Journal of Data Science 0 (0), 1–21
??? 2025

DOI: 10.6339/25-JDS1197
Data Science in Action

Predicting Stunted Growth in Two Year Old Bangladeshi Children via the Super Learner

Heather L. Cook[1,*], Jennie Z. Ma[2], Daniel M. Keenan[2], Jeffrey R. Donowitz[2], Beth D. Kirkpatrick[3], Rashidul Haque[4], Uma Nayak[2], and William A. Petri, Jr.[2]

[1] *Evansville, IN, University of Southern Indiana, United States*
[2] *Charlottesville, VA, University of Virginia, United States*
[3] *Burlington, VT, University of Vermont, United States*
[4] *Dhaka, Bangladesh, International Centre for Diarrhoeal Disease Research, Bangladesh*

Abstract

Stunted growth in children is a worldwide issue which may cause long term problems for individuals stunted as early as two years of age. However, predicting stunted growth with accuracy is quite complex, but machine learning poses a distinct advantage in this regard. While several techniques are available for predictive modeling, the Super Learner stands out as an ensemble method that integrates multiple algorithms into a single predictive model with enhanced performance. In this study, the Super Learner model, comprising generalized linear model, bagged trees, random forests, conditional random forest, stochastic gradient boosting, Bayesian additive regression trees, neural networks, and model averaged neural networks, achieved high performance with high area under the receiver operating characteristic curve, Brier Score, and the minimum of precision and recall values. However, after analyzing the results from cross validation, the final model selected was the Bayesian additive regression trees. Within the final model, the height-for-age z-score at one year, income, expenditure, anti-lipopolysaccharide antibody at week 6 and at week 18, plasma retinol binding protein at week 6, plasma soluble cluster designation 14 at week 18, fecal Reg 1B at week 12, vitamin D at week 18, mother's weight and height at enrollment, fecal calprotectin at week 12, fecal myeloperoxidase at week 12, number of days of diarrhea through the first year of life, and the number of days of exclusive breastfeeding through the first year of life emerged as the top important variables for predicting stunted growth at two years of age.

Keywords *children's health; classification; ensemble method; machine learning*

1 Introduction

Childhood stunting and undernutrition is a global problem, affecting approximately 165 million children worldwide and contributing to 45% of deaths among children under the age of five (Prendergast and Humphrey, 2014). The first few years of a child's life are critical as they strongly predict future success and health. This makes early childhood a crucial window for intervention in malnourished children. Preventing stunting and malnutrition during this period can lead to significant improvements in overall well-being for a child (Martorell and Zongrone, 2012). Previous studies have shown that children who experienced stunted growth within the

[*] Corresponding author. Email: hlcook1@usi.edu.

© 2025 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.

Received November 14, 2024; Accepted July 2, 2025

first two years of life may succumb to irreversible damage such as lower weight in childhood, smaller adult stature, lower educational attainment, reduced income in adulthood (Victora et al., 2008; Dewey and Begum, 2011; Hoddinott et al., 2008). The ambitious World Health Assembly has created a target of reducing stunting from 2010 to 2025 by forty percent (Prendergast and Humphrey, 2014).

Malnutrition is multifactorial, influenced by factors such as maternal health variables, enteric and systemic inflammation, enteric infection, and lack of nutrient intake. The pathogenesis of malnutrition is not fully explained, and it used to be thought that even with the implementation of all known interventions, growth stunting would only be reduced by 36% (Bhutta et al., 2008). Complicating research to understand stunting is the fact that many potential contributing variables are interrelated. Thus, in clinical research to understand the origins of malnutrition, it is critical to accurately predict stunting with the most appropriate method(s) such that important risk factors can be identified and a suitable predictive model can be established.

Many statistical methods can be used to develop a predictive model, each with different assumptions and varying accuracy. For example, multiple linear regression analysis is a traditional method for predictive modeling, but only evaluates the linear relationship between the response and independent predictors. With the development of modern statistical computing, tree-based statistical methods and machine learning (ML) algorithms have gained popularity in clinical studies for their nonparametric nature, capability to assess nonlinear relationships, and particularly their greater predictability.

ML is the science of inducing computers to act without being explicitly programmed, i.e., using computer algorithms to parse data, learn from it, and then make a determination or prediction about something in the world. Many ML algorithms, especially the tree-based methods, are of particular interest for their nonparametric nature and better predictability, such as bagged trees (Bagg) from Breiman (1996), random forests (RF) from Breiman (2001), conditional random forests (CRF) from Strobl et al. (2008), stochastic gradient boosting (GBM) from Friedman (2001), Bayesian additive regression trees (BART) from Chipman et al. (2010), neural networks (NNet) originally introduced in the 1960s by Widrow and Hoff (1960) and Rosenblatt (1962) but further explained in Chapter 5 of Ripley (1996), and model averaged neural networks (MANN) based on averaging many NNet (Ripley, 1995; Perrone and Cooper, 1993).

However, since each method has its own limitations and one size can't fit all, the Super Learner (SL) algorithm was developed based on a collection of algorithms. The SL is a prediction method designed to identify the optimal single algorithm or combination of these algorithms that minimizes cross validated risk (van der Laan et al., 2007). It has been applied extensively in various clinical studies, including predicting the in vitro phenotypic susceptibility of HIV, thromoboembolic stroke, and animal behavior, as well as interpreting genotypic resistance testing and estimating the propensity score (Sinisi et al., 2006; Brooks, 2012; Ladds et al., 2017; Houssaini et al., 2012; Pirracchio and Carone, 2016). These studies shown that the SL performed at least as well as the individual methods in its library, and prediction with individual algorithms or ML techniques can be improved by using the SL method. However, different datasets lead to distinct models with some methods performing better on particular datasets, but the SL can help choose the most effective method for a given dataset (Ju et al., 2016).

Few studies in malnutrition research have used the SL, and none of them have predicted stunted growth in the manner presented here (Mertens et al., 2023; Butzin-Dozier et al., 2025; Mursil et al., 2023). Thus, utilizing the data from the Performance of Rotavirus and Oral polio Vaccines In DEveloping countries (PROVIDE) birth cohort, this study aims to not only predict stunting in early life, but also to evaluate the relative performance of several ML algorithms.

The PROVIDE study was designed to evaluate oral vaccine efficacy and the impact of environmental enteropathy (EE) on vaccine failure and malnutrition in Bangladeshi children (Kirkpatrick et al., 2015). EE, a common syndrome in children from low- and middles-income countries, is characterized by small intestine inflammation with reduced villi height, intestinal barrier dysfunction, and reduced nutrient absorption (Naylor et al., 2015). This is a subclinical condition caused by continuous exposure to food or water contaminated by fecal matter, often in areas where poor sanitary conditions are prevalent, and diarrhea is among the leading causes of death for children under 5 years old.

The PROVIDE birth cohort data consisted of clinical characteristics (birth anthropometry, family socioeconomic status, and maternal health), and a comprehensive set of biomarkers from early-stage fecal and blood samples. While the PROVIDE data was previously explored to predict height-for-age z-score (HAZ) with CRF and a form of linear regression, the prediction of stunting has not been evaluated (Donowitz et al., 2018). The objective of this study was to develop a predictive model using identified clinical characteristics and biomarkers associated with malnutrition and stunted growth. Given that many advanced statistical and ML methods are available, we evaluated the relative performance of several representative methods and assessed their prediction accuracy for stunted growth at two years of age. Ultimately, a predictive model with high accuracy can enhance our understanding of early life risk factors and biomarkers associated with future stunting, paving the way for development of effective malnutrition intervention strategies.

2 Data

The PROVIDE study was a randomized controlled clinical trial with a 2-by-2 factorial design to investigate the efficacy of Rotavirus and Oral Polio Vaccines. Detailed study design, recruitment and follow-up of the PROVIDE cohort were described previously in Kirkpatrick et al. (2015) and Donowitz et al. (2018). Briefly, the PROVIDE birth cohort consisted of 700 newborns from Mirpur, a densely populated urban neighborhood in Dhaka, Bangladesh from May 2011 to November 2014. Most children were from lower socioeconomic backgrounds. Children were enrolled within 7 days after birth and followed over a two-year period by twice weekly home visits and scheduled clinical visits. This study was approved by the Ethical Review Board of the ICDDR,B (FWA 00001468) and the Institutional Review Boards of the University of Virginia (FWA 00006183) and the University of Vermont (FWA 00000727).

