Dean&Francis

ISSN 2959-6157

Prediction and Analysis of Influencing Factors of Heart Disease

Hanwei Cao[1, *], **Chang Liu**[2], **Bowen Wu**[3] and **Yilin Wu**[4]

[1]Department of Statistics, Columbia University, New York, NY, 10027, United States
[2]School of Nursing and Health, Zhengzhou University, Zhengzhou, 450001, China
[3]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China
[4]Yilin Wu, College of Basic Medical Sciences, Jilin University, Changchun, 130021, China

*Corresponding author: Hc3356@Columbia.edu

Abstract:

Cardiovascular disease, the leading cause of morbidity and mortality, affects more than 523 million people worldwide. Cardiovascular disease characterized by an insidious onset, long latency period, and complex, variable presentation, poses significant challenges to recovery post-onset. Patients often suffer from irreversible organ damage due to delayed treatment, which not only severely threatens their lives but also exacerbates medical and economic burdens. In this research, the methods of logistic regression and random forest were used to identify the influencing factors. The dataset includes 303 clinical cases and 14 variables such as chest pain type, fasting blood sugar level, serum cholesterol, and so on. The outcomes demonstrate that the overall prediction accuracy of the research model is 85.15%. Thus, this model will assist primary care personnel to recognize heart disease more easily, accurately, and earlier, and to provide a scientific basis for clinical prevention and treatment, reducing the medical and economic burden on patients.

Keywords: Heart disease; influencing factor; logistic regression; random forest.

1. Introduction

Cardiovascular disease (CVD) is a major public health issue that poses a threat to human life and health [1]. Cardiovascular disease, the leading cause of morbidity and mortality, affects more than 523 million people worldwide [2]. Due to the insidious onset of CVD, long latency period, and complex and variable condition, it is difficult to recover after the onset of CVD. Patients with CVD often face serious threats to their lives and aggravate their medical and economic burdens due to delayed treatment and irreversible organ damage. The latest statistics related to heart disease, stroke, and cardiovascular risk factors are reported annually by both the AHA and the National Institutes of Health. The American Heart Association (AHA), in conjunction with the National Institutes of Health, reports annually on the most recent statistics related to heart disease, stroke, and cardiovascular risk factors, including core health behaviors (smoking, exercise, sleep, and obesity) and health factors contributing to cardiovascular health (cholesterol, blood pressure, and metabolic syndrome) [3]. To better understand these disparities,

it is crucial to examine the various factors contributing to the risk of coronary heart disease and its progression.

Dyslipidemia is one of the risk factors of coronary heart disease, which can cause lipoprotein deposition in the endothelium of coronary arteries, thereby damaging the vascular endothelial cells, increasing the permeability of the cell membrane, impairing the function of the receptors, accelerating the progression of coronary artery disease, and also affecting the prognosis of percutaneous coronary intervention [4]. Coronary atherosclerotic heart disease (CHD) is most common in the elderly and can be associated with a variety of serious complications in the later stages of the disease, such as heart failure and cardiac arrhythmia [5]. Dyslipidemia has an insidious onset and affects a wide range of ages, so the clinic must take active measures to intervene and prevent it. In previous studies, the age, gender, body mass index (BMI), heart rate, alcohol consumption status, drinking status, and exercise of control group and observation group were compared. In multivariate logistic regression analysis, risk factors for chest pain type and dyslipidemia were identified by including the variables that had significance in univariate analysis [6].

The previous study indicated that abnormal blood sugar is also one of the risk factors for heart disease [7]. Over the past 4 decades, the United States has seen a significant increase in the prevalence of diabetes, with over 34.2 million Americans suffering from the condition. Type 2 diabetes (T2D) is characterized by over 90% to 95% of the diabetic population [8]. Patients with type 2 diabetes mellitus (T2DM) have a 2-4 times higher risk for cardiovascular disease compared to the general population. Coronary atherosclerosis in patients with T2 DM is characterized by severe and diffuse atherosclerosis, with a large atherosclerotic tissue area and a higher percentage of macrophage infiltration, which clinically reveals severe stenosis of the coronary arteries and the generation of more thrombi [9]. In many studies, researchers have used correlation analysis and logistic multivariate regression analysis to explore the relationship between diabetes and CVD and concluded that the duration of diabetes, BMI, hypertension and HbA1c are risk factors for CVD.

