Original
ISSN 2413-0516

# Prevalence and Associated Factors of Hypertension Among Adolescents: A Machine Learning Approach

Mohammed Saad Abdullah*, Kameran Hassan Ismail

Department of Community Medicine, College of Medicine, Hawler Medical University, Erbil, Iraq.
*Correspondence to: Mohammad Saad Abdullah (E-mail: mohammedsaad0093@gmail.com)
*(Submitted: 13 December 2025 – Revised version received: 07 January 2026 – Accepted: 16 January 2026 – Published online: 26 February 2026)*

**Abstract**

**Objective:** This study aims to estimate the prevalence of HTN among high school students in Erbil City, Kurdistan Region, Iraq, and to identify associated factors using a machine learning approach.

**Methods:** A school-based cross-sectional study was conducted ($n = 1619$). Eight distinct supervised learning algorithms included Logistic Regression (LR), k-Nearest Neighbors (kNN), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), Naïve Bayes (NB), and Support Vector Machine (SVM) were implemented to compare the predictive performance. The performance of each model was assessed using the following metrics: Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Area Under the Precision-Recall Curve (AUC-PR), Balanced Accuracy, Precision, Recall, Specificity, and F1 score. The SHapley Additive exPlanations (SHAP) were employed for model's interpretability. Data were analyzed using R (version 4.5.1).

**Results:** The prevalence of HTN was 11.3% (95% CI: 9.8%–13.0%). The highest (AUC-ROC) value (0.8306) was observed for the LR model. However, the XGBoost, RF, and GBM models exhibited slightly lower AUC-ROC values (0.8173, 0.8078, and 0.8041, respectively). Key factors for HTN prediction included higher salt intake, higher BMI, older age, higher sedentary behavior (hr/day), lower physical activity (days/week), positive family history of HTN, female sex, lower vegetable intake, lower sleep duration (hr/night), and lower physical activity (min/day).

**Conclusion:** This study demonstrated that machine learning algorithms, particularly LR and XGBoost, provide good discriminative power for predicting adolescent hypertension. Future research should focus on validating these models across diverse geographic cohorts to ensure generalizability.

**Keywords:** Prevalence, risk factors, hypertension, adolescents, machine learning

## Introduction

Globally, the prevalence of hypertension (HTN) is increasing owing to an aging population and increased exposure to lifestyle risk factors such as unhealthy diets and insufficient physical activity.[1] Over the past two decades, hypertension prevalence has been higher in low-income and middle-income countries than in high-income countries.[2] Consequently, these differences in trends suggest that the healthcare systems in these countries will face a rising burden of hypertension and associated cardiovascular diseases.[3]

Hypertension is associated with adverse cardiovascular events, including morbidity and mortality worldwide and remains a primary contributor to the burden of disability-adjusted life-years.[4] Moreover, hypertension in childhood tends to track into adulthood and is associated with adverse cardiovascular outcomes in adulthood.[5,6] Consequently, the early identification and management of hypertension in youth is essential in the primordial and primary prevention of cardiovascular disease. Unfortunately, hypertension is frequently under-recognized in youth, as they are generally healthy and rarely visit a physician unless there is an apparent illness.[7,8]

Nowadays, artificial intelligence (AI) has increasingly become a transformative component of modern healthcare, demonstrating strong potential for the management and prediction of chronic diseases, such as HTN.[9] Machine learning (ML), a core subset of AI, is a scientific discipline that focuses on how computers learn from data. It arises at the intersection of statistics, and computer science, with an emphasis on efficient computing algorithms.[10]

There is limited research on the prevalence of hypertension and its associated risk factors among adolescents in the Kurdistan region of Iraq. Previous studies have focused primarily on the adult population, relied on a single blood pressure measurement, or had a small sample size, with no studies particularly targeting the adolescent population.[11,12] From a public health perspective, accurate estimates of childhood and adolescent hypertension are crucial for effective prevention and treatment strategies and to guide evidence-based health resource allocation and policymaking.[13] Therefore, the present study aimed to estimate the prevalence of hypertension among high school students aged 15–19 years in Erbil City, Kurdistan Region of Iraq, and to identify the associated factors using machine learning algorithms.

## Materials and Methods

### *A. Study Design and Population*

This school-based cross-sectional study was conducted from October 1 to December 30, 2024, in Erbil City, Kurdistan Region, Iraq. The study was reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.[14] Eligible participants were students aged 15–19 years old in the grade (10th, 11th, and 12th). Students who refused to provide verbal consent, had a serious systemic disease (such as renal or hepatic dysfunction), or used medications known to elevate blood pressure, such as glucocorticoids or glycyrrhiza were excluded.

## B. Sample Size and Sampling Procedure

Using the Epi Info software, the required sample size was calculated based on the 5.8% hypertension prevalence based on previous study.[11] With a 95% confidence level, 2% margin of error, and design effect of 2, the minimum required sample size was 1,038 students. Accounting for a 20% non-response rate, the target sample size was 1,246. Ultimately, 1,619 students were enrolled in this study. Response rate was 99.75%.

A multi-stage cluster sampling method was employed: first stage, twenty-two schools were randomly selected; second stage, classes were stratified by grade, with one class per grade was selected randomly by a person other than the researcher. Only those students who provided consent were included in the final sample.

## C. Data Collection Instrument

Data were collected using a structured, self-administered questionnaire adapted from the WHO STEPwise Approach to Surveillance (STEPS).[15] Five academic and clinical experts confirmed the instrument's content validity. The questionnaire was translated from English to Kurdish by a professional translator and back-translated into English to ensure accuracy and consistency. The instrument consisted of five sections.

### *Section 1: Demographics and History of Hypertension*

Collected variables included age (years), sex, self-reported previous diagnosis of hypertension, and self-reported family history of hypertension in at least one first-degree relative.

