International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Predictive Modelling of Hypertension Risk Using Lifestyle and Clinical Biomarkers

ABDULLAH[1], ABDUL TABISH[2], AYAN AHMAD[3]

1Student, Dept. of CSE, Integral University, Lucknow, Uttar Pradesh, India
2Student, Dept. of CSE, Integral University, Lucknow, Uttar Pradesh, India
3Student, Dept. of CSE, Integral University, Lucknow, Uttar Pradesh, India

Under the Guidance of: Aaftab Alam, Assistant Professor, Dept. of CSE, Integral University, Lucknow, Uttar Pradesh, India

---

Abstract - Hypertension remains one of the most widespread chronic conditions globally, yet it is often undiagnosed because it typically develops without noticeable symptoms. Many individuals only become aware of their condition after severe complications such as cardiovascular or kidney-related events occur. This study explores the use of supervised machine learning models to identify individuals at risk of hypertension based on demographic, clinical, and lifestyle factors. Two algorithms were implemented: Logistic Regression, chosen for its interpretability, and XGBoost, selected for its ability to capture complex nonlinear relationships. Both models were trained using an 80/20 stratified split and evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. Results showed XGBoost consistently outperformed Logistic Regression, particularly in recall. Feature importance analysis revealed that modifiable lifestyle factors such as BMI, physical activity, diet, and smoking significantly influence prediction outcomes.

Key Words: Hypertension Prediction, Machine Learning, XGBoost, Logistic Regression, Lifestyle Factors, Healthcare Analytics

1.INTRODUCTION

Hypertension is a major public health concern due to its high prevalence and silent progression. Unlike many diseases that present noticeable symptoms, high blood pressure gradually damages vital organs such as the heart, kidneys, and brain over time without clear warning signs. As a result, many individuals remain undiagnosed until serious complications arise.

According to World Health Organization (2023), hypertension is one of the leading causes of preventable deaths worldwide. Its development is influenced by both non-modifiable factors (such as age and gender) and modifiable lifestyle behaviours (such as diet, physical activity, and smoking).

Traditional risk prediction methods often rely on linear assumptions and may fail to capture complex interactions between risk factors. Machine learning models, particularly ensemble methods like XGBoost, are better suited to handle such nonlinear relationships. This study compares Logistic Regression and XGBoost to evaluate whether advanced models provide meaningful improvements in hypertension prediction.

2.CONCEPTUAL RISK FRAMEWORK

Hypertension risk can be understood through three main categories:

2.1 Demographic Factors

Age is a strong predictor due to natural vascular aging, leading to reduced elasticity in blood vessels. Gender also plays a role, with hormonal differences influencing blood pressure regulation.

2.2 Clinical Measurements

Variables such as systolic and diastolic blood pressure and cholesterol levels provide direct insights into cardiovascular health and often reflect early stages of disease progression.

2.3 Lifestyle and Behavioral Factors

This includes BMI, smoking habits, physical activity, diet, and sleep patterns. These are especially important because they can be modified through intervention, making them key targets for prevention.

3.LITERATURE REVIEW

The scholarly foundation for hypertension risk research spans several decades of converging epidemiological evidence. Ezzati et al. (2015) demonstrated that blood pressure elevation accounts for a greater share of

preventable cardiovascular deaths than any other single modifiable risk factor, reframing hypertension as a global health policy priority.

Tzoulaki et al. (2016) challenged the assumption that hypertension risk operates as a binary threshold, revealing a continuous dose-response relationship with cardiovascular morbidity beginning well within the pre-hypertensive range.

Lloyd-Jones et al. (2010) demonstrated that simultaneous adoption of multiple lifestyle modifications produced synergistic blood pressure reductions exceeding the sum of individual interventions — a critical methodological insight for predictive modelling, since models assuming predictor independence systematically underestimate composite risk.

Warren et al. (2017) identified over one hundred genetic loci associated with blood pressure variability, yet their collective explanatory power remained modest, justifying emphasis on lifestyle features over genomic variables in the present model.

Dritsas and Trigka (2023) collectively suggested that gradient-boosted tree methods outperform logistic classifiers across health risk prediction tasks. Sudlow et al. (2015) and dietary research by Chan et al. (2016) further support the predictive value of richly annotated lifestyle features.

