CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 2566–2583
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Explainable Machine Learning for Hypertension Status Classification Using NHANES: Performance and Cross-Model SHAP Analysis

Diska Nur Isna Apriani[1] and A'yunin Sofro*[2]

[1] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Indonesia*
[2] *Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Indonesia*

Abstract

Hypertension is a major noncommunicable disease and an important public health concern. Its health burden is relevant to public-health risk management and broader disaster-risk contexts, supporting Sustainable Development Goal 3 (SDG 3), particularly Target 3.4 on reducing premature mortality from noncommunicable diseases. This study compares Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) for classifying reported hypertension status, evaluates hyperparameter optimization, and examines predictor contributions using SHapley Additive exPlanations (SHAP). Using 5,008 NHANES 2021–2023 respondents and eight predictors, hypertension status was defined from BPQ020 as self-reported physician diagnosis. The data were divided using an 80:20 stratified split into training and independent test sets, with five-fold cross-validation applied to the training set for hyperparameter optimization. At a fixed threshold of 0.50, LR achieved the highest baseline ROC-AUC (0.7921), while XGBoost achieved the highest baseline Recall (0.5968) and F1-Score (0.6233). Tuning improved RF Recall from 0.5332 to 0.6048 and F1-Score from 0.5793 to 0.6316. SHAP identified Age, Waist circumference, and BMI as the top three predictors in both tuned models, with a Spearman rank correlation of 0.9286 and 100% top-three agreement. Bootstrap analysis yielded a median correlation of 0.9286 (95% CI: 0.9048–1.0000), with 100% top-three agreement across 2,000 replications. At the fixed threshold, tuned RF achieved the highest Recall and F1-Score, whereas LR yielded the highest ROC-AUC.

Keywords: Hypertension; Machine Learning; NHANES; SHAP; SDG 3.

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (https://creativecommons.org/licenses/by-sa/4.0)

1. Introduction

Hypertension is one of the major noncommunicable diseases worldwide and an important risk factor for cardiovascular disease, stroke, kidney disorders, and premature mortality [1, 2]. Its substantial public health burden highlights the importance of prevention, risk assessment, and continuity of health-system capacity. Within a broader disaster-risk management perspective, these considerations are relevant because health systems are required to maintain prevention,

*Corresponding author. E-mail: ayuninsofro@unesa.ac.id

Received: August 24, 2026 Revised: October 22, 2026 Accepted: October 28, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.45787

preparedness, and essential health services in the presence of population-level health risks. Hypertension should therefore be understood as a public-health risk within this broader context rather than as a non-natural disaster itself. This perspective is aligned with Sustainable Development Goal 3 (SDG 3), particularly Target 3.4, which calls for reducing premature mortality from noncommunicable diseases through prevention and treatment [3].

Hypertension is associated with demographic, socioeconomic, and anthropometric characteristics [1, 2]. Logistic Regression (LR) provides an interpretable approach for binary health outcomes [4], but its additive linear structure may not adequately capture nonlinear relationships and interactions. Machine learning methods offer more flexible alternatives, with Random Forest (RF) and Extreme Gradient Boosting (XGBoost) showing potential for hypertension and related health prediction tasks [5–7]. Recent studies have also combined machine learning with explainable artificial intelligence (XAI), including applications of XGBoost and SHAP to NHANES data [8], interpretable machine learning models for hypertension using NHANES data [9], and comparative machine learning approaches with SHAP-based identification of important predictors [10].

Despite these developments, predictor importance is generally interpreted within individual models, with limited quantitative assessment of whether different ensemble algorithms identify consistent predictor rankings when applied to the same population and predictor set. Sofro et al. [11], for example, compared LR and machine learning approaches using sociodemographic and anthropometric characteristics among athletes, but predictive performance differences alone do not establish whether different models rely on similar predictors. This motivates a cross-model analysis that considers both classification performance and quantitative consistency of predictor importance.

Accordingly, this study uses the 2021–2023 National Health and Nutrition Examination Survey (NHANES), a population-based survey containing demographic, socioeconomic, and health-related information [12]. Logistic Regression is used as a conventional statistical baseline, while RF and XGBoost are evaluated before and after hyperparameter optimization. Hypertension status is defined using BPQ020 as self-reported physician diagnosis. SHapley Additive exPlanations (SHAP) are then applied to the optimized RF and XGBoost models to examine predictor contributions, with cross-model consistency quantified using Spearman rank correlation and top-ranked feature agreement [13].

Accordingly, this study aims to: (1) compare the classification performance of LR, RF, and XGBoost for reported hypertension status; (2) evaluate the effect of hyperparameter optimization on RF and XGBoost; and (3) examine the consistency of predictor importance between the optimized RF and XGBoost models using SHAP-based interpretation and quantitative rank agreement. The study integrates comparative model evaluation, hyperparameter optimization, and cross-model SHAP analysis while identifying predictor patterns that remain consistent across the optimized ensemble models. Because the outcome is based on self-reported physician diagnosis, the findings are interpreted as classification of reported hypertension status rather than clinical diagnosis or direct blood-pressure screening.

2. Methods

This section describes the data and variable construction used for hypertension status classification. The statistical models include Logistic Regression, Random Forest, and XGBoost, followed by hyperparameter optimization for the two ensemble models. Performance evaluation, uncertainty quantification, and SHAP-based interpretation were then used to assess predictive performance and cross-model consistency of predictor importance.

2.1. Research Data and Variables

This study used secondary data from the National Health and Nutrition Examination Survey (NHANES) 2021–2023, integrating the Demographics, Body Measures, Blood Pressure, and Blood

Pressure Questionnaire components using the respondent sequence number (SEQN) [12]. The integration yielded 5,010 observations for respondents aged ≥ 18 years with available hypertension status based on BPQ020. After excluding two observations with missing Education values, 5,008 respondents remained for analysis. Eight predictors were considered: Age, Gender, Education, Income-to-poverty ratio, Height, Weight, BMI, and Waist circumference. The data integration and participant selection process is summarized in Figure 1, and the handling of missing values and variable coding is described in Section 2.2.