### *Section 2: Dietary Behaviors*

Three items were used to collect information on dietary behaviors:[16]

Fruits: Participants were asked, "During the past week, how many times per day did you usually eat fruit, such as 'an apple?" The response options were 1 = I did not eat fruit during the past seven days, 2 = less than once per day, 3 = once per day, 4 = two times per day, 5 = three times per day, 6 = four times per day, and 7 = five or more times per day.

Vegetables: "During the past week, how many times per day did you usually eat vegetables, such as 'cucumbers and eggplant etc.?" The response options were identical to those used for fruit intake.

For the final analysis, fruit and vegetable intake was reclassified into three levels: low (<1 time/day), medium (1–2 times/day), and high (≥3 times/day).

Salt intake was assessed using the question "Do you add salt to your food?" Response options were "always," "sometimes," or "never."

### *Section 3: Physical Activity, Sedentary Behavior, and Sleep Duration*

Physical activity (PA) was evaluated using two questions:[17]

1. "On average, how many days per week do you engage in moderate-to-vigorous PA (e.g., brisk walking)?"
2. "On average, how many minutes per day do you engage in PA?"

Sedentary behavior was evaluated using the following question: "On a typical day, how much time do you spend sitting and watching television, playing computer games, talking with friends, or doing other sitting activities?".[16] Sleep duration was measured using the question "How many hours do you usually sleep at night?".[18]

### *Section 4: Anthropometric Measurements*

Height was measured to the nearest 0.1 cm using a portable stadiometer, with participants standing barefoot, heels together, knees straight, back against the board, and head in the Frankfort horizontal plane. Weight was measured to the nearest 0.1 kg using a calibrated portable digital scale, with participants wearing light clothing and no footwear.

Body mass index-for-age Z-scores (BAZ) were calculated using the 2007 WHO growth reference standards (via the R package "anthro") and classified as thin (<−2 SD), normal (−2 to 1 SD), overweight (>1 to 2 SD), and obese (>2 SD).[19]

### *Section 5: Blood Pressure Measurements*

Blood Pressure (BP) was measured using a validated OMRON® automated digital sphygmomanometer. Two trained health professionals conducted measurements in a quiet setting, with participants seated, feet flat, and back supported, after resting for 15 minutes. The cuff was placed on the non-dominant arm, positioned at the heart level, using an appropriately sized cuff based on the arm circumference. Students were advised to avoid smoking or stimulants for at least one hour prior. Two BP measurements were taken during the first visit, 3–5 minutes apart. Within 1–2 weeks, the same procedure was applied at the second visit. The means of the first and second average visits were used in the final analysis, as shown in Figure 1. The 2017 American Academy of Pediatrics (AAP) Clinical Practice Guidelines were used to define hypertension.[20]

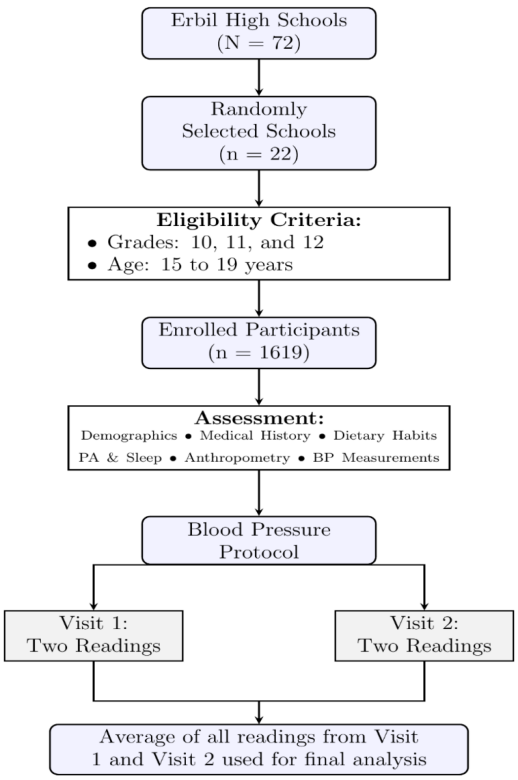


Fig. 1 **Study flowchart of school selection and students' recruitment.**

- **Normal**: <120/80 mmHg
- **Elevated**: Systolic BP 120–129 mmHg and diastolic BP <80 mmHg
- **Stage 1 HTN**: Systolic 130–139 mmHg or diastolic 80–89 mmHg
- **Stage 2 HTN**: Systolic BP ≥140 mmHg or diastolic BP ≥90 mmHg

## *D. Data Analysis and Machine Learning*

Data were analyzed using the R software (version 4.5.1). There were no missing data. The normality of the distribution was assessed using Q-Q plots, Shapiro–Wilk, and Kolmogorov-Smirnov tests. Continuous variables are presented as mean ± SD, whereas categorical variables are reported as frequencies (%). Hypertension prevalence was defined as the proportion of students who were either previously diagnosed with hypertension or had blood pressure readings meeting the AAP (2017) criteria for hypertension during the study assessment. Prevalence estimates were reported with 95% confidence intervals (CIs). Group comparisons were performed using Mann–Whitney U tests for continuous variables and chi-squared tests for categorical variables. For the final analysis, HTN (stage 1 + stage 2 HTN) was compared with the reference group with normal/elevated blood pressure. Statistical significance was set at $P < 0.05$.

### *D.1. Pre-Processing, Feature Selection, and Class Imbalance*

A fixed random seed was applied prior to data partitioning, feature selection, SMOTE implementation, and model training to ensure the reproducibility of the machine learning results. Data splitting was performed using the rsample package (V1.3.1), with 70% of the data ($n$ = 1,133) allocated to the training set and 30% ($n$ = 486) to the testing set via stratified random sampling.

