BMC Medical Informatics and Decision Making

RESEARCH

Open Access

An interpretable machine learning model with demographic variables and dietary patterns for ASCVD identification: from U.S. NHANES 1999–2018

Qun Tang[1], Yong Wang[1] and Yan Luo[2*]

Abstract

Current research on the association between demographic variables and dietary patterns with atherosclerotic cardiovascular disease (ASCVD) is limited in breadth and depth. This study aimed to construct a machine learning (ML) algorithm that can accurately and transparently establish correlations between demographic variables, dietary habits, and ASCVD. The dataset used in this research originates from the United States National Health and Nutrition Examination Survey (U.S. NHANES) spanning 1999–2018. Five ML models were developed to predict ASCVD, and the best-performing model was selected for further analysis. The study included 40,298 participants. Using 20 population characteristics, the eXtreme Gradient Boosting (XGBoost) model demonstrated high performance, achieving an area under the curve value of 0.8143 and an accuracy of 88.4%. The model showed a positive correlation between male sex and ASCVD risk, while age and smoking also exhibited positive associations with ASCVD risk. Dairy product intake exhibited a negative correlation, while a lower intake of refined grains did not reduce the risk of ASCVD. Additionally, the poverty income ratio and calorie intake exhibited non-linear associations with the disease. The XGBoost model demonstrated significant efficacy, and precision in determining the relationship between the demographic characteristics and dietary intake of participants in the U.S. NHANES 1999–2018 dataset and ASCVD.

Keywords Atherosclerotic cardiovascular disease, Demographic variables, Dietary patterns, Machine learning, NHANES

*Correspondence:
Yan Luo
luo.yan@imicams.ac.cn
[1]Department of Cardiovascular Medicine, Wuhu City Second People's Hospital, Wuhu 241000, China
[2]Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

Introduction

Atherosclerotic cardiovascular disease (ASCVD) is a collection of conditions caused by atherosclerosis [1], a pathological process characterized by arterial inflammation and plaque formation, eventually leading to arterial narrowing and reduced blood flow, which can precipitate serious cardiovascular events, including myocardial infarction and cerebrovascular accidents [2]. Numerous risk factors have been associated with ASCVD, such as hypertension, hyperglycemia, hyperlipidemia, and obesity [3, 4]. The association between population characteristics and ASCVD has received little attention, with most studies using conventional statistical methods [3]. Using a novel analytical approach could improve the accuracy of this association.

Previous studies using conventional statistical approaches for disease detection identified numerous data standards and preparation prerequisites [5, 6]. Certain traditional statistical techniques require well-organized data with a high degree of accuracy, resulting in the exclusion of a significant amount of unstructured data. Advances in science and technology have enabled the efficient and extensive collection of data. Consequently, processing, categorizing, and interpreting large datasets present a significant research challenge for scholars investigating underlying patterns across multiple sources [7]. Machine learning (ML) algorithms in "black-box" methods require less data preparation, allowing researchers to analyze large datasets for purposes including disease diagnosis and health decision-making [8]. Recently, numerous studies have focused on applying ML techniques to predict peripheral artery disease and cardiovascular events [9, 10, 11].

In this study, we analyzed datasets from the United States National Health and Nutrition Examination Survey (U.S. NHANES, 1999–2018) to investigate the relationships between demographic variables, dietary patterns, and ASCVD. Based on population characteristics, five ML models were selected and compared to detect ASCVD.

Methods

Data source

The NHANES, conducted by the National Center for Health Statistics (NCHS), is a significant program that collects nationally representative health data on the general population in the United States. It employs a sophisticated, multistage, stratified probability sampling methodology to poll approximately 5,000 individuals annually through cross-sectional surveys, which have been conducted biennially since 1999. To ensure sample representativeness, NHANES employs oversampling techniques for specific subsets of the population.

The NHANES protocol was approved by the NCHS Research Ethics Review Board, and all participants provided informed consent. NHANES survey data, comprehensive survey operation manuals, consent documents, and brochures from each period are available on the NHANES website (https://www.cdc.gov/nchs/nhanes, accessed on 8 February 2023).

Study participants

The NHANES 1999–2018 survey initially included 116,876 participants. After excluding those under 18 years old (47,979), those with missing ASCVD values (24,891), and rows containing missing data (3,708), the final analysis included 40,298 participants. This group comprised 4,189 individuals with ASCVD and 36,109 individuals without ASCVD (Fig. 1).

Outcome ascertainment

The research identified the presence of ASCVD according to the 2013 American College of Cardiology/American Heart Association (ACC/AHA) guidelines, with at least one diagnosis of coronary heart disease, angina, heart attack, or stroke [12]. More stringent criteria were defined as a history of heart attack or stroke.

Study variables

Demographic information on the study participants was collected using the NHANES questionnaire, which included variables such as gender, age, race/ethnicity, education level, and poverty income ratio (PIR). Education level was categorized into three groups: less than high school, high school or equivalent, and more than high school. PIR was considered a continuous variable, representing the ratio of self-reported family income adjusted for family or household size, composition, and year. Additionally, dietary habits, including calorie intake, the healthy eating index (HEI-2015), smoking status (never smokers, former smokers, or current smokers), and coffee consumption, were included. Our study used the most recent version of the HEI-2015 to evaluate the relationship between diet and ASCVD [13]. The HEI-2015 is a valuable tool for assessing dietary quality based on the mainstream eating habits of 185 countries [14]. It considers 13 dietary components.

Data preprocessing

We used the criteria described in Fig. 1 to build our dataset. Participants under the age of 18 and those with missing ASCVD outcomes were excluded. Missing data are unavoidable in epidemiological and clinical research. Because the number of missing values in the dataset was small (3708 samples, 8.4% of data), we excluded the samples that contained missing values directly. Out of 40,298 records, 36,109 were classified as normal and

