ISSN 2959-6157

Comparative Analysis of Machine Learning Models and Key Risk Factors in Advancing Predictive Analytics for Coronary Heart Disease Over a Decade

Jingyang Gao

East China Normal University
Email: 10215001422@stu.ecnu.edu.cn

Abstract:

This study aimed to identify key risk factors for coronary heart disease (CHD) and assess the performance of various machine learning models in predicting 10-year CHD risk. We conducted an exploratory data analysis using data from the Framingham Heart Study, which included 4,238 participants and 15 potential risk factors, to understand the distribution of variables and relationships. We used Chi-square and Mann-Whitney U tests to identify significant associations between risk factors and coronary heart disease. Logistic regression, random forest and support vector machine (SVM) models were established and their prediction accuracy and area under receiver operating characteristic curve (AUC) were evaluated. The results showed that age, systolic blood pressure and history of hypertension were the most influential risk factors. Logistic regression accuracyand AUCwere the highest, better than random forest and SVM. This indicates that we can pay more attention to such factors as age, systolic blood pressure and history of hypertension in subsequent CHD studies, and mainly use logistic regression model to predict coronary heart disease and optimize it.

Keywords: Coronary heart disease, Risk factors, Machine learning.

1. Introduction

Coronary heart disease(CHD) is the leading cause of death in the U.S., the U.K., and the world, killing more than 73,000 and 600,000 people in the U.S. and the U.K. each year [1, 2]. Timely diagnosis of heart disease is crucial in reducing health risk and preventing cardiac arrests. An American Heart Association study projects an almost 100% increase in CHD cases by 2030 [3]. CHD also puts a huge strain on national finances [4, 5].

Although CHD is the leading cause of death and disability, it is preventable [6]. The Framingham Heart Study enrolled its first participant in 1948 and is currently studying its third generation of participants. This is the first study to explore the relationship between risk factors and CHD [7]. According to the World Health Organization, risk factors of a specific disease are any attributes that raise the probability that a person may get that disease. Risk factors for coronary heart disease include smoking, high blood pressure, diabetes, and medications associated with these diseases [8]. Cardiovascular risk factors have largely similar effects on incidence of CHD and ischemic stroke in women, although the magnitude of association varies [9]. So we wondered if stroke is also associated with heart disease. Coronary heart disease is closely associated with unhealthy eating habits, many researchers to investigate the relationship between diet and heart disease risk. High fat diet is one of the key factors inducing inflammation. In addition, obesity caused by a high-fat diet is often accompanied by a series of chronic inflammatory diseases, such as type 2 diabetes, hypertension, and so on, eventually leading to coronary heart disease [10, 11]. Body mass index (BMI) can be calculated via mathematical operations where height and weight values are used to estimate the health status of a person [12].So there may also be a link between BMI and heart disease risk.

There are examinations that can help diagnose CHD, such as X-rays, MRI scans, and angiography. In the absence of medical equipment, diagnosis is difficult. But in cardiovascular disease, time is as important as every moment of diagnosis and treatment [13]. Therefore, early diagnosis of CHD is necessary in low-income populations and in low-resource areas. The search for an early diagnosis of CHD has been ongoing for many years, and many data analysis tools have been used to help healthcare practitioners identify some of the early signs of CHD [14]. Ma-

chine learning (ML) is a branch of artificial intelligence (AI) that is increasingly utilized within the field of cardiovascular medicine [15]. In the cataloguing of genetic cardiac illnesses and control subjects, a widespread set of ML algorithms with their variations is used to predict the early stages of heart failure [13] [16]. K-nearest neighbor(KNN), DT, SVC, LR, and RF machine algorithms are examples of heart attack prediction algorithms [13]. In this paper, we will use linear regression model to explore the significant influence of existing risk factors on CHD, and then divide the dataset into training set and test set to compare the prediction accuracy of two machine learning models, random forest and SVM.

2. Materials and Methods

2.1.1 *Materials*

In this paper, R is used for data preprocessing, EDA and correlation test operations, and Rstudio is used as the software. The model prediction part is completed in python, using the software Jupyter-notebook.