Besides examining the risk factors of heart disease, many previous studies have also emphasized the protective factors of heart disease. The occurrence and development of metabolic diseases can be prevented by exercise, which is a key regulator of metabolism [10]. One of the main mechanisms is to promote positive remodeling of skeletal muscle by stimulating various exercise induced pathways, including enhancing lipid oxidation and reducing intramuscular lipid content, thereby promoting metabolic improvement related to physical exercise [11]. Reasonable and scientific exercise can reduce the incidence rate and mortality of CVD, reduce the number of repeated hospitalizations, improve exercise endurance and quality of life of patients, and reasonably control medical costs by controlling metabolic risk factors [12]. Ping showed that women have a lower incidence of heart disease than men. This is because estrogen in women's blood circulation can cause vasodilation, enhance the amount of elastic fibers in the arterial wall, and decrease the amount of collagen fibers, resulting in lower arterial stiffness and better arterial elasticity compared to men [13].

In summary, this paper will investigate the epidemiological characteristics of heart disease in adults and analyze the related influencing factors. Logistic regression and Random Forest (RF) are used to construct a prediction model, to assist primary care personnel to recognize CVD more easily, accurately, and earlier, and to provide a scientific basis for clinical prevention and treatment.

2. Methods

2.1 Data Source and Description

The research in this paper is based on data from an extensive health assessment carried out by China's National Health Commission, which included a broad and varied cross-section of adults. The assessment covered a range of topics such as personal details, past health issues, daily routines, and physical examinations, providing a complete picture of heart and blood vessel health. All participants gave their consent, and the project was given the green light by the appropriate ethics committees, following the ethical standards set out in the Declaration of Helsinki.

2.2 Indicator Selection and Explanation

To assess the risk factors and epidemiological characteristics of cardiovascular diseases (CVD), this paper selected a set of indicators based on the American Heart Association's (AHA) guidelines and the findings from the literature review presented in the introduction (Table 1).

ISSN 2959-6157

Table 1. Indicator explanation

Variable	Symbol	Range
Age	x1	29 to 77
Sex	x2	0 = male, 1 = female
Chest pain type	x3	0: Typical angina, 1: Atypical angina, 2: Non-anginal pain, 3: Asymptomatic
Resting blood pressure	x4	94 to 200 mmHg
Serum cholesterol	x5	126 to 564 mg/dl
Fasting blood sugar level	x6	categorized as above 120 mg/dl (1 = true, 0 = false)
Resting electrocardiographic results	x7	0: Normal, 1: Having ST-T wave abnormality, 2: Showing probable or definite left ventricular hypertrophy
Maximum heart rate	x8	71 to 202
Exercise-induced angina	x9	1 = yes, 0 = no
ST depression	x10	0 to 6.2
Slope of the peak exercise ST segment	x11	0: Upsloping, 1: Flat, 2: Downsloping
Number of major vessels colored by fluoroscopy	x12	0-4
Thalium stress test result	x13	0: Normal, 1: Fixed defect, 2: Reversible defect, 3: Not described

2.3 Method Introduction

This paper utilized a dual-method approach in the study, combining logistic regression to ascertain the risk factors' odds ratios for CVD and Random Forest (RF) for its proficiency in managing intricate data relationships. The logistic regression laid the groundwork for the understanding, while RF, known for its aptness in handling non-linear dynamics, was engaged to build a predictive model classifying CVD risk. The model's efficacy was gauged through accuracy, sensitivity, specificity, and the ROC curve, fortified by cross-validation for reliability. Ultimately, the goal was to engineer a model that helps health-care providers in preemptively pinpointing individuals prone to CVD, thereby enabling early action and a possible curtailment of the disease's impact.

3. Results and Discussion

3.1 Descriptive Analysis

The Figure 1 is the histogram of the age. There is a peak in the distribution around the 60s, but it is generally even.

