ORIGINAL RESEARCH ARTICLE

Development of machine learning models for fetal birth weight prediction: A single-center study in a Turkish population

DOI: 10.29063/ajrh2026/v30i20.4

Usame O. Osmanoglu[1], Ahmet Dursun[2] and Ozlem Dulger[3]*

Karamanoğlu Mehmetbey University, Faculty of Medicine, Department of Biostatistics, Karaman, Türkiye[1]; Karamanoğlu Mehmetbey University, Faculty of Medicine, Department of Anatomy, Karaman, Türkiye[2]; Karamanoğlu Mehmetbey University, Faculty of Medicine, Department of Obstetrics and Gyneacology, Karaman, Türkiye[3]

***For Correspondence**: Email: *usa.osm@gmail.com*

Abstract

Fetal birth weight (BW) is a critical indicator of the prognosis of perinatal health. Accurate estimation of BW is essential in determining the appropriate delivery method. This single-center study aims to evaluate the performance of various machine learning (ML) algorithms in estimating BW. The study population consisted of 456 pregnant women with singleton pregnancies evaluated at a single tertiary center in the Karaman region of Turkiye between 38 and 41 weeks based on their last menstrual period, who gave birth to a phenotypically normal live newborn within 7 days after the ultrasound examination. Seven out of the nine models developed using ML algorithms had the lowest mean absolute error (MAE) values ranging from 184.30 to 190.00 grams. The formula closest to these values was Hadlock III, with an MAE value of 223.33 grams. In our study, the model demonstrating the highest predictive accuracy was provided by Linear Regression algorithm with an MAE value of 184.30 grams. Our findings indicate that data-driven predictive modeling, specifically locally-trained Linear Regression, can accurately estimate BW and may outperform generalized traditional formulas *(Afr J Reprod Health 2026; 30 [20]: 44-58).*

Keywords: Fetal Birth Weight, Machine Learning, Obstetrical, Prediction, Pregnancy, Abdominal circumference

Résumé

Le poids à la naissance (BW) est un indicateur essentiel du pronostic de la santé périnatale. Une estimation précise du BW est indispensable pour déterminer le mode d'accouchement approprié. Cette étude monocentrique vise à évaluer les performances de divers algorithmes d'apprentissage automatique (ML) dans l'estimation du BW. La population étudiée était composée de 456 femmes enceintes présentant une grossesse unique, évaluées dans un centre tertiaire de la région de Karaman en Turquie entre 38 et 41 semaines de grossesse (calculées à partir de la date de leurs dernières règles), et qui ont donné naissance à un nouveau-né vivant et phénotypiquement normal dans les 7 jours suivant l'échographie. Sept des neuf modèles développés à l'aide d'algorithmes d'apprentissage automatique présentaient les valeurs d'erreur absolue moyenne (MAE) les plus faibles, comprises entre 184,30 et 190,00 grammes. La formule la plus proche de ces valeurs était Hadlock III, avec une valeur MAE de 223,33 grammes. Dans notre étude, le modèle présentant la plus grande précision prédictive était celui issu de l'algorithme de régression linéaire, avec une valeur de MAE de 184,30 grammes. Nos résultats indiquent que la modélisation prédictive basée sur les données, en particulier la régression linéaire entraînée localement, permet d'estimer avec précision le BW et peut surpasser les formules traditionnelles généralisées. *(Afr J Reprod Health 2026; 30 [20]:44-58).*

Mots-clés: Poids fœtal à la naissance, apprentissage automatique, obstétrique, prédiction, grossesse, circonférence abdominale

Introduction

Various regression modeling methods have been employed to estimate fetal birth weights (BW) through ultrasonographic measurements in recent years. Determining BW is crucial as it significantly predicts neonatal morbidity and mortality.[1] Fetal BW is a critical indicator of the prognosis of perinatal health. Accurate estimation of BW is a cornerstone of reproductive health, yet it remains a significant challenge in resource-limited settings, particularly in African countries where neonatal morbidity is high.[5] The reliance on generalized, Western-centric formulas often leads to diagnostic inaccuracies. There is an urgent need for accessible, data-driven methodologies that allow regional clinics to develop their own population-specific predictive tools using basic ultrasonographic

parameters.[22] Compared to normal-weight infants (born between 2500-4000 grams), low BW (<2500 grams) or macrosomic infants (>4000 grams) face higher risks of perinatal and postnatal complications. Low BW infants are associated with various conditions such as seizures, respiratory distress, hypoglycemia, hyperbilirubinemia, polycythemia, thrombocytopenia, and necrotizing enterocolitis.[2] Similarly, infants with high BW are at risk for potential complications, including shoulder dystocia, intrapartum asphyxia, trauma, various maternal complications, and specific metabolic issues.[3]

BW cannot be directly measured before delivery and is typically estimated based on clinicians' ultrasonographic measurement parameters. Estimating fetal and BW through ultrasonography has become integral to routine obstetrics. Most BW estimation models are derived from applying regression analysis to the ultrasonographic measurement results of pregnant women. However, significant discrepancies between estimated BW and observed BW persist. This issue arises because the regression function closely fits the data according to a specific mathematical criterion and often struggles to account for outliers that deviate from the mean value of ultrasonographically measured fetal measurements. This characteristic of regression introduces a methodological limitation in determining more accurate formulas for BW.[1]

While ML methods offer advanced computational frameworks for fetal weight estimation, their effectiveness is not inherent to the algorithms themselves but is heavily dependent on the quality and diversity of the data used for training. Rather than being a universal solution, ML should be viewed as a data-driven approach that requires rigorous validation. Without representative datasets that reflect diverse demographics, these models risk reproducing the inherent biases or 'blind spots' of the narrow settings in which they were developed. Mohammadi et al. used an artificial neural network to estimate the weight of twin fetuses.[4] Feng et al. employed a deep belief network (DBN) based on various ultrasound parameters to increase the accuracy of BW estimation and help clinicians identify potential risks before birth.[5]

Traditionally, the estimation of BW in Turkiye is based on various regression models of ultrasound measurements that use multiple parameters developed by foreign academics. Due to individual differences in different populations, using these methods to estimate the BW of fetuses in Turkiye can lead to errors. In this study, we evaluated the commonly used fetal parameters—abdominal circumference (AC), femur length (FL), and biparietal diameter (BPD)—obtained from ultrasound measurements taken within the last 7 days before delivery in a cohort of pregnant women from the Karaman region. Our primary hypothesis was that locally trained, data-driven predictive models would yield significantly lower error rates in estimating birth weight for this specific demographic compared to standard, generalized historical formulas (e.g., Hadlock, Shepard). To test this hypothesis, we developed and evaluated various ML algorithms, aiming to determine whether a locally derived regression model could provide more precise, population-specific estimations and address the inaccuracies inherent in using foreign population-based formulas.

Although many traditional fetal weight estimation models have been developed based on Western populations, their accuracy often diminishes when applied to different global demographics due to genetic, nutritional, and environmental variations. This discrepancy is particularly critical in resource-limited settings, such as many African countries, where accurate birth weight prediction is vital for reducing neonatal morbidity and mortality. In such settings, deploying low-cost, open-source ML frameworks that rely only on basic ultrasonographic parameters (AC, FL, and BPD) can bypass the need for expensive, proprietary ultrasound software. This study, while utilizing a localized dataset, aims to present a highly adaptable ML framework that can easily be replicated and trained on diverse local populations across different health systems.

