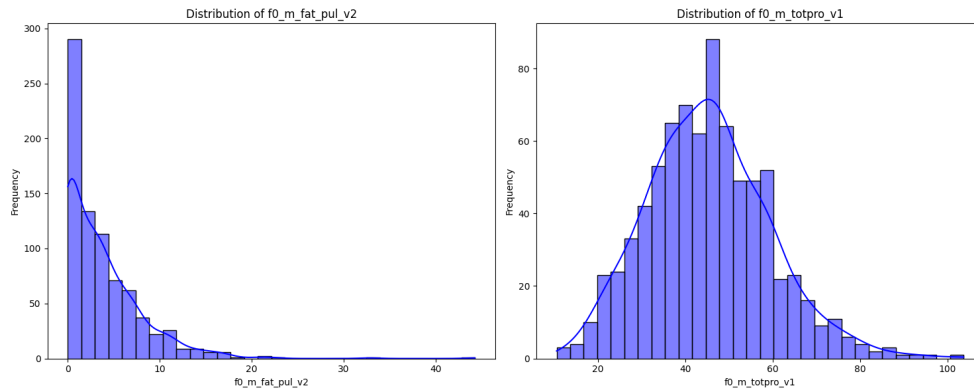


Figure 2: Distribution of Features

Exploratory Data Analysis (EDA) shows that approximately 70% of fetuses fall within the normal BW range (2,500–4,000 grams), while 29% are classified as moderately LBW (1,500–2,499 grams), as shown in Figure 3. This classification aligns with WHO and CDC guidelines.

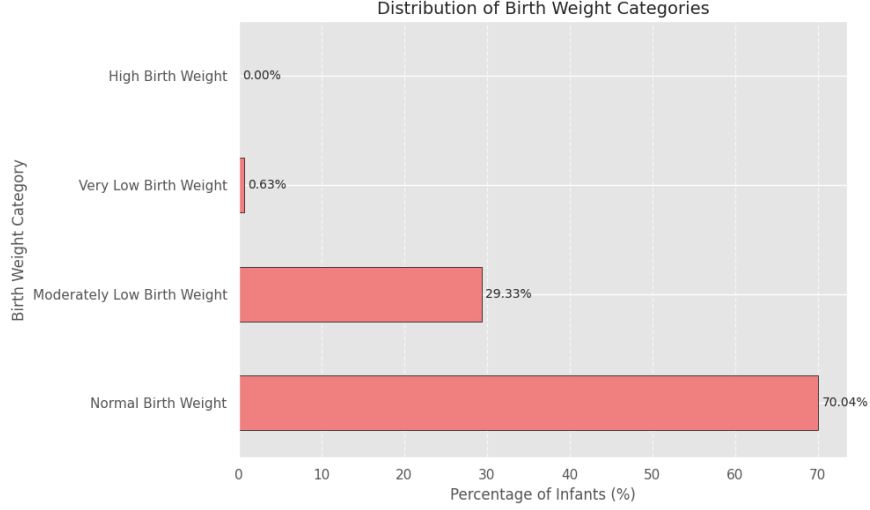


Figure 3: BW Classification on Basis of WHO Standards

## 4.2 Imputation and Feature Selection

The Multiple Imputation by Chained Equations (MICE) method performs optimally for managing missing data during feature selection and model evaluation. Both supervised and unsupervised feature selection techniques are applied, with SHAP (for feature importance) and Gradient Boosting Regression (as base regressor) playing key roles in identifying influential features. The best feature selector models are either GBR or modified version of GBR. Tree based approaches as well as Mutual Information(MI), MARS and L1 regularisation based feature selection techniques performed better as well.

Feature selection identifies critical categories, including body size, dietary intake, biochemistry, and delivery outcomes. Features from the prior study—such as *"f0\_m\_plac\_wt"*, *"f0\_m\_GA\_Del"*, *"f0\_m\_abd\_cir\_v2"*, and others—demonstrate a significant impact on model performance, underscoring the value of incorporating prior knowledge in medical prediction models. The identification of paternal features such as platelet count and head circumference states the distinction of this study from previous studies [9].