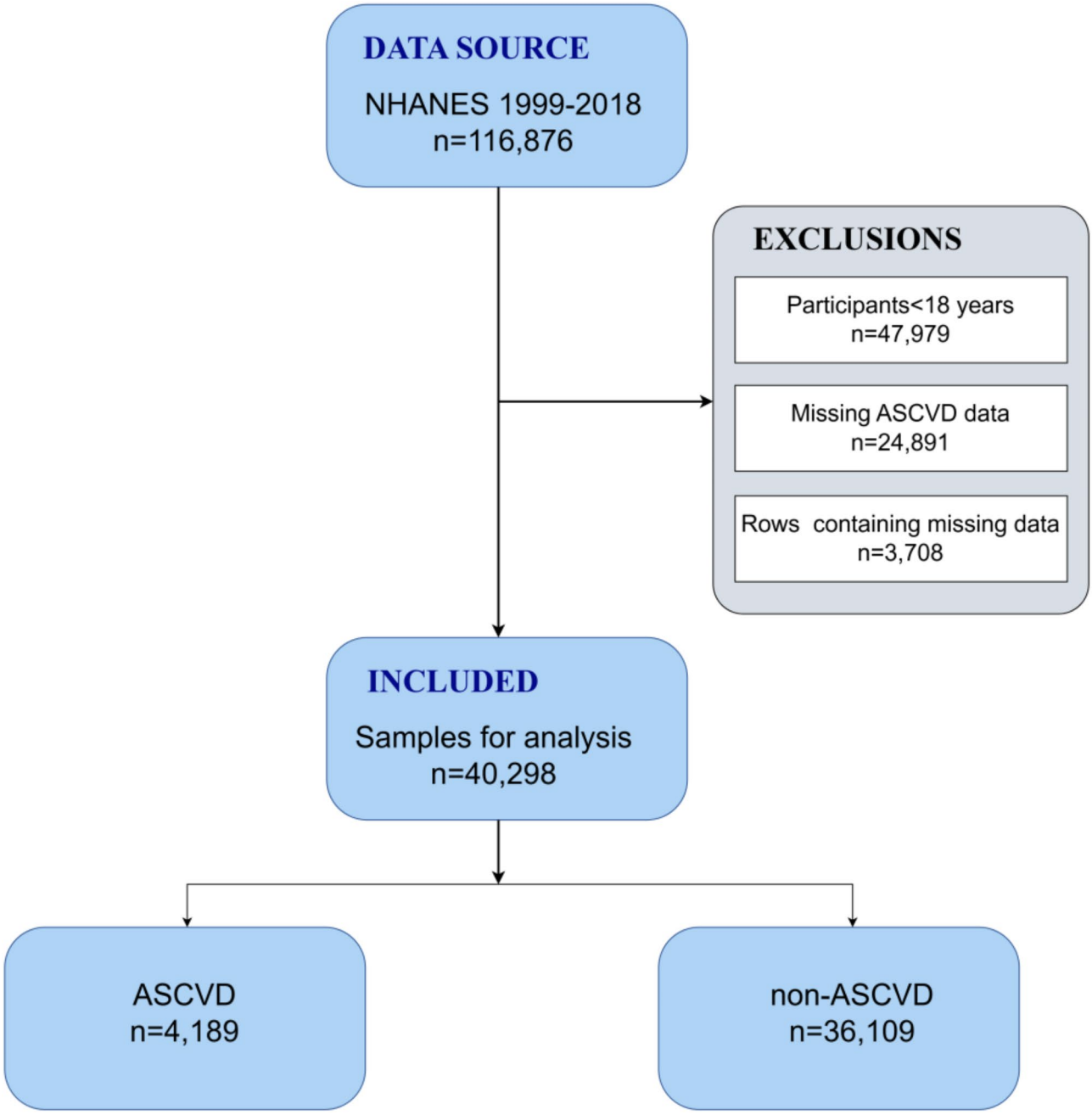

Fig. 1 Flowchart of the participants selection from NHANES 1999–2018. NHANES, national health and nutrition examination survey

4,189 as ASCVD cases (Table 1). The class distribution of our dataset was highly imbalanced, with a ratio close to 9:1. We employed the synthetic minority oversampling technique and edited the nearest neighbor (SMOTE-ENN) resampling technique to balance the data [15]. SMOTE-ENN is a commonly used method for addressing data imbalance problems. It combines oversampling of the minority class and undersampling of the majority class while maintaining the overall size of data, thereby improving model prediction accuracy. Finally, since the value range of the original data in the dataset was already determined, we used the min-max normalization method to normalize our data feature values to the range of [0, 1], thus preserving the relationships among the original data values. Min-max normalization is a linear normalization method used to scale data to a specific range, typically between 0 and 1. It is commonly used in ML and data preprocessing to prevent features with different scales from disproportionately influencing model performance.

Table 1 The demographic characteristics of the study participants were analyzed based on their ASCVD status in the NHANES dataset from 1999 to 2018

Characteristics	ASCVD		*P* value
	ASCVD(*n* = 4189)	non-ASCVD (*n* = 36109)	
HEI[1], total score, mean(SD)	53.757(13.318)	53.136(13.28)	0.004
total vegetable	3.283(1.574)	3.272(1.503)	0.640
green and bean	1.753(2.144)	1.932(2.172)	0.000
total fruit	2.709(2.013)	2.506(2.01)	0.000
whole fruit	2.85(2.177)	2.618(2.197)	0.000
whole grain	2.977(3.294)	2.462(3.043)	0.000
total dairy	4.808(3.192)	5.021(3.167)	0.000
tot prot	4.452(1.033)	4.428(1.053)	0.167
sea plant prot	2.629(2.203)	2.74(2.184)	0.002
fatty acid	5.094(3.379)	5.06(3.359)	0.536
sodium	4.184(3.435)	4.456(3.373)	0.000
add sug	6.809(3.285)	6.504(3.335)	0.000
refined grain	6.105(3.457)	5.915(3.486)	0.001
sfat	6.103(3.281)	6.22(3.281)	0.026
Energy[2], kcal, mean(SD)	3608.51(1555.138)	4156.73(1748.037)	0.000
Alcohol consumption, gm, mean(SD)	101.334(416.947)	179.1(573.37)	0.000
Coffee, gm, mean(SD)	364.325(512.788)	269.424(406.619)	0.000
Age, years, mean(SD)	66.873(12.805)	47.682(17.534)	0.000
Household_size[3], mean(SD)	2.419(1.391)	3.171(1.645)	0.000
PIR[4], mean(SD)	2.248(1.485)	2.614(1.629)	0.000
Smoke status (n,%)			0.000
current smokers	825(19.7)	7319(20.3)	
former smokers	1766(42.2)	8590(23.8)	
never smokers	1598(38.1)	20,200(55.9)	
Gender (n,%)			0.000
female	1759(42.0)	19,383(53.7)	
male	2430(58.0)	16,726(46.3)	
Citizenship (n,%)			0.000
citizen by birth or naturalization	4019(95.9)	31,250(86.5)	
not a citizen of the US	170(4.1)	4859(13.5)	
Educational attainment (n,%)			0.000
< High school	1416(33.8)	8719(24.1)	
> High school	1691(40.4)	19,134(53.0)	
High school	1082(25.8)	8256(22.9)	
Race-ethnicity (n,%)			
Mexican American	459(11.0)	6340(17.6)	
Black(non-hispanic)	830(19.8)	7438(20.6)	0.2337
White(non-hispanic)	2464(58.8)	16,535(45.8)	0.000
Other hispanic	235(5.6)	2795(7.7)	0.000
Other race	201(4.8)	3001(8.3)	0.000

Continuous variables are typically represented by means, while categorical variables are typically represented by proportions

[1] The Healthy Eating Index–2015

[2]Energy intake for two days

[3]Total number of people in the household

[4]Ratio of family income to poverty

ML model strategies

We used five ML algorithms—logistic regression (LR), artificial neural networks (ANNs), support vector machine (SVM), random forest (RF), and eXtreme gradient boosting (XGBoost)—to classify patients with ASCVD. LR is a commonly used supervised classification algorithm that predicts the probability of a target label [16]. ANNs are biologically inspired computer programs comprising hundreds of artificial neurons connected by weighted coefficients to form the neural structure [17].

