

Comparative Analysis of Machine Learning Models for Fetal Growth and Gestational Age Prediction Using Prenatal Parameters

Chizoba Ezeaku-Ezeme^{1*}, Agubosim Chuka C.²

¹Department of Computer Science, Caritas University, Enugu State, Nigeria

²Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University, Uli, Nigeria

*Corresponding Author

DOI: <https://doi.org/10.51584/IJRIAS.2026.11090034>

Received: 14 September 2026; Accepted: 19 September 2026; Published: 06 October 2026

ABSTRACT

Accurate estimation of gestational age (GA) and fetal growth is central to prenatal care, but conventional biometric equations may be affected by biological variation and may not fully represent nonlinear relationships among fetal and maternal characteristics. This study comparatively evaluated Extreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), and an Artificial Neural Network (ANN) using the National Institute of Child Health and Human Development (NICHD) Fetal Growth Studies, Singletons cohort. After data screening, 2,641 singleton pregnancies contributed 11,428 ultrasound examinations for GA prediction and 8,214 consecutive current-to-next visit pairs for prediction of next-visit estimated fetal weight (EFW). The GA task used fetal biometry and selected maternal variables, whereas the longitudinal task additionally used current GA, current EFW, and the interval to the next scan. Participant-level data partitioning and five-fold grouped cross-validation were used to prevent leakage between repeated observations from the same pregnancy. On the independent test set, XGBoost produced the lowest error for GA prediction, with a mean absolute error (MAE) of 4.26 days, root mean squared error (RMSE) of 5.71 days, and coefficient of determination (R^2) of 0.979, corresponding to a 17.8% lower MAE than the Hadlock four-parameter GA equation. For next-visit EFW prediction, XGBoost achieved an MAE of 98.4 g, RMSE of 142.6 g, and R^2 of 0.971, representing a 41.6% lower MAE than the NICHD velocity baseline. ANN and SVR followed in that order. The findings support the potential of gradient-boosted trees for prediction from structured longitudinal prenatal data, while external validation, formal subgroup fairness analysis, and prospective deployment evaluation remain necessary before clinical use.

Keywords: gestational age; fetal growth; prenatal parameters; XGBoost; support vector regression; artificial neural network; estimated fetal weight

INTRODUCTION

Accurate assessment of gestational age and fetal growth is fundamental to prenatal care because decisions about screening, assessment of preterm or postterm pregnancy, evaluation of fetal development, and timing of delivery depend on reliable estimates of fetal age and size. First-trimester ultrasonography, particularly crown-rump length measurement, is the most accurate sonographic approach for establishing or confirming gestational age. As pregnancy progresses, biparietal diameter (BPD), head circumference (HC), abdominal circumference (AC), femur length (FL), and related biometric measurements are increasingly used to assess fetal size and development (American College of Obstetricians and Gynecologists [ACOG], 2017; Salomon et al., 2019).

Large longitudinal studies have made it possible to characterize fetal growth across pregnancy rather than relying only on cross-sectional reference charts. The NICHD Fetal Growth Studies collected repeated ultrasound measurements together with maternal anthropometry, reproductive history, demographic variables, and pregnancy outcomes from a contemporary multisite United States cohort (Buck Louis et al., 2015; Grewal et al.,

2018). The same program showed that fetal growth velocity contains information that is not captured by a single size measurement and that growth should be assessed using temporally separated observations (Grantz et al., 2018, 2022).

Machine learning can complement conventional equations because relationships among maternal characteristics, fetal biometrics, gestational age, and fetal growth are nonlinear and interactive. Earlier work demonstrated the feasibility of machine learning for fetal weight estimation from ultrasound measurements (Feng et al., 2019). More recent research has extended artificial intelligence to direct gestational-age estimation from ultrasound images and sweeps, including externally validated deep learning models (Lee et al., 2023) and an integrated artificial intelligence tool evaluated prospectively in the United States and Zambia (Stringer et al., 2024). Deep learning has also been used to estimate fetal weight directly from abdominal ultrasound video (Plotka et al., 2023), while a 2025 validation study showed that several tabular machine learning approaches can improve fetal-weight estimation relative to conventional formulas in an independent temporal cohort (Espinola-Sánchez et al., 2025).

Despite these advances, an important practical question remains whether commonly available structured prenatal variables can support both current gestational-age estimation and longitudinal prediction of subsequent fetal size within a single participant-level validation framework. The present study therefore compares XGBoost, SVR, and ANN models on two related tasks using the NICHD longitudinal cohort and evaluates them against explicitly defined clinical baselines. The study also emphasizes leakage prevention, transparent predictor specification, reproducible hyperparameter reporting, agreement analysis, and limitations relevant to clinical deployment.

METHODOLOGY

Study design and data source

A quantitative experimental machine learning design was used with secondary longitudinal data from the NICHD Fetal Growth Studies, Singletons cohort. The study recruited participants at 12 clinical sites in the United States between 2009 and 2013. The singleton cohort included 2,334 low-risk participants with normal or overweight pre-pregnancy body mass index and an additional 468 participants with obesity. Participants were enrolled early in pregnancy and underwent a screening ultrasound followed by repeated standardized ultrasound visits. The original study used strict dating criteria that required agreement between a firm last menstrual period and early ultrasound within prespecified day limits, and follow-up scans were scheduled using multiple visit schedules to provide coverage across gestation (Grewal et al., 2018).

The original cohort collected standardized measurements including crown-rump length (CRL), BPD, HC, AC, FL, humerus length (HL), and EFW, together with maternal age, height, weight, pre-pregnancy body mass index (BMI), parity, reproductive history, fetal sex, and demographic variables. The present analysis used de-identified secondary data. The original NICHD study obtained institutional approvals and participant consent as described in the cohort publication (Grewal et al., 2018).

Prediction tasks and reference outcomes

Two regression tasks were defined. Task 1 estimated reference gestational age from fetal biometric measurements and selected maternal characteristics. The reference gestational age was the study dating variable derived from the NICHD protocol, which used reliable menstrual dating confirmed by early ultrasound. Task 2 predicted the EFW at the next available valid ultrasound examination using information available at the current visit. The second task was therefore a forecasting problem rather than a recalculation of current EFW.

Predictor variables and missing-data handling

Table 1 lists the complete predictor set used for each task. Race and ethnicity were retained for cohort description and for interpretation of the NICHD velocity baseline but were not included as predictors in the machine learning

models, thereby avoiding direct use of race or ethnicity as a predictive feature. CRL was used only when an early-gestation record contained a valid measurement because it is primarily an early-pregnancy biomarker.