For feature selection, a random forest-based Recursive Feature Elimination (RFE) algorithm with 10-fold cross-validation was used to identify the important features of the eleven initial candidate variables: age, sex, salt intake, fruit intake, vegetable intake, PA (min/day), PA (days/week), sedentary (hr/day), sleep (hr/night), BMI category, and family history of HTN.[21] Following the RFE, ten variables were identified as the most predictive features (fruit intake was excluded). These variables were subsequently used to construct the final predictive models for the eight machine learning classifiers evaluated in this study.

To address the class distribution imbalance, where hypertensive cases were the minority, we implemented the Synthetic Minority Over-Sampling Technique (SMOTE), which generates synthetic instances of the minority class by interpolating between existing samples, thereby providing additional variability, and strengthening the representation of the less-frequent class.[22]

### *D.2. Machine Learning Algorithms*

Eight distinct supervised learning algorithms were implemented to compare predictive performance:

1. **Logistic Regression (LR):** is the most common supervised ML-based algorithm that utilizes the idea of probability. LR is primarily used for classification.[23]
2. **k-Nearest Neighbors (kNN):** A non-parametric, distance-based algorithm known for its simplicity and ease of implementation.[23]
3. **Decision Tree (DT):** A flowchart-like model that splits data into branches based on feature values to achieve classification.
4. **Random Forest (RF):** An ensemble machine learning method for classification, regression, and unsupervised learning based on a set of multiple trees. RF is known to be robust to data with missing values, manages large and complex data, and reduce overfitting.[24]
5. **Gradient Boosting Machine (GBM):** A family of powerful ML techniques that have shown considerable success in a wide range of practical applications.
6. **Extreme Gradient Boosting (XGBoost):** A decision-tree-based ensemble ML method for classification and regression. XGBoost can control the model complexity and overfitting by adding a regular term to the loss function.[24]
7. **Support Vector Machine (SVM):** Demonstrated utility for high-dimensional data and can manage complex or non-linear relationships via kernel functions.[24]
8. **Naïve Bayes (NB):** A probabilistic classifier based on Bayes' theorem that assumes strong independence between predictors.

### *D.3. Model Training and Hyperparameter Tuning*

All models were trained using the caret package (V 7.0.1). To ensure robustness and prevent overfitting, 10-fold cross-validation was applied during the training phase. The optimal hyperparameter values for each model were determined using an automated grid search. The Area Under the Receiver Operating Characteristic curve (AUC-ROC) was prioritized as the primary metric for the model selection and tuning.

### *D.4. Classification Threshold and Model Evaluation*

Recognizing the clinical importance of balancing sensitivity and specificity, the classification threshold was not fixed at a default value of 0.5. The Youden Index (J) was used to determine the optimal cutoff for each model to maximize its ability to correctly identify hypertensive cases while maintaining high specificity.[24,25]

The performance of each model on the test dataset was assessed using the following metrics: AUC-ROC, Area Under the Precision-Recall Curve (AUC-PR), Balanced Accuracy, Precision, Recall (Sensitivity), Specificity, and F1 score. AUC-ROC and AUC-PR were derived from their respective curves, whereas the remaining metrics were calculated from the confusion matrix. Balanced accuracy was calculated specifically to account for the target imbalance in the dataset. The F1-score, calculated as the harmonic mean of precision and sensitivity, was used to assess the balance between these two metrics.[24]

### *D.5. Model Interpretability*

Interpretability of the ML model is defined as the extent to which a model can explain its output given a set of inputs. Interpretation of ML models can be either global or local in scope. Global interpretability recognizes how the model makes its predictions based on a holistic view of its features,

parameters, and structure. On the other hand, local interpretability is maintained by designing more justified model architectures that explain a single prediction.[26] The SHapley Additive exPlanations (SHAP) method provides an interpretable approach to understanding ML models. In this study, the KernelSHAP algorithm was implemented using the kernelshap (V 0.9.1) and shapviz (V 0.10.3) packages.[23,25,26]

## Results

### A. Prevalence of Blood Pressure Categories

Table 1 shows the distribution of the blood pressure categories ($N$ = 1,619). The mean systolic blood pressure (SBP) was 110.1 ± 11.6 mmHg, and the mean diastolic blood pressure (DBP) was 72.0 ± 7.0 mmHg. Males showed a higher mean SBP compared to females (113.2 ± 11.4 vs. 107.5 ± 11.2), while males showed a lower mean DBP compared to females (71.0 ± 6.8 vs. 72.9 ± 7.0).

Stage 1 HTN was found in 6.3% of males and 9.3% of females, while stage 2 HTN was found in 2.8% and 1.8%, respectively. Among all students, 22 (1.4%) were previously diagnosed with hypertension. Overall, the prevalence of HTN (stage 1 and stage 2 combined) was 11.3% (95% CI: 9.8%–13.0%) in the students.

### B. General Characteristics of Study Participants

Table 2 presents the distribution of general characteristics ($N$ = 1,619). The mean age was higher among the HTN students ($P$ < 0.001). Sex ($P$ = 0.112) and fruit intake ($P$ = 0.564) were not significantly different. Vegetable intake was significantly associated with HTN ($P$ = 0.039), but fruit intake was not ($P$ = 0.564). Furthermore, students who "always" added salt to their food (35.5%) had a higher HTN prevalence ($P$ < 0.001). The mean PA (min/day) was lower among the HTN students ($P$ = 0.002). In addition, the mean sedentary behavior (hr/day) was higher among HTN students ($P$ < 0.001). Obese students had a higher prevalence of HTN (21.1%) ($P$ < 0.001). A positive family history was significantly associated with HTN ($P$ = 0.009).

