PLOS ONE

RESEARCH ARTICLE

Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia

Md. Merajul Islam[1,2]*, **Md. Jahangir Alam**[2,3]*, **Md Maniruzzaman**[4], **N. A. M. Faisal Ahmed**[5], **Md Sujan Ali**[6], **Md. Jahanur Rahman**[2], **Dulal Chandra Roy**[2]

1 Department of Statistics, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh, **2** Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh, **3** Mainanalytics GmbH, Sulzbach/Taunus, Germany, **4** Statistics Discipline, Khulna University, Khulna, Bangladesh, **5** Institute of Education and Research, University of Rajshahi, Rajshahi, Bangladesh, **6** Department of Computer Science and Engineering, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh

* jahangir_statru63@yahoo.com (MJA); merajul.stat4811@gmail.com (MMI)

Abstract

Background and objectives

Hypertension (HTN), a major global health concern, is a leading cause of cardiovascular disease, premature death and disability, worldwide. It is important to develop an automated system to diagnose HTN at an early stage. Therefore, this study devised a machine learning (ML) system for predicting patients with the risk of developing HTN in Ethiopia.

Materials and methods

The HTN data was taken from Ethiopia, which included 612 respondents with 27 factors. We employed Boruta-based feature selection method to identify the important risk factors of HTN. The four well-known models [logistics regression, artificial neural network, random forest, and extreme gradient boosting (XGB)] were developed to predict HTN patients on the training set using the selected risk factors. The performances of the models were evaluated by accuracy, precision, recall, F1-score, and area under the curve (AUC) on the testing set. Additionally, the SHapley Additive exPlanations (SHAP) method is one of the explainable artificial intelligences (XAI) methods, was used to investigate the associated predictive risk factors of HTN.

Results

The overall prevalence of HTN patients is 21.2%. This study showed that XGB-based model was the most appropriate model for predicting patients with the risk of HTN and achieved the accuracy of 88.81%, precision of 89.62%, recall of 97.04%, F1-score of 93.18%, and AUC of 0. 894. The XBG with SHAP analysis reveal that age, weight, fat, income, body mass index, diabetes mulitas, salt, history of HTN, drinking, and smoking were the associated risk factors of developing HTN.

OPEN ACCESS

Citation: Islam M.M, Alam M.J, Maniruzzaman M, Ahmed NAMF, Ali MS, Rahman M.J, et al. (2023) Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia. PLoS ONE 18(8): e0289613. https://doi.org/10.1371/journal.pone.0289613

Editor: Ali Garavand, Lorestan University of Medical Sciences, ISLAMIC REPUBLIC OF IRAN

Received: April 16, 2023

Accepted: July 22, 2023

Published: August 24, 2023

Data Availability Statement: The data used for our paper can be accessed at: https://figshare.com/s/a709a390ecd276046607.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Conclusions

The proposed framework provides an effective tool for accurately predicting individuals in Ethiopia who are at risk for developing HTN at an early stage and may help with early prevention and individualized treatment.

Introduction

Hypertension (HTN), defined as the elevated blood pressure beyond its normal ranges, is a major public health concern with its raising prevalence and effect among the adults' overtime worldwide [1–3]. It is one of the most common serious chronic non-communicable diseases. Hypertensive people are affected by different types of cardiovascular diseases (CVDs), e.g., coronary heart disease, stroke, peripheral arterial disease, aortic disease, myocardial infarction [4–7], which are the leading cause of disability, morbidity and mortality that increase the economic burden of out-of-pocket expenditures (OOPE) [8–10]. As reported by World Health Organization (WHO), worldwide around 9.4 million people were died due to HTN every year [10]. According to Belay et al., [2022], globally the prevalence of HTN was 26% in 2000 and it was projected to reach around 1.56 billion (29.2%) by 2025 [11]. The latest estimation by WHO in 2021 revealed that about one-third (31.1%) of the world's adult population had HTN (1.39 billion); of whom 2/3 were from in low and middle-income countries (LMICs) [12]. Also, a systematic analysis of population-based studies from 90 countries, including Ethiopia estimated that HTN among adults was more prevalent in LMICs (31.5%) than the high-income countries (28.5%) [13]. Different epidemiological studies in Ethiopia reported that the prevalence of HTN was ranging from 7.7%-41.9% [14]. Moreover, the prevalence of HTN is disproportionately more prevalent and it increases alarmingly in poor resource countries, like Ethiopia [11]. But it might be helpful to mitigate and manage/control the risk of HTN if identification of HTN patients with interpretable risk factors at an early stage. Thus, early detection of HTN patients with identification of interpretable risk factors plays a key role, which could help to get the patients timely prevention and intervention. It is therefore highly essential to detect/diagnosis and identify the interpretable risk factors of HTN at an early stage.

Many convincing research and empirical studies determined several risk factors associated with HTN in LMICs countries, including Ethiopia [15–21]. Nevertheless, existing association studies had several limitations. Most importantly, previous existing studies considered traditional linear models, such as logistic regression (LR), Cox proportional hazard model, for identifying the significantly associated risk factors of HTN [22–24]. Moreover, a real data with high-dimensional non-linear pattern presents a challenge to traditional linear models, and low precision of linear models impedes patients-level use. To overcome those limitations with complex real data, machine learning (ML) might be a right choice, which is being widely used in current public health research fields. ML is a subset of artificial intelligence (AI), in which the algorithms that execute the prediction process collect the necessary information from previous experiences and/or detect patterns in data to accomplish a task, typically a classification or identification [25–28]. It can provide several advantages, including automatic specific process, reliable probabilistic estimation for uncovering hidden patterns or relationships with high accuracy while lowering labor costs and time for large amounts of data that aid in decision-making or inference, and model interpretability [29–31]. There are different types of learning algorithm in ML, among them supervised learning is the most popular and widely applicable. The supervised learning algorithm's goal is to use the dataset to build a model that can predict the system's output given new inputs. The major two types of supervised learning algorithm are regression and classification. Example of regression include linear regression and logistic regression [32]. Examples of classification include ensemble methods, decision trees (DT), k-nearest neighbors (kNN), support vector machine (SVM), Naïve Bayes (NB), artificial neural network (ANN), so on [32, 33]. The ensemble method is a machine learning technique that combine multiple models with the same learning algorithm to achieve better predictive performance [34]. Ensemble methods include eXtreme gradient boosting (XGB), adaBoost, histogram-based gradient boosting classification Tree, and random forest (RF) [25]. However, previously, some researcher's conducted their study to develop multivariable prediction models using several ML and explainable artificial intelligence methods [35–37]. Most of the existing risk prediction models were developed with limited number of risk factors that provided less accuracy for predicting HTN patient [35, 38]. However, DT and ensemble approaches have attracted a great attention in recent years for identifying individuals at risk of HTN, there is no evidence that these algorithms are successfully applied in Ethiopian clinical settings.

To the best of our knowledge, this is the first study that applied and builds a predictive model using ML algorithms for predicting the individual risk of HTN in Ethiopia. Thus, the objective of this study was to develop an efficient ensemble based explainable ML framework for predicting patients with the risk of HTN in Ethiopia.

Furthermore, we employed under-sampling and adaptive syntactic (ADASYN) class balancing strategy to enhance the confidence score of the developed prediction models. For model interpretation, we identified the key risk factors of HTN and direction of the relationship between the risk factors and HTN using SHapley Additive exPlanations (SHAP), which is a post hoc model interpretation technique viz. theoretically based on the Shapley value. The overall pipeline of the explainable machine learning based framework is displayed in Fig 1.

The layout of this paper is presented as follows: Materials and methods included data source, statistical analysis, feature selection, machine learning algorithms, performance evaluation criteria, and model interpretability. The results are presented in section 3 and discussed in section 4. Finally, conclusion is represented in section 5.

Materials and methods

Data source

The community-based cross-sectional data used in this investigation were collected in 2017 by the Hawassa city administration and made available to the public by Paulose et al. [39]. The data were collected through multistage random sampling and comprised a total of 633 respondents, ranging in age from 31 to 90, and residing in the city for at least six months. The sample size was determined by using the formula of sample size determination method, which considered the design effect of 1.5, the 95% confidence interval, the 5% margin of error, the 80% power, the proportion of 50% (to maximize sample size), and the 10% non-response rate [39]. Different levels of explanatory variables were included as individual risk factors of HTN and categorized the quantitative variables based on the previous sittings [18–20, 39]. A brief explanation of the included risk factors has been presented in Table 1. In this study, a patient with HTN is determined based on WHO cutoff ($\geq$140/90 mmHg and/or diastolic pressure $\geq$90 mmHg and/or being on medication of HTN at the time of data collection) [40]. Finally, a total of 612 respondents were incorporated in this study after eliminating all the missing values.

