Predicting Pediatric Weight Growth Trajectories Across Diverse Settings: A Machine Learning Comparison of Random Forest, SVR, and Gradient Boosting

Alemayehu Siffir Argawu[1,2*], Muniswamy Begari[1], Punyavathi Begari[1]

[1]Department of Statistics, Andhra University, Andhra Pradesh, India (530,003)
[2]Department of Statistics, Ambo University, West-Shoa Zone, Ethiopia
*Corresponding author: alex089973@gmail.com

Abstract

Purpose: This study compares three machine learning models' effectiveness in predicting non-linear pediatric weight trajectories across four developing nations, addressing limitations of traditional parametric approaches.

Method: Using longitudinal data from 7,140 children in Ethiopia, India, Peru, and Vietnam (Young Lives Study, 2002-2016), we analyzed complete weight measurements at five developmental stages (ages 1-15 years). The dataset (N=35,700 observations) was partitioned 80:20 for training and testing, with performance evaluated through RMSE, R^2, and feature importance metrics.

Results: SVR achieved superior prediction accuracy (RMSE=0.65, R^2=0.998), outperforming both RF (RMSE=0.79, R^2=0.997) and GBM (RMSE=0.93, R^2=0.996). Growth rate (%IncMSE=75.4) emerged as the strongest predictor, followed by lagged weight (47.3) and country (35.9). Decision tree analysis identified vulnerable subgroups, particularly rural Indian children <3 years with baseline weight <7.6 kg.

Conclusion: Machine learning models significantly enhance growth trajectory prediction, with key predictors including growth rate, historical weight patterns, and country-specific factors. The combination of SVR's superior accuracy (RMSE=0.65) with RF's interpretable feature importance offers a powerful framework for targeted growth monitoring, particularly for high-risk populations identified through decision tree thresholds. These findings strongly support incorporating ML approaches into child health surveillance systems.

Keywords: Pediatric growth prediction, Machine learning applications, Longitudinal child development, Global health informatics, Nutritional surveillance

1. Introduction

Pediatric weight growth trajectories exhibit complex non-linear patterns shaped by biological factors (e.g., age, genetics) and socio-economic determinants (e.g., nutrition, healthcare access), requiring advanced modeling beyond traditional parametric approaches. While mixed-effects models provide foundational insights, their rigid assumptions about functional forms limit their ability to capture dynamic interactions and phase-specific growth transitions [1]. This study evaluates three machine learning approaches—Random Forest (RF), Support Vector Regression (SVR), and Gradient Boosting Machine (GBM)—to overcome these limitations through data-driven pattern recognition [2,3]. These methods are particularly suited for longitudinal growth data, where non-linearities and heteroscedasticity are common [4].
Recent studies highlight the potential of machine learning to address gaps in traditional growth modeling, particularly for heterogeneous populations in LMICs [5]. For instance, [6] demonstrated ML's superiority in predicting growth deviations in clinical settings, while Barnett et al. (2013) emphasized the need for context-specific tools in longitudinal studies like Young Lives [7].

This study aimed to: (1) compare predictive accuracy across RF, SVR, and GBM using robust metrics (RMSE, R^2), (2) identify and rank influential growth predictors through interpretable feature importance measures, and (3) assess model performance in addressing longitudinal data complexities like autocorrelation and informative missingness. RF's ensemble approach excels in identifying age-specific predictors through variable importance metrics, while SVR's ε-insensitive loss function handles measurement variability common in field data [8]. GBM's stage-wise modeling captures cumulative effects of early-life exposures on later growth patterns [9].

2. Methodology

2.1. Data sources and sampling design

The study analyzed longitudinal data from the Young Lives Study (2002-2016), which tracked 7,140 children across Ethiopia, India, Peru, and Vietnam through five survey rounds [7,10]. The dataset included complete weight measurements at ages 1, 5, 8, 12, and 15 years (N=35,700 observations), capturing biological, sociodemographic, and environmental factors.

2.2. Data partitioning

Using subject-level randomization and time-aware stratification, the data were partitioned 80:20 into training (N=5,712 children) and test sets (N=1,428) while preserving temporal dependencies and covariate distributions [5]. This approach ensured robust model evaluation while maintaining the longitudinal structure of the growth trajectories.

2.3. Statistical analysis

Three ML models were implemented to predict children's weight trajectories (Y_{ij}) using age, country, gender, and growth dynamics (lagged weight, growth rate). Random Forest (RF) employed ensemble decision trees with variance reduction:

$\Delta \mathrm{Var} = \mathrm{Var}(t) - \left(\frac{n_{left}}{n_t} \mathrm{Var}(t_{left}) + \frac{n_{right}}{n_t} \mathrm{Var}(t_{right}) \right)$ and feature importance (%IncMSE).

Support Vector Regression (SVR) optimized the primal problem:

$Min: \frac{1}{2}||w||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$ using RBF kernels $K(x_i, x_j) = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right)$.