### C. Hypertension ML Models Comparisons

Table 3 presents hyperparameter tuning. The performance metrics for the eight ML models are summarized in Table 4. The highest value (AUC-ROC = 0.8306) was observed for the LR model. However, the XGBoost, RF, and GBM models exhibited slightly lower AUC-ROC values (0.8173, 0.8078, and 0.8041, respectively), as shown in Figure 2. In terms of balanced accuracy, the XGBoost model has the highest value (0.7753), followed by the LR model (0.7670). The SVM showed the highest recall (0.8364), although this was accompanied by a lower specificity (0.6195). In terms of the AUC-PR, the GBM model showed the highest value (0.4217), whereas the XGBoost (0.4069) and LR (0.4049) models showed slightly lower values. Overall, the gradient boosting models (GBM and XGBoost) and LR models outperformed the other models such as kNN and DT. The Logistic Regression model, given its combination of high discriminative power, was selected as the primary architecture for further clinical explanation using the SHAP analysis.

The diagnostic accuracy of the LR model on the test dataset ($n$ = 486) was further evaluated using a confusion matrix, as shown in Figure 3. The model demonstrated a sensitivity of 0.7636, correctly identified 42 out of 55 hypertensive students. The specificity was 0.7703, the model correctly classified 332 out of 431 non-hypertensive students.

### D. Model Interpretation: SHAP Analysis

The SHAP analysis of the LR model shows the mean SHAP values of the selected predictors, providing a global view of their relative impact on the model, as shown in Figure 4a. The most influential predictor of hypertension risk was the addition of salt to foods, followed by BMI, age (years), sedentary (hr/day), PA (days/week), family history of hypertension, sex, vegetable intake, sleep (hr/night), and PA (min/week). Figure 4b illustrates the summary plot highlighting the importance of these features for high risk of HTN. Key factors for HTN prediction included higher salt intake, higher BMI, older age, higher sedentary behaviors (hr/day), lower PA (days/week), positive family history of HTN, female sex, lower vegetable intake, lower sleep duration (hr/night), and lower PA (min/day).

Table 1. **Blood pressure categories distribution of the school-aged students in Erbil city, Iraq**

| Characteristics | Male $N$ = 744 | Female $N$ = 875 | Total $N$ = 1619 |
|---|---|---|---|
| **SBP (mmHg)** | | | |
| Mean ± SD (95% CI) | 113.2 ± 11.4 (112, 114) | 107.5 ± 11.2 (107, 108) | 110.1 ± 11.6 (110, 111) |
| **DBP (mmHg)** | | | |
| Mean ± SD (95% CI) | 71.0 ± 6.8 (70, 71) | 72.9 ± 7.0 (72, 73) | 72.0 ± 7.0 (72, 72) |
| **2017 AAP Guideline, *n* (%) (95% CI)** | | | |
| Normal | 545 (73.3%) (69.9, 76.4) | 723 (82.6%) (79.9, 85.0) | 1,268 (78.3%) (76.2, 80.3) |
| Elevated | 131 (17.6%) (15.0, 20.6) | 55 (6.3%) (4.8, 8.2) | 186 (11.5%) (10.0, 13.2) |
| Stage 1 HTN | 47 (6.3%) (4.7, 8.4) | 81 (9.3%) (7.5, 11.4) | 128 (7.9%) (6.7, 9.4) |
| Stage 2 HTN | 21 (2.8%) (1.8, 4.4) | 16 (1.8%) (1.1, 3.0) | 37 (2.3%) (1.6, 3.2) |
| **Previously Diagnosed with HTN, *n* (%) (95% CI)** | | | |
| No | 737 (99.1%) (98.0, 99.6) | 860 (98.3%) (97.1, 99.0) | 1,597 (98.6%) (97.9, 99.1) |
| Yes | 7 (0.9%) (0.4, 2.0) | 15 (1.7%) (1.0, 2.9) | 22 (1.4%) (0.9, 2.1) |
| **HTN (Stage 1 and Stage 2), *n* (%) (95% CI)** | 74 (9.9%) (7.9, 12.4) | 109 (12.5%) (10.4, 14.9) | 183 (11.3%) (9.8, 13.0) |