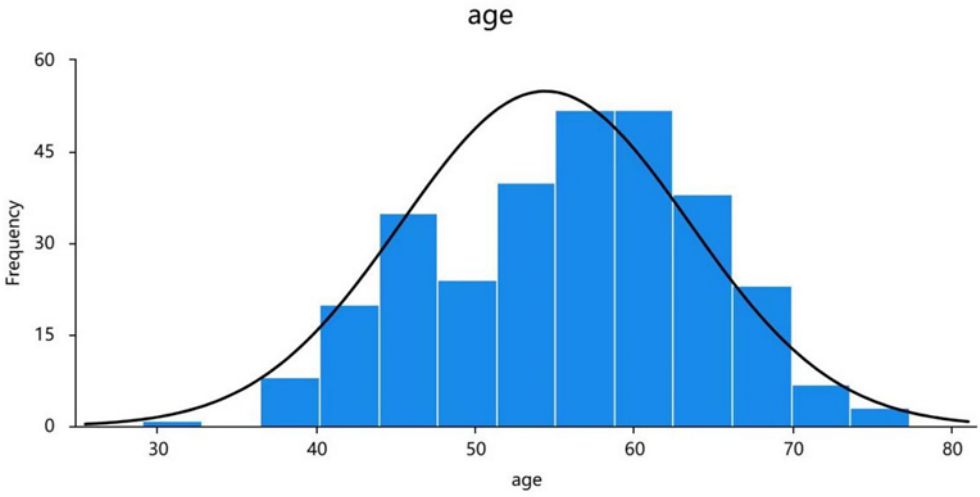

Fig. 1 The histogram of Age

The Figure 2 is the box plot of Resting Blood Pressure (trestbps). Most people's resting blood pressure ranges

from 120 to 140 mm Hg. Individuals with and without diseases have similar median resting blood pressure levels.

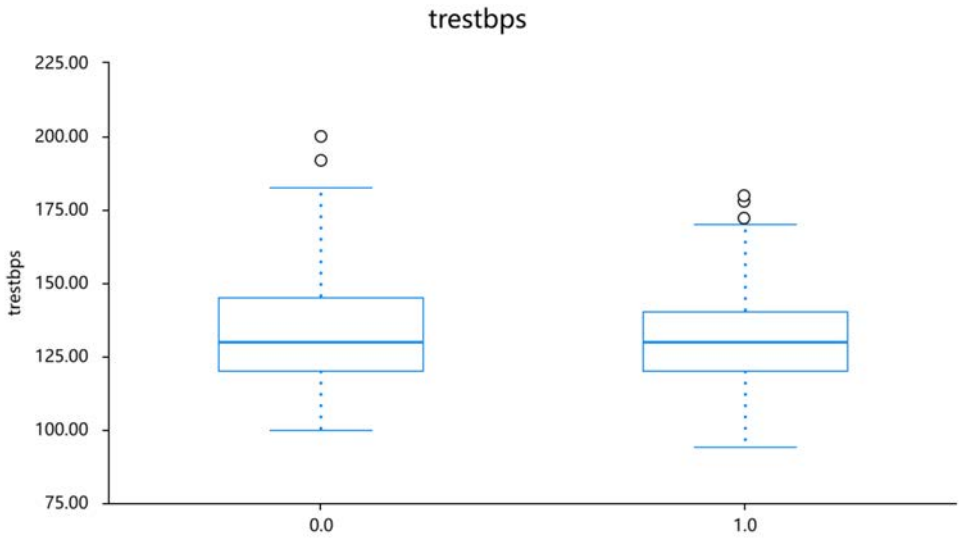

Fig. 2 The box plot of Resting Blood Pressure

The Figure 3 is the box plot of Serum Cholesterol (chol): The normal cholesterol is 0-200 mg/dl, but most individuals have the cholesterol levels between 200 and 300 mg/dl.

