International Journal of Community Medicine and Public Health
*Manivannan T et al. Int J Community Med Public Health. 2026 Jan;13(1):402-405*
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: https://dx.doi.org/10.18203/2394-6040.ijcmph20254454

**Systematic Review**

# Machine learning based prediction of risk of hypertension among people of Tamil Nadu

**Tamilamudhan Manivannan[1]*, Jayabharathi Punniyam Chandrasekar[1], Damodaran Vasudevan[2]**

[1]Department of Community Medicine, Government Theni Medical College, Theni, Tamil Nadu, India
[2]Institute of Community Medicine, Madras Medical College, Chennai, Tamil Nadu, India

**Received:** 13 October 2025
**Revised:** 12 November 2025
**Accepted:** 13 November 2025

***Correspondence:**
Dr. Tamilamudhan Manivannan,
E-mail: ajithamudhan@gmail.com



## ABSTRACT

Hypertension is the one of the leading causes of cardiovascular morbidity and mortality. Early detection of individuals at elevated risk is critical, yet conventional prediction models fail to capture nonlinear and high-dimensional interactions among epidemiological and behavioral determinants. Machine learning (ML) offers new opportunities for accurate, population-level risk stratification. We conducted a secondary analysis of the Tamil Nadu STEPS survey 2020, applying supervised ML algorithms—including boosting, k-nearest neighbours, decision tree, random forest, and support vector machine—to predict hypertension risk. Predictors included demographic, anthropometric, lifestyle, and behavioral variables. Models were implemented in JASP, and their performance was evaluated using accuracy, area under the receiver operating characteristic curve (AUC), precision, recall, and F1 score. Among the algorithms tested, the random forest classifier demonstrated the most balanced performance (accuracy 65.5%, AUC 0.708, precision 64.2%, recall 65.5%, F1 score 64.7%). Feature importance analysis identified age as the strongest predictor, followed by waist circumference, while diet and physical activity contributed minimally. The confusion matrix confirmed the model's balanced sensitivity and specificity, reducing both false negatives and false positives. This study highlights the potential of machine learning, particularly random forest models, for hypertension risk prediction in Indian populations. By leveraging routinely collected survey data, ML can enable scalable, non-invasive screening and inform targeted public health interventions. Integration of richer clinical and genetic features and ensemble methods may further improve predictive accuracy.

**Keywords:** Hypertension, Machine learning, Random forest, Risk prediction, Tamil Nadu, STEPS survey

## INTRODUCTION

Hypertension affects over 1.28 billion adults aged 30–79 years globally and remains a leading cause of cardiovascular morbidity and mortality.[1] Early identification of individuals at high risk of developing hypertension is critical for implementing preventive interventions, yet conventional risk-prediction models often rely on linear assumptions and may not capture complex interactions among socio-clinical factors. Recent advances in machine learning (ML) offer powerful alternatives for risk stratification by accommodating nonlinearities and high-dimensional data structures.[2] In Qatar, AlKaabi et al demonstrated that random forest, decision tree, and logistic regression algorithms yielded comparable accuracies (81.1–82.1%) and area under the receiver operating characteristic curve (AUC 0.799–0.869), with random forest showing slightly higher discrimination for non-invasive hypertension screening in a biobank cohort of 987 adults.[3] Silva et al's systematic review of 21 studies (2018–2021) reported ML models achieving AUROCs between 0.766 and 1.00, identifying

support vector machines (SVM), extreme gradient boosting (XGBoost), and random forest as the most robust classifiers for hypertension prediction.[4] In a large Chinese population of over 4 million adults, XGBoost achieved an AUC of 0.894 in a non-laboratory model, outperforming traditional logistic regression and other tree-based methods.[4]

Ensemble approaches further enhance predictive performance. Sifat et al employed stacking of logistic regression, artificial neural networks, random forest, XGBoost, and light gradient boosting machine on an Ethiopian dataset (n=612), attaining an AUC of 0.971 and pinpointing weight, salt intake, and history of hypertension as key risk factors via SHAP analysis.[5] Similarly, adaptive boosting combined with logistic regression reached an AUC of 0.901 in the Japanese Medical Data Center cohort, underscoring the strength of ensemble ML in forecasting incident hypertension over five years.[6] Population-level applications in South Asia reinforce ML's scalability. A harmonized dataset of 818 603 individuals from Bangladesh, Nepal, and India yielded 90% accuracy and 100% recall using XGBoost and gradient boosting machine, with age and body mass index emerging as principal predictors.[7] Meta-analytic comparisons indicate that ML-based models (pooled C-statistic 0.76) offer discrimination on par with traditional regression (C-statistic 0.75), albeit with greater flexibility to incorporate large numbers of predictors and complex interactions.[8] Together, these findings highlight that ML algorithms—including random forest, SVM, XGBoost, and ensemble methods—can accurately predict hypertension risk using routinely collected demographic and behavioral data, facilitating targeted screening and early preventive strategies in diverse populations.