2.1.2 *Data source*

This article selects the data set from Kaggle public data sets on the site https://www.kaggle.com/datasets, the data set of data from an ongoing cardiovascular research, the research object is the inhabitants of the town of framingham, Massachusetts. It included more than 4,000 records and 15 possible risk factors associated with CHD.

Description of variables:

The dataset contained a total of 4238 records with 15 risk factors. There were eight continuous variables among the risk factors: age, number of cigarettes smoked per day(cigs per day), total cholesterol levels(Tot Chol), heart rate, systolic blood pressure(Sys BP), diastolic blood pressure(Dia BP), body mass index(BMI) and glucose level. Seven categorical variables: sex, education, current smoker, take blood pressure medication, have had a stroke, have had diabetes, and 10 year risk of coronary heart disease CHD. According to the nature of the variables can be divided into the following categories: Demographic: sex, age. Behavioral: current smoker, cigs per day. Medical: BP Meds, prevalent stroke, prevalent Hyp, diabetes. Medical: Tot Chol, Sys BP, Dia BP, BMI, heart rate, glucose. Since the education variable is not defined in the data set, this variable is deleted.

Exploratory Data Analysis (EDA) :

Exploratory data analysis (EDA) is the first and important step to conduct data analysis. EDA visualizes the structure of data through some simple data analysis, shows the distribution of data, understands the potential relationship of data, and provides suggestions for subsequent data processing and modeling. The data extracted from Kaggle is not pre-processed. By looking at the data set, we could find a total of 4238 records, in current smoker, cigs per day, BPMeds, Tot Chol, BMI, heart rate and glucose. Missing values would reduce the prediction ability, the following modeling and analysis process will be biased [17, 18]. To solve this problem, we carry out missing value processing, and decide to use the mean to fill the missing value of continuous variable and the mode to fill the missing value of categorical variable. After data preprocessing, the characteristics of the data set are as follows:

Table 1 Descriptive statistics of continuous variables

	Age	Cigsperday	TotChol	SysBP	DiaBP	BMI	Heartrate	Glucose
mean	49.58	9.00	236.72	132.35	82.89	25.80	75.87	81.97
std	8.57	11.88	44.33	22.04	11.91	4.07	12.02	22.84
min	22.83	0.00	107.00	83.50	83.50	15.54	44.00	44.00
max	70.00	70.00	70.00	295.00	142.50	56.80	143.00	394.00

Table 2 Category statistics of categorical variable

	male	currentsmoker	BPMeds	prevalentStroke	prevalentHyp	diabetes	TenYearCHD
0	2419	2144	4114	4213	2922	4129	3594
1	1819	2094	124	25	1316	109	644

2.1.3 *Data visualization*

The data set contains 14 risk factors that may be associated with heart disease, and exploring the correlation between risk factors can help us identify the factors with the highest risk and provide data reference for our subsequent modeling. In order to explore the correlation be-

tween risk factors and the correlation between risk factors and whether they have CHD, We calculate the correlation coefficients of eight continuous variables and draw heat maps. The correlation coefficient reflects the linear relationship between two variables and ranges from -1 to 1. 1 indicates a completely positive correlation, -1 indicates a completely negative correlation, and 0 indicates no linear correlation.

2.2 Method

2.2.1 *Chi-square test and Mann-Whitney U test*

Chi-square test was used for subtype variables and Mann-Whitney U test for continuity variables to further explore the correlation between risk factors and the risk of CHD. Chi-square test tests the degree of deviation between the actual observed value and the theoretical inferred value of the statistical sample. Under a certain confidence level and degree of freedom, Chi-square statistics and Chi-square distribution function are compared to judge the deviation between the actual probability and the expected probability, and then analyze the correlation between variables. The chi-square statistic is calculated by the following formula:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$$