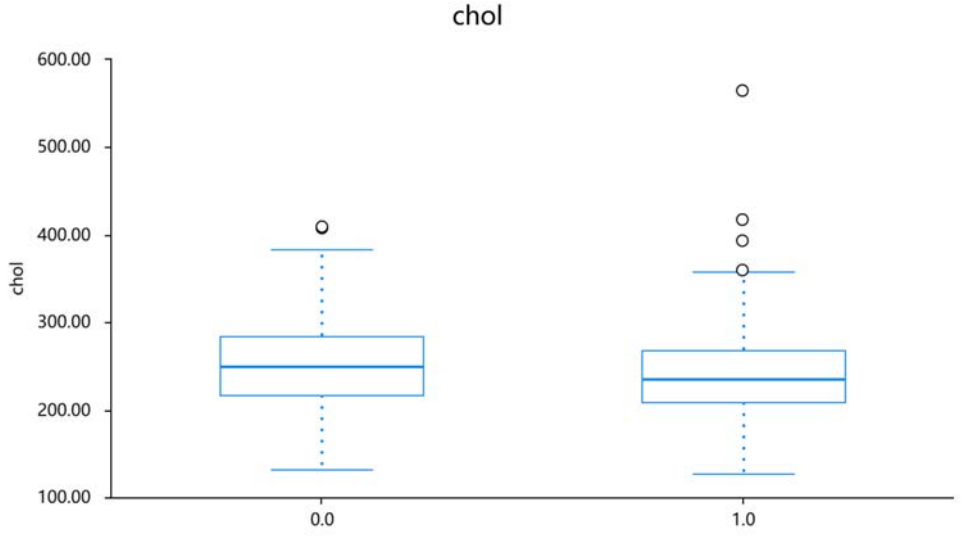

Fig. 3 The box plot of Serum Cholesterol

The Figure 4 is the violin plot of Maximum Heart Rate Achieved (thalach): The majority of people achieve a heart rate of 140 to 170 bpm. The normal maximum heart rate is 100-150, but the most patient reaches to 190bpm.

ISSN 2959-6157

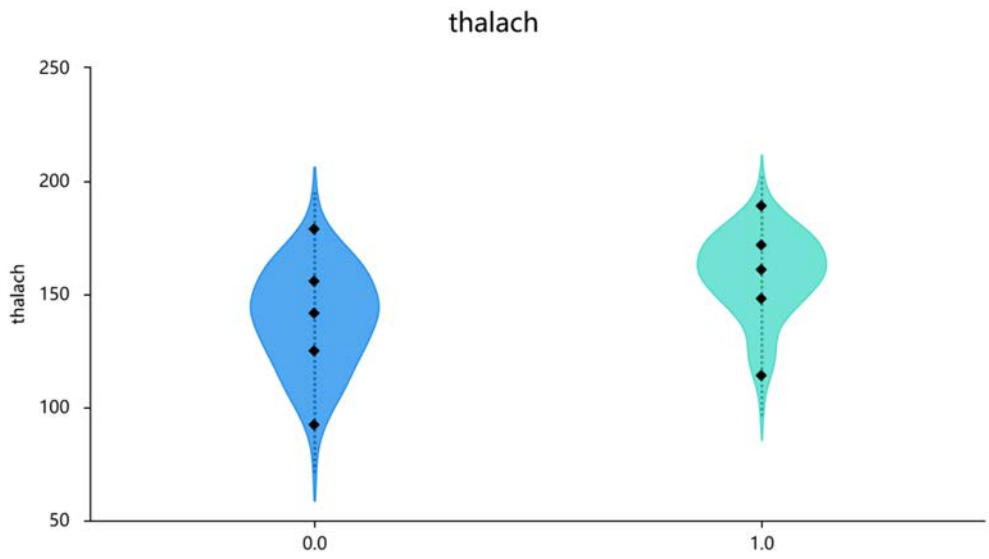

Fig. 4 The violin plot of Maximum Heart Rate Achieved

The Figure 5 is the histogram of ST Depression Induced by Exercise (oldpeak): The bulk of results were grouped at zero, showing that many people did not experience substantial ST depression when exercising.