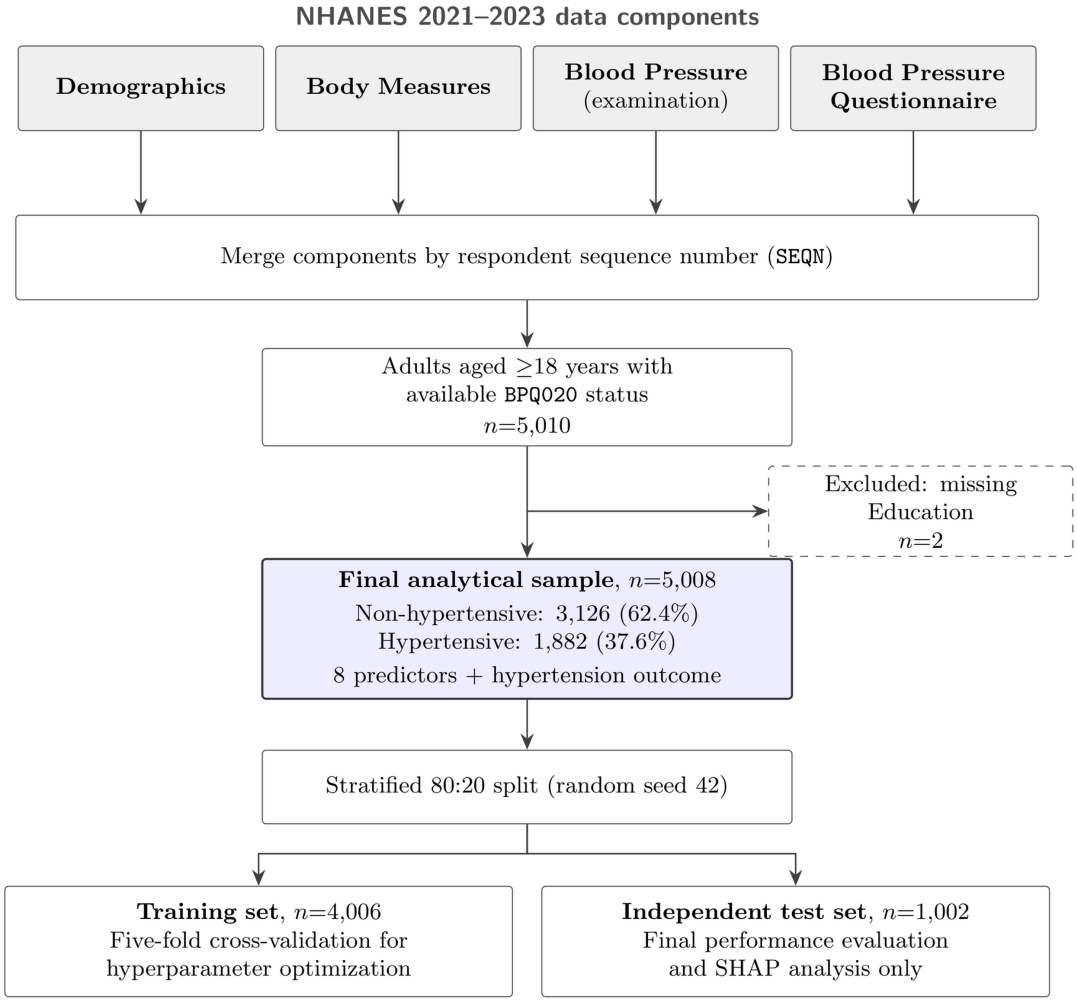


Fig. 1: Flowchart of data integration, participant selection, and data partitioning.

Hypertension status was defined as $Y = 1$ for self-reported physician-diagnosed hypertension and $Y = 0$ for no reported diagnosis, based on BPQ020: "Has a doctor or other health professional ever told you that you had hypertension, also called high blood pressure?" [12]. This questionnaire-based definition was adopted because it provides a reproducible binary outcome based on prior clinical recognition rather than a single blood-pressure measurement obtained during the survey. Measured blood pressure and antihypertensive medication variables were not incorporated so that the outcome remained consistently defined by BPQ020. Consequently, respondents with undiagnosed hypertension may be classified as non-hypertensive, whereas respondents with a previous diagnosis remain classified as hypertensive even if their blood pressure is currently controlled. Thus, the outcome represents reported diagnosed hypertension status rather than current physiological blood-pressure status.

NHANES survey weights, strata, and primary sampling unit (PSU) variables were not incorporated because the objective was to compare the internal predictive performance of classification algorithms on the observed sample rather than to estimate nationally representative

hypertension prevalence. Accordingly, the reported model performance and feature-importance results represent an internal predictive analysis within the analyzed NHANES sample. They should not be interpreted as nationally representative or survey-weighted population estimates.

2.2. Data Preprocessing

Data preprocessing included missing-value handling, validity and outlier assessment, feature scaling, and stratified partitioning. Missingness was assessed for all variables prior to analysis. The only variable with missing values was Education ($n = 2$, 0.04%), while SEQN, Age, Gender, Income-to-poverty ratio, Height, Weight, BMI, Waist circumference, and BPQ020 had no missing values. The two observations with missing Education values were excluded during data preprocessing, yielding a final analytical sample of 5,008 respondents with complete data.

Height, Weight, BMI, and Waist circumference were assessed for invalid values, and no zero, negative, or otherwise invalid values were identified. Potential outliers were assessed using the interquartile range (IQR), which is defined in Eq. (1) as

$$\mathrm{IQR} = Q_3 - Q_1. \tag{1}$$

An observation x_i was considered a potential outlier when it satisfied the lower-bound rule in Eq. (2),

$$x_i < Q_1 - 1.5(\mathrm{IQR}), \tag{2}$$

or the upper-bound rule in Eq. (3),

$$x_i > Q_3 + 1.5(\mathrm{IQR}). \tag{3}$$

Observations meeting either rule were treated as potential outliers for inspection rather than being removed automatically. Identified observations were retained when their values were considered plausible and not indicative of data-entry errors.

Gender was coded as 1 = male and 2 = female, while Education was treated as an ordinal numerical predictor coded from 1 to 5. Code 9 ("Don't know") was treated as missing and excluded. These variables retained their original coding and were not standardized. Standardization was applied only to the predictors used in Logistic Regression (Section 2.3), as given in Eq. (4),

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{4}$$

where μ_j and σ_j were estimated from the training data and subsequently applied unchanged to the test data to prevent information leakage. Random Forest and XGBoost used the original predictor scales.

The final sample comprised 3,126 respondents without reported hypertension and 1,882 respondents with reported hypertension. The data were divided using an 80:20 stratified split, yielding 4,006 training and 1,002 test observations. The test set was reserved exclusively for final evaluation, while hyperparameter optimization was conducted only on the training data, as described in Section 2.6. A fixed random seed of 42 was used for reproducibility, and no class weighting or resampling was applied. All analyses were implemented in Python using scikit-learn (version 1.9.0), XGBoost (version 3.3.0), and SHAP (version 0.52.0).

BMI, Height, Weight, and Waist circumference were retained jointly despite their related anthropometric information. This relationship was considered when interpreting feature importance and SHAP contributions, as detailed in Section 2.8. Accordingly, their contributions were interpreted as model-based predictive contributions rather than as independent or causal effects.

2.3. Logistic Regression

Logistic Regression (LR) was employed as the conventional statistical baseline for binary classification [4]. Let $Y_i \in \{0, 1\}$ denote the hypertension status of respondent i, and let

$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ denote the corresponding predictor vector, where $p = 8$. The conditional probability of hypertension is modeled in Eq. (5),

$$\pi_i = P(Y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \tag{5}$$

where the linear predictor η_i is defined in Eq. (6),

$$\eta_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}. \tag{6}$$

Equivalently, the model can be expressed through the logit transformation in Eq. (7),

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}. \tag{7}$$

Thus, Logistic Regression assumes a linear relationship between the predictors and the log-odds of hypertension, while the logistic function maps the linear predictor into the interval $(0, 1)$ [4]. Conditional on $\mathbf{x}_i$, the binary response Y_i follows a Bernoulli distribution with parameter π_i. This additive linear structure makes the model straightforward to interpret, but nonlinear relationships and interactions are not represented unless they are specified explicitly, which motivates the comparison with Random Forest and XGBoost.

For n independent observations with observed outcomes y_i, the likelihood function is given in Eq. (8),

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}, \tag{8}$$

and the corresponding log-likelihood is given in Eq. (9),

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \right]. \tag{9}$$

The regression coefficients are then estimated according to the maximum likelihood principle in Eq. (10),

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \ell(\boldsymbol{\beta}). \tag{10}$$

The estimated coefficient $\hat{\beta}_k$ represents the change in the log-odds associated with a one-unit increase in predictor X_k, conditional on the other predictors, and its corresponding odds ratio is given in Eq. (11),

$$\mathrm{OR}_k = \exp(\hat{\beta}_k). \tag{11}$$

For classification, the estimated probability $\hat{\pi}_i$ was converted into a binary prediction using the threshold $\tau = 0.50$, as defined in Eq. (12),

$$\hat{Y}_i = \begin{cases} 1, & \hat{\pi}_i \geq 0.50, \\ 0, & \hat{\pi}_i < 0.50. \end{cases} \tag{12}$$

Logistic Regression was retained as the conventional statistical baseline and was not subjected to hyperparameter optimization (Section 2.6). The standardized predictors were used for model fitting, with the transformation given in Eq. (4). The optimization procedure was allowed to run for up to 1,000 iterations, and a fixed random seed of 42 was used to ensure reproducibility. All other model settings were retained at their default values in scikit-learn.