The Mann-Whitney U test, also known as the Wilcoxon rank sum test, is a nonparametric test that does not require data to conform to a normal distribution, but in the data set we used, the correlation between continuous variables and categorical variables, that is, the ten-year risk of CHD. Our continuous variable distribution is unknown, so we choose Mann-Whitney U test, which does not require the normal distribution of data and can compare the distribution difference between two independent samples. Mann-Whitney U test first sorted all the data, then calculated the rank sum of each data, then calculated the U statistic through the rank sum, and finally calculated the significance of the test results according to the U statistic and sample size. Suppose the sizes of samples 1 and 2 are n_1 and n_2, respectively, and their rank sum is R_1 and R_2, respectively. Then the formula for calculating the U statistic is:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$U = min(U_1, U_2)$$

data type conversion:

2.2.2 *Factor transformation*

Factor conversion refers to the classification variables into factors form, convenient for the machine learning model. Our categorical variables are actually non-numerical data, such as male and female in gender, but subsequent modeling and machine learning algorithms require the input to be in numerical format, so factor conversion is required. Factor transformations label each category of a categorical variable as an integer label, such as males and females in a gender variable converted to a factor pattern where males can be labeled as 0 and females as 1.

2.2.3 *One-Hot Encoding*

One-hot Encoding is to further process the classified variable after factor conversion into a binary format acceptable to the machine learning algorithm.One-Hot Encoding first requires that categorical values be mapped to integer values, and then each integer value is represented as a binary vector, for example gender has two categories: male and female. After One-Hot Encoding, two new binary variables are generated: male and female, with the male variable taking the value 1 if it is male and 0 otherwise. The value of female variable is 1 if it is female, otherwise it is 0. In order to avoid multicollinearity problems, we usually delete the first dummy variable, such as the female variable in gender.

2.3 Prediction models:

2.3.1 *Logistic Regression*

Logistic regression is an important classification algorithm in the field of machine learning, which is often used to solve binary classification problems. Logistic regression is actually a probability estimation model that is used to predict the likelihood of an event occurring. It maps vectors to probability values in the interval (0,1) through the Sigmoid function, and judges positive and negative categories with a threshold value of 0.5. Model of the formula can be represented as:

$$\hat{p} = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8+\beta_9x_9+\beta_{10}x_{10}+\beta_{11}x_{11}+\beta_{12}x_{12}+\beta_{13}x_{13}+\beta_{14}x_{14})}}$$

$\hat{p}$ is belong to one type of probability sample, in this paper is the ten - year risk of CHD, β_0 means the intercept term in the regression, β : Partial regression coefficients of independent variables.

x_1 : Age. x_2 : Cigsperday. x_3 : TotChol. x_4 : SysBP. x_5 : DiaBP. x_6 : BMI. x_7 : Heartrate. x_8 : Glucose. x_9 : Male. x_{10}

: Currentsmoker. x_{11} : BPMeds. x_{12} : Prevalentstroke. x_{13} : Prevalenthyp. x_{14} : Diabetes.

2.3.2 *Random Forest*

Random forest is a parallel integrated algorithm composed of decision trees, which belongs to the Bagging type. Its core idea is that when the training data is fed into the model, instead of building one large decision tree with the entire training data set, the random forest builds multiple smaller decision trees with different subsets and feature attributes, and then combines them into a more powerful model. The prediction formula is as follows:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

$\hat{p}$: is belong to one type of probability sample, in this paper is the ten - year risk of CHD. N: The number of decision trees. $T_i(x)$: The first i decision tree to sample x predicted result of x (0 or 1). x : The input feature vector contains all the feature variables. According to the prediction probability $\hat{p}$, the final classification result is determined by setting a threshold of 0.5.