Methods

This is a single-center, model-development study conducted as a retrospective analysis of prospectively collected data at Karaman Education and Research Hospital to evaluate the performance of various machine learning algorithms in a specific regional cohort. The study population consisted of 456 women with singleton pregnancies who presented for routine pregnancy follow-up at

Karaman Education and Research Hospital between November 2023 and April 2024. The ultrasonographic and clinical parameters utilized in this study were systematically recorded in the institutional electronic database at the time of examination and delivery.

Only pregnant women who lived in the Karaman region were Turkish citizens and delivered at the same hospital with a gestational age of 38-41 weeks based on their last menstrual period were included in the study. The selection of AC, FL, and BPD as final predictors was guided by the principle of parsimony and the clinical goal of developing a streamlined, biometric-only predictive model. Other potential variables, such as gestational age, fetal sex, and maternal demographic factors (e.g., BMI, parity), were considered during the preliminary design phase. However, these were excluded to minimize the risk of overfitting on a single-center dataset and to ensure the model's clinical feasibility in resource-limited or high-volume settings where additional maternal data may be incomplete. Furthermore, parameters such as Head Circumference (HC) were omitted to prevent multicollinearity, as HC demonstrated a high degree of correlation with BPD in our cohort ($r>0.90$). By focusing on the 'core three' biometrics, we ensured that the resulting models remained robust, interpretable, and directly comparable to existing standard formulas.

For each patient, maternal age, sex, gestational age according to the last menstrual period, and commonly used ultrasound parameters, including AC, FL, and BPD, were measured and recorded. Cases with an interval of 7 days or less between the ultrasound measurement date and the delivery date were included in the study. Ultrasound measurements were performed once by a single obstetrician with 23 years of experience in ultrasound measurements. Each fetus was measured only once and included in the study. BPD was measured using a convex probe (Exp/SC5-1e /1-5 MHz) on a Mindray DC-8 ultrasound system at the level of the cavum septum pellucidum and falx cerebri, with the calipers placed from outer to inner. AC was measured at the level where the umbilical vein passes through the liver, ensuring symmetrical visualization of the ribs. FL was measured vertically from metaphysis to metaphysis. Newborns were weighed immediately after birth using an electronic scale, and the BW was recorded in the patient file in grams. Fetal sex was determined by ultrasound when possible and confirmed after birth.

Pregnant women with comorbid conditions that could affect fetal growth (such as diabetes mellitus, hypertension, renal disease, or thyroid disease) and those with known pregnancy complications (such as bleeding or preeclampsia) at the time of ultrasound screening were excluded from the study. Pregnant women with detected fetal malformations during ultrasound examination were also excluded. Patients with a history of obstetric complications were not included in the study.

Sample size adequacy and overfitting mitigation

The adequacy of the sample size (n=456) was evaluated based on the feature-to-sample ratio. Given that our models relied on three primary biometric predictors (AC, FL, BPD), the ratio of approximately 152 observations per predictor is well above the recommended minimum for stable machine learning modeling. To further mitigate the risk of overfitting, we implemented several strategies. Firsly, utilizing shallow architectures for the ANN and ensemble models to prevent excessive parameterization. Secondly, employing 5-fold cross-validation to ensure internal stability. And finally, assessing final performance on an independent 20% test set. The high degree of agreement between training and testing performance metrics indicates that the sample size was sufficient to capture the biological variance in fetal weight without overfitting the data.

Model development and training

All artificial intelligence processes and graphics were conducted using Python version 3.7.9 (Delaware, USA, 2020). This study, the dependent variable is the BW of the fetus. The independent numerical variables are the AC, FL, and BPD parameters. Predictions were made using supervised ML algorithms. Nine different predictions were made using Artificial Neural Networks (ANN), Decision Trees (DT), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Linear Regression (LR), Lasso Regression (LAR), Ridge Regression (RR), and Elastic Net Regression (ENR) algorithms to determine the performance of these methods. To

ensure the reproducibility of our machine learning pipeline, all data processing, algorithmic training, and visualizations were conducted using Python. Pandas (v1.3.5), numpy (v1.21.6), matplotlib (v3.5.3), seaborn (v0.12.2), statistics (v0.13.5), sklearn (v1.0.2) and keras (v2.11.0) libraries were used in Python coding. The importance of independent variables that had a statistically significant effect on the dependent variable was calculated using the Gradient Boosting Regressor model. Regarding data preprocessing, missing data handling and outlier review were primarily managed through strict clinical inclusion and exclusion criteria prior to the algorithmic phase, rather than utilizing statistical imputation. For model configuration, a significant methodological decision was made to assess the fundamental, baseline predictive capability of the algorithms. Therefore, no special hyperparameter adjustment or complex tuning (e.g., Grid Search) was performed. All nine machine learning algorithms, including ANN and LR, were implemented using their default settings provided within the sklearn and keras libraries. This approach was intentionally selected to prevent the algorithms from overfitting to our specific, single-center dataset and to demonstrate the robust, 'out-of-the-box' utility of these methods for potential clinical application in resource-limited settings where complex algorithmic tuning may not be feasible.

In the development phase, nine different algorithms were trained using fetal biometric parameters. The training process focused on identifying the optimal weights for each feature within our specific regional cohort. The workflow diagram is presented in Figure 1. All of the prediction models used in the study are machine learning algorithms and no traditional models were used.

Statistical analysis

Continuous variables were assessed for normal distribution and are expressed as mean ± standard deviation (SD). Categorical variables are expressed as frequencies and percentages. Relationships between variables are calculated using Pearson correlation analysis. The interpretation of correlation coefficients was guided by the criteria proposed by Schober et al., (2018): 0.00–0.09: no relationship; 0.10–0.39: mild association; 0.40–0.69: moderate association; 0.70–0.89: significant association; >0.90: extremely high association.[6] To evaluate intra-observer reliability, a subset of 30 randomly selected cases was re-evaluated. The same experienced senior clinician, who was blinded to the initial results, repeated the ultrasonographic measurements (AC, FL, and BPD) with a two-week interval between assessments to minimize recall bias. Statistical consistency was assessed using the Intraclass Correlation Coefficient (ICC) based on a two-way mixed-effects model. ICC interpretation drew on suggested criteria:[7] low agreement; 0.5–0.75: moderate agreement; 0.75–0.9: fair agreement; >0.90: excellent agreement. The agreement between predictions and observed results was evaluated using Bland-Altman plots and ICC. The level of statistical significance was set at $p<0.05$.