2.4. Random Forest

Random Forest (RF) is an ensemble classification method that combines multiple decision trees constructed using randomized samples of observations and predictor variables [5]. Let $h_t(\mathbf{x})$ denote the prediction of the t-th decision tree, where $t = 1, \ldots, T$ and T represents the total number of trees. Each tree is constructed from a bootstrap sample of the training data, while a randomly selected subset of predictors is considered at each splitting node. The final RF classification is determined through majority voting, as shown in Eq. (13),

$$\hat{Y}(\mathbf{x}) = \text{mode}\left\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_T(\mathbf{x})\right\}. \tag{13}$$

For binary classification, the voting rule in Eq. (13) can be written as Eq. (14),

$$\hat{Y}(\mathbf{x}) = \mathbb{I}\left(\sum_{t=1}^{T} h_t(\mathbf{x}) > \frac{T}{2}\right), \tag{14}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The randomization among observations and predictors introduces diversity among individual trees and reduces their correlation, thereby improving the stability of the aggregated prediction [5]. Unlike Logistic Regression, RF does not impose a linear relationship between predictors and the response, allowing nonlinear relationships and interactions to be represented through recursive partitioning.

In this study, RF was first evaluated using its baseline configuration to establish its initial performance and was subsequently optimized through five-fold cross-validation, as described in Section 2.6. The baseline RF configuration is reported in Table 1. A fixed random seed of 42 was used throughout the analysis to ensure reproducibility.

2.5. XGBoost

XGBoost is a gradient-boosting ensemble method that constructs decision trees sequentially, with each newly added tree contributing to the prediction of the existing ensemble [6]. Let $f_t(\mathbf{x})$ denote the tree added at iteration t and $s_i^{(t)}$ the ensemble score of observation i after t iterations. The score at iteration t is updated as in Eq. (15),

$$s_i^{(t)} = s_i^{(t-1)} + f_t(\mathbf{x}_i), \tag{15}$$

where $s_i^{(t-1)}$ denotes the score from the previous iteration and $f_t(\mathbf{x}_i)$ represents the contribution of the newly added tree. After T boosting iterations, the ensemble score is the sum of all tree contributions, as given in Eq. (16),

$$s_i = \sum_{t=1}^{T} f_t(\mathbf{x}_i). \tag{16}$$

The learning objective at iteration t is formulated in Eq. (17),

$$\mathcal{J}^{(t)} = \sum_{i=1}^{n} l\left(y_i,\, s_i^{(t)}\right) + \Omega(f_t), \tag{17}$$

where $l(\cdot)$ denotes the loss function and $\Omega(f_t)$ represents a regularization term controlling the complexity of the newly added tree [6]. For binary classification, the ensemble score is transformed into a probability using the logistic function in Eq. (18),

$$\tilde{\pi}_i = P(Y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp(-s_i)}. \tag{18}$$

XGBoost was initially evaluated using its baseline configuration and subsequently optimized through five-fold cross-validation, as described in Section 2.6. The baseline XGBoost configuration is reported in Table 1. A fixed random seed of 42 was used throughout the analysis to ensure reproducibility.

2.6. Hyperparameter Optimization

Hyperparameter optimization was performed for Random Forest (RF) and XGBoost, while Logistic Regression was retained as the baseline model without tuning [14]. Let $\mathcal{D}_{\text{train}}$ denote the training set and Θ the corresponding hyperparameter space. Five-fold cross-validation was conducted within the training set to identify the configuration that provided the highest mean ROC-AUC. For a configuration $\theta \in \Theta$, the cross-validated performance was defined in Eq. (19) as

$$\overline{\text{AUC}}(\theta) = \frac{1}{5}\sum_{k=1}^{5} \text{AUC}_k(\theta), \tag{19}$$

where $\text{AUC}_k(\theta)$ denotes the ROC-AUC obtained on the k-th validation fold. The selected configuration was therefore determined by Eq. (20),

$$\theta^* = \arg\max_{\theta\in\Theta} \overline{\text{AUC}}(\theta). \tag{20}$$

The baseline and tuned hyperparameters considered for RF and XGBoost are summarized in Table 1. Each hyperparameter was varied only within the predefined search space listed in Table 1, and the configurations selected by cross-validation are reported in the same table. Parameters not listed in Table 1 were retained at their respective library default values.

Table 1: Hyperparameter search space and selected configurations.

Model	Hyperparameter	Baseline value	Search space	Selected value
Random Forest	Number of trees	100	100, 200, 300	300
	Maximum depth	None	None, 10, 20	10
	Minimum samples for split	2	2, 5	2
	Minimum samples per leaf	1	1, 2	2
XGBoost	Number of boosting iterations	100	100, 200, 300	100
	Maximum depth	3	3, 5	3
	Learning rate	0.10	0.05, 0.10	0.05
	Subsampling ratio	1.0	0.8, 1.0	0.8

For Random Forest, 36 candidate configurations were evaluated, whereas 24 configurations were considered for XGBoost. The selected Random Forest configuration achieved a mean cross-validated ROC-AUC of 0.7580, while the selected XGBoost configuration achieved 0.7639. ROC-AUC was used as the optimization criterion because it evaluates discrimination across classification thresholds and is not restricted to a single operating point. Recall and F1-Score were retained as complementary evaluation measures, with Recall reflecting the ability to correctly classify respondents with reported hypertension and F1-Score representing the balance between Precision and Recall. The selected configurations were subsequently refitted using the complete training set and evaluated on the independent test set using the measures defined in Section 2.7.

2.7. Model Evaluation

Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Let TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative classifications, respectively. Accuracy is defined in Eq. (21), Precision in Eq. (22), Recall in Eq. (23), and F1-Score in Eq. (24):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{21}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{22}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{23}$$

and

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{24}$$

A fixed classification threshold of $\tau = 0.50$ was used to convert predicted probabilities into class labels for these threshold-dependent metrics. Accordingly, Recall and F1-Score represent model performance at this specific operating point and should not be interpreted as threshold-independent measures of overall model performance. Recall was emphasized because it measures the proportion of respondents with reported hypertension who were correctly classified as positive.

ROC-AUC was used to assess discrimination across classification thresholds. The receiver operating characteristic (ROC) curve is defined by the true-positive rate (TPR) in Eq. (25) and the false-positive rate (FPR) in Eq. (26). These two rates are calculated using the numbers of true positives, false negatives, false positives, and true negatives, as shown below:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{25}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{26}$$

Higher ROC-AUC indicates greater discrimination between respondents with and without reported hypertension. Note that TPR in Eq. (25) is numerically identical to Recall in Eq. (23), evaluated across all thresholds rather than at $\tau = 0.50$ only. Therefore, ROC-AUC summarizes discrimination across thresholds, whereas Recall at $\tau = 0.50$ measures the proportion of actual hypertensive respondents correctly identified at that specific threshold.

Uncertainty in the principal performance measures was quantified using 2,000 bootstrap resamples of the independent test set with replacement. The fitted models were held fixed, and the 2.5th and 97.5th percentiles of the bootstrap distributions were used to construct 95% percentile bootstrap confidence intervals for ROC-AUC, Recall, and F1-Score. Pairwise differences in ROC-AUC among the final models were assessed using DeLong's test with Holm adjustment for multiple comparisons. The test set was used only for final performance evaluation and was not involved in model fitting or hyperparameter optimization. The same resampling scheme was applied to the SHAP-based rankings, as described in Section 2.8.

