BMC Medical Informatics and Decision Making

RESEARCH

Open Access

A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones

Haile Mekonnen Fenta[1*], Temesgen Zewotir[2] and Essey Kebede Muluneh[3]

Abstract

Background: Undernutrition is the main cause of child death in developing countries. This paper aimed to explore the efficacy of machine learning (ML) approaches in predicting under-five undernutrition in Ethiopian administrative zones and to identify the most important predictors.

Method: The study employed ML techniques using retrospective cross-sectional survey data from Ethiopia, a national-representative data collected in the year (2000, 2005, 2011, and 2016). We explored six commonly used ML algorithms; Logistic regression, Least Absolute Shrinkage and Selection Operator (L-1 regularization logistic regression), L-2 regularization (Ridge), Elastic net, neural network, and random forest (RF). Sensitivity, specificity, accuracy, and area under the curve were used to evaluate the performance of those models.

Results: Based on different performance evaluations, the RF algorithm was selected as the best ML model. In the order of importance; urban–rural settlement, literacy rate of parents, and place of residence were the major determinants of disparities of nutritional status for under-five children among Ethiopian administrative zones.

Conclusion: Our results showed that the considered machine learning classification algorithms can effectively predict the under-five undernutrition status in Ethiopian administrative zones. Persistent under-five undernutrition status was found in the northern part of Ethiopia. The identification of such high-risk zones could provide useful information to decision-makers trying to reduce child undernutrition.

Keywords: Composite index for anthropometric failure (CIAF), Confusion matrix, Covariate selection and ranking, Multicollinearity, Receiver operating characteristics (ROC)

Background

Proper nutrition is so crucial to lead a healthy lifestyle. Malnutrition, particularly undernutrition, is a global concern for the health condition and survival of children [1–5]. Almost half of the deaths of children in developing countries were directly or indirectly linked to malnutrition [3, 6]. Malnourished children are more vulnerable to different illnesses compared to their counterparts [1–6]. A considerable number of studies investigating the issue targeting under-five children malnutrition and the risk factors associated with this age group. These studies employed classical models such as generalized linear (mixed) models [4, 5, 7–10]. The finding from the investigations, among others, showed that the nutritional status of children of this age group has gradually improved over

*Correspondence: hailemekonnen@gmail.com
[1] Department of Statistics, College of Science, Bahir Dar University, Bahir Dar, Ethiopia
Full list of author information is available at the end of the article

the last 2 decades in Ethiopia. Particularly, it has been found that the prevalence of under-five children underweight in Ethiopia was 47.1% in 2000, 38.5% in 2005, 28.8% in 2011, 23.3 in 2016, and 20.56% in 2019, while the prevalence of stunting was 51.22% in 2000, 46.5% in 2005, 44.3% in 2011, 38.3% in 2016, and 36.9% in 2019. Similarly, 10.7% of under-five children were wasted in 2000, 10.5% in 2005, 9.9% in 2011, 10.1% in 2016, and 7% in 2019. The prevalence of having at least one of the undernutrition indicators measured in terms of the composite index for anthropometric failure (CIAF) was 61.38% in 2000, 56.58% in 2005, 51.58% in 2011, 46.49% in 2016, and 42.4 in 2019. Moreover, the CIAF is computed by grouping different forms of anthropometric failure as such: B-wasting only, C-wasting and underweight, D-wasting, stunting and underweight, E-stunting and underweight, F-stunting only, and Y-underweight only. The CIAF, calculated by aggregating these six (B–Y) categories [11–15]. Most of such studies conducted in this country depicted the effects of socio-economic and demographic covariates that were associated with under-five children undernutrition status using the classical regression models [4, 5, 7, 8]. Those traditional models are widely used for causal inferences and with the selection of built-in features, with a relatively small number of covariates [16, 17]. Correlations between covariates (multicollinearity) and a large number of factors are the common analytical challenges in traditional modeling [18–21]. Moreover, as compared to those classical models, the machine learning (ML) methods have the qualities of using a larger number of predictors, requiring fewer assumptions, incorporating "multi-dimensional correlations", and producing a more flexible relationship among the predictor variables and the outcome variables [16–18, 20–22]. In addition, the ML models can create models for prediction purposes that show superiority in taking care of classification problems when compared with the classical approaches [16–18, 21, 23]. In the present paper, we focused to predict CIAF in Ethiopia using this tool drawing on the nationally representative data. Machine learning employs methods developed within the disciplines of statistics, computer sciences, mathematics, and artificial intelligence which allow the formation of algorithms that can learn from and make predictions using data [24–29]. As such, it is applicable in different disciplines, such as in medical sciences; for diagnosis and outcome prediction [23, 30–44], disease modeling [33], disease prediction [34–37], child mortality [23, 38], and it is also used in industrial applications [39–41]. Just only a few studies had investigated the role of this tool to create prediction models of childhood for malnutrition [42–44]. Moreover, the study is conducted at the administrative zones in Ethiopia. This is because, in the country, the zonal health departments have the mandate to plan, follow up, monitor, and evaluate health activities of Woreda health offices and the different Woredas in the same Zone are relatively similar in many respects. Moreover, the administrative Zones are mainly ethnic-based, and the assessment of the Zones provides cultural practices regarding staple food and the geographic environment of the community in the Zones [45–48]. Hence detecting the problems of undernutrition and its variations among administrative Zones provides deeper insight into the health priorities which helps policymakers to design focused intervention strategies. The main objective of this study was, therefore, to identify ML algorithms in predicting and identifying the important covariates that underline the spatial variations in childhood CIAF among 72 Ethiopian administrative zones.

Materials and methods

This study was carried out on the disparities of malnutrition in Ethiopia, with a surface area of 1.1 million km^2, the country shares borders with Eritrea in the north, Djibouti and Somali in the east, Sudan and South Sudan in the west, and Kenya in the south. It is divided into 11 administrative units (regions) including Addis Ababa, the capital city of the country. The regions were further divided into 72 second-level administrative boundaries called zones [49] (Fig. 1).

Data sources and analysis tools

We conducted the analysis based on the four EDHS datasets (2000, 2005, 2011, and 2016), a nationally representative household survey developed by the United States Agency for International Development (USAID) in the 1980s [50]. The outcome variable that we aimed to predict is the undernutrition status of under-five children measured in terms of the composite index for anthropometric failure (CIAF). CIAF is measured as a binary response as being nourished (coded as 0) and undernourished (coded as 1). The covariates (features) were collected from different pieces of literature [4, 5, 7–10]. All the categorical features are converted to numerical dummy variables, by mapping each unique value to a number [4, 5, 7–10]. The boundaries (shapes) were used to define the second-level administrative zones and merged with the real dataset for analysis [51].

