Available Online: 21 April 2026 DOI: 10.54254/3049-5458/2026.32940

Machine learning approaches to predict hypertensive cardiovascular vessel disease risk: a comparative study using GBD and NHANES databases

Yulong He, Yan Mao*

Department of Clinical Medicine, Hunan University of Traditional Chinese, Changsha, China

*Corresponding Author. Email: 110024@hnucm.edu.cn

Abstract. Hypertension is recognized as a major global public health concern and a leading risk factor for cardiovascular disease and mortality. However, traditional risk scoring methods have limitations in identifying high-risk populations, and there is a pressing need for more precise and rapid predictive tools. With the emergence of Artificial intelligence (AI), machines are increasingly employed to execute complex tasks, yielding substantial achievements. As the most pertinent branch of AI in medicine, Machine Learning (ML) is progressively being embedded in routine clinical practice. Therefore, optimizing ML predictive strategies to enhance their predictive accuracy, clinical utility, and generalizability is of great significance. This review summarizes the predictive performance and clinical value (early detection) of ML algorithms in predicting Hypertension-related Heart Disease (HHD) risk in the Global Burden of Disease Study (GBD) and the National Health and Nutrition Examination Survey (NHANES) databases, as well as the prospects for optimizing multimodal data fusion strategies to improve predictive accuracy and expand the application of precision prevention. We aim to compare the performance of different ML algorithms in predicting HHD risk using the Global Burden of Disease Study and the National Health and Nutrition Examination Surveys databases.

Keywords: hypertension, machine learning, GBD, NHANES, CVD

1. Introduction

Hypertension remains a major global public health concern, posing a significant threat to human health. It is estimated that the global prevalence of adult hypertension reached 1.3 billion in 2019, and this number continues to rise [1]. In China alone, approximately 330 million people suffer from cardiovascular diseases, with as many as 245 million cases attributed to hypertension [2]. As a modifiable risk factor, hypertension remains a primary cause of cardiovascular diseases [3]. HHD results in nearly 10.8 million deaths annually, accounting for a significant portion of the global disease burden [4].

Artificial Intelligence (AI) and Machine Learning (ML) have emerged in recent years as sophisticated analytical tools capable of integrating a broader range of static and time-varying data points, promising the construction of complex models that incorporate both linear and nonlinear variables. Unlike traditional

statistical methods, which aim to infer a mathematical relationship between carefully selected variables and a specific target, often with some predefined assumptions, machine learning methods focus on predicting outcomes based on learning trends and correlations from data. Prior assumptions are minimal. Once a substantial amount of data with known outcomes has been inputted (i.e., the training dataset), machine learning algorithms can be trained to infer implicit and nonlinear relationships between the data and the desired outcomes, which can subsequently be used for risk prediction without explicit programming [5].

For example, machine learning studies based on the NHANES database have demonstrated excellent predictive performance. Using data from NHANES 2017-2023, the XGBoost model achieved an accuracy rate of 82.16% and a recall rate of 86.45% in predicting cardiovascular disease risk, while the random forest model exhibited the highest AUROC (0.8139) [6]. In terms of early detection of hypertension-related heart disease, Artificial Intelligence-based Electrocardiogram (AI-ECG) screening techniques are effective in detecting cardiac contractile dysfunction, with AUROC values ranging from 0.913 to 0.961, providing a cost-effective tool for screening large populations [7]. Furthermore, the complementary strengths of GBD's global trend analysis and NHANES's individual-level risk profiling suggest significant potential for future integration, which could enhance the precision of HHD burden prediction and inform targeted prevention strategies.

Given the clinical success and predictive performance of machine learning in cardiovascular disease prediction, as well as the potential to optimize multimodal data fusion strategies to enhance predictive accuracy and expand precision prevention applications, this review summarizes the clinical value and public health significance of machine learning algorithms in predicting the risk of hypertension-related heart disease within the GBD and NHANES databases. By systematically comparing the performance differences of various algorithms such as random forests, XGBoost, and deep learning, we explore strategies for constructing multimodal predictive models that integrate clinical indicators, biomarkers, and sociodemographic characteristics. This aims to provide new technological pathways and decision support for the early identification, risk stratification, and personalized intervention of hypertension-related heart disease. Therefore, optimizing machine learning predictive strategies to improve their accuracy, clinical utility, and generalizability is of great significance.