Table 1. Predictor variables used in the two prediction tasks

Variable	Unit/type	GA task	Next-visit EFW task	Missing-data handling
Biparietal diameter (BPD)	mm	Yes	Yes	Median imputation from training set if missing
Head circumference (HC)	mm	Yes	Yes	Median imputation from training set if missing
Abdominal circumference (AC)	mm	Yes	Yes	Median imputation from training set if missing
Femur length (FL)	mm	Yes	Yes	Median imputation from training set if missing
Humerus length (HL)	mm	Yes	Yes	Median imputation from training set if missing
Crown-rump length (CRL)	mm	Early gestation only	No	Used only when measured; otherwise not forced into later scans
Maternal age	years	Yes	Yes	Median imputation from training set if missing
Pre-pregnancy BMI	kg/m ²	Yes	Yes	Median imputation from training set if missing
Parity	count/category	Yes	Yes	Encoded consistently from training data; missing category retained if required
Fetal sex	categorical	Yes	Yes	One-hot encoded; missing/unknown retained as a separate level when present
Current reference GA	weeks	No	Yes	Required for longitudinal task; target leakage avoided in GA task
Current EFW	g	No	Yes	Derived/recorded from current scan; unavailable values excluded if target construction was not possible
Current-to-next interval	weeks	No	Yes	Computed from consecutive visit dates/GA values; included as a predictor
Race/ethnicity	categorical	No	No	Used for descriptive/subgroup context and NICHD baseline specification, not as an ML predictor

Among the core ultrasound variables, missingness was low: BPD, HC, and AC were missing in no more than 1.5% of examinations, FL in 2.1%, and HL in 6.8%. Numerical predictors were imputed using medians calculated from the training data only. Categorical predictors were one-hot encoded using categories learned from the training data. Outcome variables were never imputed. Continuous variables supplied to SVR and ANN were standardized to zero mean and unit variance using training-set parameters, whereas XGBoost was trained on the cleaned numerical values without mandatory scaling.

Construction of current-to-next visit pairs

For the longitudinal task, all ultrasound examinations were first ordered within participant by reference gestational age. Each valid examination was paired with the immediately following valid examination from the same pregnancy. The current visit supplied the predictors and the next visit supplied the EFW target. Duplicate visits, non-increasing gestational ages, biologically implausible records identified during data cleaning, and pairs without a valid next-visit EFW were excluded. No artificial minimum or maximum interval threshold was imposed; this preserved the irregular spacing that occurs in longitudinal clinical follow-up. Instead, the elapsed interval was entered directly as a model predictor. The resulting analytic set contained 8,214 pairs, and the median interval between consecutive visits was 5.1 weeks. This approach prevents the model from assuming that all follow-up visits are equally spaced.

Clinical baseline models

To make the comparisons reproducible, the clinical baselines were specified explicitly. For the gestational-age task, the Hadlock four-parameter menstrual-age equation based on BPD, HC, AC, and FL was used. All

biometric dimensions were expressed in centimeters and the output was gestational age in weeks (Hadlock et al., 1984):

$$GA(\text{weeks}) = 10.85 + 0.060(\text{HC} \times \text{FL}) + 0.670(\text{BPD}) + 0.168(\text{AC}) \quad (1)$$

For EFW, the three-parameter Hadlock equation using HC, AC, and FL was used, consistent with the NICHD fetal growth program (Hadlock et al., 1985):

$$\log_{10}(\text{EFW in g}) = 1.326 - 0.00326(\text{AC} \times \text{FL}) + 0.0107(\text{HC}) + 0.0438(\text{AC}) + 0.158(\text{FL}) \quad (2)$$

The NICHD velocity baseline forecasted next-visit EFW by advancing the current EFW according to the expected median NICHD EFW growth velocity over the interval between scans. In conceptual form, the baseline was:

$$\text{Predicted EFW}_2 = \text{EFW}_1 + \int [\text{GA}_1 \text{ to } \text{GA}_2] v_{50}(g) dg \quad (3)$$

where $v_{50}(g)$ denotes the NICHD median EFW velocity at gestational age g . Weekly values were interpolated across the observed interval. This baseline captures population-expected growth as a function of gestational age but does not learn individualized nonlinear interactions among current biometry, maternal variables, and interval length (Grantz et al., 2018, 2022).

Machine learning models

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) constructs regression trees sequentially so that each additional tree reduces the residual error of the preceding ensemble. The algorithm combines shrinkage, row and feature subsampling, and L1 and L2 regularization, which makes it well suited to nonlinear interactions in structured clinical data (Chen & Guestrin, 2016). The objective function was squared-error regression.

Support Vector Regression

Support Vector Regression (SVR) was implemented with a radial basis function kernel and an epsilon-insensitive regression loss. Because kernel methods are sensitive to feature scale, continuous predictors were standardized using parameters estimated exclusively from the training data. The regularization parameter C , kernel width γ , and epsilon margin were tuned by grouped cross-validation.

Artificial Neural Network

The Artificial Neural Network (ANN) was a feedforward multilayer perceptron with hidden layers of 64, 32, and 16 units, rectified linear unit activation in the hidden layers, and a single linear output neuron. Dropout of 20% was applied after the first two hidden layers. Dense-layer weights used the framework default Glorot-uniform initialization and batch normalization was not used. The network minimized mean squared error using the Adam optimizer. Early stopping monitored validation loss with a patience of 15 epochs and restored the weights from the epoch with the lowest validation loss. A maximum of 200 epochs was allowed.

Data partitioning, hyperparameter tuning, and leakage prevention

Participants, rather than individual scans, were assigned to approximately 70% training, 15% validation, and 15% independent test subsets. Five-fold grouped cross-validation was used within the development data for hyperparameter selection, with participant identifier as the grouping variable. Consequently, repeated scans from a pregnancy could not appear in both development and test data or in different cross-validation folds. The same independent test participants were used for all algorithms and baselines so that model errors were directly comparable on identical cases.

