Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Using Machine Learning to Predict Childhood Obesity from Data from First 1000 days of Life
Principal Investigator:	Uzma Rani, Aamer Imdad, William Story, Donna Santillan
Prepared by (IIAI):	Yanan Liu
Other investigators:	Samantha Pothitakis, Yash Vora
Date:	

Were specific aims fulfilled:	Y
Readiness for extramural proposal?	Y
If yes … Planned submission date	10/2025
Funding agency	NIH
Grant mechanism	R21
If no … Why not? What went wrong?	

Brief summary of accomplished results:

The goal of this project was to develop and validate a machine learning algorithm that can accurately predict a child's BMI at 3 years of life. This prediction is based on the data from the patient's first 1,000 days of life. After cleaning the data that included information about the mother and child, features were chosen based on a LASSO and T-test pipeline from about 13,000 patients. One of the two final models had a mean absolute error of predicted BMI of 0.773 with a standard deviation of 0.656. BMI was ranging from 10.9 to 23 in our cohort. This model contained around 90 features. The other model contained 14 features, selected by domain expert from the 90-feature set based on clinical relevance. That model had a mean absolute error of predicted BMI of 0.787 with a standard deviation of 0.680. These had a 19% and 18% decrease in MAE, respectively, compared to the Cheng et al 2023 paper. The results of this study are considered successful, and the model will be tuned finer to produce better results.

In addition, machine learning model was used to predict obesity (yes/no) based on BMI>18.5 at 3 years. The prediction accuracy is 83% and F1 is 0.83 using 14 selected features.

Research report:

Aims (provided by PI):

1. To Predict Childhood Obesity from Data from First 1000 days of Life.

Data:

This project used two different datasets: the IHK and EHR. This data was collected at the University of Iowa from mothers and children who were there receiving care and consented to

having their data used for research. The IHK data contained information about the mothers' diagnoses, medications, social history, and delivery type. It had information on newborn diagnoses as well as information from Bright Future Flowsheets that tracked children's generalized habits (feeding, playing, etc.). These sheets had over 30,000 unique patients. The EHR data contained data on the children's labs, medications, diagnoses, weight, and height. This data contained information on over 27,000 unique patients. The IHK data came in 17 different csv sheets. On the sheets that were not time series data, the statistical values for numerical columns were calculated. For categorical columns, histograms were made to visualize the most common values. Dr. Imdad went through each column and commented on which would be significant enough to put into the LASSO feature selection algorithm. For time series data sheets (maternal and newborn diagnoses, screenings, social history, and medications) pivot tables were made. The unique responses from the selected variable (i.e. diagnosis code) for that sheet were made into columns and the rows contained the number of times the patient had received that response (i.e. number of times diagnosed). Once the selected sheets in the IHK data were merged, the dataset contained over 19,000 unique patients and 40,000 unique features. The EHR data contained time series data. This data was pivoted where the columns were the time points of the recorded measurement (i.e. bmi_t1, bmi_t2). The time points for each age are shown in Table 1 below.

Table 1: Bounds for classifying age for BMI

Age (years)	Lower Bound	Upper Bound
1	0.8	1.2
2	1.8	2.2
3	2.5	3.49
4	3.5	4.49
5	4.5	5.5

This was done to account for the appointments before or after a child's birthdate. The EHR and IHK dataset were merged where the final dataset contained over 19,000 unique patients. After dropping the rows where the BMI at age 3 was missing, there 13,050 unique patients remaining. The missing data in the merged file after feature selection was imputed by taking the mean value for numerical features and the mode value for categorical.

AI/ML Approach:

To guide our methodology, we adopted the approach outlined in Cheng et al. (2023) in their paper "Predicting childhood obesity using machine learning: The importance of feature selection" as a baseline, since their study addressed a similar objective—predicting childhood BMI using early life data. Using their framework allowed us to evaluate how well their methods generalized to our dataset while providing a solid foundation for model development.

i. Feature Selection Process

We implemented a multi-step feature selection strategy to enhance model performance, reduce overfitting, and ensure interpretability. This process began by combining clinically related features (e.g., diabetic, preeclampsia, maternal antibiotics, and obesity codes), resulting in an initial set of 40,176 features. As part of preprocessing, missing values were imputed using mean imputation for numerical variables and mode imputation for categorical variables, preserving as much information as possible while maintaining consistency across the dataset.

