Efficient Machine Learning for Malnutrition Prediction among under-five children in India

Saksham Jain*
Department of Electrical Engineering
Aligarh Muslim University
Aligarh, India
saksham1598@gmail.com

Tayyibah Khanam*
Department of Electrical Engineering
Aligarh Muslim University
Aligarh, India
tayyibahkhanam@zhcet.ac.in

Ali Jafar Abedi
Department of Community Medicine
Aligarh Muslim University
Aligarh, India
ajabedi@myamu.ac.in

Abid Ali Khan
Department of Mechanical Engineering
Aligarh Muslim University
Aligarh, India
abid.khan.me@amu.ac.in

2022 IEEE Delhi Section Conference (DELCON) | 978-1-6654-5883-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/DELCON54057.2022.9753080

***Abstract*—Child malnutrition is considered to be one of the leading causes of infant mortality and malnutrition. This study was aimed to leverage the advantages offered by machine learning models in terms of determining and accurately predicting significant factors of malnutrition. For this study, the Children's recode files from the Indian Demographic and Health Survey (IDHS) datasets from 2005-2006 and 2015-2016 were used. To examine the nutritional status of children aged 0-59 months, this study looks at stunting (Height-for-age), wasting (Weight-for-Height), and concurrent stunted wasting (Height-for-age-Weight-for-Height). Regular Machine Learning models, Tabular Deep Learning frameworks, H2O base models, and AutoML models are the four types of machine learning models employed in our research. This research found that Automated machine learning algorithms and Tabular Deep Learning frameworks, in general, outperformed other models in terms of speed and efficiency, as well as Accuracy (up to 96.46%) and AUC-ROC scores (up to 99.95%), which are important in classification problems like this one. Following a graphical representation of the importance of numerous drivers of malnutrition for all three anthropometric indices, we concluded our findings by comparing the performances of several models and determining the top-performing algorithms. This paper significantly contributes to the possibilities of using machine learning in identifying probable correlates of malnutrition for the effective prevention, cure, and identification of target groups.**

***Index Terms*—AutoML algorithms, Feature Importance, Malnutrition, Tabular Machine Learning**

I. Introduction

Starvation, dietary deficiency and sanitation are some of the major obstacles to achieving economic and social development goals in any developed nation. According to the World Health Organization, there were 149.2 million stunted children under the age of five, 45.4 million wasting children, and 38.9 million overweight children in 2020. With the exception of Africa, the number of stunted children is decreasing in all regions. Southern Asia is home to more than half of all wasting children, while Asia as a whole is home to more than three-quarters of all children suffering from severe wasting. In terms of the targets, the stunting aim is making the most progress at the country level, with approximately two-thirds of countries experiencing at least some progress [12].The world's most malnourished people are in developing countries in southern Asia and Africa and hence, malnutrition rates in Southern Asia and Sub-Saharan Africa are particularly high, among the highest ones in the world [5].

Thus, malnutrition among children under the age of five is a major public health issue in India and thus is also responsible for about half of all under-5 child deaths in India. The graveness of this situation is evident by the fact that India has one of the highest rates of stunted children in the world, approximately 46.6 million which is double that of Sub-Saharan Africa [14] and also accounts for one-third of the global total according to Global Nutrition Report 2018 [15]. Poor nutrition in the first few years of a child's life can also contribute to restricted growth, which is directly linked to poor cognitive abilities, reduced understanding and attention at school and thus, is a major hindrance to a child's holistic progress and development. Furthermore, in countries where malnutrition is prevalent, the implications of malnutrition extend far beyond the person, and thus contribute in reduction of total labor-force productivity and economic growth. Other significant indicators of malnutrition in children aged 0 -5 are low birth weight, diarrhoea within the last six months, and physical developmental delay in most underdeveloped nations, including India. While socioeconomic status of the family is one of the major factors contributing to health issues in a child and directly affects the child's dietary score, sanitation and lifestyle, other external factors such as environmental pollution, political commitment and the role of governmental policies can't be ignored.

To lessen the risk of malnutrition and its dire consequences in the Indian society, it is thus important to identify other significant concentrations through background factors and independent factors, the level of importance of each contributing factor and the correlation between these factors. With the

*Equal Contribution

goal of proper identification and treatment of malnutrition, researchers in the past have aimed to accurately predict the linked condition through machine learning techniques, with the primary aim of accurate malnutrition prediction to reduce the risks. While most existing research attempts to predict malnutrition through one or more anthropometric indices, we consider a concurrent stunted-wasting index based on the studies and evidence found in the literature regarding its relevance. Finally, our study extends a novel empirical approach through Machine Learning algorithms to be able to flexibly determine to what extent malnutrition is driven by several detected social, economic, and environmental factors. This also gives us the opportunity to examine the effect of other undetected factors such as culture, politics, and conflict.

The rest of the paper is sectioned as follows. Section II discusses the background behind Machine Learning and Malnutrition. Section III discusses related works and the methodology of our approach with its novel contributions. Section IV, V & VI give overview on the data-sets, discusses the implementation & results and summarize the paper findings respectively.