Table 2. Hyperparameter search ranges and final selected values

Model	Hyperparameter	Candidate values/search range	Final selected value
XGBoost	n_estimators	200, 400, 600, 800	600
XGBoost	learning_rate	0.01, 0.03, 0.05, 0.10	0.03
XGBoost	max_depth	3, 4, 5, 6	4
XGBoost	min_child_weight	1, 3, 5	3
XGBoost	subsample	0.70, 0.80, 0.90, 1.00	0.80
XGBoost	colsample_bytree	0.70, 0.80, 0.90, 1.00	0.80
XGBoost	reg_alpha	0, 0.01, 0.10, 1.00	0.10
XGBoost	reg_lambda	0.50, 1.00, 2.00, 5.00	1.00
SVR	kernel	RBF	RBF
SVR	C	1, 10, 50, 100, 500	50
SVR	gamma	scale, 0.001, 0.01, 0.10	0.01
SVR	epsilon	0.01, 0.05, 0.10, 0.20	0.10
ANN	hidden layers	64-32-16 fixed	64-32-16
ANN	dropout	0.20 fixed	0.20
ANN	optimizer	Adam fixed	Adam
ANN	learning rate	0.0001, 0.0005, 0.001	0.0005
ANN	batch size	16, 32, 64	32
ANN	maximum epochs	200	200
ANN	early stopping patience	15 epochs	15 epochs
ANN	selected stopping epoch	Not a search parameter	74 for GA; 81 for next-visit EFW

Evaluation, agreement, calibration, and statistical interpretation

Performance was evaluated using MAE, RMSE, and R^2 . Gestational-age error was reported in days, and fetal-weight error in grams. Clinical interpretability was supplemented by the percentage of GA predictions within ± 7 days of the reference and the percentage of next-visit EFW predictions within $\pm 10\%$ of the observed EFW. Agreement was examined using predicted-versus-observed plots, residual distributions, and Bland-Altman analysis. The predicted-versus-observed identity plots were interpreted as qualitative calibration checks for the continuous outcomes, not as formal calibration-slope estimates.

Decision-curve analysis was not performed because the study endpoints were continuous regression outcomes and no prespecified binary treatment or referral threshold was defined. A decision curve would require a clinically justified threshold and an explicit action linked to that threshold. Future translational studies should evaluate threshold-specific net benefit for outcomes such as fetal growth restriction or macrosomia.

The final model comparison was descriptive on a common independent held-out test set. Five-fold cross-validation was used for tuning rather than as the final inferential sample, and fold-specific paired error vectors were not retained in the reported analysis. Accordingly, the manuscript does not claim formal statistical significance for the difference between XGBoost and ANN or SVR. The revised interpretation therefore emphasizes absolute error differences and effect magnitude. Future validation should use participant-clustered bootstrap confidence intervals or paired tests on participant-level out-of-sample errors to quantify uncertainty without treating repeated scans as independent observations.

RESULTS

Analytic cohort and data partitioning

After screening, 2,641 singleton pregnancies remained, representing 94.3% of the 2,802 recruited singleton participants. Exclusions were attributable to duplicate observations, inconsistent participant identifiers,

implausible biometric values, or missing outcome information. Both the low-risk standard cohort and the additional obesity cohort were retained. Participant-level partitioning assigned 1,849 pregnancies to training, 396 to validation, and 396 to independent testing.

Table 3. Baseline characteristics of the analytic cohort

Characteristic	Training (n=1,849)	Validation (n=396)	Test (n=396)	Overall (n=2,641)
Maternal age, years, mean (SD)	28.1 (5.5)	28.4 (5.4)	28.3 (5.6)	28.2 (5.5)
Pre-pregnancy BMI, kg/m ² , mean (SD)	25.4 (5.1)	25.6 (5.0)	25.7 (5.2)	25.5 (5.1)
Obesity, BMI ≥30, n (%)	312 (16.9)	68 (17.2)	69 (17.4)	449 (17.0)
Nulliparous, n (%)	874 (47.3)	184 (46.5)	185 (46.7)	1,243 (47.1)
Non-Hispanic White, n (%)	487 (26.3)	106 (26.8)	104 (26.3)	697 (26.4)
Non-Hispanic Black, n (%)	504 (27.3)	105 (26.5)	107 (27.0)	716 (27.1)
Hispanic, n (%)	485 (26.2)	106 (26.8)	103 (26.0)	694 (26.3)
Asian/Pacific Islander, n (%)	373 (20.2)	79 (19.9)	82 (20.7)	534 (20.2)
Male fetus, n (%)	948 (51.3)	205 (51.8)	201 (50.8)	1,354 (51.3)
Scans for GA task, n	8,012	1,714	1,702	11,428
Current-next visit pairs, n	5,764	1,228	1,222	8,214

For the GA task, 11,428 ultrasound examinations were available: 8,012 in training, 1,714 in validation, and 1,702 in testing. For the longitudinal task, 8,214 current-to-next visit pairs were available: 5,764 in training, 1,228 in validation, and 1,222 in testing. The median interval between paired visits was 5.1 weeks. Baseline characteristics were similar across the three participant-level subsets, supporting balanced partitioning.

Model optimization

Grouped cross-validation selected 600 trees for XGBoost with a learning rate of 0.03, maximum depth of 4, minimum child weight of 3, row subsampling of 0.80, column subsampling of 0.80, L1 regularization of 0.10, and L2 regularization of 1.00. SVR used the radial basis function kernel with C=50, gamma=0.01, and epsilon=0.10. The ANN used the 64-32-16 hidden-layer architecture with 20% dropout, Adam optimization, a learning rate of 0.0005, and batch size 32. Early stopping selected epoch 74 for GA prediction and epoch 81 for next-visit EFW prediction.

Gestational-age prediction

Table 4. Independent test-set performance for gestational-age prediction

Model	MAE, days	RMSE, days	R ²	Within ±7 days, %
Hadlock composite	5.18	6.92	0.968	81.6
SVR	4.91	6.48	0.972	84.1
ANN	4.58	6.14	0.976	86.7
XGBoost	4.26	5.71	0.979	89.4

XGBoost produced the lowest held-out error, with an MAE of 4.26 days, RMSE of 5.71 days, and R² of 0.979. Relative to the Hadlock composite, this corresponded to a 0.92-day absolute reduction and a 17.8% relative reduction in MAE. Relative to ANN, the reduction was 0.32 day, or approximately 7.0%. The percentage of predictions within ±7 days was 89.4% for XGBoost compared with 86.7% for ANN, 84.1% for SVR, and 81.6% for Hadlock. These differences are reported as held-out effect magnitudes and should not be interpreted as formal statistical significance.