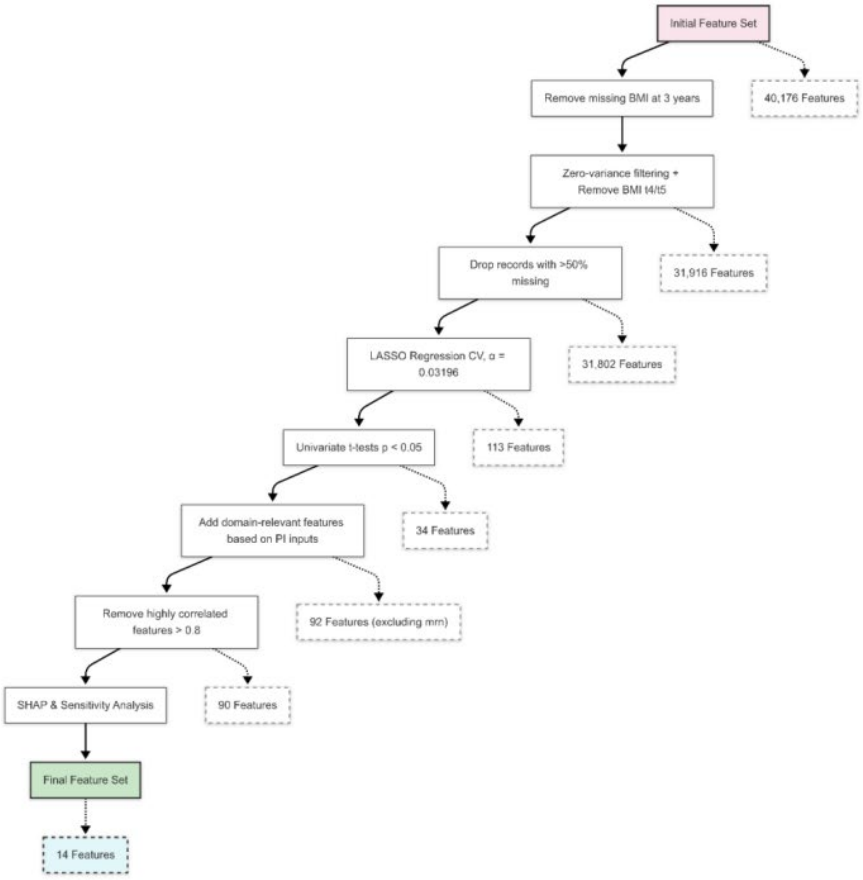

Figure 1. Feature Selection Workflow - A stepwise reduction from an initial set of 40,176 features to a final set of 14 using a combination of data cleaning, zero-variance filtering, LASSO regression, univariate testing, domain knowledge, correlation filtering, SHAP and sensitivity analysis.

This combination of data-driven techniques and expert input ensured a robust, interpretable feature set tailored to both the dataset characteristics and domain-specific insights.

ii. Modeling Approach

We implemented Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel, following a strategy like Cheng et al. SVR was chosen for its robustness and ability to model complex, non-linear relationships in continuous outcomes such as BMI. Two models were developed:

• One with the 90 selected features

• One with a reduced set of 14 clinically relevant features and which didn't act as confounders

Both models were trained and evaluated using the same pipeline for comparability. The 14 feature model contained the following features: BMI at year 1, BMI at year 2, birth weight, large gestational age for the child, maternal obesity diagnosis, maternal placenta insufficiency, maternal antibiotic prescription, maternal preeclampsia, child diabetes, child Amoxicillin prescription, food other than breastmilk or formula for 4 and 6 month olds, types of milk eaten for 12 month olds, and the number of servings of milk eat for 12 month olds.

iii. Sensitivity Analysis

The model using 90 features achieved a mean absolute error (MAE) of predicted BMI of 0.773 (SD = 0.656). A second model, built using a curated set of 14 features, produced a comparable MAE, indicating that the excluded 76 features had minimal impact on overall performance.