2.3.3 *Support Vector Machine(SVM)*

SVM is a machine learning algorithm for classification and regression analysis that efficiently handles both linearly separable and linearly indistinguishable data and constructs optimal decision boundaries in high-dimensional spaces. The training process of SVM is a convex optimisation problem with the objective of minimising the structural risk of the model. During the solution process, the SVM focuses only on those samples that lie near the decision boundary, which are called support vectors. This property makes the SVM robust and generalisable. The decision function can be expressed as:

$$f(x) = sign(w \bullet x + b)$$

$f(x)$: The predicted result of the Input sample x . w : The normal vector of the decision hyperplane. x : The input feature vector contains all the feature variables. b :

The bias term, representing the intercept of the model. For non-linearly separable data, Support Vector Machines (SVM) use a kernel function to map the data into a higher-dimensional space where it becomes linearly separable. The kernel function is represented as:

$$K(x_i, x_j) = \phi(x_i) \bullet \phi(x_j)$$

$\phi(x_i)$: The mapping function that maps the input features into a higher-dimensional space.

2.4 Assessment indicators

2.4.1 *Receiver Operating Characteristic Curve(ROC) and Area Under the Curve(AUC)*

ROC is a curve drawn on a two-dimensional plane with false positive rate(FPR) and true positive rate(TPR). For a classifier, we can get a pair of TPR and FPR points based on its performance on the test sample. In this way, the classifier can be mapped to a point on the ROC plane. True positive rate (TPR) is the proportion of all samples that are actually positive that are correctly classified as positive. The false positive rate (FPR) is the proportion of all samples that are actually negative that are misclassified as positive.

$$TPR = TP / (TP + FN)$$

$$FPR = FP / (FP + TN)$$

AUC is one of the important indexes to evaluate the performance of binary classification model. The AUC value is the area under the ROC curve that represents the overall classification performance of the model. The larger the value of AUC, the better the classification performance of the model. When the AUC value is 0.5, the model behaves the same as the random guess. When the AUC value is less than 0.5, the model performs worse than random guesses.

2.4.2 *Accuracy*

The formula used for calculating the accuracy of a machine-learning model is given below:

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$

TN: True Negative. FN: False Negative. TP: True Positive. FP : False Positive.

3. Results

3.1 Correlation research results

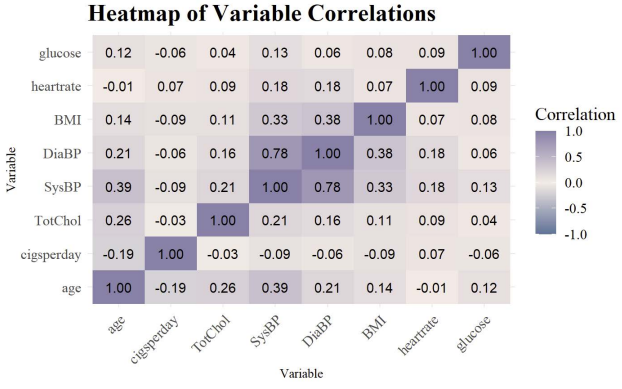

Figure 1 Heatmap of Variable Correlations

Heat maps show correlations between multiple continuous variables in a data set. The results are as follows: The

correlation coefficient between age and daily smoking volume is -0.2, indicating a weak negative correlation between age and daily smoking volume, and daily smoking volume may decrease slightly with increasing age. The correlation coefficient between age and total cholesterol (TotChol) was 0.3, indicating a moderately positive relationship between age and total cholesterol, with total cholesterol levels likely to rise with increasing age. The correlation coefficient between age and systolic blood pressure (SysBP) is 0.4, indicating a moderate positive relationship between age and systolic blood pressure, with systolic blood pressure likely to increase with age. The correlation coefficient between daily smoking and systolic blood pressure (sysBP) was -0.1, indicating that there is little linear relationship between daily smoking and systolic blood pressure. The correlation coefficient between daily smoking and diastolic blood pressure (DiaBP) is -0.1, indicating that there is little linear relationship between daily smoking and diastolic blood pressure. The correlation coefficient between total cholesterol and systolic blood pressure (SysBP) was 0.2, indicating a weak positive correlation between total cholesterol and systolic blood pressure. The correlation coefficient between total cholesterol and diastolic blood pressure (DiaBP) was 0.2, indicating a weak positive correlation between total cholesterol and diastolic blood pressure. The correlation coefficient between systolic and diastolic blood pressure (DiaBP) was 0.8, indicating a strong positive correlation between systolic and diastolic blood pressure. The correlation coefficient between systolic blood pressure and body mass index (BMI) was 0.3, indicating a weak positive correlation between systolic blood pressure and BMI. The correlation coefficient between diastolic blood pressure and body mass index (BMI) was 0.4, indicating a moderate positive correlation between diastolic blood pressure and BMI. The correlation coefficient between HeartRate and other variables is low, indicating that there is almost no correlation between heart rate and other variables.

