JOURNAL OF
EMERGING INVESTIGATORS

Article

Risk assessment modeling for childhood stunting using automated machine learning and demographic analysis

Aditya Sirohi[1], Jason H. Moore[2,3]

[1]Conestoga High School, Berwyn, Pennsylvania

[2]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

[3]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, California

SUMMARY

Over the last few decades, childhood stunting has persisted as a major global challenge. More than 154 million children under five are stunted across the world, with 95% of them residing in Asia and Africa alone. Apart from the small stature, stunting poses additional impediments including delayed cognitive development and health complications. Furthermore, the involvement of several demographic factors and limited timeframe of intervention effectiveness poses challenges to existing treatments. Thus, understanding the impact of demographic conditions on stunting and developing predictive methods is vital for advancing future research developing predictive models. In this study, we hypothesized that TPTO (Tree-based Pipeline Optimization Tool), an AutoML (automated machine learning) tool, would outperform all pre-existing machine learning models and reveal the positive impact of economic prosperity, strong familial traits, and resource attainability on reducing stunting risk. The TPOT pipelines, for data sets from Kenya and Bangladesh, outscored previously reported ML models with most performance metrics improving by 0.1 or more. Feature correlation plots revealed that maternal height, wealth indicators, and parental education were universally important features for determining stunting outcomes approximately two years after birth. Additionally, the machine learning models trained by TPOT resulted in strong overall performances with the best metrics scoring over 0.8. These results help inform future research by highlighting how demographic, familial, and socio-economic conditions influence stunting and providing medical professionals with a deployable risk assessment tool for predicting childhood stunting.

INTRODUCTION

Childhood stunting is a global issue impacting over 154.8 million children. A child is considered stunted when their height or length is two standard deviations or more below the World Health Organization Child Growth Standard median (3). The critical timeframe where a child is susceptible to stunting starts during pregnancy and continues until around 24 months of age. This is also the timeframe where interventions are most effective (4). In addition to a short stature, children with stunting also face long-term deterioration of the quality of life (5). Stunted women shorter than 145cm impose general health difficulties on their children and lower birth survival rates. Additionally, stunting is frequently associated with delayed cognitive development and even impairment, leading to lower academic performance and wages (3-5). Expanding the effects of stunting across generations, children with stunted parents had lower-performing developmental quotients and cognitive scores as well as shorter height when compared to those without stunted parents (2, 6). Thus, the major challenge now is understanding, predicting, and treating stunting at a global scale.

Of the 154.8 million stunted children, the majority are present in Asia (87 million) and Africa (59 million), where stunting rates rise above 30% of the population (1). In addition, more than 23% of the population in many major cities live in impoverished, densely populated urban areas without proper access to food, decent housing, sanitation, clean water, and essential infrastructure services (2). Although the frequency and prevalence of stunting is decreasing on a global scale, many low-income regions in West Africa and Southern Asia face stunting rates 50% higher than those of other countries (1, 3). Therefore, understanding childhood stunting — the causes, preventive timeframe, and measurement challenges — is vital for preventive action.

Machine learning is a powerful tool to leverage in clinical/medical applications. Recently, ML (machine learning) classifiers have gained traction within risk prediction due to the new abundance of publicly available records and studies (7). While traditional statistical analysis techniques exponentially scale in complexity as the data dimensionality increases, ML can be used to process and analyze this complicated

information quickly and efficiently. Furthermore, clinical trial data compiled over the past decades provide opportunities to better understand the relationship between environmental factors and stunting, along with advancements in prognostic modeling. Recent studies developed predictive ML models based on childhood stunting in West Africa and South Asia. A study in Bangladesh utilized classic linear regression models and yielded varied success with an accuracy below 70% (8). Another study using stunting data from Ethiopia developed five ML models with accuracy scores between 63.7% and 67.7% (9). These past studies have produced models with low scoring accuracy metrics because the complexity of medical/clinical data caused a significant predictive challenge for these models. Furthermore, there are countless combinations of preprocessing and parameter tuning that previous studies have not considered.

AutoML (automated machine learning) is a new solution for generating more accurate models in healthcare domains. AutoML methodically proceeds through the most challenging steps in creating ML models, which include feature engineering/processing, model training, and hyperparameter optimization, to produce an effective model for any given dataset (10). Tree-based Pipeline Optimization Tool (TPOT) is a powerful AutoML tool built upon existing ML frameworks that generates optimal pipelines for given data using genetic programming. Conventionally, a data scientist dedicates considerable time processing features, selecting models, and optimizing hyperparameters. Proceeding through these three phases of the pipeline in a timely manner requires substantial background and experience with building ML models. TPOT streamlines feature processing, model selection, and hyperparameter optimization by the means of genetic programming (GP), a computational technique used for automatically constructing programs. Feature processing includes encoding large numeric data points into a smaller range and reducing data dimensionality. Model building automatically trains different ML models on the provided dataset. Hyperparameter optimization searches for each algorithms' ideal run parameters to maximize the score on a specified metric. The GP process enables TPOT to explore thousands of pipelines through multiple generations (iterations). The best performing pipelines from each generation are used to build the next generation until the run completes. TPOT has shown promise in previous benchmarking biomedical analyses and has tremendous possibility when applied to childhood stunting (11, 12).

