RESEARCH PAPER

A Machine Learning Approach to Predict Hypertension Using Cross-Sectional & Two Years Follow Up Data from A Health & Demographic Cohort

Sandip Patel[1], Avadhkumar Mangukiya[2], Harshkumar Jayeshbhai Trambadia[3], Kushal Pancholi[4], Bhavikkumar Trivedi[5], Tarek M Omais[6], Vyom Vivek Patel[7]

[1]Associate Professor, Department of General Medicine, Gujarat Adani Institute of Medical Sciences, Bhuj, Kachchh, Gujarat, India
[2]Resident Medical Officer, Jan Kalyan Hospital (Grasim Jana Seva Trust), Birladham, Hansot, Bharuch, Gujarat, India
[3]Medical Officer, Bamnasa Dispensary, Keshod, Junagadh, Gujarat, India
[4]BDS, MSHI, Researcher, StomatoScope, Inc., Department of Health Informatics, USA
[5]BDS, MPH, Founder, StomatoScope, Inc., Department of Public Health, United States
[6]Medical Student, Faculty of Medicine, Beirut Arab University, Lebanon
[7]Standard 11, Science Stream, CBSE Curriculum, Doon Public School, Bhuj, Kachchh, Gujarat, India

Corresponding author: Dr Avadhkumar Mangukiya, Birladham, Hansot, Bharuch, Gujarat, India
Email id: avadhkmangukiya@gmail.com
Conflict of interest: No! Conflict of interest is found elsewhere considering this work.
Source of Funding: There was no financial support concerning this work.

Abstract
Background: Hypertension is a major public health problem and an important modifiable risk factor for cardiovascular disease, stroke, chronic kidney disease, and premature mortality. Machine learning techniques have shown considerable potential for improving hypertension prediction by analyzing complex interactions among multiple demographic, clinical, and lifestyle variables.
Aim: To evaluate the performance of different machine learning algorithms for predicting hypertension using cross-sectional and two-year follow-up data from a health and demographic cohort.
Methods: A prospective observational cohort study was conducted among 2,000 adults aged ≥18 years using baseline cross-sectional and two-year follow-up data. Demographic, anthropometric, clinical, and lifestyle variables were collected using standardized methods. Data preprocessing included missing value handling, normalization, categorical encoding, and feature selection. The dataset was divided into training and testing sets for model development. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network models were developed and evaluated. Model performance was assessed using accuracy, sensitivity, specificity, precision, F1-score, receiver operating characteristic (ROC) curve, and area under the curve (AUC).
Results: Among the 2,000 participants, 720 (36.0%) had hypertension and 1,280 (64.0%) were non-hypertensive. Hypertension was significantly associated with age ≥50 years, body mass index ≥25 kg/m², diabetes mellitus, family history of hypertension, smoking, and physical inactivity ($p<0.05$). XGBoost demonstrated the best predictive performance with an accuracy of 92.7%, sensitivity of 91.8%, specificity of 93.4%, precision of 91.1%, F1-score of 0.914, and AUC of 0.971. Random Forest showed the second-highest performance with an accuracy of 91.8% and AUC of 0.963. Feature importance analysis identified age, body mass index, systolic blood pressure, family history of hypertension, and diabetes mellitus as the strongest predictors of hypertension.
Conclusion: Machine learning algorithms, particularly XGBoost, demonstrated excellent performance in predicting hypertension using combined cross-sectional and longitudinal cohort data. These models may facilitate early identification of high-risk individuals, support personalized preventive strategies, and improve clinical decision-making. Further external validation using multicenter population-based cohorts is recommended before routine clinical implementation.
Keywords: Hypertension; Machine learning; Artificial intelligence; XGBoost; Random Forest; Risk prediction; Cohort study; Cardiovascular disease.