SVM classifies data by separating classes using a boundary, such as a line or multi-dimensional hyperplane [18]. RF and XGBoost are ensemble prediction models based on decision trees, making them suitable for classification tasks [19].

Performance metrics

To validate the stability and reproducibility of the models, a 10-fold cross-validation resampling was performed to ensure consistent model performance across different data subsets, thereby reducing the likelihood of overfitting. The area under the curve (AUC) of the receiver operating characteristic curve was used as a statistical measure to evaluate the performance of the models. Additionally, indicator accuracy, precision, recall, and F1-score were used to analyze the performance of different classifiers [20]. SHapley Additive exPlanations (SHAP) is a game-theoretic approach used to explain the outputs of ML models. SHAP uses SHapley values to determine how and to what extent each feature contributes to the final result [21]. Accordingly, we employed SHAP to analyze the performance of the best ML model.

Statistical analysis

The descriptive statistical analysis included a summary of the demographic characteristics of the study participants, stratified by the presence of ASCVD. Independent samples t-tests were used for continuous variables with homogeneity of variance, whereas Welch's t-tests were used for variables without homogeneity of variance. Differences in categorical variables were evaluated using Pearson's chi-square test.

Results

Demographic characteristics of the study participants

The demographic profile of the study participants is presented in Table 1. The mean age of individuals with ASCVD was 66, whereas that of individuals without ASCVD was 47. In the ASCVD group, 58% were males, and 42% were females. The primary racial demographic was White (non-Hispanic), accounting for 58.8% of the population.

ML to predict outcomes

The evaluation results of ML classifiers are presented in Table 2. All five models demonstrated good efficiency and stability. The XGBoost model performed the best with an AUC of 0.8143, followed by the ANNs (AUC = 0.8135), SVM (AUC = 0.8111), LR (AUC = 0.8088), and RF (AUC = 0.8075) models. The SVM model exhibited the highest precision, with a value of 0.8721. The XGBoost model achieved the highest F1-score with values of 0.8631. The F1-score represents the weighted harmonic mean of precision and recall, effectively capturing the model's overall performance in terms of both recall and precision metrics. Given the stability of the evaluation metrics for the XGBoost model and the potential heightened significance of the F1-score in clinical contexts, we have chosen this model for the identification of ASCVD.

Comparison with public prediction performance

We conducted a systematic review of the literature on ASCVD risk prediction published in PubMed between 2019 and 2024 using the search keywords "NHANES" and "ASCVD." Following a comprehensive filtering process based on title, topic, abstract, and full-text examination, we identified five articles that used ML methodologies to predict ASCVD risk (Table 3).

Explaining the XGBoost model using SHAP

Experimental results revealed that the XGBoost model outperformed others, especially in terms of F1-score. Therefore, SHAP analysis was conducted using the XGBoost model. We used SHAP to visually explain the selected variables and their relationships with ASCVD occurrence. Figure 2 displays the top 20 most important features in our model. The points in the plots represent different cases, color-coded based on the value of the corresponding variable on the y-axis and the SHapley value on the x-axis. The colors indicate whether the variable is high (red) or low (blue), with the value associated with a higher or lower prediction. Consequently, based on the values of the variables, we obtained information on the magnitude and direction of their contributions to the model.

As illustrated in Fig. 2, age had the highest value across all characteristic horizons, followed by PIR, smoking

Table 2 Shows the performance (mean of ten-fold cross-validation) for each performance parameter with 95% confidence intervals based on the LOS class

Classifiers	Accuracy [95% CI]	Precision [95% CI]	Recall [95% CI]	F1-Score [95% CI]	AUC [95% CI]
LR	0.8952 (0.8948, 0.8956)	0.8356 (0.8298, 0.8414)	0.8952 (0.8948, 0.8956)	0.8477 (0.8474, 0.8481)	0.8088 (0.8020, 0.8157)
ANNs	0.8240 (0.8182, 0.8299)	0.8718 (0.8695, 0.8741)	0.8240 (0.8182, 0.8299)	0.8434 (0.8400, 0.8468)	0.8135 (0.8082, 0.8188)
SVM	0.8216 (0.8154, 0.8277)	0.8721 (0.8701, 0.8742)	0.8216 (0.8154, 0.8277)	0.8419 (0.8376, 0.8463)	0.8111 (0.8058, 0.8165)
RF	0.8863 (0.8837, 0.8889)	0.8487 (0.8450, 0.8523)	0.8863 (0.8837, 0.8889)	0.8601 (0.8578, 0.8623)	0.8075 (0.8037, 0.8114)
XGBoost	0.8841 (0.8831, 0.8852)	0.8590 (0.8584, 0.8596)	0.8841 (0.8831, 0.8852)	0.8631(0.8616, 0.8646)	0.8143 (0.8094, 0.8191)

Table 3 Comparison with state-of-the-art ASCVD prediction using NHANES data

Researches	Purpose	Features	Sample size	Algorithms	AUC	Year of publication
Chen et al. [22]	Identify chronic heart disease	20 variables	14,971	support vector machine (SVM)	0.898	2023
Inoue et al. [23]	Predict low HbA1c levels and all-cause or cardiovascular mortality	72 variables	39,453	SuperLearner	0.90	2020
Li et al. [24]	Heavy metals' exposure to identify coronary heart disease	13 heavy metals	12,554	random forest (RF)	0.827	2023
Martin-Morales et al. [25]	Predict cardiovascular disease mortality	59 variables	9,706	RF	0.88	2023
Wang et al. [26]	Predict coronary heart disease risk in patients with periodontitis	29 variables	3,245	K-nearest neighbor	0.977	2023

status, household size, gender, and race. Furthermore, increases in age exerted a positive impact and pushed the prediction toward ASCVD, whereas increases in PIR exerted a negative impact and pushed the prediction toward non-ASCVD. Further investigation into the relationship between ASCVD and diet revealed that coffee consumption, alongside moderate intake of dietary salt and vegetables, is beneficial in preventing ASCVD. The recommended intake of refined grains by the HEI does not apply to patients with ASCVD.