This study was conducted to: [1] assess the accuracy of TPOT's models, [2] analyze TPOT's applicability in childhood stunting, and [3] understand the subtle correlations between socio-economic status, familial conditions, growth/sanitation interventions, and stunting. We hypothesized that predictive models generated by the TPOT AutoML method would perform better than those generated by conventional methods due to TPOT's ability to consider many different models and parameter settings. Additionally, we hypothesized that economic prosperity, access to resources, and positive familial demographics would negatively correlate to the risk of stunting. When compared to ML models from previous studies that used traditional techniques, we observed a 5%-10% increase in the compared scoring metrics which proved TPOT's accuracy and applicability in stunting problem domains. Additionally, feature correlation scores of the ML models demonstrated how Maternal Education, Maternal height, and assets score (an indicator of wealth) were universally deterministic of stunting risk while the interventions generally had no impact on stunting risk.

RESULTS

AutoML is a powerful tool for developing predictive models and uncovering complex associations present in data. We applied AutoML to understand and predict childhood stunting using socio-economic, familial, and environmental conditions. Once the data was preprocessed and the models built, we extracted feature importance, feature correlation, and model performance to explain our hypothesis. Each run of the AutoML produced a best-performing ML model, which we denote by the letter P, and the random state of that pipeline yielding that model (e.g., P12). Our data originated from two regions, Kenya and Bangladesh.

Preliminary data analysis

Prior to conducting TPOT runs, we analyzed the correlations between the individual features to understand how the demographic conditions of a target child related to one another. The Kenya clinical trial features did not correlate as only one pair of features had an absolute Pearson correlation value above 0.2. Conversely, the Bangladesh clinical trial recorded over nine absolute correlation values above 0.25 (**Figure 1**). For Kenya and Bangladesh, the number of children in the household and the maternal age exhibited positive correlations of 0.74 and 0.52, respectively (**Figure 1**). Interestingly, the Assets Score (a metric we developed to indicate socioeconomic wealth) for the Kenya data had an insignificant relationship to maternal education level (**Figure 1a**). However, parental education in Bangladesh was moderately related to the Assets Score with a correlation value of 0.39 for Maternal Education and 0.43 for Paternal Education (**Figure 1b**). Generally, higher parental education in these two studies produced a higher wealth indicator for the target child. In addition, parental education levels in Bangladesh produced a generally lower hunger/food insecurity score when compared to Kenya. The Bangladesh heatmap revealed that mothers with a higher level of education were less likely to have more children as Maternal Education and Birth Order had a correlation score of -0.4 (**Figure 1b**). Similarly, Paternal Education was negatively related to the Birth Order with a score of -0.27. Maternal and paternal education levels in Bangladesh were significantly and positively correlated with a score of 0.62, signifying a tendency to marry people with similar education

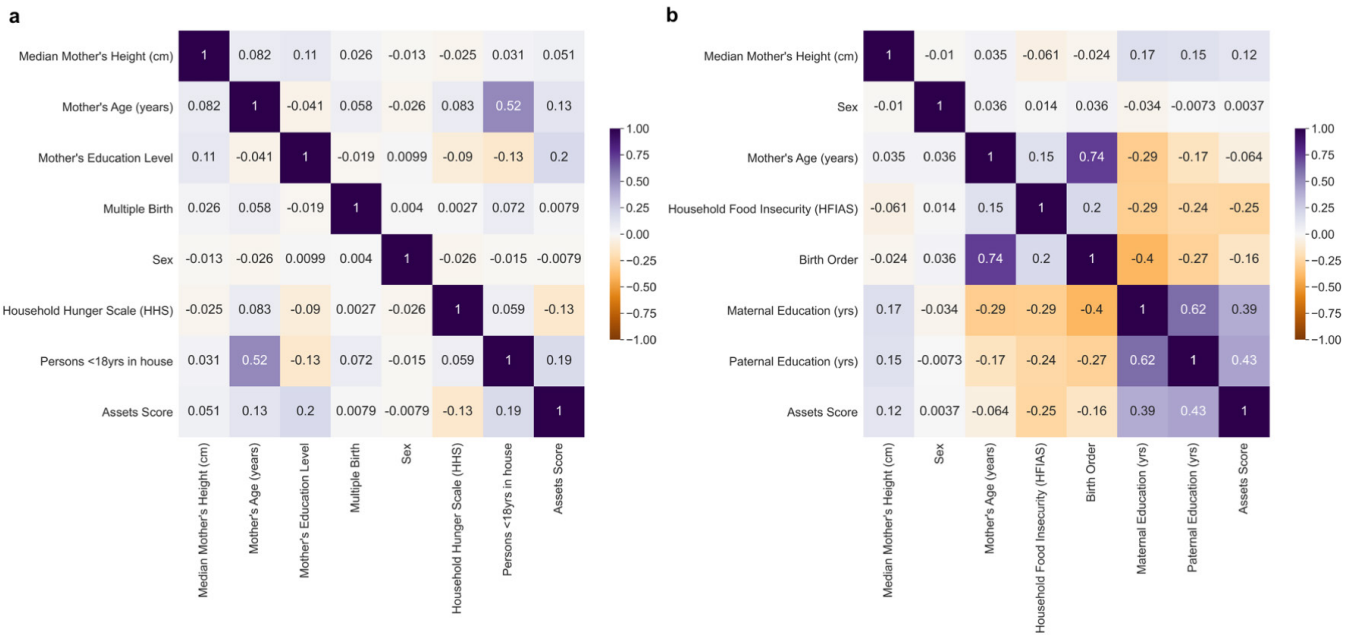

Figure 1. Feature correlation heatmaps for (a) Kenya and (b) Bangladesh. Heatmap showing the correlation value between two features generated using Pearson correlation coefficient. The lowest value of -1 denotes a negative linear relationship, 0 denotes no relationship, and the highest value of 1 denotes a positive linear relationship.

levels (**Figure 1b**).