Statistical analysis

The baseline and demographic characteristics of the patients were presented in percentage (%) for categorical and mean ± SD (standard deviation) for continuous data. Pearson χ^2-test was employed to determine the association between categorical risk factors and HTN, whereas for

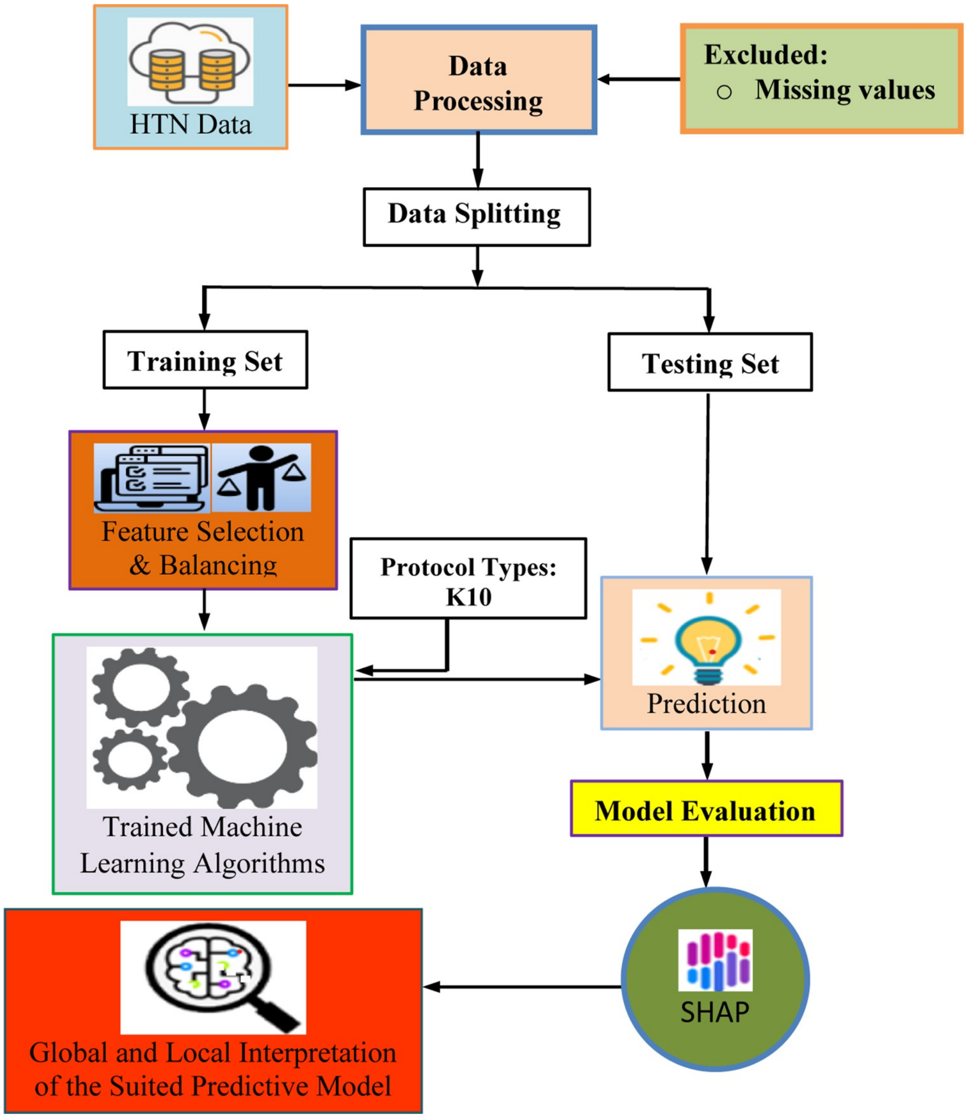

Fig 1. Workflow of the proposed ML-based methodology for predicting risk of HTN.

https://doi.org/10.1371/journal.pone.0289613.g001

continuous risk factors, independent sample t-test was used to examine the mean difference between the HTN groups (HTN vs. non-HTN) for normally distributed data. Two-sided test was performed and a p-value of <0.05 was considered statistically significant for all the tests. Data analysis was performed by SPSS (version-27.0) and R (version-4.2.2).

Feature selection

Feature selection (FS), or risk factor identification is also known as variable selection, or subset selection in statistics and ML. The identification of risk factors is a method for selecting the relevant features by removing the irrelevant or redundant features from the dataset. In this study, Boruta-based feature selection method (FSM) was adopted to identify the relevant features. Boruta is a wrapper-based feature selection method that employs the random forest classifier algorithm. This method has a wider range of applications and performs better than others as it is unbiased and steady [41].

Table 1. Name, description, and categorization of the selected factors.

SN	Name	Description	Categorization
1	Residence	Permanent Residence	Urban, Semi-urban
2	Sex	Sex of the respondents	Male, Female
3	Age	Age of the respondents	Continuous variable (year)
4	MS	Marital status	Single, Married, Divorced, Widowed
5	Religion	Religion status	Protestant, Orthodox, Catholic, Muslim, Others
6	Ethnicity	Ethnicity	Sidama, Walayita, Kembata, Guraga, Amahra, Oromo, Hadiya
7	Education	Education level	Cannot read and write, Read and write only, Primary, Secondary, Diploma and above
8	Occupation	Occupation status	Employee, Daily-labor, Merchant, Housewife, Retired, Others
9	FM	Family member	1–3, 4–6, 7 or 7+
10	Income	Average monthly income	Continuous variable (Birr)
11	PA	Physical activity	Yes, No
12	Walking	Walking at least 10 minutes	Yes, No
13	Diabetes	Having diabetes mellitus	Yes, No
14	Height	Height of the respondents	Continuous variable (cm)
15	Weight	Weight of the respondents	Continuous variable (kg)
16	BMI	Body mass status	Underweight, Normal, Overweight, Obese
17	Smoking	Smoking status	Yes, No
18	Drinking	Drinking alcohol	Yes, No
19	Kchat	Ever chew kchat	Yes, No
20	Fruit	Eat fruit at least per week	Yes, No
21	Vegetable	Eat fruit at least per week	Yes, No
22	Fat	Having fat	Yes, No
23	Salt	Eating habit salt	Yes, No
24	Transport	Mode of transport	On foot/pedal bicycle, Engine
25	HD	History of diabetes	Yes, No
26	Wealth	Wealth status	Poorest, Very poor, Poor, Less poor, Least poor
27	HHTN	History of hypertension	Yes, No

https://doi.org/10.1371/journal.pone.0289613.t001

Machine learning algorithms

This study used three different types of supervised ML algorithms for predicting patients with the risk of HTN (Table 2).

Logistic regression

Logistic regression (LR) is a most popular supervised ML-based algorithm that leverages the idea of probability. Logistic regression (LR) is a most popular supervised ML algorithm mainly used for classification task [42]. The LR model employs the logistic function to estimate the probability of the response variable (HTN and non-HTN) in terms of one or more input

Table 2. Different machine learning algorithms with types.

Types	Algorithms
Classical	Logistic regression (LR)
Non-linear	Artificial neural network (ANN)
Ensemble	Random forest (RF) and extreme gradient boosting (XGB)

https://doi.org/10.1371/journal.pone.0289613.t002 features. The logistic function can be represented as follows

$$\mathrm{logit}(p_j) = \log_e\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 \mathrm{x}_{1j} + \beta_2 \mathrm{x}_{2j} + \ldots \quad .. \quad + \beta_k \mathrm{x}_{kj} + \epsilon_j, j = 1, 2, \ldots, n \quad (1)$$

where, p_j denote the probability of HTN and $(1-p_j)$ denote the probability of non-HTN for jth individual; X_{kj} is the kth input feature of the jth individual and β_k is the kth regression coefficients.

The above Eq (1) can be expressed as

$$p = \frac{\exp(\beta_0 + \beta_1 \mathrm{x}_{1j} + \beta_2 \mathrm{x}_{2j} + \ldots \quad .. \quad + \beta_k \mathrm{x}_{kj})}{1 + \exp(\beta_0 + \beta_1 \mathrm{x}_{1j} + \beta_2 \mathrm{x}_{2j} + \ldots \quad .. \quad + \beta_k \mathrm{x}_{kj})} \quad (2)$$

and odds as

$$\frac{p}{1-p} = \exp\left(\beta_0 + \beta_1 \mathrm{x}_{1j} + \beta_2 \mathrm{x}_{2j} + \ldots \quad .. \quad + \beta_k \mathrm{x}_{kj}\right) \quad (3)$$

If $\frac{p}{1-p} > 1$, then we classify as HTN, while $\frac{p}{1-p} < 1$, then we classify as non-HTN.

Artificial neural network

Artificial neural network (ANN) is a non-linear modeling algorithm that is inspired by the structure and function of human brain. It consists of interconnected processing nodes that are organized by three different types of layers: input, hidden, and output. The input layer is connected to hidden layer with updated weight, and hidden layer is connected to the output. In this method, $X = x_1,\ldots,x_k$ are used as the input vector in back propagation (BP) algorithm for learning as well as mapping the relationship between input features and outcome variable. The BP algorithm propagates the error between the input risk factors and outcome variable by adjusting weights of hidden layers via backward direction with non-linear sigmoid activation function [43]. The sigmoid activation function is defined as

$$Sigmoid(x) = \frac{1}{1+e^{-x}}. \quad (4)$$

This procedure is repeated iteratively until no change iteration values or not getting the minimum error.