Performance evaluation

Mean absolute error (MAE) characterizes the variation between observed (Y_i) and predicted ($\hat{Y}_i$) values and is derived as the mean of the total variation of the dataset. MAE is calculated by dividing the sum of absolute errors (i.e., Manhattan distance) by the sample size:[8]

$$MAE = \frac{\sum_i^n |Y_i - \hat{Y}_i|}{n}$$

Root mean squared error (RMSE) is one of the two closely related and frequently used measures that quantify the difference between observed and predicted values. The RMSE of a sample is the square root of the average of squared differences between observed and predicted values:[8]

$$RMSE = \sqrt{\frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n}}$$

Mean absolute percentage error (MAPE) is a measure of prediction accuracy in statistics. It is often used as a loss function in regression problems and model evaluation due to its intuitive interpretation in terms of relative error. MAPE measures the average magnitude of errors produced by a model or how far predictions are, on average, from the observed values:[8]

$$MAPE = \frac{\sum_i^n \left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|}{n} x100$$

Mean error (ME) is the average difference between observed and predicted values, calculated as the

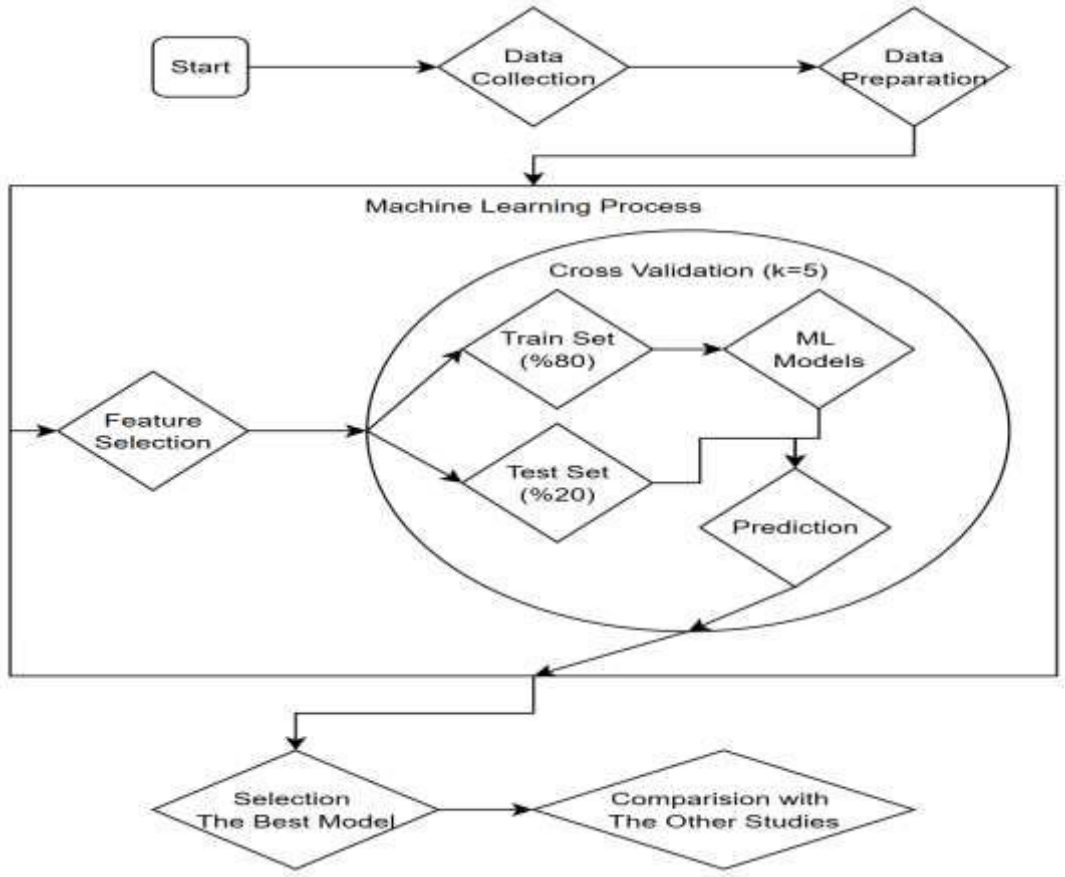


Figure 1: Systematic workflow of the machine-learning-based fetal weight prediction process

sum of errors divided by the number of observations, indicating model bias:[8]

$$ME = \frac{\sum_i^n (Y_i - \hat{Y}_i)}{n}$$

Max error (MaxE) is a metric that represents the largest absolute difference between the predicted and observed values, measuring the worst-case performance of the model. Coefficient of Determination (R^2): Indicates the proportion of the variance in the dependent variable (birth weight) that is predictable from the independent variables (AC, FL, BPD).[8] The success rate in predicting a continuously structured dependent variable can be evaluated based on error terms. Performances were evaluated by comparing the parameters of MAE, MAPE, RMSE, ME and MaxE. The model with the lowest MAE value was considered the model demonstrating the highest predictive accuracy.

Validation strategy

To ensure robust model evaluation and to prevent the variance associated with a single train-test split, a 5-fold cross-validation strategy was implemented. The entire dataset (n=456) was randomly partitioned into five equal subsets (folds). In each iteration, four folds (80%) were used to train the algorithms, while the remaining one-fold (20%) served as the validation set. This process was repeated five times, allowing every data point to be used for validation exactly once.

The final performance metrics reported in the Results section and associated tables (e.g., MAE, RMSE, and R^2) represent the calculated averages of these five independent iterations. This cross-validation approach maximizes the utility of our sample size and provides a highly stable and unbiased estimate of the models' true predictive capabilities.

Comparative analysis

Once the algorithm with the lowest mean error was identified and validated, its predictive accuracy was benchmarked against traditional formulas commonly used in clinical practice, including

Hadlock III and Shepard. The comparison was based on performance metrics such as MAE and RMSE.

Ethical consideration

Ethical approval was obtained from Karamanoglu Mehmetbey University Faculty of Medicine Local Scientific Medical Research Ethics Committee (E-11095095-050.01.04-162429) on 27.11.2023. Participation was voluntary, and participants could withdraw at any time without penalty. Written informed consent was obtained prior to data collection. Ethical principles of respect, beneficence, and justice were upheld, and confidentiality was maintained throughout.

Results

A total of 456 cases were deemed suitable for final analysis. The mean ± standard deviation of BW was 3255.50 ± 397.33 grams. Among the newborns, 50.20% were male and 49.80% were female. The maternal age ranged from 18 to 44 years, averaging 28.02 ± 5.25 years. The gestational age of the fetuses at the time of measurement ranged from 38 to 41 weeks based on the last menstrual period, with an average gestational age of 38.64 weeks. Details regarding maternal age and neonatal measurement results are presented in Table 1.

The BW of the cases in our study ranged from 2195 to 4610 grams. Eleven cases (2.40%) had low BW below 2500 grams. Fourteen cases (3.10%) had high BW above 4000 grams. The BW of 431 cases (94.50%) was between 2500 and 4000 grams, considered normal BW.

Intra-observer reliability analysis demonstrated excellent consistency across all biometric parameters. The calculated ICC values were 0.94 (95% CI: 0.84–0.96) for BPD, 0.91 (95% CI: 0.92–0.98) for AC, and 0.96 (95% CI: 0.82–0.96) for FL. These high coefficients indicate that the biometric data utilized to train the machine-learning algorithms were highly stable and reproducible.