Feature importance and feature correlation

We next used Shapely Additive Explanations (SHAP) to determine the impact of individual features on the ML model's output and extract feature importance. The SHAP summary plot presents information in four distinct categories. The first is feature importance. Variables are placed in descending importance starting from the top of the y-axis. The impact of a feature's value on the model's output is denoted by the location of each respective dot (a target child in the training data set) on the x-axis. The color of each dot denotes whether the value of the feature was high or low. The initial distinction between the two plots was observed in the domain of SHAP values on the x-axis. Features in the Kenya models recorded many absolute SHAP values below 1.5 (**Figure 2a**). Conversely, most stunting features from Bangladesh produced absolute SHAP values above 1.5 (**Figure 2b**). According to this result, the features from Bangladesh generated a stronger push on the TPOT models' predictions when compared to the features from Kenya. Additionally, a larger domain in the positive x-axis for both ML models' summary plots revealed that the features increased stunting risk with a higher magnitude than they reduced the risk (**Figure 2**).

Feature importance rankings – determined by subtracting the performance of the models in the absence of the feature from the performance of the model in the presence of it – stressed similarities between the top four most influential features for TPOT models in both countries. Median mother's height universally ranked as the most important feature for determining stunting. Similarly, the Assets Score ranked as the second most important feature for determining the ML output (**Figure 2**). All parental education features also ranked in the top four in importance. According to these rankings, maternal height, wealth indicators, and parental education levels were strongly indicative of stunting risk across all regions. Interestingly, gender/sex ranked third for the Kenya data but ninth for Bangladesh's data. This indicates that gender plays a prominent role in stunting in Kenya but not in Bangladesh. The intervention arms were ranked at the bottom of feature importance and were not very indicative of the prediction as they recorded low SHAP values (**Figure 2**).

By considering the distribution of individual values of each feature (denoted by a circle and plotted horizontally along the graph) on the SHAP plot, we observed correlations between features and stunting. For both models, the clustering of higher (red) maternal height readings with large, negative, SHAP values emphasized that maternal height is negatively correlated with stunting (**Figure 2**). Therefore, taller mothers were less likely to have stunted children. Interestingly, a shorter mother in Bangladesh produced a stronger impact on stunting than taller mothers as lower height produced higher SHAP scores. We also observed negative associations between Assets Score, parental education, and Mother's Age on stunting as higher values of these features scored negative SHAP values (**Figure 2**). The separation of values based on sex/gender across the x-axis showed that females (red) were less likely to be stunted than males (blue). The number of children in the household—Birth Order, Persons <18 in house, and Multiple Births—had high feature values plotted on positive SHAP values. As the number of children in a household increased, so did the likelihood of stunting in

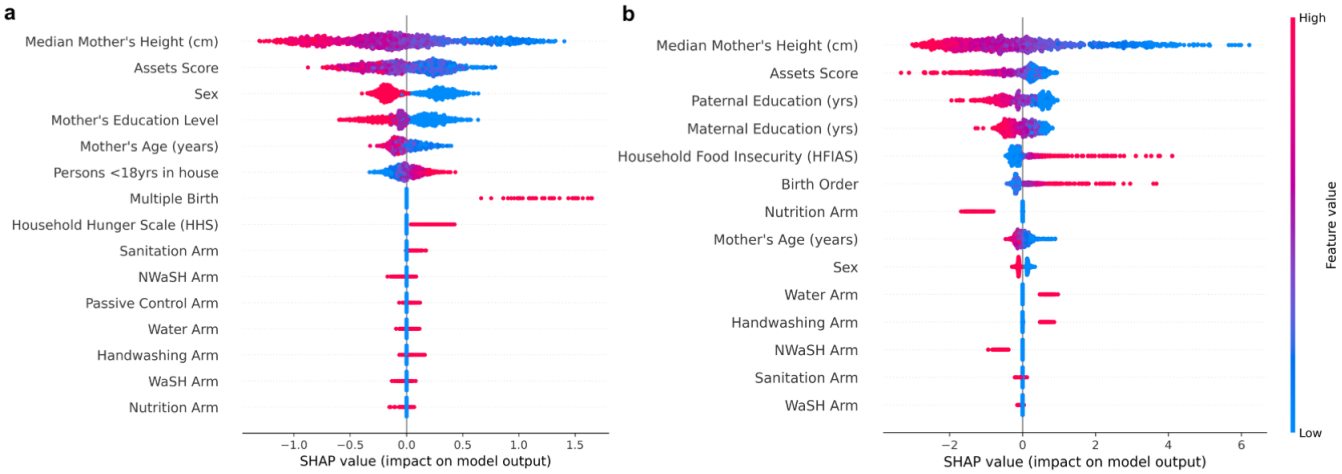

Table 2. SHAP summary plots for the soft voting ensembles for (a) Kenya and (b) Bangladesh. The SHAP summary plots represent the feature importance in descending order and the correlation between feature value and stunting. The SHAP library was used to generate the summary plot using models, maskers, and feature values. Negative SHAP values denote a reduction in the prediction (non-stunted) and positive values increase prediction (stunted).

the TPOT models (**Figure 2**). Similar positive relationships between features and stunting were recorded for the household hunger/food insecurity scores in both countries.

According to the TPOT ML model analysis, most interventions in Kenya had no correlation with the growth outcome as implemented interventions (denoted by the red) were symmetrically distributed across a SHAP score of 0. Sanitation Arm was the only intervention in Kenya that slightly increased stunting risk in the models (**Figure 2a**). The Nutrition and NWaSH (Nutrition, Water, Sanitation, and Handwashing) intervention Arm in Bangladesh appeared effective at reducing stunting due to the negative SHAP values recorded for these interventions. However, the Water and Handwashing interventions increased the likelihood of stunting as the "hits" produced positive SHAP values (**Figure 2b**).