Analysis of potential interactions between features based on SHAP values

Figure 2 provides global insights into our XGBoost model. For local model explanations, other outputs from the SHAP framework focused on individual or pairs of variables. As illustrated in Fig. 3, we selected the top three dietary-related features from the HEI index system, alongside the "kcal" (total calorie intake over two days) intake from the NHANES database, for further analysis of the relationship between diet and model output. We observed two different trends in the recommended appropriate intake (higher scores indicating reduced intake) of salt and refined grains from the HEI score. While reduced salt intake was correlated with lower ASCVD risk, the same was not true for refined grains. Reduced consumption of refined grains did not correlate with a decreased ASCVD risk (Fig. 3A-B). Figure 3C indicates a positive association between the recommended intake of dairy products and ASCVD prevention. Higher scores indicate higher intake. Figure 3D presents the non-linear relationship between the "kcal" feature and its corresponding SHAP values. A total calorie intake of 5,000 kcal over two days appears to be an ideal value.

We identified potential interactions between features by analyzing the interaction plots using SHAP values. We enhanced the age-dependence scatter plot by incorporating color coding based on another feature (Fig. 4). We analyzed the pattern and trend of the relationship between age and the model's output while considering different levels of "kcal" and PIR. Individuals who consumed more than 5,000 kcal were aged 20 to 45 (Fig. 4A). This age range typically represents the primary workforce in society, thus having higher energy demands. The plot indicates that even though these individuals had a higher calorie intake every two days, their likelihood of developing ASCVD remained relatively low. Figure 4B suggests that individuals under the age of 60 viewed PIR as a protective factor against ASCVD, whereas those aged 60 and above viewed it as a negative influence.

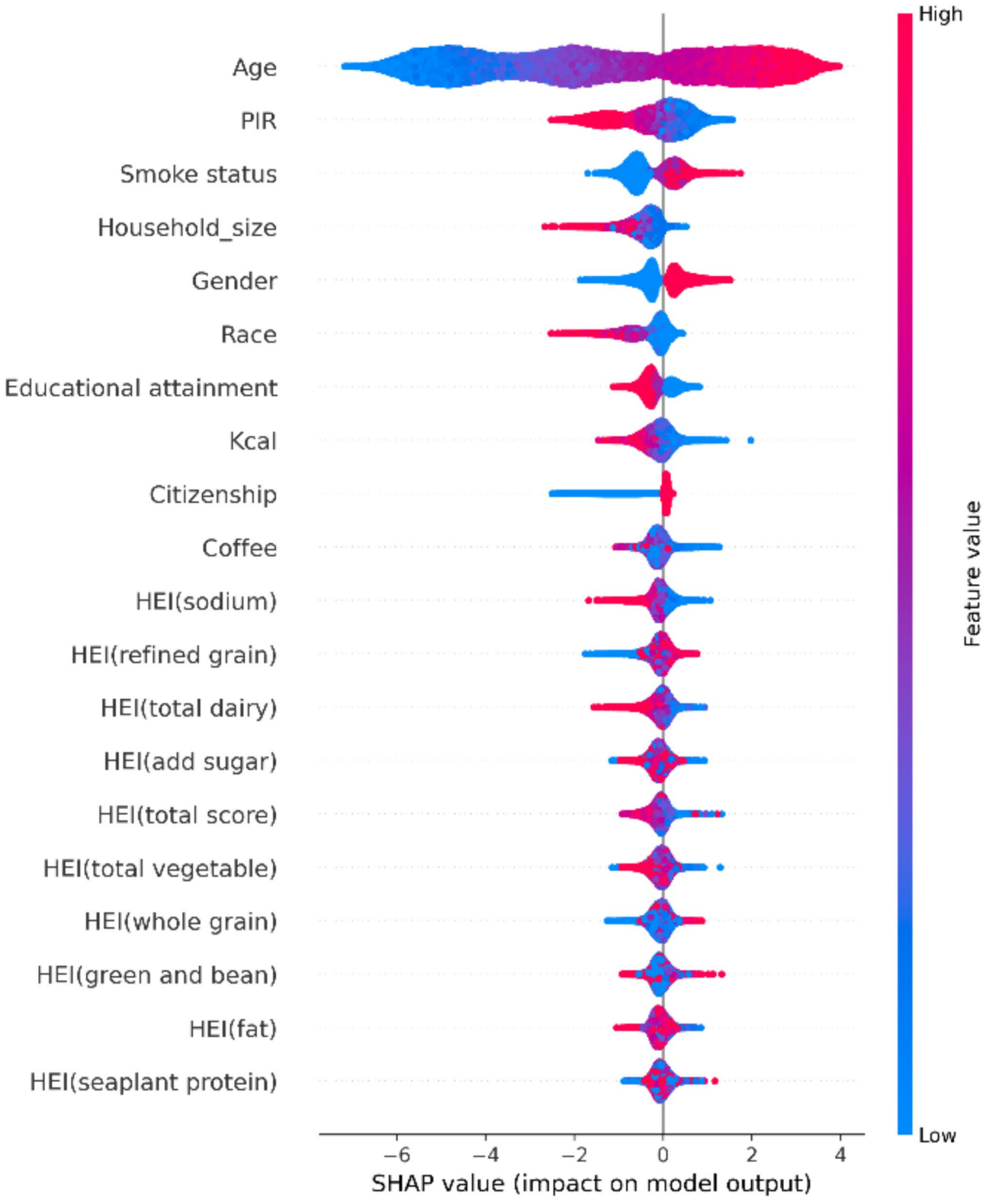

Fig. 2 SHAP summary plot for model interpretation. The importance ranking of the top 20 risk factors with stability and interpretation using the optimal model. The higher SHAP value of a feature is given, the higher risk of a patient developing ASCVD. The red part in feature value represents higher value

