**Protective Predictors of Cardiovascular Disease: an Explainable AI Approach**

**ABSTRACT**

**Purpose**  
To develop interpretable machine learning (ML) models using nationally representative survey data to identify protective factors against cardiovascular disease (CVD), addressing gaps in traditional clinical risk scores across diverse populations.

**Methods**  
We analyzed 116,608 adult records after rigorous data cleaning. Three ML models - XGBoost, convolutional neural network (CNN), and random forest - were trained on 11 demographic and behavioral features, including age, sex, race/ethnicity, income, smoking, alcohol use, depression, diabetes, insurance status, and fruit and vegetable intake. Performance was assessed using precision, recall, F1-score, AUROC, and AU-PRC. SHapley Additive exPlanations (SHAP) were applied to interpret model predictions.

**Results**  
XGBoost outperformed other models, achieving a precision of 0.90, recall of 0.82, F1-score of 0.86, AUROC of 0.76, and AU-PRC of 0.95. SHAP analysis identified younger age, higher income, insurance coverage, and absence of diabetes or depression as strong protective predictors.

**Conclusion**  
This explainable XGBoost model predicts cardiovascular resilience by emphasizing diabetes absence, mental health stability, socioeconomic advantage, and younger age. It shifts focus from disease detection to proactive, equitable prevention and efficient resource allocation in CVD care.

**Keywords**

CVD, machine learning, SHAP, prevention.

**INTRODUCTION**

Cardiovascular disease (CVD) is the foremost cause of death globally, causing about 20 million deaths and nearly 400 million DALYs each year (1, 2). Primary prevention lags behind therapeutic advances, particularly in socially disadvantaged groups where modifiable risks cluste (3).

Traditional CVD risk models like logistic regression (LR) and Cox regression rely on assumptions (e.g., linearity, proportionality) that limit their ability to capture complex interactions among social, behavioral, and clinical factors. These models often exclude key contextual variables such as income, mental health, and access to care, which may result in underestimating risk in disadvantaged groups. ML methods, by contrast, have shown superior performance: in a Korean cohort, ML outperformed the Pooled Cohort Equations (PCEs) (C-statistic 0.751 vs. 0.738) (4) .In a large clinical cohort, an XGBoost ensemble model surpassed LR in CVD risk prediction (AUROC 0.76 vs. 0.74) (5). This study benchmarks interpretable ML to advance equitable CVD prevention.

Although tools like the Framingham Risk Score and PCEs have advanced CVD risk assessment (6, 7), they often overlook broader social and behavioral factors such as diet, mental health, and socioeconomic status, which play critical roles in cardiovascular outcomes (3, 8). ML models are better equipped to capture these nonlinear relationships (9, 10). Yet, despite their improved performance, many ML models have seen limited clinical adoption due to their opaque, “black-box” nature (11). This highlights the need for interpretable ML approaches that balance predictive accuracy with transparency in clinical decision-making.

Explainable AI (XAI) originated from early symbolic systems (1970s - 1990s) like MYCIN (12), GUIDON (13), SOPHIE (13), and PROTOS (14), which could both reason and explain their logic. Modern methods, such as SHAP (15), enhance model interpretability by quantifying how features contribute to individual predictions. This enables transparency in high stakes settings like healthcare.

To address these gaps, this study utilizes the 2021 Behavioral Risk Factor Surveillance System (BRFSS), a nationally representative U.S. dataset, to develop and evaluate explainable ML models based on simple demographic and lifestyle variables. The aim is to identify protective cardiovascular profiles and highlight factors linked to lower CVD risk. By focusing on individuals without diagnosed CVD, this study shifts from a disease-centered to a resilience based prevention model, offering transparent tools adaptable to diverse and resource limited settings.