ML model performance metrics

A rigorous analysis of each optimized TPOT pipeline and the soft voting ensemble provided greater understanding of the model performances. This study accounted for the differing metric outputs by measuring each model across seven performance statistics: Area Under the Receiver Operating Characteristic Curve (ROC AUC), Area Under Precision v. Recall Curve (PR AUC), Accuracy Score (AS), Balanced Accuracy (BA), Precision Score (PS), Recall Score (RS), and F1 Score (F1S). ROC AUC demonstrates the tradeoffs between the true positive and false positive rates and PR AUC demonstrates how accuracy changes as precision changes. AS, BA, and F1S are all variations of accuracy metrics using different methods to gauge overall performance. Additionally, a confusion matrix was generated for each TPOT model and soft voting ensemble to summarize performance. Confusion matrices provide greater context into the performance of models such as what values are being predicted correctly and which ones incorrectly. Using the confusion matrix, we specifically further analyzed the RS and PS. RS, calculated using the following equation:

$$RS = \frac{TP}{TP + FN}$$

which explains how well a model can identify stunting from all stunting cases. The PS, determined by:

$$PS = \frac{TP}{TP + FP}$$

shows how many stunting predictions were correct. There is a general tradeoff between the PS and RS which helps us better gauge a model's usability and overall performance.

The pipeline for the Kenya clinical trial yielded more predictive models than those generated by the Bangladesh pipelines (**Table 1**). The Kenya models outperformed the Bangladesh models by a score of ~0.1 on a scale of 0-1 on four metrics: ROC AUC, PR AUC, AS and PS. The Kenya TPOT pipeline run with a random seed 72 (P72) was the outlier as its final model performed comparable to the Bangladesh models, differing in metric scores by ~0.05 or less. Kenya's P72 had the greatest RS, BA, and F1S of all models across both studies. However, high RS, BA, and F1S scores led to underperformance in the other metrics. The BA for both trials was comparable, with most scores ranging between 0.60 and 0.65 (**Table 1**).

The confusion matrices elaborate on the PS, RS, and the type of error made by the models. The recall score measures the model's ability to identify stunting from all stunting cases in the data. The Kenya models recorded twice larger false negative (FN) counts than true positive (TP) counts, yielding lower recall scores (**Figure 3a**). Conversely, the Bangladesh

Table 1. Performance Metrics for TPOT models and soft voting (SV) Ensemble.

Metrics		ML Pipeline											
		SV	P12	P24	P34	P44	P50	P68	P72	P75	P100	P124	AVG
ROC AUC	KE	0.828	0.894	0.818	0.803	0.869	0.798	0.799	0.736	0.858	0.818	0.769	0.817
	BD	0.670	0.693	0.698	0.695	0.674	0.684	0.691	0.690	0.697	0.643	0.684	0.684
PR AUC	KE	0.681	0.796	0.659	0.641	0.745	0.632	0.628	0.574	0.726	0.666	0.593	0.667
	BD	0.588	0.564	0.586	0.593	0.596	0.550	0.576	0.580	0.582	0.585	0.544	0.577
AS	KE	0.770	0.802	0.756	0.743	0.774	0.751	0.751	0.709	0.773	0.759	0.740	0.757
	BD	0.644	0.616	0.643	0.648	0.652	0.634	0.636	0.641	0.639	0.652	0.644	0.641
BA	KE	0.667	0.702	0.643	0.620	0.660	0.632	0.635	0.704	0.659	0.642	0.617	0.653
	BD	0.638	0.605	0.646	0.643	0.647	0.608	0.626	0.641	0.638	0.631	0.624	0.632
PS	KE	0.743	0.850	0.724	0.705	0.810	0.724	0.716	0.524	0.802	0.754	0.695	0.731
	BD	0.553	0.521	0.547	0.556	0.562	0.552	0.544	0.546	0.543	0.575	0.562	0.551
RS	KE	0.395	0.439	0.347	0.296	0.357	0.319	0.328	0.691	0.357	0.334	0.292	0.378
	BD	0.605	0.554	0.660	0.618	0.616	0.478	0.577	0.639	0.634	0.520	0.522	0.584
F1S	KE	0.516	0.579	0.469	0.417	0.496	0.443	0.450	0.596	0.494	0.463	0.411	0.485
	BD	0.577	0.537	0.598	0.585	0.588	0.512	0.560	0.589	0.585	0.546	0.541	0.565

NOTE: *To represent the models based on country and the performance succinctly in the table, we used abbreviations for many terms: Kenya models (KE), Bangladesh models (BD), Area Under the Receiver Operating Characteristic Curve (ROC AUC), Area Under Precision v. Recall Curve (PR AUC), Accuracy Score (AS), Balanced Accuracy (BA), Precision Score (PS), Recall Score (RS), F1 Score (F1S), and Average (AVG).*

models labeled stunting well as the TP was, on average, 63 counts greater than the FN which led to a higher recall score (**Figure 3b**). Thus, the Bangladesh models were better able to detect stunting compared to the Kenya models. The precision score measures how many of the children labeled stunted were stunted. The Kenya ML models, except for P72, had a TP to false positive (FP) ratio greater than 2 (<66 FP count and >138TP count) when compared to the Bangladesh models, which led to a higher precision score (**Figure 3a**). The Bangladesh ML models had an average of 0.18 lower precision scores, with FP counts over 100 individuals higher than the Kenya models (**Figure 3b**). In context, a stunting label by Kenya ML models was considered accurate with high confidence, and a stunting prediction by the Bangladesh ML models was considered accurate with moderate confidence. Overall, the TPOT models worked well for both datasets