The final feature set was led by prior BMI measurements, which are strong indicators of future BMI, along with other variables reflecting established clinical patterns related to obesity and features that were of interest to the PIs. This result highlights the strength of the selected features and confirms that a more streamlined, interpretable model can achieve similar predictive performance.

Experimental methods, validation approach:

After the feature selection process, 90 features were put into the model. The model used a 5-fold cross validation technique to determine the best parameters for the model. The folds were randomly generated equal portions of the data. The C and gamma values were chosen by testing 16 combinations of the parameters. The four gamma test values were 0.0001, 0.001, 0.01, and 0.1. The four C test values were 1, 10, 100, and 1,000. After running the full dataset on this model with a different combination every time, the best parameter combination was found to be gamma = 0.0001 and C = 100. For predicting BMI, a new model was then run with 80% of the data used for training, and 20% used for testing. After the model with 90 features was run, the Principal Investigators chose 14 features to put into another model. This model was trained and validated in the same way as the larger feature model. For predicting obesity, Random Forest model is used with 14 selected features. BMI was ranging from 10.9 to 23 in our cohort.

Results:

To evaluate our 14-feature SVR model, we plotted actual vs. predicted BMI values in Figure 2 and overlaid a linear regression for interpretability $\hat{y}$ = 8.69 + 0.47x (from linear fit), R2=0.46 (from SVR predictions). This equation approximates the SVR output but does not reflect its underlying RBF kernel, which lacks a closedform expression. Since SVR minimizes a margin-based loss rather than R2, lower R2 values can occur even with strong predictive performance. Our model explained 46% of BMI variance and achieved a test MAE of predicted BMI of 0.787 (SD = 0.680.)—an ~18% improvement over Cheng et al. (2023). This suggests our SVR model provides more precise individual predictions, valuable for early obesity risk detection. While an

R^2 of 0.46 may appear modest, it is important to recognize that BMI is influenced by a complex interplay of genetic, behavioral, and environmental factors. Moreover, mean absolute error is often a more clinically relevant metric than R^2, as it directly reflects how close predictions are to the true values in real units. In this context, an average error of less than around 0.8 can meaningfully support early risk stratification and intervention planning. The fact that our SVR model achieves this level of precision—using only 14 features—demonstrates its strength as a baseline model for early BMI prediction and risk assessment.