## 4.3 ML Model Evaluation

Table 2 represents a detailed comparison of regression models, combining various imputation techniques, feature selection methods, and ML algorithms. The optimal combination—MICE imputation, SHAP-based feature selection, and Bayesian Additive Regression Trees (BART)—delivers the highest predictive performance, achieving an R-squared( $R^2$ ) of 0.6572, Mean Squared Error(MSE) of 24,891.37, and Root Mean Squared Error(RMSE) of 157.77.

The second-best performance is obtained with MICE imputation, SHAP-based feature selection, and CatBoost, which yields an  $R^2$  of 0.6454, MSE of 57,083.76, and RMSE of 238.92. Albeit its strong results, CatBoost demonstrates reduced accuracy comparing with BART,

Table 2: Regression Model Evaluation

Supervised Feature Selector	Model	RMSE	R <sup>2</sup>
Shap	BART	157	0.657
Shap	CatBoost	238	0.645
Gradient Boosting	CatBoost	239	0.643
Shap	Gradient Boosting Regression	245	0.625

particularly when features from prior studies are excluded, emphasizing the importance of incorporating domain-specific knowledge into the model.

Fine-tuning further improves CatBoost’s performance (1,748 iterations, depth = 4, learning rate = 0.0075) but does not surpass BART’s predictive accuracy (1,200 iterations, 100 trees). These findings highlight the critical influence of imputation and feature selection techniques on model outcomes and the variability in performance across algorithms.

#### 4.4 Feature Importance

Feature importance is evaluated using SHAP values with both the BART and CatBoost models. In the BART model, "*f0\_m\_plac\_wt*" has the highest importance (17.60%), followed by "*f0\_m\_GA\_Del*" (13.60%) and "*f0\_m\_fundal\_ht\_v2*" (7.20%), as shown in figure 4.

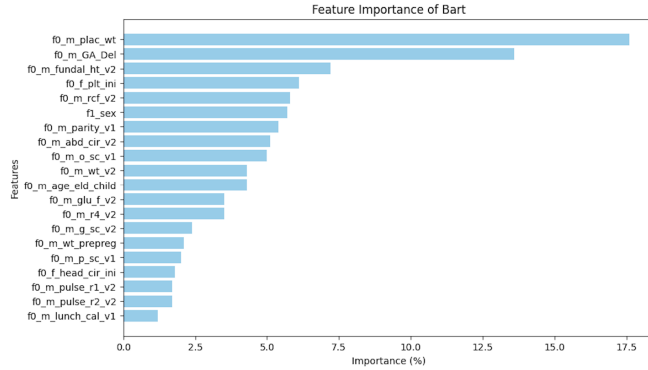


Figure 4: Feature Importance BART

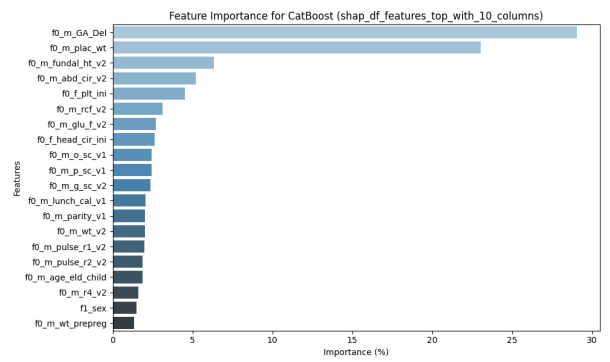


Figure 5: Feature Importance CatBoost

In the CatBoost model, "*f0\_m\_GA\_Del*" exhibits the highest importance (28%), followed by "*f0\_m\_plac\_wt*" (21%), as shown in figure 5. These findings highlight the consistent predictive influence of Gestoplac, which includes gestational age at delivery and placental weight across both models. The results align with established medical literature on the critical role of these factors in determining birth outcomes, further validating the predictive capability of the models (Table 3 ).

While features such as gestoplac consistently ranked as the most important while others, including paternal features, shows lower influence. These findings provide insights into the relative contributions of different factors, emphasizing the centrality of maternal and fetal health parameters in BW prediction.

Table 3: Best Features Categorized by Domain and Count.