Methodology

Model building The ML models have shown superiority in taking care of classification problems when compared with the traditional models (like generalized linear mixed models). The raw data are usually not found in the form and shape that is required for optimal performance of the machine learning algorithms. The algorithms that

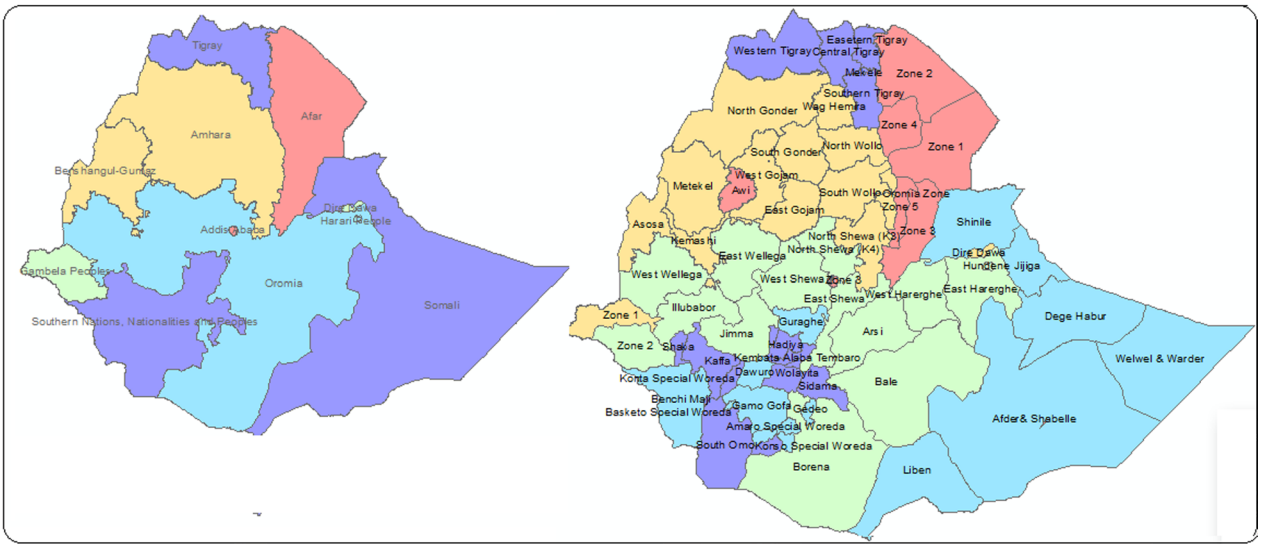

Fig. 1 Map of Ethiopia with regions and zones: **A** the 11 regions of the country and **B** the 72 administrative zones

would be implemented in ML are only numerical values and therefore it is important to transform the categorical variables into numerical values. Hence, the preprocessing step is the most important aspect in the ML model applications [21, 23, 52–54]. The categorical features of the dataset are encoded to transform these features into numerical values and the continuous data in this study were normalized. For ML approaches, the dataset is randomly split into two: a training dataset which trains the model, and a test dataset where we predict the response variable and check whether the predicted outcome is similar to the actual outcomes, and the validation dataset is considered for the parameter estimates to be incorporated in the training models [24–29]. Influence of different training and testing ratios on the performance of the given ML models were checked. This study (train/test: 80/20, and 70/30) was implemented to divide the datasets into the training and testing datasets for performance assessment of models. Popular statistical indicators have been employed to evaluate the predictive capability of the models under different training and testing ratios. The results revealed that the train-test 70–30% split were more advantageous to undernutrition classification than their counterparts (80/20). A variety of supervised ML algorithms including Logistic Regression (LR) [55], Ridge regression [56], Least Absolute Shrinkage and Selection Operator (LASSO) regression [57], Elastic Net [27, 58], Artificial Neural Network (ANN) [59, 60] and Random Forest (RF) [27, 61] were included in the analysis.

The Ridge, Lasso, and Elastic Net are very similar to LR, except that we have an additional penalty term called regularization to estimate the regression coefficients [26, 27] to reduce the over-fitting and the adverse effects of multicollinearity [26–28, 62]. The advantage of ridge, lasso and elastic net modeling over the classical statistical methods is that, in addition to fitting optimized models, a penalty is applied to predictors in the model, causing covariates with little impact on the outcome variable to be minimized or dropped from the final model. This reduces the model's complexity while increasing its generalizability.

Logistic regression (LR) LR is a widely applied statistical model for classification problems. This model applies the maximum likelihood estimation procedure to estimate the parameter of interest. Let y_i be the response variable for the ith child, and it is Bernoulli distributed and takes on the value 1 with a probability of $\pi_i = P(y_i = 1|\boldsymbol{x}_i)$, where $\boldsymbol{x}_i = \left(x_1, ..., x_p\right)^T$ is the ith child's covariate vector, and value 0 with probability $1-\pi_i$. Then the logistic regression model with the logit link function can be given as:

$$\pi_i = \frac{\exp\left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)}{1 + \exp\left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)} \tag{1}$$

where β_0 is the intercept term, and $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_p\right)^T$ is a $p \times 1$ vector of estimated regression parameters on the logit scale. If parameter $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta})^T$, then the corresponding log-likelihood function is given by the following equation as it was also shown by [55]:

$$l_{\boldsymbol{\theta}} = \sum_{i=1}^{n} \left[y_i \log\left(\pi_i\right) + \left(1 - y_i\right) \log\left(1 - \pi_i\right)\right] \tag{2}$$

By replacing π_i from Eq. 1 in Eq. 2, we have:

$$l_{\theta} = \sum_{i=1}^{n} \left[y_i \left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} \right) - \log \left(1 + \exp(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} \right)) \right] \tag{3}$$

In the maximum likelihood method, the goal is finding a set of $\boldsymbol{\theta}$ that can maximize Eq. (3). When we have a large number of features (dimensionality), the traditional LR has a few problems: over-fitting, multicollinearity, and computational difficulties. To address this problem, we used regularization which is a GLM that imposes a penalty on the parameters to shrink towards zero [27, 55–58, 63].

The ridge regression (L2 regularization, which shrinks coefficients of correlated covariates towards each other) is obtained by maximizing the function with a penalized parameter λ applied for all the parameters except the constant (intercept) [55, 56]. The penalized likelihood formulation for ridge regression is given by (4)

$$l_{\lambda}^{\mathrm{R}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) - \log \left(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta} \right)) \right] - \lambda \sum_{j=1}^{p} \boldsymbol{\beta}_j^2 \tag{4}$$

When the λ values are too large ($\lambda \to \infty$), the coefficients of all the parameters tend to be zero, but when $\lambda = 0$, the ridge regression is equal to the traditional approach.