2.8. SHAP Analysis

SHapley Additive exPlanations (SHAP) was used to interpret the optimized Random Forest (RF) and XGBoost models obtained in Section 2.6 [13]. SHAP values corresponding to the positive class (class 1) were obtained using tree-based explainers appropriate for the fitted models. For a model output $f(\mathbf{x})$, the SHAP decomposition is expressed in Eq. (27),

$$f(\mathbf{x}) = E[f(\mathbf{X})] + \sum_{j=1}^{p} \phi_j, \tag{27}$$

where $E[f(\mathbf{X})]$ is the expected model output and ϕ_j denotes the contribution of predictor j. Global feature importance was measured using the mean absolute SHAP value in Eq. (28),

$$I_j = \frac{1}{n} \sum_{i=1}^{n} |\phi_{ij}|, \tag{28}$$

where I_j represents the average magnitude of predictor j's contribution across the n test observations. The same independent test-set observations were used for both models to ensure comparable feature rankings. Predictors were then ordered by decreasing I_j within each model, and these orderings formed the basis of the cross-model comparison described below.

For RF, SHAP values were evaluated on the probability scale, whereas for XGBoost they were evaluated on the default raw output scale (log-odds/margin). Therefore, absolute SHAP

magnitudes were not directly compared between models; cross-model consistency was assessed quantitatively using feature rankings and top-three agreement, while contribution directions were compared descriptively. Ranking agreement was quantified using the Spearman rank correlation coefficient in Eq. (29),

$$\rho_s = 1 - \frac{6 \sum_{j=1}^{p} d_j^2}{p(p^2 - 1)}, \tag{29}$$

where d_j denotes the difference between the ranks assigned to predictor j by the two models. Top-k agreement was assessed using Eq. (30),

$$A_k = \frac{|S_{\mathrm{RF}}^{(k)} \cap S_{\mathrm{XGB}}^{(k)}|}{k}, \tag{30}$$

where $S_{\mathrm{RF}}^{(k)}$ and $S_{\mathrm{XGB}}^{(k)}$ denote the sets of the k highest-ranked predictors. In this study, agreement was evaluated for $k = 3$, so that $A_3 = 1$ indicates that both models selected exactly the same three predictors. Values of ρ_s close to 1 indicate similar complete rankings, whereas A_k focuses specifically on the most influential predictors.

To assess the robustness of the cross-model SHAP ranking, 2,000 bootstrap resamples of the independent test set were generated with replacement while keeping the fitted models fixed. The same bootstrap indices were applied to the SHAP values of both models to preserve paired observations. For each resample, mean absolute SHAP values were used to obtain predictor rankings and their Spearman rank correlation. The 2.5th and 97.5th percentiles of the bootstrap distribution defined the 95% percentile bootstrap interval. Top-three agreement and feature-level rank stability were also assessed across the bootstrap replications.

Because Height, Weight, BMI, and Waist circumference are interrelated anthropometric predictors, SHAP values were interpreted as model-based predictive contributions rather than independent or causal effects. In particular, BMI is mathematically derived from Height and Weight, while Waist circumference provides related anthropometric information. Therefore, SHAP contributions may reflect shared or overlapping information.

3. Results and Discussion

This section presents the results of hypertension status classification using NHANES 2021–2023 data. The analysis covers the characteristics of the study data, baseline performance of the three models, and the effect of hyperparameter tuning together with bootstrap and DeLong comparisons. Cross-model SHAP results are then used to examine the consistency of predictor importance at the global and individual levels.

3.1. Characteristics of the Study Data

Before evaluating the classification models, the characteristics of the analytical dataset were examined to describe the study population and distribution of the outcome variable. The final dataset consisted of 5,008 respondents from the 2021–2023 NHANES cycle after the data construction and preprocessing described in Sections 2.1 and 2.2. The outcome variable represented reported hypertension status based on self-reported physician diagnosis. Among the respondents, 3,126 individuals (62.4%) were classified as non-hypertensive, whereas 1,882 individuals (37.6%) were classified as hypertensive, as presented in Table 2. Thus, the dataset contained a moderate imbalance between the two outcome classes, with the non-hypertensive group representing the majority of observations. The predictor variables comprised eight demographic, socioeconomic, and anthropometric characteristics: Age, Gender, Education, Income-to-poverty ratio, Height, Weight, BMI, and Waist circumference. These variables were subsequently used as predictors in the comparative classification analysis using Logistic Regression, Random Forest, and XGBoost.

Table 2: Distribution of hypertension status in the study data.

Hypertension Status	Frequency	Percentage
Non-hypertensive	3,126	62.4
Hypertensive	1,882	37.6
Total	5,008	100.0

3.2. Baseline Comparison of Logistic Regression, Random Forest, and XGBoost

Table 3 presents the performance of Logistic Regression, Random Forest, and XGBoost using their baseline configurations (Section 2.6) on the independent test set. Logistic Regression achieved the highest Accuracy (0.7335), Precision (0.6809), and ROC-AUC (0.7921), whereas XGBoost obtained the highest Recall (0.5968) and F1-Score (0.6233). Random Forest yielded the lowest ROC-AUC (0.7611) and F1-Score (0.5793).

Table 3: Comparison of model performance before hyperparameter tuning.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.7335	0.6809	0.5491	0.6079	0.7921
Random Forest	0.7086	0.6341	0.5332	0.5793	0.7611
XGBoost	0.7285	0.6522	0.5968	0.6233	0.7883

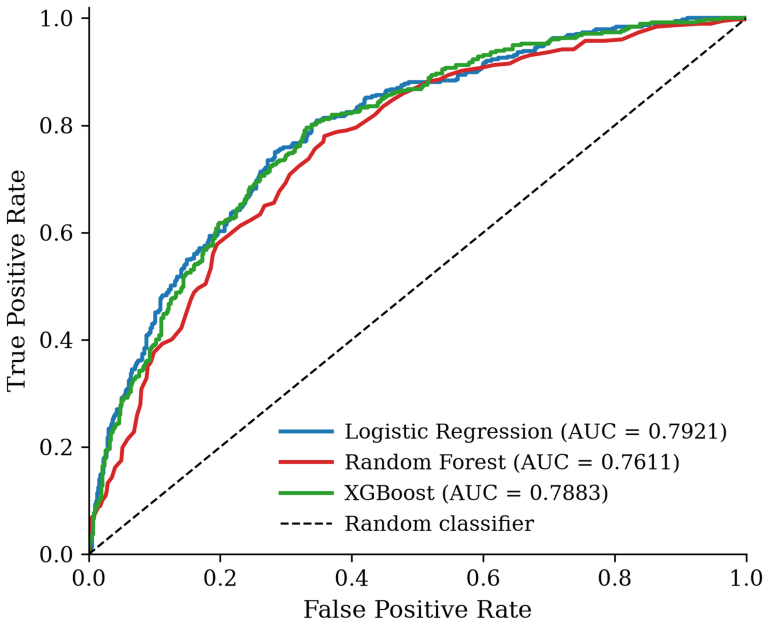


Fig. 2: ROC curves of Logistic Regression, Random Forest, and XGBoost before hyperparameter tuning.