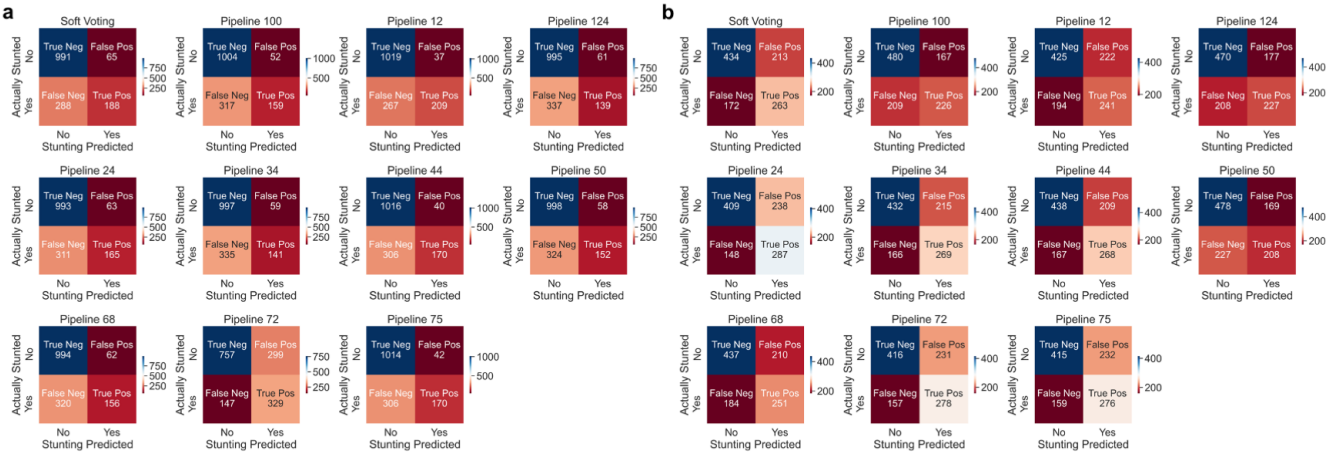

Figure 3. Confusion matrices for ML models for (a) Kenya and (b) Bangladesh. The true positive (TP), false positive (FP), true negative (TN), and false negative (FN) count for each ML model is plotted. The predicted label for given features were compared to the actual label in the validation dataset to generate the confusion matrix values.

in predicting childhood stunting and produced favorable performance scores given the complex problem domains.

DISCUSSION

The ML analysis results supported our hypothesis in the emphasis of the negative correlations between demographic and socio-economic features on stunting risk and the ranking of maternal height, household wealth, and parental education as the most prominent features. Additionally, the outperformance of previous ML models by TPOT affirmed TPOT's applicability in childhood stunting domains. Our analysis first consisted of an initial correlation analysis using heat maps to better understand the features and their linear relationship with one another. Familial and socio-economic conditions in Kenya were not correlated with one another as most pairwise correlations had absolute values below 0.15 (**Figure 2a**). Therefore, higher maternal education levels did not necessarily lead to a higher economic wealth (Assets Score) or a higher Assets Score to lower Household Hunger Scale. Features for Bangladesh recorded greater correlation with one another, as many absolute correlation values were greater than or equal to 0.3 (**Figure 2b**). Interestingly, parental education levels were strong indicators of wealth, Household Food Insecurity Access Scale values, and the number of children of a given family in Bangladesh. These results for Bangladesh aligned with results from previous studies that related education, wealth, and food access (14-15). The difference between the two countries' results stressed the nuance when working with clinical trial data. A generalization gathered from Bangladesh about stunting would not necessarily hold true for Kenya or another country (16). Thus, we did not draw any initial conclusions from the correlation heat maps, relying on TPOT's ML model performance to gain greater insight into the data. The data gathered here was then used in conjugation with the results from the model analysis to explain/interpret performance.

Compared to previously constructed ML models using conventional techniques, the TPOT models scored higher across multiple metrics. A study in Ethiopia that developed predictive ML models for childhood stunting based on environmental factors reported accuracy below 70% and varying sensitivity (RS) and specificity rates (9). Our models for Kenya yielded accuracy scores above 80% and consistent specificity (true negatives divided by the sum of true negatives and false positives) above 1.5 on a smaller dataset. Additionally, our ML models for Bangladesh improved on previous studies by increasing the sensitivity rate and specificity by over 0.1 while maintaining almost identical accuracy scores (8). The improvement over previous ML models in this work confirmed that TPOT developed more accurate models for predicting childhood stunting. We believe that TPOT outperformed traditional, non-automated pipelines because of the thorough search it conducts when finding the optimized pipeline. In the same timeframe that a traditional pipeline builds and compares a small quantity of models, TPOT can construct an order of magnitude more. This difference in efficiency is because computers progress through feature engineering, model selection, and hyperparameter optimization at a faster pace than humans can. With 100 generations and populations in one TPOT run, we tested over 10,000 pipeline models. Ten runs of TPOT on a dataset effectively examined 100,000 pipelines. While an increased use of AutoML like TPOT in healthcare applications is pivotal, manual processes for effective data cleaning, model validation, and overall supervision are still vital as they require the expertise of a data scientist. The only limitation of our conclusion is the subtle variations in data used by this study and the Bangladesh and Ethiopia studies using traditional methods. Sampling methods and the year in which the data was collected can lead to fluctuations in model accuracy. Given that all the models compared in this study used demographic data collected between 2010 and 2016, we assumed that the model performance variation caused by variable sampling times was negligible.