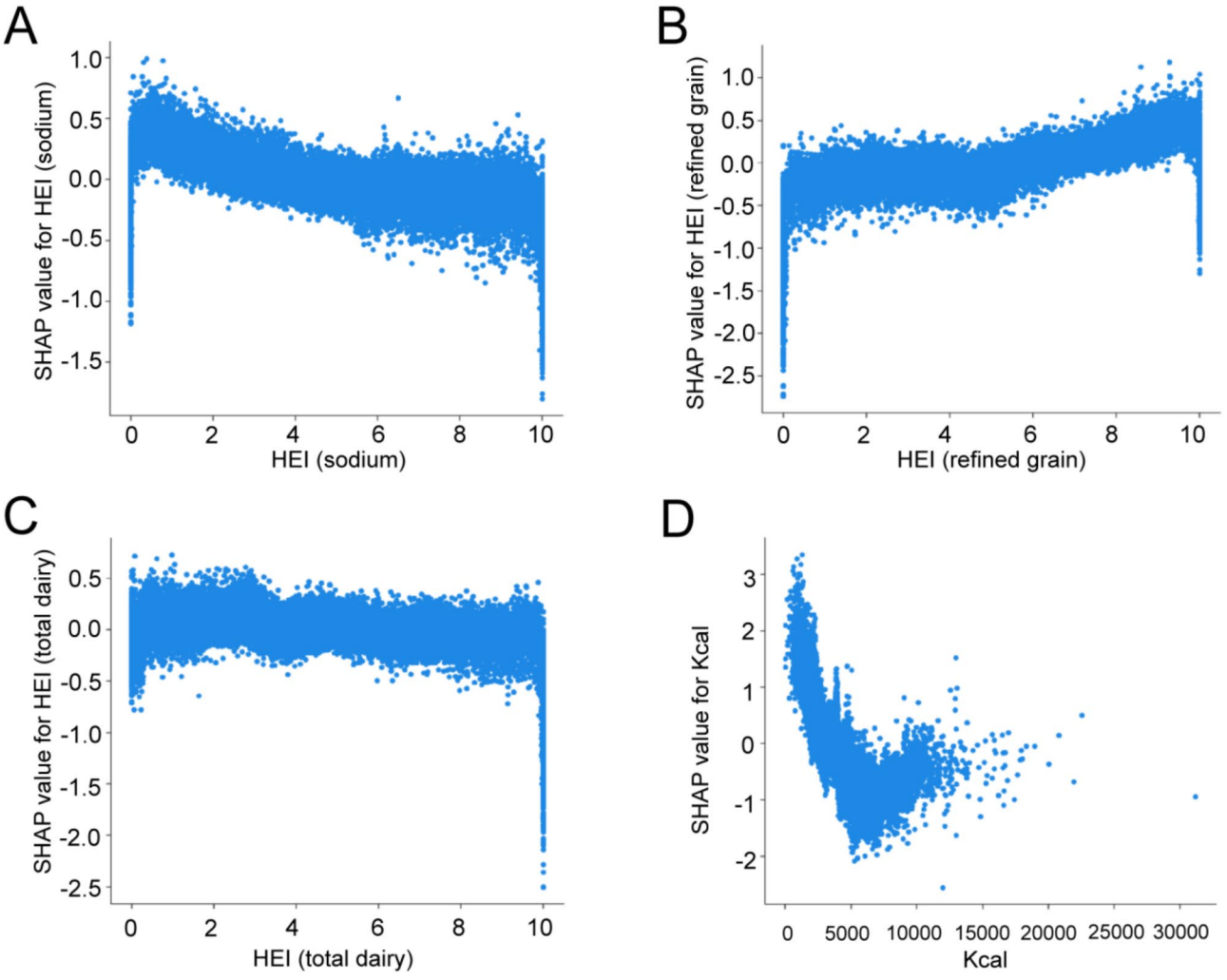

Fig. 3 SHAP dependence plot for model interpretation. The x-axis displays four distinct feature values, while the y-axis represents the SHAP values associated with each of these individual feature values

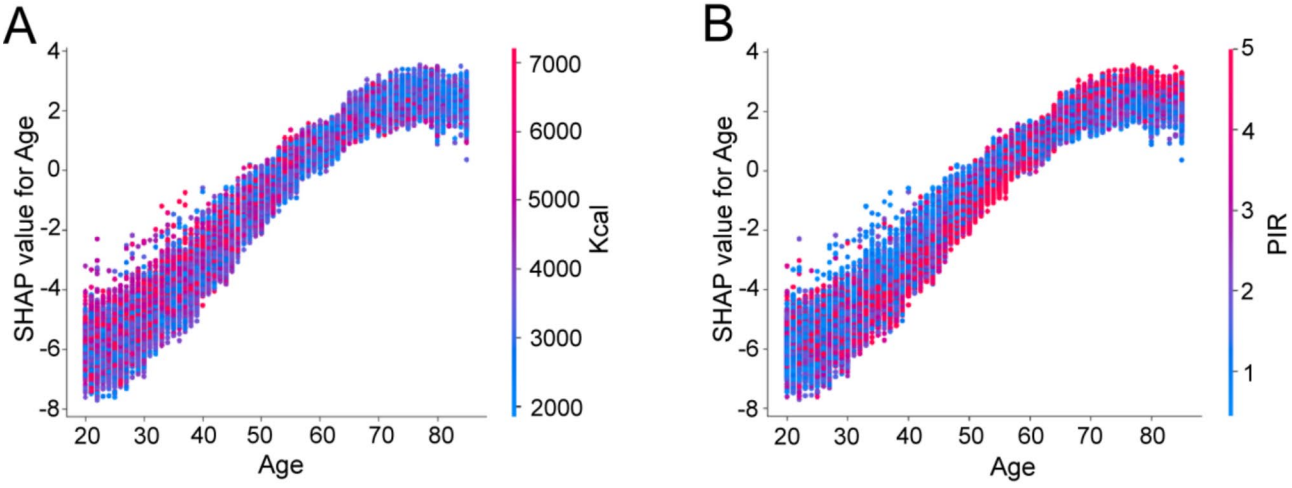

Fig. 4 SHAP interactive plot for model interpretation. The x-axis represents the range of values for the Age feature, while the y-axis represents the SHAP values for the AGE feature, indicating the change it brings to the model's output. On the right side are the contrasting features Kcal (**A**) and PIR (**B**), where red represents the high-score section and blue represents the low-score section

Discussion

This study employed ML to analyze how demographic and dietary data in the U.S. NHANES dataset from 1999 to 2018 relate to ASCVD. Among the five ML models evaluated, the XGBoost model demonstrated the best performance in classification and stability, with an AUC of 0.8143 and a F1-score of 0.8631.

We used SHAP to visually explain the selected variables and their relationships with ASCVD occurrence. Positive SHAP values imply that the feature values associated with it contributed to an elevated risk of ASCVD throughout the 20-year period of the U.S. NHANES survey, whereas negative SHAP values indicate a reduced risk. Our study ranked the features based on the average absolute value of SHapley scores and identified age, PIR, and smoking status as the most significant factors impacting the models. Specifically, advanced age, a low PIR, and smoking are associated with an increased ASCVD risk (Fig. 2). Furthermore, larger household sizes and higher education levels effectively reduced ASCVD risk. Figures 3 and 4 present the dependency graphs, which illustrate the marginal effects of one (Fig. 3) or two (Fig. 4) features on the predicted outcomes of the ML model. These graphs indicate whether the relationship between the target variable and the features is linear, monotonous, or more complicated. As illustrated in Fig. 3A, a lower sodium HEI score (less than 2) was associated with an increased likelihood of ASCVD risk, whereas Fig. 3B depicts a reversed trend. When the HEI score for refined grains exceeded 8, there was a positive correlation with the incidence of ASCVD. Figure 3C demonstrates that a total dairy HEI score of 10 significantly reduced ASCVD risk. This suggests that higher intake levels, as indicated by higher scores, can help reduce ASCVD risk. Figure 3D illustrates a non-linear relationship between the "kcal" feature and its SHAP value. A kcal value below 2,500 significantly increased ASCVD risk, while 5,000 kcal optimally reduced this risk, implying that it is the ideal total calorie intake every two days. In Fig. 4, the x-axis represents the range of characteristic values for the age feature, and the y-axis denotes the SHapley value of the age feature. Figure 4A-B depict the interaction between age on the x-axis and kcal and PIR on the y-axis, specifically for age values exceeding 70 years. The SHapley value remained relatively constant, suggesting that age exerts little influence on the target variable predicted by the model within this range. Furthermore, color variations were used to analyze the interaction effect between the primary and interaction features. In Fig. 4A, the majority of the red points are below $y = 0$, indicating that consuming 5,000 to 6,000 kcal every 2 days had a negative effect on individuals under the age of 60. Figure 4B demonstrates that for individuals under the age of 60, a high PIR exerted a more significant negative effect on the model's output. Notably, PIR exerted a more significant positive effect on individuals over 60 years old. A key limitation of SHAP is its high computational complexity. The computation of SHapley values requires the consideration of all potential feature combinations, making it computationally intensive, especially for large datasets and complex models. Furthermore, the order in which features are evaluated can have an impact on the interpretability of SHAP values, potentially resulting in unstable outcomes.