A graph showing the importance of the variables in the prediction model on the dependent variable is provided in Figure 2a. Our study initially found that the most influential parameter was AC, while the parameter with the least effect was FL. The correlation between the parameters used in the study is shown in Figure 2b. The performance results of the BW prediction using artificial intelligence algorithms are given in Table 2. The performance of nine different algorithms in predicting BW was evaluated by comparing parameters such as MAE, MAPE, RMSE, and ME. The model with the lowest MAE value was considered the best model (Table 2). LR algorithm, with an MAE value of 184.30 grams, provided the best prediction performance when Estimation Birth Weight (EBW) was estimated using AC, FL, and BPD parameters. The prediction model created using the LR method is as follows.

$$\boldsymbol{Estimation\ Birth\ Weight} = -\mathbf{3869.31} + \mathbf{28.75} \times (\boldsymbol{BPD}) + \mathbf{21.89} \times (\boldsymbol{FL}) + \mathbf{8.79} \times (\boldsymbol{AC})$$

BW was also estimated using 10 different formulas from the literature that predict BW using our study's measurement parameters. These formulas are listed in Table 3. The LR model, which provided the best result among our ML prediction models, had a MAPE value of 5.67%, making it the most successful method (Table 4). Bland-Altman analysis was performed to evaluate the distribution of differences and the level of agreement between the prediction and observation. The plot below presents the mean difference and 95% CI (Figure 3a). According to the Bland-Altman results, BW value obtained with the prediction model can be found to be 430.88 lower and 482.26 higher than the mean value obtained with the observation.

It can be stated that the differences in BW values of these two methods are randomly distributed around zero (not systematically distributed), so there is a good agreement between the two methods examined according to the Bland-Altman method. Moreover, the mean difference was negative, indicating that the prediction model obtained generally higher values than the observation. ICC was calculated as 0.85 (95% CI: 0.81–0.89), indicating excellent absolute agreement between the estimated and observed birth weights, rather than a mere linear association. As a result of the study, it was determined that the correlation between the predicted and observed birth weight of the LR model, which gave the highest predictive accuracy, was a strong positive relationship ($r=0.80$, $R^2=0.64$) (Figure 3b). In addition, the scattering of the residuals around zero is also shown in Figure 3b.

Table 1: Descriptive statistics of fetal biometric parameters and observed BW (n=456)

Parameter	Mean	Standard Deviation	Median	Q1-Q3	Minimum-Maximum
BW (grams)	3255.50	397.33	3240.00	3000.00-3540.00	2195-4610
AC (mm)	330.67	20.85	332.00	318.25-343.00	223-374
FL (mm)	71.65	3.48	71.00	69.00-74.00	62-93
BPD (mm)	92.32	3.89	92.00	90.00-95.00	80-102
Maternal age (year)	28.02	5.25	27.00	24.00-31.00	18-44

BW: Birth weight (grams); AC: Abdominal circumference (mm); FL: Femur length (mm); BPD: Biparietal diameter (mm)

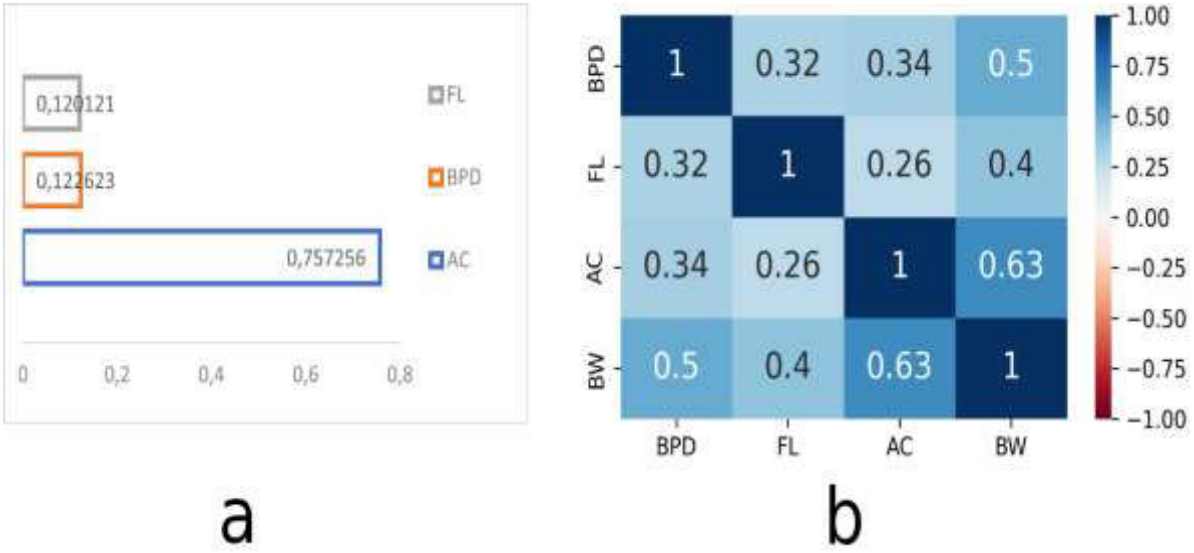


Figure 2: (a) Importance values of variables in the prediction model affecting the fetal biometric parameters, (b) Correlation heatmap plot of the fetal biometric parameters

Table 2: Comparative predictive performance of machine learning algorithms

Algorithms	MAPE (%)	MAE (g)	RMSE (g)	MaxE (g)
LR	5.674	**184.30***	234.35	698
LAR	5.675	184.31	234.36	699
RR	5.675	184.31	234.36	699
ANN	5.675	184.31	234.36	699
ENR	5.675	184.32	234.37	704
RF	5.742	185.13	238.50	783
KNN	5.828	190.00	243.00	760
SVM	8.649	279.92	353.81	905
DT	9.268	299.62	378.29	1010

MAPE: Mean Absolute Percent Error; MAE: Mean Absolute Error; RMSE: Root Mean Squared Error; MaxE: Maximum Error; LR: Linear Regression; SVM: Support Vector Machine; DT: Decision Tree; ENR: Elastic Net Regression; LAR: LASSO Regression; RR: Ridge Regression; ANN: Artificial Neural Network; RF: Random Forest; KNN: K-Nearest Neighbors; *The lowest MAE value.

The Bland-Altman analysis revealed Limits of Agreement (LoA) ranging from [-482.26] to [430.88] g. While these limits may appear wide, they are consistent with the reported margins of error in current obstetric literature for third-trimester ultrasound.[15] Clinically, a discrepancy within this range is critical for the management of macrosomia. Our findings suggest that while the proposed ML model offers high average accuracy, clinicians should interpret individual estimates

Table 3: Established fetal weight estimation formulas used for comparative performance and external validation

Study	Parameters	Formulas
This study	AC, BPD, FL	$-3869.31 + 28.75 * (BPD) + 21.89 * (FL) + 8.79 * (AC)$
Hadlock III [15]	AC, BPD, FL	$10^{(1.335-0.000034*AC*FL+0.00316*BPD+0.00457*AC+0.01623*FL)}$
Shepard et al [16]	AC, BPD	$1000 * 10^{(-1.7492+0.0166*BPD+0.0046*AC-0.00002646*AC*BPD)}$
Merz II [17]	AC, BPD	$-3200.40479 + 15.707186 * AC + 0.1590391 * BPD^2$
Hadlock II [15]	AC, FL	$10^{(1.304+0.005281*AC+0.01938*FL-0.00004*AC*FL)}$
Campbell and Wilkin [18]	AC	$1000 * e^{(-4.564+0.0282*AC-0.0000331*AC^2)}$
Hadlock IV [15]	AC	$e^{(2.695+0.0253*AC-0.0000275*AC^2)}$
Vintzileos et al [19]	AC, BPD	$10^{(1.879+0.0084*BPD+0.0026*AC)}$
Esinler et al. [10]	AC, FL	$1073.4 + 0.016 * AC^2 + 0.371 * FL^2 + 20.187 * (AC/FL) * (AC/FL)$
Warsof et al [20]	AC, BPD	$10^{(1.401+0.0144*BPD+0.0032*AC-0.000000111*AC*BPD^2)}$
Merz I [17]	AC	$AC^3/10000$
Turnau et al [21]	AC, BPD	$-229.076 + 0.09337 * AC * BPD$

AC: Abdominal circumference; FL: Femur length; BPD: Biparietal diameter.