Table 2. **Characteristics of the school-aged students in Erbil City, Iraq**

| Characteristics | Blood Pressure Category | | Total (*N* = 1619) | *P*-value* |
|---|---|---|---|---|
| | HTN (*N* = 183) | Non-HTN (*N* = 1436) | | |
| **Age (years), *n* (%)** | | | | **<0.001** |
| 15 | 23 (6.5%) | 332 (93.5%) | 355 (100.0%) | |
| 16 | 35 (6.5%) | 507 (93.5%) | 542 (100.0%) | |
| 17 | 41 (10.8%) | 337 (89.2%) | 378 (100.0%) | |
| 18 | 38 (17.6%) | 178 (82.4%) | 216 (100.0%) | |
| 19 | 46 (35.9%) | 82 (64.1%) | 128 (100.0%) | |
| Mean ± SD | 17.3 ± 1.4 | 16.4 ± 1.1 | 16.5 ± 1.2 | **<0.001** |
| **Sex, *n* (%)** | | | | 0.112 |
| Male | 74 (9.9%) | 670 (90.1%) | 744 (100.0%) | |
| Female | 109 (12.5%) | 766 (87.5%) | 875 (100.0%) | |
| **Fruit Intake, *n* (%)** | | | | 0.564 |
| Low | 40 (12.9%) | 271 (87.1%) | 311 (100.0%) | |
| Medium | 87 (11.3%) | 685 (88.7%) | 772 (100.0%) | |
| High | 56 (10.4%) | 480 (89.6%) | 536 (100.0%) | |
| **Vegetable Intake, *n* (%)** | | | | **0.039** |
| Low | 92 (13.2%) | 606 (86.8%) | 698 (100.0%) | |
| Medium | 57 (8.9%) | 586 (91.1%) | 643 (100.0%) | |
| High | 34 (12.2%) | 244 (87.8%) | 278 (100.0%) | |
| **Adding Salt, *n* (%)** | | | | **<0.001** |
| Never | 20 (6.5%) | 290 (93.5%) | 310 (100.0%) | |
| Sometimes | 60 (5.9%) | 959 (94.1%) | 1,019 (100.0%) | |
| Always | 103 (35.5%) | 187 (64.5%) | 290 (100.0%) | |
| **PA (days/week)** | | | | 0.154 |
| Mean ± SD | 2.9 ± 1.8 | 3.0 ± 1.7 | 3.0 ± 1.8 | |
| **PA (min/day)** | | | | **0.002** |
| Mean ± SD | 31.5 ± 26.7 | 40.1 ± 36.4 | 39.1 ± 35.5 | |
| **Sedentary Behaviors (hr/day)** | | | | **<0.001** |
| Mean ± SD | 2.9 ± 2.1 | 2.2 ± 1.7 | 2.3 ± 1.7 | |
| **Sleep Duration (hr/night)** | | | | 0.482 |
| Mean ± SD | 8.0 ± 1.6 | 8.1 ± 1.7 | 8.1 ± 1.7 | |
| **BMI Category, *n* (%)** | | | | **<0.001** |
| Thin (<−2 SD) | 6 (10.5%) | 51 (89.5%) | 57 (100.0%) | |
| Normal (−2 to 1 SD) | 85 (8.3%) | 936 (91.7%) | 1,021 (100.0%) | |
| Overweight (>1 to 2 SD) | 48 (14.5%) | 284 (85.5%) | 332 (100.0%) | |
| Obese (>2 SD) | 44 (21.1%) | 165 (78.9%) | 209 (100.0%) | |
| **FH HTN, *n* (%)** | | | | **0.009** |
| No | 102 (9.8%) | 941 (90.2%) | 1,043 (100.0%) | |
| Yes | 81 (14.1%) | 495 (85.9%) | 576 (100.0%) | |

**P*-values are based on Mann- Whitney U test for continuous variables, and Pearson's Chi-squared test for categorical variables. Bold *P*-value indicates significant difference. FH, Family History; PA, Physical Activity.

## Discussion

This study aimed to estimate the prevalence of HTN among public high school students in Erbil City and identify its associated factors using the ML approach. The results revealed that approximately 11.3% of students had HTN. We trained eight supervised ML algorithms and the LR model demonstrated a higher AUC-ROC. The SHAP plot revealed that the key factors for HTN prediction included higher salt intake, higher BMI, older age, higher sedentary behaviors (hr/day), lower PA

Table 3. **Hyperparameter tuning value for ML-based models**

| Model | Parameter | Optimized value |
|---|---|---|
| **LR** | N/A | (None) |
| **kNN** | Number of Neighbors (k) | 23 |
| **DT** | Complexity Parameter (cp) | 0.01791 |
| **RF** | Variables sampled at each split (mtry) | 2 |
| **GBM** | n.trees, depth, shrinkage, n.minobs | 50, 3, 0.1, 10 |
| **XGBoost** | rounds, depth, eta, gamma, colsample, min_child_weight, subsample | 50, 3, 0.1, 0, 0.8, 1, 0.5 |
| **NB** | Laplace, Usekernel, adjust | 0, True, 1 |
| **SVM** | Sigma, C | 0.034858, 0.25 |

Table 4. **Performance metrics of machine learning models**

| Model | AUC-ROC | AUC-PR | Balanced Accuracy | Precision | Recall | Specificity | F1-Score | Youden Index |
|---|---|---|---|---|---|---|---|---|
| **LR** | 0.8306 | 0.4049 | 0.7670 | 0.2978 | 0.7636 | 0.7703 | 0.4286 | 0.5339 |
| **XGBoost** | 0.8173 | 0.4069 | 0.7753 | 0.3280 | 0.7455 | 0.8051 | 0.4556 | 0.5506 |
| **RF** | 0.8078 | 0.3548 | 0.7550 | 0.2603 | 0.8000 | 0.7100 | 0.3929 | 0.5100 |
| **GBM** | 0.8041 | 0.4217 | 0.7542 | 0.2763 | 0.7636 | 0.7448 | 0.4058 | 0.5084 |
| **NB** | 0.7807 | 0.3359 | 0.7385 | 0.3837 | 0.6000 | 0.8770 | 0.4681 | 0.4770 |
| **SVM** | 0.7785 | 0.3455 | 0.7279 | 0.2190 | 0.8364 | 0.6195 | 0.3472 | 0.4559 |
| **kNN** | 0.7422 | 0.2635 | 0.6938 | 0.3061 | 0.5455 | 0.8422 | 0.3922 | 0.3877 |
| **DT** | 0.7412 | 0.1175 | 0.7428 | 0.3500 | 0.6364 | 0.8492 | 0.4516 | 0.4856 |

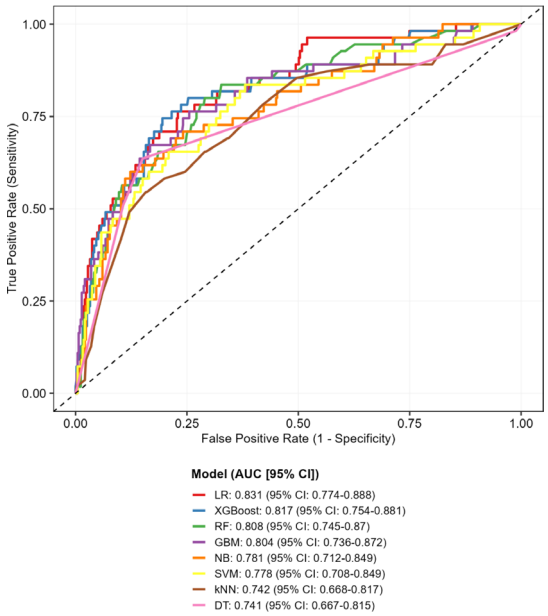


Fig. 2 **ROC curve for ML models in identifying hypertension among adolescents. The figure illustrates the discriminative ability with area under ROC curve (AUC) of eight machine learning algorithms.**