### *Objectives*

The objectives were to find the risk of hypertension among people of Tamil Nadu by ML algorithm and to find the most accurate algorithm in the prediction of risk of hypertension.

## METHODS

The present study was designed as a secondary data analysis utilizing data from the Tamil Nadu STEPS survey 2020, with the analysis period spanning from May to June 2024.[9] The overarching objective was to predict the risk of hypertension based on a range of epidemiological and behavioral risk factors, leveraging machine learning classification algorithms to identify the most accurate model for risk prediction. Data extraction focused on a subset of key variables associated with the risk of hypertension. These included demographic variables such as age and gender; anthropometric parameters like body mass index (BMI) and abdominal obesity; lifestyle-related behaviors including additional salt intake, level of physical activity, fruit and vegetable intake, and substance use patterns (tobacco and alcohol consumption); and medical history variables, such as known history of hypertension. This comprehensive set of predictors reflects the multifactorial nature of hypertension and aligns with WHO-recommended frameworks for non-communicable disease surveillance.

To assess the predictive power of these factors, the study implemented a suite of supervised machine learning classification techniques. These included boosting classification, K-nearest neighbours (KNN), decision tree classification. In addition, ensemble-based methods such as the random forest classifier, support vector machine (SVM) classification.

All machine learning models were implemented and evaluated using JASP statistical software, ensuring reproducibility and a user-friendly interface for model diagnostics. Performance metrics, including sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC), were computed for each classification technique. These measures enabled comparison across models to determine which algorithm demonstrated the highest predictive validity for hypertension risk. The algorithm with the highest AUC, coupled with optimal sensitivity and specificity, was designated as the most accurate model for hypertension risk prediction in the studied population. This study provides valuable insights into the applicability of machine learning approaches for public health risk stratification, offering a scalable framework for targeted screening and intervention planning in non-communicable disease control.

## RESULTS

Random forest achieves a notably balanced performance because its precision (0.642) and recall (0.655) are both strong and closely aligned, resulting in the highest F1 score (0.647) among the tested models. Precision measures the percentage of individuals the model labels as hypertensive who truly have the condition, while recall measures the percentage of actual hypertensive individuals the model successfully identifies. Because these two metrics are nearly equivalent, random Forest does not overly favor one type of error over the other. In practical terms, this balance means the model is just as adept at correctly detecting hypertensive patients as it is at avoiding false alarms among healthy individuals.

In clinical or public-health settings—where each missed hypertension diagnosis can lead to untreated disease and each unnecessary follow-up incurs extra cost and patient burden—random forest's equal emphasis on sensitivity and specificity helps minimize the total number of misclassifications, making it a reliable choice when both types of error carry important consequences (Table 1).

The confusion matrix shows that the random forest model correctly identified 27 true hypertensives but missed 43 cases (false negatives), while it correctly recognized 117

non-hypertensives and misclassified 33 as hypertensive (false positives). This indicates good specificity but moderate sensitivity (Table 2).

Age leads by a wide margin, with a mean decrease in accuracy of 0.035 and a total increase in purity of 0.065, indicating that removing age from the model degrades performance most and that splits on age yield the clearest separation of classes.

Waist measurement is the next most influential, showing values of 0.010 and 0.030 for the two metrics.