II. Background

Considering India's situation, the irony is that India, the world's second largest food producer, also has the highest number of undernourished children. Several factors tend to explain such an ironical situation, one of them being the distribution of wealth and hence resources among citizens of the country, which directly leads to malnutrition being a concentrated phenomenon instead of being evenly spread across the country. [7] observes states, districts and villages being one of the concentrations with only 5 states and 50% of villages accounting for about 80% of the malnutrition burden. Because Machine Learning techniques are widely employed in determining the indicators of malnutrition, analyzing the trends and building predictive models to ameliorate the risks of malnutrition, we attempt to give a brief background on both - Machine Learning and Malnutrition.

A. *Machine Learning*

Machine Learning algorithms learn by searching for patterns in huge amounts of data, and updating the program when it finds one in order to reflect the "reality" of what it discovered. Deep learning algorithms, a subset of Machine Learning algorithms have emerged as successful alternatives to traditional machine learning models, especially when dealing with big data since they provide an exceptionally sophisticated approach to learning and are poised to solve these difficulties. These are deep, multi-layered, neural networks that allow data to undergo non-linear transformation. As researchers our goal while designing a machine learning algorithm is to maintain the trade-off between optimization and generalization to obtain decent accuracies on both, the training set and the test set.

B. *Malnutrition*

Malnutrition encompasses both under-nutrition and over-nutrition, and refers to deficiencies, excesses, or imbalances in a person's intake of energy and nutrients. In general, malnutrition refers to three distinct conditions: Undernutrition, Micronutrition and Overweight.

Of the three conditions, under-nutrition makes children in particular much more vulnerable to disease and death [1]. Each of the three subcategories in undernutrition attempts at explaining a different malnourished condition.

- *Wasting* refers to the condition where a child is too thin for his or her height. Wasting (WH) is caused by a recent rapid weight loss or an inability to gain weight.
- A child who is too short for his or her age is said to be *stunting*. These children may suffer permanent physical and cognitive harm as a result of their stunted growth. Stunting's (HA) devastation can last a lifetime and even damage future generations.
- An *underweight* (WA) child tends to have a lower weight for his/her age. This results in poor stamina and weak immune system, leaving him/her open to infections and diseases.

Additionally, Body Mass Index (BMI) is also commonly used as the fourth anthropometric index in various studies. BMI reflects a similar health index idea by categorizing a child as overweight, underweight, or normal weight based on equation 1.

$$BMI = \frac{W}{H^2} \tag{1}$$

Where W represents the weight of a child in kilograms & H represents the height of a child in meters.

III. Related Work & Methodology

One of the most recent works in the field of malnutrition detection and prediction was done by Khare et al. [9] utilizing the Indian Demographic and Health Surveys. They designed a prediction model for malnutrition using the four main factors of study - BMI (Body Mass Index), HAZ (Height for Age Z score), WAZ (Weight for Age Z score) WHZ (Weight for Height Z score) to compare the features identified by Machine Learning models with the ones identified in the literature. Their experiments run in two phases, where Phase 1 corresponds to a three step procedure.

- Data Pre-processing: To manage class imbalances through techniques like SMOTE (Synthetic Minority Oversampling Technique) for re-sampling of the lesser frequent class.
- Evaluating Gain Ratio Features for calculating the degree to which various features are able to distinguish between HIGH, MEDIUM & LOW of each of the dependent variables.
- Feature Selection methods and Decision Tree Classifier for final model prediction.

Phase 2 on the other hand deals with feature extraction depending on the level of significance and odd ratio parameters and construction of a Logistic Regression Model for all four dependent variables.

Given that malnutrition in children is a major concern, studies in the past have attempted to narrow their research on children, especially infants. While Talukdar & Ahammed [16] aimed at predicting malnutrition particularly in children under the age of five in Bangladesh, Briend et al. in their study [11] target the age group between 5-18 in their study. [16] employed several Machine Learning algorithms and limited their study to the predictions of only one anthropometric index, i.e., WAZ. Performance parameters of Accuracy, Specificity and Sensitivity were utilized to calculate the predictive performances of these algorithms, whereas the discriminative precision of these algorithms were determined utilizing Cohen's k statistic. [11] not only aimed at determining the factors responsible for the other two anthropometric indices of WHZ & HAZ in Ethopian street children. These factors and their prevalence were identified through bi-variate and multi-variate logistic regression analyses.

Another study [13] utilized machine learning techniques for an individual decision support system for personalized nutritional treatment of individuals. Leveraging the ability of machine learning algorithms to recognize malnutrition, they intend to use this as a screening tool for their nutritional treatment plan.

Methodology: Our study overlaps between the works done in [16], [9] & [11]. While our goal is similar to that of [16], i.e., determining crucial factors of malnutrition prediction in children under the age of five in India through several machine learning algorithms, we are also keen on utilizing techniques for feature learning and re-sampling as demonstrated in [9]. Further, unlike the existing research that utilized regular shallow and deep machine learning algorithms such as (K Nearest Neighbours (KNN), Logistic regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN)) we introduce several other highly performative AutoML (Automated Machine Learning) algorithms and specifically designed libraries for tabular data-sets in our study. Through these algorithms and libraries, the process of data pre-processing, feature learning and model selection becomes automated. Lastly, most studies use accuracy as their metric for binary classification, and we note that accuracy alone cannot be utilized to evaluate the performance of a model. Thus, we introduced AUC-ROC (Area under the Receiver Operating Characteristic Curve) score for our evaluated models.