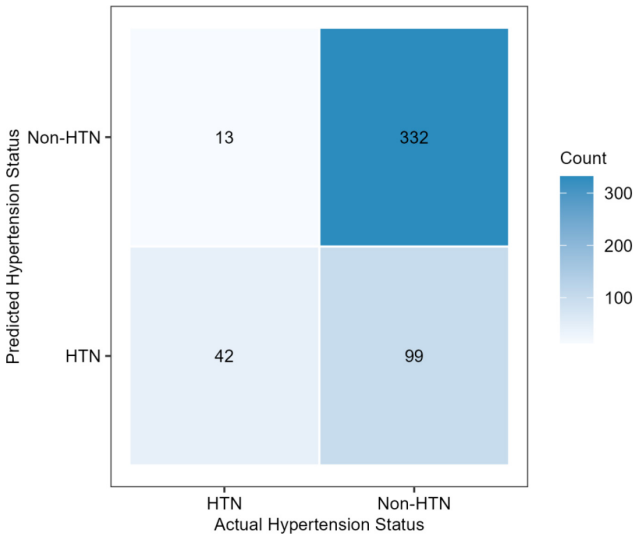


Fig. 3 **Confusion matrix for the logistic regression model. The heatmap illustrates the classification performance. The x-axis shows actual hypertension status, and the y-axis represents the model's prediction.**

(days/week), positive family history of HTN, female sex, lower vegetable intake, lower sleep duration (hr/night), and lower PA (min/day).

A direct comparison of the prevalence of hypertension with other studies is limited because of the differences in the devices used, hypertension classification guidelines, number of measurements on a single occasion, number of occasions when measurements were taken, and time intervals between these measurements.[27] Nevertheless, our prevalence is comparable to that reported in a systematic review of Arab countries (12.6% hypertension, 13.9% prehypertension) and a study of Serbian schoolchildren, which reported a 10.5% prevalence.[28]

Several studies conducted in Iraq reported lower HTN prevalence. In Ramadi City, 5.7% of hypertension cases and 9.8% of prehypertension cases were found in intermediate schools,[29]

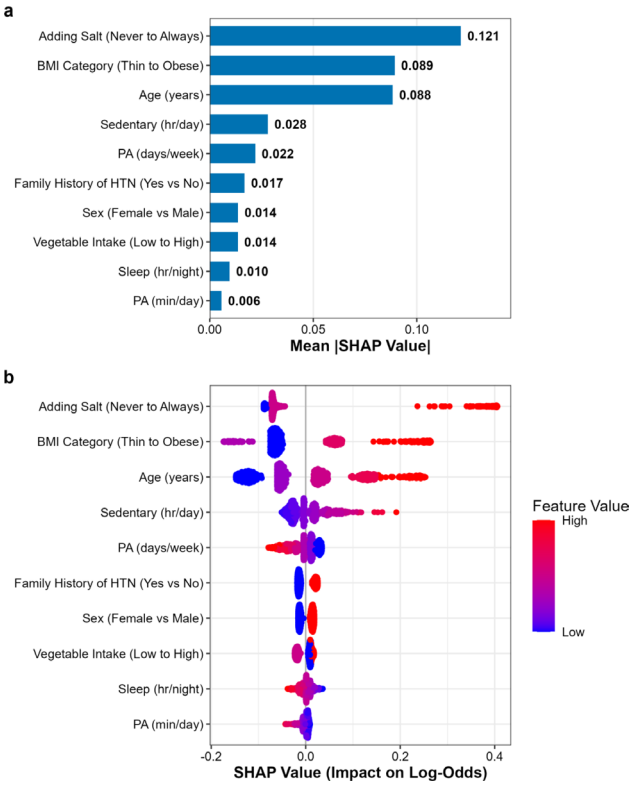


Fig. 4 **SHAP values for logistic regression model. (a) Mean absolute SHAP values (b) Local explanation summary.**

whereas in Duhok City, a study reported a prevalence of 5.84% among students aged 13–18 years.[11] Lower prevalence was also observed in Yemen (2.4% hypertension, 8.2% prehypertension) when using WHO adult thresholds (>140/90 mmHg),[30] which may underestimate pediatric hypertension. In contrast, higher prevalence rates have been reported in Tunisia (15.4%),[31] Saudi Arabia (17.2%),[32] and the United Arab Emirates (15.4% in boys and 17.8% in girls).[33] The relatively high prevalence of HTN observed in our study might be explained by the higher prevalence of modifiable cardiovascular risk factors such as low physical activity, poor diet, and obesity, as previously documented in the Kurdistan region of Iraq.[34,35]

This study identified several predictors of HTN. These findings align with those of previous studies showing that the risk of HTN increases in adolescents.[36] Specifically, higher fruit and vegetable intake is associated with a reduced risk of hypertension,[37] and a meta-analysis confirmed the harmful effect of sodium intake on adolescent blood pressure.[38] A previous study also showed a protective effect of physical activity,[39] whereas each additional sedentary hour has been shown to increase blood pressure in youth.[40] Furthermore, these results are consistent with those of numerous studies that link adiposity to adolescent hypertension via sympathetic activation and hormonal mechanisms.[41]

A key finding of the current study was the LR model showed the highest AUC-ROC, whereas the XGBoost, RF, and GBM models demonstrated slightly lower AUC-ROC values. On the other hand, the highest balanced accuracy was noticed for the XGBoost model, followed by the LR model. Previous studies using ML models have similarly reported that the LR model outperformed more complex algorithms.[25,42] In addition, a study using six different ML models to identify the most salient features contributing to hypertension found that an ensemble of Adaptive Boosting and Logistic Regression yielded superior balanced accuracy (0.812, sensitivity 0.806, specificity 0.818, and AUC-ROC 0.901).[43]