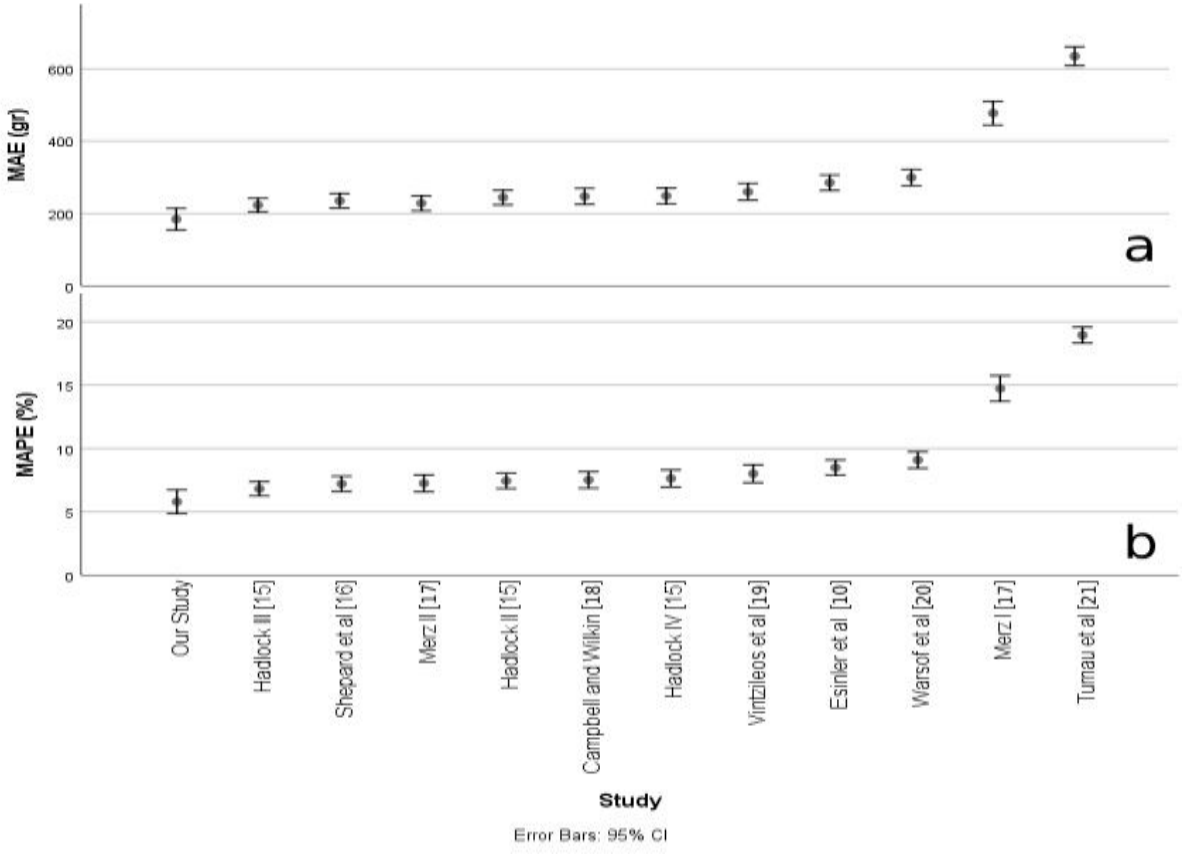


Figure 4: (a) The mean absolute error (MAE) of the proposed model as compared to other methods, (b) The mean absolute percent error (MAPE) of the proposed model as compared to other methods

Table 4: Comprehensive performance metrics of the evaluated models for external validation on the independent cohort (n=456)

Study	Parameter	RMSE (g)	MAPE (%) [95% CI]	MAE (g) [95% CI]	ME (g) [95% CI]	MaxE (g)
This study	AC, FL, BPD	234.35	5.67±4.54 [4.85 to 6.74]	184.30±144.75 [154.17 to 214.45]	-25.69±232.94 [-74.20 to 22.81]	697.72
Hadlock III [15]	AC, BPD, FL	303.46	6.81±6.14 [6.25 to 7.38]	223.33±205.46 [204.40 to 242.26]	106.08±284.32 [79.89 to 132.28]	2340.95
Shepard et al [16]	AC, BPD	319.92	7.21±6.53 [6.60 to 7.81]	235.30±216.75 [215.33 to 255.26]	91.97±306.41 [63.75 to 120.20]	2342.85
Merz II [17]	AC, BPD	320.87	7.25±7.19 [6.59 to 7.91]	228.26±225.51 [207.48 to 249.04]	-95.75±306.25 [-123.98 to -67.50]	3377.89
Hadlock II [15]	AC, FL	328.99	7.44±6.53 [6.83 to 8.04]	244.23±220.42 [223.93 to 264.54]	135.09±299.98 [107.45 to 162.72]	2497.17
Campbell and Wilkin [18]	AC	344.99	7.51±7.10 [6.86 to 8.17]	247.18±240.67 [225.00 to 269.35]	133.85±317.97 [104.55 to 163.14]	3382.33
Hadlock IV [15]	AC	345.96	7.63±7.25 [6.96 to 8.30]	248.59±240.60 [223.43 to 270.76]	106.93±329.02 [76.62 to 137.24]	3265.61
Vintzileos et al [19]	AC, BPD	359.59	8.01±7.63 [7.30 to 8.71]	259.97±248.43 [237.09 to 282.86]	-55.17±355.33 [-87.91 to -22.44]	2530.01
Esinler et al. [10]	AC, FL	367.39	8.49±6.51 [7.89 to 9.09]	285.14±232.16 [263.75 to 306.52]	230.24±286.69 [203.83 to 256.65]	2334.16
Warsof et al [20]	AC, BPD	384.58	9.08±7.10 [8.43 to 9.74]	299.39±241.38 [277.16 to 321.63]	242.43±298.55 [214.92 to 269.93]	2448.15
Merz I [17]	AC	595.77	14.73±11.01 [13.71 to 15.74]	477.06±356.86 [444.19 to 509.94]	-400.83±440.77 [-441.44 to -360.22]	3298.91
Turnau et al [21]	AC, BPD	693.16	18.94±6.90 [18.31 to 19.58]	634.26±279.62 [608.50 to 660.02]	631.76±285.22 [605.48 to -658.04]	2646.02

AC: Abdominal circumference; FL: Femur length; BPD: Biparietal diameter; MAPE: Mean absolute percent error; MAE: Mean absolute error; RMSE: Root mean squared error; MaxE: Maximum error; ME: Mean error; CI: Confidence interval. Means are presented along with standard deviations.