Random forest

Random forest is a popular machine learning algorithm that developed by Leo Breiman and widely used in classification and regression problems [44]. It is based on the concept of ensemble learning algorithm that trains multiple decision tree on random subsets of the data to solve the problem. The RF-based model is constructed by using the following steps:

Step1: The given training data set (X_{ij}, i = 1, 2. . . k, j = 1, 2. . . n), select randomly risk factors from training dataset by using bootstrap sampling procedure.

Step 2: Built a decision tree (DT) for creating new subset.

Step3: Repeat Step1 and Step2, until construct many trees and consist of a forest.

Steps 4: Consider the prediction result from each created DT and select final prediction with the help of majority voting.

Extreme gradient boosting

Extreme gradient boosting (XGB) is an efficient ensemble-based machine learning algorithm that uses decision trees and gradient boosting algorithm. It is highly adaptable and working in most classification problem, especially HTN disease prediction [45]. Boosting is a learning algorithm, which attempts to create a strong classifier based on weak learners or classifiers. The weak and strong classification models mention to the correlation of predicted and actual class. By adding classifiers on top of each other iteratively, the next classifier can modify the errors of the earlier one. This procedure is repeated until the training data set accurately predicts the membership class label of the target variable.

Data partition and balancing

We randomly divided the whole dataset into two sets as 70% training set [HTN: 91 (21.2%), non-HTN: 338 (78.8)] and 30% testing set HTN: 39 (21.3%), non-HTN: 144 (78.7)] using stratified sampling procedure [46]. Membership class label of the data was imbalance i.e., skewed class distribution of observations. Imbalance class problem of a data provided a biased result for the majority class of the response variable in classification task [47, 48]. To deal this problem, several data balancing strategy are widely applicable. Among them, under-sampling and Adaptive synthetic (ADASYN) balancing strategy were executed in the training set to balance the data. ADASYN is the newly generalized version of synthetic minority oversampling technique (SMOTE) and generates new sample for the minority class using a weighted distribution [49].

Cross validation and tune hyperparameters

The mentioned above four ML algorithms (LR, ANN, RF, and XGB) have other parameters, called hyperparameters. Hyperparameters are those parameters that the user explicitly defines before the learning process to improve the model performance. The grid search method with repeated10-fold (K10) cross-validation protocol was used to tune the hyperparameter values in the training set. The training dataset is divided into a 9:1 ratio as a training subset and a verification set to perform the K10 protocol. The caret package (version 6.0-93) in R was used to generate the optimal hyperparameter values for four models, which are displayed in Table 3.

Performance evaluation criteria

The performance of selected four ML models was evaluated by five popular evaluation criteria: accuracy, precision, recall, F-score, and area under the curve (AUC). The values of performance evaluation criteria were calculated from the confusion matrix by four measures (Table 4):

True positive (t_p): model predicted the disease group as HTN where actual group was HTN,

Table 3. The value of hyperparameter for ML-based models.

Models	Hyperparameter
LR	c = (0.001, 0.01, 0.1, 1, 10,100,1000)
ANN	size = (1, 2, 3, 4, 5, 6, 7,8, 9, 10); decay = (0, 0.1, 0.01, 0.001, 0.001)
RF	mtry = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
XGB	nrounds = (100, 500); max_depth = (10, 15, 20, 25, 30); colsample_bytree = seq (0.5, 0.9, length.out = 5); eta = 0.1; gamma = 0; min_child_weight = 1; subsample = 1.

https://doi.org/10.1371/journal.pone.0289613.t003

Table 4. Confusion matrix.

		Actual class	
	Total cases/population = HTN + non-HTN	**HTN**	**non-HTN**
Predicted class	HTN	True positive (t_p)	False positive (f_p)
	non-HTN	False negative (f_n)	True negative (t_n)

https://doi.org/10.1371/journal.pone.0289613.t004

False positive (f_p): model predicted the disease group as HTN where actual group was non-HTN,

False negative (f_n): model predicted the control group non-HTN where actual class was HTN,

True negative (t_n): model predicted the control group non-HTN where actual group was non-HTN.

Accuracy. It is used to assess the overall accuracy for the models. It is defined as the ratio of the sum of true cases (t_p and t_n) against total number of cases. Accuracy is defined mathematically as

$$\text{Accuracy}(\%) = \left(\frac{\mathrm{t_p + t_n}}{\mathrm{t_p + t_n + f_p + f_n}}\right) \times 100 \tag{5}$$

Precision. It is the ratio of $\mathrm{t_p}$ cases against the predicted positive (DR) cases. It is also called positive predictive value and used to assess the reliability for predicting the model as positive. Precision is defined mathematically as

$$\text{Precision}(\%) = \left(\frac{\mathrm{t_p}}{\mathrm{t_p + f_p}}\right) \times 100 \tag{6}$$

Recall. It is the ratio of t_p cases against the actual positive cases (DRs). Model with high recall indicates low f_n. It's also called sensitivity or true positive rate (TPR). Recall is defined mathematically as

$$\text{Recall}(\%) = \left(\frac{\mathrm{t_p}}{\mathrm{t_p + f_n}}\right) \times 100 \tag{7}$$

F1-score. It is a harmonic mean of precision and recall. F-score is defined mathematically as

$$\text{F1}-\text{score}(\%) = \left(\frac{\mathrm{2t_p}}{\mathrm{2t_p + f_p + f_n}}\right) \times 100 \tag{8}$$

Area under the curve

The AUC is defined as an integral of the receiver operating characteristic (ROC) function over the given range and used to assess the quality of the built predictive model. The mathematical formula of AUC is as follows

$$AUC = \int_{x=0}^{1} TPR(FPR^{-1}(x))dx \tag{9}$$

A ROC curve is a plot of TPR or *sensitivity* on the y axis against false positive rate (FPR) or *1-specificity* on the x axis for different cutoff values. The ROC curve is broadly used in medical diagnosis as another single-number measure for evaluating the predictive validity of ML-based model [50]. ROCs generate an AUC value from 0 to 1.

Model interpretability

Shapley additive explanations (SHAP) is an interpretability visualization approach, which is constructed based on Shapley values. This method was introduced by Lundberg and Lee (2017), and widely used to explain the local and global importance using SHAP value by computing the contribution of each risk factor in the ML-based prediction model [51]. The explanation value of SHAP was initially established from coalitional game theory, where each predictor is used as an individual player in a game or coalition. SHAP values framework offers a fair solution for each player in a model outcome, and provides a series of desirable properties/axioms, including consistency, efficiency, dummy, and additively [52]. The efficiency property of SHAP method provided better reliable results compared to another methods, for example local interpretable model-agnostic explanations [53]. Risk factors contribute to the model's outcome or prediction with different magnitude and sign, which is accounted for by Shapley values. Accordingly, Shapley values represent estimates of feature importance magnitude of the contribution and its direction (sign). Risk factors with positive SHAP value contribute to predict patent with HTN in the model, whereas risk factors with negative SHAP value contribute to predicting patients with control in the model. Particularly, the importance of each risk factor, say k^{th} risk factor, is measured by the Shapley value defined by the following formula

$$\emptyset_k(v) = \frac{1}{M!}\sum_{S\subseteq M\backslash\{k\}} |S|!(M-|S|-1)![v(S\cup\{k\}) - v(S)] \tag{10}$$

where, S denotes the subset of risk factors, that does not include the risk factor for which we are calculating the value of $\emptyset_k(v)$; $S\cup\{k\}$ is the subset of risk factors, that includes in S and the k^{th} risk factor; $v(S)$ corresponds to the outcome of the ML-based model that explain using the risk factors of S; $S\subseteq M\backslash\{k\}$ represents all sets of S that are subsets of the full set of M risk factors, excluding the k^{th} risk factor.

Results

Baseline characteristics

This study enrolled 612 participants (HTN: 130, 21.2% and non-HTN: 482, 78.8%) with 27 HTN-related predictor variables (Table 5). About 53.4% respondents were male and more than half of the respondents living in urban areas. The average age of the participants was 47.56.20 ±13.40 years, height 165.20±8.87 cm, and weight 66.589±8.769 kg. Obese respondents showed higher prevalence rate of HTN than normal (50.0% vs. 13.4%). Patients having diabetes (47.5% vs. 30.0%) and smoking (50.4% vs. 23.8%) were more prevalent to HTN. The prevalence of HTN was greater among the respondents who had family history of diabetes (41.8% vs. 11.2%) and HTN (60.3% vs. 21.9%). The result of association showed that residence, sex, age, occupation, income, PA, walking, diabetes, height, weight, BMI, smoking, drinking, vegetable, fat, salt, transport, HD, wealth, HHTN were significantly associated with HTN (P-value<0.005).

Risk factors selection using Boruta

The result of Boruta based feature selection method is presented in Fig 2. The method showed that age, occupation, PA, walking, diabetes, height, weight, BMI, smoking, drinking, vegetable,

Table 5. Baseline characteristic of the respondents.