To the best of our knowledge, this is the first study in Erbil City to estimate the prevalence of, and factors associated with HTN among adolescents using ML algorithms. The results revealed several modifiable risk factors for hypertension. Hence, these findings provide a foundation for targeted prevention strategies in school health programs. However, this study had some limitations. First, the cross-sectional design precludes the inference of causality. Second, the sample was limited to an urban geographical area (Erbil City), which may have affected the generalizability of the results to the entire Kurdistan Region. Third, blood pressure was measured only during two visits, with two readings per visit, which may have overestimated the true levels due to the white coat effect or temporary stress. Fourth, self-reported dietary habits and physical activity measures may have been subject to a recall bias.

## Conclusion

This study highlights the high prevalence of hypertension among high school students in Erbil City. Furthermore, machine learning models, particularly Logistic Regression and XGBoost, demonstrated high discriminative ability in predicting adolescent hypertension. The most significant predictors identified were adding salt to food, followed by BMI, age, sedentary behavior, physical activity, family history of hypertension, sex, vegetable intake, and sleep duration. Future research should focus on validating these models across diverse geographic cohorts to ensure their generalizability to school-based and clinical screening programs.

## Ethics Statement

Approval for conducting the study was obtained from the research ethics committee at Hawler Medical University/College of Medicine, which confirmed that the research complied with the ethical standards regarding human subjects (paper code:40, dated: 22/9/2024).

## Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

## Funding Statement

This study was not supported by any sponsor or funder.

## Data Availability Statement

The data supporting the findings of this study are available on request from the corresponding author.

## Disclosure Statement

AI-assisted tools were used to improve clarity and readability, whereas the scientific content remains entirely the work of the researchers.