The LASSO regression uses the L-1 penalty for variable selection and shrinkage. As such, if the λ is large enough, it forces the coefficient to be zero which provides a lesser number of predictors [57]. The function for the lasso regression is given by (5)

$$l_{\lambda}^{\mathrm{L}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) - \log \left(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta} \right)) \right] - \lambda \sum_{j=1}^{p} \left| \boldsymbol{\beta}_j \right| \tag{5}$$

The term λ allows the lasso model to carry out much iteration for a given function and find the optimum values for all coefficients. The optimal regularization parameter (λ) was determined using the nfold cross-validation techniques. The smaller the λ value, the more the effect of regularization upon the number of covariates (features) in the model and their respective coefficients [26–28]. Thus, variables with non-zero estimates are considered the important covariates for the outcome variable of interest.

The elastic net regularization is a combination of both (3) and (4) penalties [27, 58]. This method can effectively control the group of correlated features and also shrink the coefficients of non-informative features to zero [27, 58, 63, 64]. The elastic net regression is given by (5)

$$l_{\alpha}^{\mathrm{El}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) - \log \left(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta} \right)) \right] + \alpha \sum_{j=1}^{p} \boldsymbol{\beta}_j^2 + (1 - \alpha) \sum_{j=1}^{p} \left| \boldsymbol{\beta}_j \right| \tag{5}$$

All the ML algorithms including the logistic regression were performed with R statistical software R and the packages glmnet, pROC, caret, random forest, ggplot, and ROCit were included in the analysis [65–69]. In this paper, we trained the generalized linear model (GLM) estimators with common α values from the set {0, 0.5, 1}, where ($\alpha = 0.0$, 0.5 and 1.0 respectively refers to the ridge, elastic net and lasso penalty) [27, 58, 63].

The Random forest (RF) is the popular supervised ML approach in applied statistics because of its applicability in both classification and regression [70–72]. It is also used for variable screening for dimension reduction. It is a "tree-based" technique in which several decision trees are constructed from a random set of covariates and used to predict an outcome label for a subset of samples. It builds multiple trees (called the forest) and the decision is based on the majority votes over all the trees in the forest [70–73].

The Neural Network (NN) is a type of ML algorithm that is made up of layers of nodes, the most important of which are an input layer [74], hidden layers, and output layers. It is set up with several input neurons (X) that represent the information extracted from each feature in the dataset. Back-propagation is a process used in recurrent NN in which prediction errors are fed back through the NN before modifying the weights of each neural connection until the error level is minimized [59, 60].

$$y = activation\left(\sum (weight + input \right) + bias)$$

Model evaluation

Model performance The performances of the given ML models are evaluated using different model performance approaches including sensitivity, specificity, and accuracy [24–29, 75] which are calculated using the observed data as the gold standard. The model sensitivity and specificity relationship are expressed using the Receiver operating characteristics (ROC) curves (Fig. 2).

All the curves which are plotted to the left of the diagonal line are performing better than chance. The area under each curve (AUC) gives an aggregated value which explains the probability that a random sample

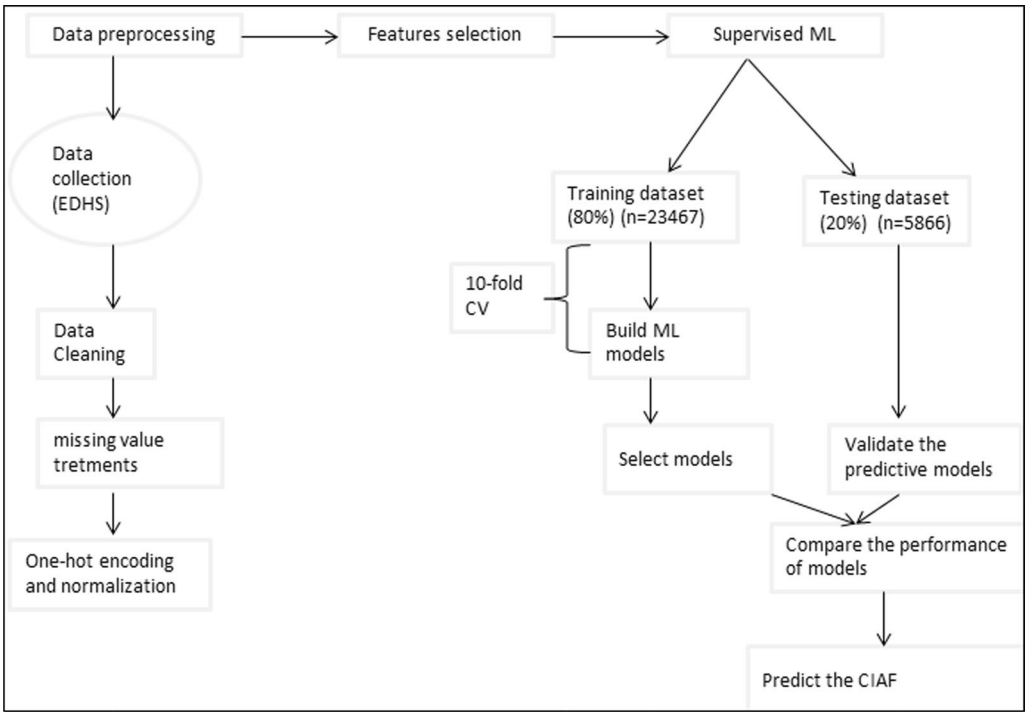

Fig. 2 Overview flow chart of the methodologies used

would be correctly classified by each of the ML algorithms [25, 76]. The AUC of the ROC curve averaged over 10 cross-validation folds (ten repeats) [25], which partitions the original sample into ten disjoint subsets, uses nine of those subsets in the training process, and then makes predictions about the remaining subset. Then the identified best-fit model is used to predict the undernutrition in another dataset, known as the test dataset [24–29].

Covariate selection and ranking Covariate selection is very important for prediction and interpretations, especially for high-dimensional datasets. To assess the importance of predictors in the selected model, the study employed two important measures; Mean Decreases Accuracy (MDA) and Mean Decrease Gini (MDG). The highest decrease in the accuracy and Gini values of the model implies the best predictive and the most important variable respectively [77] for the successful classifications (Table 1).