Risk factors	Total	HTN	Non-HTN	P-value
Overall, n (%)	612	130 (21.2)	482 (78.8)	
Residence, n (%)				
Urban	408(66.7)	100(24.5)	308(75.5)	0.006
Semi-urban	204(33.3)	30(14.7)	174(85.3)	
Sex, n (%)				
Male	327(53.4)	85(26.0)	242(74.0)	0.002
Female	285(46.6)	45(15.8)	240(84.2)	
Age, mean (SD)	47.56(13.40)	54.66 (14.04)	45.64(12.56)	<0.001
MS, n%				
Single	29(4.74)	4(13.8)	25(86.2)	0.795
Married	500(81.70)	108(21.6)	392(78.4)	
Divorced	24(3.92)	5(20.8)	19(79.2)	
Widowed	59(9.64)	13(22.0)	46(78.0)	
Religion, n (%)				
Protestant	337(55.07)	62(18.4)	275(81.6)	0.147
Orthodox	181(29.58)	50(27.6)	131(72.4)	
Catholic	38(6.21)	6(15.8)	32(84.2)	
Muslim	50(8.17)	11(22.0)	39(78.0)	
Other	6(0.98)	1(16.7)	5(83.3)	
Ethnicity, n (%)				
Sidama	222(36.27)	43(19.4)	179(80.6)	0.285
Walayita	146(23.86)	26(17.8)	120(82.2)	
Kembata	105(17.16)	24(22.9)	81(77.1)	
Guraga	58(9.48)	14(24.1)	44(75.9)	
Amahra	44(7.19)	15(34.1)	29(65.9)	
Oromo	27(4.41)	7(25.9)	20(74.1)	
Hadiya	10(1.63)	1(10.0)	9(90.0)	
Education, n (%)				
Cannot read and write	139(22.71)	33(23.7)	106(76.3)	0.595
Read and write only	98(16.01)	16(16.3)	82(83.7)	
Primary education (1–8)	84(13.73)	16(19.0)	68(81.0)	
Secondary education (9–12)	108(17.65)	22(20.4)	86(79.6)	
Diploma and above	183(29.90)	43(23.5)	140(76.5)	
Occupation, n (%)				
Employee	199(32.52)	32(16.1)	167(83.9)	0.033
Daily-laborer	53(8.66)	9(17.0)	44(83.0)	
Merchant	165(26.96)	43(26.1)	122(73.9)	
Housewife	118(18.28)	22(18.6)	96(81.4)	
Retired	59(9.64)	20(33.9)	39(66.1)	
Others	18(2.94)	4(22.2)	14(77.8)	
FM, n (%)				
1–3	138(22.55)	22(15.9)	116(84.1)	0.062
4–6	281(45.92)	57(20.3)	224(79.7)	
7 or 7+	193(31.54)	51(26.4)	142(73.6)	
Income, mean (SD)	3169.69 (1999.468)	3664.23 (2503.960)	3036.31(1820.149)	0.001
PA, n (%)				

(*Continued*)

Table 5. (Continued)

Risk factors	Total	HTN	Non-HTN	P-value
Yes	371(60.6)	41(11.1)	330(88.9)	<0.001
No	241(39.4)	89(36.9)	152(63.1)	
Walking, n (%)				
Yes	471(77.0)	63(13.4)	408(86.6)	<0.001
No	141(23.0)	67(47.5)	74(52.5)	
Diabetes, n (%)				
Yes	561(91.67)	98(17.5)	463(82.5)	<0.001
No	51(8.33)	32(62.7)	19(37.3)	
Height, mean (SD)	165.20(8.871)	168.98(8.062)	164.18(8.811)	<0.001
Weight, mean (SD)	66.59(8.769)	73.07(9.462)	64.84(7.698)	<0.001
BMI, n (%)				
Underweight	10(1.63)	2(10.0)	8(80.0)	<0.001
Normal	366((59.80)	49(13.4)	317(86.6)	
Overweight	212(34.64)	67(31.6)	145(68.4)	
Obese	24(3.92)	12(50.0)	12(50.0)	
Smoking, n (%)				
Yes	74(12.1)	29(29.2)	45(60.8)	<0.001
No	538(87.9)	101(18.8)	437(81.7)	
Drinking, yes, n (%)				
Yes	141(23.0)	51(36.2)	90(63.8)	<0.001
No	471(77.0)	79(16.8)	392(83.2)	
Kchat, yes, n (%)				
Yes	526(85.9)	107(20.3)	419(79.7)	0.200
No	86(14.1)	23(26.7)	63(73.3)	
Fruit, no, n (%)				
Yes	429(70.1)	89(20.7)	340(79.3)	0.667
No	183(29.9)	41(22.4)	142(77.6)	
Vegetable, no, n (%)				0.011
Yes	524(85.6)	102(19.5)	422(80.6)	
No	88(14.4)	28(31.8)	60(68.2)	
Fat, yes, n (%)				<0.001
Yes	188(69.3)	61(32.4)	127(67.6)	
No	424(30.7)	61(32.4)	127(67.6)	
Salt, yes, n (%)				<0.001
Yes	81(13.2)	31(38.3)	50(61.7)	
No	531(86.8)	99(18.6)	432(81.4)	
Transport, n (%)				
On foot/pedal bicycle	313(51.1)	51(16.3)	262(83.7)	0.003
Engine	299(48.9)	79(26.4)	220(73.6)	
HD, n (%)				
Yes	51(8.3)	32(62.7)	19(37.3)	<0.001
No	561(91.7)	98(17.5)	463(82.5)	
Wealth, n (%)				

(*Continued*)

Table 5. (Continued)

Risk factors	Total	HTN	Non-HTN	P-value
Poorest	170(27.78)	21(12.4)	149(87.6)	0.011
Very poor	103(16.83)	24(23.3)	79(76.7)	
Poor	117(19.12)	29(24.8)	88(75.2)	
Less poor	161(26.31)	37(23.0)	124(77.0)	
Least poor	61(9.97)	19(31.1)	42(68.9)	
HHTN, n (%)				
Yes	84(13.7)	45(53.6)	39(46.4)	<0.001
No	528(86.3)	85(16.1)	443(83.9)	

https://doi.org/10.1371/journal.pone.0289613.t005

fat, transport, HD, wealth, and HHTN were the important risk factors of HTN. The selected risk factors were included to construct the ML-based model for prediction of HTN status (HTN or non-HTN).

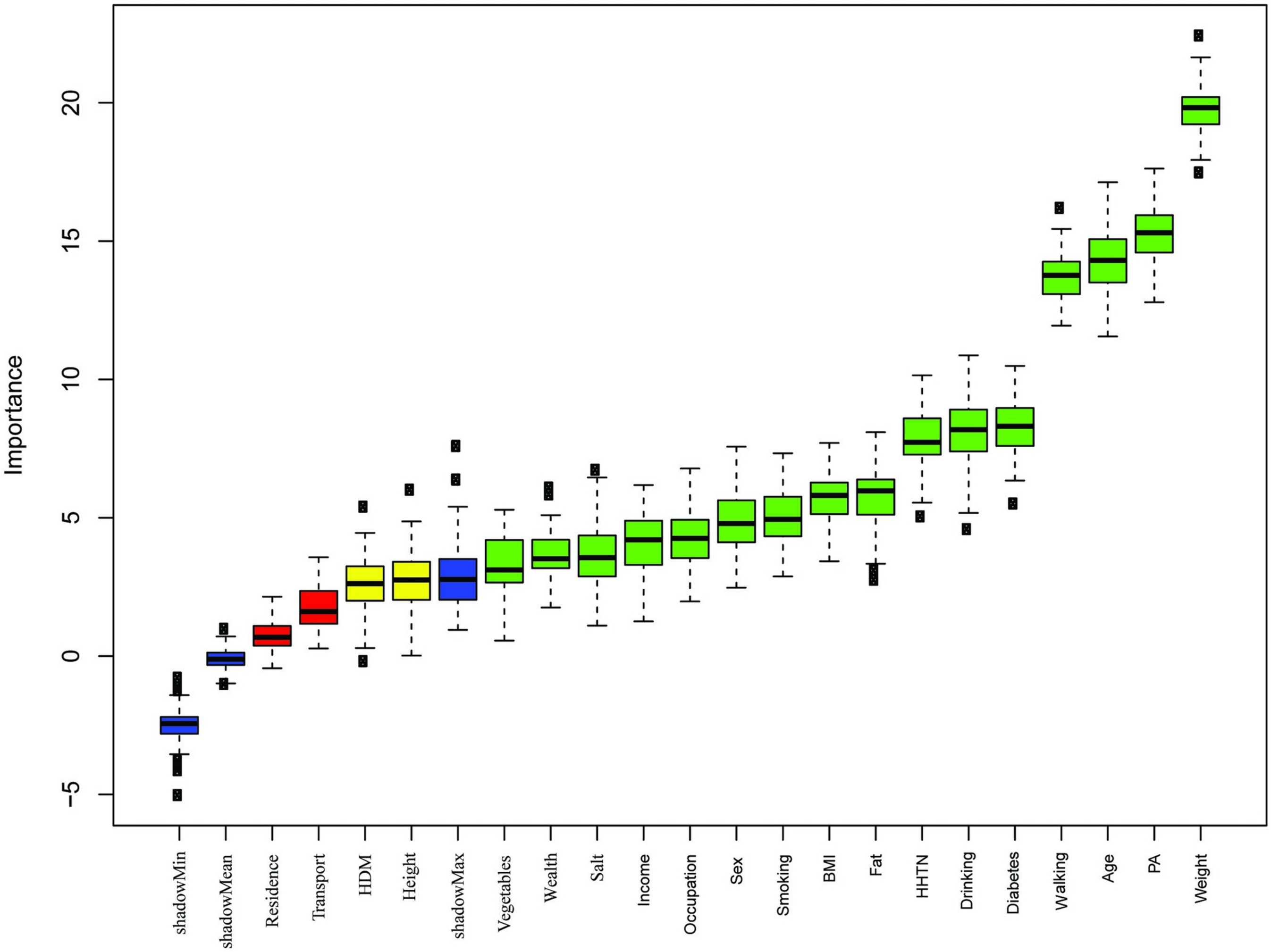

Fig 2. Risk factors selection using Boruta based feature selection method.