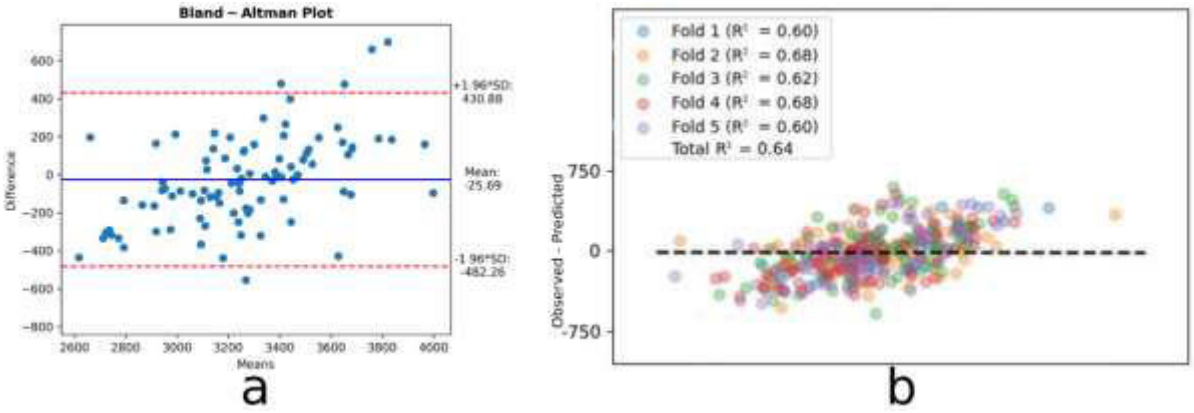


Figure 3: (a) Bland–Altman plot illustrating the agreement between observed and predicted birth weights. The X-axis represents the mean of the observed and predicted values (g), while the Y-axis represents the difference (error) between the two methods (g). The solid horizontal line indicates the mean bias (-25.69 g), representing the systematic overestimation or underestimation. The dashed horizontal lines represent the 95% limits of agreement (Mean Bias ± 1.96 SD), ranging from -482.26 to 430.88 g. Most data points fall within these limits, suggesting [good/excellent] agreement for clinical fetal weight estimation. (b) Residual scatter plot for the optimized Linear Regression model. The X-axis shows the predicted birth weight (g), and the Y-axis shows the residuals (observed minus predicted values in g). The central horizontal line at zero represents the baseline for perfect prediction. The random distribution of residuals around the zero line indicates homoscedasticity, suggesting that the model maintains consistent predictive accuracy and stability across the entire birth weight spectrum without systematic bias at lower or higher weight extremes

within the context of these limits, particularly when the EFW is near the thresholds for elective delivery interventions.

Discussion

This study demonstrates that localized predictive models, specifically calibrated to a regional cohort, significantly enhance the accuracy of fetal birth weight estimation compared to generalized global equations. Our primary finding—that a locally derived linear model and specific machine-learning architectures (e.g., ANN) reduced the mean absolute error (MAE) compared to Hadlock III—highlights the clinical necessity of population-specific tools. These results suggest that the integration of local biometric data can effectively mitigate the systematic biases inherent in universal formulas when applied to diverse obstetric populations.

The fundamental issue in developing an accurate formula is that data from one population may not be suitable for another. From our perspective, the best approach for clinicians is to evaluate the performance of the formulas they choose within their populations. The use of BW formulas should not be arbitrary.[10]

In our study, we kept the time interval between the ultrasound and birth dates to 7 days or less. It has been shown that if this period is extended, the MAPE value of the formulas increases. Basha et al. investigated the effect of the time interval between ultrasound and delivery on MAPE values. They reported that the MAPE for BW estimation by ultrasound (according to the Hadlock I formula) was 6.5% for all babies in their study; however, if the fetus was born within seven days or 8-14 days after the ultrasound, these rates were 6.1% and 8.2%, respectively [11]. This study's maximum interval between the ultrasound and birth dates was seven days. The average interval between the ultrasound and birth dates was 2.76 days.

The decision to restrict the study cohort to uncomplicated term pregnancies (38–41 weeks) with a maximum seven-day interval between ultrasound and delivery was a deliberate methodological choice to ensure high internal validity. This restriction minimizes the impact of fetal growth that occurs between the final scan and birth, which can significantly skew the accuracy of prediction models. While this approach enhances the precision of our localized Linear Regression model, we recognize that it limits its applicability for early obstetric decisions or preterm birth management. Our model should be utilized as a specialized tool for pre-delivery weight assessment, aimed at assisting clinicians in evaluating delivery risks (such as macrosomia) at term. Further research is necessary to evaluate the performance of these machine-learning frameworks in preterm

populations, where early management decisions are critically dependent on accurate weight estimation.

This study employed nine different ML algorithms trained with our dataset using the AC, FL, and BPD parameters from the literature. The reason for using these parameters in our study was that Hammami et al. showed in their 2018 study that these three parameters were the most effective.[12] Attempts to improve BW estimation by adding maternal characteristics such as height, weight, age, and racial origin to fetal biometric measurements have not been successful.[13] A study on more than 9000 singleton pregnancies investigated the effects of maternal age, weight, height, parity, diabetes, fetal sex, presentation, amniotic fluid index, and sonographer experience; despite some factors having a significant effect on EFW, their contributions were found to be small and of questionable clinical significance.[14]

The best model identified among ML algorithms was the LR model. The BW prediction of the LR model using AC, FL, and BPD parameters was compared with the formulas frequently used in our country: Hadlock III,[15] Shepard et al.,[16] Merz II,[17] Hadlock II,[15] Campbell and Wilkin,[18] Hadlock IV,[15] Vintzileos et al.,[19] Esinler et al.,[1] Warsof et al.,[20] Merz I,[17] and Thurnau et al..[21] The comparison results of the literature formulas and our developed ML algorithm are presented in Table 4. The primary accuracy criterion of this study was the lowest MAE value. According to this criterion, the model developed with ML method provided the best result with an MAE of 184.31 grams. The closest formula to our result was the Hadlock III formula, with an MAE of 223.33 grams. The worst result was the Turnau et al. formula with an MAE of 634.26 grams (Figure 4a).

The highest predictive performance of the Linear Regression model in our study suggests that the clinical gain is derived primarily from the use of population-specific data rather than the complexity of the algorithm itself. This supports our initial goal of developing a localized prediction framework to overcome the limitations of universal formulas when applied to distinct demographics.

While our study evaluated a diverse range of algorithms, including more complex models such as ANN, RF, and SVM, the LR model demonstrated the best predictive performance with the lowest MAE of 184.30 grams. This finding is highly significant, as it suggests that the relationship between fundamental fetal biometric parameters (AC, FL, BPD) and birth weight within the 38–41 week gestational window is predominantly linear. In this context, highly sophisticated artificial intelligence methods may be prone to overfitting or unnecessary computational complexity. Conversely, a straightforward, classical linear model—when adequately trained and cross-validated on a local dataset—offers robust, highly accurate predictions without the need for advanced algorithmic infrastructure.