**METHODS**

**Study population and data cleaning**

This study analyzed data from the 2021 BRFSS, a nationwide telephone-based survey that collects information on health-related behaviors, chronic conditions, and access to healthcare among non-institutionalized U.S. adults aged 18 years and older. Participants were selected using a multistage random-digit-dialing method covering both landline and cellular telephones. One adult was randomly selected per household for inclusion. From an initial sample of 438,694 respondents, we excluded records with missing responses for any of the key predictor or outcome variables, as well as those with implausible values (e.g., extreme age or income), resulting in a final analytic dataset of 116,608 fully de-identified individuals. The workflow as shown in **Figure 1.**

**Variable selection and outcome definition**

A total of 11 variables were selected for model development based on theoretical and empirical relevance to CVD risk within a public health context. These predictors were grouped into three domains: (1) demographic and socioeconomic characteristics, including age group, sex, race/ethnicity, income category, and insurance status; (2) behavioral and lifestyle factors, including smoking status, alcohol use in the past 30 days, and fruit/vegetable intake frequency; and (3) clinical and mental health indicators, including diagnosed diabetes (type 1, type 2, or gestational) and history of depressive disorder. Although exercise frequency and urban/rural status were initially assessed, they were excluded from final model training due to low predictive contribution. However, these variables are retained in **Table 1** to fully characterize the sample.

The primary outcome was a binary indicator of CVD, defined by a self-reported diagnosis of coronary artery disease, myocardial infarction, or stroke provided by a physician. Participants were coded as CVD-positive if they reported any of these conditions.

**Data preprocessing**

We randomly split the dataset into training (80%) and test (20%) sets using stratified sampling to preserve class proportions. The training set contained 93,286 records and the test set contained 23,322 records. Given the skewed class distribution (CVD:non-CVD ≈ 1:5.76), we applied a combined sampling strategy to mitigate class imbalance. Specifically, CVD-positive cases in the training set were upsampled threefold using random sampling with replacement, while non-CVD cases were downsampled to a 5:5.76 ratio. This produced a final training sample of 110,394 individuals with a more moderate imbalance (69,000 non-CVD vs. 41,394 CVD).

For model validation, five-fold stratified cross-validation was employed during training, and ten-fold cross-validation was implemented during hyperparameter tuning with Optuna (see **Supplementary Table 2**).

**Model architecture and development**

We evaluated six baseline algorithms: SVM, MLP, LR, RF, XGBoost (XGB), and CNN, which included two convolutional layers (64 and 128 filters) followed by three fully connected layers. Models were trained using 10-fold cross-validation with class balancing (random upsampling of CVD cases and downsampling of non-CVD cases). Three models were developed: XGB, CNN, and RF. The CNN used ReLU activation in hidden layers, a sigmoid activation for binary classification, and a dropout rate of 0.494 to prevent overfitting. RF served as a benchmark for interpretability.

**Training procedure**

The XGBoost model was selected for hyperparameter tuning and final evaluation. We used the Adam optimizer with an initial learning rate of 0.00123 and a focal loss function to manage class imbalance. Adaptive alpha and gamma parameters were tuned within a range of 2.0 to 10.0. Early stopping was applied with a patience of 50 epochs. Learning rate scheduling was performed using ReduceLROnPlateau with a reduction factor of 0.1, patience of 25 epochs, and a minimum learning rate threshold of 0.00001.

**Hyperparameter optimization**

Hyperparameter tuning was conducted using the Optuna framework with 20 trials per model. The objective was to maximize the specificity metric using ten-fold stratified cross-validation within each trial. For XGBoost, optimized parameters included estimators, tree depth, learning rate, subsample, colsample\_bytree, and gamma.

**Performance evaluation**

Once the model was fully trained on the training data, its performance was evaluated on the test set. Predicted outcomes were compared against the observed CVD status to calculate various performance metrics as follows:

* Precision (Positive Predictive Value):  
  This metric quantifies the proportion of individuals predicted to have CVD who truly have the condition.
* Recall (Sensitivity):  
  This indicates the proportion of actual CVD cases that the model correctly identified as positive.
* F1-Score:  
  Representing the harmonic mean of precision and recall, the F1-score provides a balance between these two metrics.
* Accuracy:  
  This is the overall proportion of correctly classified cases (both CVD and non-CVD).
* Area Under the ROC Curve (AUROC):  
  The AUROC measures the model’s ability to discriminate between CVD and non-CVD cases across various decision thresholds.
* Positive Predictive Value (PPV)

PPV measures the proportion of individuals predicted to have CVD who actually have the condition, indicating the reliability of positive predictions.

* Negative Predictive Value (NPV)

NPV measures the proportion of individuals predicted not to have CVD who are truly free of the condition, reflecting the reliability of negative predictions.