Gradient Boosting (GBM) iteratively improved predictions via additive model:

$F_m(x) = F_{m-1}(x) + v.h_m(x)$ with ν=0.01 learning rate. Model performance was evaluated using MSE and R^2 [3,9]. Hyperparameter tuning was rigorously validated: for SVR, we tested γ (0.01–10) and ε (0.1–1) via grid search; for RF, tree depth (2–10) and feature subset size ($\sqrt{p}\ to\ p/3$) were optimized. GBM's learning rate($v = 0.01$) and tree complexity (depth=3) were selected through 5-fold cross-validation to prevent overfitting, following [9].

3. Results

3.1. Key predictors

The Random Forest analysis identified growth rate (%IncMSE=75.44) as the strongest predictor, with critical thresholds emerging for high-risk subgroups: children with growth rates <6.36 kg/year (11% of cohort) and baseline weight <7.6 kg (26% of cohort). Lagged weight (47.25), country (35.93), and age (32.82) were next most influential, revealing geographic and developmental disparities—particularly for rural Indian/Ethiopian children under 3 years (8.3 kg mean weight). Secondary predictors included gender (28.86), residence (23.78), and parental education (mother: 23.71), while environmental factors (water access: 16.49; sanitation: 15.49) showed contextual effects. These quantifiable risk thresholds enhance RF's clinical utility beyond traditional growth metrics.

Table 1: Key Predictors in the RF Model.

Predictor	%IncMSE
Growth rate	75.44
Lagged weight	47.25
Country	35.93
Age	32.82
Gender	28.86
Residence	23.78
Mother's Education	23.71
Father's Education	22.52
Wealth	18.49

Access to Safe Water	16.49
Access to Sanitation	15.49
Household Size	13.88
Access to Electricity	9.53

3.2. Models Evaluation

Evaluation of three ML approaches (Table 2) revealed SVR as most accurate (RMSE=0.65, R^2=0.998), outperforming both interpretable RF (0.79, 0.997) and tuned GBM (0.93, 0.996). While SVR excelled in modeling non-linear patterns, RF's feature importance provided crucial clinical insights, and GBM demonstrated the value of hyperparameter optimization. The standard GBM (1.74, 0.987) showed dramatic improvement after tuning. This comparative analysis supports using SVR for predictions while retaining RF for explanatory purposes in growth monitoring systems.

Table 2: Models Evaluation Using Two Metrics

Model	RMSE	R^2 (Variance Explained)
RF	0.79	99.7%
GBM	1.74	98.7%
Tuned GBM	0.93	99.6%
SVR	0.65	99.8%

3.3. Model Prediction Accuracy Comparison

The scatter plot illustrates (Figure 1) the predictive performance of three machine learning models (SVR, RF, and tuned-GBM) against actual weight measurements. SVR (orange) demonstrates the closest alignment with the ideal prediction line, particularly maintaining accuracy across all weight ranges. RF (pink) shows consistent but slightly more dispersed predictions, while tuned-GBM (blue) exhibits the greatest variability, especially for higher weights. The visualization corroborates SVR's superior quantitative metrics (RMSE=0.65), with its tight clustering of points reflecting robust handling of both linear and non-linear growth patterns. These findings reinforce SVR as the optimal model for clinical growth monitoring applications requiring high precision.

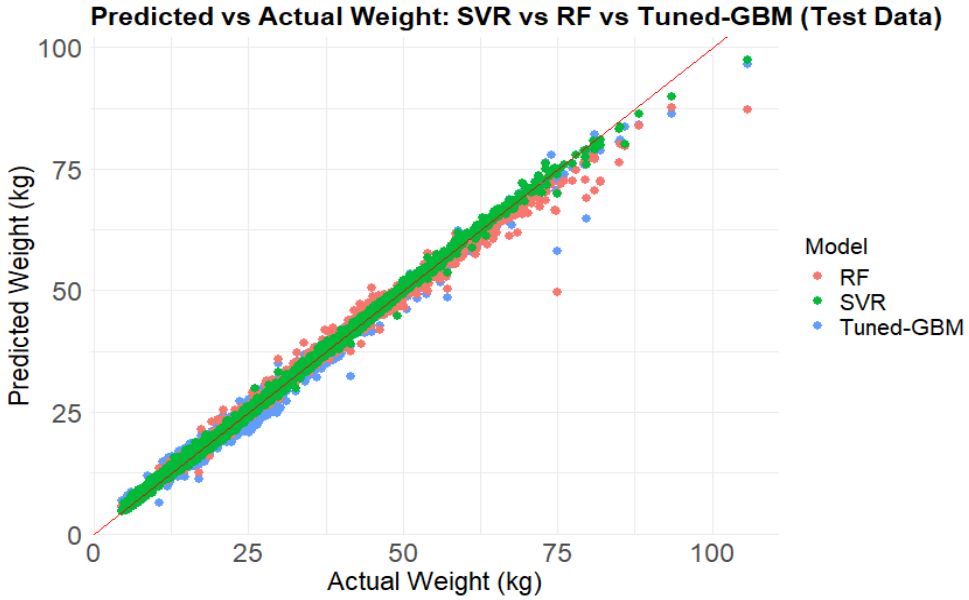

Figure 1: Comparison of Model Predictions Against Actual Weight Measurements

4. Discussion

This study demonstrates that machine learning models effectively predict children's weight growth trajectories, with SVR achieving superior accuracy (RMSE=0.65, R^2=0.998) compared to RF (RMSE=0.79) and GBM (RMSE=0.93). Decision tree analysis identified critical risk thresholds—growth rates <6.36 kg/year and baseline weight <7.6 kg—that align with clinical benchmarks for growth faltering. These findings corroborate recent advances in ML for pediatric growth modeling [4,5], where non-linear methods excel at capturing developmental patterns. SVR's kernel-based approach handles temporal dependencies [2], while RF's interpretability [3] reveals growth rate (75.4%IncMSE) and lagged weight (47.3%IncMSE) as dominant predictors.