https://doi.org/10.1371/journal.pone.0289613.g002

Table 6. Performance of four models with two class balancing methods.

Balancing methods	Models	Accuracy	Precision	Recall	F1-score	AUC
Under-sampling	LR	83.61(77.43–88.66)	87.08	96.44	91.52	0.848(0.789–0.929)
	ANN	84.15(78.04–89.12)	86.02	96.04	91.20	0.843(0,788–0.928)
	RF	85.81(79.88–90.50)	87.42	96.13	92.02	0.863(0.797–0.931)
	XGB	86.89(81.12–91.41)	88.23	96.35	92.51	0.871 (0.794–0.937)
ADASYN	LR	86.43(81.74–91.86)	88.05	97.22	92.41	0.863(0.809–0.941)
	ANN	85.25(79.26–90.05)	87.80	96.67	92.02	0.874(0.816–0.943)
	RF	87.88(84.41–90.81)	88.65	97.04	92.66	0.880(0.806–0.895)
	XGB	**88.81(85.44–91.63)**	**89.62**	**97.04**	**93.18**	**0.894(0.827–0.961)**

https://doi.org/10.1371/journal.pone.0289613.t006

Performance comparisons of ML-based models

The performance of four ML-based models with under-sampling and ADASYN shown in Table 6 and S1 Fig. It is to be noticed that XGB model with ADASYN balancing method achieved the highest predictive discrimination ability with the accuracy of 88.81% (95% CI: 85.44–91.63), precision of 89.62, recall of 97.04, F1-score of 93.18, and AUC of 0.894 (95% CI: 0.827–0.961) compared to others.

The corresponding ROC curves and precision recall curves of four predictive models with ADASYN displayed in Fig 3. The ROC curves and precision recall curves also indicated that the XGB model reached significantly better than other models as LR, ANN, and RF. Therefore, in comparison to other models, our results showed that the XGB-based model with ADASYN performed well.

Interpretable risk factors of hypertension

SHAP analysis was executed to determine the interpretable predictive risk factor of HTN for the suited prediction model (XGB) based on the SHAP values. Fig 4(A) explains the global

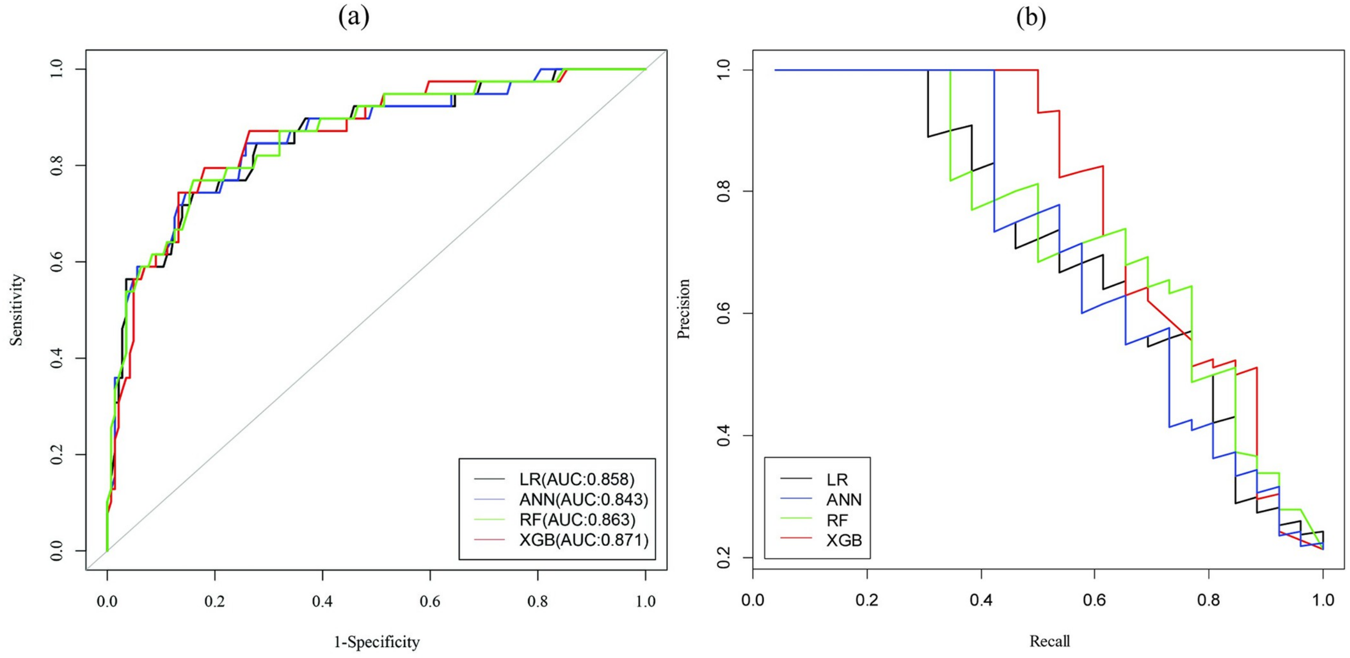

Fig 3. (a) ROC curves and (b) Precision vs. recall curves of four predictive models.

https://doi.org/10.1371/journal.pone.0289613.g003

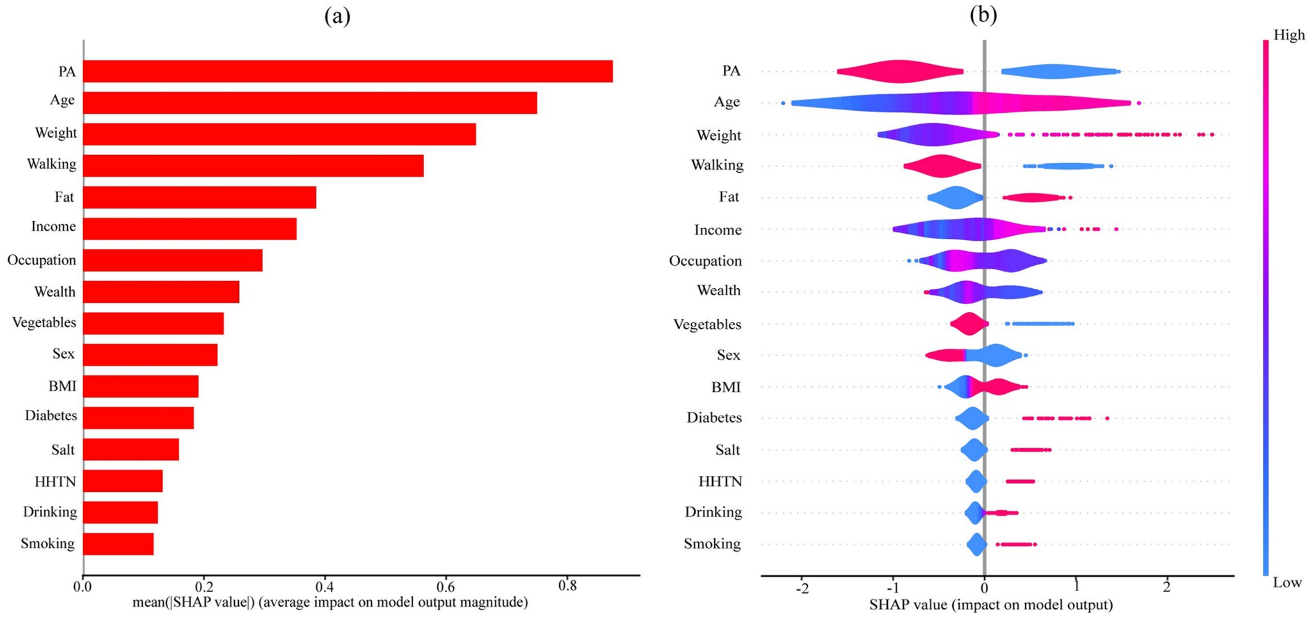

Fig 4. Importance of risk factors based on SHAP values. (A) Mean absolute SHAP values, to explain global risk factor importance, (B) Local explanation summary, to reveal the direction of the relationship between a risk factor and game outcome.

https://doi.org/10.1371/journal.pone.0289613.g004

importance of each risk factor of XGB-based model. The importance plots only show the global influence of each feature on the prediction. However, the global importance plot does not indicate which risk factors affect positively (HTN) or negatively (non-HTN) on the prediction. For that reason, summary plots are executed, which provide a global macro-level explanation of how the input risk factors contribute to the prediction. Fig 4(B) represents the summary plot indicating the importance, impact, original value, and correlation of the risk factors to high risk of HTN. Particularly, the effect [positive (HTN) vs. negative (non-HTN)] is shown on the *x*-axis. The color signifies the value of a specific risk factor, wherein red indicates a high value and blue indicate a low value. However, XGB-based model showed that age, weight, fat, income, BMI, diabetes, salt, HHTN, drinking, and smoking were the high interpretable risk factors on the predication of HTN.