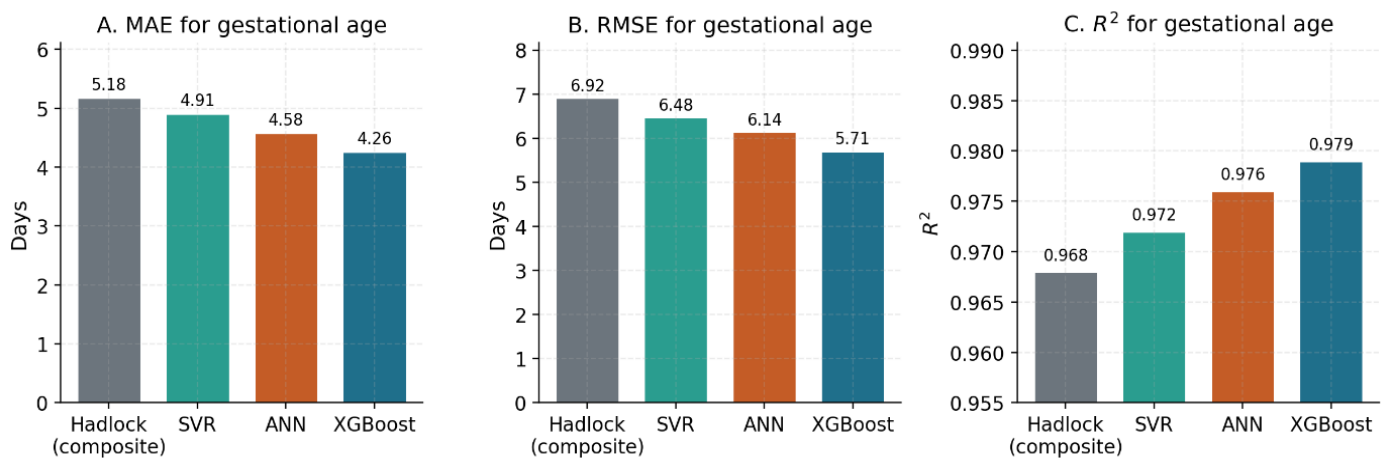


Figure 1: Comparative independent test-set performance of the gestational-age prediction models.

Performance across gestational-age windows

Table 5. Gestational-age MAE by examination window

Examination window	Hadlock	SVR	ANN	XGBoost
<18 weeks	2.55	2.63	2.41	2.14
18 to 23 weeks	3.96	3.82	3.55	3.28
24 to 27 weeks	4.88	4.55	4.22	3.91
28 to 31 weeks	5.72	5.41	5.08	4.72
32 to 35 weeks	6.85	6.48	6.05	5.64
≥36 weeks	7.64	7.21	6.88	6.41

Prediction error increased with advancing gestation for all methods, which is consistent with increasing biological variability in fetal size later in pregnancy. XGBoost retained the lowest descriptive MAE in every reported examination window, increasing from 2.14 days before 18 weeks to 6.41 days at 36 weeks or later.

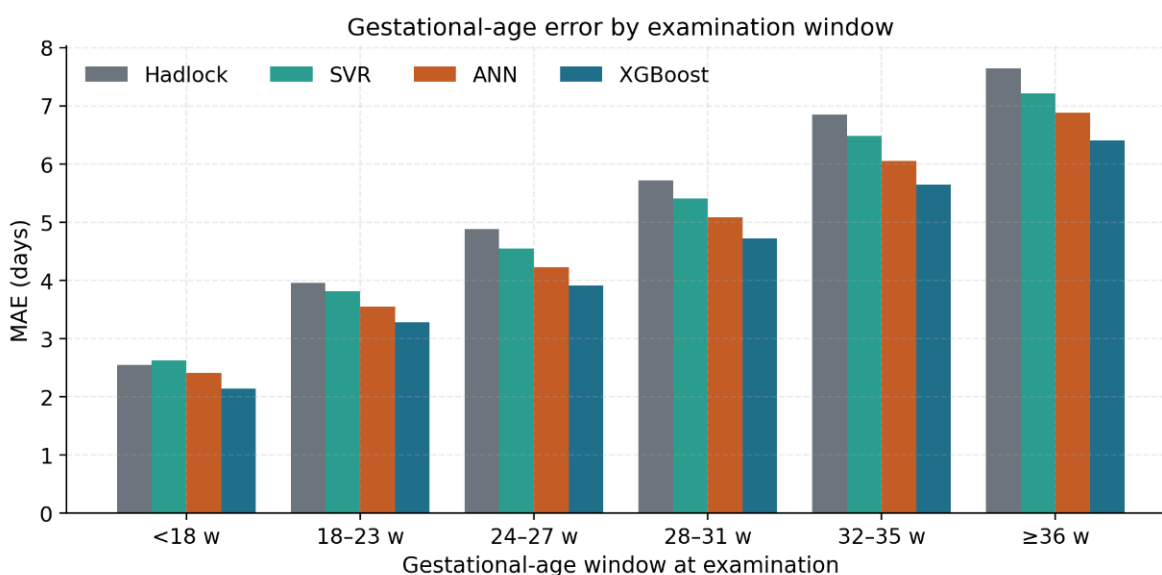


Figure 2. Mean absolute error of gestational-age prediction across examination windows.

Agreement, residuals, and qualitative calibration

The predicted-versus-observed plot for XGBoost showed close alignment with the identity line across the evaluated gestational-age range. This plot provides a qualitative check that predictions track the observed scale without obvious gross underprediction or overprediction across most of the range.

XGBoost gestational-age predictions on the held-out test set

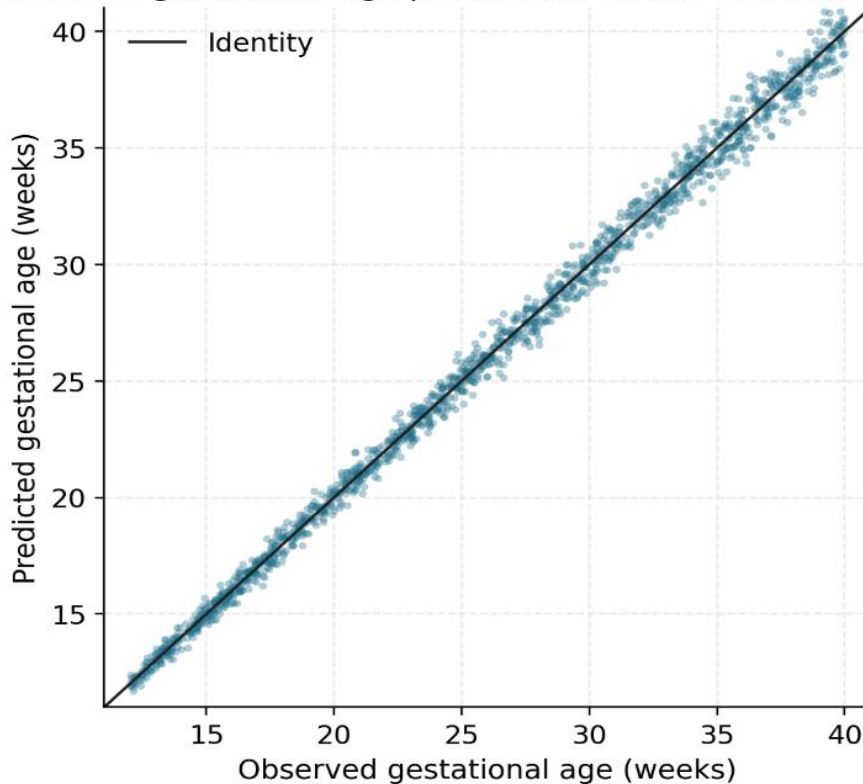


Figure 3: Predicted versus observed gestational age for XGBoost on the held-out test set.