The primary outcome of interest in this analysis was stunting at two years, defined as a HAZ at two years old below -2. HAZ as a continuous response was explored as the secondary response with results included in Section 3 of the Supplemental Material. HAZ is a measure of child height normalized by age and gender by the World Health Organization (WHO) Multicenter Growth Reference Study Child Growth Standards and is a commonly used measurement for malnutrition since it captures cumulative effects through childhood. If a subject had incomplete data, then they were excluded from the analysis. Additionally, outliers in predictors defined as any value over five standard deviations from the mean were excluded from the analysis. Outliers for the HAZ were not removed prior to creating the stunted growth binary response. After dealing with missing values and outliers, 391 children were included in our analysis and then the quantitative variables were standardized.

A comprehensive set of biomarkers collected from fecal and blood samples around the time of primary vaccination were included in the PROVIDE data. In addition to the biomarkers of nutrition and systemic inflammation, variables measuring socioeconomic status and maternal

Table 1: Enrollment characteristics with mean ± standard deviation or count (percent).

Variable	Overall	Stunted	Not Stunted
	$n = 391$	$n = 116$	$n = 275$
Child HAZ at Enrollment	−0.90 ± 0.87	−1.39 ± 0.75	−0.69 ± 0.83
Child WAZ at Enrollment	−1.26 ± 0.84	−1.67 ± 0.80	−1.11 ± 0.81
Child WHZ at Enrollment	−1.30 ± 1.14	−1.27 ± 1.19	−1.31 ± 1.12
Expenditure (Thousands of Taka[1])	11.98 ± 7.77	10.11 ± 6.58	12.77 ± 8.10
Income (Thousands of Taka[1])	13.30 ± 1.00	10.82 ± 8.58	14.34 ± 10.38
Mother's Weight at Enrollment (kg)	49.71 ± 9.35	45.41 ± 6.92	51.52 ± 9.65
Mother's Height at Enrollment (cm)	150.49 ± 5.73	148.64 ± 5.23	151.26 ± 5.77
No Maternal Education	127 (32.48%)	53 (45.69%)	74 (26.91%)
Male	196 (50.13%)	55 (47.41%)	141 (51.27%)
Treated Water	241 (61.64%)	55 (47.41%)	186 (67.64%)
Septic Tank/Toilet Facility	212 (54.22%)	59 (50.86%)	153 (55.64%)
No Open Drain Outside Home	234 (59.85%)	64 (55.17%)	170 (61.82%)
No Shared Toilet Facility with Other Households	62 (15.86%)	14 (12.07%)	48 (17.45%)

[1] 1 Taka = 0.0082 USD

health were included, resulting in a total of 45 predictors (Naylor et al., 2015). The full list of variables collected along with summary statistics are listed in Tables S1 and S2. The C-reactive protein (CRP) index was created from longitudinal CRP measurements to assess sustained inflammation. It was defined as the total number of times an individual CRP measurement was in the top 50th percentile at 6, 18, 40, and 53 weeks of age. Given our goal of evaluating the impact of elevated biomarkers on stunted growth, skewed biomarkers were categorized based on their distributions. For example, the cytokines (such as ILs, MIP1β, and TNFα) were classified into three categories: lower 50th percentile, 50th to 75th percentile, and upper 25th percentile (75th to 100th percentile) due to their extreme skewness.

Table 1 shows the descriptive statistics about enrollment characteristics by the stunting status at two years. On average, children in this cohort had lower HAZ, WAZ (weight-to-age z-score), and WHZ (weight-to-height z-score) at enrollment compared to WHO median (i.e., their mean values < 0). Further, those who were stunted at two years had even worse HAZ and WAZ scores, a higher percentage of mothers with no education, and a lower percentage of households with access to water treatment. A complete list of predictors included for the analyses with numerical summaries is included in Tables S1 and S2.

As mentioned previously, some variables were highly correlated by Spearman's correlation. See the correlation heatmap in Figure S1. Income and expenditure were highly correlated at 0.97 while the correlation between mother's weight and height was 0.48. Due to the high correlations among HAZ, WAZ, and WHZ at one year (ranging from 0.4 to 0.88), only HAZ at one year is used as a predictor. Mother's weight at enrollment was fairly correlated (0.41) with HAZ at one year. Other notable correlations were between calprotectin and myeloperoxidase (MPO) at week 12 (0.36), and between neopterin and alpha-1 antitrypsin at week 12 (0.31). Interestingly, exclusive breastfeeding was negatively correlated with vitamin D (−0.26 and −0.25 at weeks 6 and 18 respectively). Given the correlated nature of these clinical characteristics and biomarkers, we aim to develop a meaningful predictive model for stunted growth with high accuracy and performance with modern statistical and ML approaches.

3 Methods

In this section, we introduce several advanced statistical and ML methods for predictive modeling. We then describe the evaluation methods used to assess predictive performance in Section 3.2. The application of these prediction models and their performance evaluation are presented in Section 4, followed by a discussion in Section 5.

3.1 ML Methods for Classification

Many modern statistical and ML methods are available for predictive modeling. Given the large volume of data and complex intercorrelated predictors in the PROVIDE study, ML methods are natural options for predicting stunting outcome due to their ability to "learn" information and natural patterns directly from data without relying on predetermined formulas or equations. However, there is no best method or one size fits all; this is where the SL proves useful.

Although many methods can be implemented within the SL algorithm, we chose flexible nonparametric nonlinear methods due to the complex structure of our data along with a basic classical approach of a generalized linear model (GLM). The ML algorithms of interest included GLM, Bagg, RF, CRF, GBM, BART, NNet, and MANN. These methods are summarized in Table S3 and explained in detail below. All of these methods except GLM, NNet, and MANN are tree-based, where multiple decision trees are built then aggregated. For the analyses, R (R Core Team, 2023, v4.3.0) was used and R packages included `SuperLearner` (Polley et al., 2024, v2.0-29), `bartMachine` (Kapelner and Bleich, 2016, v1.3.4.1), `gbm` (Greg and Developers, 2024, v2.1.9), `nnet` (Venables and Ripley, 2002, v7.3-19), `randomForest` (Liaw and Wiener, 2002, v4.7-1.1), `party` (Strobl et al., 2008, v1.3-15), `ipred` (Peters and Hothorn, 2023, v0.9-14), and `caret` (Kuhn, 2008, v6.0-94). The `caret` (Classification And Regression Training) package was used to train ML algorithms in the SL library.

Additionally, per each method, the variable importance (VIMP) can be explored, however each method has it's own approach for calculating VIMP, as explained below.

3.1.1 Generalized Linear Models

GLM is a family of models but in our study, we will use multiple logistic regression to classify the observations. Our goal is to predict the probability that an observation belongs to a particular class. Define p as the probability of an event and thus $p/(1-p)$ is the odds of the event. With multiple logistic regression, the log odds of the event is modeled as a linear function of predictors. We can estimate p from a nonlinear sigmoidal function of that constrains the probability estimates between zero and one. Although p is a nonlinear function of the predictors, it still produces linear classification boundaries. Maximum likelihood estimation is used to estimate the parameters just like in ordinary least squares regression. A strength of this method compared to others in the SL library is the ease of interpreting the significance of predictors (Kuhn and Johnson, 2016, pgs. 282–287).

Given the large number of predictors, it is important to avoid including all of them in a single model and instead applying variable selection approaches. In this case, we applied stepwise selection with logistic regression using Akaike Information Criterion (AIC) as the criterion for adding or dropping predictors. According to AIC, the best-fitting model explains the most variation with the fewest predictors, so lower AIC is preferred. Stepwise selection starts with an empty model and adds the predictor that lowers AIC the most, and continues adding or removing predictors based on AIC until no further improvement is possible. The VIMP is then determined by the coefficient of the selected variables in the GLM model.