Discussion

In this study, we investigate several ML-based algorithms to propose an explainable framework for predicting the risk of HTN in Ethiopia. We trained up four ML algorithms (ANN, SVM, RF, and XGB) to predict HTN, using 16 risk factors obtained from Boruta feature selection method. The performance of the developed models compared by accuracy, precision, recall, F1-score, and ROC curve with AUC value on testing set. Based on performance measurements, we proposed XGB model as the most appropriate candidate classifier for predicting HTN.

Several studies were conducted using ML framework to predict the risk of HTN. A comparison of the present study with the existing studies is presented in Table 7. Chowdhury et al. [54] proposed a system on 18,322 respondents with 24 candidate risk factors in Canada. Before constructing the models, they applied five top FSM for selecting the significant risk factors and adopted five ML algorithms LASSO, Elastic Net, random survival forest (RSF), and gradient boosting, with the conventional Cox proportional hazard model for predicting HTN. They

Table 7. Comparative performance of the proposed study with the existing studies.

Authors	Year	Country	Data size	# of risk factors	Algorithms	AUC	SHAP
Chowdhury et al. [54]	2023	Canadian	18,322	24	penalized regression Ridge, Lasso, Elastic Net (EN), random survival forest (RSF), and GB	NA	No
Pratiwi [35]	2022	Indonesia	30,320	11	DT, RF, GB, **LR**	0.829	No
Oanh & Tung [55]	2022	Vietnam	2509	10	Naïve Bayes, MLP, Decision Tree, kNN and SVM; and **RF**, boosting and voting	NA	No
Islam et al. [38]	2022	Bangladesh, Nepal, India	818603	7	GT, RF, GBM, **XGB**, LR, LDA	NA	No
Chai et al. [56]	2022	Malaysia	2461	11	LR, DT, RF, SVM, NB, kNN, MLP, GBM, XGB, **LightGBM**, CatBoost, AdaBoost, LogitBoost	0.686	No
Islam et al. [57]	2021	Bangladesh	6965	13	ANN, DT, RF, **GB**	0.669	No
Zheng et al. [58]	2021	USA	500	17	LR, SMV, DTR, GPR, **ANN**	NA	No
AlKaabi et al. [59]	2020	Qatar	987	12	DT, **RF**, LR	NA	No
Proposed		Ethiopia	612	27	LR, ANN, RF, **XGB**	0.894	Yes

https://doi.org/10.1371/journal.pone.0289613.t007

measure the performance of the models by C-index for each model. Pratiwi OA [35] applied four ML algorithms such as DT, RF, GB, and LR for predicting individual risk of HTN in Indonesia. He developed the model by K10 protocol based on training set and prediction performance of these models was measure on testing set in terms of accuracy, precision, recall, F1-score, and AUC. He indicated LR is the best performer marginally compared to others with AUC (0.829). Oanh and Tung [55] suggested a ML based model to predict patient with the risk of HTN in Vietnam. The model was developed by Naïve Bayes (NB), multilayer perceptron (MLP), decision tree (DT), k-nearest neighbors (kNN), SVM, and ensemble algorithms: bagging (RF), boosting and voting based on training set. The performance of the models was assessed by testing set in terms of F1-score, precision, and recall. Islam et al. [38] conducted a study on three countries such as Bangladesh, Nepal, and India. They included 818603 respondents with seven risk factors and performed GT, RF, GBM, XGB, LR, LDA algorithms for predicting HTN patients. They focused that XGB achieved the best performance score than others. Chai et al. [56] used Malaysian data with 2461 respondents and 11 covariates to develop a system for diagnosing HTN patients by 3 different types of algorithms, including neural network (MLP), classical model (LR, DT, NB, k-NN), and ensemble model (RF, SVM, GB, XGB, LightGBM, CatBoost, AdaBoost, and LogitBoost). Before building the model, they adopted correlation-based FSM to select a set of leading features and utilized SMOTE technique to balance membership class label of the data. They evaluate the predictive ability of the models by sensitivity, specificity, accuracy, precision, F1-score, misclassification rate, and AUC on testing set and found that LightGBM based model acquired the best accuracy with 74.39%. Islam et al. [57] used nationally representative HTN data in Bangladesh. The data consisted of 6965 subjects with 13 risk factors. They determine the prominent risk factors of HTN by two popular FSM such as LASSO and SVMRFE in Bangladesh. They utilized then K10 protocol to construct model using four ML algorithms on training set and measured the performance of the models on testing set using accuracy, precision, recall, F1- score and AUC. Overall experimental sittings demonstrated that gradient boosting model attained the best score of AUC (0.669). Zheng et al. [58] explored a system for predicting HTN patients using several ML techniques in USA. No feature selection method had used to select the prominent features of HTN before constructing ML-based system. They found that ANN model reached the maximum performance score. Alkaabi et al. [59] utilized HTN data in Qatar. The dataset comprised of 987 respondents with 12 risk factors. They adopted 3 ML-based algorithms including DT, RF, and

LR. Overall experimental results anticipated that RF model provided better generalization predictive ability than others.

Thus, the comparative results suggested that our proposed XGB framework can predict HTN with higher AUC (Table 7). Moreover, SHAP analysis with the proposed method revealed that age, weight, fat, income, diabetes, BMI, height, salt, smoking, and HHTN were the associated risk factors for developing HTN. Local explanation summary plot showed that age is the 1st leading risk factor of HTN in Ethiopia. A study conducted by Belay et al., [2022] in Ethiopia found that a patient with age>60 years was two times more likely to have HTN than those with age 18–40 years [11]. This result also supported by several systematic review and meta-analysis studies [60, 61]. The vascular system of our body changes in arteries, particularly with large artery stiffness caused by older age. Weight and fat are the 2nd and 3rd leading drivers of HTN. This finding supports the conclusions of earlier investigations [62]. Excess body weight increases visceral and retroperitoneal fat, which can contribute to the development of HTN. Household income is linked to the risk of HTN, which was in line with the prior investigations [63]. Due to a number of reasons, including the ongoing nutritional transition, rising trends in sedentary lifestyle, and other modifiable risk factors, people from low-income families may have a greater burden from the disease [64]. BMI is another gradient of HTN which is corroborated with the earlier studies [65]. BMI might be a cause of HTN and other cardiovascular disease by stimulating the renin-aldosterone system and endothelial dysfunction [66]. Diabetes is another important marker of HTN. The two medical conditions diabetes and HTN may cause each other and share common risk factors. HHTN is another important covariate of HTN. This result is also coincided with the previous studies conducted in Ethiopia and other countries [67]. This might be as family member share same genetic factors, behaviors, mostly similar lifestyle, and environments related factor that could influence the risk of HTN disease. Additionally, other risk factors such as salt, drinking alcohol, and smoking were found to be an important contributing risk factors of HTN, which is similar with other studies in literature [68, 69]. Although this work has many strengths, it also has some limitations, such as the sample only included permanent the residents of the city administration who had lived in the area for more than six months and were older than 30. Additionally, it did not measure the amount of alcohol, cigarettes, fruits, vegetables, fats, and salts that were consumed in measurable units.

Conclusions

In this study, we adopted four different machine learning algorithms to build the most appropriate predictive model for classification of HTN. Overall experimental results anticipated that, among four models, the XGB model is the most appropriate model for predicting patient with the risk of HTN. The SHAP analysis revealed that age, weight, fat, income, BMI, diabetes, salt, HHTN, drinking, and smoking are the high contributing risk factors for developing HTN. Therefore, the proposed integrating system can be conveniently utilized as a useful tool in clinical sittings to accurately identify the patients with the risk of HTN at an early stage. With the help of this information, a doctor can make decisions that will reduce healthcare costs and time while also enabling individualized interventions and targeted treatment to minimize the burden of HTN in Ethiopia.