2. Methods

We conducted targeted literature searches using PubMed and MeSH terms, including those related to machine learning, hypertensive heart disease, GBD, NHANES, cardiovascular risk prediction, and related concepts. We employed a three-phase standard methodology (phase 1: title review, phase 2: abstract synthesis, phase 3: full-text synthesis) to determine the comparative performance of current machine learning algorithms in predicting the risk of hypertensive heart disease within the GBD and NHANES databases. Decisions on final inclusion were made, and reasons for exclusion were documented (e.g., lack of specific HHD data, data overlap, non-machine learning models). Standardized data extraction tables were developed and tested prior to the full extraction process. To link the research findings to relevant policy contexts, we reviewed cardiovascular prevention guidelines from selected countries (e.g., the United States, China, the United Kingdom, Germany, Japan) and various recommendations and policy documents from the World Health Organization (WHO). These materials were compared with published economic evidence to identify gaps between research outputs and policy needs [4, 8].

3. Result

3.1. Evidence of machine learning applications in the prediction and monitoring of cardiovascular diseases related to hypertension by the Global Burden of Disease (GBD) project

The Global Burden of Disease Study (GBD) provides the most comprehensive data base for understanding the flow characteristics of Hypertensive Heart Disease (HHD). In recent years, the integration of machine learning techniques with GBD data has significantly improved the accuracy of burden of disease and trend forecast. A groundbreaking study based on GBD 2021 data analyzed the burden of HHD in 204 countries and territories worldwide between 1990 and 2021. The projections to 2050 show that although the age-standardized mortality rate is decreasing, the absolute number of deaths continues to rise due to population ageing and growth, and is expected to continue to rise through 2050 [4]. In terms of predictive model construction, Artificial Neural Networks (ANNs) have been successfully applied to cardiovascular disease risk prediction in stratified populations, with AUC values of 0.77-0.86 for severe heart failure and coronary heart disease [9]. These models effectively capture the non-linear relationship between population transition, metabolic risk factors and HHD outcomes, overcoming the limitations of traditional statistical methods. In addition, Bayesian age-period-cohort models combined with integrated nested Laplace approximation have demonstrated enhanced predictive accuracy for global HHD burden estimation [10], while ensemble learning methods such as random forest and XGBoost require further validation in this context. It is worth noting that deep learning algorithms are increasingly used in cardiovascular risk prediction, automatically learning complex feature representations through multi-layer neural networks, significantly better than traditional logistic regression models [11]. The combination of longitudinal trend characteristics of GBD data with machine learning predictive models enables policymakers to anticipate future resource allocation needs and assess the effectiveness of interventions across different healthcare systems. However, data heterogeneity and the ability of models to generalize in low resource areas remain challenges that need to be addressed.

3.2. Evidence from the National Health and Nutrition Examination Survey on machine learning models for individualized identification of cardiovascular disease risk associated with hypertension

The US National Health and Nutrition Examination Survey (NHANES) is the cornerstone database for developing personalized machine learning predictive models. Using NHANES data from 1999 to 2023, researchers constructed robust cardiovascular risk assessment tools. A comprehensive analysis covering 49,490 participants was selected by random forest and XGBoost variables. Six key predictors, such as age, creatinine, platelets, glycosylated hemoglobin, uric acid, and red blood cell distribution width variability factors, were identified, and the logistic regression model was constructed with an AUC of 0.841 and good calibration [12]. For predictions of people at high risk of heart failure, a random forest model based on NHANES 2007-2018 showed superior differentiation ability in pre-diabetes or diabetes in middle-aged and older adults with an AUC of 0.978 [13]. In high blood pressure prediction, the artificial neural network used NHANES 2013-2016 data combined with demographic, lifestyle, and laboratory indicators to achieve predictive performance of AUC 0.77, superior to traditional logistic regression (AUC 0.73) [14]. The XGBoost model combined with the SHAP interpretability analysis used NHANES 2017-2020 data to identify key predictors such as age (53.1 percent contribution), poverty (4.33 percent contribution) and ethnicity (4.18 percent contribution). Nutrition contributed 37 percent overall [15]. In the study of the association of environmental chemical exposure with hypertension, the Support Vector Machine (SVM) integrated 23