4.METHODOLOGY

4.1 Study Design

A supervised classification approach was employed to predict hypertension status, as described by Rahman et al. (2022). Two models were compared:

- Logistic Regression (baseline, interpretable)
- XGBoost (advanced, nonlinear model)

4.2 Variables

Demographic: Age, Gender | Clinical: SBP, DBP, Cholesterol | Lifestyle: BMI, Smoking, Physical Activity, Diet, Sleep | Target: Hypertension (Yes/No)

4.3 Data Preprocessing

- Outliers handled using IQR capping
- Categorical variables encoded using one-hot encoding
- Features scaled using Min-Max normalization
- Data split: 80% training, 20% testing

Table -1: Study Variables by Domain

Domain	Variable	Type	Role
Demographic	Age	Continuous	Predictor
Demographic	Biological Sex	Categorical (Binary)	Predictor
Clinical	Systolic BP (SBP)	Continuous	Predictor
Clinical	Diastolic BP (DBP)	Continuous	Predictor
Clinical	Total Serum Cholesterol	Continuous	Predictor
Lifestyle/Metabolic	BMI	Continuous	Predictor
Lifestyle/Metabolic	Tobacco Use Status	Categorical (Multi)	Predictor
Lifestyle/Metabolic	Physical Activity Level	Ordinal	Predictor
Lifestyle/Metabolic	Dietary Quality Score	Continuous	Predictor
Lifestyle/Metabolic	Avg. Sleep Duration	Continuous	Predictor
Outcome	Hypertension Diagnosis	Binary (0/1)	Target

5.MODELLING PIPELINE ARCHITECTURE

The end-to-end workflow followed these five steps:

1. Data collection and inspection
2. Cleaning and encoding
3. Feature scaling
4. Model training
5. Performance evaluation

Table -2: End-to-End Modelling Pipeline

Stage	Process	Key Action
1	Data Collection & Inspection	Variable validation, distributional analysis, completeness verification
2	Cleaning & Encoding	IQR-based outlier capping; one-hot and binary encoding
3	Feature Scaling	Min-max normalisation of continuous variables to [0,1]

4	Model Training	Logistic Regression (L2) and XGBoost (k-fold CV) on 80% partition
5	Performance Evaluation	Accuracy, Precision, Recall, F1-Score, AUC-ROC on 20% test set

6.MODEL DEVELOPMENT

6.1 Logistic Regression

Logistic Regression remains a foundational tool in clinical risk modeling, valued for its transparency and computational tractability (Hosmer et al., 2013). It estimates the log-odds of a binary outcome as a linear combination of input features, converted to class probability via the sigmoid transformation:

$\sigma(z) = 1/(1+e^{-}_{r})$, $z = \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n$

Each coefficient βi quantifies the change in log-odds for a unit of increase in the corresponding predictor. L2 regularization was applied to prevent overfitting.

6.2 XGBoost (Extreme Gradient Boosting)

XGBoost builds its predictive capacity incrementally by adding shallow decision trees, each trained on the residual prediction errors of the cumulative ensemble (Chen and Guestrin, 2016). The objective function contains a prediction loss term plus a complexity penalty, providing built-in regularisation that resists overfitting.

The clinical relevance lies in its capacity to model conditional interactions. An elevated BMI in a physically active individual may carry a materially different risk profile than the same BMI in a sedentary one interactions XGBoost captures naturally.

7.EVALUATION METRICS

The F1-score combines precision and recall into a single balanced measure. AUC-ROC evaluates the model's ability to distinguish between classes across all thresholds, ranging from 0.5 (random) to 1.0 (perfect) (Fawcett, 2006).

- Accuracy → overall correctness
- Precision → correctness of positive predictions
- Recall → ability to detect true cases
- F1-Score → balance between precision and recall
- AUC-ROC → overall discriminative performance

Table -3: Evaluation Metric Definitions

Metric	Formula	Clinical Significance
Accuracy	TP+TN/(TP+TN+FP+FN)	Overall correctness; limited with imbalanced classes
Precision	TP/(TP+FP)	Reliability of positive predictions; relevant when false alarms are costly
Recall (Sensitivity)	TP/(TP+FN)	Detection rate of true cases; critical for screening
F1-Score	2*(P*R)/(P+R)	Balanced summary when both FP and FN matter
AUC-ROC	Area under TPR vs. FPR	Threshold-independent discriminative capacity

8.RESULTS

XGBoost outperformed Logistic Regression across all evaluation metrics as shown in Table -4:

Table -4: Comparative Performance Results

Metric	Logistic Regression	XGBoost	Improvement (Δ)
Accuracy	0.81 (81.0%)	0.85 (85.0%)	+4.0 pp
Precision	0.79 (79.0%)	0.84 (84.0%)	+5.0 pp
Recall	0.80 (80.0%)	0.86 (86.0%)	+6.0 pp
F1-Score	0.79 (79.0%)	0.85 (85.0%)	+6.0 pp
AUC-ROC	0.82	0.87	+0.05

pp = percentage points

9.KEY FINDINGS AND FEATURE IMPORTANCE ANALYSIS

- Age and systolic blood pressure were the strongest predictors
- Among modifiable factors, BMI and physical activity had the highest impact
- Lifestyle factors collectively contributed significantly to predictions
- XGBoost captured variable interactions better than Logistic Regression

The XGBoost framework reveals amplification effects at the intersection of exposures. The compound risk of an overweight, physically inactive smoker reflects synergistic physiology reported by Lloyd-Jones et al. (2010), confirming that additive modelling systematically understates composite vulnerability.