3.1.2 Bagged Trees

We also implement methods such as Bagg that aggregate multiple individual trees to reduce variability. Bagg is an ensemble method that aggregates the decision trees generated from bootstrapped samples, as described by Breiman (1996). Each of the trees is grown deep with low bias but high variability between the trees. Since there are no parameters to be trained, trees are grown deep, and each split considers all predictors, the model tends to overfit. However, the variability is reduced when the results of individual trees are aggregated. For a binary outcome, each tree casts a vote for the predicted class. The predicted probability for a new observation is the proportion of votes over all the trees, and the observation is classified to a category based on a decision threshold, e.g. 0.5.

One general downfall to aggregating trees is the loss of easy interpretations. Another issue, particularly for Bagg, is that the trees are not completely independent of one another since all predictors are considered at each split for every single tree in the ensemble. Tree correlation may prevent the method from optimally reducing the variance of the predictions since each tree can have comparable structures due to the underlying relationships as discussed in (Kuhn and Johnson, 2016, pgs. 198–199).

VIMP can be calculated for a Bagg model by tracking the overall reduction in the optimization criteria for each predictor. Here, the Gini index, a measure of impurity, can be used. The overall reduction in the Gini index is averaged across all trees in the ensemble to calculate a single VIMP value per predictor. For the binary case, the Gini index for a given node is defined as $p_1(1-p_1)+p_2(1-p_2)$ where p_1 and p_2 are the class probabilities. The Gini index is minimized when either class probability is zero (least impure) and maximized when the class probabilities are equal (most impure) (Kuhn and Johnson, 2016, pgs. 370–371).

3.1.3 Random Forests and Conditional Random Forests

RF improves upon Bagg by using a randomly selected subset of predictors for each split instead of choosing from all predictors per each split. This random selection reduces the correlations between trees. Each tree is grown to the maximal depth and contributes equally to the final model. Because trees are less correlated, the number of trees created does not attribute much to overfitting (Breiman, 2001).

When predictors are highly correlated, the importance of predictors in RF may be overestimated, causing the correlated predictors to appear more important than those uncorrelated (Strobl et al., 2008, 2009; Boulesteix et al., 2012). The CRF method takes into account the correlations among predictors using an unbiased splitting criterion, and thus appropriately evaluates the impact of a single variable in predicting the outcome.

A tuning parameter for both RF and CRF is the number of predictors to be randomly chosen at each decision node. It is important to tune this parameter since a small number of predictors can lead to choosing variables that are sub optimal at each split with a loss of information (Boulesteix et al., 2012; Strobl et al., 2008). In any situation, subsets of informative predictors are preferred to get the best predictions.

For RF and CRF, predicted probabilities are calculated in the same way as for Bagg: the proportion of votes over all the trees. While the predicted probabilities for RF and CRF are calculated in the same manner, the VIMP values may not be. Similar to Bagg, the VIMP for RF is calculated by averaging the total decrease in the Gini index for each predictor across all trees. Note that the decrease in Gini index is only calculated for a predictor when that predictor is used for a split.

However, for CRF, VIMP is calculated using the mean decrease in accuracy, either with or without being conditioned on related covariates. In the case of no conditioning, the out-of-bag (OOB) observations (those left out during construction of a tree in the forest) are used for each variable. First, the OOB values are passed through a tree to set the baseline accuracy for each predictor. Then, the OOB values are permuted for a single predictor and rerun through the same tree to give the permuted OOB accuracy for that predictor. The difference between the baseline accuracy and permuted OOB accuracy is taken per tree and averaged across all trees for each predictor. This gives the VIMP for the ensemble as the mean decrease in accuracy. Conditioning can be incorporated for CRF VIMP where the values of a particular predictor are only permuted within particular groups of observations based on a conditioning grid from the predictors which are correlated with that particular predictor preserving the correlation structure between the predictors (Strobl et al., 2009).

3.1.4 Stochastic Gradient Boosting

GBM is another tree-based method, but unlike RF, the individual trees are shallow and built sequentially, with each tree depending on the previous ones. As a result, the trees do not contribute equally to the final model (Friedman, 2001, 2002). The basic idea behind GBM is taking numerous weak classifiers, such as a tree with minimal depth, and combining them into a strong classifier.

The GBM algorithm for classification begins with initializing all predictions to the sample log odds. Then, over a set number of iterations, the residual (gradient) is calculated, a random sample is taken from the training data, a tree is created on that random sample using the residuals as the outcome, then the new residuals are computed, and finally, the model is updated by adding the new predicted value to the previous predicted value for each observation (Kuhn and Johnson, 2016, pgs. 390–392). The predicted probabilities are calculated similar to logistic regression. As mentioned, the initial estimate can be the sample log odds from the training set, $f_i^{(0)} = \log \frac{\hat{p}}{1-\hat{p}}$ where $\hat{p}$ is the probability from the sample (training set) of the event occurring. For binary classification, the Bernoulli distribution is used (Kuhn and Johnson, 2016, pgs. 390–392).

During the last step of updating the model, shrinkage or model regularization can constrain the learning process that helps avoid overfitting. This shrinkage parameter is a value between zero and one, and it determines the fraction of the current prediction will be added to the previous predicted value. This parameter must be selected when configuring the GBM (Kuhn and Johnson, 2016, pgs. 390–392). VIMP, referred to as the relative importance, is calculated as the average improvement across the ensemble. Such improvement is based on the splitting criteria for each predictor within each tree (Kuhn and Johnson, 2016, pgs. 390–392).

3.1.5 Bayesian Additive Regression Trees

The BART method is an ensemble of trees that incorporate a Bayesian model. There are three main components in BART including the regularization prior, the backfitting Markov chain Monte Carlo (MCMC) algorithm (this is a sampling method for any distribution), and the sum of trees. Its goal is to approximate a function for the expected outcome given the predictors through a sum of trees, while each tree models a distinct portion of the relationship and the summed model is additive with Bayesian methods used to control the fit (Chipman et al., 2010). The regularization prior prevents an individual tree from governing the entire fit, similar to the shrinkage parameter in GBM (Chipman et al., 2010; Kapelner and Bleich, 2016). With the backfitting MCMC algorithm, a sampler is engaged where draws are generated from the posterior distribution giving the predictions. Each tree is created in order to capture the fit that remains unexplained over numerous iterations of the MCMC algorithm.

In the case of classification, BART does not use a voting system to generate predicted probabilities. Instead, it uses the sum of trees model to estimate the conditional probit, which is then transformed into a conditional probability of a particular class (Chipman et al., 2010; Kapelner and Bleich, 2016). With BART, VIMP values here are measured by the inclusion proportions, assessed via the splitting rules in the trees across the post-burn-in MCMC iterations (Chipman et al., 2010; Kapelner and Bleich, 2016). The inclusion proportion represents how often a predictor is used as a splitting variable relative to all splits in the ensemble within each posterior Gibbs sample. Since BART generates many posterior samples, the VIMP is calculated as the posterior mean of the inclusion proportions across all of the posterior samples (Bleich et al., 2014).

3.1.6 Neural Networks and Model Averaged Neural Networks

NNet were introduced in the 1960s by Widrow and Hoff (1960) and Rosenblatt (1962). Here we use the ideas explained in Chapter 5 of Ripley (1996). With this method, the outcome is modeled by hidden units, which are linear combinations of the original predictors without constraints. A nonlinear transformation, usually a sigmoidal function, on the linear combination of hidden units is used to predict each class and transform the predictions between zero and one. However, these predictions need to be further softmax transformed since they do not necessarily add up to one. The final model usually includes several hidden units but may be subject to overfitting. In order to avoid this, a tuning parameter called weight decay is employed to regularize the model and maintain smoother boundaries (Kuhn and Johnson, 2016, pgs. 333–337).

Model averaging is another technique to avoid overfitting, in that the predictions from several NNet models are averaged to produce more stable predictions, resulting in the MANN introduced by Ripley (1995); Perrone and Cooper (1993). MANN is preferred as a single NNet may only find locally optimal estimates, rather than globally optimal. One downside to NNet is that highly correlated predictors can significantly reduce the prediction performance, so it is critical to apply feature selection or extraction beforehand if such predictors are present (Kuhn and Johnson, 2016, pgs. 333–337). For NNet, each class is predicted by the linear combination of the hidden units that have been transformed via a sigmoidal function and then via the softmax transformation to output the predicted probabilities per sample. Thus, the class with the highest predicted value can be used to classify an observation. Since MANN is average over many NNet, the predicted probabilities are averaged across the NNet and then these averaged probabilities can be used to classify each observation.