Figure 2 shows that Logistic Regression achieved the highest baseline ROC-AUC (0.7921), indicating the highest numerical discrimination among the baseline models on the independent test set. The baseline ROC-AUC values were broadly comparable with previous hypertension prediction studies. Huang and Huang [8], using NHANES 2017–2020 data with XGBoost and SHAP interpretation, reported an AUROC of 0.84, while a study using the China Health and Nutrition Survey reported an AUC of 0.784 for Logistic Regression [10]. The ROC-AUC values in the present study (0.7611–0.7921) fall within a similar range. However, direct numerical comparisons should be interpreted cautiously because of differences in populations, outcome definitions, predictor sets, and validation procedures.

Although XGBoost obtained a slightly lower ROC-AUC than Logistic Regression (0.7883 versus 0.7921), it achieved higher Recall (0.5968 versus 0.5491) and F1-Score (0.6233 versus 0.6079). Thus, the baseline results varied according to the evaluation criterion rather than

indicating a uniformly superior model. As defined in Eq. (23), Recall measures the proportion of respondents with reported hypertension who were correctly classified as positive. XGBoost achieved a higher Recall at the fixed classification threshold of $\tau = 0.50$, indicating that it correctly classified a larger proportion of respondents with reported hypertension as positive. This difference represents performance at a specific operating point and should not be interpreted as threshold-independent evidence of overall model performance. Because the outcome represents reported physician-diagnosed hypertension, these results should be interpreted as empirical classification performance rather than evidence of clinical screening suitability. The baseline results provide a basis for examining whether hyperparameter optimization improves the performance of the ensemble models. Accordingly, Random Forest and XGBoost were subsequently optimized using the predefined search spaces and ROC-AUC as the model-selection criterion.

3.3. Effect of Hyperparameter Tuning and Model Comparison

The performance of the baseline and tuned models was evaluated on the independent test set to assess the effect of the hyperparameter optimization described in Section 2.6. Accuracy, Precision, Recall, F1-Score, and ROC-AUC were considered to provide a comprehensive assessment of predictive performance across the models. The comparative results of the baseline and tuned models are presented in Table 4.

Table 4: Comparative performance of baseline and tuned classification models on the independent test set.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.7335	0.6809	0.5491	0.6079	0.7921
Random Forest	0.7086	0.6341	0.5332	0.5793	0.7611
Tuned Random Forest	0.7345	0.6609	0.6048	0.6316	0.7823
XGBoost	0.7285	0.6522	0.5968	0.6233	0.7883
Tuned XGBoost	0.7275	0.6512	0.5942	0.6214	0.7905

Hyperparameter tuning produced a more pronounced improvement for Random Forest, particularly in Recall and F1-Score. Its Accuracy increased from 0.7086 to 0.7345, Recall from 0.5332 to 0.6048, F1-Score from 0.5793 to 0.6316, and ROC-AUC from 0.7611 to 0.7823. In contrast, tuning XGBoost resulted in only a marginal increase in ROC-AUC from 0.7883 to 0.7905, while its threshold-dependent measures changed slightly. These results indicate that the effect of hyperparameter optimization varied across models and that improvement in one performance criterion did not necessarily translate into improvement across all criteria.

Among the final models, tuned Random Forest achieved the highest Accuracy (0.7345), Recall (0.6048), and F1-Score (0.6316), whereas Logistic Regression obtained the highest Precision (0.6809) and ROC-AUC (0.7921). Thus, the observed model performance varied across evaluation criteria rather than indicating a single uniformly superior model. The higher Recall of tuned Random Forest indicates that, at the fixed classification threshold of 0.50, it correctly classified a greater proportion of respondents with reported hypertension as positive. This result represents performance at a specific operating point and should not be interpreted as threshold-independent evidence of overall model performance.

The slightly higher ROC-AUC of Logistic Regression compared with the tuned ensemble models indicates that, for this dataset and the evaluated configurations, the ensemble approaches did not achieve higher numerical discrimination. The larger performance change observed for Random Forest after tuning suggests that its performance was more affected by the evaluated hyperparameter configuration in this dataset than that of XGBoost. These findings are specific to the present dataset and should not be interpreted as evidence of general superiority of one modeling approach.

To quantify uncertainty in the selected performance measures, 95% percentile bootstrap confidence intervals based on 2,000 resamples of the independent test set (Section 2.7) are

presented in Table 5. The point estimates correspond to the observed performance on the independent test set. The intervals were computed for ROC-AUC, Recall, and F1-Score because these measures represent the principal comparison criteria of the study.

Table 5: Bootstrap 95% confidence intervals for selected performance metrics.

Model	ROC-AUC (95% CI)	Recall (95% CI)	F1-Score (95% CI)
Logistic Regression	0.7921 (0.7648–0.8187)	0.5491 (0.4975–0.6005)	0.6079 (0.5598–0.6533)
Tuned Random Forest	0.7823 (0.7543–0.8089)	0.6048 (0.5575–0.6540)	0.6316 (0.5826–0.6769)
Tuned XGBoost	0.7905 (0.7637–0.8171)	0.5942 (0.5467–0.6430)	0.6214 (0.5733–0.6659)

The bootstrap intervals showed substantial overlap across the final models. The 95% confidence intervals for ROC-AUC were 0.7648–0.8187 for Logistic Regression, 0.7543–0.8089 for tuned Random Forest, and 0.7637–0.8171 for tuned XGBoost. Pairwise DeLong tests with Holm adjustment further indicated no statistically significant differences in ROC-AUC, with adjusted p-values of 0.6513 for Logistic Regression versus tuned Random Forest, 1.0000 for Logistic Regression versus tuned XGBoost, and 0.4363 for tuned Random Forest versus tuned XGBoost. Therefore, the observed numerical differences in ROC-AUC were not statistically distinguishable in the present test set.

The threshold-dependent measures in Table 4 were evaluated using a fixed threshold of $\tau = 0.50$, representing performance at a specific operating point. Threshold optimization and calibration analysis were not performed; therefore, the observed Recall and F1-Score values may vary under different classification thresholds. Future studies may examine threshold selection and model calibration. Given that the outcome was based on self-reported physician diagnosis, the models classify reported hypertension status rather than provide a clinical diagnosis or evidence of early-detection screening.

3.4. Cross-Model SHAP Analysis and Individual-Level Explanation

Previous studies have applied machine learning and SHAP-based approaches to hypertension modeling [7, 8, 10, 11]. However, predictive performance alone does not indicate whether models with different learning mechanisms identify similar influential predictors. Therefore, the optimized Random Forest and XGBoost models were further examined using SHAP, as described in Section 2.8, to assess the consistency of predictor contributions across models. Figures 3 and 4 show the global SHAP feature importance for the tuned Random Forest and XGBoost models, respectively.

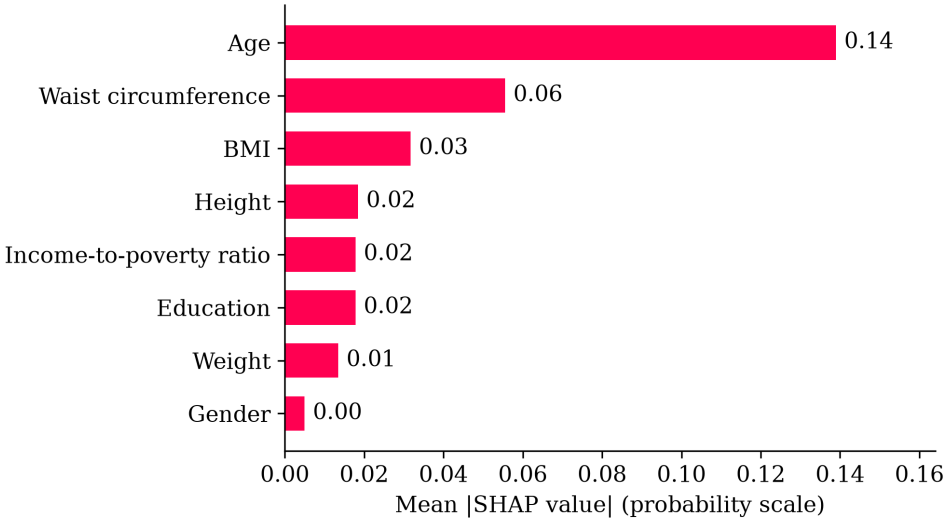


Fig. 3: Global SHAP feature importance of the tuned Random Forest. SHAP values are presented on the model probability scale.