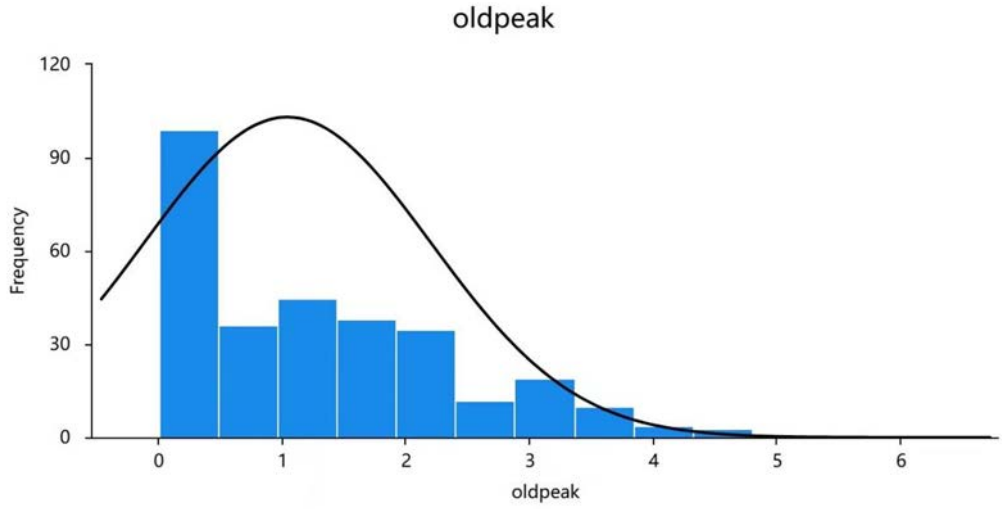

Fig. 5 The histogram of ST Depression induced by Exercise

3.2 Spearman Correlation Analysis

Correlation analysis was used to investigate the correlation between target and x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12 and x13, a total of 13 items, and the strength of the association was shown using Spearman's correlation coefficient (Figure 6). Target exhibits a strong negative association with x1, x2, x4, x5, x9, x10, x12 and x13. A noteworthy positive association has been observed between the target and x3, x7, x8 and x11. There is no correlation between target and x6.

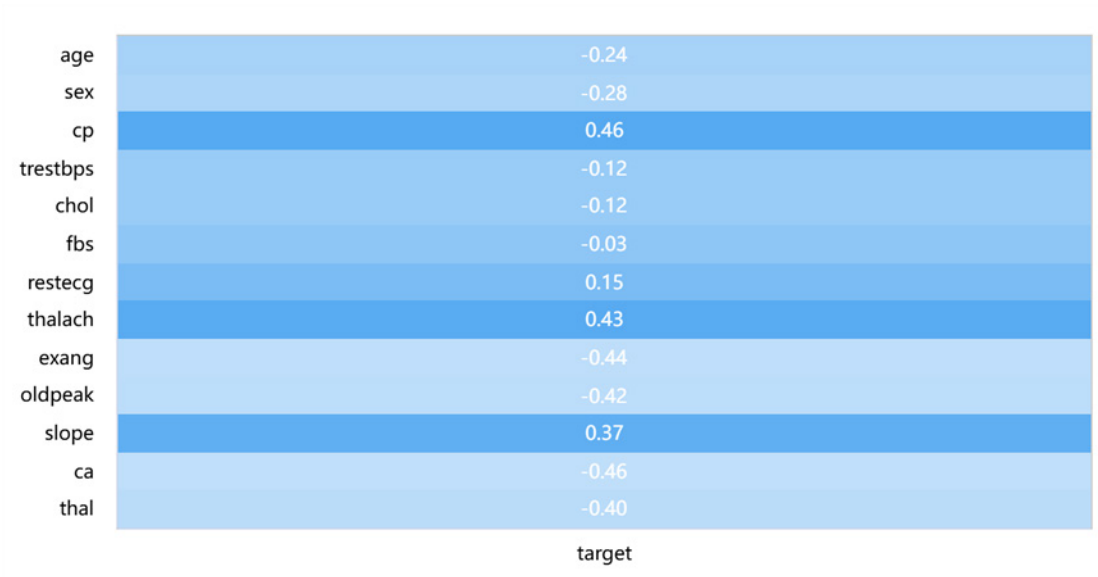

Fig. 6 Spearman correlation visualization

It can be seen that a total of 12 items, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12 and x13, were correlated with x1 using correlation analysis, and the link's strength was shown by the Spearman's correlation coefficient. (Figure 7).