Category	Best Features	Original Feature Names	Count
Obstetric History	f0_m_parity_v1	Mother's parity at visit 1	1
Body Size	f0_m_wt_prepreg f0_m_fundal_ht_v2 f0_m_abd_cir_v2 f0_m_wt_v2	Mother's weight (kg) at prepregnancy visit Fundal height (cm) at visit 1 Mother's abdominal circumference (cm) at visit 2 Mother's weight (kg) at visit 1	4
Dietary Intake	f0_m_r4_v2 f0_m_lunch_cal_v1 f0_m_p_sc_v1 f0_m_o_sc_v1	Green chilli consumed by family weekly or monthly Calories during lunch at visit 1 Festival food consumption score : visit 1 Sweet snack consumption score : visit 1	4
Blood Pressure	f0_m_pulse_r1_v2 f0_m_pulse_r2_v2	Mother's pulse (/min) reading 1 at visit 2 Mother's pulse (/min) reading 2 at visit 2	2
Biochemistry	f0_m_glu_f_v2 f0_m_rcf_v2 f0_m_g_sc_v2	Mother's fasting glucose (mg%) at visit 2 Mother's red cell folate (ng/mL) at visit 2 Glv consumption score : visit 1	3
Delivery Outcome	f0_m_plac_wt f0_m_GA_Del	Placental weight recorded by ANM (gm) Mother's gestation age in days of delivery	2
Paternal Features	f0_f_head_cir_ini f0_f_plt_ini	Father's head circumference in cm at initial visit Father's platelet count initial visit	2
Outcomes	f1_sex	-	1
Elder Child Age	f0_m_age_eld_child	Age (yrs) of elder child	1

The Partial Dependence Plots (PDPs) for placental weight (Figure 6) and gestational age (Figure 7) show nonlinear positive relationships with predicted BW. Moderate placental weights (200–300g) and mid-range gestational ages (260–270 days) have the most substantial positive impacts, with predictions stabilizing in these ranges. Low and very high values of both features exhibit higher uncertainty.