environmental chemistries and 18 covariates from NHANES 2003-2016 with an AUC of 0.822, where chemicals such as lead, phthalates and polycyclic aromatic hydrocarbons contributed significantly to the predictive model [16]. In addition, the machine learning model successfully identified cardiovascular mortality risk, Triglyceride-Glucose Index (TyG), and its obesity-related derivatives (TyG-BMI), in NHANES data. TyG-WC and TyG-WHtR were shown to be independent predictors of all-cause and cardiovascular death, with each unit increase in TyG-WHtR increasing the risk of all-causing death by 41.7% and cardiovascular mortality by 48.1% [17]. The strength of NHANES data lies in its population-representative and comprehensive phenotypic characteristics, enabling models to achieve precise individual risk stratification. However, these models primarily reflect U.S. population patterns, limiting their direct outreach to global populations with different genetic backgrounds and environmental exposures.

3.3. Comparison and integration of evidence on machine learning methods for predicting hypertension-related cardiovascular disease from the project of GBD and the NHANES

Comparative analysis reveals that the GBD and NHANES-driven machine learning approaches have complementary advantages in addressing different dimensions of hypertensive cardiovascular disease prediction. GBD derived models are adept at capturing macro-population trends, predicting long-term disease burden and informing global health policy, while the NHANES foundational model provides a refined individual risk stratification suitable for clinical decision-making [4, 12]. Methodologically, both databases tend to use ensemble learning methods—random forests and XGBoost consistently outperform single classifiers in different contexts, although the GBD application emphasizes time series prediction while the NHANES study focuses on cross-sectional risk differentiation [10, 13, 18]. In terms of data features, GBD provides standardized epidemiological estimates for 204 countries and regions around the world, which is suitable for building a universal prediction framework [19]. NHANES provides detailed individual-level laboratory testing, imaging, and nutritional assessment data to support refined phenotypic risk modeling [17, 20]. The potential of this fusion approach has been validated in initial explorations: adapting GBD-trained global trend models to NHANES native data through transfer learning techniques, Alternatively, optimizing risk stratification algorithms for GBD using predictors validated by NHANES allows for the construction of a "global-local" forecasting tool that offers both global information breadth and local calibration accuracy. Key methodological considerations include standardization of variable definitions across databases, time alignment between periodic GBD updates and continuous NHANES cycles, and the development of transfer learning techniques adapted to different populations. Future research should prioritize multi-database validation studies to establish robust performance benchmarks and identify optimal algorithm configurations in diverse medical scenarios. Combining the epidemiological breadth of GBD with the clinical depth of NHANES ultimately provides a path to precision public health for hypertension and heart disease—while achieving the twin goals of population-level predictability and individual-level intervention guidance.

4. Discussion

This review systematically compares the application of machine learning in Global Disease Burden Studies (GBD) and the United States National Health and Nutrition Examination Survey (NHANES) databases for predicting the risk of Hypertension-related Heart Disease (HHD). Key findings indicate that ensemble learning methods, particularly random forests and XGBoost, demonstrate exceptional predictive performance, with studies based on NHANES data showing significant improvements over traditional statistical models [6]. In the GBD database, the Bayesian age-period-cohort model, combined with the Integrated Nested Laplace

Approximation (INLA) algorithm, effectively addresses the computational challenges of traditional Markov chain Monte Carlo sampling, enabling precise predictions of global HHD burdens [10]. In the NHANES database, machine learning models, by integrating detailed individual-level laboratory test results, imaging examinations, and nutritional assessments, achieve fine-grained phenotype-based risk stratification. The random forest model achieves an AUC of 0.978 in specific high-risk populations [13], while the XGBoost model, combined with SHAP explainability analysis, reveals the roles of age (contribution: 53.1%), poverty status (4.33%), and ethnicity (4.18%) as key predictive factors [15], with similar algorithmic approaches confirming age and waist circumference as predominant risk factors among Korean postmenopausal women [21].