The VIMP per predictor are calculated using the Olden method (Olden et al., 2004). This method calculates the VIMP as the product of the raw input-hidden and hidden-output connection weights between each input and output neuron. Then, these products are summed across all hidden neurons maintaining the magnitude and sign of the relative contributions.

3.1.7 Super Learner

Due to the high cost of international birth cohort studies, such as the PROVIDE study, it is critical to optimize the analysis with available data to predict stunted growth. Typically, the primary prespecified analyses are relatively straightforward, focusing on key variables of interest and a limited number of potential confounders. However, the challenge lies in determining which ML techniques perform well. In this study, we demonstrate that SL is an effective statistical technique to generate predictions and insights into which methods would be most effective or whether a combination of methods is preferred.

The SL is an ensemble method designed to identify the optimal method or weighted combination of several algorithms for prediction, and it is guaranteed to perform at least as well as the best method in the ensemble (Polley and van der Laan, 2010). The goal of the SL is to estimate the expected outcome given the covariates through optimization, with performance measured by a loss function. The non-negative least squares (NNLS) based on the Lawson-Hanson algorithm is used for continuous or categorical responses, while the Nelder-Mead method which minimizes the rank loss may be used for only categorical responses. Here, we choose to focus on the rank loss option as advised by Phillips et al. (2023).

The SL requires specifying a library of algorithms, which should be diverse and tailored to the study dataset and its features. The SL then creates a linear combination of the algorithms in the library with selected weights or coefficients for the various library algorithms. In this study, our library of algorithms outlined in Table S3 includes GLM, Bagg, RF, CRF, GBM, BART, NNet, and MANN. In general, it is best to include a wide variety of methods in the library to achieve optimal performance for the SL (van der Laan et al., 2007; Phillips et al., 2023,?).

Since our data are mixed with both categorical and continuous predictors, we have chosen several different tree-based methods along with NNet and the classic approach of GLM. These tree-based and NNet methods are nonparametric in that there are no distributional assumptions about data. To mitigate overfitting, the models can be trained in order to choose the optimal values of the parameters through cross-validation (CV) using the `caret` package (Kuhn, 2008). The SL process is outlined below.

Following Naimi and Balzer (2018), the SL process for binary classification is outlined as:

1. Split the observed data $\mathbf{X} = \{X_1, \ldots, X_p\}$ and the outcome variable Y into V folds.
2. For each fold $v = 1, 2, \ldots, V$:
   (a) Define the observations in fold v as the validation set and the rest as the training set.
   (b) Let m be the number of algorithms in the library. Fit each of the m algorithms on the training set.
   (c) Use each fitted algorithm to predict the outcome for the validation set.
   (d) For each algorithm, estimate the risk using the rank loss. For each validation set and each algorithm, we estimate the AUC and the expected loss (risk) is $1 - \text{AUC}$.
3. For each algorithm, average the estimated risks across the folds to obtain one measure of performance.
   - We could select the algorithm with the smallest CV risk estimate called the discrete SL.
   - We continue by combining the CV predictions to create the ensemble SL. Denote the CV predictions per algorithm as $\hat{Y}_{1,\text{CV}}, \ldots, \hat{Y}_{m,\text{CV}}$.
4. The weights or coefficients, α_k, of each algorithm for the SL are calculated using the rank loss function to maximize the AUC. The coefficients, α_k, are estimated using the constrained regression

$$\Pr(Y = 1 | \hat{Y}_{1,\text{CV}}, \ldots, \hat{Y}_{m,\text{CV}}) = \alpha_1 \hat{Y}_{1,\text{CV}} + \cdots + \alpha_m \hat{Y}_{m,\text{CV}}$$

where $\alpha_k \geqslant 0$ and $\sum_{k=1}^{m} \alpha_k = 1$ such that $1 - \text{AUC}$ is minimized when comparing the SL predicted probabilities to the observed outcomes. Denote the estimated coefficients as $\hat{\alpha}_1, \ldots, \hat{\alpha}_m$ and Y as the binary outcome variable.

5. Create the SL. Refit the m algorithms on the observed data, $\mathbf{X}$, and calculate the predicted outcomes denoted as $\hat{Y}_1, \ldots, \hat{Y}_m$. Then, combine these predictions with the estimated coefficients as:
$$\Pr(Y = 1 | \hat{Y}_1, \ldots, \hat{Y}_m) = \hat{\alpha}_1 \hat{Y}_1 + \cdots + \hat{\alpha}_m \hat{Y}_m.$$
6. Evaluate the algorithms and/or SL.

For this analysis, we used $V = 20$ folds and had $m = 8$ algorithms with $p = 45$ predictors for a single SL model. The choice of $V = 20$ was chosen based on the recommendations of Phillips et al. (2023) given that the effective sample size in our situation is the number of observations $n = 391$, which is less than 500. Furthermore, because the outcome is binary, stratified sampling was used to preserve the class ratio within each training and validation set across the CV folds.

The SL can also be implemented under CV framework, resulting in multiple SL models. In this study, we used 10-fold CV, creating 10 separate SL models with the same settings as described above, one for each CV fold. The CV process for the SL enabled us to view the discrete SL for each fold, helping us determine whether one method prevails or if the ensemble SL should be selected as the final model.

3.2 Evaluation Methods

Since stunting is a binary outcome, these algorithms are actually classifying children to be stunted or not at two years of age with estimated probabilities. The performance of classification can be evaluated by the AUC, BS adjusted, and Min(R,Pr). The ROC curve plots the recall (true positive rate or sensitivity) on the y-axis and one minus the specificity (false positive rate) on the x-axis. Recall is the proportion of individuals with that outcome of interest correctly identified (R $= TP/(TP + FN)$) while specificity is the proportion of individuals without the outcome of interest correctly identified (Sp $= TN/(TN + FP)$) where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives, and FP is the number of false positives (Fawcett, 2006). Sp is additionally one minus the false positive rate.

The ROC curve shows the tradeoff between R and Sp, meaning that as R increases, Sp typically decreases. An AUC of 90%-100% represents an excellently performing method while a value in the 80%-90% range is still a well performing method (Fawcett, 2006). ROC curves and AUC evaluation methods typically do not take into account the incidence rate of the outcome of interest.

An alternative approach was introduced in the 2012 PhysioNet Challenge (Silva et al., 2012). This innovative metric, referred to as Min(R,Pr), is calculated using R and the precision (Pr $= TP/(TP + FP)$), also known as the positive predictive value. As the decision threshold varies, R and Pr are calculated at each threshold value. When R and Pr are plotted across the decision thresholds, the optimal decision cutoff is identified as the point where the R and Pr are closest to each other, as shown in Figure 1. This value of Min(R,Pr) at this threshold is used as the performance measure. This metric is considered a “reasonable tradeoff between accuracy of discrimination and prognostic value”, especially in settings with a low incidence rate (Silva et al., 2012). A higher value indicates better model performance.

The F-measure also known as F1 score or F-score, is the harmonic mean between Pr and R (Christen et al., 2023). When both Pr and R are high, the F-measure would also be high with an

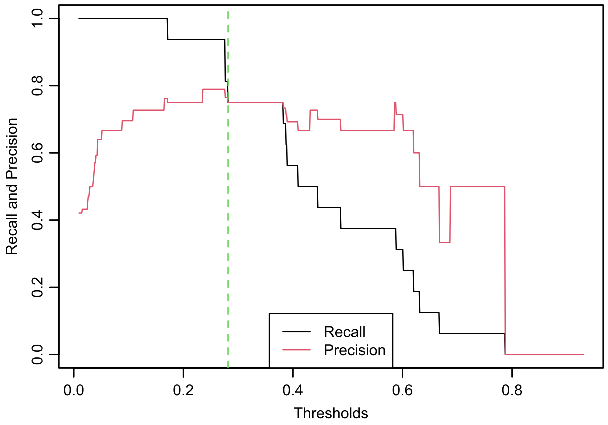

Figure 1: Recall and Precision values over a range of decision thresholds for the predictions using the SL model from the second fold of CV. This leads to a Min(R,Pr) value of 0.75 at the selected threshold indicated by the dotted line.

upper limit of one (all classifications correct) and a lower limit of zero (no classifications correct). While many use the F-measure as an evaluation metric, there are some critiques outlined by Christen et al. (2023). In particular, the F-measure is a pragmatic measure, in that while it provides a useful numerical summary, it does not represent any particular objective feature of the model. Additionally, the F-measure ignores true negatives, has asymmetric behavior for varying thresholds, and can result in the same value for different values of R and Pr (Christen et al., 2023). Hence, we have chosen to focus on the other evaluation measures mentioned in this section.