We believe that the pipeline performance for both countries can be further improved. Currently, the Kenya ML models performed extremely well, with scores near 0.80 and 0.75 on metrics like AS and the ROC AUC. However, the Kenya models underperformed on RS and F1S, with scores averaging 0.378 and 0.485, respectively (**Table 1**). The Bangladesh models also performed respectably on all seven metrics, with the highest score in ROC AUC and the lowest in PS. We believe that the Kenya ML models had a higher performance because of the low feature correlation. As the correlation begins to approach ±1, the features determine each other. Therefore, highly related features do not provide new predictive information to the ML model and can be substituted by one another. As Bangladesh features had higher correlation, the familial and socio-economic indicators provided less information about the target child's environment to the models. For example, we could have substituted maternal with paternal education in Bangladesh without significantly altering the performance of the models. In future Bangladesh studies, we could consider substituting some of the familial and socio-economic indicators. Viable options include replacing Paternal Education and Birth Order—two variables having large positive correlations to more influential ones—with Union Council (location), parent's employment type, and Caregiver Weight. These new features would improve our models and even reveal new relationships between the added features and stunting.

According to the confusion matrices for both countries, the ML models most accurately identified non-stunted children with 400+ TN for Bangladesh and 990+ for Kenya (except for P72). While Bangladesh had an equal number of FN and FP, Kenya recorded high FN and low FP (**Figure 2**). We believe the high number of FN predictions for Bangladesh and Kenya are caused by class imbalances in the data. A higher number of non-stunted children in the training data caused the ML models to become insensitive to stunted

children. We propose two potential solutions that could improve the RS and PS in future models. The first solution is resampling the dataset by either adding duplicates of existing stunted children or deleting non-stunted children. The drawback of this solution is propagating biases within the data, as important correlations/relationships can be missed by the model. The second method to balance the classes is synthetically generating samples. Although this will generate non-duplicate data points, nonlinear relationships may not be preserved. Both techniques can be of interest in further studies as they could significantly increase the ML model's accuracy.

Based on improved performance metrics when compared to past studies from Bangladesh and East Africa, we believe that the TPOT models are deployable algorithms for predictive risk assessments (8, 9). Therefore, they can be used by medical professionals in assessing a household's risk of having stunted children. A potential use case is as follows: medical professionals would initially record the socio-economic and familial features used by the ML model for each family. The model would then be run to generate predictions where the active intervention feature (e.g., water, handwashing, NWaSH, etc.) was altered each time. Risk of having a stunted child for a household would range from substantial risk, a stunting outcome in each ML prediction, to minimal risk, no stunting outcome in each ML prediction. This is not the only or best deployment method, but we believe that it will produce the most reliable results and help reduce stunting through preemptive action in these two countries.

The SHAP correlation plots for the TPOT ML models also improved our understanding of various interventions and their impact on stunting. We observed that most interventions in Kenya were not determinant of stunting. However, in Bangladesh, the Water and Handwashing Arms were likely to reduce stunting, while the Nutrition and NWaSH Arms increased the likelihood of stunting (**Figure 3**). The varied results for intervention success in this study is not necessarily indicative of effectiveness. The clinical trials minimized the contact between the intervention provider and the target child's family (17). Thus, we speculate that the limited communication about proper intervention usage, yearly surveying, and self-reporting of medical symptoms caused many families to insufficiently maximize an intervention's positive impact. Future studies conducting clinical trials with stronger interventions are necessary to properly understand how a given intervention impacts stunting outcomes. However, we determined that intervention success is nuanced and varies by region and country. Therefore, no generalized conclusion could be reached about the most effective intervention. A case-by-case basis should be considered for intervention implementation in the future.

We believe that the negative correlation between median mother's height and stunting has two components. The first component is genetics. Taller mothers likely pass forward different height alleles to their children, leading to taller children (18). The second factor is the birth complication, which increases as maternal height decreases, leading to children with delayed structural and neurological growth (5). Additionally, we observed higher parental education correlating with lower stunting. Educated parents are more likely to make informed decisions regarding nutrition and sanitation of their child, which promotes healthy growth in children (3, 18). The negative association between Assets Score and stunting demonstrated that material wealth indicators are indicative of the overall family's economic stability. Therefore, a household with commodity items like TVs and cars would also likely provide a healthy growth environment for a child. However, an increased number of children in the household would counteract the positive impacts of greater economic wealth on stunting. The cost of providing nutritious food and proper sanitation would increase, and each member of the family would receive a smaller portion of limited resources (17). We additionally observed that infant males are at higher risk of stunting as the blue values for the Sex feature—representing males—scored positive SHAP values while the female counterparts—represented by red feature values—scored negative SHAP values. Thus, male children face a higher risk of stunting and should receive more attention during trials or intervention programs.

Features in this study, and thus the feature importance results, do not reflect changes in wealth, access to resources, and sanitary conditions from the baseline survey when the demographic conditions were recorded to the second study timepoint. The presence of the intervention administrator could have caused variations in the family's daily habits such as seeking nutritious food or improving overall sanitation. To account for these changes in the future, a baseline demographic survey could be conducted each year, which would further validate our results.

In future studies we could expand the data used when constructing the ML models, along with potential options that include adding sanitation features into the dataset or considering clinical trials from more countries. Expanding this project could help us better understand the regional variance between feature-stunting correlation, helping better inform policy decisions and underlining effective solutions to stunting regionally. Overall, we believe that the results presented here shed light into the socio-economic and familial conditions that influence stunting. We anticipate our results will guide future studies and combative measures taken against childhood stunting to reach the United Nations' goal of a 40% reduction in stunting by 2025.