In addition, a confusion matrix was used to visually summarize the number of true positives, true negatives, false positives, and false negatives observed in the test set.

**Software and statistical analysis**

All analyses were conducted using Python 3.12.7. Data processing, model training, and evaluation employed standard machine learning libraries. Visualizations, including confusion matrices and SHAP plots, were created using established graphical tools. Statistical significance was assessed using two-sided tests with a p-value threshold of 0.05.

**RESULTS**

**Study population and data distribution**

CVD prevalence rose with age (1.3% at 18–24 to 24.6% at ≥65), higher in males (17.2%) than females (12.2%). Black non-Hispanics had the highest (16.7%), Hispanics lowest (9.6%) prevalence. Lower income correlated with higher CVD (23.6% vs. 7.5%).

Regarding behavioral and clinical variables, adults with diabetes (30.9%) and depressive disorder (16.9%) increased CVD risk. Daily smokers had higher rates (14.5%) than occasional smokers (12.2%). Interestingly, alcohol consumption in the past 30 days was associated with lower CVD prevalence (11.1%) than abstention (19.4%), possibly reflecting reverse causation or confounding by health status. Daily vegetable intake is linked to lower CVD (14.3%). Health insurance was associated with lower CVD prevalence (7.4% uninsured vs. 15.2% insured).

Data split into 80% training (93,286) and 20% testing (23,322) with imbalanced CVD distribution. As shown, the distribution between non-CVD and CVD cases was substantially imbalanced in both subsets (see in **Table 1**).

**Baseline algorithm comparison**

We applied 10-fold CV to each algorithm, using random under- and over-sampling to balance the class ratio at 1.67 non-CVD to 1 CVD (see in **Table 2**). We evaluated performance based on confusion matrices, specificity, precision, recall, macro F1, accuracy, and AUROC. XGBoost emerged as the most suitable model for fine-tuning due to its balanced performance across multiple metrics. Although RF and LR had slightly higher precision, XGBoost offered better recall (74.54%) and F1 score (77.28%), with lower standard deviation, ensuring stable performance. Despite CNN’s higher AUROC, XGBoost’s consistent performance led to its selection for further tuning.

**Optimize model by Optuna**

We set up 20 trials with 10-fold CV (on preprocessed training data) for each trial in Optuna to automate XGBoost hyper-parameter finetuning **(Supplementary Table 2)**. Each set of parameters in each trial was validated on the test set to compare the specificity metrics.

The explored parameters included the number of estimators (50 to 500), maximum tree depth (3 to 10), learning rate (0.01 to 0.30), subsample ratio (0.60 to 1.00), colsample\_bytree (0.50 to 1.00), and gamma (0.00 to 1.00). The model was configured with a binary logistic objective, and a fixed random seed (random\_state = 42) was used to ensure reproducibility. The optimal parameter set obtained in trial 14th – which included n\_estimators of 53, max\_depth of 8, learning\_rate of 0.08, subsample of 0.90, colsample\_bytree of 0.61, gamma of 0.83 – achieved the highest 0.82 on specificity.

We also compared the Optuna-optimized XGBoost with other algorithms on the original training data using 10-fold CV, to make a fair comparison with other baseline models. The result in Table 2 illustrated the improved F1 score and AUROC over other algorithms, which highlighted the capacity of the optimized XGBoost in classifying non-CVD and CVD cases.

**Model performance metrics**

The XGBoost model achieved a PPV of 32.36% and a NPV of 90.49%, based on confusion matrix counts (TP = 1,746; FP = 3,649; TN = 16,223; FN = 1,704) **(Supplementary Figure 1).** These results indicate strong performance in correctly identifying individuals without CVD, as reflected by the high NPV. However, the relatively low PPV suggests that many individuals predicted to have CVD were false positives, implying a tendency to overestimate risk in some non-CVD cases. This pattern is typical of imbalanced datasets, where the number of non-CVD cases far exceeds positive cases. While the model excels at ruling out CVD, further refinement and supplementary clinical screening could enhance precision and reduce unnecessary follow-up for false positives. The model's discriminative ability was further validated using 10-fold cross-validation, yielding an average AUROC of 0.7616 and a high AUPRC of 0.9469, demonstrating its robustness and reliability in predicting Non-CVD cases **(Supplementary Figure 2).**

**Feature importance and explainability**

To improve interpretability, we applied XAI techniques using SHAP values and traditional feature importance from the random forest model. These approaches allowed us to quantify and rank the influence of each predictor on the XGBoost model’s output.