ROC curves and AUC are discrimination methods, meaning they assess a model's ability to distinguish between observations with the outcome of interest versus those without. Another measure of accuracy is calibration which measures prediction accuracy across the range of predictive values. The BS is a combination of both discrimination and calibration Kramer (2016). This score is calculated as the sum of the squared differences between the observed outcomes (Y_i) and the predicted outcomes ($\hat{Y}_i$) divided by the sample size (n),

$$\mathrm{BS} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}.$$

For example, if every patient outcome is treated as a coin flip, then the BS would be 0.25 since every prediction would have a probability of 0.5 and then every patient would have a score of 0.25. This BS is linked to the incidence rate (γ) of the outcome in the dataset. If every subject is naively given the probability of the outcome equal to the incidence rate, then the BS reduces to, what we will call the baseline BS

$$\mathrm{BS}_{baseline} = \gamma(1-\gamma)^2 + \gamma^2(1-\gamma) = \gamma(1-\gamma).$$

So, if the incidence rate is 0.04, then the baseline BS is 0.0384. If the actual or regular BS is 0.0192, then the model has reduced the variation or uncertainty in prediction by 50% ($1 - (0.0192/0.0384)$).

Thus, this score needs to be adjusted for the incidence rate of the outcome leading to the BS adjusted

$$\mathrm{BS}_{adjusted} = \frac{\gamma(1-\gamma) - \mathrm{BS}}{\gamma(1-\gamma)} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{baseline}}.$$

By using the BS adjusted of each model, we can compare the classification methods in a similar approach to AUC, where a higher value indicates better model performance (Kramer, 2016). Further, the upper limit for the BS adjusted is 1 indicating a relative perfect performance, the lower limit is negative infinity indicating a relative poor performance, and a value of zero indicates the performance is the same as using the incidence rate as the predicted probability. The BS adjusted is negative when the original BS is larger than the baseline BS, BS > $\text{BS}_{baseline}$. The data in this study has an incidence rate for stunted growth of 0.2967 and thus $\text{BS}_{baseline} = 2.776 \times 10^{-17}$. For our data then, if a model has poor performance in terms of the BS, the $\text{BS}_{adjusted}$ will be negative.

4 Results

The results for CV of the SL are explored first followed by the final model selected. In the CV results, the coefficients were examined across the CV folds per algorithm along with the evaluation measures. For the final prediction model, the evaluation measures were explored then the important variables were inspected. Partial dependence plots for the top important variables were used to study the direction of relationships between the important predictors and outcome.

4.1 Super Learner Models with Cross-Validation

We performed ten-fold CV, resulting in 10 different SL models built for the rank loss optimization method for the stunted response. Table 2 summarizes the coefficients for each SL model built for each CV fold, highlighting how the contribution of individual algorithms varied across folds and offering insight into the consistency and stability of the ensemble weights. The NNLS optimization method was also implemented for the continuous response of HAZ at two years of age. These results are summarized in Section 3 of the Supplemental Material and discussed in Section 5.

Table 2: Summary of weights/coefficients for the 10 CV folds for the algorithms in the SL library with rank loss optimization of the stunted at two years outcome.

CV Fold	GLM	Bagg	RF	CRF	GBM	BART	NNet	MANN
1	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
2	0.263	0.130	0.114	0.142	0.139	0.071	0.045	0.096
3	0.463	0.000	0.052	0.006	0.296	0.000	0.102	0.081
4	0.157	0.000	0.213	0.189	0.190	0.098	0.144	0.008
5	0.562	0.064	0.093	0.068	0.043	0.160	0.003	0.007
6	0.239	0.073	0.146	0.150	0.083	0.230	0.000	0.079
7	0.131	0.094	0.137	0.143	0.156	0.167	0.064	0.109
8	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
9	0.219	0.000	0.021	0.124	0.113	0.188	0.093	0.243
10	0.284	0.103	0.104	0.104	0.081	0.193	0.000	0.131
Mean	0.257	0.071	0.113	0.118	0.135	0.136	0.070	0.100
SD	0.148	0.054	0.052	0.050	0.070	0.067	0.056	0.067
Number Nonzero	10	7	10	10	10	9	8	10

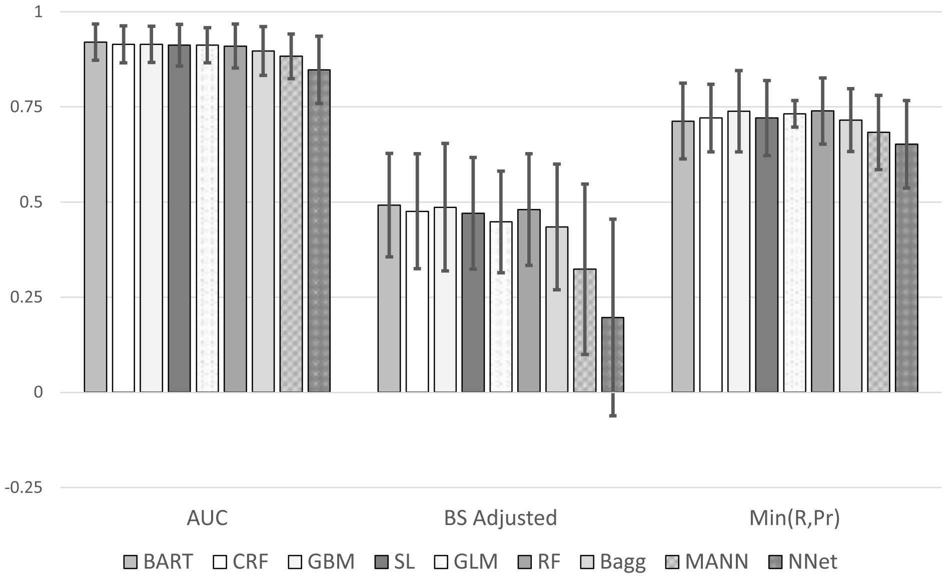

Figure 2: Average AUC, BS adjusted, and Min(R,Pr) values with plus and minus one standard deviation per algorithm over the 10 CV folds for the rank loss optimization method with the stunted at two years outcome ordered by decreasing AUC.

Among the 10 folds of CV, GLM, RF, CRF, GBM, and MANN were utilized all 10 times, BART was used 9 times, NNet 8 times, and Bagg only 7 times. Even though these methods were included in the SL models for each CV fold, some contributed minimally due to low weighting or small coefficient. For example, Bagg was used in 7 of the 10 CV folds, however it had an average weight of only 0.071 and thus had little contribution to the SL models and predictions on average. In contrast, GLM had an average coefficient of 0.257 (ranging from 0.125 to 0.562) with all CV folds. Clearly, GLM is used more (with highest average coefficient and highest coefficient in 5 of the 10 folds) in the SL models than other methods. Thus, GLM, BART, GBM, CRF, and RF are on average more preferred in the SL models (ordered by average coefficient), closely followed by MANN. Bagg and NNet almost tied for lowest average coefficient. For a future analysis, it may be appropriate to drop those methods with low or zero coefficients such as Bagg and NNet.