Except for SVM and DT, all other ML algorithms used in our study had lower MAE values than the Hadlock III formula (Table 3). These findings indicate that ML methods can provide better results than traditional methods.In our study, all formulas showed a deviation of over 5% between the observed BW and the predicted BW values. The deviation in our developed formula was 5.67%. Nine out of the eleven formulas we compared showed a deviation of less than 10%. The Merz I and Turnau et al. formulas had deviations of 14.73% and 18.94%, respectively (Figure 4b).[17, 21] For more details, please refer to Table 4.

The proposed Linear Regression model significantly outperformed established global formulas, such as Hadlock, within our cohort. This performance gap suggests that localized calibration accounts for regional biometric variations more effectively, leading to a notable reduction in prediction error that is critical for late-term clinical assessments.

The highest predictive performance of our proposed models compared to established global formulas should be interpreted with caution. While our results show significantly lower error rates, this comparison is primarily an external validation of formulas like Hadlock and Shepard within a specific Turkish term population. The performance gap highlights the biological and regional diversity that universal formulas may fail to capture, rather than serving as a general proof of the absolute robustness of machine-learning algorithms. The inherent advantage of our models stems from their training on the same local demographic in which they were evaluated. Thus, our findings advocate for the development of population-specific clinical tools to mitigate the systematic biases introduced by utilizing generalized global models in diverse regional cohorts.

A strength of our study is the use of prospectively collected data, which ensures that ultrasonographic measurements were recorded systematically during routine clinical workflow, thereby reducing the information bias typically associated with retrospective designs. However, the analysis remains retrospective, and further prospective validation in different cohorts is necessary to confirm the generalizability of the developed models. It is known that intra-observer variability is high in ultrasonographic evaluations. In this study, a senior obstetrician with 23 years of ultrasound experience scanned all fetuses. All fetal biometric measurements were performed by a single experienced senior clinician to minimize intra-observer variability and ensure high procedural consistency. Various ultrasound devices are reported in the literature for scanning. This variability can lead to measurement errors, as the accuracy of different ultrasound devices may vary. This study used a single type of ultrasound device to scan all fetuses.

We acknowledge a crucial limitation regarding the generalizability of our findings. The proposed models were developed and internally validated using a narrow dataset collected solely from the Karaman region. This highlights a broader, fundamental challenge in medical AI: models that demonstrate high accuracy on localized datasets often carry substantial blind spots when applied to broader, more heterogeneous populations. Because our data lacks diverse, multi-center representation, we cannot claim that this algorithm is universally clinically ready for generalized obstetric practice. Before any predictive machine learning system can be safely and broadly integrated into clinical workflows, it must be rigorously trained, tuned, and externally validated on diverse and representative datasets that account for varying racial, socio-economic, and environmental factors. Until such rigorous external validation is performed, our findings should be interpreted primarily as a localized proof-of-concept. It demonstrates that population-specific modeling can be highly effective for a targeted demographic, but it is not a substitute for universally validated clinical tools.

A notable limitation of this study is that the dataset originates from a single center in Türkiye. However, the primary translational value of these findings lies in the methodology. In resource-limited obstetric environments and diverse African health systems, access to advanced ultrasonography infrastructure can be frequently lacking. Our findings demonstrate that simple, computationally efficient ML algorithms, specifically Linear Regression, can accurately predict birth weight using only three fundamental parameters (AC, FL, BPD). This indicates that clinicians in developing regions can seamlessly integrate such algorithms into mobile applications or low-cost portable ultrasound devices. Moreover, rather than relying on generalized traditional formulas like Hadlock — which may not accurately reflect the anthropometric characteristics of African populations— researchers in these regions can adopt the ML framework proposed in this study to train and deploy highly precise, population-specific models using their own local datasets.

While our data were collected in a Turkish center, the implications of this study are highly relevant to sexual and reproductive health priorities in African countries. In environments where high-end ultrasound machines with integrated software are scarce, our findings show that a simple, open-source computational model (such as Linear Regression) can provide high-precision results using only fundamental biometry (AC, FL, BPD). This approach empowers local health systems to move away from 'one-size-fits-all' global formulas and instead adopt localized digital health solutions. Such tools can be integrated into mobile health (mHealth) applications, providing frontline healthcare workers in diverse African settings with a reliable, low-cost method for identifying obstetric risks and improving perinatal outcomes.

The promising results of our localized models must be interpreted with caution. A significant challenge in medical AI is the 'brittleness' of models trained on single-center datasets. Because our study utilized a specific cohort from a single Turkish center, the resulting algorithms may have inherited localized clinical patterns that might not exist in broader populations. For AI systems to achieve true clinical readiness, they must move beyond narrow developmental silos and undergo validation on diverse, multi-center datasets to ensure that they do not merely replicate regional biases but instead offer robust, generalizable predictive power.

A primary limitation of the current study is the use of a highly selected, low-risk cohort consisting solely of uncomplicated singleton term

pregnancies. While this homogeneity allowed us to assess the performance of locally-trained machine learning models without the confounding influence of maternal or fetal pathologies, it inherently introduces a selection bias. In clinical practice, precise birth weight estimation is most crucial for high-risk pregnancies, including preterm deliveries, multiple gestations, and cases of intrauterine growth restriction or macrosomia. Consequently, our findings should be interpreted as a foundational baseline rather than a universally applicable clinical tool. Future research should focus on incorporating more diverse and high-risk datasets to validate and refine these algorithms, ensuring they remain robust in the complex scenarios that characterize routine obstetric care.

In evaluating our results, we compared the performance of our models primarily against traditional clinical equations (e.g., Hadlock and Shepard) rather than prior machine-learning architectures. While our findings show lower MAE and MAPE values compared to these established benchmarks, these results should not be interpreted as a universal superiority of one mathematical approach over another. Instead, they underscore the limitation of using generalized formulas developed decades ago on external populations for contemporary regional cohorts. The improved accuracy in our study likely reflects the benefits of localized calibration—where the models (both linear and ML-based) are trained on the unique biometric distribution of our specific population—rather than a direct algorithmic competition between different study designs.

When evaluating the performance of our localized models against established global equations like Hadlock III, it is imperative to distinguish between statistical reduction in error and actual clinical utility. For a standard term fetus weighing approximately 3300 g, an absolute error reduction of 50-80 g may not alter the clinical management plan (i.e., the decision for a vaginal delivery remains unchanged). However, the clinical value of this improved accuracy becomes significant at the decision boundaries. Generalized formulas often exhibit systematic biases, such as overestimating birth weight in certain local populations. In clinical practice, an overestimation near the macrosomia threshold (e.g., predicting 4100 g instead of a true weight of 3850 g) can trigger a cascade of 'defensive medicine' practices, leading to unnecessary elective cesarean deliveries or premature inductions of labor. By providing a tighter margin of error (lower MAE and narrower limits of agreement), our localized algorithm enhances the clinician's diagnostic confidence. Consequently, the primary real-world implication of this model is its potential to reduce false-positive categorizations at the extremes of the normal curve, thereby supporting safer and more evidence-based delivery planning.