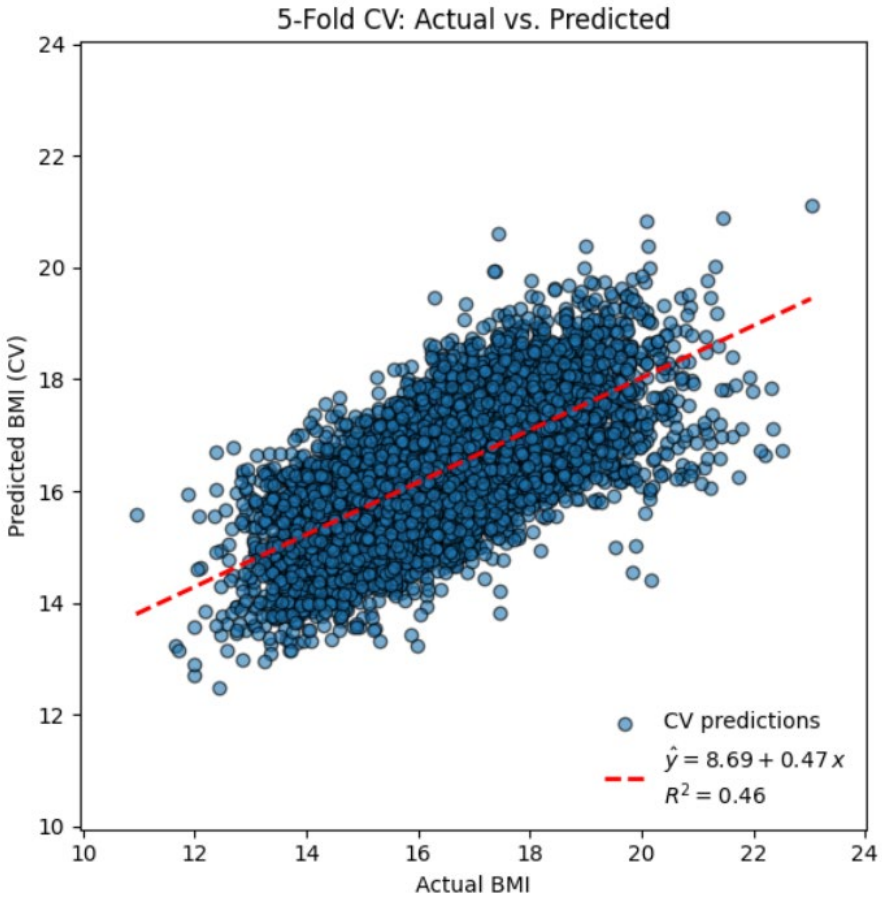

Figure 2. Actual vs. Predicted BMI with equation of fit.

However, like prior studies, our model underpredicts at the high end of the BMI range, likely due to fewer extreme cases in the training data. This trade-off reflects the model's focus on minimizing overall MAE, favoring mid-range accuracy over extremes. Nonetheless, its strong performance across the majority of the BMI spectrum makes it a valuable starting point, with room for further refinement through targeted sampling or tailored modeling for high-risk subgroups.

To enhance model transparency and understand feature influence on BMI predictions, we used SHAP (Shapley Additive Explanations) to measure each feature's contribution to individual predictions.

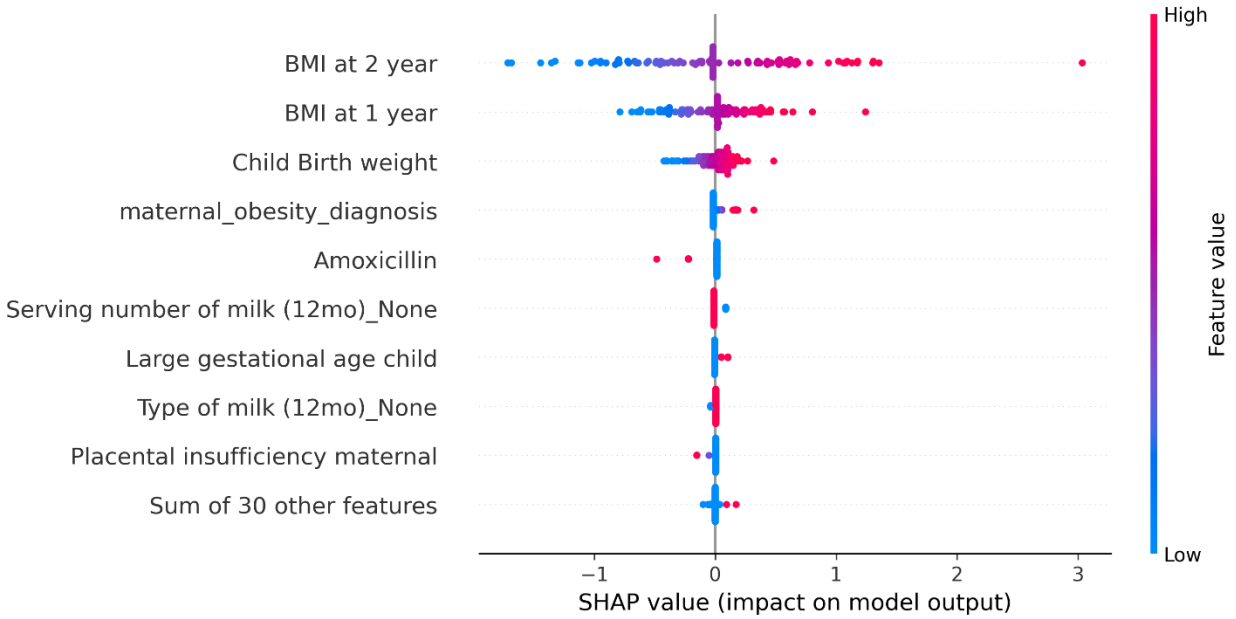

Figure 3. SHAP Beeswarm Plot - Top features ranked by impact on model output, with color showing feature value (red = high, blue = low).