Further, each of these studies revolve around one or more of the four basic anthropometric indices. To this end, we utilize only two of these existing indices in our study: Wasting and Stunting, due to their stated importance in previous studies over the other two anthropometric indices, and also test our model on a hybrid index of HAZ & WHZ, known as Concurrent Stunted-Wasting (HAWHZ). Although Stunting and Wasting are defined as separate parameters of child malnutrition, various studies manifest that they are closely related and can be grouped together. Moreover, the occurrence of stunting and wasting is usually observed in the same populations and often in the same children [4]. A possible connection between these two forms of malnutrition is the decreased muscle mass in the body. A study [3] suggests that muscle mass plays a vital role in overall health and survival of the child. While both stunting and wasting contribute to increased child mortality, the increase in mortality is tremendous when both are simultaneously present in the same child. Children with severe wasting are often stunted, suggesting that wasting and stunting have a common cause or that one form of malnutrition can contribute to the development of the other. Therefore, it is reasonable to group these two parameters together, and this hybrid form of under-nutrition is commonly referred to as "Concurrent Stunted-Wasting", represented by HAWH. This novel hybrid index refers to an under-nutrition condition in which the child is stunted as well as wasted, resulting in high mortality rates of children and thus it is relevant to group these two parameters together to identify children at greater risk of mortality make policies to alleviate the morality rates in children under the age of five in India. Finally, the three forms of undernutrition that have been used in this study are - Wasting (WH), Stunting (HA) & Concurrent Stunted-Wasting (HAWH).

Thus, our novel contributions in this paper can be summarized as follows:

- *Our Goal*: We introduce a hybrid index: **Stunted-Wasting** as HAWH, due to its significance as discussed above.
- *Method*: Our study moves past the common Machine Learning Algorithms for constructing a predictive model and leverages open-source libraries and AutoML algorithms to make the process quicker, efficient and satisfactory. The proposed method consistently outperforms prior work by large margins, and we present our compelling results for the same. Intuitively, our approach reduces the need to do laborious data pre-processing, feature learning and model selection tasks, thus allowing the developer to focus on the data rather than model building.
- *Metrics*: Along with classification accuracy as the main metric, we utilize the AUC-ROC metric for determining the performance of our algorithms.

To the best of our knowledge, this is the first study of its kind, and these novel contributions have not been studied in any of the existing literature.

IV. Data

A. *Dataset Details*

Our study utilized two datasets obtained from the Indian Demographic and Health Surveys (DHS) for the years 2005-2006 and 2015-2016. From the DHS datasets, the 'Children Recode' files were examined and used for further analysis. Children Recode files contained questionnaire for the mother for each child born in the month of interview and the 59 months preceding. These questionnaires were broadly related to three subjects -

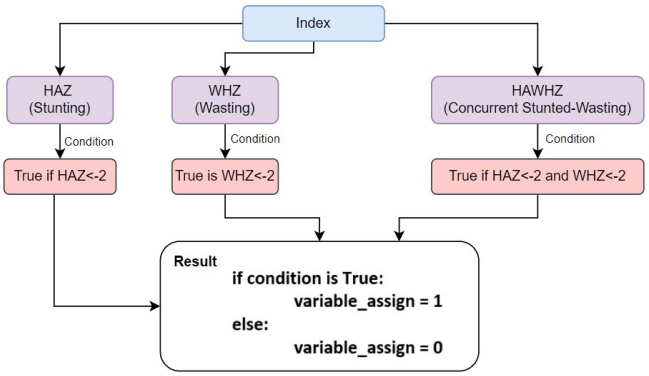
- General household and hygiene characteristics: Area of residence, state, Wealth index, Toilet facility, Source of

drinking water, Father's education Occupation, Mother's education Occupation, etc.
- Maternal characteristics: Age, Body Mass Index, Number of living children, Work status, etc.
- Child characteristics: Age, birth weight, sex, birth order , immunization and other factors such as history of illness and breastfeeding

B. *Data Pre-processing*

These tabular data-sets contained about 24 and 28 features respectively, each corresponding to one of three categories defined in the previous section. The raw data obtained needed extensive pre-processing techniques to get rid of flagged features, null features, and merged features. Additionally, each feature column also needs to be modified and values need to be classified distinctively into one of the many feature sub-classes. To create a binary classification dataset, a threshold of -2 was applied at both indices of Stunting and Wasting. Further the concurrent stunted wasting index was created by a simple 'AND' operation on both independent variables of Stunting and Wasting. Table I gives an overview of the features in each of the two data-sets, the type of feature values and pre-processing technique (if any).