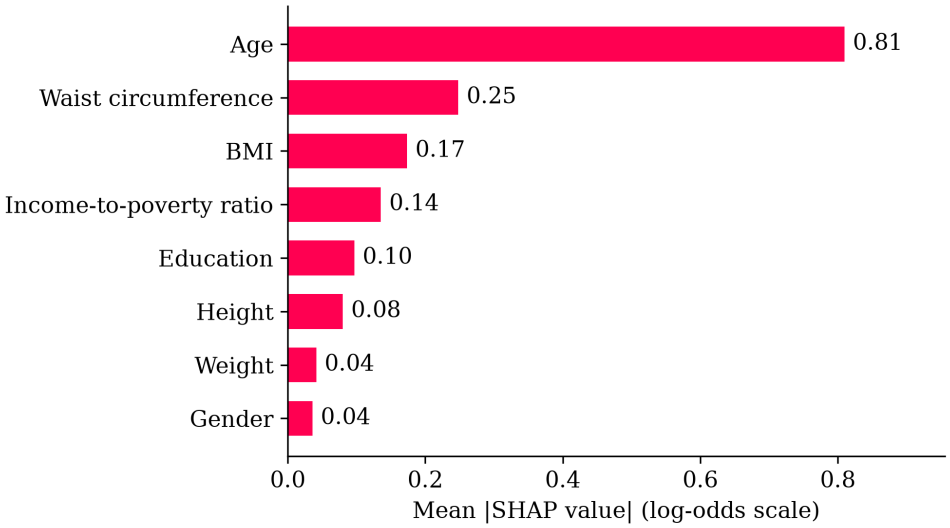


Fig. 4: Global SHAP feature importance of the tuned XGBoost. SHAP values are presented on the model raw-output (log-odds) scale.

Both models identified Age, Waist circumference, and BMI as the three most influential predictors. The complete feature rankings showed a Spearman rank correlation of $\rho_s = 0.9286$, as defined in Eq. (29), while the three highest-ranked predictors were identical in both models, corresponding to 100% top-three agreement as defined in Eq. (30). These results indicate a high degree of consistency in predictor-importance rankings despite the different learning mechanisms of the two ensemble models. The importance values underlying these rankings correspond to the mean absolute SHAP values defined in Eq. (28).

To assess the robustness of this cross-model agreement, 2,000 bootstrap resamples of the independent test set were generated with replacement while keeping the fitted models fixed (Section 2.8). The median Spearman rank correlation across the bootstrap replications was 0.9286, with a 95% bootstrap percentile interval of 0.9048–1.0000. The exact top-three predictor set was identical between the two models in 100% of the bootstrap replications. At the feature level, Age, Waist circumference, and BMI retained median ranks of 1, 2, and 3, respectively, and each appeared among the top three predictors in 100% of the bootstrap replications for both models. These findings indicate that the observed cross-model SHAP agreement was stable under bootstrap resampling of the analyzed test sample.

The prominence of Age, Waist circumference, and BMI is also consistent with previous machine-learning studies of hypertension. Zhang et al. [10], using China Health and Nutrition Survey data, identified Age, Waist circumference, and BMI among the important predictors in their hypertension prediction model. Similarly, a recent Indonesian hypertension screening study reported Age, Waist circumference, and BMI among the important predictors in its machine-learning models [15]. Although the outcome definitions, populations, predictor sets, and modeling procedures differed from those of the present study, the recurrence of these predictors provides contextual support for their importance in hypertension-related classification.

Figures 5 and 6 present the SHAP beeswarm plots for the tuned Random Forest and XGBoost models, respectively, and further illustrate the direction and distribution of individual feature contributions. Higher values of Age were generally associated with positive SHAP contributions, while Waist circumference and BMI also showed substantial contributions across observations. Income-to-poverty ratio and Education exhibited more heterogeneous contribution patterns. Because the two models use different SHAP output scales, their absolute SHAP magnitudes were not directly compared. Instead, cross-model interpretation focused on the consistency of feature rankings and contribution directions.

The prominence of Age, Waist circumference, and BMI should not be interpreted as evidence

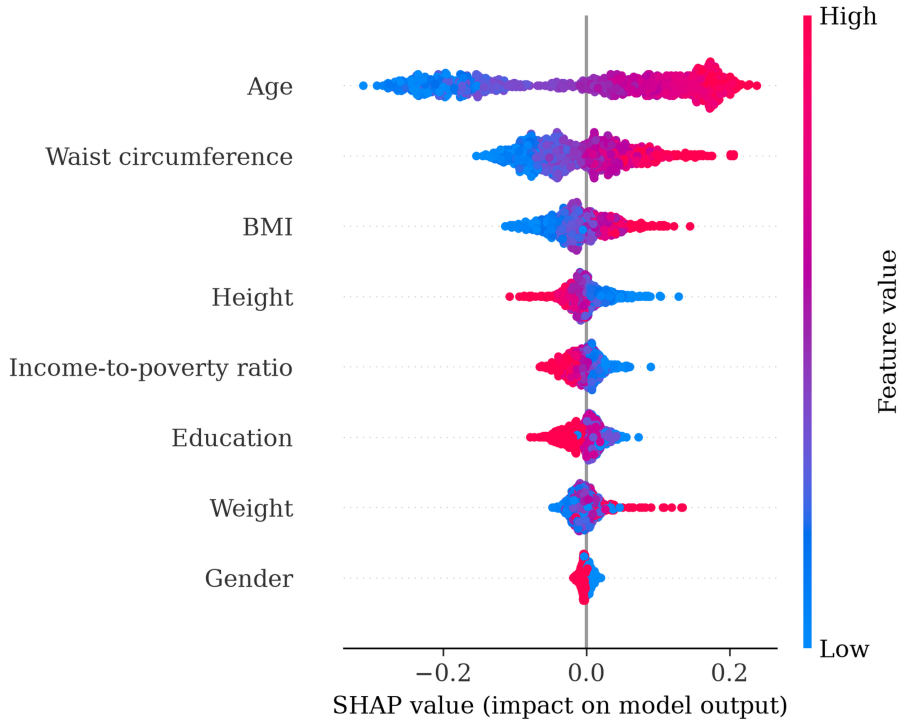


Fig. 5: SHAP beeswarm plot of the tuned Random Forest. SHAP values are presented on the model probability scale.