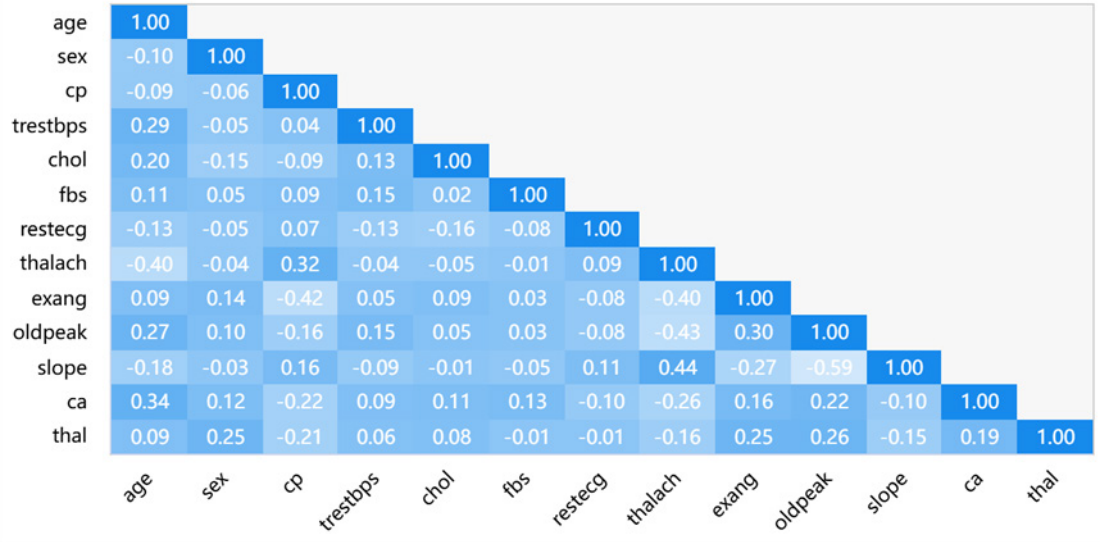

Fig. 7 Spearman correlation visualization

3.3 Logistic Regression Results

For binary logistic regression analysis, it is evident that x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12 and x13 are considered independent variables while target is considered dependent (Table 1). As can be seen from the preceding table, the 0.49 difference in goal can be explained by x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12 and x13.

Additionally, the relationships x2, x9, x10, x12 and x13 will have a big negative influence on the target, whereas x3 and x8 will have a significant positive effect. But x1, x4, x5, x6, x7 and x11 will not have an influence relationship on target (Table 2).

ISSN 2959-6157

Table 2. Results from binary logistic regression analysis

item	regression coefficient	standard error	z-value	Wald $\chi 2$	p-value	OR-value	OR-value95% CI
age	-0.005	0.023	-0.212	0.045	0.832	0.995	0.951 ~ 1.041
sex	-1.758	0.469	-3.751	14.067	0.000	0.172	0.069 ~ 0.432
cp	0.860	0.185	4.638	21.510	0.000	2.363	1.643 ~ 3.398
trestbps	-0.019	0.010	-1.884	3.549	0.060	0.981	0.961 ~ 1.001
chol	-0.005	0.004	-1.224	1.499	0.221	0.995	0.988 ~ 1.003
fbs	0.035	0.529	0.066	0.004	0.947	1.036	0.367 ~ 2.923
restecg	0.466	0.348	1.339	1.793	0.181	1.594	0.805 ~ 3.155
thalach	0.023	0.010	2.219	4.924	0.026	1.023	1.003 ~ 1.045
exang	-0.980	0.410	-2.391	5.719	0.017	0.375	0.168 ~ 0.838
oldpeak	-0.540	0.214	-2.526	6.383	0.012	0.583	0.383 ~ 0.886
slope	0.579	0.350	1.656	2.742	0.098	1.785	0.899 ~ 3.543
ca	-0.773	0.191	-4.051	16.414	0.000	0.461	0.317 ~ 0.671
thal	-0.900	0.290	-3.104	9.634	0.002	0.406	0.230 ~ 0.718
intercept	3.450	2.571	1.342	1.800	0.180	31.515	0.204 ~ 4868.206
remark: implicit variable = target							
McFadden R2= 0.494							
Cox & Snell R2= 0.494							
Nagelkerke R2= 0.660							

Using the model prediction accuracy to determine the quality of the model fit, the research model has an overall prediction accuracy of 85.15%, indicating acceptable model fit. Prediction accuracy is 76.81% when the true value is 0. Prediction accuracy is 92.12% when the true value is 1 (Table 3).