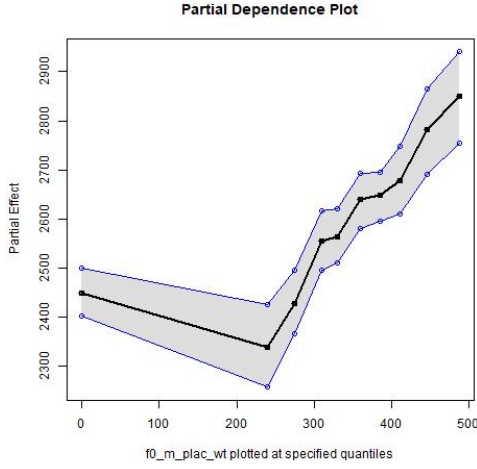


Figure 6: PDP Plot for Placental Weight

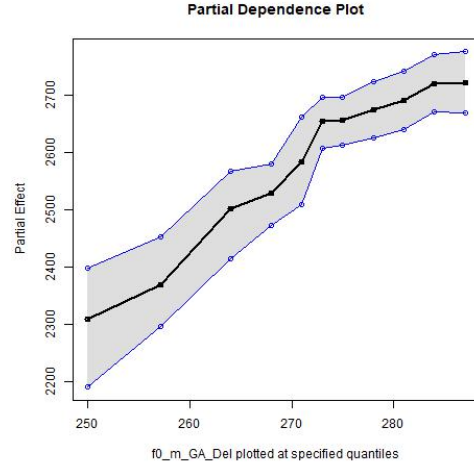


Figure 7: PDP Plot for Gestational Age

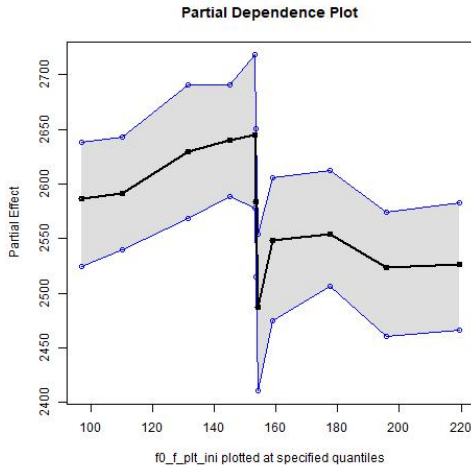


Figure 8: PDP Plot for Paternal Platelet Count

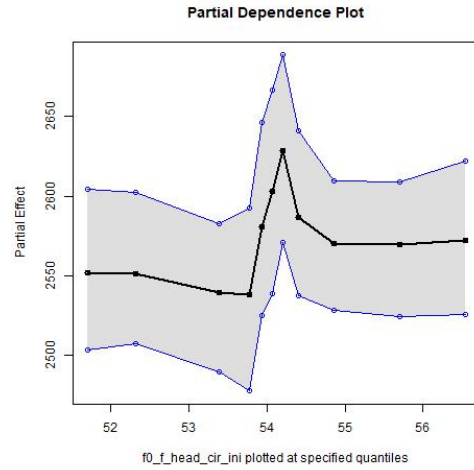


Figure 9: PDP Plot for Paternal Head Circumference

The PDPs for paternal head circumference (Figure 8) and platelet count (Figure 9) highlight complex relationships. Head circumference shows a non-monotonic pattern, with the greatest impact near 54 cm. Platelet count demonstrates an inflection point around 150, transitioning from a positive to a negative effect. Both features provide additional insights, although their influence is more variable compared to maternal factors.

## 4.5 Residual Analysis

Residual analysis focuses on the performance of the BART model to assess prediction accuracy, error margins, and potential biases. The residuals ranges from -600 to 400 grams, with the majority of predictions clustering tightly around actual BW's, demonstrating the

model's robustness.

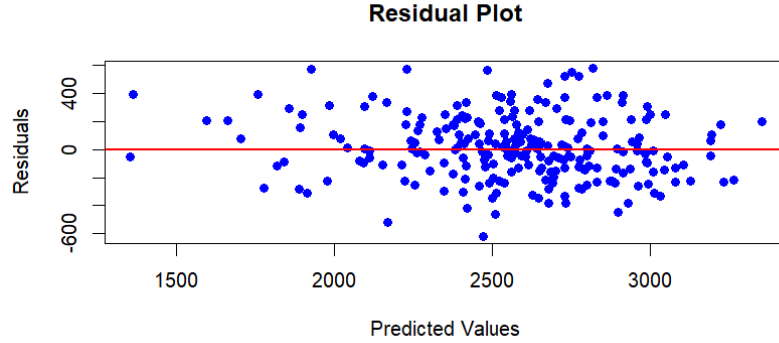


Figure 10: Residual vs predicted scatter Plot for BART

In the critical range of 2,500 to 3,000 grams, a slight increase in residual variability is observed, reflecting minor challenges in predicting mid-range BW's. However, extreme outliers are sparse, and no systematic biases are identified, indicating strong alignment between predicted and actual values. 10 illustrates the residual versus predicted values for the BART model, highlighting close clustering and minimal outliers. The distribution of residuals confirms that the error margins are well-contained, with minimal deviations affecting the reliability of the model.

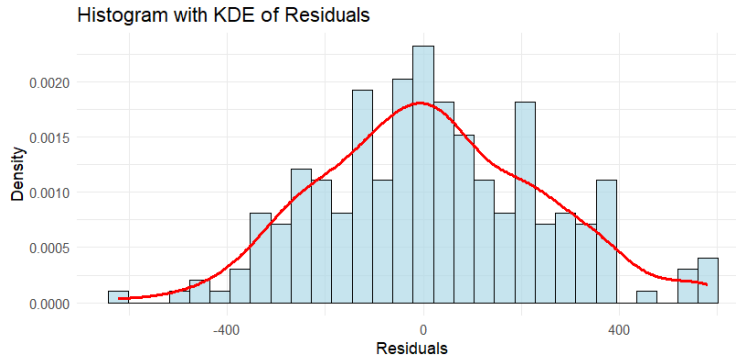


Figure 11: Histogram and KDE Plot for BART Residuals

The histogram and KDE plots (see figure 11) for the BART reveal unimodal, bell-shaped residual distributions that are approximately normal. Distributions exhibit slight positive skewness, indicating a minor tendency for the residuals to lean towards larger positive values. For the BART model, the residual distribution is slightly better centered, demonstrating its robustness in capturing variability across the majority of the data points.

The QQ plots for the BART model, residuals show fewer deviations in the tails, suggesting a higher capacity for managing extreme predictions with minimal error. While a few outliers are present, they are less pronounced (figure 12).