With 10 CV folds of the SL, we calculated the averages and standard deviations of the AUC, Min(R,Pr), and BS adjusted for all algorithms and the SL, using the observations that were not in the CV fold to build the models. Figure 2 depicts the average evaluation values with their standard deviations ordered by AUC. BART had the highest mean AUC at 92.1%, closely followed by CRF, GBM, SL, GLM, and RF all with an average AUC around 91%. Bagg, MANN, and NNet still fair well with 89.7%, 88.3%, and 84.8% respectively. BART also had the highest mean BS adjusted at 0.492 followed by GBM, RF, CRF, SL, GLM, Bagg, MANN, and NNet in that order. RF had the highest mean Min(R,Pr) of 0.739 followed by GBM, GLM, CRF, SL, Bagg, BART, MANN, and NNet. Thus, in terms of the evaluation values, BART was the best performing algorithm on average, while NNet was consistently the worst.

Additionally, the CV risk was explored to determine whether a single method consistently prevailed or outperformed the others across the folds. BART again emerged as the best performer, achieving the lowest CV risk in 8 of the 10 folds, while GLM had the lowest CV risk in the other 2 folds.

4.2 Final Prediction Model

While the SL itself demonstrates good evaluation measure values, if we were to choose one method, BART would be selected as the final prediction model based on the evaluations and discussion from the 4.1 section. The full dataset was randomly split into 80% training and 20% testing data. The BART model was created on the training data and then evaluated using the testing data. The evaluation metrics are summarized in Table 3. Based on the BS adjusted, BART reduced the predictive uncertainty by 58.3%. It also achieved a high AUC of 94.9% and a strong Min(R,Pr) value of 0.792. These consistently high values across AUC, BS adjusted, and Min(R,Pr) indicate that the BART model exhibits excellent calibration or discrimination performance.

The VIMP for BART (Figure 3) identified HAZ at one year as the most important predictor of stunting at two years, as might be expected. For this final model, due to the large number of variables, we have chosen to focus on the top 15 predictors, those with a mean inclusion proportion, VIMP, of around 0.02 or higher. Following HAZ at one year, the next most important predictors were income, expenditure, anti-lipopolysaccharide (LPS) antibody at week 6, plasma retinol binding protein (RBP) at week 6, plasma soluble cluster designation 14 at week 18 (sCD14), fecal Reg 1B at week 12, vitamin D at week 18, mother's weight at enrollment, number of days of diarrhea through the first year of life, fecal MPO at week 12, mother's height at

Table 3: Final BART model evaluation measures using the test data.

Algorithm	AUC	BS Adjusted	Min(R,Pr)
BART	0.949	0.583	0.792

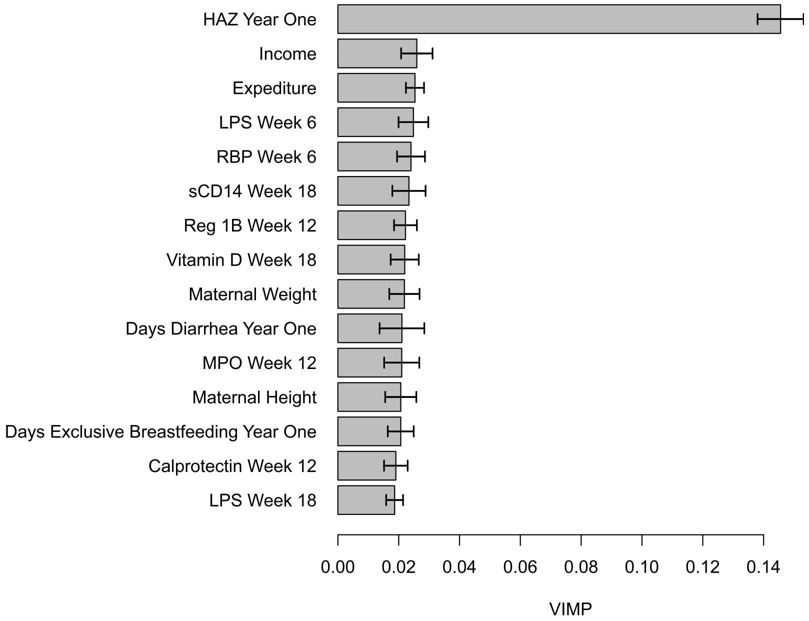

Figure 3: VIMP plot of the top 15 important variables with mean inclusion proportion plus and minus one standard deviation from the final BART model.

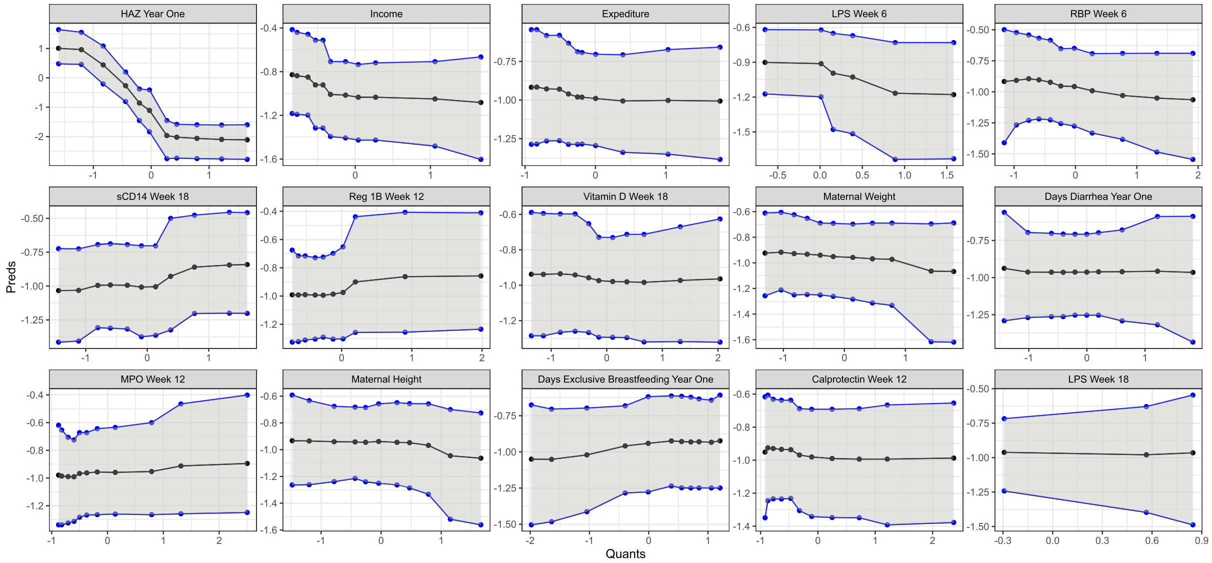

Figure 4: Partial dependence plots of the top 15 important variables from the final BART model.

enrollment, the number of days of exclusive breastfeeding through the first year of life, fecal calprotectin at week 12, and LPS at week 18, in that order. While the exact ranking may vary with different BART models or even different runs of the BART model with the same parameters, these variables consistently appear among the top 15 predictors.

We can also examine the partial dependence plots for these top 15 variables to explore their relationship with the response in terms of probits. As shown in Fig. 4, HAZ at one year, income, expenditure, anti-LPS antibody, RBP, vitamin D, mother's weight, mother's height, and calprotectin all have positive effects (downward trending plots) on the stunting response, suggesting that higher values in these predictors are associated with a decreased probability of stunting. In contrast, sCD14, Reg 1B, MPO, and number of days exclusive breastfeeding have a negative effect (upward trending plots) on stunting as their increases are associated with an increasing risk of stunting. The number of days of diarrhea does not clearly have a positive or negative relationship with stunting based on the partial dependence plot in Figure 4.

5 Discussion

In this study, we applied several ML algorithms, along with an ensemble of these algorithms, to predict the stunting outcome at two years of age in a Bangladesh birth cohort. Our goal was to evaluate which ML algorithm would be the best predictive method for this data. After identifying BART as the most effective ML algorithm, we further assessed the relative importance of individual predictors.

Implementing the SL with CV is advantageous for identifying whether a single algorithm or a combination of algorithms yields the best predictive performance, though it comes at the cost of increased computational time. For particularly complex datasets, an ensemble like SL may offer more accurate and robust prediction than any individual algorithm. By incorporating a diverse set of flexible and nonparametric methods, we were able to identify which techniques performed best for prediction in this data, ultimately leading to our conclusion to select BART.