1. ***SHAP summary bar***

**Supplementary Figure 3** displays a SHAP summary bar plot, ranking features by their mean absolute SHAP values. This visualization highlights each variable’s overall impact on model predictions, offering a clear and interpretable overview of feature importance.

1. ***SHAP summary dot***

**Figure 2** shows a SHAP dot plot, which combines feature importance with directionality. Each dot represents a participant, with color indicating the feature’s value and horizontal position showing whether it increases or decreases predicted CVD risk. This plot enables insight into how individual-level feature values influence predictions.

Across the population, older age, lower income, diabetes, and depressive disorder were associated with higher predicted risk, while higher fruit and vegetable intake shifted predictions toward lower risk. These results reinforce the model’s alignment with established clinical and social determinants of cardiovascular disease.

**DISCUSSION**

XGBoost outperformed CNN and RF models in identifying individuals without CVD, achieving precision of 0.90, recall of 0.82, and an F1-score of 0.86. Its high negative predictive value and a precision-recall AUC of 0.95 underscore its strength in ruling out disease, even in imbalanced datasets. By leveraging demographic, behavioral, and nutritional data from a nationally representative sample (n = 116,608), the model demonstrates strong generalizability and supports its potential utility in preventive public health screening.

**Key predictors and cardiovascular protection**

**Age** emerged as the strongest predictor of cardiovascular resilience. Younger individuals who have had less cumulative exposure to risk factors dominated the no-CVD class. This pattern reflects longstanding evidence that age is a major CVD risk factor, but its impact is attenuated in the absence of hypertension, hyperlipidemia, diabetes, and smoking (16). Thus, our model’s low-risk assignment for younger adults, and for older adults without comorbidities, supports the concept that absence of risk factors confers substantial protection (16, 17).

**Diabetes** status was equally impactful. Individuals without diabetes were far more likely to be classified as no-CVD. This is consistent with prior research showing a 2–4-fold increase in CVD risk among those with diabete**s** (18). Our findings reinforce that glucose control whether through lifestyle or pharmacologic means is essential for cardiovascular prevention (18).

**Depressive disorder** also emerged as a relevant determinant. The absence of a depressive disorder was strongly associated with lower predicted CVD risk. Depression has been shown to significantly increase the incidence of heart disease and is now recognized as a marker of future cardiovascular events (19). Mental health influences CVD both directly and indirectly through behavioral factors such as physical inactivity, poor diet, and nonadherence to medical therapy (19).

**Socieconomic status** as captured by household income, showed a graded inverse relationship with CVD risk. Higher income individuals were substantially overrepresented in the no-CVD group. This echoes evidence linking lower income to increased prevalence of hypertension, diabetes, and cardiovascular events (20). Socioeconomic advantage likely improves access to healthcare, nutrition, and healthier living environments factors that collectively reduce disease burden (20-22).

**Insurance coverage** similarly predicted cardiovascular protection. Those with health insurance were more often classified as CVD-free, underscoring the role of access in early detection and management of risk factors. Studies show that uninsured adults have worse control of blood pressure and cholesterol (23) while those with insurance benefit from routine checkups and preventive care (21, 24).

Dietary factors showed low importance in our SHAP analysis and were not emphasized, likely because other variables dominated the predictive signal. While healthy eating remains important for cardiovascular health, it was not a key predictor in this model.

**Public health model vs AI models – Trade-offs**

Traditional models such as logistic regression and Cox regression have long supported CVD research but rely on assumptions of linearity, proportionality, and additivity that limit their ability to capture complex, non-linear relationships among behavioral, social, and clinical factors (29, 30). Often developed in homogenous cohorts, they may overlook key contextual variables like income or mental health, underestimating risk in marginalized populations (31). In contrast, ML especially XGBoost handles high-dimensional, non-linear data effectively (32), while SHAP enables transparent, case-level interpretation (33).