Fig. 1. Anthropometric Index Transformations

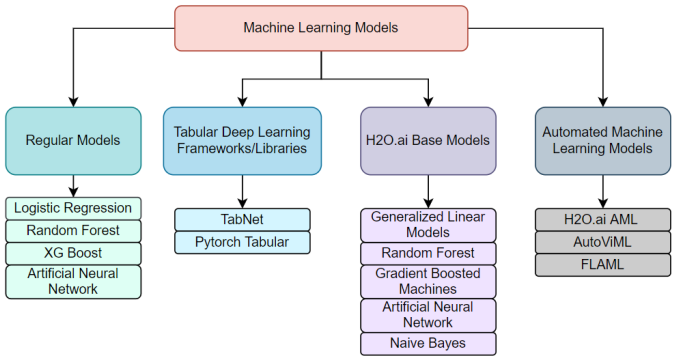


V. Classification Models Implemented

This work considers four categories of Machine Learning models - Regular models, Popular Tabular Deep Learning frameworks (TDL), H2O base models Automated Machine Learning models. The basic model functionalities and key deciding factors for each of them have been discussed in this section. Because existing research utilize three regular machine learning models: Logistic Regression, Random Forest and Artificial Neural Network, these three remain the top choices for the regular base models considered in this study. Additionally, XG-Boost and similar tree algorithms have demonstrated significantly better results on tabular data-sets are also included in the first category of regular models. Popular libraries and frameworks specifically developed for fast and scalable Tabular Deep Learning such as Pytorch Tabular and TabNet form the second category of models considered in our study. The third category of models - H2O base models, utilize regular models, but trained in H2O environments for optimization. Finally, this study utilizes three different Automated Machine Learning algorithms, namely - H2O AML, AutoViML FLAML that form the fourth category. Figure 2 demonstrates the four model categories and their sub-models considered in this study.

1) ***Logistic Regression****:* Logistic Regression is named after the function used at the core of this method- the Logistic Function, also known as the Sigmoid function. The sigmoid function is an S-shaped curve that can map any real-valued number to a value between 0 and 1, but never exactly between those limits. This algorithm thus classifies the dependent variable by predicting its probability and comparing with a threshold value (0.5 by default). It is widely used for binary classification problems since the output is between 0 & 1, each serving as one of the two classes.

Fig. 2. Classification Models implemented in this study



A. *Regular Models*

1) ***Random Forest****:* Random Forest employs a large number of decision trees (operating as ensembles) on different subsets of the available data to attain the best performances. The fundamental concept behind ensembling these decision trees together is that by intuition, we can expect a large number of relatively uncorrelated models (trees) acting as a group will outperform any of the constituent models individually. Thus, all these decision trees cast their vote of prediction for classification of the dependent variable and based on the majority of votes, random forest predicts the final output.

2) ***Artificial Neural Networks****:* An Artificial Neural Network (ANN) is a sequence of an algorithm that attempts at recognizing the underlying relationship in a data set through an iterative process. These networks play a crucial role in deep learning. Neural networks learn by analysing samples that have a known "input" and "output", creating probability-weighted connections between the two, which are stored inside the net's data structure. A neural network is typically trained from a given example by calculating the difference between the network's processed output (often a prediction) and a target output.

TABLE I
DATA DESCRIPTION

Features	Data 2005-06	Data 2015-16	Description
State	Yes	Yes	Different states and Union Territories were divided into the following classes- North, South, East, West, Central, North-eastern India.
Area	Yes	Yes	Whether Urban or Rural
Religion	Yes	Yes	Different religions were divided into the following classes- Hindu, Muslim, Sikh, Christian and others (included the classes-Buddhists, Jains, Jews, and Parsis , no religion) from the original data
Caste	Yes	Yes	Different castes were divided into the following categories - SC, ST, OBC & others (included the class "don't know" and "none of these") from the original data
Type of Family	Yes	**No**	Nuclear, Joint, Three-generation or broken
Father's Education	Yes	**No**	Classified as one of the following: Primary education, Secondary education, and No education.
Father's Occupation	Yes	**No**	Classified as one of the following: Unemployed, Skilled labor, Sales, Services, Clerical & Professional.
Mother's Education	Yes	Yes	Classified as one of the following: Primary education, Secondary education, and No education.
Mother's Occupation	Yes	**No**	Classified as one of the following: Unemployed, Skilled labor, Sales, Services, Clerical & Professional.
Total Family Members	Yes	Yes	Number of members sharing the same kitchen
Mother's Exposure to Mass Media	Yes	Yes	Labelled as yes if the mother had exposure to any one of these (newspapers or magazines, television or radio) at-least once a week
Source of Drinking Water	Yes	Yes	Categorized into Improved sources of water safe for drinking or Unimproved Sources
Toilet Facility	Yes	Yes	Categorized into Improved Toilet facilities and Unimproved Toilet Facilities
Mother's Age	Yes	Yes	Classified in one of the following groups: Less than 19 years, 20-29 years, 30-39 years, greater than 40 years
Mother's BMI	Yes	Yes	Classified in one of the following groups: 18, 18-25, 60
Child's Age	Yes	Yes	Child's age in months was divided into the following categories - Less than 6, 6-12, 12-18, 18-24,24-36,36-48,48-60 months
Sex of Child	Yes	Yes	Male or Female
Birth-weight	Yes	Yes	Birth weight was converted into kilograms and divided into the following categories- less than 2.5, more than 2.5 and not weighed at birth.
Initiation to Breastfeeding	Yes	Yes	Labelled as early if the child was put to Breastfeeding within 1 hour of birth, late in all other cases
Months of Breastfeeding	Yes	Yes	Divided into the following categories: Less than 2 years and greater than 2 years
Wealth Index	Yes	Yes	Categorized into one of the following: Poorest, Poorer, Middle, Richer, and Richest
History of Illness	Yes	Yes	Labelled as yes if the child had any one of the following- diarrhea, fever or cough
Immunization	Yes	Yes	Categories: Fully immunized, partially immunized, and Not Immunized
Index to Birth History	Yes	Yes	Numeric Values
Number of Living Children	Yes	Yes	Numeric Values
Number of Family Members	Yes	Yes	Numeric Values
Sex of Household Head	Yes	Yes	Male or Female
Dietary Score	Yes	Yes	Labelled as adequate dietary score if the child ate at least 4 of these products (Grains, roots and tubers, legumes and nuts ,dairy products (milk, yogurt, cheese), flesh foods (meat, fish, poultry and liver/organ meats), eggs, vitamin-A rich fruits and vegetables)