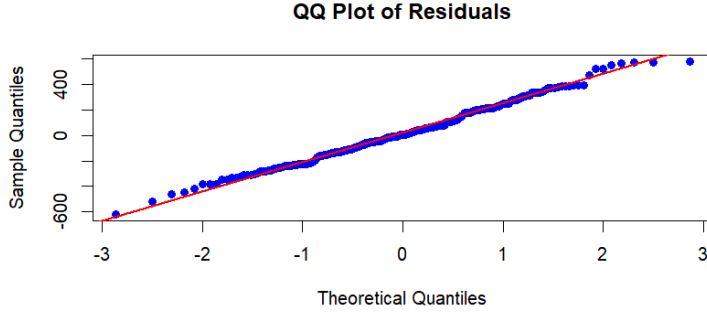


Figure 12: Residual vs. Predicted Values scatter plots for BART

These results validate the BART model’s robustness in capturing variability across the dataset and highlight its effectiveness for clinical applications in fetal BW prediction.

## 5 Discussion

This study builds upon the foundation of a prior investigation that models the fetal BW using fewer maternal features, reaffirming the critical roles of placental weight and gestational age [26]. The earlier findings, grounded in robust biomedical theory, are validated and extended through the inclusion of additional predictors and refined analysis in this study [5].

### 5.1 Significance of Longitudinal Data

While the data reflects conditions from rural India three decades ago, its relevance remains significant due to the large rural population (64% of India) that shares similar socioeconomic conditions even today [11]. While modern advancements in healthcare and lifestyle may limit the direct generalizability to current urban populations, the historical perspective gained from this data is invaluable in understanding hereditary and environmental influences on fetal BW [5].

### 5.2 ML Models and Techniques

SHAP-based importance scores, applied with gradient boosting algorithms, effectively prioritize features from a set of 853 variables, identifying the 20 most relevant predictors of fetal BW [2]. Both BART and CatBoost regression models demonstrate robust nonlinear modeling capabilities, capturing complex feature interactions [25]. The unimodal, near-normal residual distributions validate the models’ predictive alignment, with BART exhibiting superior performance through a more tightly centered distribution and reduced tail deviations [27].

### 5.3 Feature Relevance and Inter Relationships

The study underscores the clinical significance of key predictors such as mother’s placental weight, father’s head circumference, gestoplac, pulse, etc. Placental weight, which demon-



strated the highest feature importance at 17.6%, reflects its critical role in nutrient and oxygen transfer to the fetus [28]. The results align with the *Placental Efficiency Theory* and *Barker’s hypothesis*, emphasizing the placenta’s central role in fetal programming [15]. Similarly, gestational age, with an importance of 13.6%, exhibited a strong monotonic relationship with BW, highlighting its role in determining fetal maturity and growth potential. Fundal height, contributing 7.2% to the prediction, serves as a practical measure of uterine growth and amniotic fluid volume, aiding in the early identification of high-risk pregnancies. Together, these features represent a hierarchical yet interconnected framework of fetal growth regulation [16, 17]. Maternal factors such as pre-pregnancy weight, dietary habits, and blood pressure, sugar level also play a role in BW predictions. Fetus sex also has a significant effect on BW, with male fetus being heavier than female counterparts [3].

The inclusion of novel paternal factors, such as platelet count and head circumference, provides valuable insights into paragenetics and epigenetic influences on fetal growth [6]. Paternal platelet count affects fetal development through platelet-derived growth factor (PDGF) signaling, which induces epigenetic modifications that enhance nutrient utilization and placental development [6]. Similarly, paternal head circumference is linked to growth-promoting genes and hormonal factors, such as insulin-like growth factors (IGFs), which improve fetal metabolic efficiency [6]. These findings illustrate the complex interplay of maternal, paternal, and environmental factors in shaping fetal growth trajectories.

## 5.4 Novelty

The study’s novelty lies in its ability to achieve accurate BW predictions using just 20 features, maintaining an error margin of approximately 200 grams. This compact yet effective feature set, paired with the interpretability of PDP analyses, emphasizes the practical utility of the model in clinical and resource-constrained environments. The findings suggest potential applications in prenatal care, including personalized warnings for at-risk pregnancies based on paternal and maternal factors.

## 5.5 Limitations

Although its strengths, this study has notable limitations that warrant careful consideration to provide a balanced discussion. One key limitation is the absence of sonographic data [31] and proportional measurements, which could have significantly enhanced the predictive modeling by introducing more detailed biometric parameters. Relying solely on tabular data excludes the potential benefits of imaging modalities, such as ultrasound, which can offer deeper insights and augment the accuracy of ML approaches.[2] The integration of imaging data in future work could bridge this gap and provide a more holistic understanding of the factors influencing fetal outcomes [28].