Results

This analysis consisted of data from 29,333 children of age 0–59 months. Of these, 15,281 (52.09%) had at least one form of the undernutrition indicators (stunting, wasting, and underweight) measured in terms of CIAF. We examined the prevalence of CIAF of U5C experience across different child and mother-household level covariates. The prevalence of CIAF was more common among parents with no formal education compared to parents with secondary and post-secondary levels of educations. Most of the undernourished children were from rural areas. Also, the prevalence of undernourished children was reported from the lower wealth index of households, from mothers having no media exposure, from unimproved toilets and sanitation compared with their counterparts. Covariates that were significant in the Chi-square statistics were used to develop the ML algorithms on the training dataset (Table 2).

Figures in the supplementary documents indicated the effects of different levels of the log of the regularization parameter (λ) for the ridge, elastic net, and lasso regression using the dotted vertical lines (here at $x=-4.51$, $x=-7.84$, and $x=-8.71$) respectively, which indicates the accuracy of the prediction maximization. The coefficients for the given model features were indicated for different values of log (λ) that minimizes a mean squared error (MSE) of coefficients established during the cross-validation. The graph shows that as the log (λ) value decreases, the number of the variables included in the model (those with nonzero coefficients) increases (Additional file 1).

Performance comparisons The accuracy and AUC were implemented to evaluate the efficiency of ML algorithms.

Table 1 The description of the response variable and the respective covariates included in the model

	Descriptions
Childhood undernutrition using CIAF (outcome variable)	$yi = \begin{cases} 1: \textit{ if a child i had at least one form of undernutrion (CIAF)} \\ 0: \textit{ if child i is nourished} \end{cases}$
Children level covariates	
Sex	Sex of a child (female vs male)
Age (months)	Age of a child in months
Vitamin A (VA)	Yes/no
Birth order (BO)	1, 2–3 4 + (birth order number)
Breastfeeding (BF)	Yes/no
Child comorbidity status (CO)	(Presence of diarrhea, fever, ALRI in last 2 weeks before the survey): (No vs yes)
Types of birth (TB)	(Multiple vs singleton)
Size of the child at birth (SC)	(Smaller than average, average, larger than average)
Dietary diversity score (DDS)	Below minimum requirement/ minimum requirement
Maternal/household-level covariates	
Mother's age	15–24, 25–34 and 35–49 (respondents current age 15–49)
Residence (PR)	Rural/urban
Mother's educational level (ME)	No formal education, primary and secondary and above
Father's educational level (FE)	No formal education, primary and secondary and above
Women's autonomy tertiles (WA)	Low, medium, and high
Toilet facility (TF)	Improved and unimproved
Source of drinking water (SDW)	Improved and unimproved
BMI	Body mass index of mothers (< 18.5 kg/m^2, 18.5–24.9 kg/m^2 and ≥ 25 kg/m^2)
Number of children under five (NUFC)	Number of children under the age of 5 (0–1, 2, 3 or more)
Survey year (SY)	Years of the survey (2000, 2005, 2011, and 2016)
Media exposure	Yes/no
Working status of the mother (WS)	Not working/working
Household size	< 4, 5–9, 10 + (continuous)
Wealth quantile (WQ)	Poorest, poor, middle, richer, and richest
Geospatial covariates	
Precipitation (precp)	The average precipitation measured within the 10 km (rural) or 2 km (urban)
Aridity index	The ratio of annual precipitation to annual potential evapotranspiration (10 km × 10 km)
Evaporation	Global elevation above earth's sea level
Maximum temperature (MaxT)	The average annual maximum temperature within the 10 km (rural) or the 2 km (urban)
Minimum temperature (MinT)	The average annual minimum temperature within the 10 km (rural) or the 2 km (urban)
Potential evaporation (pet)	The average annual pet within the 10 km (rural) or the 2 km (urban)
Proximity to water (proxtmty)	Straight-line distance to the nearest major water body
Urban–rural settlement (UR)	This is the urban–rural population classification of the area within the 10 km (rural) or the 2 km (urban)
Population density (popD)	Estimates of human population density is the number of persons/km^2
Enhanced vegetation index (EVI)	The average vegetation index value within the 10 km (rural) or the 2 km (urban)
Cluster altitude (Alt)	Cluster altitude
Wet days (WetD)	The average number of days receiving rainfall within the 10 km (rural) or 2 km (urban)

The comparison of the efficiency of ML algorithms with the traditional LR was depicted in Fig. 3 and Table 3. All the ML algorithms considered in this study perform better than those of the classical logistic regression model to predict the undernutrition status. More detail is given in the Additional file 1.

A comparison of 70% training and 30% validation, 80% training and 20% validation was performed respectively to examine the six models' behaviors with some statistical measures and area under the receiver operating characteristic curve. Although all the models with the two train-test splits ratio had almost identical performances evaluation metrics, the 70–30% split was chosen as the most

Table 2 Sample characteristics (n = 29,333)

Variables	Categories	Nourished (%)	CIAF (%)	X^2 test statistic	*p* values
Sex of child	Male	44.74	55.26	29.96	< 0.001
	Female	47.93	52.07		
Age of a child (months)	< 23	55.40	44.60	652.83	< 0.001
	24–59	40.25	59.78		
Vitamin A	Yes	43.75	56.25	75.02	< 0.001
	No	48.79	51.21		
Birth order	1	50.36	49.64	67.89	< 0.001
	2–3	47.55	52.45		
	4 +	44.17	55.83		
Breastfeeding	Yes	46.21	53.79	0.14	0.707
	No	46.59	53.41		
Comorbidity	Yes	43.17	56.83	58.25	< 0.001
	No	47.88	52.12		
Size of the child at birth	Smaller than average	46.30	53.7	268.357	< 0.001
	Average	48.06	51.94		
	Larger than average	50.9	49.10		
Dietary diversity score (DDS)	Below minimum	46.56	53.44	2.432	0.349
	Minimum	45.51	54.49		
Types of birth	Singleton	46.62	53.38		
	Multiple	29.17	70.83		
Mother's age	15–24	48.51	51.49	31.06	< 0.001
	25–34	46.46	53.54		
	35–49	43.91	56.09		
Place of residence	Rural	44.61	55.39	285.50	< 0.001
	Urban	60.58	39.42		
Mother's education	No formal education	42.98	57.02	510.57	< 0.001
	Primary	51.81	48.19		
	Secondary and above	69.85	30.15		
Father's education	No formal education	41.09	58.91	475.61	< 0.001
	Primary	49.68	50.32		
	Secondary and above	59.81	40.19		
Woman's autonomy tertiles	Low autonomy	44.11	55.89	49.84	< 0.001
	Middle autonomy	47.60	52.40		
	High autonomy	48.84	51.16		
Source of drinking water	Unimproved	44.77	55.23	20.04	0.009
	Improved	47.41	52.59		
Toilet facilities	Unimproved	41.29	58.71	442.18	< 0.001
	Improved	53.77	46.23		
BMI	Underweight	40.46	59.54	278.85	< 0.001
	Normal	46.85	53.15		
	Overweight	66.01	33.99		
Household number	Less than 4	48.42	51.58	18.74	0.017
	5–9	45.48	54.52		
	≥10	47.22	52.78		
Number of under-five children in HH	1	47.19	52.81	62.44	< 0.001
	2	44.23	55.77		
	3 or more	50.43	49.57		
Media exposure	No	43.03	56.97	205.71	< 0.001
	Yes	51.63	48.37		