*3) **XG Boost & Gradient Boosting Machines**:* Boosting is an ensembling process in which new models are introduced to correct the faults generated by existing models. Models are introduced in a logical order until no more advancements are possible. Subsequently, Gradient boosting is a method for creating new models that predict the residuals or mistakes of earlier models, which are then combined to form the final prediction. Gradient boosting derives its name from the fact that it employs a gradient descent approach to minimise loss while adding new models. XG Boost actually stands for **Extreme Gradient Boosting**. Introduced first in [6] is a scalable tree boosting algorithm widely employed for tabular datasets. It is actually an implementation of gradient boosted decision trees designed for speed and performance in particular. Additionally, it offers advantages such as sparse optimization, parallel training, multiple loss functions, regularization, bagging, and early stopping. The XG Boost library publicly available allows the three main forms of gradient boosting.

- Gradient Boosting algorithm also called gradient boosting machine including the learning rate.
- Stochastic Gradient Boosting with sub-sampling at the row, column and column per split levels.
- Regularized Gradient Boosting with both L1 and L2 regularization.

B. Tabular ML frameworks

*1) **TabNet**:* TabNet is a high performance, interpretable canonical deep tabular data learning architecture. The architecture developed by researchers at Google [2], attempts to learn the most salient features from raw input data through sequential attention. Thus, it chooses which features to reason from at every decision step enabling interpretability and better learning capacity. Additionally, TabNet supports two types of interpretabilities: local interpretability, which visualizes the relevance of features and how they interact with one another; and global interpretability, that quantifies each feature's contribution to the trained model.

*2) **Pytorch Tabular**:* Pytorch Tabular is a ready-to-use deep learning library developed by Manu Joseph [8] at Facebook, built on top of gaints such as Pytorch, Pytorch Lightening and Pandas. Till date, Gradient Boosted Decision Trees defeated deep learning models on shallow datasets. Hence, the motivation behind the development of this library was to have a scalable, easy to customize and easier to deploy library for tabular datasets. Thus, making the software engineering part of working with neural networks effortless, and allowing users to focus on the models.

Pytorch Tabular is driven by five configs - *DataConfig*, *ModelConfig*, *TrainerConfig*, *OptimizerConfig* & *ExperimentConfig*. Interested readers are suggested to visit the official documentation*.

C. H2O Base models

H2O is a distributed in-memory machine learning platform with linear scalability that is completely open source. It supports the implementation of a wide range of supervised and unsupervised algorithms for optimized training procedures. Among these modes, this study utilizes five base models trained inside the H2O environment, namely - Generalized Linear Models, Random Forest, Gradient Boosted Machine, Neural Network Model, and Naive Bayes. Out of these five, Random forests, Gradient Boosted Machines and Neural networks are described previously. Thus, this section further briefs about the remaining two algorithms.

*1) **Generalized Linear Models**:* Even when the underlying relationship is not linear, Generalized Linear Models (GLMs) allow us to construct a linear relationship between the response and predictors. This is accomplished by employing a link function, which connects the response variable to a linear model. In contrast to Linear Regression models, the response variable's error distribution does not have to be regularly distributed. The mistakes in the response variable are expected to be distributed in an exponential family (i.e., normal, binomial, Poisson, or gamma distributions). Because we are attempting to generalize a linear regression model that can also be used in various instances, the term Generalized Linear Models was coined.

*2) **Naive Bayes**:* Naive Bayes (NB) is actually a family of probabilistic classification algorithms sharing the same principal of Bayes Theorem. This theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge. This classifier assumes that each feature makes an independent and equal contribution towards the output prediction and is highly scalable. The training procedure consists of calculation and storage of class probabilities and conditional probabilities for different inputs. Hence the training time is significantly reduced since no coefficients need to be fitted for optimization procedures.

D. Auto Machine Learning Platforms

The process of automating model selection, feature generation, hyper-parameter tuning, iterative modelling, and model assessment is known as Auto Machine Learning (AutoML). AutoML simplifies the training and evaluation of machine learning models by automating repetitive operations, allowing individuals to concentrate on the data and the challenges they are attempting to address. The goal of an AutoML algorithm can be summarized as - Given a training dataset and an error metric, use low computational cost to search for learner and hyperparameter choices and produce models optimizing the error metric in short time.

*1) **H2O AutoML Platform**:* H2O AutoML is a highly scalable automated machine learning algorithm included inside the H2O framework that supports supervised training of regression, binary classification, and multi-class classification models on tabular datasets. Specifically, we utilize H2O's Stacked Ensemble method, which is a supervised ensemble machine learning algorithm that finds the optimal combination of a collection of prediction algorithms using a process called stacking [10] depending on the input base models. Thus, the algorithm that learns the optimal combination of these base learners is called the meta-learning algorithm or meta-learner.

*https://pytorch-tabular.readthedocs.io/en/latest/

In our study, we have used the H2O base models individually with different as well as learned a meta-learner through a stacked ensemble method.