Another notable challenge lies in the issue of missing data within the dataset. This shortfall inevitably affects the robustness and reliability of the analysis, underscoring the importance of comprehensive and well-curated datasets for future research.[29] Moreover, while the study’s findings hold particular relevance for rural healthcare settings, their applicability to urban populations remains less certain. Urban areas exhibit distinct patterns

of healthcare access, socioeconomic conditions, and lifestyle factors, which could limit the generalizability of our results.[9]

Ethical considerations also played a significant role in shaping the study’s methodology. The sensitive nature of prenatal data precluded the use of synthetic augmentation techniques, a common approach to addressing data scarcity. While this decision underscores our commitment to ethical research practices, it also highlights the pressing need for the collection of representative and ethically sourced data to unlock the full potential of machine learning in maternal and fetal health.

## 5.6 Future Studies and Directions

Future directions involve expanding this study’s framework to support early detection of fetal growth-related conditions and enhance prenatal care through personalized risk assessments. Integrating multimodal data, such as imaging, clinical records, and genetic profiles, could improve model accuracy and interpretability [31]. These advancements may serve as decision-support tools for clinicians, aiding in timely interventions and tailored healthcare strategies [34]. Furthermore, the application of ML in the biomedical domain holds potential to bridge gaps in resource-limited settings, providing accessible and data-driven insights to healthcare professionals and patients alike[36].

## 6 Conclusion

This study represents a significant advancement in the prediction of fetal BW, using a compact and interpretable set of 20 features derived from an initial pool of 853 variables. By integrating maternal, paternal, and environmental factors, it presents a holistic framework for understanding fetal growth dynamics. The results validate the pivotal roles of gestoplac ie placental weight, gestational age, and novel paternal factors, such as head circumference and platelet count, highlighting their interdependent contributions to fetal development. Working with state-of-the-art ML models like BART and CatBoost, the study shows robust predictive performance with an error margin of approximately 200 grams. The integration of SHAP-based feature importance and partial dependence analysis improves the interpretability of the models, making them practical for clinical use. This approach underscores the potential of data-driven methodologies to transform prenatal care, particularly in resource-constrained environments. Although its strengths, the study also acknowledges key limitations, such as the absence of sonographic data and challenges posed by missing entries. These limitations indicate essential areas for future research, such as the incorporation of multimodal data and the development of more comprehensive datasets. The ethical considerations that underpin this work also underscore the need to balance technological advancement with the responsible application of sensitive prenatal data. This study not only advances the determinants of BW but also lays the foundation for personalized prenatal care strategies. By bridging gaps in data accessibility and interpretability, it paves the way for future innovations in maternal and fetal health, fostering equitable and data-informed healthcare solutions.