Recently, ML has rapidly advanced as a predictive classifier that uses historical data to forecast future outcomes. Previous research used ML techniques to develop predictive models for various risk factors linked to ASCVD [27, 28, 29]. The researchers employed several ML models, including RFs, SVMs, and decision trees, to investigate the correlation between heavy metals and hypertension. Such models are adept at handling complex non-linear relationships, thereby improving the precision of predictions [27]. Several ML models, including LR, RF, and XGBoost, were used to forecast ASCVD risk over five years. The findings suggest that the ML models can provide improved risk predictions across a wider patient population, potentially aiding ASCVD management and prevention [28]. Diverse omics and clinical data can be combined using artificial intelligence and ML methodologies to develop personalized diagnostic and treatment strategies for ASCVD [29]. However, there is limited research on predictive models that examine the relationship between dietary habits and ASCVD.

In this study, the top three dietary-related features (sodium, refined grains, and total dairy) from the HEI index system and the "kcal" intake from the NHANES database were selected for further examination to analyze the correlation between diet and model output. Sodium plays a critical role in fluid balance and cellular homeostasis. Excess sodium is linked to heightened arterial stiffness and unfavorable ventricular remodeling, thereby significantly contributing to ASCVD development on a larger scale [30, 31, 32]. Our findings indicate a correlation between reduced salt consumption and decreased ASCVD risk, consistent with the findings of De and Shan et al. [33, 34]. Whole grains are often recommended as a healthy dietary option because they contain beneficial nutrients and phytochemicals, which are mostly found in the outer layers of the grains. These layers are removed during milling to produce refined grain products. Prior research indicated that the consumption of certain whole grains, such as whole grain bread, whole grain breakfast cereals, and added bran, is associated with a reduced risk of cardiovascular diseases [35, 36]. However, there is limited evidence establishing a connection between refined grains, white rice, total rice, or total grain consumption and reduced risk of cardiovascular diseases. In our study, a significant finding was that reducing refined grain

intake within a specific range did not exhibit a linear correlation with reduced ASCVD risk.

Dairy products are a good source of essential nutrients, such as energy, calcium, and protein; however, they may contain high levels of saturated fatty acids. The idea that saturated fats contribute to ASCVD risk has influenced dietary guidelines [37]. However, some studies suggest that the impact of saturated fats on this risk may vary depending on the source [38, 39]. Because dairy products are a good source of essential nutrients, there is a tendency to increase their consumption within dietary regimens. Our findings support our initial hypothesis that adhering to the recommended dairy product intake can help prevent ASCVD.

Our analysis of the NHANES data focused on two days of total energy consumption. Prior research suggested a correlation between increased energy consumption in older adults or the general population and reduced mortality rates [40, 41]. A 25-year research in Cuba discovered that reducing energy intake while maintaining proper nutrition could decrease the risks of diabetes and cardiovascular diseases. During the 1997–2002 economic crisis, lower energy consumption resulted in a lower incidence of diabetes and heart disease. However, as energy consumption increased in subsequent years, heart disease mortality rates increased slightly [42]. In addition to demonstrating a non-linear relationship between kcal and SHAP values, our findings suggest that 5,000 kcal every two days is the optimal daily calorie intake.

PIR, which measures the relationship between family income and poverty threshold, was used as a socioeconomic status index. Numerous studies have indicated that individuals with lower socioeconomic status are more likely to develop cardiovascular diseases and experience all-cause mortality [43, 44, 45]. In our experimental findings, a high PIR appeared to protect against ASCVD in individuals below 60 years old, whereas it exerted a negative impact on those aged 60 and above. This trend may be associated with the higher dietary energy intake requirements of individuals below 60 with a high PIR. However, this finding requires further investigation through additional data, such as information on the specific occupations of these demographic subsets, for example, engagement in physical labor.

Limitations

This study has several limitations. First, we did not break down race, age, or other characteristics in the subgroup analyses owing to computational constraints when analyzing protected health-sensitive data. Second, ASCVD diagnosis partly relied on self-reported data collected using an interview questionnaire in the U.S. NHANES. This might have introduced information bias due to potential cognitive limitations and recall biases of respondents. Additionally, some participants might have experienced multiple ASCVD events or taken the survey multiple times, complicating the process of linking their dietary information to the nearest ASCVD event. Consequently, any misclassification of ASCVD could have impacted the accuracy of the ML models in detecting ASCVD. Moreover, approximately 10% of cases had missing variables when used in the ML models, leading to biases in the analysis. Furthermore, the findings derived from the SHAP analysis in this study are based on a single training/test partition. Consequently, these interpretations may not be applicable to all ten replicates, and variations in feature importance may exist across different partitions. Lastly, the complexity and interpretive challenges of the models might have affected their reproducibility.

Conclusions

The XGBoost model demonstrated significant efficacy and precision in determining the relationship between the demographic characteristics and dietary intake of participants in the U.S. NHANES 1999–2018 dataset and ASCVD. Besides, the model showed a positive correlation between male sex and ASCVD risk, while age and smoking also exhibited positive associations with ASCVD risk, dairy intake displaying a negative correlation, and PIR and calorie intake revealing non-linear associations with the disease.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02937-5.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

Qun Tang and Yan Luo wrote the main manuscript text and prepared figures. Yong Wang provided financial support. All authors reviewed the manuscript.

Funding

This research was funded by the Wuhu Science and Technology Bureau (2022cg26). Research Projects of Wuhu Municipal Health Commission (WHWJ2023z016).

Data availability

All analysis data in this article are from publicly available databases. Users can download relevant data for free for research and publish relevant articles. (https://www.cdc.gov/nchs/nhanes/index.htm)

Declarations

Ethics approval and consent to participate

The present study utilized data from the National Health and Nutrition Examination Survey (NHANES), which is conducted by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC). NHANES is a publicly available database that collects health information from a nationally representative sample of the US population.