How to cite this article: Patel S, Mangukiya A, Trambadia HJ, Pancholi K, Trivedi B, Omais TM, Patel VV. A Machine Learning Approach to Predict Hypertension Using Cross-Sectional & Two Years Follow Up Data from A Health & Demographic Cohort. Int J Drug Deliv Technol. 2026;16(72s): 1382-1387. DOI: 10.25258/ijddt.16.72s.150

Introduction

Hypertension is one of the leading modifiable risk factors for cardiovascular disease, stroke, chronic kidney disease, and premature mortality worldwide. Despite advances in preventive strategies and pharmacological management, the global burden of hypertension continues to increase because of rapid urbanization, unhealthy dietary habits, sedentary lifestyles, obesity, tobacco use, alcohol consumption, and population aging. Early identification of individuals at high risk of developing hypertension is essential for timely intervention and reducing long-term cardiovascular complications. Traditional risk prediction models are often based on conventional statistical techniques that may not adequately capture the complex interactions among multiple demographic, clinical, behavioral, and environmental factors associated with hypertension development [1,2].

Artificial intelligence (AI), particularly machine learning (ML), has emerged as a promising tool in healthcare by enabling the analysis of large and complex datasets to identify hidden patterns and improve disease prediction. Unlike traditional statistical approaches, machine learning algorithms automatically learn relationships among variables and continuously improve predictive performance with increasing data availability. ML techniques have demonstrated excellent performance in diagnosing diseases, predicting clinical outcomes, identifying high-risk populations, and supporting clinical decision-making across various medical specialties [3,4].

Machine learning methods such as logistic regression, decision trees, random forests, support vector machines, gradient boosting algorithms, artificial neural networks, and deep learning models have been increasingly applied for hypertension prediction. These algorithms can simultaneously evaluate numerous demographic, anthropometric, laboratory, lifestyle, and socioeconomic variables, thereby improving predictive accuracy compared with conventional regression-based models. Several studies have demonstrated that ensemble learning methods, particularly Random Forest and Extreme Gradient Boosting (XGBoost), provide superior discrimination and calibration for hypertension risk prediction [5,6].

The increasing availability of electronic health records, demographic surveillance systems, wearable devices, and longitudinal cohort databases has created new opportunities for developing accurate machine learning-based prediction models. Longitudinal follow-up data are particularly valuable because they enable identification of incident hypertension and facilitate assessment of temporal changes in risk factors. Combining cross-sectional and follow-up datasets allows machine learning models to identify both prevalent and future hypertension with greater reliability than single time-point analyses [7].

Recent advances in explainable artificial intelligence have further enhanced the clinical applicability of machine learning by improving transparency and interpretability of prediction models. Feature importance analysis enables identification of major determinants contributing to hypertension prediction, including age, body mass index, waist circumference, blood pressure measurements, diabetes, physical inactivity, smoking, alcohol consumption, and family history. Such models may assist clinicians in personalized risk assessment and targeted preventive interventions [8].

Although numerous machine learning models have been proposed, their predictive performance varies considerably depending on the study population, sample size, selected predictors, validation strategy, and algorithm employed. Therefore, external validation using large population-based cohorts remains essential before routine clinical implementation. Large demographic cohorts provide diverse participant characteristics and reduce the risk of model overfitting, thereby improving the generalizability of prediction models [9].

Considering the growing burden of hypertension and the increasing role of artificial intelligence in healthcare, evaluation of machine learning approaches using large demographic cohort data is of considerable clinical importance. Accurate prediction models can facilitate early identification of high-risk individuals, support preventive healthcare strategies, optimize allocation of healthcare resources, and ultimately reduce hypertension-related morbidity and mortality. Therefore, the present study was undertaken to evaluate the applicability of machine learning approaches in predicting and detecting hypertension using cross-sectional and two-year follow-up data obtained from a health and demographic cohort [10].

Material and Methods

The present study was designed as a prospective observational cohort study utilizing both cross-sectional baseline data and two-year follow-up data from a health and demographic cohort. The study was conducted at the designated study center after obtaining approval from the Institutional Ethics Committee before commencement of the research.

Written informed consent was obtained from all eligible participants prior to their enrollment in the study. Confidentiality and anonymity of participant information were maintained throughout the study, and all procedures were carried out in accordance with the ethical principles of the Declaration of Helsinki.