## References

- [1] M. Mursil, H. A. Rashwan, P. Cavallé-Busquets, L. A. Santos-Calderón, M. M. Murphy, and D. Puig, “Maternal nutritional factors enhance birthweight prediction: A super learner ensemble approach,” *Information*, vol. 15, no. 11, 2024. [Online]. Available: <https://www.mdpi.com/2078-2489/15/11/714>
- [2] J. Tao, Z. Yuan, L. Sun, K. Yu, and Z. Zhang, “Fetal birthweight prediction with measured data by a temporal machine learning method,” *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 26, 2021. [Online]. Available: <https://doi.org/10.1186/s12911-021-01388-y>
- [3] Y. Lu, X. Fu, F. Chen, and K. K. L. Wong, “Prediction of fetal weight at varying gestational age in the absence of ultrasound examination using ensemble learning,” *Artificial Intelligence in Medicine*, vol. 102, p. 101748, 2020. [Online]. Available: <https://doi.org/10.1016/j.artmed.2019.101748>
- [4] T. B. Reza and N. Salma, “Prediction and feature selection of low birth weight using machine learning algorithms,” *Journal of Health, Population and Nutrition*, vol. 43, p. 157, 2024. [Online]. Available: <https://doi.org/10.1186/s41043-024-00647-8>
- [5] J. Allotey, L. Archer, K. Snell, D. Coomar, J. Masse, L. Sletner, H. Wolf, G. Daskalakis, S. Saito, W. Ganzevoort, A. Ohkuchi, H. Mistry, D. Farrar, F. Mone, J. Zhang, P. Seed, H. Teede, F. Da Silva Costa, A. Souka, M. Smuk, S. Ferrazzani, S. Salvi, F. Prefumo, R. Gabbay-Benziv, C. Nagata, S. Takeda, E. Sequeira, O. Lapaire, J. Cecatti, R. Morris, A. Baschat, K. Salvesen, L. Smits, D. Anggraini, A. Rumbold, M. van Gelder, A. Coomarasamy, J. Kingdom, S. Heinonen, A. Khalil, F. Goffinet, S. Haqnawaz, J. Zamora, R. Riley, and S. Thangaratinam, “Development and validation of a prognostic model to predict birth weight: individual participant data meta-analysis,” *BMJ Medicine*, vol. 3, no. 1, Aug. 2024.
- [6] A. Libretti, F. Savasta, A. Nicosia, C. Corsini, A. D. Pedrini, L. Leo, A. S. Laganà, L. Troia, M. Dellino, R. Tinelli *et al.*, “Exploring the father’s role in determining neonatal birth weight: A narrative review,” *Medicina*, vol. 60, p. 1661, 2024. [Online]. Available: <https://doi.org/10.3390/medicina60101661>
- [7] Q. Wu, H.-Y. Zhang, L. Zhang, Y.-Q. Xu, J. Sun, N.-N. Gao, X.-Y. Qiao, and Y. Li, “A new birthweight reference by gestational age: A population study based on the generalized additive model for location, scale, and shape method,” *Frontiers in Pediatrics*, vol. 10, p. 810203, 2022. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/35386253/>
- [8] D. Varma, C. S. Yajnik, A. Thorave, and N. Sharma, “Handling missing data in longitudinal anthropometric data using multiple imputation method,” *Data Management, Analytics and Innovation*, 2024. [Online]. Available: <https://easychair.org/publications/preprint/vbF7>

- [9] N. Kapure, H. Joshi, R. Mistri, P. Kumari, M. Mali, S. Purohit, N. Sharma, M. Panday, and C. S. Yajnik, “Predicting fetal birthweight from high-dimensional data using advanced machine learning,” *arXiv*, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2502.14270>
- [10] Z. Liu, N. Han, T. Su, Y. Ji, H. Bao, S. Zhou, S. Luo, H. Wang, J. Liu, and H.-J. Wang, “Interpretable machine learning to identify important predictors of birth weight: A prospective cohort study,” *Frontiers in Pediatrics*, vol. 10, p. 899954, 2022. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/36440327/>
- [11] Macrotrends, “Rural population - india,” 2024, accessed: 2024-12-04. [Online]. Available: <https://www.macrotrends.net/global-metrics/countries/ind/india/rural-population>
- [12] R. Wahab, J. Scholing, and R. Gaillard, “Maternal early pregnancy dietary glycemic index and load, fetal growth, and the risk of adverse birth outcomes,” *European Journal of Nutrition*, vol. 60, no. 3, pp. 1301–1311, 2021. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/32666314/>
- [13] H. Okubo, S. Crozier, N. Harvey *et al.*, “Maternal dietary glycemic index and glycemic load in early pregnancy are associated with offspring adiposity in childhood: the southampton women’s survey,” *American Journal of Clinical Nutrition*, vol. 100, no. 2, pp. 676–683, 2014. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/24944056/>
- [14] F. Atkinson, K. Foster-Powell, and J. Brand-Miller, “International tables of glycemic index and glycemic load values 2021: a systematic review,” *American Journal of Clinical Nutrition*, vol. 114, no. 5, pp. 1625–1632, 2021. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/34257626/>
- [15] M. Janthanaphan, O. Kor-Anantakul, and A. Geater, “Placental weight and its ratio to birth weight in normal pregnancy at songkhlanagarind hospital,” *Journal of the Medical Association of Thailand*, vol. 89, no. 2, pp. 130–137, 2006. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/16623040/>
- [16] E. Prior, S. Uthaya, and E. Harding, “Predicting birth weight at booking,” *BMJ Medicine*, vol. 3, p. e001018, 2024. [Online]. Available: <https://bmjmedicine.bmj.com/content/3/1/e001018>
- [17] C. Li, Y. Peng, B. Zhang *et al.*, “Birth weight prediction models for the different gestational age stages in a chinese population,” *Scientific Reports*, vol. 9, p. 10834, 2019. [Online]. Available: <https://www.nature.com/articles/s41598-019-47056-0>
- [18] I. Papastefanou, U. Nowacka, A. Syngelaki, V. Dragoi, G. Karamanis, D. Wright, and K. Nicolaides, “Competing-risks model for prediction of small-for-gestational-age neonate from estimated fetal weight at 19–24 weeks’ gestation,” *Ultrasound in Obstetrics & Gynecology*, vol. 57, no. 6, pp. 917–924, 2021. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/33955292/>