**Benchmark comparison of traditional vs. ML models**

To contextualize model performance, we reviewed recent studies comparing traditional statistical methods with ML approaches in cardiovascular risk prediction. Tree-based models like XGBoost consistently outperformed logistic regression and Cox models in AUROC, F1 score, and sensitivity, especially in heterogeneous populations. **Supplementary Table 3** highlights these advantages. Notably, our use of SHAP values addresses the common critique of ML opacity, enhancing transparency. These findings underscore the value of interpretable ML for equitable, population-level risk stratification.

**Prevention and resilience perspective**

These findings highlight a profile of cardiovascular resilience, with the model excelling in identifying individuals without major risk factors, particularly those with socioeconomic advantages. Managing modifiable risks, such as preventing diabetes and addressing depressive disorders, along with enhancing social support, helps maintain low-risk status. The model’s high negative predictive value ensures accurate identification of low-risk individuals, guiding early interventions and encouraging healthy habits. This interpretable model shifts focus from diagnosis to health promotion, reinforcing the importance of minimizing risks and enhancing social support for better cardiovascular outcomes, even in older adults (25). It fosters a preventive approach by emphasizing protective factors and minimizing risks (19, 25).

**Strengths**

Studying has several strengths. It uses a nationally representative BRFSS 2021 sample of over 116 000 adults, ensuring strong statistical power and external validity. The model incorporates SHAP-based explainability, providing transparency for clinicians and patients. Its preventive focus is clear, with high NPV and class-0 recall above 0.82, making it reliable for ruling out disease. Additionally, the model was validated through ten-fold cross-validation, focal loss optimization, and hyperparameter tuning, ensuring stable performance despite class imbalance.

**Limitations**

This study has several limitations. Its cross-sectional design limits causal interpretation and may reflect reverse causation. Self-reported lifestyle and health data introduce recall and social desirability biases. Nutrition was measured in a binary way, limiting detail. While class-imbalance methods improved sensitivity, they reduced the model’s precision, with a 32% positive predictive value. Lastly, the model’s U.S.-based data limits its generalizability without further validation.

**Public health implementation and future directions**

From an implementation perspective, the model is well-suited for rapid screening and triage in primary care or community settings. A brief survey on mobile devices could efficiently identify low-risk adults, directing more intensive tests to higher-risk individuals. Incorporating SHAP dashboards into mobile health apps would offer personalized insights. Future steps include testing in longitudinal cohorts, refining the model with additional features, and conducting cost-effectiveness analyses to support broader adoption and policy integration.

**Conclusion**

This explainable XGBoost model identifies adults likely to be absent of cardiovascular disease by emphasizing factors like no diabetes, stable mental health, socioeconomic advantage, and younger age. It promotes proactive health maintenance and supports equitable, large-scale CVD prevention.

**Ethics Statement**

This study used publicly available, de-identified data from the 2021 Behavioral Risk Factor Surveillance System (BRFSS). Ethical review and informed consent were not required.

**Funding Statement**

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Conflict of Interest**

The authors declare no conflicts of interest.

**Data Availability**

The BRFSS 2021 dataset is publicly available from the U.S. Centers for Disease Control and Prevention.

**REFERENCES**

1. Lindstrom M, DeCleene N, Dorsey H, Fuster V, Johnson CO, LeGrand KE, et al. Global Burden of Cardiovascular Diseases and Risks Collaboration, 1990-2021. J Am Coll Cardiol. 2022;80(25):2372-425.

2. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. Journal of the American college of cardiology. 2020;76(25):2982-3021.

3. Stringhini S, Carmeli C, Jokela M, Avendaño M, Muennig P, Guida F, et al. Socioeconomic status and the 25× 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1· 7 million men and women. The Lancet. 2017;389(10075):1229-37.

4. Cho S-Y, Kim S-H, Kang S-H, Lee KJ, Choi D, Kang S, et al. Pre-existing and machine learning-based models for cardiovascular risk prediction. Scientific reports. 2021;11(1):8886.

5. Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. European Heart Journal-Digital Health. 2025;6(1):7-22.

6. D’Agostino Sr RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53.

7. Goff Jr DC, Lloyd-Jones DM, Bennett G, Coady S, D’agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129(25\_suppl\_2):S49-S73.