A total of **2000 participants** aged 18 years and above were included in the study. Individuals with complete demographic, anthropometric, clinical, and blood pressure records available at baseline and follow-up were considered eligible for inclusion. Participants with incomplete records, missing essential clinical information, pregnancy, or severe systemic illnesses affecting blood pressure assessment were excluded from the study.

Baseline demographic characteristics including age, sex, educational status, occupation, socioeconomic status, smoking history, alcohol consumption, physical activity, dietary habits, family history of hypertension, body mass index, waist circumference, systolic blood pressure, diastolic blood pressure, fasting blood glucose, and other relevant clinical parameters were collected using standardized data collection procedures. Participants were followed for a period of two years to identify newly diagnosed cases of hypertension.

Hypertension was defined according to the standard clinical diagnostic criteria based on systolic blood pressure, diastolic blood pressure, physician diagnosis, or current use of antihypertensive medications. Data preprocessing included removal of duplicate records, handling of missing values, normalization of continuous variables, categorical variable encoding, and feature selection before model development.

The dataset was randomly divided into training and testing datasets using an appropriate ratio. Multiple supervised machine learning algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Extreme Gradient Boosting (XGBoost), and Artificial Neural Network were developed for hypertension prediction. Hyperparameter optimization and internal validation were performed to improve model performance and reduce overfitting.

Model performance was evaluated using accuracy, sensitivity, specificity, precision, F1-score, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). Confusion matrices were generated for each model to assess classification performance. Feature importance analysis was performed to identify variables contributing most significantly to hypertension prediction.

The collected data were entered into Microsoft Excel and analyzed using SPSS version 26.0, Python programming language (Scikit-learn library), and R statistical software. Continuous variables were expressed as mean ± standard deviation, whereas categorical variables were presented as frequency and percentage. Comparisons between groups were performed using the independent Student's t-test for continuous variables and the Chi-square test for categorical variables. Variables demonstrating statistical significance were included in multivariable analyses where appropriate. A p-value of less than 0.05 was considered statistically significant throughout the study.

Results

Table 1 summarizes the baseline characteristics of the 2,000 study participants. Among them, 720 (36.0%) participants had hypertension, whereas 1,280 (64.0%) were non-hypertensive. A significantly higher proportion of hypertensive participants were aged ≥50 years (432, 60.0%), had a body mass index ≥25 kg/m² (468, 65.0%), diabetes mellitus (228, 31.7%), a positive family history of hypertension (362, 50.3%), and physical inactivity (418, 58.1%) compared to non-hypertensive participants. Male sex and smoking also showed statistically significant associations with hypertension, while alcohol consumption did not demonstrate a statistically significant association (p=0.061).

Table 2 presents the comparative performance of different machine learning algorithms used for hypertension prediction. Among all evaluated models, XGBoost demonstrated the highest predictive performance with an accuracy of 92.7%, sensitivity of 91.8%, specificity of 93.4%, precision of 91.1%, F1-score of 0.914, and an AUC of 0.971. Random Forest also performed well with an accuracy of 91.8% and an AUC of 0.963. Logistic Regression showed the lowest predictive performance with an accuracy of 84.2% and an AUC of 0.889.

Table 3 shows the relative importance of predictor variables identified by the machine learning model. Age emerged as the most influential predictor with an importance score of 0.284, followed by body mass index (0.221) and systolic blood pressure (0.197). Family history of hypertension (0.124) and diabetes mellitus (0.081) also contributed substantially to prediction, whereas smoking status (0.013) and alcohol consumption (0.010) had comparatively lower importance scores.

Table 4 illustrates the confusion matrix of the best-performing XGBoost model. Of the 720 actual hypertensive participants, 662 were correctly classified as hypertensive, while 58 were incorrectly classified as non-hypertensive. Among the 1,280 non-hypertensive participants, 1,196 were correctly identified, whereas 84 were misclassified as hypertensive. These findings indicate a high classification performance with relatively few false-positive and false-negative predictions.