*2) **AutoViML**:* Automatically Build Variant Interpretable Machine Learning models (AutoViML) [*] is a library primarily built for interpretabilty and high performance on large datasets. Its variability is attributed to the technique of experimenting with multiple models and multiple features on a particular dataset till it finds the best performing model. Similarly, it enforces interpretability through the library SHAP [*] (Shapley Additive Explanations) by selecting the least number of features necessary to build a simple model.

*3) **FLAML**:* A Fast and Lightweight AutoML library, FLAML [17] is the most recent AutoML platform in our study that is designed to perform efficiently and robustly without relying on meta-learning or ensemble at first order, for several usability reasons. FLAML allows a user to plug in the library in new applications without requiring the user to collect diverse meta-training datasets before being able to use it. Additionally, a user can easily customize learners, search spaces and optimization metrics and use FLAML directly after customization.

VI. Evaluation Metrics

*1) **Accuracy**:* Accuracy is the base metric for evaluation of classification algorithms. It is actually the ratio of correct predictions of the model to the total number of predictions made. Generally, accuracy is a reasonable evaluation metric when dealing with a class balanced dataset.

$$Accuracy = \frac{Number of Correct Predictions}{Total Number of Predictions made} \quad (2)$$

When dealing with binary classification, we can also represent accuracy in terms of positives and negatives. Intuitively, each classification can be distinguished into one of the four categories -

- **T**rue **P**ositives: Outcome when the model correctly predicts the positive class.
- **F**alse **P**ositives: Outcome when the model gives incorrect prediction about the negative class to be positive.
- **T**rue **P**Negatives: Outcome when the model correctly predicts the negative class.
- **F**alse **N**egatives: Outcome when the model gives incorrect prediction about the postive class to be negative.

Thus, accuracy in terms of positives and negatives is described by -

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

[*]https://github.com/AutoViML
[*]https://github.com/slundberg/shap

*2) **AUC-ROC Curve**:* Area Under the Curve of Receiver Operator Characteristic, is a popular method of visualizing the performance of our machine learning algorithm and is mainly used in binary classification problems. The Receiver Operator Characteristic (ROC) is a probability curve that plots the True Positive Rates (TPRs) against False Positive Rates (FPRs) at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. When 0.5 <AUC <1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives. Since AUC-ROC is not the best metric in case of an imbalanced data-set, we have oversampled our original data-sets to overcome the imbalance of binary classes.

VII. Experiments & Results

We evaluate the metrics for both our data-set through the 4 different machine learning approaches across 14 different models. Our intention behind experimenting with different models was not to compare these models, but to understand which model(s) have the best performance on our data-sets. Tables II to VII demonstrate Accuracy scores and AUC-ROC scores for all three anthropometric indices considered in this study. Values in bold signify to the top scores achieved.

A. *Results*

- **Stunting**: Top accuracy and AUC-ROC scores for the index HA are achieved on the 2015-16 DHS datasets. With the TabNet architecture for interpretable tabular deep learning, this study was able to achieve as high as **74.53%** accuracy and **82.55%** AUC-ROC scores as shown in Table II and Table III.
- **Wasting**: Unlike Stunting, top accuracy and AUC-ROC scores for the index WH are achieved on different datasets. While the highest accuracy score of **86.30%** is achieved With the TabNet architecture for interpretable tabular deep learning as shown in Table IV, an AUC-ROC score of **99.95%** was achieved with the random forest base model in the H2O environment (Table V).
- **Stunted-Wasting**: The highest scores follow a similar trend as in the previous two cases for the concurrent stunted wasting index. Models TabNet achieves the highest accuracy score of **96.46%**, while Random Forest base model in the H2O environment achieves the highest AUC-ROC score of **98.53%** as shown in Tables VI and VII.

Noticeably, the TabNet architecture and the Random Forest H2O base model provide compelling results on all three anthropometric indices considered in this study. Similarly, most algorithms considered in this study give reasonable

performances such as GBM, GLM and ANN in H2O environment. However, Naive Bayes consistently gives lesser predictive performances as compared to other models for both metrics in nearly all indices. Since Naive Bayes makes an assumption that each feature is independent, the results on NB are now an indicator that several features considered in this study must be correlated. Alongside, this study demonstrates significant contribution of a AutoML algorithms. While H2O AML and FLAML are also among the top 3 performing models, AutoViML's interpretative functionality also enabled us to generate feature importance plots for all three metrics. The plots are described in Figure 3, 4 & 5 for HA, WH HAWH respectively.

TABLE II
ACCURACY RESULTS ON HA (STUNTING)

Algorithm	Type	2005-06	2015-16
Logistic Regression	Regular	64.82	64.20
Random Forest	Regular	64.98	**66.29**
ANN	Regular	62.99	54.76
XG-Boost	Regular	65.11	65.11
TabNet	TDL	64.25	**74.53**
Pytorch Tabular	TDL	64.42	66.33
GLM	H2O base model	58.49	56.87
Random Forest	H2O base model	62.97	66.21
GBM	H2O base model	60.34	63.00
ANN	H2O base model	60.10	58.69
NB	H2O base model	55.90	57.60
H2O AML	AutoML	61.23	**70.63**
AutoViML	AutoML	65.00	65.00
FLAML	AutoML	65.39	65.98