3.2 Correlation test result

According to the classification of the chi-square test and continuous variables the Mann - Whitney U test, it is concluded that the following 14 independent variables with Ten years of coronary heart disease Risk (TenYearCHD Risk) the correlation between the test results. The p value of age (1.17e-47) showed a significant correlation. The p value of daily smoking amount (cigsperday) was 4.71e-03, which showed significant correlation. The P-value of total cholesterol (TotChol) was 2.44e-07, showing a significant correlation. The p value of systolic blood pressure (SysBP) was 9.82e-37, and there was a significant correlation. The p value of diastolic blood pressure (DiaBP) was 1.00e-17, and there was a significant correlation. The P-value of body mass index (BMI) was 1.77e-06, and there was a significant correlation. The P-value of heartrate (heartrate) was 2.40e-01, with no significant correlation. The p value of glucose was 7.49e-04, and there was a significant correlation. The p value of male was 1.11e-08, indicating a significant correlation. The p value of currentsmoker was 2.21e-01, and there was no significant correlation. The p value of whether to take antihypertensive drugs (BPMeds) was 3.16e-08, and there was significant correlation. The p value of prevalentStroke was 1.81e-04, and there was a significant correlation. History of hypertension (prevalentHyp) had a p value of 1.09e-30, showing a significant correlation.

3.3 Prediction model results

With three different prediction models: Random Forest, support vector machine (SVM), and Logistic Regression, we get the following results. The accuracy of the random forest model is 83.65%, and the AUC is 0.68. The accuracy of the SVM model is also 83.65%, but the AUC is only 0.59. In contrast, the logistic regression model performed best, with an accuracy of 84.43% and an AUC of 0.72.

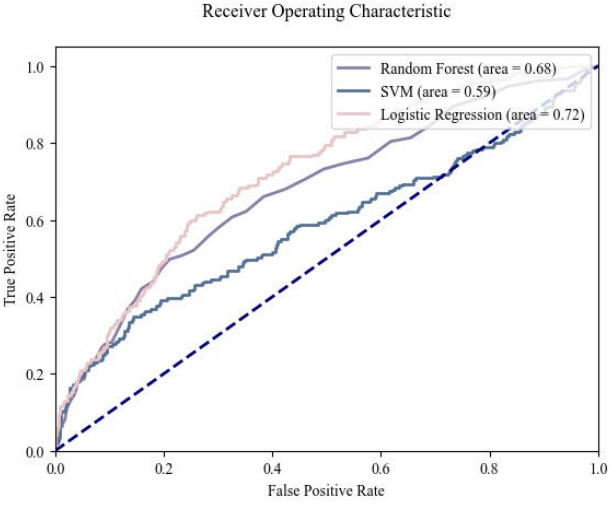

Figure 2 ROC of three models

4. Discussion

4.1 Discussion of risk factors

According to the heat map above, we can observe the three significant correlation relationships: between SysBP and DiaBP very strong positive correlation, the correlation coefficient of 0.8: blood pressure is in the process of blood flow in blood vessels of the lateral pressure of blood vessel walls, it depends on the interaction of thrust and drag blood flow. When the heart contracts, the highest pressure produced by pumping blood into the arteries is called systolic blood pressure (high pressure), and when