The SHAP beeswarm plot revealed:

- Higher values of historical BMI (bmi_t2 and bmi_t1) strongly push predictions toward higher BMI, indicating a child's previous BMI is the most critical predictor.
- Higher birth weight is associated with increased BMI predictions, aligning with known risk factors for early childhood obesity.
- Maternal health indicators (obesity diagnosis) and certain prescriptions (AMOXICILLIN) impact predictions to a lesser extent.
- While we initially selected 14 key features, the SHAP plot reflects a larger number (e.g., "Sum of 31 other features") due to one-hot encoding of categorical variables, which expands single features into multiple binary columns. The spread of points for these features indicates variability in how they influence individual cases— sometimes increasing, sometimes decreasing predicted BMI depending on context.

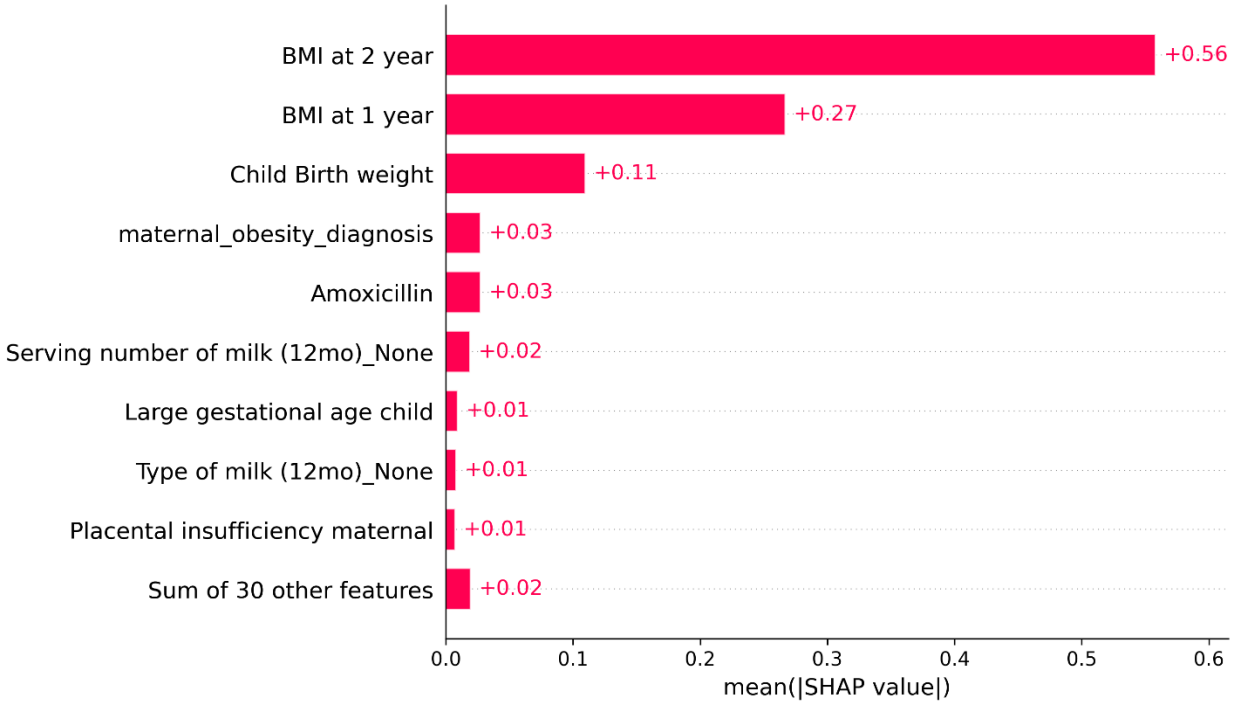

Figure 4. SHAP Bar Plot - Mean absolute SHAP values highlight the most influential features on model predictions, led by BMI at timepoints 2 years (t2) and 1 year (t1).

The SHAP bar plot reinforces our findings by ranking features based on their average impact. It clearly shows that past BMI values (bmi_t2 and bmi_t1) overwhelmingly dominate the prediction process, followed by birth weight. The remaining features, including maternal diagnoses and medications, contribute modestly but highlight the multifactorial nature of BMI development, where both biological history and environmental/medical factors interact.