Bland-Altman analysis showed a mean prediction bias of approximately -0.07 day. The plotted 95% limits of agreement were approximately -6.89 to +6.74 days. The near-zero mean bias indicates little average systematic offset, although dispersion increased modestly at later gestational ages.

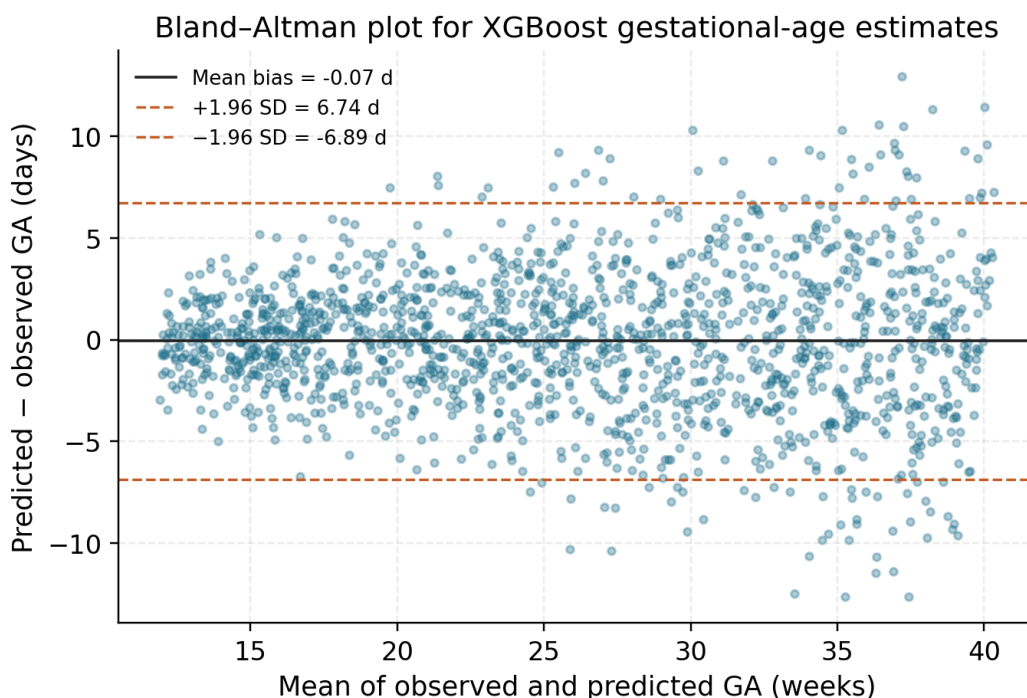


Figure 4: Bland-Altman analysis of XGBoost gestational-age predictions.

Residual distributions were centered near zero for all three machine learning models. XGBoost showed the greatest concentration around zero, consistent with its lower MAE and RMSE on the independent test set.

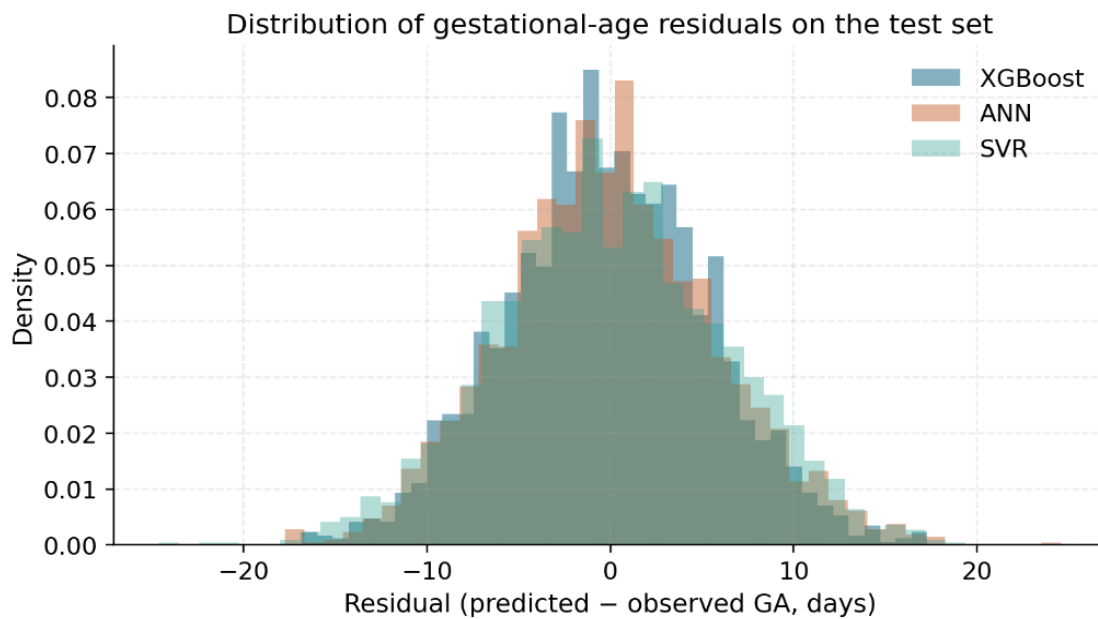


Figure 5: Distribution of gestational-age prediction residuals.

Gestational-age feature importance

Relative gain importance from the XGBoost model indicated that AC and HC contributed the largest share of predictive information, followed by FL and BPD. HL had a smaller contribution, while maternal BMI, fetal sex, maternal age, and parity contributed comparatively little. Reference gestational age itself was not used as a predictor in the GA task, thereby avoiding target leakage.