Table 2 (continued)

Variables	Categories	Nourished (%)	CIAF (%)	X^2 test statistic	*p* values
Mother's working status	Unemployed	48.14	51.86	65.67	< 0.001
	Employed	43.27	56.73		
Wealth quintile	Poorest	40.31	59.69	343.16	< 0.001
	Poorer	42.56	57.44		
	Middle	45.92	54.05		
	Richer	48.56	51.44		
	Richest	56.67	43.33		
EDHS	2000	38.81	61.19	394.42	< 0.001
	2005	43.62	56.38		
	2011	48.54	51.46		
	2016	53.34	46.66		

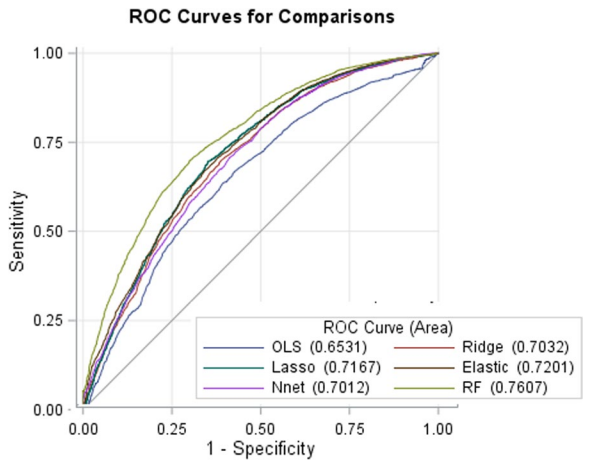

Fig. 3 ROC curves for LR, L-1 regularization, L-2 regularization, elastic net regularization, ANN, and RF in predicting undernutrition among under-five children

appropriate model to undernutrition classification. Moreover, it was noticed that the prediction model based on RF demonstrated the best-performed model, with AUC up to 0.761, followed by LASSO (AUC = 0.717), while the perdition model using the traditional model (LR) is the least efficient (AUC = 0.653). Hence RF model was chosen as the classification engine to construct the perdition model for under-five undernutrition in Ethiopian administrative zones (Table 3).

In machine learning prediction, identifying important attributes is also crucial. The importance of each aspect for a tree's decision is represented by feature importance rates. The random forest (best algorithm for childhood undernutrition in our study gives the MDA and MDG measures of the relative importance of covariates in the model which are summarized in Fig. 4. The factors include urban–rural settlement (ur), the total number of under-five population, the BMI, literacy rates of parents

Table 3 The performance of the prediction models based on different classifications on the independent tests for two ratios

Train/test ratios	Algorithms	Sensitivity	Specificity	Precision	F1	AUC (95% CI)	Accuracy (95% CI)
80/20	GLM	0.585	0.169	0.399	0.475	0.630 (0.619, 0.641)	0.371 (0.359, 0.383)
	Ridge	0.503	0.789	0.683	0.580	0.699 (0.686, 0.713)	0.645 (0.633, 0.658)
	Lasso	0.484	0.814	0.711	0.576	0.711 (0.698, 0.724)	0.654 (0.641, 0.666)
	elastic-net	0.484	0.802	0.697	0.572	0.701 (0.689, 0.714)	0.647 (0.635, 0.660)
	NN	0.499	0.785	0.686	578	0.697 (0.684, 0.711)	0.646 (0.634, 0.658)
	RF	0.524	0.819	0.732	0.611	0.756 (0.744, 0.769)	0.676 (0.663, 0.688)
70/30	GLM	0.601	0.189	0.361	0.445	0.653 (0.639, 0.667)	0.356 (0.344, 0.369)
	Ridge	0.510	0.804	0.743	0.604	0.703 (0.690, 0.717)	0.649 (0.636, 0.661)
	Lasso	0.516	0.819	0.698	0.593	0.717 (0.704, 0.730)	0.683 (0.671, 0.695)
	Elastic-net	0.527	0.824	0.717	0.608	0.720 (0.707, 0.733)	0.682 (0.670, 0.694)
	NN	0.499	0.785	0.751	0.621	0.701 (0.688, 0.715)	0.656 (0.644, 0.668)
	RF	0.524	0.819	0.715	0.595	0.761 (0.749, 0.773)	0.688 (0.676, 0.700)

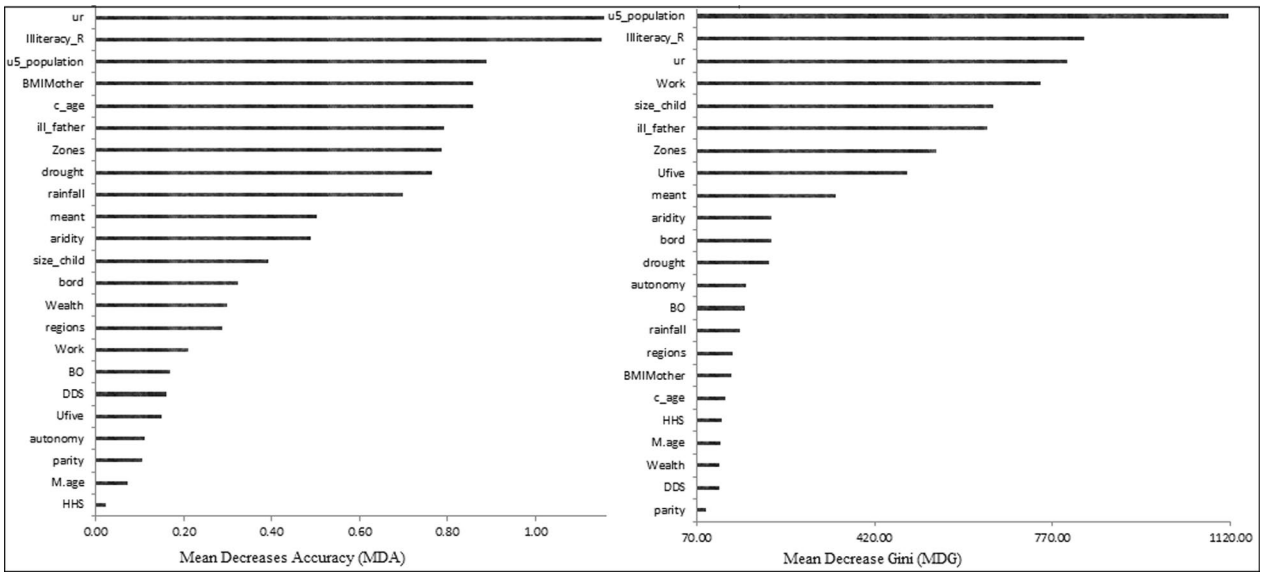

Fig. 4 Relative Variable importance from the best model (random forest)