Ethical approval for NHANES was obtained by NCHS, and all participants provided written informed consent prior to their participation in the survey. The consent form explained the purpose of the survey, procedures involved, potential risks and benefits, confidentiality measures, and the right to withdraw from the survey at any time without penalty. Participants were also informed that their data would be kept confidential and used only for research purposes.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 25 June 2024 / Accepted: 18 February 2025
Published online: 03 March 2025

References

1. Lindbohm JV, Sipilä PN, Mars NJ, Pentti J, Ahmadi-Abhari S, Brunner EJ, Shipley MJ, Singh-Manoux A, Tabak AG, Kivimäki M. 5-year versus risk-category-specific screening intervals for cardiovascular disease prevention: a cohort study. Lancet Public Health. 2019;4:e189–99. https://doi.org/10.1016/s2468-2667(19)30023-4.
2. Murray KN, Parry-Jones AR, Allan SM. Interleukin-1 and acute brain injury. Front Cell Neurosci. 2015;9:18. https://doi.org/10.3389/fncel.2015.00018.
3. Cainzos-Achirica M, Glassner K, Zawahir HS, Dey AK, Agrawal T, Quigley EMM, Abraham BP, Acquah I, Yahya T, Mehta NN, Nasir K. Inflammatory bowel disease and atherosclerotic cardiovascular disease: JACC review topic of the week. J Am Coll Cardiol. 2020;76:2895–905. https://doi.org/10.1016/j.jacc.2020.10.027.
4. Wong ND, Budoff MJ, Ferdinand K, Graham IM, Michos ED, Reddy T, Shapiro MD, Toth PP. Atherosclerotic cardiovascular disease risk assessment: an American society for preventive cardiology clinical practice statement. Am J Prev Cardiol. 2022;10:100335. https://doi.org/10.1016/j.ajpc.2022.100335.
5. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inf Decis Mak. 2019;19:211. https://doi.org/10.1186/s12911-019-0918-5.
6. Jun S, Cowan AE, Dodd KW, Tooze JA, Gahche JJ, Eicher-Miller HA, Guenther PM, Dwyer JT, Potischman N, Bhadra A, Forman MR, Bailey RL. Association of food insecurity with dietary intakes and nutritional biomarkers among US children, National health and nutrition examination survey (NHANES) 2011–2016. Am J Clin Nutr. 2021;114:1059–69. https://doi.org/10.1093/ajcn/nqab113.
7. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. NPJ Digit Med. 2020;3:30. https://doi.org/10.1038/s41746-020-0229-3.
8. Alber M, Buganza Tepole A, Cannon WR, De S, Dura-Bernal S, Garikipati K, Karniadakis G, Lytton WW, Perdikaris P, Petzold L, Kuhl E. Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. NPJ Digit Med. 2019;2:115. https://doi.org/10.1038/s41746-019-0193-y.
9. Ho V, Brown Johnson C, Ghanzouri I, Amal S, Asch S, Ross E. Physician- and patient-Elicited barriers and facilitators to implementation of a machine Learning-Based screening tool for peripheral arterial disease: preimplementation study with physician and patient stakeholders. JMIR Cardio. 2023;7:e44732. https://doi.org/10.2196/44732.
10. Milosevic M, Jin Q, Singh A, Amal S. Applications of AI in multi-modal imaging for cardiovascular disease. Front Radiol. 2023;3:1294068. https://doi.org/10.3389/fradi.2023.1294068.
11. Omiye JA, Ghanzouri I, Lopez I, Wang F, Cabot J, Amal S, Ye J, Lopez NG, Adebayo-Tijani F, Ross EG. Clinical use of polygenic risk scores for detection of peripheral artery disease and cardiovascular events. PLoS ONE. 2024;19:e0303610. https://doi.org/10.1371/journal.pone.0303610.
12. Stone NJ, Robinson JG, Lichtenstein AH, Bairey Merz CN, Blum CB, Eckel RH, Goldberg AC, Gordon D, Levy D, Lloyd-Jones DM, McBride P, Schwartz JS, Shero ST, Smith SC Jr., Watson K, Wilson PW. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American college of cardiology/american heart association task force on practice guidelines. J Am Coll Cardiol. 2014;63:2889–934. https://doi.org/10.1016/j.jacc.2013.11.002.
13. Krebs-Smith SM, Pannucci TE, Subar AF, Kirkpatrick SI, Lerman JL, Tooze JA, Wilson MM, Reedy J. Update of the healthy eating index: HEI-2015. J Acad Nutr Diet. 2018;118:1591–602. https://doi.org/10.1016/j.jand.2018.05.021.
14. Miller V, Webb P, Cudhea F, Shi P, Zhang J, Reedy J, Erndt-Marino J, Coates J, Mozaffarian D. Global dietary quality in 185 countries from 1990 to 2018 show wide differences by Nation, age, education, and urbanicity. Nat Food. 2022;3:694–702. https://doi.org/10.1038/s43016-022-00594-9.
15. Lin M, Zhu X, Hua T, Tang X, Tu G, Chen X. Detection of ionospheric scintillation based on XGBoost model improved by SMOTE-ENN technique. Remote Sens. 2021;13:2577.
16. Qin Y, Wu J, Xiao W, Wang K, Huang A, Liu B, Yu J, Li C, Yu F, Ren Z. Machine learning models for Data-Driven prediction of diabetes by lifestyle type. Int J Environ Res Public Health. 2022;19. https://doi.org/10.3390/ijerph192215027.
17. Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal. 2000;22:717–27. https://doi.org/10.1016/s0731-7085(99)00272-1.
18. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. https://doi.org/10.1007/BF00994018.
19. Zhao M, Wan J, Qin W, Huang X, Chen G, Zhao X. A machine learning-based diagnosis modelling of type 2 diabetes mellitus with environmental metal exposure. Comput Methods Programs Biomed. 2023;235:107537. https://doi.org/10.1016/j.cmpb.2023.107537.
20. Goutte C, Gaussier E. A Probabilistic Interpretation of precision, recall and F-Score, with implication for evaluation. Heidelberg: Springer Berlin Heidelberg, Berlin; 2005. pp. 345–59.
21. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions, Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, California, USA, 2017, pp. 4768–4777.
22. Chen X, Guo D, Wang Y, Qu Z, He G, Sui C, Lan L, Zhang X, Duan Y, Meng H, et al. Using machine learning algorithms to identify chronic heart disease: National Health and Nutrition Examination Survey 2011–2018. J Cardiovasc Med. 2023;24(7):461–6.
23. Inoue K, Nianogo R, Telesca D, Goto A, Khachadourian V, Tsugawa Y, Sugiyama T, Mayeda ER, Ritz B: Low HbA1c levels and all-cause or cardiovascular mortality among people without diabetes: the US National Health and Nutrition Examination Survey 1999–2015. Int J Epidemiol. 2021;50(4):1373–83.
24. Li X, Zhao Y, Zhang D, Kuang L, Huang H, Chen W, Fu X, Wu Y, Li T, Zhang J, et al. Development of an interpretable machine learning model associated with heavy metals' exposure to identify coronary heart disease among US adults via SHAP: findings of the US NHANES from 2003 to 2018. Chemosphere. 2023;311:137039.
25. 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.
26. Wang Y, Ni B, Xiao Y, Lin Y, Jiang Y, Zhang Y. Application of machine learning algorithms to construct and validate a prediction model for coronary heart disease risk in patients with periodontitis: a population-based study. Front Cardiovasc Med. 2023;10:1296405.
27. Li W, Huang G, Tang N, Lu P, Jiang L, Lv J, Qin Y, Lin Y, Xu F, Lei D. Effects of heavy metal exposure on hypertension: A machine learning modeling approach. Chemosphere. 2023;337:139435. https://doi.org/10.1016/j.chemosphere.2023.139435.
28. Ward A, Sarraju A, Chung S, Li J, Harrington R, Heidenreich P, Palaniappan L, Scheinker D, Rodriguez F. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. NPJ Digit Med. 2020;3:125. https://doi.org/10.1038/s41746-020-00331-1.
29. Sopic M, Vilne B, Gerdts E, Trindade F, Uchida S, Khatib S, Wettinger SB, Devaux Y, Magni P. Multiomics tools for improved atherosclerotic cardiovascular disease management. Trends Mol Med. 2023;29:983–95. https://doi.org/10.1016/j.molmed.2023.09.004.
30. Farquhar WB, Edwards DG, Jurkovitz CT, Weintraub WS. Dietary sodium and health: more than just blood pressure. J Am Coll Cardiol. 2015;65:1042–50. https://doi.org/10.1016/j.jacc.2014.12.039.
31. Graudal N.A., Hubeck-Graudal T, Jurgens G. Effects of low sodium diet versus high sodium diet on blood pressure, Renin, aldosterone, catecholamines, cholesterol, and triglyceride. Cochrane Database Syst Rev. 2020;12:Cd004022. https://doi.org/10.1002/14651858.CD004022.pub5.