For the first time, we have systematically integrated the macro-epidemiological trends of the Global Burden of Disease Study (GBD) with the micro-individual risk characteristics of the National Health and Nutrition Examination Survey (NHANES). GBD provides standardized epidemiological estimates for 204 countries and regions worldwide, with its longitudinal data structure being particularly well-suited for constructing a universal predictive framework. NHANES, on the other hand, offers detailed individual-level data, including laboratory tests, imaging examinations, and nutritional assessments, supporting the development of refined phenotype-based risk modeling [12]. The methodological complementarity between these two large datasets lies in the fact that GBD's applications emphasize time-series predictions and global trend forecasts, while NHANES research focuses on cross-sectional risk stratification and individual identification [10, 13].

At the algorithmic level, the consistent superiority of ensemble learning methods over individual classifiers has been well-documented. Random forests minimize errors by averaging the predictions of multiple decision trees and enhance the model's generalization capabilities. XGBoost handles high-dimensional structured data through regularization and gradient optimization [18]. Deep learning models, particularly Convolutional Neural Networks (CNNs), significantly outperform traditional logistic regression models in cardiovascular risk prediction (AUC 0.932 vs. 0.655) [11]. Explainable AI (XAI) techniques, specifically SHAP (SHapley Additive exPlanations), provide transparency to "black box" models, enabling clinicians to understand the mechanisms by which features contribute to predictions [22, 23].

5. Limitations and future directions

This review has the following limitations. Firstly, the majority of the included studies were retrospective in design, lacking large-scale prospective validation, which may introduce selection bias and confounding factors. Secondly, there are differences between the GBD and NHANES databases in terms of data collection standards, variable definitions, and population representation, necessitating careful interpretation of results when comparing across databases. Thirdly, the "black box" nature of machine learning models limits causal inference capabilities, and most studies have not undergone external validation, making their generalizability uncertain. Additionally, the search strategy may have missed non-English publications or gray literature, introducing a risk of publication bias. Finally, since there have been few studies that integrate the two databases directly, the construction of "global-local" predictive tools remains at a conceptual level, and their actual effectiveness needs to be validated through future research.

Future research should prioritize exploring strategies for integrating multiple databases and employing transfer learning techniques. Transfer learning techniques have been validated in epidemiological contexts for adapting region-specific training models to limited local datasets [24], providing a methodological foundation for integrating the GBD global trend model with NHANES local data. By adapting the globally trained GBD

trend model to NHANES's local data or using NHANES-validated predictive factors to optimize the GBD's risk stratification algorithm, it is possible to construct a "glocal" predictive tool that combines global information breadth with local calibration precision.

Data fusion and multimodal integration are another crucial area of focus. Future research should integrate data from genomics, proteomics, metabolomics, and imaging sciences, alongside phenotypic characteristics such as coronary artery imaging, to achieve a comprehensive understanding of individual HHD phenotypes [25].

6. Conclusion

This review summarizes the clinical success and predictive performance of machine learning algorithms in predicting the risk of hypertension-related heart disease using the GBD and NHANES databases, as well as the potential for optimizing multi-modal data fusion strategies to enhance predictive accuracy and expand the application of precision prevention. Both databases have their strengths: GBD provides global macro-trends, while NHANES offers local micro-precision. Ensemble learning methods such as random forests and XGBoost consistently outperform single classifiers in various contexts, and the integration of explainable artificial intelligence technologies enhances the clinical credibility of the models. Despite challenges such as data silos, model generalization, algorithmic bias, and regulatory frameworks hindering clinical translation, machine learning holds promise for providing a pathway to precision public health for hypertension-related heart disease by enabling cross-database integration, multi-modal data fusion, and prospective clinical validation through transfer learning. Improving machine learning prediction strategies to enhance their predictive accuracy, clinical utility, and generalizability thus holds significant importance, offering new technological pathways and decision support for early identification, risk stratification, and personalized interventions for hypertension-related heart disease.