For prediction obesity, we used Random Forest, and the prediction accuracy is 83% and F1 is 0.83 using 14 selected features.

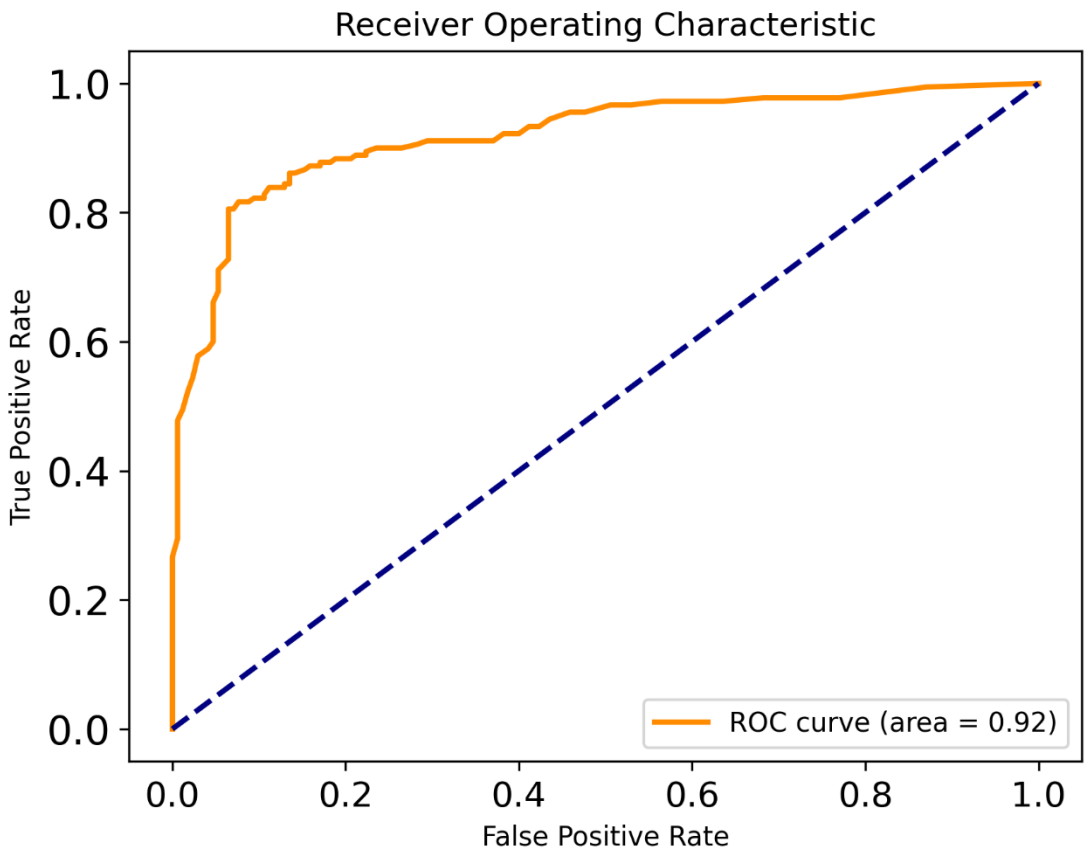

Figure 5. ROC curve of Random Forest.

Ideas/aims for future extramural project:

Future work should focus on improving prediction accuracy for higher BMI values, as this is critical for obesity risk identification. Techniques such as data balancing, weighted loss functions, complicated imputation methods to impute the missing data, or switching to a classification approach could be explored. Given the strong baseline model developed, trying alternative models could be valid, but meaningful improvements would likely require additional data preprocessing, better feature engineering, and potentially richer data sources.

Publications resulting from project:

NASPGHAN abstract submitted

References:

1. Cheng, E. R., Cengiz, A. Y., & Ben Miled, Z. (2023). Predicting body mass index in early childhood using data from the first 1000 days. Scientific Reports, 13, 8781

2. Memon, S. M. Z., Wamala, R., & Kabano, I. H. (2023). A comparison of imputation methods for categorical data. Informatics in Medicine Unlocked, 42, 101382

3. Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence, 33(10), 913–933