Table -5: Feature Importance Rankings (XGBoost)

Rank	Feature	Domain	Modifiable?	Importance
1	Age	Demographic	No	Very High
2	Systolic Blood Pressure	Clinical	Partially	Very High
3	BMI	Lifestyle/Metabolic	Yes	High
4	Physical Activity Level	Lifestyle/Metabolic	Yes	High
5	Diastolic Blood Pressure	Clinical	Partially	Moderate-High
6	Total Serum Cholesterol	Clinical	Partially	Moderate
7	Dietary Quality Score	Lifestyle/Metabolic	Yes	Moderate
8	Tobacco Use Status	Lifestyle/Metabolic	Yes	Moderate
9	Sleep Duration	Lifestyle/Metabolic	Yes	Low-Moderate
10	Biological Sex	Demographic	No	Low

10.DISCUSSION

The results invite reflection on two related questions: what they reveal about the nature of hypertension risk, and what they suggest about the practical utility of machine learning in preventive clinical practice.

The variable importance study reinforces that hypertension is not simply an inevitable product of biology and age but is largely driven and modifiable by behaviors and environmental factors (Ezzati et al., 2015). The high predictive power of BMI, physical exercise, nutrition, and smoking status speaks directly to the causal architecture of blood pressure dysregulation (Lloyd-Jones et al., 2010).

From a technical perspective, XGBoost provides better predictive performance in identifying high-risk individuals (Chen and Guestrin, 2016), while Logistic Regression remains valuable for clinical interpretability (Hosmer et al., 2013). Both models can be deployed together:

- Logistic Regression → for clinical explainability
- XGBoost → for accuracy and large-scale screening

11.CONCLUSION

This study demonstrates that machine learning, particularly XGBoost (Chen and Guestrin, 2016), improves hypertension risk prediction compared to traditional methods. The model not only identifies at-risk individuals but also highlights actionable lifestyle factors for intervention.

Limitations include the cross-sectional design, which prevents examination of how predicted risk evolves over time. External validation across diverse demographic groups is needed. Post-hoc interpretability tools such as SHAP values (Lundberg and Lee, 2017) are recommended to build clinician trust.

Future work should include:

- Longitudinal data analysis
- External validation on diverse populations
- Use of explainability tools like SHAP (Lundberg and Lee, 2017)
- Integration with wearable health data

ACKNOWLEDGEMENT

This work would not have been completed without the support of Mr. Aaftab Alam, Assistant Professor, Department of CSE, Integral University, Lucknow. He gave us both direction and confidence throughout this journey, and we are genuinely thankful for his time and effort.

REFERENCES

1. World Health Organization.: Hypertension: Key Facts. WHO Global Report on Hypertension, Geneva (2023). Available: https://www.who.int/news-room/fact-sheets/detail/hypertension
2. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, pp. 785–794 (2016)
3. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters. 27(8), 861–874 (2006)
4. Lloyd-Jones, D.M., Hong, Y., Labarthe, D., Mozaffarian, D., Appel, L.J., Van Horn, L. et al.: Defining and setting national goals for cardiovascular health promotion and disease reduction. Circulation. 121(4), 586–613 (2010)
5. Ezzati, M., Obermeyer, Z., Tzoulaki, I., Mayosi, B.M., Elliott, P., Leon, D.A.: Contributions of risk factors and medical care to cardiovascular mortality trends. Nature Reviews Cardiology. 12(9), 508–530 (2015)
6. Tzoulaki, I., Elliott, P., Kontis, V., Ezzati, M.: Worldwide exposures to cardiovascular risk factors and associated health effects. Circulation. 133(23), 2314–2333 (2016)
7. Warren, H.R., Evangelou, E., Cabrera, C.P. et al.: Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. Nature Genetics. 49(3), 403–415 (2017)
8. Dritsas, A., Trigka, M.: Efficient data-driven machine learning models for cardiovascular diseases risk prediction. Sensors. 23(3), 1161 (2023)
9. Sudlow, C., Gallacher, J., Allen, N. et al.: UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine. 12(3), e1001779 (2015)
10. Chan, Q., Stamler, J., Griep, L.M., Daviglus, M.L., Van Horn, L., Elliott, P.: An update on nutrients and blood pressure. Journal of Atherosclerosis and Thrombosis. 23(3), 276–289 (2016)
11. Rahman, M.M., Hossain, M.I., Islam, M.S.: A machine learning approach for the prediction of hypertension using clinical and lifestyle features. IEEE Access. 10, 94596–94611 (2022)
12. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. 3rd edn. Wiley, Hoboken, NJ (2013)
13. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30 (2017)