While other methods could be included in the SL to diversify its library, this study covers a wide range of methods, from simple to complex, each with distinct strengths. Although the SL results are fairly specific to the data and situation, our approach itself is broadly applicable. The SL framework can be implemented across various datasets and the results can be interpreted in a consistent and generalizable manner.

Based on the model performance and variable importance, our findings show that HAZ at one year is the most important predictor of stunted growth at two years, which aligns well with biological expectations. Overall, stunted growth appears to be positively affected (meaning a lower likelihood of stunted growth) by HAZ at one year, income, expenditure, anti-LPS antibody, RBP, vitamin D, mother's weight and height, and calprotectin. Calprotectin is a gastrointestinal tract inflammatory marker that was previously found to be positively associated with growth in the PROVIDE study by Donowitz et al. (2018). The positive effect of calprotectin may be partially explained by elevated levels in breastfed children, suggesting it could serve as potential surrogates for improved nutrition (Dorosko et al., 2008; Davidson and Lönnerdal, 1990). RBP is a biomarker of vitamin A status, which is essential for growth (Goodman, 1980). The number of days exclusively breastfeeding in the first year of life was recently found to be protective against stunting by Campos et al. (2020) and Hadi et al. (2021). An increase in anti-LPS antibody is typically associated with growth faltering (McDonald et al., 2016; Syed et al., 2018).

Stunted growth may be negatively impacted (meaning higher likelihood of stunted growth) by increased levels in Reg 1B and sCD14. Reg 1B has been previously associated with stunting in Bangladesh and Peru (Peterson et al., 2013), while an elevated sCD14 has been linked to increased risk of stunted growth in Peruvian infants (Zambruni et al., 2019). The number of days of diarrhea seems to have a minor effect, not clearly showing a negative or positive effect by the partial dependence plot in Figure 4. Notably, some variables, including several cytokines like IL-4 appeared to have minimal importance (VIMP less than 0.01). In further explorations, these less informative variables may be excluded from analysis to focus on the most influential variables identified above.

This study focused on the categorical response of stunted growth, nevertheless, the continuous response of HAZ was additionally explored. For the continuous response, GLM had the highest weights in the SL models (mean coefficient of 0.577) and was used in all 10 CV folds, followed by RF (0.112), CRF (0.107), and MANN (0.106), while the other algorithms have a mean coefficient less than 0.1. In terms of root mean square error (RMSE), GLM achieved the lowest value followed by CRF, Bagg, RF, BART, GBM, with NNet and MANN having the highest RMSE. Notably, NNet was not used across any SL CV fold for the continuous response while MANN was only used twice, RF and GBM three times, and BART four times. Given that GLM was used with the highest weight on average, while other methods had a relatively lower weight or much less often (Bagg and CRF were used seven times out of the 10 CV folds which was the next highest), GLM was selected as the final model for the continuous response. The final model and relevant details can be found in Section 3 of the Supplemental Material.

Across both categorical and continuous responses, GLM was simplistic yet most useful and one of the best performers in the SL library. While GLM was selected as the final model for the continuous response, BART performed better for the categorical response and thus was selected as the final model to predict stunted growth. The SL method also performed well since it was a combination of all the methods in the library. However, it was not selected to be the final model for either response type, as a single algorithm prevailed/outperformed in each case.

One limitation from this analysis includes the inability to incorporate more algorithms for various reasons. For example, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were considered as possible candidate methods. However, both rely on assumptions of multivariate normality, which are not well-suited for datasets with the categorical predictors. Further, many continuous predictors in our study, such as cytokines, are highly skewed. While transformations could be applied, concerns remain about whether transformed predictors would truly follow a multivariate normal distribution, which is especially critical for QDA. Additionally, LDA assumes equal covariance matrices across groups, which is not satisfied in our data. Thus, these methods were not considered in our library for the SL and subsequently excluded from the analysis. Many other methods exist and were briefly considered. However, we focused on the selected methods due to their nonparametric nature (with some exceptions) to ensure a diverse yet practical spread of algorithms.

In addition, time and computer memory constraints were also considered when selecting the set of algorithms and their tuning parameters. A common tuning parameter for tree-based methods is the number of trees which was not included in this analysis. Including it may have improved the predictability of these methods but would have further increased computational demands considerably. Another challenge was missing data in the PROVIDE study. For illustration purpose, subjects with missing values in predictors or response were excluded from the analysis. This reduced the sample size from the original 700 children to 391, due to dropouts or incomplete measurements over the two year period. Further work could explore strategies for handing missing data and incorporating additional predictors to better utilize the available information for growth prediction. Perhaps the biggest limitation from using the CV SL is the computational cost with respect to time and storage size. As the size of data and number of algorithms in the library increase, the required computational resources can growth substantially, leading to much longer processing time. This may pose challenges for researchers without access to high-performance computing resources. Nevertheless, in most cases, building a single SL model should suffice can still provide strong predictive performance with reduced computational burden, following the guidance proposed by Phillips et al. (2023).

Supplementary Material

The supplementary material consists of the Supplementary Tables and Analyses PDF, a README file, R scripts to run all analyses, RData file with the data appropriately formatted for analyses, and RData files with the corresponding models. The Supplementary Tables and Analyses PDF file that contains data descriptions, summary of methods, and results for the NNLS optimization for the continuous outcome. The README file briefly explains each file.

References

Bhutta ZA, Ahmed T, Black RE, Cousens S, Dewey K, Giugliani E, et al. (2008). What works? Interventions for maternal and child undernutrition and survival. *Lancet*, 371(9610): 417–440.

Bleich J, Kapelner A, George EI, Jensen ST (2014). Variable selection for BART: An application to gene regulation. *Annals of Applied Statistics*, 8(3): 1750–1781. https://doi.org/10.1214/14-AOAS755

Boulesteix AL, Janitza S, Kruppa J, Konig IR (2012). Overview of random forest methodology and practical guidance with emphasis on comutaional biology and bioinformatics. *WIREs Data Mining and Knowledge Discovery*, 2(6): 493–507. https://doi.org/10.1002/widm.1072

Breiman L (1996). Bagging predictors. *Machine Learning*, 24: 123–140. https://doi.org/10.1023/A:1018054314350

Breiman L (2001). Random forests. *Machine Learning*, 45: 5–32. https://doi.org/10.1023/A:1010933404324

Brooks J (2012). Super Learner and Targeted Maximum Likelihood Estimation for Longitudinal Data Structures with Applications to Atrial Fibrillation, Dissertation/Thesis.

Butzin-Dozier Z, Ji Y, Coyle J, Malenica I, McQuade ETR, Grembi JA, et al. (2025). Treatment heterogeneity of water, sanitation, hygiene, and nutrition interventions on child growth by environmental enteric dysfunction and pathogen status for young children in Bangladesh. In: PLOS Neglected Tropical Diseases.

Campos AP, Vilar-Compte M, Hawkins SS (2020). Association between breastfeeding and child stunting in Mexico. *Annals of Global Health*, 86(1): 1–14. https://doi.org/10.5334/aogh.2836

Chipman HA, George EI, McCulloch RE (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1): 266–298. https://doi.org/10.1214/09-AOAS285

Christen P, Hand DJ, Kirielle N (2023). A review of the F-measure: Its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3): 1–24.

Davidson LA, Lönnerdal B (1990). Fecal alpha 1-antitrypsin in breast-fed infants is derived from human milk and is not indicative of enteric protein loss. *Acta Paediatrica Scandinavica*, 79(2): 137–141. https://doi.org/10.1111/j.1651-2227.1990.tb11429.x

Dewey KG, Begum K (2011). Long-term consequences of stunting in early life. *Maternal and Child Nutrition*, 7(Suppl 3): 5–18. https://doi.org/10.1111/j.1740-8709.2011.00349.x

Donowitz JR, Cook H, Alam M, Tofail F, Kabir M, Colgate ER, et al. (2018). Role of maternal health and infant inflammation in nutritional and neurodevelopmental outcomes of two-year-old Bangladeshi children. *PLOS Neglected Tropical Diseases*, 12(5): 1–20. https://doi.org/10.1371/journal.pntd.0006363

Dorosko SM, MacKenzie T, Connor RI (2008). Fecal calprotectin concentrations are higher in exclusively breastfed infants compared to those who are mixed-fed. *Breastfeeding Medicine*, 3(2): 117–119. PMID: 18564000. https://doi.org/10.1089/bfm.2007.0036

Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Friedman JH (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5): 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman JH (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4): 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Goodman DS (1980). Plasma retinol-binding protein. *Annals of the New York Academy of Sciences*, 348: 378–390. https://doi.org/10.1111/j.1749-6632.1980.tb21314.x

Greg R, Developers G (2024). gbm: Generalized Boosted Regression Models. R package version 2.1.9.