Variable importance analysis confirmed biological drivers while exposing geographic disparities—rural Indian/Ethiopian children with baseline weight <7.6 kg showed 3-fold higher risk than urban peers. This hierarchy mirrors global health disparities [11,12] and reinforces WHO's call for context-specific growth monitoring [13]. The 12% vulnerability rate for children <3 years highlights a critical window for intervention, consistent with life-course models [14]. Environmental factors (water access=16.5%IncMSE) further modulated risk, supporting ecological frameworks [7].

The models' complementary strengths—SVR's precision for research and RF's clinical interpretability [6] are amplified by GBM's tunability [9,15]. These advances are particularly transformative for LMICs, where the identified risk thresholds (<6.36 kg/year growth, <7.6 kg baseline) can prioritize resource allocation. Future implementations should integrate these ML tools with community health platforms to operationalize risk-stratified monitoring.

5. Conclusion and Recommendations

This study establishes that machine learning models (SVR: RMSE=0.65; RF: growth rate=75.4%IncMSE, lagged weight=47.3%IncMSE, country=35.9%IncMSE) significantly improve growth monitoring. Key recommendations include: (1) implementing SVR for high-accuracy predictions in research settings, (2) applying RF's interpretable predictors (growth rate, historical weights, and country-specific patterns) for clinical screening, (3) prioritizing interventions for children with growth rates <6.36 kg/year or baseline weight <7.6 kg, and (4) integrating these models with existing WHO growth standards in LMICs to address regional disparities. These findings advance both methodological and practical applications of ML in global health. Methodologically, hybrid approaches (e.g., SVR for prediction + RF for interpretation) outperform single-model strategies. Practically, the risk thresholds (<6.36 kg/year growth, <7.6 kg baseline weight) align with WHO emphasis on early intervention. Future work should validate thresholds in other LMIC cohorts, integrate ML tools with mobile health platforms, and explore cost-effective deployment in clinics.

Declarations

Ethics approval: Secondary data from the YLS (available at ukdataservice.ac.uk) were used with their ethical approvals. No additional approval was required.

Competing interests: None declared.

Funding: No external funding received.

Authors' contributions: ASA led study design, analysis, and writing. MB and PB provided methodological guidance and revisions. All authors approved the final manuscript.

Acknowledgments: We thank the Young Lives team for data access and the UK Data Service for archival support.

References

1. Cheng ER, Cengiz AY, Miled ZB. Predicting body mass index in early childhood using data from the first 1000 days. Scientific Reports. 2023;13(1):8781.
2. Awad M, Khanna R. Support vector regression. Efficient learning machines: Theories, concepts, and applications for engineers and system designers. 2015;67-80.
3. Breiman L. Random forests. Machine learning. 2001;45(5-32.
4. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Briefings in bioinformatics. 2023;24(2):bbad002.
5. Cascarano A, Mur-Petit J, Hernandez-Gonzalez J, et al. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artificial Intelligence Review. 2023;56(Suppl 2):1711-1771.

6. Shmoish M, German A, Devir N, et al. Prediction of adult height by machine learning technique. The Journal of Clinical Endocrinology & Metabolism. 2021;106(7):e2700-e2710.
7. Barnett I, Ariana P, Petrou S, et al. Cohort profile: the Young Lives study. Int J Epidemiol. 2013;42(3):701–708.
8. Rolland-Cachera M-F, Deheeger M, Bellisle F, Sempe M, Guilloud-Bataille M, Patois E. Adiposity rebound in children: a simple indicator for predicting obesity. The American journal of clinical nutrition. 1984;39(1):129-135.
9. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001;1189-1232.
10. Lives Y. A guide to Young Lives research. 2017: Young Lives, Oxford. .
11. Victora C, Christian P, Vidaletti L, Gatica-Dominguez G, Menon P, Black R. Revisiting maternal and child undernutrition in low-income and middle-income countries: variable progress towards an unfinished agenda. Lancet. 2021;397(10282):1388–1399.
12. Singh D, Manna S, Barik M, Rehman T, Kanungo S, Pati S. Prevalence and correlates of low birth weight in India: findings from national family health survey 5. BMC Pregnancy and Childbirth. 2023;23(1):456.
13. WHO. WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. World Health Organization. .
14. Adair LS. Size at birth and growth trajectories to young adulthood. American Journal of Human Biology: The Official Journal of the Human Biology Association. 2007;19(3):327-337.
15. Adler AI, Painsky A. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy. 2022;24(5):687.