TABLE III
AUC-ROC RESULTS ON HA (STUNTING)

Algorithm	Type	2005-06	2015-16
Logistic Regression	Regular	64.47	64.05
Random Forest	Regular	64.52	66.15
ANN	Regular	63.6	53.4
XGBoost	Regular	69.30	71.50
TabNet	TDL	67.79	**82.55**
Pytorch Tabular	TDL	65.00	67.00
GLM	H2O base model	66.91	67.91
Random Forest	H2O base model	70.38	**76.11**
GBM	H2O base model	69.84	**71.77**
ANN	H2O base model	68.12	67.55
NB	H2O base model	64.85	66.53
H2O AML	AutoML	70.63	70.63
AutoViML	AutoML	71.00	70.00
FLAML	AutoML	65.10	66.00

TABLE IV
ACCURACY RESULTS ON WH (WASTING)

Algorithm	Type	2005-06	2015-16
Logistic Regression	Regular	71.99	65.32
Random Forest	Regular	80.10	72.81
ANN	Regular	72.16	66.53
XG-Boost	Regular	72.9	66.53
TabNet	TDL	72.41	**86.30**
Pytorch Tabular	TDL	71.32	67.10
GLM	H2O base model	68.46	58.79
Random Forest	H2O base model	82.52	74.01
GBM	H2O base model	76.13	65.43
ANN	H2O base model	67.00	55.37
NB	H2O base model	64.23	58.26
H2O AML	AutoML	**84.10**	84.10
AutoViML	AutoML	66.00	73.00
FLAML	AutoML	**83.29**	70.44

Results also demonstrate the significantly higher accuracy and AUC-ROC scores of the Concurrent Stunted-Wasting index as compared to the two variables considered independently. Thus, Stunted-Wasting is the highest performative anthropoemtric index in predicting malnutrition in under-five children. It is then also found that Stunting as an independent variable is the least performative in predicting malnutrition.

TABLE V
AUC-ROC RESULTS ON WH (WASTING)

Algorithm	Type	2005-06	2015-16
Logistic Regression	Regular	71.98	65.35
Random Forest	Regular	80.10	72.82
ANN	Regular	72.1	66.60
XG-Boost	Regular	77.9	72.3
TabNet	TDL	79.93	89.01
Pytorch Tabular	TDL	73.00	67.00
GLM	H2O base model	76.82	68.70
Random Forest	H2O base model	**99.95**	**97.12**
GBM	H2O base model	84.34	74.94
ANN	H2O base model	75.59	64.58
NB	H2O base model	73.69	67.48
H2O AML	AutoML	**92.05**	**92.05**
AutoViML	AutoML	75.00	85.00
FLAML	AutoML	84.60	70.7

TABLE VI
ACCURACY RESULTS ON HAWH (STUNTED-WASTING)

Algorithm	Type	2005-06	2015-16
Logistic Regression	Regular	78.48	72.79
Random Forest	Regular	91.02	89.10
ANN	Regular	81.37	77.27
XG-Boost	Regular	79.03	74.57
TabNet	TDL	81.22	**96.46**
Pytorch Tabular	TDL	78.44	74.21
GLM	H2O base model	74.08	70.20
Random Forest	H2O base model	92.80	89.88
GBM	H2O base model	86.12	77.40
ANN	H2O base model	86.53	75.96
NB	H2O base model	69.5	66.20
H2O AML	AutoML	**94.69**	85.24
AutoViML	AutoML	74.00	81.00
FLAML	AutoML	**93.49**	92.53

TABLE VII
AUC-ROC RESULTS ON HAWH (STUNTED-WASTING)

Algorithm	Type	2005-06	2015-16
Logistic Regression	Regular	78.48	72.79
Random Forest	Regular	91.04	89.10
ANN	Regular	81.50	77.3
XG-Boost	Regular	84.20	84.80
TabNet	TDL	89.18	**97.19**
Pytorch Tabular	TDL	77.00	74.00
GLM	H2O base model	83.05	78.56
Random Forest	H2O base model	**97.81**	95.70
GBM	H2O base model	93.47	85.66
ANN	H2O base model	93.50	82.83
NB	H2O base model	79.29	76.44
H2O AML	AutoML	**98.53**	92.82
AutoViML	AutoML	86.00	94.00
FLAML	AutoML	94.00	93.50

VIII. CONCLUSION

This research presents evidence for combining wasting and stunting as a hybrid index of Stunted-Wasting, that demonstrates significantly better results than stunting and wasting considered individually.Even though Machine Learning algorithms are currently in practice, by utilizing open-source tabular deep learning libraries and AutoML algorithms, we attempted to fasten up the process while also making it much more efficient. Ultimately, we also demonstrate results of the AUC-ROC scores to give a complete idea of the performances of these models. We conclude that the best performances are achieved on TabNet, an interpretable Tabular Deep Learning framework, followed by H2O Random Forest base model and AutoML algorithms. Some of the notably important features present in our dataset found by the feature importance plots in Figures 3,4 & 5 are - Child's age, Mother's BMI, Mother's age, Toilet Facility, Wealth Index, Birth weight, Mother's Education, Sex of Child, State, Religion, Caste, Exposure to Mass Media and Immunization. With this background, we can now predict a child's health beforehand given information about important health, family, and personal factors.