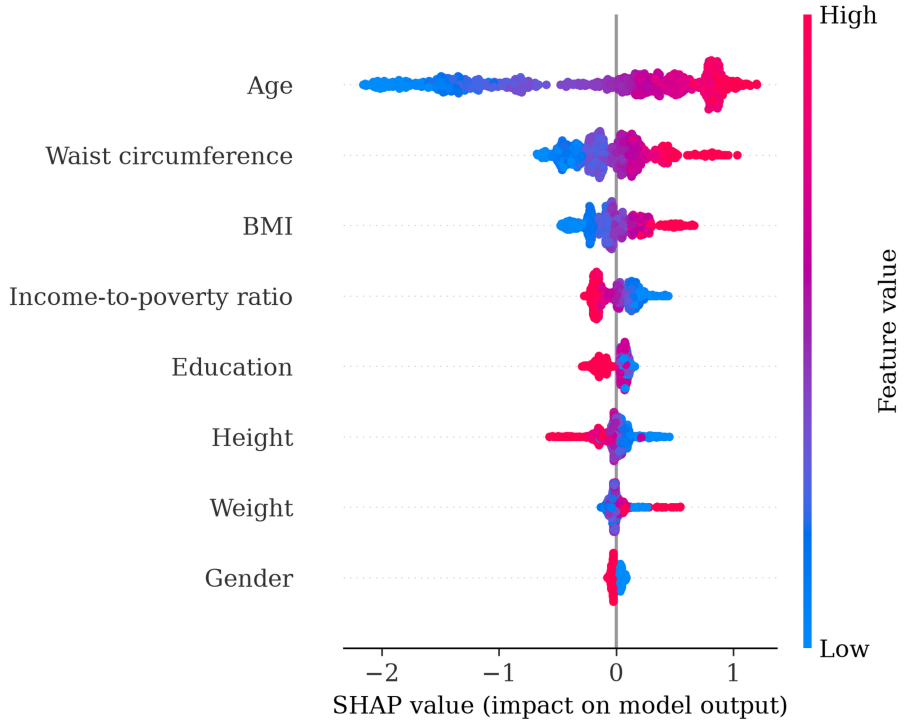


Fig. 6: SHAP beeswarm plot of the tuned XGBoost. SHAP values are presented on the model raw-output (log-odds) scale.

of independent or causal effects. BMI is mathematically derived from Height and Weight, while Waist circumference provides related anthropometric information. Consequently, SHAP contributions may reflect shared or overlapping information among anthropometric predictors. The results therefore describe how the fitted models use the available predictors for classification rather than estimating isolated causal effects.

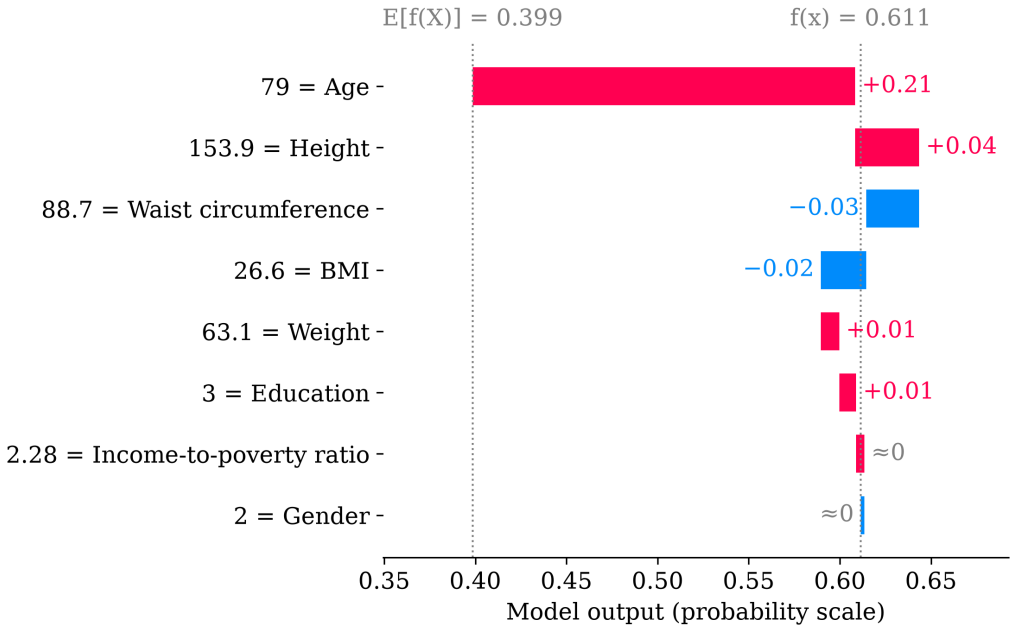


Fig. 7: SHAP waterfall plot of the tuned Random Forest for a single test-set observation. SHAP values are presented on the model probability scale. Contributions that round to zero at two decimals are shown as ≈ 0, and displayed values may not sum exactly to the model output because of rounding.

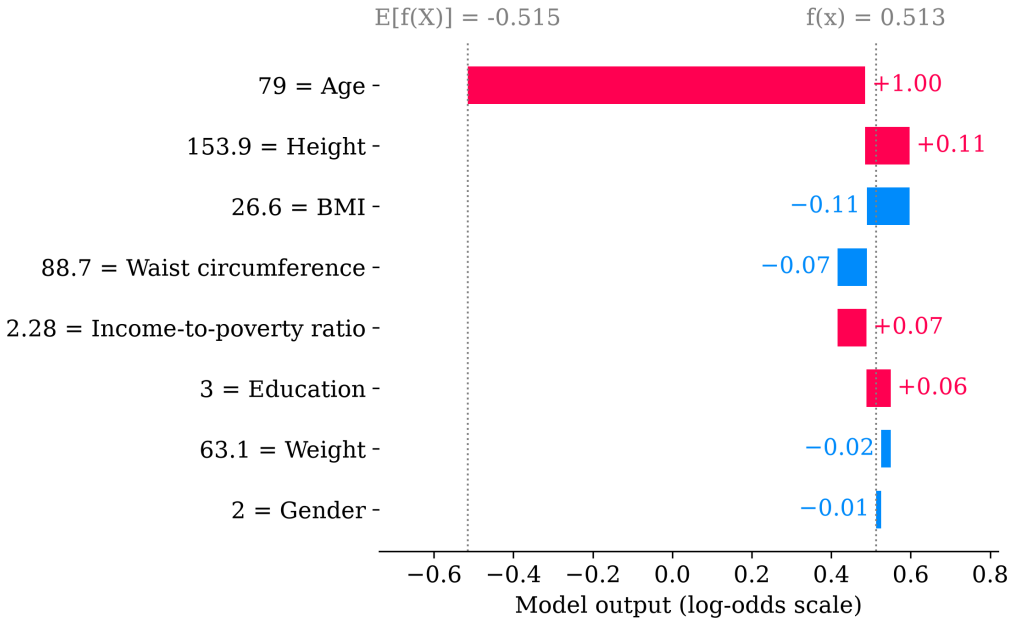


Fig. 8: SHAP waterfall plot of the tuned XGBoost for the same observation. SHAP values are presented on the model raw-output (log-odds) scale. The raw model output of 0.513 corresponds to a predicted probability of approximately 0.625 after the logistic transformation.

Figures 7 and 8 present SHAP waterfall plots for the same test-set observation, enabling comparison of the predictor contributions assigned by the two models. The observation was selected because both tuned models classified it as hypertensive despite its observed status being non-hypertensive, representing a false-positive case. The respondent was 79 years old, with a height of 153.9 cm, weight of 63.1 kg, waist circumference of 88.7 cm, and BMI of 26.6. The tuned Random Forest and XGBoost models assigned predicted probabilities of 0.611 and 0.625, respectively. Both probabilities exceeded the fixed classification threshold of 0.50, resulting in a positive classification despite the observed non-hypertensive status.

Age provided the largest positive contribution in the Random Forest explanation (+0.21), while Age also showed the dominant contribution in the XGBoost explanation. This individual-level pattern is consistent with the first-ranked position of Age in the global SHAP analysis. For XGBoost, the waterfall plot is expressed on the raw model-output scale rather than the probability scale. Thus, the raw model output of 0.513 shown in Figure 8 corresponds to a predicted probability of approximately 0.625 after the logistic transformation in Eq. (18). In both waterfall plots, the observation-level contributions sum to the difference between the model output and the expected output, consistent with the SHAP decomposition in Eq. (27).