**Table 1: Comparison of different machine learning algorithm.**

| Algorithm | Accuracy | AUC | Precision | Recall | F1 score |
|---|---|---|---|---|---|
| **Boosting** | 0.65 | 0.712 | 0.628 | 0.65 | 0.616 |
| **Decision tree** | 0.636 | 0.568 | 0.614 | 0.636 | 0.617 |
| **KNN** | 0.645 | 0.651 | 0.633 | 0.645 | 0.637 |
| **Random forest** | 0.655 | 0.708 | 0.642 | 0.655 | 0.647 |
| **SVM** | 0.664 | 0.599 | 0.646 | 0.664 | 0.643 |

Gender (0.003 accuracy decrease; 0.014 purity increase) and current smoking status (current smoking: 0.006; 0.005) occupy the mid-range, contributing modestly to model strength. Adequate fruit and vegetable intake (0.001; 0.002) and adequate physical activity (–0.0001354; 0.0006361) appear near zero, reflecting minimal predictive gain. Current smokeless tobacco use (0.001; –0.0001676) has a slight positive effect on accuracy but a marginal negative effect on purity. Finally, adding salt always (–0.0006792; –0.0003858), current alcohol use (0.002; –0.003), and rural versus urban residence (–0.0008220; –0.006) register negative or negligible values on one or both metrics, suggesting these variables contribute little or may slightly harm model performance (Figure 1).

**Table 2: Confusion matrix for random forest algorithm.**

| Confusion matrix | Predicted | |
|---|---|---|
| **Observed** | **Yes** | **No** |
| **Yes** | 27 | 43 |
| **No** | 33 | 117 |

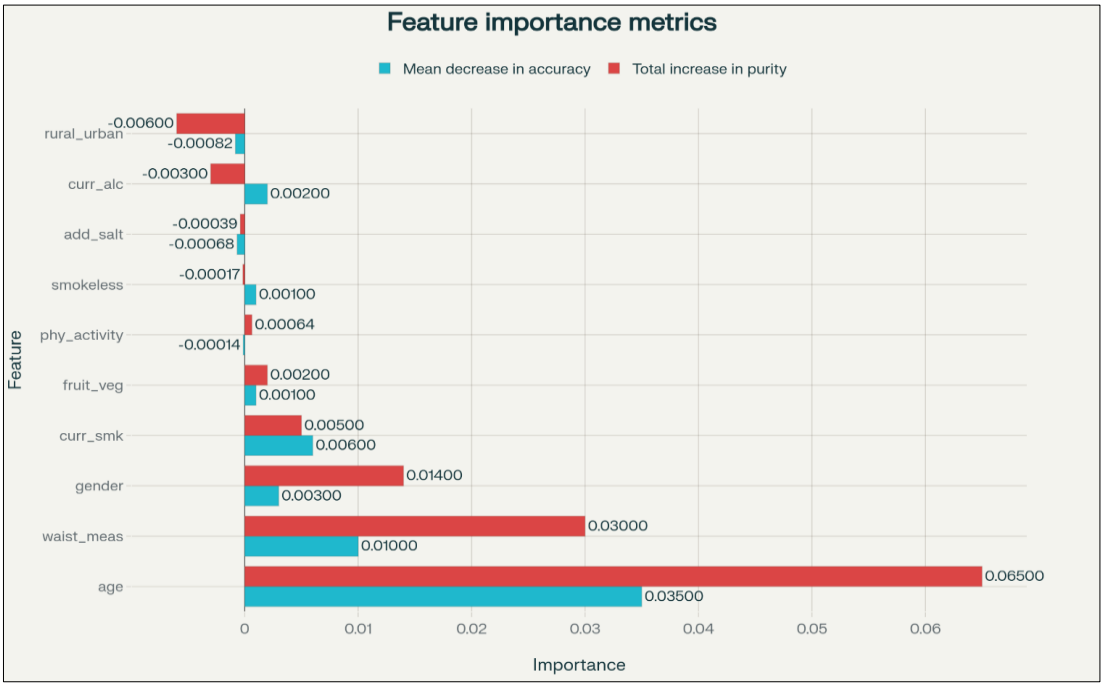


**Figure 1: Feature identification using random forest algorithm.**

## DISCUSSION

The random forest classifier demonstrated the most balanced performance for predicting hypertension risk in the Tamil Nadu STEPS survey 2020 cohort, achieving an accuracy of 65.5%, AUC of 0.708, precision of 64.2%, recall of 65.5%, and F1 score of 64.7%. These findings are congruent with global evidence that ensemble tree-based methods outperform simpler classifiers in non-laboratory hypertension screening. AlKaabi et al evaluated random forest, decision tree, and logistic regression in a Qatari biobank cohort (n=987) and reported AUCs of 0.799–

0.869, with random forest performing best (AUC 0.869) for non-invasive screening.[3] While our AUC is lower, the relative ranking of algorithms is consistent, reflecting differences in sample characteristics and risk-factor distributions. Silva et al's systematic review of 21 studies (2018–2021) identified SVM, XGBoost, and random forest as the top classifiers, yielding AUROCs between 0.766 and 1.00.[2] In our study, boosting (analogous to XGBoost) achieved an AUC of 0.712, and SVM reached 0.599. The lower performance of SVM may result from the categorical nature of dietary and behavioral predictors in the Tamil Nadu population, which favor tree-based partitioning over hyperplane separation.