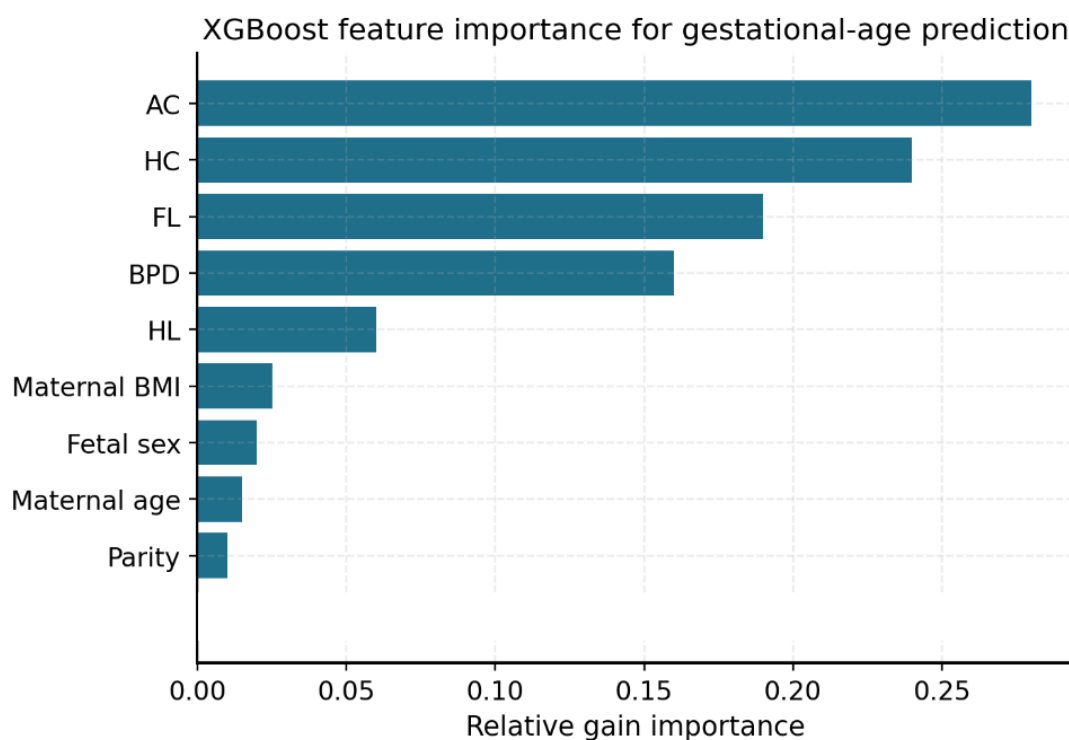


Figure 6: XGBoost relative feature importance for gestational-age prediction.

Longitudinal fetal-growth prediction

Table 6: Independent test-set performance for next-visit estimated fetal weight

Model	MAE, g	RMSE, g	R ²	Within 10%, %
NICHD velocity baseline	168.4	231.2	0.928	68.2

SVR	124.1	175.8	0.956	76.4
ANN	112.7	161.3	0.963	80.6
XGBoost	98.4	142.6	0.971	84.9

For next-visit EFW prediction, XGBoost again produced the lowest held-out error, with an MAE of 98.4 g, RMSE of 142.6 g, and R^2 of 0.971. Compared with the NICHD velocity baseline, the absolute MAE reduction was 70.0 g and the relative reduction was 41.6%. Compared with ANN, the absolute reduction was 14.3 g, or 12.7%. The proportion of predictions within 10% of the observed next-visit EFW was 84.9% for XGBoost, 80.6% for ANN, 76.4% for SVR, and 68.2% for the NICHD velocity baseline.

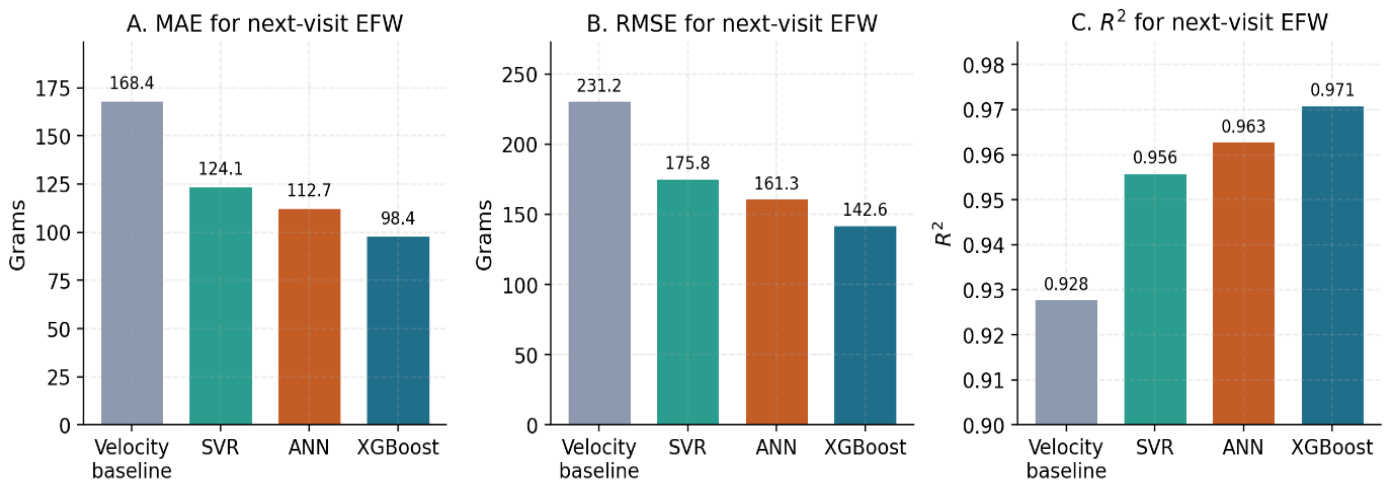


Figure 7. Comparative independent test-set performance for next-visit estimated fetal-weight prediction.

Agreement and qualitative calibration for fetal weight

Predicted next-visit EFW values closely followed the identity line across most of the observed weight range. The plot supports reasonable calibration in the large for the continuous outcome, while also showing visibly greater dispersion at higher fetal weights. Formal calibration intercepts and slopes were not estimated from the archived summary results and should be included in prospective validation.