8. Aune D, Giovannucci E, Boffetta P, Fadnes LT, Keum N, Norat T, et al. Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality—a systematic review and dose-response meta-analysis of prospective studies. International journal of epidemiology. 2017;46(3):1029-56.

9. Martin-Morales A, Yamamoto M, Inoue M, Vu T, Dawadi R, Araki M. Predicting cardiovascular disease mortality: Leveraging machine learning for comprehensive assessment of health and nutrition variables. Nutrients. 2023;15(18):3937.

10. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. Journal of the American College of Cardiology. 2018;71(23):2668-79.

11. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347-58.

12. Shortliffe E. Computer-based medical consultations: MYCIN: Elsevier; 2012.

13. Clancey WJ. Knowledge-based tutoring: The GUIDON program: MIT press; 1987.

14. Bareiss ER, Porter BW, Wier CC. Protos: An exemplar-based learning apprentice. Machine learning: Elsevier; 1990. p. 112-27.

15. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. Redox Biology. 2025;79:103470.

16. Lloyd-Jones DM, Leip EP, Larson MG, d’Agostino RB, Beiser A, Wilson PW, et al. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation. 2006;113(6):791-8.

17. Terry DF, Pencina MJ, Vasan RS, Murabito JM, Wolf PA, Hayes MK, et al. Cardiovascular risk factors predictive for survival and morbidity‐free survival in the oldest‐old Framingham Heart Study participants. Journal of the American Geriatrics Society. 2005;53(11):1944-50.

18. Dal Canto E, Ceriello A, Rydén L, Ferrini M, Hansen TB, Schnell O, et al. Diabetes as a cardiovascular risk factor: An overview of global trends of macro and micro vascular complications. European journal of preventive cardiology. 2019;26(2\_suppl):25-32.

19. Barger SD, Struve GC. Association of depression with 10-year and lifetime cardiovascular disease risk among US adults, national health and nutrition examination survey, 2005–2018. Preventing Chronic Disease. 2022;19:E28.

20. Minhas AMK, Jain V, Li M, Ariss RW, Fudim M, Michos ED, et al. Family income and cardiovascular disease risk in American adults. Scientific reports. 2023;13(1):279.

21. Wang T, Li Y, Zheng X. Association of socioeconomic status with cardiovascular disease and cardiovascular risk factors: a systematic review and meta-analysis. Journal of Public Health. 2024;32(3):385-99.

22. Abdalla SM, Yu S, Galea S. Trends in cardiovascular disease prevalence by income level in the United States. JAMA network open. 2020;3(9):e2018150-e.

23. Brooks EL, Preis SR, Hwang S-J, Murabito JM, Benjamin EJ, Kelly-Hayes M, et al. Health insurance and cardiovascular disease risk factors. The American journal of medicine. 2010;123(8):741-7.

24. Hendriks ME, Wit FW, Akande TM, Kramer B, Osagbemi GK, Tanović Z, et al. Effect of health insurance and facility quality improvement on blood pressure in adults with hypertension in Nigeria: a population-based study. JAMA internal medicine. 2014;174(4):555-63.

25. Dhingra R, Vasan RS. Age as a risk factor. Medical Clinics. 2012;96(1):87-91.

26. Teshale AB, Htun HL, Owen A, Gasevic D, Phyo AZZ, Fancourt D, et al. The role of social determinants of health in cardiovascular diseases: an umbrella review. Journal of the American Heart Association. 2023;12(13):e029765.

27. Brandt EJ, Tobb K, Cambron JC, Ferdinand K, Douglass P, Nguyen PK, et al. Assessing and addressing social determinants of cardiovascular health: JACC state-of-the-art review. Journal of the American College of Cardiology. 2023;81(14):1368-85.

28. Corwin T, Ozieh MN, Garacci E, Walker RJ, Egede LE. Association of social risk domains with poor cardiovascular risk factor control in US adults with diabetes, from 2006 to 2016. JAMA Network Open. 2022;5(9):e2230853-e.

29. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, et al. Deep learning for cardiovascular medicine: a practical primer. European heart journal. 2019;40(25):2058-73.

30. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology. 2019;110:12-22.