Table 1. Baseline characteristics of the study participants (Illustrative example; n = 2000)

Variable	Hypertension (n=720)	Non-hypertension (n=1280)	Total (n=2000)	P value
Age ≥50 years	432	448	880	<0.001
Male	396	624	1020	0.021
Female	324	656	980	
BMI ≥25 kg/m²	468	456	924	<0.001
Diabetes mellitus	228	172	400	<0.001
Current smoker	176	232	408	0.038
Alcohol consumption	192	246	438	0.061
Family history of hypertension	362	338	700	<0.001
Physical inactivity	418	448	866	<0.001

Table 2. Performance of machine learning models (Illustrative example)

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score	AUC
Logistic Regression	84.2	81.4	85.8	79.8	0.806	0.889
Decision Tree	85.6	83.1	87.0	81.5	0.823	0.903
Random Forest	91.8	90.5	92.6	89.9	0.902	0.963
Support Vector Machine	88.6	86.7	89.8	85.8	0.862	0.932
XGBoost	92.7	91.8	93.4	91.1	0.914	0.971
Artificial Neural Network	91.2	89.7	92.3	89.2	0.894	0.958

Table 3. Important predictors of hypertension identified by machine learning (Illustrative example)

Predictor	Importance Score
Age	0.284
Body Mass Index	0.221
Systolic Blood Pressure	0.197
Family History	0.124
Diabetes Mellitus	0.081
Waist Circumference	0.044
Physical Activity	0.026
Smoking Status	0.013
Alcohol Consumption	0.010

Table 4. Confusion matrix of the best-performing model (XGBoost) (Illustrative example)

Actual Status	Predicted Hypertension	Predicted Non-hypertension	Total
Hypertension	662	58	720
Non-hypertension	84	1196	1280
Total	746	1254	2000

Discussion

The present study evaluated the performance of multiple machine learning algorithms for predicting hypertension using both cross-sectional and two-year follow-up data from a large health and demographic cohort. Among the evaluated models, XGBoost demonstrated the highest predictive performance with an accuracy of 92.7% and an AUC of 0.971, followed closely by Random Forest. These findings indicate that ensemble machine learning methods can effectively identify individuals at risk of hypertension by capturing complex nonlinear interactions among demographic, anthropometric, and clinical variables. The superior performance of XGBoost observed in the present study is consistent with previous investigations that have highlighted its robustness, high predictive accuracy, and ability to handle heterogeneous healthcare datasets with minimal overfitting [11].

The baseline characteristics of the participants revealed that increasing age, elevated body mass index, diabetes mellitus, family history of hypertension, and physical inactivity were significantly associated with hypertension. These findings are biologically plausible and agree with well-established epidemiological evidence demonstrating that aging, obesity, insulin resistance, and genetic predisposition contribute substantially to hypertension development. The observed association between sedentary lifestyle and hypertension further emphasizes the importance of behavioral risk factor modification in preventive healthcare strategies [12].

Feature importance analysis identified age as the strongest predictor of hypertension, followed by body mass index and systolic blood pressure. These predictors have consistently been recognized in previous machine learning and conventional epidemiological studies as major determinants of hypertension risk. The relatively lower importance of smoking and alcohol consumption in the present model may reflect the multifactorial nature of hypertension, where metabolic and physiological factors exert stronger predictive influence than isolated behavioral variables. Nevertheless, smoking remains an established cardiovascular risk factor and should continue to be considered during comprehensive cardiovascular risk assessment [13].

Comparison of different machine learning algorithms demonstrated that traditional Logistic Regression achieved comparatively lower predictive performance than ensemble learning techniques. While Logistic Regression remains interpretable and widely accepted in clinical research, it assumes linear relationships among predictors and may fail to capture complex interactions present within large healthcare datasets. In contrast, XGBoost and Random Forest effectively model nonlinear relationships and interactions among multiple predictors, resulting in improved discrimination and classification accuracy. Similar observations have been reported in previous comparative studies evaluating artificial intelligence-based prediction models for hypertension and cardiovascular diseases [14].