the heart diastates, the blood pressure inside the artery drops to its lowest value, called diastolic blood pressure (low pressure). The difference between them is called pulse pressure, and too high or too low pulse pressure has certain health risks, too high pulse pressure may lead to heart attack, coronary heart disease, and too low pulse pressure may be due to obesity, alcoholism and other factors [19]. As two indicators that are highly correlated with heart health, there must be a strong positive correlation between them.There is a moderately strong correlation between Age and SysBP, with a correlation coefficient of 0.4: Some current studies have shown that systolic blood pressure is highly correlated with age, and the existing literature shows that after 30-40 years old, systolic blood pressure increases linearly with age and reaches a peak in the later years [20, 21]. However, there is also some literature showing that in Mexico and China, hypertension has been trending younger in recent years [22, 23]. There is a medium strong correlation between Age and TotChol, with a correlation coefficient of 0.3: with the increase of age, human metabolism slows down, which leads to the weakening of cholesterol metabolism and the rise of cholesterol in blood pressure. It has also been shown that postmenopausal women have higher cholesterol levels than premenopausal women [24].

In the hypothesis test results of the dependent variable and the independent variable, we can see that there are three variables with strong correlation: the age variable has the strongest correlation with CHD within ten years, and the p-value is 1.17e-47: Age has always been an important factor in CHD, which is associated with an increased likelihood of developing any other cardiac risk factor [25]. Some studies have shown that older people are at high risk for heart disease, and the associated risk of cardiovascular disease increases with age in both men and women, which corresponds to an overall decline in sex hormones [26]. The elderly population should be the focus of prevention and treatment of CHD, we should pay more attention to the physical indicators of the elderly, and effectively reduce the incidence of coronary heart disease. Systolic blood pressure (SysBP) is the second highest ranked variable in terms of correlation with Ten-Year CHD Risk, and the p-value is 9.82e-37:Many studies have shown that aggressive antihypertensive therapy reduces the incidence of CHD [27], and the Systolic Blood Pressure Intervention Trial (SPRINT) study demonstrated a significant benefit in reducing cardiovascular events in older hypertensive patients with lower systolic blood pressure targets (< 120 mmHg) compared with a higher systolic target (<140 mmHg) for elderly hypertensive patients had a significant benefit in reducing cardiovascular events [28, 29]. This study further validates the correlation study in this article that systolic blood pressure is significantly associated with CHD risk. Appropriate lowering of systolic blood pressure may reduce the risk of CHD, but it is also important to be aware of the problem of hypotension caused by excessive lowering of blood pressure. There was also an extremely significant association between a history of hypertension and the ten-year risk of coronary heart disease, and the p-value is 1.09e-30: Although hypertension does not cause cancer like smoking, it is the strongest risk factor or one of the strongest risk factors for virtually different cardiovascular diseases, including coronary heart disease, left ventricular hypertrophy and valvular heart disease, and arrhythmias (including atrial fibrillation) [30]. Hypertension has multiple damages to blood vessels and heart, so the history of hypertension is significantly correlated with Ten-Year CHD Risk. Therefore, the control of hypertension is of great significance for the prevention of coronary heart disease.

4.2 Discussion of prediction model

According to our experimental results, the prediction accuracy and AUC performance of Logistic Regression are the best, indicating that it can better capture the linear relationship between variables when dealing with this binary classification problem, and has good performance, which is helpful for us to understand the risk factors of CHD risk. Although the random forest model can handle complex nonlinear relationships and interaction effects, it is not as effective as Logistic Regression in this data set. In the correlation hypothesis test, there were two factors that were not associated with Ten-Year CHD Risk, heartrate and currentsmoker. If some features have a strong correlation with the dependent variable, and some have a small correlation with the dependent variable, the random forest model may rely too much on certain features, which affects the overall performance of the model. However, the AUC of SVM model is only 0.59, close to the probability of random guess, which may be due to the sensitivity of SVM to parameter and data preprocessing, and the need to select appropriate kernel functions and parameters when processing high-dimensional data. In this data set, SVM fails to show its advantages. In this dataset, all the models are not up to the ideal level. We need to further optimize the model parameters and fit the model with more data in the following work. Moreover, further optimization is needed in data preprocessing to improve the classification effect. In addition, for example, remove some variables with poor correlation. we can also try other ensemble learning algorithms or deep learning algorithms to capture complex patterns and relationships in the data and improve predictive performance.