32. Selvaraj S, Djoussé L, Aguilar FG, Martinez EE, Polsinelli VB, Irvin MR, Arnett DK, Shah SJ. Association of estimated sodium intake with adverse cardiac structure and function: from the hypergen study. J Am Coll Cardiol. 2017;70:715–24. https://doi.org/10.1016/j.jacc.2017.06.036.
33. De Pergola G, D'Alessandro A. Influence of mediterranean diet on blood pressure. Nutrients. 2018;10. https://doi.org/10.3390/nu10111700.
34. Shan Z, Li Y, Baden MY, Bhupathiraju SN, Wang DD, Sun Q, Rexrode KM, Rimm EB, Qi L, Willett WC, Manson JE, Qi Q, Hu FB. Association between healthy eating patterns and risk of cardiovascular disease. JAMA Intern Med. 2020;180:1090–100. https://doi.org/10.1001/jamainternmed.2020.2176.
35. Aune D, Keum N, Giovannucci E, Fadnes LT, Boffetta P, Greenwood DC, Tonstad S, Vatten LJ, Riboli E, Norat T. Whole grain consumption and risk of cardiovascular disease, cancer, and all cause and cause specific mortality: systematic review and dose-response meta-analysis of prospective studies. BMJ. 2016;353:i2716. https://doi.org/10.1136/bmj.i2716.
36. Wu H, Flint AJ, Qi Q, van Dam RM, Sampson LA, Rimm EB, Holmes MD, Willett WC, Hu FB, Sun Q. Association between dietary whole grain intake and risk of mortality: two large prospective studies in US men and women. JAMA Intern Med. 2015;175:373–84. https://doi.org/10.1001/jamainternmed.2014.6283.
37. DeSalvo KB, Olson R, Casavale KO. Dietary guidelines for Americans. JAMA. 2016;315:457–8. https://doi.org/10.1001/jama.2015.18396.
38. Astrup A, Magkos F, Bier DM, Brenna JT, de Oliveira Otto MC, Hill JO, King JC, Mente A, Ordovas JM, Volek JS, Yusuf S, Krauss RM. Saturated fats and health: A reassessment and proposal for Food-Based recommendations: JACC State-of-the-Art review. J Am Coll Cardiol. 2020;76:844–57. https://doi.org/10.1016/j.jacc.2020.05.077.
39. de Oliveira Otto MC, Mozaffarian D, Kromhout D, Bertoni AG, Sibley CT, Jacobs DR Jr., Nettleton JA. Dietary intake of saturated fat by food source and incident cardiovascular disease: the Multi-Ethnic study of atherosclerosis. Am J Clin Nutr. 2012;96:397–404. https://doi.org/10.3945/ajcn.112.037770.
40. Lee PH, Chan CW. Energy intake, energy required and mortality in an older population. Public Health Nutr. 2016;19:3178–84. https://doi.org/10.1017/s1368980016001750.
41. Willcox BJ, Yano K, Chen R, Willcox DC, Rodriguez BL, Masaki KH, Donlon T, Tanaka B, Curb JD. How much should we eat? The association between energy intake and mortality in a 36-year follow-up study of Japanese-American men. J Gerontol Biol Sci Med Sci. 2004;59:789–95. https://doi.org/10.1093/gerona/59.8.b789.
42. Franco M, Orduñez P, Caballero B, Tapia Granados JA, Lazo M, Bernal JL, Guallar E, Cooper RS. Impact of energy intake, physical activity, and population-wide weight loss on cardiovascular disease and diabetes mortality in Cuba, 1980–2005. Am J Epidemiol. 2007;166:1374–80. https://doi.org/10.1093/aje/kwm226.
43. Kirzner RS, Robbins I, Privitello M, Miserandino M. Listen and learn:' participant input in program planning for a low-income urban population at cardiovascular risk. BMC Public Health. 2021;21:504. https://doi.org/10.1186/s12889-021-10423-6.
44. Shen R, Zhao N, Wang J, Guo P, Shen S, Liu D, Liu D, Zou T. Association between socioeconomic status and arteriosclerotic cardiovascular disease risk and cause-specific and all-cause mortality: data from the 2005–2018 National health and nutrition examination survey. Front Public Health. 2022;10:1017271. https://doi.org/10.3389/fpubh.2022.1017271.
45. Zhang YB, Chen C, Pan XF, Guo J, Li Y, Franco OH, Liu G, Pan A. Associations of healthy lifestyle and socioeconomic status with mortality and incident cardiovascular disease: two prospective cohort studies. BMJ. 2021;373:n604. https://doi.org/10.1136/bmj.n604.

Publisher's note