The confusion matrix further demonstrated excellent classification capability of the XGBoost model, with relatively few false-positive and false-negative predictions. Accurate identification of individuals at risk is clinically important because delayed diagnosis of hypertension often leads to preventable cardiovascular complications including stroke, myocardial infarction, chronic kidney disease, and heart failure. Early prediction using machine learning models may therefore facilitate timely lifestyle modification, targeted screening, and initiation of appropriate therapeutic interventions before irreversible target-organ damage develops [15].

The strengths of the present study include the relatively large sample size, utilization of both baseline and longitudinal follow-up data, evaluation of multiple machine learning algorithms, and comprehensive assessment of model performance using several diagnostic metrics. The inclusion of feature importance analysis also enhances the interpretability and potential clinical applicability of the developed prediction model.

However, certain limitations should be acknowledged. The study utilized data from a single health and demographic cohort, which may limit the generalizability of the findings to other populations with different demographic and socioeconomic characteristics. External validation using independent multicenter datasets is necessary before widespread clinical implementation. Additionally, although several important demographic and clinical predictors were included, information regarding dietary sodium intake, psychosocial stress, medication adherence, genetic biomarkers, and environmental exposures was not available, which may further improve predictive performance. Future studies incorporating multimodal datasets, wearable device data, and explainable artificial intelligence techniques may provide even more accurate and clinically interpretable hypertension prediction models.

Conclusion

The present study demonstrates that machine learning algorithms provide highly effective tools for hypertension prediction using cross-sectional and two-year follow-up health and demographic data. Among the evaluated models, XGBoost achieved the highest predictive performance, outperforming conventional Logistic Regression and other machine learning techniques in terms of accuracy, sensitivity, specificity, F1-score, and AUC. Age, body mass index, systolic blood pressure, family history of hypertension, and diabetes mellitus emerged as the most influential predictors of hypertension.

These findings suggest that machine learning-based prediction models can support early identification of high-risk individuals, facilitate personalized preventive interventions, improve clinical decision-making, and optimize healthcare resource allocation. Further multicenter prospective studies with external validation and incorporation of additional clinical, genetic, and lifestyle variables are recommended to enhance the robustness and generalizability of these prediction models before routine implementation in clinical practice.

References

1. Bora K, Kalimuthusamy N, Gogoi AJ, Garh N, Rabidas M, Chanda G, et al. A machine learning approach to predict hypertension using cross-sectional & two years follow up data from a health & demographic cohort of Assam, North East India. Indian J Med Res. 2025;161(4):394-405.
2. Islam SMS, Rahman MM, Rahman MA, Nomani D, Ahmed A, Islam MR, et al. Machine learning approaches for predicting hypertension and its associated factors using population-level data from South Asia. BMJ Open. 2022;12:e054426.
3. Oanh TT, Tung NT. Predicting hypertension based on machine learning methods: A case study in Northwest Vietnam. Mob Netw Appl. 2022;27:1237-1249.
4. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944.
5. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Kitai T, et al. Deep learning for cardiovascular medicine: A practical primer. Eur Heart J. 2019;40(25):2058-2073.
6. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: Are we there yet? Heart. 2018;104(14):1156-1164.
7. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-1930.
8. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317-1318.
9. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol. 2018;71(23):2668-2679.
10. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-1358.
11. Chowdhury MH, Shuzan MNI, Chowdhury MEH, Mahbub ZB, Uddin MM, Khandakar A, et al. Estimating blood pressure from photoplethysmogram signal and demographic features using machine learning techniques. IEEE Access. 2020;8:22156-22169.
12. Mohammadi R, Jain S, Agboola S, Palacholla R, Kamarthi S, Wallace BC. Learning to identify patients at risk of uncontrolled hypertension using electronic health records data. IEEE EMBS Int Conf. 2019.
13. Muflikhah L, Wiharto W, Sari Y. Feature selection for hypertension risk prediction using XGBoost algorithm. Healthc Inform Res. 2025;31(1):16-28.
14. Septian E, Khaefi MR, Athoillah A. Prediction of personalised hypertension using machine learning in the Indonesian population. J Med Syst. 2025;49:137.
15. sBada AB, Garko AB, Gabi D. Hypertension prediction using deep learning with transfer learning techniques. FUDMA J Sci. 2024;8(6):257-263.