5. Conclusion

Age, systolic blood pressure, a history of high blood pressure is very important key risk factors for coronary heart disease (CHD). Logistic regression is superior to other models in predicting the 10-year risk of coronary heart disease, and careful model selection and data preprocessing are required in predictive health analysis. These findings help to develop targeted prevention and intervention strategies to reduce the burden of coronary heart disease.

References

1. Yang H, Garibaldi JM: A hybrid model for automatic identification of risk factors for heart disease. *J Biomed Inform* 2015, 58 Suppl(Suppl):S171-S182.

2. Murphy SL, Xu J, Kochanek KD: Deaths: final data for 2010. *Natl Vital Stat Rep* 2013, 61(4):1-117.

3. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR *et al*: Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation* 2019, 139(10):e56-e528.

4. Cui H, Liu Q, Wu Y, Cao L: Cumulative triglyceride-glucose index is a risk for CVD: a prospective cohort study. *Cardiovasc Diabetol* 2022, 21(1):22.

5. Zhao D, Liu J, Wang M, Zhang X, Zhou M: Epidemiology of cardiovascular disease in China: current features and implications. *Nat Rev Cardiol* 2019, 16(4):203-212.

6. Brown JC, Gerhardt TE, Kwon E: Risk Factors for Coronary Artery Disease. In: *StatPearls*. edn. Treasure Island (FL) ineligible companies. Disclosure: Thomas Gerhardt declares no relevant financial relationships with ineligible companies. Disclosure: Edward Kwon declares no relevant financial relationships with ineligible companies.; 2024.

7. Hajar R: Risk Factors for Coronary Artery Disease: Historical Perspectives. *Heart Views* 2017, 18(3):109-114.

8. Houssein EH, Mohamed RE, Ali AA: Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. *Sci Rep* 2023, 13(1):7173.

9. Leening MJG, Cook NR, Franco OH, Manson JE, Lakshminarayan K, LaMonte MJ, Leira EC, Robinson JG, Ridker PM, Paynter NP: Comparison of Cardiovascular Risk Factors for Coronary Heart Disease and Stroke Type in Women. *J Am Heart Assoc* 2018, 7(19):e007514.

10. Sletten AC, Peterson LR, Schaffer JE: Manifestations and mechanisms of myocardial lipotoxicity in obesity. *J Intern Med* 2018, 284(5):478-491.

11. Machate DJ, Figueiredo PS, Marcelino G, Guimaraes RCA, Hiane PA, Bogo D, Pinheiro VAZ, Oliveira LCS, Pott A: Fatty Acid Diets: Regulation of Gut Microbiota Composition and Obesity and Its Related Metabolic Dysbiosis. *Int J Mol Sci* 2020, 21(11).

12. Oniszczenko W, Stanislawiak E: Association between sex and body mass index as mediated by temperament in a nonclinical adult sample. *Eat Weight Disord* 2019, 24(2):291-298.

13. Hassan CAU, Iqbal J, Irfan R, Hussain S, Algarni AD, Bukhari SSH, Alturki N, Ullah SS: Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. *Sensors (Basel)* 2022, 22(19).

14. Narasimhan G, Victor A: Analysis of computational intelligence approaches for predicting disease severity in humans: Challenges and research guidelines. *J Educ Health Promot* 2023, 12:334.

15. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, Zhang H, Kaplin S, Narasimhan B, Kitai T *et al*: Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep* 2020, 10(1):16057.

16. Juhola M, Joutsijoki H, Penttinen K, Shah D, Polonen RP, Aalto-Setala K: Data analytics for cardiac diseases. *Comput Biol Med* 2022, 142:105218.

17. Rasmy L, Nigo M, Kannadath BS, Xie Z, Mao B, Patel K, Zhou Y, Zhang W, Ross A, Xu H *et al*: Recurrent neural network models (CovRNN) for predicting outcomes of patients with COVID-19 on admission to hospital: model development and validation using electronic health record data. *Lancet Digit Health* 2022, 4(6):e415-e425.