MATERIALS AND METHODS

Data acquisition and feature extraction

The data used for this study were accessed from ClinEpiDB, an open-access online database containing various epidemiological clinical trials (20). The specific data considered in this study consisted of the WASH Benefits Kenya and WASH Benefits Bangladesh Cluster Randomized

Trials conducted between 2012 and 2015. These trials focused on the individual and combined impacts of water, sanitation, hygiene, and nutritional interventions on diarrhea and infant growth. Participants for the Bangladesh and Kenya region were selected from rural regions associated with low sanitation and water quality and equally split into eight clusters, each representing a different intervention arm. This study focused on the children in utero during the enrollment and born within six months of the baseline survey (target children). Enrolled target children in Kenya (n = 8000) and Bangladesh (n = 5760) were measured for diarrhea and growth outcomes at 12 months (timepoint 1) and 24 months (timepoint 2) after the baseline enrollment. One-fourth of the children were in the control group, while the remaining 3/4 were equally distributed between 6 different interventions. The baseline enrollment survey recorded information on the community (e.g., intervention cluster arm), household (e.g., sanitation measures, utilities, and wealth indicators), participant (parental information, sex, and enrollment/survey flags), and observations (disease indicators and growth indicators recorded during timepoints 1 and 2).

The target variable stunting was donated by a yes or no, where an infant was considered stunted if the BMI-for-age z-score was lower than -2. Feature selection for the ML models excluded recorded observations from time point 1 and 2 for two main reasons: a) stunting indicator highly correlates with the growth z-scores, which would positively skew the models' accuracy, and b) recorded observations occurred during intervention implementation and did not fit this study's goal of predicting stunting and assessing stunting risk before birth. Selected features provided insight into the impact of socio-economic status, paternal/maternal education level, genetics, and other environmental factors on infant stunting and intervention effectiveness. Slight variations between the Kenya and Bangladesh features occurred due to the different demographic settings, but the overall theme and structure was identical.

The open-source Python library *pandas* was used to extract, visualize, and encode the desired data from the clinical trials (21). Individuals with missing data values were excluded from analysis as some of the ML models used in our analysis do not support missing values and standard imputation methods cannot always accurately represent the missing value. Simple binary features (yes/no or male/female) were binary encoded to values of 0 or 1. The intervention arm feature with 7 different interventions were one-hot encoded into 6 dummy variables with binary values, where a 1 represented a use of the given intervention. The assets score features combined individual economic indicators by multiplying the value of each feature to a weight between 0 and 2 with a higher weight signifying greater utility and cost. The higher the score, the more economically advantaged the household.

ML model building

This study used TPOT Classifier and provided a configuration dictionary to customize the algorithms, transformers, and hyperparameters used. This configuration dictionary was tailored to classification problems by narrowing the scope of classification models, preprocessors, and selectors used during pipeline exploration. We used naive Bayes (Gaussian, Bernoulli, and multinomial), decision trees, extra tree, random forest, gradient boosting, extreme gradient boosting, *k*-nearest neighbor, logistic regression, and multi-layer perceptron classifiers considered in the TPOT pipeline optimization process. These were used because they are well established in biomedical applications and are known to be good classifiers.

The Kenya and Bangladesh data were divided into a 75%/25% training/testing set using a constant randomization seed of 44. TPOT was then run 10 times for the datasets with 10 different random seeds (12, 24, 34, 44, 50, 68, 72, 75 ,100, 124), each time using 100 generations and 100 populations. Given that TPOT uses randomization to generate new pipelines through the run, 10 runs of TPOT increased the working sample size, reducing the biases of using only one randomized run.

Ensemble voting classifiers combine distinct and similar ML models via majority/plurality voting schema. Hard voting ensembles use majority voting, where the class label ($\hat{y}$) is predicted by taking the mode of each individual classifier's C_j prediction:

$$\hat{y} = mode\{C_1(x), C_2(x), \dots, C_m(x)\}.$$

Soft voting ensembles use the model's predicted probabilities (p) from each respective classifier and apply a weight (w_j) to reach a prediction. The following equation provides the mathematical construct for weighted soft voting ensembles:

$$\hat{y} = arg\ max\ \sum_{j=1}^{m} w_j p_{ij}\ .$$

The soft voting ensemble more accurately combines the models produced for each dataset as it leverages individual predicted probabilities and applies weightage to each prediction. Thus, the 10 optimized models from each TPOT run for both datasets were combined using a soft voting ensemble from the *MLxtend* library (22). This study utilized weights of 1 for each model with the intent of equalizing the influence of each prediction probability on the output.

Result and data analysis

A preliminary analysis of the features occurred before building the TPOT and soft voting ensembles. The six dummy variables representing the intervention arm were excluded from the data as the interventions were randomly assigned. Correlations were effectively visualized using heatmaps. Heatmaps were generated using the built-in functionality in the *pandas* library and *seaborn* (23).

The final evaluation conducted on the ML models

consisted of extracting feature importance. Feature importance is a technique used to understand the importance of a given feature over a model's prediction. This analysis tool is particularly useful for providing data understanding and model interpretability (globally and locally). SHAP summary plots for the soft voting ensembles were generated to gain interpretability into the models and understand the impact of feature values on the output. Generating SHAP plots is computationally demanding; we optimized this process by taking the SHAP values for the Soft Voting Ensembles, which leverages each TPOT model's weighted prediction to arrive at one final classification. Correlation is interpreted by analyzing the distribution of feature values with respect to the SHAP value. TPOT models were excluded from feature importance extraction as they would yield almost identical results to the soft voting ensembles, providing no new insight.

ACKNOWLEDGEMENTS

Special thanks to Dr. Elisabetta Manduchi for mentoring A.S. throughout the period of research and proofreading the manuscript.