31. Havranek EP, Mujahid MS, Barr DA, Blair IV, Cohen MS, Cruz-Flores S, et al. Social determinants of risk and outcomes for cardiovascular disease: a scientific statement from the American Heart Association. Circulation. 2015;132(9):873-98.

32. Islam MM, Alam MJ, Maniruzzaman M, Ahmed NF, Ali MS, Rahman MJ, et al. Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia. PLoS One. 2023;18(8):e0289613.

33. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

**A diagram of a data analysis

AI-generated content may be incorrect.FIGURE 1 -** **Workflow of Study**

A graph of different colored lines

AI-generated content may be incorrect.

**FIGURE 2 - SHAP Summary Plot For Feature Effects**

**TABLE 1 -** **Baseline Characteristics of Study Participants**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Total** | **Non-CVD** | **CVD** | **p-value** |
| **DEMOGRAPHIC AND SOCIOECENOMIC FACTORS** | | | | |
| **Age group (n,%)** | | | | |
| 18-24 | 2175 (1.9%) | 2147 (98.7%) | 28 (1.3%) | <0.0001 |
| 25-34 | 10313 (8.8%) | 10081 (97.8%) | 232 (2.2%) |  |
| 35-44 | 17056 (14.6%) | 16372 (96.0%) | 684 (4.0%) |  |
| 45-54 | 17855 (15.3%) | 16275 (91.2%) | 1580 (8.8%) |  |
| 55-64 | 24860 (21.3%) | 21029 (84.6%) | 3831 (15.4%) |  |
| 65+ | 44349 (38.0%) | 33456 (75.4%) | 10893 (24.6%) |  |
| **Gender (n,%)** | | | | |
| Male | 60029 (51.5%) | 49705 (82.8%) | 10324 (17.2%) | <0.0001 |
| Female | 56579 (48.5%) | 49655 (87.8%) | 6924 (12.2%) |  |
| **Race/ethnicity (n,%)** | | | | |
| White, non-Hispanic | 94958 (81.4%) | 80659 (84.9%) | 14299 (15.1%) | <0.0001 |
| Black, non-Hispanic | 6710 (5.8%) | 5590 (83.3%) | 1120 (16.7%) |  |
| Other race, non-Hispanic | 5440 (4.7%) | 4660 (85.7%) | 780 (14.3%) |  |
| Multiracial, non-Hispanic | 2846 (2.4%) | 2434 (85.5%) | 412 (14.5%) |  |
| Hispanic | 6654 (5.7%) | 6017 (90.4%) | 637 (9.6%) |  |
| **Income level (n,%)** | | | | |
| < $15,000 | 8642 (7.4%) | 6601 (76.4%) | 2041 (23.6%) | <0.0001 |
| 15000-25000 | 13863 (11.9%) | 10827 (78.1%) | 3036 (21.9%) |  |
| 25000-35000 | 16533 (14.2%) | 13530 (81.8%) | 3003 (18.2%) |  |
| 35000-50000 | 17472 (15.0%) | 14777 (84.6%) | 2695 (15.4%) |  |
| 50000-100000 | 36236 (31.1%) | 31734 (87.6%) | 4502 (12.4%) |  |
| 100000-200000 | 19382 (16.6%) | 17748 (91.6%) | 1634 (8.4%) |  |
| >= 200000 | 4480 (3.8%) | 4143 (92.5%) | 337 (7.5%) |  |
| **Health insurance (n,%)** | | | | |
| Yes | 109856 (94.2%) | 93105 (84.8%) | 16751 (15.2%) | <0.0001 |
| No | 6752 (5.8%) | 6255 (92.6%) | 497 (7.4%) |  |
| **Urban/Rural status (n, %)** | | | | |
| Urban | 97928 (84.0%) | 83883 (85.66%) | 14045 (14.34%) | <0.0001 |
| Rural | 18680 (16.0%) | 15477 (82.85%) | 3203 (17.15%) |  |
| **BEHAVIORAL AND LIFESTYLE FACTORS** | | | | |
| **Smoking status (n, %)** | | | | |
| Everyday | 27057 (23.2%) | 23137 (85.5%) | 3920 (14.5%) | <0.0001 |
| Some days | 10183 (8.7%) | 8937 (87.8%) | 1246 (12.2%) |  |
| Not at all | 79368 (68.1%) | 67286 (84.8%) | 12082 (15.2%) |  |
| **Alcohol consumption (n, %)** | | | | |
| Yes | 65320 (56.0%) | 58045 (88.9%) | 7275 (11.1%) | <0.0001 |
| No | 51288 (44.0%) | 41315 (80.6%) | 9973 (19.4%) |  |
| **Exercise (past 30 days) (n,%)** | | | | |
| Yes | 84759 (72.7%) | 74280 (87.6%) | 10479 (12.4%) | <0.0001 |
| No | 31849 (27.3%) | 25080 (78.7%) | 6769 (21.3%) |  |
| **Fruit consumption less than once per day (n, %)** | | | | |
| >= 1 day | 66667 (57.2%) | 56824 (85.2%) | 9843 (14.8%) | 0.7704 |
| < 1 day | 49941 (42.8%) | 42536 (85.2%) | 7405 (14.8%) |  |
| **Vegetable consumption less than once per day (n, %)** | | | | |
| >= 1 day | 95246 (81.7%) | 81672 (85.7%) | 13574 (14.3%) | <0.0001 |
| < 1 day | 21362 (18.3%) | 17688 (82.8%) | 3674 (17.2%) |  |
| **CLINICAL AND MENTAL HEALTH FACTORS** | | | | |
| **Depressive disorder status (n, %)** | | | | |
| No | 86510 (74.2%) | 74363 (86.0%) | 12147 (14.0%) | <0.0001 |
| Yes | 30098 (25.8%) | 24997 (83.1%) | 5101 (16.9%) |  |
| **Diabetes status (n, %)** |  |  |  |  |
| Yes | 17858 (15.3%) | 12339 (69.1%) | 5519 (30.9%) | <0.0001 |
| Yes, but female told only during pregnancy | 988 (0.8%) | 920 (93.1%) | 68 (6.9%) |  |
| No | 94852 (81.3%) | 83752 (88.3%) | 11100 (11.7%) |  |
| **Ever told you had a depressive disorder (n, %)** | | | | |
| Yes | 30098 (25.8%) | 24997 (83.1%) | 5101 (16.9%) | <0.0001 |
| No | 86510 (74.2%) | 74363 (86.0%) | 12147 (14.0%) |  |