18. Liu M, Li S, Yuan H, Ong MEH, Ning Y, Xie F, Saffari SE, Shang Y, Volovici V, Chakraborty B *et al*: Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artif Intell Med* 2023, 142:102587.

19. Dai N, Deng Y, Wang B: Association between human blood metabolome and the risk of hypertension. *BMC Genom Data* 2023, 24(1):79.

20. Satoh M, Metoki H, Asayama K, Murakami T, Inoue R, Tsubota-Utsugi M, Matsuda A, Hirose T, Hara A, Obara T *et al*: Age-Related Trends in Home Blood Pressure, Home Pulse Rate, and Day-to-Day Blood Pressure and Pulse Rate Variability Based on Longitudinal Cohort Data: The Ohasama Study. *J Am Heart Assoc* 2019, 8(15):e012121.

21. Ntineri A, Stergiou GS, Thijs L, Asayama K, Boggia J, Boubouchairopoulou N, Hozawa A, Imai Y, Johansson JK, Jula AM *et al*: Relationship between office and home blood pressure with increasing age: The International Database of HOme blood pressure in relation to Cardiovascular Outcome (IDHOCO). *Hypertens Res* 2016, 39(8):612-617.

22. Castro-Porras LV, Rojas-Martinez R, Aguilar-Salinas CA, Bello-Chavolla OY, Becerril-Gutierrez C, Escamilla-Nunez C: Trends and age-period-cohort effects on hypertension mortality rates from 1998 to 2018 in Mexico. *Sci Rep* 2021, 11(1):17553.

23. Hosseini M, Yousefifard M, Baikpour M, Rafei A, Fayaz M, Heshmat R, Koohpayehzadeh J, Asgari F, Etemad K, Gouya MM *et al*: Twenty-year dynamics of hypertension in Iranian adults: age, period, and cohort analysis. *J Am Soc Hypertens*

2015, 9(12):925-934.
24. Song DK, Hong YS, Sung YA, Lee H: The effect of menopause on cardiovascular risk factors according to body mass index in middle-aged Korean women. *PLoS One* 2023, 18(3):e0283393.
25. Global Burden of Metabolic Risk Factors for Chronic Diseases C, Lu Y, Hajifathalian K, Ezzati M, Woodward M, Rimm EB, Danaei G: Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: a pooled analysis of 97 prospective cohorts with 1.8 million participants. *Lancet* 2014, 383(9921):970-983.
26. Rodgers JL, Jones J, Bolleddu SI, Vanthenapalli S, Rodgers LE, Shah K, Karia K, Panguluri SK: Cardiovascular Risks Associated with Gender and Aging. *J Cardiovasc Dev Dis* 2019, 6(2).
27. Lee DH, Lee JH, Kim SY, Lee HY, Choi JY, Hong Y, Park SK, Ryu DR, Yang DH, Hwang SJ *et al*: Optimal blood pressure target in the elderly: rationale and design of the HOW to Optimize eLDerly systolic Blood Pressure (HOWOLD-BP) trial. *Korean J Intern Med* 2022, 37(5):1070-1081.
28. Williamson JD, Supiano MA, Applegate WB, Berlowitz DR, Campbell RC, Chertow GM, Fine LJ, Haley WE, Hawfield AT, Ix JH *et al*: Intensive vs Standard Blood Pressure Control and Cardiovascular Disease Outcomes in Adults Aged >/=75 Years: A Randomized Clinical Trial. *JAMA* 2016, 315(24):2673-2682.
29. Group SR, Lewis CE, Fine LJ, Beddhu S, Cheung AK, Cushman WC, Cutler JA, Evans GW, Johnson KC, Kitzman DW *et al*: Final Report of a Trial of Intensive versus Standard Blood-Pressure Control. *N Engl J Med* 2021, 384(20):1921-1930.
30. Kjeldsen SE: Hypertension and cardiovascular risk: General aspects. *Pharmacol Res* 2018, 129:95-99.