Overall, the SHAP analysis demonstrated strong cross-model agreement in the identification of influential predictors. Age, Waist circumference, and BMI consistently occupied the top three positions, with a Spearman rank correlation of $\rho_s = 0.9286$ and 100% top-three agreement in the original test set. The bootstrap analysis further showed that this agreement remained stable under resampling of the analyzed test sample, with a median Spearman rank correlation of 0.9286 (95% CI: 0.9048–1.0000) and 100% top-three agreement across 2,000 replications. Nevertheless, these findings represent model-based predictive contributions rather than causal effects or clinical diagnostic evidence. Because the outcome was based on self-reported physician diagnosis rather than direct blood-pressure measurement, the models should be interpreted as classifiers of reported hypertension status rather than tools for clinical diagnosis or early detection.

4. Conclusion

This study compared Logistic Regression (LR), Random Forest (RF), and XGBoost for classifying reported hypertension status using NHANES 2021–2023 data. No single model showed uniformly higher performance across all evaluation metrics. LR achieved the highest ROC-AUC (0.7921), whereas tuned RF achieved the highest Recall (0.6048) and F1-Score (0.6316) at the fixed classification threshold of $\tau = 0.50$. Pairwise DeLong tests with Holm adjustment showed that the observed differences in ROC-AUC were not statistically distinguishable.

Cross-model SHAP analysis showed consistent predictor rankings, with Age, Waist circumference, and BMI occupying the top three positions in both tuned RF and XGBoost ($\rho_s = 0.9286$; 100% top-three agreement). This agreement remained stable under bootstrap resampling of the analyzed test sample, with a median Spearman rank correlation of 0.9286 (95% CI: 0.9048–1.0000) and 100% top-three agreement across 2,000 replications. These results support the consistency of the observed predictor-ranking patterns across the two tree-based models.

SHAP contributions should be interpreted as model-based predictive contributions rather than causal effects, particularly because the anthropometric predictors are interrelated. Because the outcome was based on self-reported physician diagnosis, the models classify reported hypertension status rather than provide clinical diagnosis or early detection. Future research should incorporate measured blood pressure and treatment information, consider NHANES survey design when population-level inference is intended, and evaluate calibration, threshold selection, repeated data partitions, and external validation.

CRediT Authorship Contribution Statement

Diska Nur Isna Apriani: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Visualization, Writing–Original Draft Preparation.
A'yunin Sofro: Conceptualization, Methodology, Validation, Supervision, Writing–Review & Editing.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this manuscript, ChatGPT (OpenAI) was used as an AI-assisted tool for language editing, improving clarity and academic writing, and providing assistance with

LaTeX formatting. The authors reviewed, verified, and took full responsibility for the content of the manuscript. An AI-writing detection report with a score below 25% is provided as required by the journal.

Declaration of Competing Interest

The authors declare no competing interests.

Ethics Approval and Consent to Participate

This study used secondary data from the National Health and Nutrition Examination Survey (NHANES) and involved no direct interaction with human participants or collection of new personal data by the authors. Therefore, the study did not involve participant recruitment or primary-data collection by the authors.

Funding and Acknowledgments

This research received no external funding.

Data and Code Availability

The NHANES dataset used in this study is publicly available through the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The analysis code used in this study is publicly available at the GitHub repository: https://github.com/Diska74/NHANES-Hypertension-ML, to support the reproducibility of the reported analyses.

References

[1] S. M. S. Islam, A. Talukder, M. A. Awal, M. M. U. Siddiqui, M. M. Ahamad, M. M. Ahammed, B. Ahammed, L. B. Rawal, R. Alizadehsani, J. Abawajy, L. Laranjo, C. K. Chow, and R. Maddison. "Machine learning approaches for predicting hypertension and its associated factors using population-level data from three South Asian countries." *Frontiers in Cardiovascular Medicine* 9 (2022), p. 839379. Doi: 10.3389/fcvm.2022.839379.

[2] N. Chen, F. Fan, J. Geng, Y. Yang, Y. Gao, H. Jin, Q. Chu, D. Yu, Z. Wang, and J. Shi. "Evaluating the risk of hypertension in residents in primary care in Shanghai, China with machine learning algorithms." *Frontiers in Public Health* 10 (2022), p. 984621. Doi: 10.3389/fpubh.2022.984621.

[3] United Nations Department of Economic and Social Affairs. "Goal 3: Ensure healthy lives and promote well-being for all at all ages." *United Nations Sustainable Development Goals*. [Online]. Available: https://sdgs.un.org/goals/goal3. Accessed: Sep. 20, 2026.

[4] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2013. Doi: 10.1002/9781118548387.

[5] L. Breiman. "Random forests." *Machine Learning* 45.1 (2001), pp. 5–32. Doi: 10.1023/A:1010933404324.

[6] T. Chen and C. Guestrin. "XGBoost: A scalable tree boosting system." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794. Doi: 10.1145/2939672.2939785.

[7] I. K. Sifat and M. K. Kibria. "Optimizing hypertension prediction using ensemble learning approaches." *PLOS ONE* 19.12 (2024), p. e0315865. Doi: 10.1371/journal.pone.0315865.

[8] A. A. Huang and S. Y. Huang. "Shapely additive values can effectively visualize pertinent covariates in machine learning when predicting hypertension." *The Journal of Clinical Hypertension* 25.12 (2023), pp. 1135–1144. Doi: 10.1111/jch.14745.

[9] W. Li, G. Huang, N. Tang, P. Lu, L. Jiang, J. Lv, Y. Qin, Y. Lin, F. Xu, and D. Lei. "Effects of heavy metal exposure on hypertension: A machine learning modeling approach." *Chemosphere* 337 (2023), p. 139435. Doi: 10.1016/j.chemosphere.2023.139435.

[10] M. Zhang, X. Xia, Q. Wang, Y. Pan, G. Zhang, and Z. Wang. "Application of machine learning algorithms in predicting new onset hypertension: a study based on the China Health and Nutrition Survey." *Environmental Health and Preventive Medicine* 30 (2025), p. 3. Doi: 10.1265/ehpm.24-00270.

[11] A. Sofro, A. Maharani, M. E. Mustafidah, K. N. Khikmah, A. Oktaviarina, and D. Ariyanto. "Traditional logistic regression and machine learning approaches of sociodemographic and anthropometric factors influencing hypertension in athletes." *BAREKENG: Jurnal Ilmu Matematika dan Terapan* 20.2 (2026), pp. 1125–1138. Doi: 10.30598/barekengvol20iss2pp1125-1138.

[12] Centers for Disease Control and Prevention, National Center for Health Statistics. "About NHANES." *National Health and Nutrition Examination Survey*. [Online]. Available: https://www.cdc.gov/nchs/nhanes/about/index.html. Accessed: Sep. 20, 2026.

[13] S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*, vol. 30 (2017), pp. 4765–4774. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[14] P. Probst, A.-L. Boulesteix, and B. Bischl. "Tunability: Importance of hyperparameters of machine learning algorithms." *Journal of Machine Learning Research* 20.53 (2019), pp. 1–32. [Online]. Available: https://jmlr.org/papers/v20/18-444.html.

[15] R. I. Estiko, B. Widyantoro, D. A. Juzar, R. M. Yusup, I. F. Rakhmat, and E. Rijanto. "Screening hypertension using non-laboratory risk factors with machine learning: a retrospective cross-sectional study in Indonesia." *BMJ Open* 15.8 (2025), p. e092364. Doi: 10.1136/bmjopen-2024-092364.