APPENDIX

GitHub assess to data, code, and additional files: github.com/AdiSir05/TPOT-Stunting-JEI

Received: March 24, 2022
Accepted: August 14, 2022
Published: September 25, 2022

REFERENCES

1. World Health Organization., "What Do We Need to Know About Child Stunting?" Equity Considerations for Achieving the Global Nutrition Targets 2025, World Health Organization, 2018, pp. 4–5.
2. Lucci, Paula, *et al*., "Are We Underestimating Urban Poverty?" *World Development*, vol. 103, 2018, pp. 297-310, doi:10.1016/j.worlddev.2017.10.022.
3. de Onis, Mercedes, and Francesco, Branca. "Childhood Stunting: A Global Perspective." *Maternal & Child Nutrition,* vol. 12, Suppl 1, 2016, pp. 12-26, doi:10.1111/mcn.12231
4. Victora, Cesar Gomes *et al*., "Worldwide Timing of Growth Faltering: Revisiting Implications for Interventions." *Pediatrics,* vol. 125, no. 3, 2010, pp. e473-480, doi:10.1542/peds.2009-1519.
5. Dewey, Kathryn G, and Khadija Begum. "Long-term Consequences of Stunting in Early Life." *Maternal & child nutrition,* vol. 7 Suppl 3, 2011, pp 5-18, doi:10.1111/j.1740-8709.2011.00349.x.
6. Walker, Susan P., *et al*. "Early Childhood Stunting is Associated with Lower Developmental Levels in the Subsequent Generation of Children." *The Journal of nutrition,* vol. 145, no. 4, 2015, pp. 823-828, doi:10.3945/jn.114.200261.
7. Obermeyer, Ziad, and Ezekiel J. Emanuel. "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." *The New England journal of medicine*, vol. 375, no. 13, 2016, pp. 1216-9, doi:10.1056/NEJMp1606181.
8. Mansur, Mohaimen *et al*., "Sociodemographic Risk Factors of Under-five Stunting in Bangladesh: Assessing the Role of Interactions Using a Machine Learning Method." *PloS one,* vol. 16, no. 8, Aug. 2021, doi:10.1371/journal.pone.0256729.
9. Bitew, Fikrewold H *et al*., "Machine Learning Algorithms for Predicting Undernutrition Among Under-five Children in Ethiopia." *Public health nutrition*, Oct. 2021, pp. 1-12, doi:10.1017/S1368980021004262.
10. Waring, Jonathan *et al*., "Automated Machine Learning: Review of the State-of-the-art and Opportunities for Healthcare." *Artificial Intelligence in Medicine,* vol. 104, 2020, pp. 101822, doi:10.1016/j.artmed.2020.101822.
11. Le, Trang T *et al*., "Scaling Tree-based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector." *Bioinformatics (Oxford, England),* vol. 36, no. 1, 2020, pp. 250-256, doi:10.1093/bioinformatics/btz470.
12. Olson, Randal S. *et al*., "Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science." *Proceedings of the Genetic and Evolutionary Computation Conference,* July 2016, pp. 485-492, doi:10.1145/2908812.2908918.
13. Lundberg, Scott M. *et al*., "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4763-4774, doi:10.3389/frai.2021.752558.
14. Knueppel, Danielle *et al*., "Validation of the Household Food Insecurity Access Scale in Rural Tanzania." *Public health nutrition,* vol. 13, no. 3, 2010, pp. 360-367, doi:10.1017/S1368980009991121.
15. Anand, S *et al*., "Urban Food Insecurity and its Determinants: A Baseline Study of Bengaluru." *Environment and Urbanization*, vol. 3, no. 2, 2019, pp. 421–442, doi:10.1177/0956247819861899.
16. Hossain, Muttaquina *et al*., "Evidence-based Approaches to Childhood Stunting in Low- and Middle-Income Countries: A Systematic Review." *Archives of disease in childhood* vol. 102, no. 10, 2017, pp. 903-909, doi:10.1136/archdischild-2016-311050.
17. Arnold, Benjamin F *et al*., "Cluster-randomised Controlled Trials of Individual and Combined Water, Sanitation, Hygiene and Nutritional Interventions in Rural Bangladesh and Kenya: The WASH Benefits Study Design and Rationale." *BMJ open,* vol. 3, no. 8, Aug. 2013, pp. e003476, doi:10.1136/bmjopen-2013-003476.
18. Phiri, Thokozani. "Review of Maternal Effects on Early Childhood Stunting." *Grand Challenges Canada Economic Returns to Mitigating Early Life Risks Project Working Paper Series*, 2014-18.
19. Guerrant, R L *et al*., "Diarrhea as a Cause and an Effect of Malnutrition: Diarrhea Prevents Catch-up Growth and Malnutrition Increases Diarrhea Frequency and Duration." *The American journal of tropical medicine and hygiene,* vol. 47, no. 1 Pt 2, 1992, pp. 28-35, doi:10.4269/ajtmh.1992.47.28.
20. Ruhamyankaka, Emmanuel *et al*., "ClinEpiDB: an Open-access Clinical Epidemiology Database Resource Encouraging Online Exploration of Complex Studies." *Gates open research,* vol. 3, no. 1661. Apr. 2020, doi:10.12688/

gatesopenres.13087.2.
21. McKinney, Wes *et al*., "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference,* 2010, pp. 56-61, doi: 10.25080/Majora-92bf1922-00a.
22. Raschka, Sebastian. "MLxtend: Providing Machine Learning and Data Science Utilities and Extensions to Python's Scientific Computing Stack." *The Journal of Open-Source Software*, vol. 3, no. 24, April 18, doi: 10.21105/joss.00638.
23. Waskom, Michael L. "Seaborn: Statistical Data Visualization." *Journal of Open-Source Software*, vol. 6, no. 60, 2021, pg. 3021, doi: 10.21105/joss.03021.