## References


1. Laatikainen T, Nissinen A, Kastarinen M, Jula A, Tuomilehto J. Blood Pressure, Sodium Intake, and Hypertension Control: Lessons From the North Karelia Project. Glob Heart. 2016 Jun 1;11(2):191.
2. Zhou B, Perel P, Mensah GA, Ezzati M. Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension. Nat Rev Cardiol. 2021 Nov 28;18(11):785–802.
3. Mills KT, Bundy JD, Kelly TN, Reed JE, Kearney PM, Reynolds K, et al. Global Disparities of Hypertension Prevalence and Control. Circulation. 2016 Aug 9;134(6):441–50.
4. Daniels SR. Understanding the Global Prevalence of Hypertension in Children and Adolescents. JAMA Pediatr. 2019 Dec 1;173(12):1133.
5. Fishman B, Bardugo A, Zloof Y, Bendor CD, Libruder C, Zucker I, et al. Adolescent Hypertension Is Associated With Stroke in Young Adulthood: A Nationwide Cohort of 1.9 Million Adolescents. Stroke. 2023 Jun;54(6): 1531–7.
6. Hisamatsu T, Kinuta M. High blood pressure in childhood and adolescence. Vol. 47, Hypertension Research. Springer Nature; 2024. p. 203–5.
7. Khoury M, Urbina EM. Hypertension in adolescents: diagnosis, treatment, and implications. Lancet Child Adolesc Health. 2021 May;5(5):357–66.
8. National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents. The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. Pediatrics. 2004 Aug;114(2 Suppl 4th Report):555–76.
9. Araujo-Moura K, Souza L, de Oliveira TA, Rocha MS, De Moraes ACF, Chiavegatto Filho A. Prediction of Hypertension in the Pediatric Population Using Machine Learning and Transfer Learning: A Multicentric Analysis of the SAYCARE Study. Int J Public Health. 2025 Mar 11;70:1607944.
10. Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920–30.
11. Abdul Majeed AA, Haleem AA. Prevalence of hypertension and its associated risk factors among secondary school students in Duhok City. Healthcare in Low-Resource Settings [Internet]. 2024 Mar 21;12(2):12073 Available from: https://www.pagepressjournals.org/hls/article/view/12073.
12. Saka M, Shabu S, Shabila N. Prevalence of hypertension and associated risk factors in older adults in Kurdistan, Iraq. Eastern Mediterranean Health Journal. 2020;26(3):265–72.
13. Song P, Zhang Y, Yu J, Zha M, Zhu Y, Rahimi K, et al. Global Prevalence of Hypertension in Children. JAMA Pediatr. 2019 Dec 1;173(12):1154.
14. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. PLoS Med. 2007;4(10):e297.
15. WHO. Noncommunicable Disease Surveillance, Monitoring and Reporting [Internet]. [cited 2024 Sep 3]. Available from: https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/steps
16. Peltzer K, Pengpid S. Fruits and Vegetables Consumption and Associated Factors among In-School Adolescents in Five Southeast Asian Countries. Int J Environ Res Public Health. 2012 Oct 11;9(10):3575–87.
17. Wattanapisit A, Ng CJ, Angkurawaranon C, Wattanapisit S, Chaovalit S, Stoutenberg M. Summary and application of the WHO 2020 physical activity guidelines for patients with essential hypertension in primary care. Vol. 8, Heliyon. Elsevier Ltd; 2022.
18. Kuciene R, Dulskiene V. Associations of short sleep duration with prehypertension and hypertension among Lithuanian children and adolescents: a cross-sectional study. BMC Public Health. 2014 Dec 15;14(1):255.
19. WHO. BMI-for-age (5–19 years) [Internet]. [cited 2025 Feb 10]. Available from: https://www.who.int/tools/growth-reference-data-for-5to19-years/indicators/bmi-for-age
20. Flynn JT, Kaelber DC, Baker-Smith CM. Clinical Practice Guideline for Screening and Management of High Blood Pressure in Children and Adolescents [Internet]. Vol. 140, Pediatrics. 2017. Available from: http://publications.aap.org/pediatrics/article-pdf/140/3/e20171904/1104403/peds_20171904.pdf
21. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems. 2006 Sep;83(2): 83–90.
22. Martínez-García M, Gutiérrez-Esparza GO, Márquez MF, Amezcua-Guerra LM, Hernández-Lemus E. Machine learning analysis of emerging risk factors for early-onset hypertension in the Tlalpan 2020 cohort. Front Cardiovasc. Med. 2025 Jan 17;11:1434418.
23. Islam MdM, Alam MdJ, Maniruzzaman M, Ahmed NAMF, Ali MS, Rahman MdJ, et al. Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia. PLoS One. 2023 Aug 24;18(8):e0289613.
24. Seo JW, Lee S, Yim MH. Machine Learning Approach for Predicting Hypertension Based on Body Composition in South Korean Adults. Bioengineering. 2024 Sep 14;11(9):921.
25. Pandey S, Pandey A, Neupane A, Subedi D, Guragain A. Development of a Hypertension Risk Prediction Model using Nationally Representative Survey Data: A Machine Learning Approach and Web Application Deployment. medRxiv. 2025.
26. Salah H, Srinivas S. Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. Sci Rep. 2022 Dec 19;12(1):21905.
27. Petracco AM, Mattiello R, Bortolotto CC, Ferreira RW, Matijasevich A, de Barros FCLF, et al. Prevalence of and Factors Associated With High Blood Pressure at 15 Years of Age: A Birth Cohort Study. J Am Heart Assoc. 2023 Dec 5;12(23):e029627.
28. Maric GD, Dusanovic MG, Kostic A V., Pekmezovic TD, Kisic-Tepavcevic DB. Prevalence of hypertension in a sample of schoolchildren in the Belgrade district. Blood Press Monit. 2016 Jun;21(3):155–9.
29. Saeed NY, Al-Ani MM, Khudhur WY. Prevalence of hypertension among intermediate school children in Ramadi city, west of Iraq. Journal of Emergency Medicine, Trauma and Acute Care. 2022 Dec 1;2022(6):14.
30. Badi MAH, Garcia-Triana BE, Suarez-Martinez R. Overweight/obesity and hypertension in schoolchildren aged 6–16 years, Aden Governorate, Yemen, 2009. Eastern Mediterranean Health Journal. 2012;18(7):718–22.
31. Soua S, Ghammam R, Maatoug J, Zammit N, Ben Fredj S, Martinez F, et al. The prevalence of high blood pressure and its determinants among Tunisian adolescents. J Hum Hypertens. 2022 Apr 8;38(4):371–9.
32. Bandy A, Qarmush MM, Alrwilly AR, Albadi AA, Alshammari AT, Aldawasri MM. Hypertension and its risk factors among male adolescents in intermediate and secondary schools in Sakaka City, Aljouf Region of Saudi Arabia. Niger J Clin Pract. 2019;22(8):1140–1146.
33. Abdulle A, Al-Junaibi A, Nagelkerke N. High Blood Pressure and its Association with Body Weight among Children and Adolescents in the United Arab Emirates. PLoS One. 2014 Jan 20;9(1):e85129.
34. Murad NS, Miro SS, Ismael VAH, Abdulah DM. Modifiable risk factors for cardiovascular disease in Iraqi Kurdistan population: a large epidemiological study. Healthcare in Low-Resource Settings. 2024 Dec 19;12(2):12087.
35. Qadir MS, M. Weli S. Prevalence of Cardiovascular Disease Risk Factors Among Secondary School Pupils in Sulaimani City Kurdistan-Iraq. A Cross-Sectional Study. J Fac Med Baghdad. 2023 Jul 1;65(2):129–34.
36. Sudikno S, Mubasyiroh R, Rachmalina R, Arfines PP, Puspita T. Prevalence and associated factors for prehypertension and hypertension among Indonesian adolescents: a cross-sectional community survey. BMJ Open. 2023 Mar 23;13(3):e065056.
37. Madsen H, Sen A, Aune D. Fruit and vegetable consumption and the risk of hypertension: a systematic review and meta-analysis of prospective studies. Eur J Nutr. 2023 Aug 27;62(5):1941–55.
38. Leyvraz M, Chatelan A, da Costa BR, Taffé P, Paradis G, Bovet P, et al. Sodium intake and blood pressure in children and adolescents: a systematic review and meta-analysis of experimental and observational studies. Int J Epidemiol. 2018 Dec 1;47(6):1796–810.
39. Liu X, Zhang D, Liu Y, Sun X, Han C, Wang B, et al. Dose–Response Association Between Physical Activity and Incident Hypertension. Hypertension. 2017 May;69(5):813–20.
40. Lee PH, Wong FKY. The Association Between Time Spent in Sedentary Behaviors and Blood Pressure: A Systematic Review and Meta-Analysis. Sports Medicine. 2015 Jun 8;45(6):867–80.
41. Hall JE, da Silva AA, do Carmo JM, Dubinion J, Hamza S, Munusamy S, et al. Obesity-induced Hypertension: Role of Sympathetic Nervous System, Leptin, and Melanocortins. Journal of Biological Chemistry. 2010 Jun;285(23):17271–6.

42. 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 (IJ-AI). 2023 Jun 1;12(2):776.

43. Hwang SH, Lee H, Lee JH, Lee M, Koyanagi A, Smith L, et al. Machine Learning–Based Prediction for Incident Hypertension Based on Regular Health Checkup Data: Derivation and Validation in 2 Independent Nationwide Cohorts in South Korea and Japan. J Med Internet Res. 2024 Nov 5;26:e52794.