References

[1] NCD Risk Factor Collaboration (NCD-RisC). (2021). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. *The Lancet*, *398*(10304), 957–980. https: //doi.org/10.1016/S0140-6736(21)01330-1

[2] In China The Writing Committee of the Report on Cardiovascular Health and Diseases, & Hu, S. S. (2023). Report on cardiovascular health and diseases in China 2021: an updated summary. *Journal of Geriatric Cardiology*, *20*(6), 399–430. https: //doi.org/10.26599/1671-5411.2023.06.001

[3] Si, F., Liu, Q., & Yu, J. (2025). A prediction study on the occurrence risk of heart disease in older hypertensive patients based on machine learning. *BMC Geriatrics, 25*(1), 27. https: //doi.org/10.1186/s12877-025-05679-1

[4] Lee, C., Hwang, S. H., Cho, J., Lee, S., Hong, S., Kim, T. H., Lee, H., Lee, J., Pizzol, D., Smith, L., Hwang, J., Yang, S. Y., & Yon, D. K. (2025). Global, regional, and national burden of hypertensive heart disease in 1990-2021, with forecasts to 2050: a Global Burden of Disease Study 2021. *Clinical Hypertension*, *31*, e36. https: //doi.org/10.5646/ch.2025.31.e36

[5] Gautam, N., Mueller, J., Alqaisi, O., Gandhi, T., Malkawi, A., Tarun, T., Alturkmani, H. J., Zulqarnain, M. A., Pontone, G., & Al'Aref, S. J. (2023). Machine Learning in Cardiovascular Risk Prediction and Precision Preventive Approaches. *Current Atherosclerosis Reports*, *25*(12), 1069–1081. https: //doi.org/10.1007/s11883-023-01174-3

[6] Ahiduzzaman, M., & Hasan, M. N. (2025). Interpretable machine learning for cardiovascular risk prediction: Insights from NHANES dietary and health data. *PLOS ONE*, *20*(11), e0335915. https: //doi.org/10.1371/journal.pone.0335915

[7] Cho, J., Lee, B., Kwon, J. M., Lee, Y., Park, H., Oh, B. H., Jeon, K. H., Park, J., & Kim, K. H. (2021). Artificial Intelligence Algorithm for Screening Heart Failure with Reduced Ejection Fraction Using Electrocardiography. *ASAIO Journal*, *67*(3), 314–321. https: //doi.org/10.1097/MAT.0000000000001218

[8] Smith, D. S., Postma, M., Fisman, D., & Mould-Quevedo, J. (2025). Cost-effectiveness models assessing COVID-19 booster vaccines across eight countries: A review of methods and data inputs. *Vaccine*, *51*, 126879. https: //doi.org/10.1016/j.vaccine.2025.126879

[9] Taha, K., Ross, H. J., Peikari, M., Mueller, B., Fan, C. S., Crowdy, E., Moayedi, Y., Billia, F., & Manlhiot, C. (2025). Predicting the future risk and outcomes of severe heart failure and coronary artery disease with machine learning in the UK Biobank Cohort. *PLOS ONE*, *20*(9), e0329461. https: //doi.org/10.1371/journal.pone.0329461

[10] Liu, F., Pan, H. W., Li, Y. Y., Zhao, X. J., Hong, X. Q., Liu, Z. Y., & You, Y. Y. (2025). Trends analysis of the global burden of hypertensive heart disease from 1990 to 2021: a population-based study. *BMC Public Health*, *25*(1), 2233. https: //doi.org/10.1186/s12889-025-23389-6

[11] Lee, S. J., Lee, S. H., Choi, H. I., Lee, J. Y., Jeong, Y. W., Kang, D. R., & Sung, K. C. (2022). Deep Learning Improves Prediction of Cardiovascular Disease-Related Mortality and Admission in Patients with Hypertension: Analysis of the Korean National Health Information Database. *Journal of Clinical Medicine*, *11*(22), 6677. https: //doi.org/10.3390/jcm11226677

[12] Lu, J., Hu, H., Xiu, J., Yang, Y., Zhu, Q., Dai, H., Liu, X., & Wang, J. (2024). Machine learning-driven risk assessment of coronary heart disease: Analysis of NHANES data from 1999 to 2018. *Journal of Central South University (Medical Sciences)*, *49*(8), 1175–1186. https: //doi.org/10.11817/j.issn.1672-7347.2024.240394