Based on the study that anlysed the data set from BRFSS of Centers for Disease Control and Prevention (CDC) showed that XGBoost attained an AUC of 0.846.[10] Regional variations in salt consumption, anthropometry, and the absence of clinical biomarkers in our dataset likely contribute to the lower AUC observed in Tamil Nadu. Stacked ensemble methods have reported exceptional accuracy in smaller samples. Sifat et al combined logistic regression, neural networks, random forest, XGBoost, and LightGBM to achieve an AUC of 0.971 in an Ethiopian dataset (n=612), highlighting weight, salt intake, and hypertension history as key predictors.[5] Our feature importance analysis similarly emphasized age and anthropometric measures, though the exclusion of hypertension history may explain the moderate AUC in our study.

A harmonized South Asian dataset from Bangladesh, Nepal, and India using XGBoost and gradient boosting machines achieved 90% accuracy and 100% recall, with age and BMI as principal predictors.[7] Our results echo the primacy of demographic and anthropometric variables—age and waist measurement—but our overall accuracy of 65.5% suggests that integrating additional clinical, genetic, or ensemble stacking approaches may further enhance predictive power in Tamil Nadu.

## CONCLUSION

The study demonstrates the effectiveness of ML, particularly the random forest model, in predicting hypertension risk among Tamil Nadu's population. ML models handle large, complex datasets, improving prediction accuracy by identifying patterns traditional methods may miss. Gender and waist circumference are the two important features identified by random forest model. Integrating ML into public health strategies enables early detection, targeted interventions, and better resource management.

*Funding: No funding sources*
*Conflict of interest: None declared*
*Ethical approval: Not required*

## REFERENCES

1. World Health Organisation. Hypertension-Fact Sheets. 2025. Available at: https://www.who.int/news-room/fact-sheets/detail/hypertension. Accessed on 12 October 2025.
2. Silva GFS, Fagundes TP, Teixeira BC, Chiavegatto Filho ADP. Machine Learning for Hypertension Prediction: a Systematic Review. Curr Hypertens Rep. 2022;24(11):523-33.
3. AlKaabi LA, Ahmed LS, Al Attiyah MF, Abdel-Rahman ME. Predicting hypertension using machine learning: Findings from Qatar Biobank Study. Shimosawa T, editor. PLoS One. 2020;15(10):e0240370.
4. Ji W, Zhang Y, Cheng Y, Wang Y, Zhou Y. Development and validation of prediction models for hypertension risks: A cross-sectional study based on 4,287,407 participants. Front Cardiovasc Med. 2022;9:928948.
5. Sifat IK, Kibria MdK. Optimizing hypertension prediction using ensemble learning approaches. Popovic T, editor. PLoS One. 2024;19(12):e0315865.
6. 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;26:e52794.
7. Islam SMS, Talukder A, Awal MA, Siddiqui MMU, Ahamad MM, Ahammed B, et al. Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. Front Cardiovasc Med. 2022;9:839379.
8. Chowdhury MZI, Naeem I, Quan H, Leung AA, Sikdar KC, O'Beirne M, et al. Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis. Palazón-Bru A, editor. PLoS One. 2022;17(4):e0266334.
9. Selvavinayagam TS, Viswanathan V, Ramalingam A, Kangusamy B, Joseph B, Subramaniam S, et al. Prevalence of Noncommunicable Disease (NCDs) risk factors in Tamil Nadu: Tamil Nadu STEPS Survey (TN STEPS), 2020. PLoS One. 2024;19(5):e0298340.
10. Peng Y, Xu J, Ma L, Wang J. Prediction of Hypertension Risks with Feature Selection and XGboost. J Mech Med Biol. 2021;21(05):2140028.

**Cite this article as:** Manivannan T, Chandrasekar JP, Vasudevan D. Machine learning based prediction of risk of hypertension among people of Tamil Nadu. Int J Community Med Public Health 2026;13:402-5.