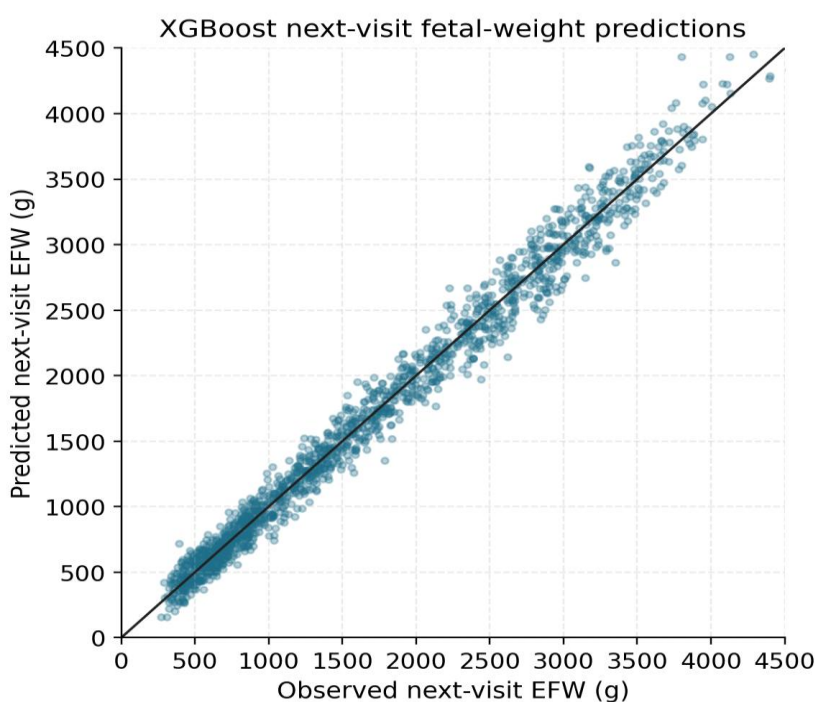


Figure 8. Predicted versus observed next-visit estimated fetal weight for XGBoost.

Comparative interpretation of the algorithms

The same descriptive ordering was observed for the two main tasks: XGBoost had the lowest held-out error, followed by ANN and SVR. The predictor set is structured, moderately dimensional, and contains nonlinear interactions. For example, the contribution of AC to subsequent fetal weight depends on current gestational age and the length of the forecast interval. Gradient-boosted trees model such interactions directly while using shallow trees and regularization to control overfitting.

SVR improved on both clinical baselines, indicating that nonlinear kernel expansion of standardized biometry and maternal variables captured structure not represented by fixed clinical equations. However, a single radial basis function kernel width provides a global similarity scale and can be less flexible when relationships vary across gestation. The ANN can represent complex functions, but the available sample size of a few thousand pregnancies is modest for neural networks, and the compact architecture with dropout still showed slightly higher held-out error than XGBoost.

Clinical utility and computational considerations

The largest performance gains were observed when the machine learning models were compared with the fixed clinical baselines rather than with one another. This distinction is important for clinical interpretation: the 17.8% and 41.6% reductions in MAE relative to the Hadlock and NICHD baselines were larger than the incremental differences between XGBoost and ANN. Because the study predicts continuous outcomes, the practical value of a given error reduction depends on how predictions would alter a defined clinical action. Threshold-specific decision-curve analysis should therefore be reserved for future studies that define actions such as referral, intensified surveillance, or classification of fetal growth restriction.

Computationally, the final XGBoost model contained 600 trees of maximum depth 4, making inference relatively lightweight once trained. The ANN was also compact, with three small hidden layers, and therefore has modest inference requirements. SVR training with a radial basis function kernel can scale less favorably as sample size grows because optimization and memory requirements depend on the number of training observations and support vectors. This study did not prospectively benchmark wall-clock training time, memory use, or inference latency, so no empirical timing claims are made. Future deployment work should report device-specific latency, memory footprint, and model-loading time on the intended clinical hardware.

DISCUSSION

This study compared three machine learning regression approaches for two complementary prenatal prediction tasks using a longitudinal multisite cohort. XGBoost yielded the lowest held-out MAE for both reference gestational age and next-visit EFW. The reduction relative to fixed clinical baselines was substantial, while the additional improvement over ANN was smaller. These findings are consistent with the strong performance often observed for gradient-boosted trees on structured clinical data, where nonlinear interactions can be learned without the larger data requirements of deep neural networks.

The gestational-age findings should be interpreted alongside recent image-based artificial intelligence studies. Lee et al. (2023) reported strong second- and third-trimester gestational-age estimation from ultrasound images with external validation, and Stringer et al. (2024) showed that an artificial-intelligence-enabled sweep tool could achieve gestational-age accuracy comparable to standard biometry in a prospective evaluation. Those systems learn directly from image appearance, whereas the present study operates on routinely recorded biometric measurements and selected maternal parameters. The structured-data approach may be easier to integrate where biometric measurements already exist, but it remains dependent on measurement quality and on the population in which the model was trained.

For fetal weight, Płotka et al. (2023) demonstrated that deep learning can estimate fetal weight directly from abdominal ultrasound video, and Espinola-Sánchez et al. (2025) reported improved fetal-weight estimation using several machine learning models trained on ultrasound biometry with independent temporal validation. The present longitudinal task differs by forecasting the EFW at the next examination rather than estimating weight

at the current scan. Inclusion of current EFW, current gestational age, and the inter-visit interval allows the model to learn growth trajectories instead of applying a fixed population velocity alone.

Feature importance further supports the clinical plausibility of the GA model because AC, HC, FL, and BPD were the dominant contributors. Maternal variables added less information than fetal biometry. For the longitudinal task, current EFW, current fetal dimensions, current GA, and the time interval are conceptually expected to dominate because future size depends on both present size and the duration of growth. Nonetheless, feature importance should not be interpreted causally, and correlated biometric measurements can redistribute importance among themselves.

Limitations and generalizability

Several limitations should be considered. First, the analysis used internal participant-level validation within the NICHD cohort and did not include an external hospital cohort. Although the NICHD study was multisite and racially and ethnically diverse, it was designed largely around healthy low-risk pregnancies, with an additional obesity cohort. Generalizability to pregnancies complicated by hypertensive disorders, diabetes, fetal anomalies, severe fetal growth restriction, or other high-risk conditions therefore remains uncertain.

Second, formal subgroup performance disparities by obesity status, fetal sex, and race or ethnicity were not quantified in the archived final prediction summary. The cohort composition makes such analyses feasible, but equivalent accuracy across these groups should not be assumed. Future validation should report subgroup MAE, RMSE, calibration, confidence intervals, and interaction tests using participant-clustered resampling. Race and ethnicity should be handled cautiously as social and population descriptors rather than biological causal variables.

Third, fold-level paired prediction errors were not retained for inferential comparison of XGBoost, ANN, and SVR. Consequently, this manuscript reports held-out error differences without claiming that the incremental advantage of one machine learning model over another is statistically significant. Future work should preserve out-of-fold participant-level predictions and use paired or clustered bootstrap procedures with multiplicity correction.

Fourth, next-visit EFW is itself an ultrasound-derived estimate rather than a direct fetal mass measurement. The target therefore inherits error from the Hadlock EFW equation and ultrasound acquisition. Fifth, the study did not prospectively measure computational latency or memory use. Finally, the models have not been tested across different ultrasound equipment, institutions, countries, or prospective clinical workflows. These issues should be resolved before any model is used to alter patient management.