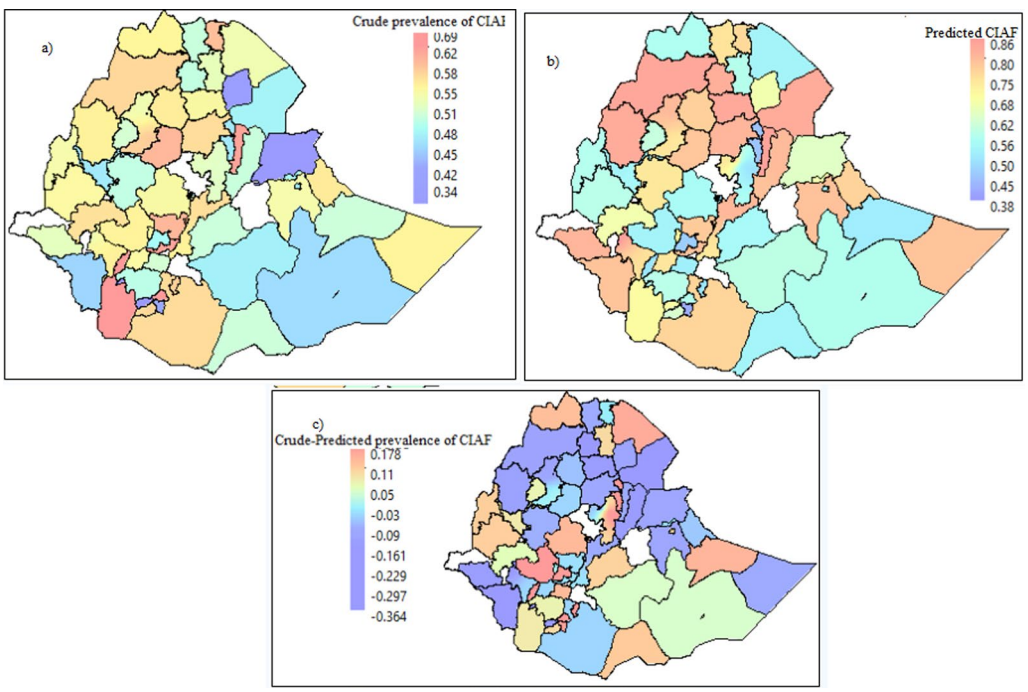

Fig. 5 mapping the predicted and actual prevalence of undernutrition outcomes based on the test data

and zones were the most important predictors of CIAF, but household size, age of mother, parity, and autonomy were the lowest predictive variables in our model (Fig. 4).

The predicted values with the actual values of undernutrition among the 72 administrative areas were mapped in Fig. 5. Having the best predictive model (RF) that yielded the highest AUC, we further predicted the undernutrition status of under-five children by the administrative zones. Both the crude and predicted undernutrition values were merged with the second-level administrative level (zones) shapefiles. A visual comparison confirms that while discrepancies did exist between few zones, the overall patterns of the observed

prevalence were in line with the patterns of the predicted prevalence of undernutrition. The degrees of agreement between the actual and predicted values indicated that the two variables are strongly correlated. Moreover, the third map reveals that the difference. Further, it is between the crude and predicted CIAF of U5C in some zones that have a positive difference indicated that the crude prevalence is less than the predicted value and vice versa (Fig. 5).

Discussions

Previous studies carried out on this subject reported that Ethiopia is one of the countries with the highest number of under-five undernourished children in the world [2, 4, 8, 78, 79]. Further, the studies indicated that, while the prevalence of under-five undernutrition has declined in the nation from time to time, more effort is needed to facilitate this decline and to contain the negative consequences of the phenomena. In this study, we briefly described spatial disparities in under-five undernutrition and predicted under-five undernutrition among Ethiopian administrative zones. The spatial maps show evidences of considerable zonal disparities in under-five undernutrition rates in the administrative zones similar to what has been reported in different countries [80–82]. The continuous data in this study were normalized and the categorical variables were encoded. The machine learning models are known as advanced approaches and techniques for quick and accurate prediction of real-world problems. In this paper, the ML techniques are analyzed by investigating the influence of training/testing ratio on the performance of the six popular ML models to predict the undernutrition of under-five children. The performance of the ML models was slightly changed under the two different ratios. The result revealed that the ratio 70/30 was the most suitable ratio for the training and validating ML models. This study is in line with previously published studies [18, 23, 30–44, 83–86]. The ML tool can offer insight into the identification of novel factors associated with under-five undernutrition that can serve as targets for intervention. Among the six predictive models built using these techniques, the Random Forest (RF) model reveals a higher predictive power as compared to other ML models including the logistic regression. The RF model reveals that urban–rural settlement ratio, the literacy level of parents, under five populations, BMI of mothers, locations (zones, place of residence), and rainfall distributions were the top important predictors of under-five undernutrition in Ethiopia. This study is consistent with previous studies [4, 42, 79, 81]. Moreover, the selected ML algorithm reveals consistent effects of the covariates with the classical generalized linear model which shows that the educational level of parents, the age of the child, sex of the child, birth order, dietary diversity, types of the birthplace of residence, women's autonomy, household sanitation, and a clean water supply were the most significant variables for undernutrition [4, 6, 7, 10, 21, 79–82]. The child's residence (zones) was one of the important risk factors for the U5C CIAF rate which varied significantly across spatial zones. Moreover, this paper briefly explored the spatial variation in under-five child undernutrition and the predicted under-five undernutrition risk factors in Ethiopia using the different machine learning approaches. Hence, we explored a spatial map for the crude prevalence and predicted (from RF) rate of under-five undernutrition by zones in Ethiopia to document the zonal disparities in under-five undernutrition in the country.