[13] Wang, Y., Hou, R., Ni, B., Jiang, Y., & Zhang, Y. (2023). Development and validation of a prediction model based on machine learning algorithms for predicting the risk of heart failure in middle-aged and older US people with prediabetes or diabetes. *Clinical Cardiology*, *46*(10), 1234–1243. https: //doi.org/10.1002/clc.24104

[14] AlKaabi, L. A., Ahmed, L. S., Al Attiyah, M. F., & Abdel-Rahman, M. E. (2020). Predicting hypertension using machine learning: Findings from Qatar Biobank Study. *PLOS ONE*, *15*(10), e0240370. https: //doi.org/10.1371/journal.pone.0240370

[15] Huang, A. A., & Huang, S. Y. (2023). Shapely additive values can effectively visualize pertinent covariates in machine learning when predicting hypertension. *The Journal of Clinical Hypertension*, *25*(12), 1135–1144. https: //doi.org/10.1111/jch.14745

[16] Guo, K., Ni, W., Du, L., Zhou, Y., Cheng, L., & Zhou, H. (2024). Environmental chemical exposures and a machine learning-based model for predicting hypertension in NHANES 2003-2016. *BMC Cardiovascular Disorders*, *24*(1), 544. https: //doi.org/10.1186/s12872-024-04216-z

[17] Li, C., Zhang, Z., Luo, X., Xiao, Y., Tu, T., Liu, C., Liu, Q., Wang, C., Dai, Y., Zhang, Z., Zheng, C., & Lin, J. (2025). The triglyceride-glucose index and its obesity-related derivatives as predictors of all-cause and cardiovascular mortality in hypertensive patients: insights from NHANES data with machine learning analysis. *Cardiovascular Diabetology*, *24*(1), 47. https: //doi.org/10.1186/s12933-025-02591-1

[18] Hassan, F. H., Wang, S., & Miron, A. (2026). Machine Learning for Predicting Coronary Heart Disease Risk in Patients with Hypertension: An Ensemble Modeling Approach. *Healthcare Informatics Research*, *32*(1), 28–37. https: //doi.org/10.4258/hir.2026.32.1.28

[19] GBD 2021 Forecasting Collaborators. (2024). Burden of disease scenarios for 204 countries and territories, 2022-2050: a forecasting analysis for the Global Burden of Disease Study 2021. *The Lancet*, *403*(10440), 2204–2256. https: //doi.org/10.1016/S0140-6736(24)00685-8

[20] Dillon, C. F., & Weisman, M. H. (2018). US National Health and Nutrition Examination Survey Arthritis Initiatives, Methodologies and Data. *Rheumatic Diseases Clinics of North America*, *44*(2), 215–265. https: //doi.org/10.1016/j.rdc.2018.01.010

[21] Kim, H., Khomidov, M., & Lee, J. H. (2025). XGBoost and SHAP-Based Analysis of Risk Factors for Hypertension Classification in Korean Postmenopausal Women. *Bioengineering*, *12*(6), 659. https: //doi.org/10.3390/bioengineering12060659

[22] Cheng, Y., Guo, Y., Zhao, Y., Wang, C., Zhao, X., Yu, Q., Huang, J., Zhang, Y., Zhang, J., Liu, X., Cai, P., Zhang, C., Wu, B., & Guo, Y. (2026). Development and validation of a machine learning model to predict 30-day mortality in ischemic stroke patients with consciousness impairment: Insights from MIMIC-IV database and multicenter ICU data in China. *International Journal of Medical Informatics*, *207*, 106203. https: //doi.org/10.1016/j.ijmedinf.2025.106203

[23] Hossain, M. K., Ashraf, A., Islam, M. M., Sourav, S. H., & Shimul, M. M. H. (2025). Optimizing Alzheimer's disease prediction through ensemble learning and feature interpretability with SHAP-based feature analysis. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *17*(3), e70162. https: //doi.org/10.1002/dad2.70162

[24] Lueangwitchajaroen, P., Anupong, S., Winalai, C., & Chadsuthi, S. (2026). Leveraging universal and transfer learning models for influenza prediction in Thailand. *Scientific Reports*, *16*(1), 6668. https: //doi.org/10.1038/s41598-026-37855-7

[25] Banerjee, T., & Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. *Turkish Journal of Biology*, *49*(5), 600–634. https: //doi.org/10.55730/1300-0152.2766