**TABLE 2 -** **Comparative of Baseline Performance of Algorithms On The Original Training Data Using 10-Fold Cv**.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithms** | **Precision**  **(Weighted)** | **Precision**  **(Class 0)** | **Recall**  **(Weighted)** | **Recall**  **(Class 0)** | **F1 score**  **(Weighted)** | **AUROC** |
| **XGBoost**  **(optimized)** | 81.94 ± 0.34 | 90.61 ± 0.29 | 76.60 ± 0.32 | 80.93 ± 0.37 | 78.69 ± 0.28 | 76.16 ± 0.49 |
| **XGBoost (original)** | 81.97 ± 0.34 | 90.94 ± 0.33 | 74.54 ± 0.38 | 77.88 ± 0.65 | 77.28 ± 0.28 | 75.28 ± 0.52 |
| **SVM** | 81.55 ± 0.31 | 90.19 ± 0.32 | 76.98 ± 0.45 | 81.90 ± 0.82 | 78.83 ± 0.29 | 74.30 ± 0.51 |
| **CNN** | 81.99 ± 0.56 | 90.67 ± 0.88 | 76.32 ± 2.46 | 80.55 ± 4.12 | 78.43 ± 1.57 | 76.01 ± 0.39 |
| **MLP** | 81.54 ± 0.45 | 90.61 ± 0.59 | 73.84 ± 1.64 | 77.34 ± 2.70 | 76.66 ± 1.09 | 74.18 ± 0.49 |
| **RF** | 83.78 ± 0.19 | 93.76 ± 0.20 | 64.99 ± 0.51 | 63.12 ± 0.69 | 70.06 ± 0.43 | 75.95 ± 0.49 |
| **LR** | 83.73 ± 0.25 | 93.66 ± 0.25 | 65.64 ± 0.30 | 64.01 ± 0.41 | 70.60 ± 0.25 | 75.87 ± 0.40 |