Limitations

Since there are no regression coefficients and no directional effects in ML algorithms, the parameters are difficult to be interpreted [21, 23, 87]. In the current study, ML models only predict or classify certain variables depending on the importance of their contribution in determining under-five undernutrition instead of causal inferences. More types of classification ML algorithms could also have been used [21, 23, 28, 38, 59].

Conclusions

The main objective of this study was to compare and evaluate the performance of different machine learning (ML) algorithms considering the influence of two train-test splits ratios in predicting the undernutrition under-five classification. Popular statistical indicators, such as accuracy and area under the curve were employed to evaluate the predictive power of the ML models under different testing and training ratios. The higher the accuracy the model had, the better was the performance of the model. Our results confirm that ML models can effectively predict the under-five undernutrition status and hence may be useful for concerned body decision tools. The best model was the RF, with accuracy and AUC of (68.2%, 76.2%) respectively. The findings from this paper showed that considerable zonal disparities in the under-five undernutrition status persist in the northern part of Ethiopia. When implementing health policies aimed at the redaction of child undernutrition in Ethiopian administrative zones, the zone characteristics must be taken into account.

Abbreviations

AUC: Area under the curve; EDHS: Ethiopian Demographic and Health Survey; BMI: Body mass index; ML: Machine learning; RF: Random forest; ANN: Artificial neural network; ROC: Receiver operating characteristic; U5C: Under-five children; LR: Logistic regression.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-021-01652-1.

Additional file 1: Implementation of different Supervised Machine Learning (SML) using R statistical software.

Acknowledgements
The datasets used in this study were obtained from the DHS program thanks to the authorization received to download the dataset on the website**.**

Authors' contributions
HMF was involved in this study from data management, data analysis, drafting, and revising the final manuscript. TZ and EKM contributed to the conception, design, and interpretation of data, as well as to manuscript reviews and revisions. All authors have read and approved the manuscript.

Funding
Not applicable.

Availability of data and materials
The dataset used and analyzed during the current study is available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Ethics approval and consent to participate Institutional review board of Macro International and USAID ethically approved the data utilized on this study. Authorization to make use of the data was formally applied through online registration on the MEASURE DHS website. The study protocol was submitted. Thus, approval was sought to use the datasets.

Consent for publication
Not applicable.

Competing interests
We, the authors, declare that we have no competing interests.

Author details
[1]Department of Statistics, College of Science, Bahir Dar University, Bahir Dar, Ethiopia. [2]School of Mathematics, Statistics and Computer Science, College of Agriculture Engineering and Science, University of KwaZulu-Natal, Durban, South Africa. [3]School of Public Health, College of Medicine and Health Sciences, Bahir Dar University, Bahir Dar, Ethiopia.

Received: 18 August 2021 Accepted: 4 October 2021
Published online: 24 October 2021

References

1. Phalkey RK, et al. Systematic review of current efforts to quantify the impacts of climate change on undernutrition. Proc Natl Acad Sci. 2015;112(33):E4522–9.
2. Organization WH. The state of food security and nutrition in the world 2019: safeguarding against economic slowdowns and downturns, vol 2019. Food & Agriculture Org; 2019.
3. El-Ghannam AR. The global problems of child malnutrition and mortality in different world regions. J Health Soc Policy. 2003;16(4):1–26.
4. Fenta HM, et al. Determinants of stunting among under-five years children in Ethiopia from the 2016 Ethiopia demographic and Health Survey: application of ordinal logistic regression model using complex sampling designs. Clin Epidemiol Glob Health. 2020;8(2):404–13.
5. Kassie GW, Workie DL. Determinants of under-nutrition among children under five years of age in Ethiopia. BMC Public Health. 2020;20:1–11.
6. Pelletier DL, Frongillo EA. Changes in child survival are strongly associated with changes in malnutrition in developing countries. J Nutr. 2003;133(1):107–19.
7. Degarege D, Degarege A, Animut A. Undernutrition and associated risk factors among school age children in Addis Ababa, Ethiopia. BMC Public Health. 2015;15(1):1–9.
8. Takele K, Zewotir T, Ndanguza D. Understanding correlates of child stunting in Ethiopia using generalized linear mixed models. BMC Public Health. 2019;19(1):1–8.
9. Suriyakala V et al. Factors affecting infant mortality rate in India: an analysis of Indian states. In: The international symposium on intelligent systems technologies and applications. Springer; 2016.
10. Habyarimana F, Zewotir T, Ramroop S. A proportional odds model with complex sampling design to identify key determinants of malnutrition of children under five years in Rwanda. Mediterr J Soc Sci. 2014;5(23):1642–1642.
11. Nandy S, Svedberg P. The composite index of anthropometric failure (CIAF): an alternative indicator for malnutrition in young children. In: Handbook of anthropometry. Springer, pp 127–137; 2012.
12. Rasheed W, Jeyakumar A. Magnitude and severity of anthropometric failure among children under two years using Composite Index of Anthropometric Failure (CIAF) and WHO standards. Int J Pediatr Adolesc Med. 2018;5(1):24.
13. Shit S, et al. Assessment of nutritional status by composite index for anthropometric failure: a study among slum children in Bankura, West Bengal. Indian J Public Health. 2012;56(4):305.
14. Mandal G, Bose K. Assessment of overall prevalence of undernutrition using composite index of anthropometric failure (CIAF) among preschool children of West Bengal, India; 2009.
15. Sen J, Mondal N. Socio-economic and demographic factors affecting the Composite Index of Anthropometric Failure (CIAF). Ann Hum Biol. 2012;39(2):129–36.
16. Knol MJ, et al. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. Am J Epidemiol. 2008;168(9):1073–81.
17. Gu W, et al. Use of random forest to estimate population attributable fractions from a case-control study of *Salmonella enterica* serotype Enteritidis infections. Epidemiol Infect. 2015;143(13):2786–94.
18. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017;38(23):1805–14.
19. Ambale-Venkatesh B, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res. 2017;121(9):1092–101.
20. Adler ED, et al. Improving risk prediction in heart failure using machine learning. Eur J Heart Fail. 2020;22(1):139–47.
21. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–30.
22. Shameer K, et al. Machine learning in cardiovascular medicine: are we there yet? Heart. 2018;104(14):1156–64.
23. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng. 2007;160(1):3–24.
24. Quinlau R. Induction of decision trees. Mach Learn. 1986;1(1):S1–106.
25. Gareth J, et al. An introduction to statistical learning: with applications in R. Berlin: Spinger; 2013.
26. Molina M, Garip F. Machine learning for sociology. Annu Rev Sociol. 2019;45:27–45.
27. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media; 2019.
28. Marsland S. Machine learning: an algorithmic perspective. Boca Raton: CRC Press; 2015.
29. Zhang H. The optimality of Naïve Bayes. FLAIRS2004 conference. 2004.
30. Esteva A. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
31. Anderson JP, et al. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. J Diabetes Sci Technol. 2016;10(1):6–18.
32. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29.