Hadi H, Fatimatasari F, Irwanti W, Kusuma C, Alfiana RD, Asshiddiqi MIN, et al. (2021). Exclusive breastfeeding protects young children from stunting in a low-income population: A study from eastern Indonesia. *Nutrients*, 13(12): 1–14.

Hoddinott J, Maluccio JA, Behrman JR, Flores R, Martorell R (2008). Effect of a nutrition intervention during early childhood on economic productivity in guatemalan adults. *Lancet*, 371(9610): 411–416. https://doi.org/10.1016/S0140-6736(08)60205-6

Houssaini A, Assoumou L, Marcelin AG, Molina JM, Calvez V, Flandre P (2012). Investigation of super learner methodology on HIV-1 small sample: Application of jaguar trial data. *AIDS*

Research and Treatment, 2012(1): 1–7.

Ju C, Combs M, Lendle SD, Franklin JM, Wyss R, Schneeweiss S, et al. (2016). Propensity Score Prediction for Electronic Healthcare Databases using Super Learner and High-Dimensional Propensity Score Methods, *Technical report, The Berkeley Electronic Press. Working Paper 351.*

Kapelner A, Bleich J (2016). BartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4): 1–40. https://doi.org/10.18637/jss.v070.i04

Kirkpatrick BD, Colgate ER, Mychaleckyj JC, Haque R, Dickson DM, Carmolli MP, et al. (2015). The "performance of rotavirus and oral polio vaccines in developing countries" (PROVIDE) study: Description of methods of an interventional study designed to explore complex biologic problems. *The American Journal of Tropical Medicine and Hygiene*, 92(4): 744–751. https://doi.org/10.4269/ajtmh.14-0518

Kramer AA (2016). Which statistic can be either the worst or best metric for assessing a predictive model? Prescient News.

Kuhn M (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5): 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn M, Johnson K (2016). *Applied Predictive Modeling.* Springer Science+Business Media LLC.

Ladds MA, Thompson AP, Kadar JP, Hocking DP, Harcourt RG (2017). Super machine learning: Improving accuracy and reducing variance of behaviour classification from accelerometry. *Animal Biotelemetry*, 5(8): 1–9.

Liaw A, Wiener M (2002). Classification and regression by randomForest. *R News*, 2(3): 18–22.

Martorell R, Zongrone A (2012). Intergenerational influences on child growth and undernutrition. *Paediatric and Perinatal Epidemiology*, 26: 302–314. https://doi.org/10.1111/j.1365-3016.2012.01298.x

McDonald CM, Manji KP, Gosselin K, Tran H, Liu E, Kisenge R, et al. (2016). Elevations in serum anti-flagellin and anti-LPS igs are related to growth faltering in young Tanzanian children. *The American Journal of Clinical Nutrition*, 103(6): 1548–1554. https://doi.org/10.3945/ajcn.116.131409

Mertens A, Benjamin-Chung J, Colford JM Jr, Coyle J, van der Laan MJ, Hubbard AE, et al. (2023). Causes and consequences of child growth faltering in low-resource settings. *Nature*, 621: 568–576. https://doi.org/10.1038/s41586-023-06501-x

Mursil M, Rashwan HA, Santos-Calderon L, Murphy M, Valls DSP (2023). Maternal nutritional factors enhance birthweight prediction: A super learner ensemble approach. *SSRN.*

Naimi AI, Balzer LB (2018). Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, 33: 459–464. https://doi.org/10.1007/s10654-018-0390-z

Naylor C, Lu M, Haque R, Mondal D, Buonomo E, Nayak U, et al. (2015). Environmental enteropathy, oral vaccine failure and growth faltering in infants in Bangladesh. *eBioMedicine*, 2(11): 1759–1766. https://doi.org/10.1016/j.ebiom.2015.09.036

Olden JD, Joy MK, Death RG (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3): 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013

Perrone M, Cooper L (1993). When networks disagree: Ensemble methods for hybrid neural networks. *Neural Networks for Speech and Image Processing.*

Peters A, Hothorn T (2023). ipred: Improved predictors. R package version 0.9-14.

Peterson KM, Buss J, Easley R, Yang Z, Korpe PS, Niu F, et al. (2013). REG1B as a predictor of childhood stunting in Bangladesh and Peru. *The American Journal of Clinical Nutrition*, 97(5): 1129–1133. https://doi.org/10.3945/ajcn.112.048306

Phillips RV, van der Laan MJ, Lee H, Gruber S (2023). Practical considerations for specifying a super learner. *International Journal of Epidemiology*, 52(4): 1276–1285. https://doi.org/10.1093/ije/dyad023

Pirracchio R, Carone M (2016). The balance super learner: A robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical Methods in Medical Research*, 27(8): 2504–2518. https://doi.org/10.1177/0962280216682055

Polley E, LeDell E, Kennedy C, van der Laan M (2024). SuperLearner: Super learner prediction. R package version 2.0-29.

Polley EC, van der Laan MJ (2010). Super Learner in Prediction, *Technical report, The Berkeley Electronic Press. Working Paper 266.*

Prendergast AJ, Humphrey JH (2014). The stunting syndrome in developing countries. *Paediatrics and International Child Health*, 34(4): 250–265. https://doi.org/10.1179/2046905514Y.0000000158

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripley BD (1995). Statistical ideas for selecting network architectures. In: *Neural Networks: Artificial Intelligence and Industrial Applications* (B Kappen, S Gielen, eds.), 183–190. Springer, London, London.

Ripley BD (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rosenblatt F (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.

Silva I, Moody G, Scott DJ, Celi LA, Mark RG (2012). Predicting in-hospital mortality of ICU patients: The PhysioNet computing in cardiology challenge 2012. *Computing in Cardiology*, 39: 245–248.

Sinisi SE, Petersen ML, van der Laan MJ (2006). Super Learning: An Application to Prediction of HIV-1 Drug Susceptibility, *Technical report, The Berkeley Electronic Press. Working Paper 206.*

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307): 1–11.

Strobl C, Hothorn T, Zeileis A (2009). Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package, *Technical report*, Department of Statistics. Number 50.

Syed NT, Sand Iqbal, Sadiq K, Ma JZ, Akhund T, Xin W, Moore SR, et al. (2018). Serum Anti-flagellin and Anti-lipopolysaccharide Immunoglobulins as Predictors of Linear Growth Faltering in Pakistani Infants at Risk for Environmental Enteric Dysfunction. *PLOS One*, 13(3): 1–13.

van der Laan MJ, Polley EC, Hubbard AE (2007). Super Learner, *Technical report, The Berkeley Electronic Press. Working Paper 222.*

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Springer, New York.

Victora CG, Adair LS, Fall CHD, Hallal PC, Martorell R, Richter L, et al. (2008). Maternal and child undernutrition: Consequences for adult health and human capital. *The Lancet*, 371(9609): 340–357. https://doi.org/10.1016/S0140-6736(07)61692-4

Widrow B, Hoff ME (1960). Adaptive switching circuits. *1960 IRE WESCON Convention Re-*

ocrd, 4: 96–104.

Zambruni M, Ochoa TJ, Somasunderam A, Cabada MM, Morales ML, Mitreva M, et al. (2019). Stunting is preceded by intestinal mucosal damage and microbiome changes and is associated with systemic inflammation in a cohort of Peruvian infants. *The American Journal of Tropical Medicine and Hygiene*, 101(5): 1009–1017. https://doi.org/10.4269/ajtmh.18-0975