Methodological distinctions: linear modeling vs. complex ML

In this study, we evaluated a spectrum of predictive approaches ranging from traditional regression-based modeling to complex ML architectures. While AI encompasses the broad field of computational intelligence, Machine Learning specifically refers to the data-driven algorithms used to identify patterns. It is important to note that Multiple Linear Regression (MLR) serves as a foundational supervised learning algorithm in the ML hierarchy. The superior performance of the MLR model in our specific dataset highlights a crucial clinical insight: the morphometric growth of fetal parameters in a healthy term population likely follows a predominantly linear trajectory. While non-linear ML models like ANN or Random Forest offer higher flexibility, they may be prone to overfitting or increased variance in homogeneous datasets. Therefore, our results suggest that for localized fetal weight estimation, the clinical interpretability and stability of linear-based ML models provide a distinct advantage over more complex, non-linear alternatives.

Conclusion

In conclusion, this study demonstrates that locally calibrated predictive models, particularly linear regression, can improve fetal birth weight estimation accuracy within a narrow, highly selected cohort of uncomplicated term pregnancies. However, these findings do not imply a universal superiority of machine-learning architectures over traditional formulas, nor do they support the immediate integration of these models into general obstetric practice. Given the reliance on a single-center dataset, the absence of external validation, and inherent limitations in our retrospective

methodological reporting, the proposed algorithms must be viewed strictly as a localized proof-of-concept. These models are not yet ready for clinical adoption. Future multi-center, prospective studies with rigorous methodological transparency and diverse populations are absolutely essential to validate their safety, generalizability, and true clinical utility before routine application.

Conflict of ınterest

The authors declare that they have no conflict of interest.

Funding

No funding was received for this study.

Data availability

Anonymized data can be retrieved by contacting the corresponding author and providing a compelling reason.

Contribution of authors

UO Osmanoglu: Project development, Data analysis, Manuscript writing/editing
A Dursun: Project development, Manuscript writing/editing
O Dulger: Project development, Data Collection, Manuscript editing.

References

1. Miyagi Y and Miyake T. Potential of Artificial Intelligence for Estimating Japanese Fetal Weights. Acta Med Okayama 2020; 74: 483–493. doi: 10.18926/AMO/61207
2. Tao J, Yuan Z, Sun L, Yu K and Zhang Z. Fetal birthweight prediction with measured data by a temporal machine learning method. BMC Med Inform Decis Mak 2021; 21. doi: 10.1186/s12911-021-01388-y
3. Özdamar O, Gün I, Öner O and Özden O. Evaluation of the accuracy of ultrasonography fetal weight estimation models; Assessing regression formulae in a Turkish population. J Obstet Gynaecol (Lahore) 2015; 35: 437–441. doi: 10.3109/01443615.2014.968111
4. Mohammadi H, Nemati M, Allahmoradi Z, Raissi HF, Esmaili SS and Sheikhani A. Ultrasound estimation of fetal weight in twins by artificial neural network. J Biomed Sci Eng 2011; 04: 46–50. doi: 10.4236/jbise.2011.41006
5. Ugwa EA, Gaya S and Ashimi A. Estimation of fetal weight before delivery in low-resource setting of North-west Nigeria: can we rely on our clinical skills?. The Journal of Maternal-Fetal & Neonatal Medicine 2015; 28(8): 949-953. doi: 10.3109/14767058.2014.938627
6. Schober P, Boer C and Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 2018; 126(5): 1763-1768. doi: 10.1213/ANE.0000000000002864
7. Koo TK and Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine 2016; 15(2): 155-163. doi: 10.1016/j.jcm.2016.02.012
8. Willmott CJ and Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 2005; 30(1): 79-82. doi: 10.3354/cr030079
9. Seman NM, Adem HM, Disasa FA and Simegn GL. Development of birth weight estimation model for Ethiopian population from sonographic evaluation. BMC Pregnancy Childbirth 2023; 23. doi: 10.1186/s12884-023-06145-9
10. Esinler D, Bircan O, Esin S, Sahin EG, Kandemir O and Yalvac S. Finding the Best Formula to Predict the Fetal Weight: Comparison of 18 Formulas. Gynecol Obstet Invest 2015; 80: 78–84. doi: 10.1159/000365814
11. Basha AS, Abu-Khader IB, Qutishat RM and Amarin ZO. Accuracy of sonographic fetal weight estimation within 14 days of delivery in a Jordanian population using Hadlock formula 1. Medical Principles and Practice 2012; 21: 366–369. doi: 10.1159/000335539
12. Hammami A, Mazer Zumaeta A, Syngelaki A, Akolekar R and Nicolaides KH. Ultrasonographic estimation of fetal weight: development of new model and assessment of performance of previous models. Ultrasound in Obstetrics and Gynecology 2018; 52: 35–43. doi: 10.1002/uog.19066
13. Balsyte D, Schäffer L, Burkhardt T, Wisser J and Kurmanavicius J. Sonographic prediction of macrosomia cannot be improved by combination with pregnancy-specific characteristics. Ultrasound in Obstetrics and Gynecology 2009; 33: 453–458. doi: 10.1002/uog.6282
14. Barel O, Maymon R, Vaknin Z, Tovbin J and Smorgick N. Sonographic fetal weight estimation - is there more to it than just fetal measurements? Prenat Diagn 2014; 34: 50–55. doi: 10.1002/pd.4250
15. Hadlock FP, Harrist RB, Sharman RS, Deter RL and Park SK. Estimation of fetal weight with the use of head, body, and femur measurements—A prospective study. Am J Obstet Gynecol 1985; 151: 333–337.
16. Shepard MJ, Richards VA, Berkowitz RL, Warsof SL and Hobbins JC. An evaluation of two equations for predicting fetal weight by ultrasound. Am J Obstet Gynecol 1982; 142: 47–54. doi: 10.1016/S0002-9378(16)32283-9
17. Merz E, Lieser H, Schicketanz KH and Härle J. Intrauterine fetal weight assessment using ultrasound. a comparison of several weight assessment methods and development of a new formula for the determination of fetal weight. Ultraschall in Der Medizin 1988; 9: 15–24.

18. Campbell S and Wilkin D. Ultrasonic measurement of fetal abdomen circumference in the estimation of fetal weight. British Journal of Obstetrics 1975; 82: 689–697.
19. Vintzileos AM, Campbell WA, Rodis JF, Bors-Koefoed R and Nochimson DJ. Fetal weight estimation formulas with head, abdominal, femur, and thigh circumference measurements. Am J Obstet Gynecol 1987; 157: 410–414.
20. Warsof SL, Gohari P, Berkowitz RL and Hobbins JC. The estimation of fetal weight by computer-assisted analysis. Am J Obstet Gynecol 1977; 128: 881–892. doi: 10.1016/0002-9378(77)90058-8
21. Thurnau GR, Tamura RK, Sabbagha R, Depp OR, Dyer A, Larkin R, Lee T and Laughlin C. A simple estimated fetal weight equation based on real-time ultrasound measurements of fetuses less than thirty-four weeks' gestation. Am J Obstet Gynecol 1983; 145: 557–561. doi: 10.1016/0002-9378(83)91195-X
22. Islam MN, Mustafina SN, Mahmud T and Khan NI. Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda. BMC pregnancy and childbirth 2022; 22(1): 348. doi: 10.1186/s12884-022-04594-2