33. Ayer T, et al. Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics. 2010;30(1):13–22.
34. Farran B, et al. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. BMJ Open. 2013;3(5):e002457.
35. Aneja S, Lal S. Effective asthma disease prediction using naive Bayes—Neural network fusion technique. In: 2014 international conference on parallel, distributed and grid computing. 2014. IEEE.
36. Behroozi M, Sami A. A multiple-classifier framework for Parkinson's disease detection based on various vocal tests. Int J Telemed Appl. 2016;2016:6837498.
37. Weiss JC, et al. Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. AI Mag. 2012;33(4):33–33.
38. Methun MIH, et al. A machine learning logistic classifier approach for identifying the determinants of under-5 child morbidity in Bangladesh. Clin Epidemiol Glob Health. 2021;12:100812.
39. Bertolini M et al. Machine Learning for industrial applications: a comprehensive literature review. Expert Syst Appl; 2021: 114820.
40. Schmidt J, et al. Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater. 2019;5(1):1–36.
41. Wuest T, et al. Machine learning in manufacturing: advantages, challenges, and applications. Prod Manuf Res. 2016;4(1):23–45.
42. Talukder A, Ahammed B. Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. Nutrition. 2020;78:110861.
43. Khare S, et al. Investigation of nutritional status of children based on machine learning techniques using Indian demographic and health survey data. Procedia Comput Sci. 2017;115:338–49.
44. Rahman SJ, et al. Investigate the risk factors of stunting, wasting, and underweight among under-five Bangladeshi children and its prediction based on machine learning approach. PLoS ONE. 2021;16(6):e0253172.
45. Gebreyesus SH, et al. Local spatial clustering of stunting and wasting among children under the age of 5 years: implications for intervention strategies. Public Health Nutr. 2016;19(8):1417–27.
46. Collaborators GRF. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet (London, England). 2016;388(10053):1659.
47. Corsi DJ, et al. Shared environments: a multilevel analysis of community context and child nutritional status in Bangladesh. Public Health Nutr. 2011;14(6):951–9.
48. Griffiths P, et al. A tale of two continents: a multilevel comparison of the determinants of child nutritional status from selected African and Indian regions. Health Place. 2004;10(2):183–99.
49. Fetene N, et al. The Ethiopian health extension program and variation in health systems performance: what matters? PLoS ONE. 2016;11(5):e0156438.
50. Croft TN et al. Guide to DHS statistics. Rockville, Maryland, USA: ICF; 2018.
51. Esri, ArcGIS Version 10.1. ESRI; 2010.
52. Ibeji JU, et al. Modelling children ever born using performance evaluation metrics: a dataset. Data Brief. 2021;36:107077.
53. Raschka S. Python machine learning. Birmingham: Packt publishing ltd; 2015.
54. Seger C. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing; 2018.
55. Yu H-F, Huang F-L, Lin C-J. Dual coordinate descent methods for logistic regression and maximum entropy models. Mach Learn. 2011;85(1–2):41–75.
56. Arthur EH, Robert WK. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.
57. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol). 1996;58(1):267–88.
58. Zou H, Hastie T. Addendum: regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol). 2005;67(5):768–768.
59. Hecht-Nielsen R. Theory of the backpropagation neural network. In: Neural networks for perception. Elsevier. p. 65-93; 1992.
60. Abdelhafiz D, et al. Deep convolutional neural networks for mammography: advances, challenges and applications. BMC Bioinform. 2019;20(11):1–20.
61. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data mining (New York, NY, USA, 2016), KDD '16, ACM; 2016.
62. Garg A, Tai K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. Int J Model Identif Control. 2013;18(4):295–312.
63. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.
64. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol). 2005;67(2):301–20.
65. Yuan G-X, Ho C-H, Lin C-J. An improved glmnet for l1-regularized logistic regression. J Mach Learn Res. 2012;13(1):1999–2030.
66. Genuer R, Poggi J-M, Tuleau-Malot C. VSURF: an R package for variable selection using random forests. R J. 2015;7(2):19–33.
67. Robin X, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform. 2011;12(1):1–8.
68. Khan MRAA. ROCit-An R package for performance assessment of binary classifier with visualization; 2019.
69. Wickham H, Chang W, Wickham MH. Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. Version. 2016; 2(1): 1–189.
70. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
71. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. Pattern Recogn Lett. 2010;31(14):2225–36.
72. Janitza S, Tutz G, Boulesteix A-L. Random forest for ordinal responses: prediction and variable selection. Comput Stat Data Anal. 2016;96:57–73.
73. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.
74. Liang N-Y, et al. A fast and accurate online sequential learning algorithm for feedforward networks. IEEE Trans Neural Netw. 2006;17(6):1411–23.
75. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;327(8476):307–10.
76. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
77. Han H, Guo X, Yu H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In: 2016 7th IEEE international conference on software engineering and service science (ICSESS). IEEE; 2016.
78. Gebre A et al. Prevalence of malnutrition and associated factors among under-five children in pastoral communities of Afar Regional State, Northeast Ethiopia: a community-based cross-sectional study. J Nutr Metab. 2019;2019.
79. Kassie GW, Workie DL. Determinants of under-nutrition among children under five years of age in Ethiopia. BMC Public Health. 2020;20(1):1–11.
80. Spray AL, et al. Spatial analysis of undernutrition of children in leogane Commune, Haiti. Food Nutr Bull. 2013;34(4):444–61.
81. Simler KR. Nutrition mapping in Tanzania: an exploratory analysis. IFPRI Food Consumption and Nutrition Division Discussion Paper, 2006(204).
82. Khan J, Mohanty SK. Spatial heterogeneity and correlates of child malnutrition in districts of India. BMC Public Health. 2018;18(1):1–13.
83. Pham BT, et al. Spatial prediction of rainfall-induced landslides using aggregating one-dependence estimators classifier. J Indian Soc Remote Sens. 2018;46(9):1457–70.
84. Verma C, Illés Z. Attitude prediction towards ICT and mobile technology for the real-time: an experimental study using machine learning. In: The international scientific conference elearning and software for education. 2019. "Carol I" National Defence University.
85. Van Dao D, et al. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. CATENA. 2020;188:104451.
86. Nguyen PT, et al. Soft computing ensemble models based on logistic regression for groundwater potential mapping. Appl Sci. 2020;10(7):2469.
87. Bitew FH, et al. Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey. Genus. 2020;76(1):1–16.

Publisher's Note