Fig. 3. Feature Importances for predicting HA

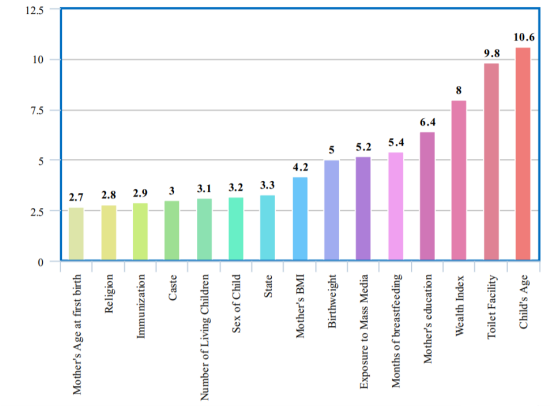


Fig. 4. Feature Importances for predicting WH

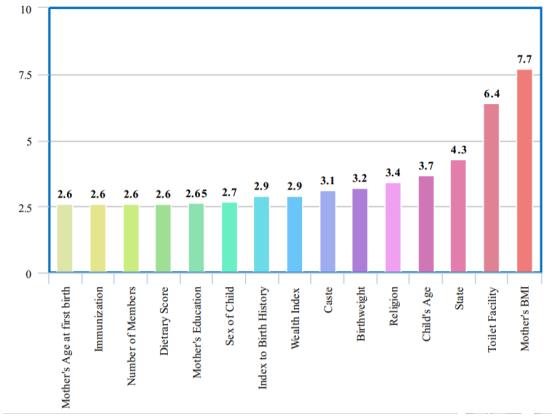
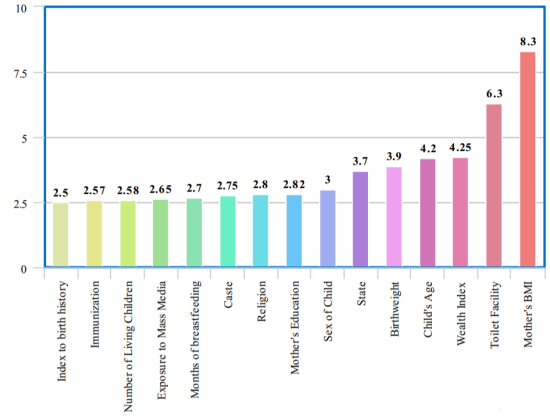


Fig. 5. Feature Importances for predicting HAWH



REFERENCES

[1] Fact sheets - malnutrition.

[2] Sercan O Arık and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *arXiv*, 2020.

[3] André Briend, Michel Garenne, Bernard Maire, Olivier Fontaine, and K Dieng. Nutritional status, age and survival: the muscle mass hypothesis. *European Journal of Clinical Nutrition*, 43(10):715–726, 1989.

[4] André Briend, Tanya Khara, and Carmel Dolan. Wasting and stunting—similarities and differences: policy and programmatic implications. *Food and nutrition bulletin*, 36(1_suppl1):S15–S23, 2015.

[5] Published by M. Szmigiera and Jul 30. Malnutrition: share of people by region 2020, Jul 2021.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] Michele Gragnolati et al. *India's undernourished children: a call for reform and action*. World Bank Publications, 2006.

[8] Manu Joseph. Pytorch tabular: A framework for deep learning with tabular data. *arXiv preprint arXiv:2104.13638*, 2021.

[9] Sangita Khare, S Kavyashree, Deepa Gupta, and Amalendu Jyotishi. Investigation of nutritional status of children based on machine learning techniques using indian demographic and health survey data. *Procedia computer science*, 115:338–349, 2017.

[10] Erin LeDell and Sebastien Poirier. H2o automl: Scalable automatic machine learning. In *Proceedings of the AutoML Workshop at ICML*, volume 2020, 2020.

[11] Nega Mulu, Bekrie Mohammed, Haile Woldie, and Kegnie Shitu. Determinants of stunting and wasting among street children in northwest ethiopia: A community-based study. *Nutrition*, page 111532, 2021.

[12] World Health Organization et al. Levels and trends in child malnutrition: Unicef/who/the world bank group joint child malnutrition estimates: key findings of the 2021 edition. In *Levels and trends in child malnutrition: UNICEF/WHO/The World Bank Group joint child malnutrition estimates: key findings of the 2021 edition*. 2021.

[13] Orit Raphaeli and Pierre Singer. Towards personalized nutritional treatment for malnutrition using machine learning-based screening tools, 2021.

[14] Swaroop Kumar Sahu, S Ganesh Kumar, B Vishnu Bhat, KC Premarajan, Sonali Sarkar, Gautam Roy, and Nitin Joseph. Malnutrition among under-five children in india and strategies for control. *Journal of natural science, biology, and medicine*, 6(1):18, 2015.

[15] Abhishek Singh. Childhood malnutrition in india. *Perspective of Recent Advances in Acute Diarrhea*, 2020.

[16] Ashis Talukder and Benojir Ahammed. Machine learning algorithms for predicting malnutrition among under-five children in bangladesh. *Nutrition*, 78:110861, 2020.

[17] Chi Wang, Qingyun Wu, Markus Weimer, and Erkang Zhu. Flaml: A fast and lightweight automl library. In A. Smola, A. Dimakis, and I. Stoica, editors, *Proceedings of Machine Learning and Systems*, volume 3, pages 434–447, 2021.