Supporting information

S1 Fig. ROC curve of four models with two class balancing methods, (a) under-sampling and (b) ADASYN.
(DOCX)

Acknowledgments

Authors would like to thanks the PLOS ONE's editor and reviewers for their valuable comments and suggestions to improve the quality of the manuscript.

Author Contributions

Conceptualization: Md. Merajul Islam, Md. Jahangir Alam, Md Maniruzzaman, Md. Jahanur Rahman.

Data curation: Md. Merajul Islam.

Formal analysis: Md. Merajul Islam, Md. Jahangir Alam.

Investigation: Md. Jahangir Alam.

Methodology: Md. Merajul Islam, Md. Jahangir Alam.

Resources: Md. Merajul Islam.

Software: Md. Merajul Islam, Md. Jahangir Alam.

Supervision: Md. Jahangir Alam, Md Maniruzzaman, Md. Jahanur Rahman, Dulal Chandra Roy.

Validation: Md. Jahangir Alam, Md Maniruzzaman.

Visualization: Md. Merajul Islam.

Writing – original draft: Md. Merajul Islam, Md. Jahangir Alam.

Writing – review & editing: Md. Merajul Islam, Md. Jahangir Alam, Md Maniruzzaman, N. A. M. Faisal Ahmed, Md Sujan Ali, Md. Jahanur Rahman, Dulal Chandra Roy.

References

1. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. Nature Reviews Nephrology. 2020; 16(4):223–37. https://doi.org/10.1038/s41581-019-0244-2 PMID: 32024986
2. GBD 2017 Risk Factor Collaborators. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018; 392:1923–94. https://doi.org/10.1016/S0140-6736(18)32225-6 PMID: 30496105
3. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018; 392:1736–88. https://doi.org/10.1016/S0140-6736(18)32203-7 PMID: 30496103
4. Gupta R, Xavier D. Hypertension: the most important non communicable disease risk factor in India. Indian heart journal. 2018; 70(4):565–72. https://doi.org/10.1016/j.ihj.2018.02.003 PMID: 30170654
5. Fuchs FD, Whelton PK. High blood pressure and cardiovascular disease. Hypertension. 2020; 75 (2):285–92. https://doi.org/10.1161/HYPERTENSIONAHA.119.14240 PMID: 31865786
6. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. Journal of the American College of Cardiology. 2020; 76(25):2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010 PMID: 33309175
7. Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah AD, Denaxas S, et al. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. The Lancet. 2014; 383(9932):1899–911.
8. Sorato MM, Davari M, Kebriaeezadeh A, Sarrafzadegan N, Shibru T. Societal economic burden of hypertension at selected hospitals in southern Ethiopia: a patient-level analysis. BMJ open. 2022; 12 (4):e056627. https://doi.org/10.1136/bmjopen-2021-056627 PMID: 35387822
9. Mehta R, Mantri N, Goel AD, Gupta MK, Joshi NK, Bhardwaj P. Out-of-pocket spending on hypertension and diabetes among patients reporting in a health-care teaching institute of the Western Rajasthan.

Journal of Family Medicine and Primary Care. 2022; 11(3):1083. https://doi.org/10.4103/jfmpc.jfmpc_998_21 PMID: 35495832

10. Berek PA, Irawati D, Hamid AY. Hypertension: A global health crisis. Ann Clin Hypertens. 2021; 5:8–11.
11. Belay DG, Fekadu H, Molla MD, Chekol HA, Adugna DG, Melese E, et al. Prevalence and associated factors of hypertension among adult patients attending the outpatient department at the primary hospitals of Wolkait tegedie zone, Northwest Ethiopia. Frontiers in Neurology. 2022; 13:943595. https://doi.org/10.3389/fneur.2022.943595 PMID: 36034276
12. Mamdouh H, Alnakhi WK, Hussain HY, Ibrahim GM, Hussein A, Mahmoud I, et al. Prevalence and associated risk factors of hypertension and pre-hypertension among the adult population: findings from the Dubai Household Survey, 2019. BMC Cardiovascular Disorders. 2022; 22(1):18. https://doi.org/10.1186/s12872-022-02457-4 PMID: 35090385
13. Tesfa E, Demeke D. Prevalence of and risk factors for hypertension in Ethiopia: A systematic review and meta-analysis. Health Science Reports. 2021; 4(3):e372. https://doi.org/10.1002/hsr2.372 PMID: 34589614
14. Anjulo U, Haile D, Wolde A. Prevalence of Hypertension and Its Associated Factors Among Adults in Areka Town, Wolaita Zone, Southern Ethiopia. Integrated Blood Pressure Control. 2021; 14:43–54. https://doi.org/10.2147/IBPC.S295574 PMID: 33758539
15. Damtie D, Bereket A, Bitew D, Kerisew B. The prevalence of hypertension and associated risk factors among secondary school teachers in Bahir Dar City administration, Northwest Ethiopia. International Journal of Hypertension. 2021; 2021:525802. https://doi.org/10.1155/2021/5525802 PMID: 33953969
16. Asresahegn H, Tadesse F, Beyene E. Prevalence and associated factors of hypertension among adults in Ethiopia: a community based cross-sectional study. BMC research notes. 2017; 10:1–8.
17. Khanam R, Ahmed S, Rahman S, Al Kibria GM, Syed JR, Khan AM, et al. Prevalence and factors associated with hypertension among adults in rural Sylhet district of Bangladesh: a cross-sectional study. BMJ open. 2019; 9(10):e026722. https://doi.org/10.1136/bmjopen-2018-026722 PMID: 31662350
18. Matsuzaki M, Sherr K, Augusto O, Kawakatsu Y, Ásbjörnsdóttir K, Chale F, et al. The prevalence of hypertension and its distribution by sociodemographic factors in Central Mozambique: a cross sectional study. BMC public health. 2020; 20:1–9.
19. Sharma JR, Mabhida SE, Myers B, Apalata T, Nicol E, Benjeddou M, et al. Prevalence of hypertension and its associated risk factors in a rural black population of Mthatha town, South Africa. International Journal of Environmental Research and Public Health. 2021; 18(3):1215. https://doi.org/10.3390/ijerph18031215 PMID: 33572921
20. Manios Y, Androutsos O, Lambrinou CP, Cardon G, Lindstrom J, Annemans L, et al. A school-and community-based intervention to promote healthy lifestyle and prevent type 2 diabetes in vulnerable families across Europe: design and implementation of the Feel4Diabetes-study. Public Health Nutrition. 2018; 21(17):3281–90. https://doi.org/10.1017/S1368980018002136 PMID: 30207513
21. Hong K, Yu ES, Chun BC. Risk factors of the progression to hypertension and characteristics of natural history during progression: A national cohort study. Plos one. 2020; 15(3):e0230538. https://doi.org/10.1371/journal.pone.0230538 PMID: 32182265
22. Chowdhury MZ, Naeem I, Quan H, Leung AA, Sikdar KC, O'Beirne M, et al. Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis. Plos one. 2022; 17(4):e0266334. https://doi.org/10.1371/journal.pone.0266334 PMID: 35390039
23. Chowdhury MZ, Leung AA, Sikdar KC, O'Beirne M, Quan H, Turin TC. Development and validation of a hypertension risk prediction model and construction of a risk score in a Canadian population. Scientific Reports. 2022; 12(1):12780. https://doi.org/10.1038/s41598-022-16904-x PMID: 35896590
24. Ghosh S, Kumar M. Prevalence and associated risk factors of hypertension among persons aged 15–49 in India: a cross-sectional study. BMJ open. 2019; 9(12):e029714. https://doi.org/10.1136/bmjopen-2019-029714 PMID: 31848161
25. Baştanlar Y, Özuysal M. Introduction to machine learning. miRNomics: MicroRNA biology and computational analysis. Humana Press. 2014:105–28.
26. Ghaderzadeh M, Asadi F, Hosseini A, Bashash D, Abolghasemi H, Roshanpour A. Machine learning in detection and classification of leukemia using smear blood images: a systematic review. Scientific Programming. 2021; 2021:1–4.
27. Ghaderzadeh M, Rebecca FE, Standring A. Comparing performance of different neural networks for early detection of cancer from benign hyperplasia of prostate. Applied Medical Informatics. 2013; 33 (3):45–54.
28. Salehnasab C, Hajifathali A, Asadi F, Parkhideh S, Kazemi A, Roshanpoor A, et al. An Intelligent Clinical Decision Support System for Predicting Acute Graft-versus-host Disease (aGvHD) following

Allogeneic Hematopoietic Stem Cell Transplantation. Journal of Biomedical Physics & Engineering. 2021; 11(3):345. https://doi.org/10.31661/jbpe.v0i0.2012-1244 PMID: 34189123

29. Kruppa J, Liu Y, Biau G, Kohler M, Koenig IR, Malley JD, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. Biometrical Journal. 2014; 56 (4):534–63. https://doi.org/10.1002/bimj.201300068 PMID: 24478134

30. Garavand A, Salehnasab C, Behmanesh A, Aslani N, Zadeh AH, Ghaderzadeh M. Efficient model for coronary artery disease diagnosis: a comparative study of several machine learning algorithms. Journal of Healthcare Engineering. 2022;2022. https://doi.org/10.1155/2022/5359540 PMID: 36304749

31. Nadim K, Ragab A, Ouali MS. Data-driven dynamic causality analysis of industrial systems using interpretable machine learning and process mining. Journal of Intelligent Manufacturing. 2023; 34(1):57–83.

32. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc. 2022.

33. Rezaianzadeh A, Dastoorpoor M, Sanaei M, Salehnasab C, Mohammadi MJ, Mousavizadeh A. Predictors of length of stay in the coronary care unit in patient with acute coronary syndrome based on data mining methods. Clinical Epidemiology and Global Health. 2020; 8(2):383–8.

34. Kumar A, Mayank J. Ensemble learning for AI developers. BA press: Berkeley, CA, USA. 2020.

35. Kurniawan R, Utomo B, Siregar KN, Ramli K, Besral B, Suhatril RJ, et al. Hypertension prediction using machine learning algorithm among Indonesian adults. IAES International Journal of Artificial Intelligence. 2023; 12(2): 776–84.

36. Visco V, Izzo C, Mancusi C, Rispoli A, Tedeschi M, Virtuoso N, et al. Artificial Intelligence in Hypertension Management: An Ace up Your Sleeve. Journal of Cardiovascular Development and Disease. 2023; 10(2):74. https://doi.org/10.3390/jcdd10020074 PMID: 36826570

37. Alsaleh MM, Allery F, Choi JW, Hama T, McQuillin A, Wu H, et al. Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: A systematic review. International Journal of Medical Informatics. 2023; 175:105088. https://doi.org/10.1016/j.ijmedinf.2023.105088 PMID: 37156169

38. Islam SM, Talukder A, Awal MA, Siddiqui MM, Ahamad MM, Ahammed B, et al. Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data from Three South Asian Countries. Frontiers in Cardiovascular Medicine. 2022; 9:839379. https://doi.org/10.3389/fcvm.2022.839379 PMID: 35433854

39. Paulose T, Nkosi ZZ, Endriyas M. Prevalence of hypertension and its associated factors in Hawassa city administration, Southern Ethiopia: Community based cross-sectional study. Plos one. 2022; 17(3): e0264679. https://doi.org/10.1371/journal.pone.0264679 PMID: 35231073

40. Park S. Ideal target blood pressure in hypertension. Korean Circulation Journal. 2019; 49(11):1002–9. https://doi.org/10.4070/kcj.2019.0261 PMID: 31646769

41. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics. 2022; 2:927312. https://doi.org/10.3389/fbinf.2022.927312 PMID: 36304293

42. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: logistic regression. Perspectives in clinical research. 2017; 8(3):148. https://doi.org/10.4103/picr.PICR_87_17 PMID: 28828311

43. Montesinos López OA, Montesinos López A, Crossa J. Fundamentals of Artificial Neural Networks and Deep Learning. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction. Cham: Springer International Publishing. 2022 (pp. 379–425).

44. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32.

45. Guang P, Huang W, Guo L, Yang X, Huang F, Yang M, et al. Blood-based FTIR-ATR spectroscopy coupled with extreme gradient boosting for the diagnosis of type 2 diabetes: A STARD compliant diagnosis research. Medicine. 2020; 99(15). https://doi.org/10.1097/MD.0000000000019657 PMID: 32282717

46. May RJ, Maier HR, Dandy GC. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Networks. 2010; 23(2):283–94. https://doi.org/10.1016/j.neunet.2009.11.009 PMID: 19959327

47. Thabtah F.; Hammoud S.; Kamalov F.; Gonsalves A. Data imbalance in classification: Experimental evaluation. Inf. Sci. 2020; 513:429–441.

48. Buda M.; Maki A.; Mazurowski M.A. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks. 2018; 106:249–259. https://doi.org/10.1016/j.neunet.2018.07.011 PMID: 30092410

49. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE. 2008 (pp. 1322–1328).

50. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine. 2013; 4(2):627. PMID: 24009950

51. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

52. Shapley LS. 17. A value for n-person games. InContributions to the Theory of Games (AM-28). Princeton University Press. 2016: 307–318.

53. Palatnik de Sousa I, Maria Bernardes Rebuzzi Vellasco M, Costa da Silva E. Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors. 2019; 19(13):2969. https://doi.org/10.3390/s19132969 PMID: 31284419

54. Chowdhury MZ, Leung AA, Walker RL, Sikdar KC, O'Beirne M, Quan H, et al. A comparison of machine learning algorithms and traditional regression-based statistical modeling for predicting hypertension incidence in a Canadian population. Scientific Reports. 2023; 13(1):1–3.

55. Oanh TT, Tung NT. Predicting Hypertension Based on Machine Learning Methods: A Case Study in Northwest Vietnam. Mobile Networks and Applications. 2022; 27(5):2013–23.

56. Chai SS, Goh KL, Cheah WL, Chang YH, Ng GW. Hypertension Prediction in Adolescents Using Anthropometric Measurements: Do Machine Learning Models Perform Equally Well? Applied Sciences. 2022; 12(3):1600.

57. Islam MM, Rahman MJ, Roy DC, Tawabunnahar M, Jahan R, Ahmed NF, et al. Machine learning algorithm for characterizing risks of hypertension, at an early stage in Bangladesh. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2021; 15(3):877–884. https://doi.org/10.1016/j.dsx.2021.03.035 PMID: 33892404

58. Zheng J, Yu Z. A novel machine learning-based systolic blood pressure predicting model. Journal of Nanomaterials. 2021; 2021:1–8.

59. AlKaabi LA, Ahmed LS, Al Attiyah MF, Abdel-Rahman ME. Predicting hypertension using machine learning: Findings from Qatar Biobank Study. Plos One. 2020; 15(10):e0240370. https://doi.org/10.1371/journal.pone.0240370 PMID: 33064740

60. Legese N, Tadiwos Y. Epidemiology of hypertension in Ethiopia: a systematic review. Integrated blood pressure control. 2020; 13:135–43. https://doi.org/10.2147/IBPC.S276089 PMID: 33116810

61. Koya SF, Pilakkadavath Z, Chandran P, Wilson T, Kuriakose S, Akbar SK, et al. Hypertension control rate in India: Systematic review and meta-analysis of population-level non-interventional studies, 2001–2022. The Lancet Regional Health-Southeast Asia. 2023; 9:100113. https://doi.org/10.1016/j.lansea.2022.100113 PMID: 37383035

62. Solomon M, Shiferaw BZ, Tarekegn TT, GebreEyesus FA, Mengist ST, Mammo M, et al. Prevalence and Associated Factors of Hypertension Among Adults in Gurage Zone, Southwest Ethiopia, 2022. SAGE Open Nursing. 2023; 9:2377960823115347 https://doi.org/10.1177/23779608231153473 PMID: 36761364

63. Qin Z, Li C, Qi S, Zhou H, Wu J, Wang W, et al. Association of socioeconomic status with hypertension prevalence and control in Nanjing: a cross-sectional study. BMC Public Health. 2022; 22(1):1–9.

64. Ranzani OT, Kalra A, Di Girolamo C, Curto A, Valerio F, Halonen JI, et al. Urban-rural differences in hypertension prevalence in low-income and middle-income countries, 1990–2020: A systematic review and meta-analysis. Plos Medicine. 2022; 19(8):e1004079. https://doi.org/10.1371/journal.pmed.1004079 PMID: 36007101

65. Hall JE, do Carmo JM, da Silva AA, Wang Z, Hall ME. Obesity, kidney dysfunction and hypertension: mechanistic links. Nature reviews nephrology. 2019; 15(6):367–85. https://doi.org/10.1038/s41581-019-0145-4 PMID: 31015582

66. Imai Y. A personal history of research on hypertension from an encounter with hypertension to the development of hypertension practice based on out-of-clinic blood pressure measurements. Hypertension Research. 2022; 45(11):1726–42. https://doi.org/10.1038/s41440-022-01011-1 PMID: 36075990

67. Mayl JJ, German CA, Bertoni AG, Upadhya B, Bhave PD, Yeboah J, et al. Association of alcohol intake with hypertension in type 2 diabetes mellitus: The ACCORD Trial. Journal of the American Heart Association. 2020; 9(18):e017334. https://doi.org/10.1161/JAHA.120.017334 PMID: 32900264

68. Nguyen TT, Nguyen MH, Nguyen YH, Nguyen TT, Giap MH, Tran TD, et al. Body mass index, body fat percentage, and visceral fat as mediators in the association between health literacy and hypertension among residents living in rural and suburban areas. Frontiers in Medicine. 2022;9. https://doi.org/10.3389/fmed.2022.877013 PMID: 36148456

69. Choi JW, Han E, Kim TH. Risk of Hypertension and Type 2 Diabetes in Relation to Changes in Alcohol Consumption: A Nationwide Cohort Study. International Journal of Environmental Research and Public Health. 2022; 19(9):4941. https://doi.org/10.3390/ijerph19094941 PMID: 35564335