- [19] S. Mizuno, S. Nagaie, G. Tamiya *et al.*, “Establishment of the early prediction models of low-birth-weight reveals influential genetic and environmental factors: a prospective cohort study,” *BMC Pregnancy and Childbirth*, vol. 23, p. 628, 2023. [Online]. Available: <https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-023-05919-5>
- [20] G. Cohen, H. Shalev-Ram, H. Schreiber *et al.*, “Factors affecting clinical over and underestimation of fetal weight: A retrospective cohort,” *Journal of Clinical Medicine*, vol. 11, no. 22, p. 6760, 2022. [Online]. Available: <https://www.mdpi.com/2077-0383/11/22/6760>
- [21] N. D’souza, R. V. Behere, B. Patni, M. Deshpande, D. Bhat, A. Bhalerao, S. Sonawane, R. Shah, R. Ladkat, P. Yajnik, S. K. Bandyopadhyay, K. Kumaran, C. Fall, and C. S. Yajnik, “Pre-conceptional maternal vitamin b12 supplementation improves offspring neurodevelopment at 2 years of age: Priya trial,” *Frontiers in Pediatrics*, vol. 9, 2021. [Online]. Available: <https://doi.org/10.3389/fped.2021.755977>
- [22] G. E. Batista and M. C. Monard, “A study of k-nearest neighbour as an imputation method,” in *Proceedings of the International Conference on Health Information Science*, 2002. [Online]. Available: [https://www.researchgate.net/publication/220981745\\_A\\_Study\\_of\\_K-Nearest\\_Neighbour\\_as\\_an\\_Imputation\\_Method](https://www.researchgate.net/publication/220981745_A_Study_of_K-Nearest_Neighbour_as_an_Imputation_Method)
- [23] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: what is it and how does it work?” *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011. [Online]. Available: <https://doi.org/10.1002/mpr.329>
- [24] E. Liu, P. X. Lin, Q. Wang, and K. C. Feng, “Feature selection approaches for newborn birthweight prediction in multiple linear regression models,” *arXiv preprint arXiv:2411.11167*, 2024. [Online]. Available: <https://arxiv.org/abs/2411.11167>
- [25] D. A. Alabbad, S. Y. Ajibi, R. B. Alotaibi, N. K. Alsqer, R. A. Alqahtani, N. M. Felemban, A. Rahman, S. S. Aljameel, M. I. B. Ahmed, and M. M. Youldash, “Birthweight range prediction and classification: A machine learning-based sustainable approach,” *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 770–788, 2024. [Online]. Available: <https://doi.org/10.3390/make6020036>
- [26] X. Bai, Z. Zhou, M. Su, Y. Li, L. Yang, K. Liu, H. Yang, H. Zhu, S. Chen, and H. Pan, “Predictive models for small-for-gestational-age births in women exposed to pesticides before pregnancy based on multiple machine learning algorithms,” *Frontiers in Public Health*, vol. 10, p. 940182, 2022. [Online]. Available: <https://doi.org/10.3389/fpubh.2022.940182>
- [27] J. Hill, A. R. Linero, and J. S. Murray, “Bayesian additive regression trees: A review and look forward,” *Annual Review of Statistics and Its Application*, vol. 7, pp. 251–278, 2020. [Online]. Available: <https://doi.org/10.1146/annurev-statistics-031219-041110>

- [28] S. M. Leyto and K. U. Mare, “Association of placental parameters with low birth weight among neonates born in the public hospitals of hadiya zone, southern ethiopia: An institution-based cross-sectional study,” *International Journal of General Medicine*, vol. 15, pp. 5005–5014, 2022. [Online]. Available: <https://doi.org/10.2147/IJGM.S354909>