Table 3. Logistic Regression Prediction Accuracy

0		predicted value		Predictive accuracy	Predictive error rate
		1			
real value	0	106	32	76.81%	23.19%
	1	13	152	92.12%	7.88%
total				85.15%	14.85%

3.4 Random Forest Results

Overall, the weighted average Precision is 0.55, Recall is 0.61, and F1-score is 0.54, indicating the model's overall predictive capability, but with significant differences in predicting performance. The macro average Precision, Recall, and F1-score are 0.53, 0.51, and 0.47, respectively, reflecting the overall average performance across factor (Table 4).

Table 4. Forest Prediction Performance

	precision	recall	f1-score	support
0	0.43	0.13	0.20	23
1	0.63	0.89	0.74	38
accuracy			0.61	61
macro avg	0.53	0.51	0.47	61
weighted avg	0.55	0.61	0.54	61

4. Conclusion

In conclusion, this study highlights the complexity of predicting cardiovascular disease (CVD), with the logistic regression model identifying significant predictors such as chest pain type, resting blood pressure, fasting blood sugar level, serum cholesterol. The random forest classifier revealed performance disparities among target, showing strong predictive accuracy for factors. Future efforts should focus on inclusive healthcare policies, improving preventive care access, and raising awareness of CVD risk factors tailored to diverse populations. Notably, the findings suggest that in are linked to increased CVD risk in the study population, warranting further research to understand the implications for cardiovascular health management.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Ciumărnean L, Milaciu M V, Negrean V, et al. Cardiovascular Risk Factors and Physical Activity for the Prevention of Cardiovascular Diseases in the Elderly. Int J Environ Res Public Health, 2021, 19(1): 207.

[2] Nedkoff L, Briffa T, Zemedikun D, Herrington S, Wright F L. Global Trends in Atherosclerotic Cardiovascular Disease. Clin Ther, 2023, 45(11): 1087-1091.

[3] Martin S S, Aday A W, Almarzooq Z I, et al. Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. Circulation, 2024, 7: 1164

[4] Rui Q. Effects of different doses of fluvastatin sodium extended-release tablets on lipid metabolism and cardiac function in patients with dyslipidemia with coronary heart disease. Contemporary medicine, 2021, 27, 72-74.

[5] Yao Y S, Li T D, Zeng Z H. Mechanisms underlying direct actions of hyperlipidemia on myocardium: an updated review. Lipids Health Dis, 2020, 19(1): 23.

[6] Xiufang Z, Xiaoyu C. Clinical observation on the treatment of coronary atherosclerotic heart disease combined with dyslipidemia by combining self-modeled lipid-regulating soup and atorvastatin calcium tablets. Chinese folk therapy, 2022, 30(7): 93-96.

[7] Ritchie R H, Abel E D. Basic Mechanisms of Diabetic Heart Disease. Circ Res, 2020, 126(11): 1501-1525.

[8] Joseph J J, Deedwania P, Acharya T, et al. Comprehensive Management of Cardiovascular Risk Factors for Adults With Type 2 Diabetes: A Scientific Statement From the American Heart Association. Circulation, 2022, 145(9): e722-e759.

[9] Zhang X, Liu J, Wang M, et al. Twenty-year epidemiologic study on LDL-C levels in relation to the risks of atherosclerotic event, hemorrhagic stroke, and cancer death among young and middle - aged population in China. J Clin Lipodol, 2018, 12(5): 1179-1189.

[10] Egan B, Zierath J R. Exercise metabolism and the molecular regulation of skeletal muscle adaptation. Cell Metab, 2013, 17(2): 162-184.

[11] Gan Z, Fu T, Kelly D P, et al. Skeletal muscle mitochondrial remodeling in exercise and diseases. Cell Res, 2018, 28(10): 969-980.

[12] Yida T. Reducing the burden of metabolic cardiovascular disease starts with scientific exercise. Chinese Journal of Health Management, 2024, 18(1): 3-6.

[13] Ping D, Zhengqiu Z, Han W. The impact of gender differences on the correlation between ultra fast pulse wave velocity and cardiovascular disease risk factors. Chongqing Medical, 2020, 49 (16): 2642-2645.