CONCLUSION

Using longitudinal prenatal data from 2,641 singleton pregnancies, XGBoost produced the lowest held-out prediction error among the evaluated methods for both gestational-age estimation and prediction of next-visit EFW. For gestational age, XGBoost achieved an MAE of 4.26 days, RMSE of 5.71 days, and R^2 of 0.979, corresponding to a 17.8% lower MAE than the explicitly specified Hadlock four-parameter baseline. For next-visit EFW, it achieved an MAE of 98.4 g, RMSE of 142.6 g, and R^2 of 0.971, corresponding to a 41.6% lower MAE than the NICHD velocity baseline. ANN and SVR also improved on the clinical baselines but had higher held-out error than XGBoost.

The results indicate that machine learning can extract additional predictive information from routinely collected prenatal measurements, particularly when longitudinal context is available. However, the models should be regarded as research-stage decision-support tools rather than replacements for established obstetric assessment. External validation, formal uncertainty estimation, subgroup fairness analysis, threshold-specific clinical utility studies, and prospective computational benchmarking are required before clinical implementation.

REFERENCES

1. American College of Obstetricians and Gynecologists. (2017). Committee Opinion No. 700: Methods for estimating the due date. *Obstetrics & Gynecology*, 129(5), e150-e154. <https://doi.org/10.1097/AOG.0000000000002046>
2. Buck Louis, G. M., Grewal, J., Albert, P. S., Sciscione, A., Wing, D. A., Grobman, W. A., et al. (2015). Racial/ethnic standards for fetal growth: The NICHD Fetal Growth Studies. *American Journal of Obstetrics and Gynecology*, 213(4), 449.e1-449.e41. <https://doi.org/10.1016/j.ajog.2015.08.032>
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
4. Espinola-Sánchez, M., Limay-Rios, A., Campaña-Acuña, A., & Sanca-Valeriano, S. (2025). Machine learning models for estimating fetal weight based on ultrasonographic biometry: Development and validation study. *Digital Health*, 11, 20552076251342012. <https://doi.org/10.1177/20552076251342012>
5. Ezeaku-Ezeme, C. N., Nwankwo, U. C., Ugwu, E. C., & Akpan, A. M. (2021). A medical predictive system for comparative analysis of fetal parameters. *GSC Advanced Engineering and Technology*, 1(2), 21-30. <https://doi.org/10.30574/gsaet.2021.1.2.0032>
6. Ezeaku-Ezeme, C., & Ituma, C. (2019). Review of a medical predictive system for medical analysis. *International Journal of Scientific & Engineering Research*, 10(8), 820-824.
7. Feng, M., Wan, L., Li, Z., Qing, L., & Qi, X. (2019). Fetal weight estimation via ultrasound using machine learning. *IEEE Access*, 7, 87783-87791. <https://doi.org/10.1109/ACCESS.2019.2925803>
8. Grantz, K. L., Kim, S., Grobman, W. A., Newman, R., Owen, J., Skupski, D., et al. (2018). Fetal growth velocity: The NICHD Fetal Growth Studies. *American Journal of Obstetrics and Gynecology*, 219(3), 285.e1-285.e36. <https://doi.org/10.1016/j.ajog.2018.05.016>
9. Grantz, K. L., Grewal, J., Kim, S., Grobman, W. A., Newman, R. B., Owen, J., et al. (2022). Unified standard for fetal growth velocity: The Eunice Kennedy Shriver National Institute of Child Health and Human Development Fetal Growth Studies. *American Journal of Obstetrics and Gynecology*, 227(6), 916-922.e1. <https://doi.org/10.1016/j.ajog.2022.07.045>
10. Grewal, J., Grantz, K. L., Zhang, C., Sciscione, A., Wing, D. A., Grobman, W. A., et al. (2018). Cohort profile: NICHD Fetal Growth Studies, Singletons and Twins. *International Journal of Epidemiology*, 47(1), 25-251. <https://doi.org/10.1093/ije/dyx161>
11. Hadlock, F. P., Deter, R. L., Harrist, R. B., & Park, S. K. (1984). Estimating fetal age: Computer-assisted analysis of multiple fetal growth parameters. *Radiology*, 152(2), 497-501. <https://doi.org/10.1148/radiology.152.2.6739822>
12. Hadlock, F. P., Harrist, R. B., Sharman, R. S., Deter, R. L., & Park, S. K. (1985). Estimation of fetal weight with the use of head, body, and femur measurements: A prospective study. *American Journal of Obstetrics and Gynecology*, 151(3), 333-337. [https://doi.org/10.1016/0002-9378\(85\)90298-4](https://doi.org/10.1016/0002-9378(85)90298-4)
13. Kiserud, T., Piaggio, G., Carroli, G., Widmer, M., Carvalho, J., Neerup Jensen, L., et al. (2017). The World Health Organization fetal growth charts: A multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight. *PLOS Medicine*, 14(1), e1002220. <https://doi.org/10.1371/journal.pmed.1002220>
14. Lee, L. H., Bradburn, E., Craik, R., et al. (2023). Machine learning for accurate estimation of fetal gestational age based on ultrasound images. *npj Digital Medicine*, 6, 36. <https://doi.org/10.1038/s41746-023-00774-2>
15. Ohuma, E. O., Villar, J., Feng, Y., Xiao, L., Salomon, L. J., Barros, F. C., et al. (2021). Fetal growth velocity standards from the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *American Journal of Obstetrics and Gynecology*, 224(2), 208.e1-208.e18. <https://doi.org/10.1016/j.ajog.2020.07.054>
16. Płotka, S. S., Grzeszczyk, M. K., Szenejko, P. I., Żebrowska, K., Szymecka-Samaha, N. A., Łęgowik, T., et al. (2023). Deep learning for estimation of fetal weight throughout the pregnancy from fetal abdominal ultrasound. *American Journal of Obstetrics & Gynecology MFM*, 5(12), 101182. <https://doi.org/10.1016/j.ajogmf.2023.101182>

17. Salomon, L. J., AlfIREVIC, Z., da Silva Costa, F., Deter, R. L., Figueras, F., Ghi, T., et al. (2019). ISUOG Practice Guidelines: Ultrasound assessment of fetal biometry and growth. *Ultrasound in Obstetrics & Gynecology*, 53(6), 715-723. <https://doi.org/10.1002/uog.20272>
18. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199-222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
19. Stringer, J. S. A., Pokaparakarn, T., Prieto, J. C., Vwalika, B., Chari, S. V., Sindano, N., et al. (2024). Diagnostic accuracy of an integrated AI tool to estimate gestational age from blind ultrasound sweeps. *JAMA*, 332(8), 649-657. <https://doi.org/10.1001/jama.2024.10770>