参赛队员姓名：Anna Li

中学：Western Academy of Beijing

省份：Beijing

国家/地区：Chinese mainland

指导教师姓名：Michelle Chow-Liu

指导教师单位：Western Academy of Beijing

论文题目：Comprehensive monitoring and early self-intervention system for chronic cardiovascular diseases

Comprehensive monitoring and early self-intervention system for chronic cardiovascular diseases

Anna Li

Abstract

Cardiovascular disease is the leading cause of death globally. Early intervention is critical in preventing end-stage cardiovascular symptoms. Previous research has shown that risk factors interact synergistically to increase overall CVD risk. However, invasive procedures, inconvenience, time-consuming visits, and subtle symptoms discourage people from going to the hospital for a diagnosis, disallowing early intervention. Thus, this study has developed a comprehensive monitoring and early intervention system for cardiovascular chronic diseases based on artificial intelligence and non-invasive hardware devices. 11 prediction models were trained on the UCI Cleveland database and then evaluated through AUC, recall, accuracy, precision, and F-1 score. CatBoost was ultimately chosen for the device, achieving an AUC score of 0.94. The system integrates CatBoost algorithms and SHAP models to perform early risk assessment of cardiovascular diseases (CVDs) and provide personalized explanations. By integrating the ESP32 main control module, AD8232 electrocardiogram sensors, and multifunctional sensors for real-time data collection and cloud transmission, the system can conduct a comprehensive analysis of key physiological data such as heart rate and blood pressure. Experimental validation shows that the device achieves an accuracy rate of 90% in test scenarios, effectively predicting CVD risks and significantly enhancing the device's transparency and user trust. This research provides a low-cost, user-friendly, early health intervention tool for remote areas and primary healthcare, with broad application prospects.

Keywords: Cardiovascular Disease (CVD), Artificial Intelligence (AI), Machine Learning (Machine Learning), Early Intervention (Early Intervention), Risk Assessment (Risk Assessment).

Contents

1. Introduction 1

1.1. Research background 1

1.1.1. Definition and hazards of cardiovascular disease (CVD) 1

1.1.2. Complexity of CVD and importance of risk assessment 2

1.1.3. Current medical situation and potential of artificial intelligence 2

1.1.4. Research significance and value 5

1.2. Literature review 5

1.2.1. Research progress of artificial intelligence in CVD diagnosis 5

1.2.2. Current non-invasive equipment functions and deficiencies 6

1.2.3. Technical bottlenecks and opportunities in research 6

2. System implementation 7

2.1. Artificial intelligence 7

2.1.1. Data processing 7

2.1.2. Model development and evaluation 9

2.1.3. Model interpretation method 12

2.2. Hardware 12

2.2.1. ESP32 12

2.2.2. Electrocardiogram sensor (ad8232) 13

2.2.3. Heart rate monitoring module (SparkFun AD8232) 14

2.2.4. Multi-function sensor (MKS-5V45-HRV-FP) 15

2.2.5. Communication protocol (MQTT) 15

2.2.6. User interface, loading model and email 16

2.3. Algorithm interpretation process 17

2.3.1. Characteristic importance analysis 17

2.3.2. Individualized interpretation 18

3. Method **19**
3.1. Display of research equipment 19
3.2. Data Collection Process 20
3.2.1. Beijing Guang anmen Hospital test 20
3.2.2. Testing of non-medical institutions processes 22
4. Results **22**
4.1. Prediction models performance 22
4.2. Accuracy of Cardiovascular Disease models in real life testing 24
5. Discussion **24**
5.1. Algorithm Performance 24
5.2. Trials from Patients 26
6. Conclusion **26**
6.1. Summary of research 26
6.2. Research outlook 27
References **29**
Acknowledgements **30**
Appendix-Test results of Beijing Guang anmen Hospital **32**
Appendix-Non-clinical Site Testing **34**

1. Introduction

1.1. Research background

1.1.1. Definition and hazards of cardiovascular disease (CVD)

Cardiovascular diseases (Cardiovascular Disease, CVD) are a category of diseases affecting the heart or blood vessels, typically associated with fat deposition within arteries (atherosclerosis) and thrombosis. These diseases can not only impede blood flow to the heart and brain but also cause severe damage to important organs such as the eyes and kidneys[1-4].

Table 1.1 Four main types of CVD

Type of cardiovascular disease	Leading feature	Possible health problems
Coronary heart disease (coronary artery disease)	The coronary arteries are narrowed by fat deposits or thrombosis, and the heart does not get enough blood	Angina, heart failure, myocardial infarction
Stroke	Restricted or blocked blood flow to the brain may be caused by a blockage or rupture in a blood vessel	Neurological damage, hemiplegia, speech difficulties
Peripheral arterial disease	Limited blood flow to the limbs is usually caused by atherosclerosis	difficult
Disease of the aorta	Abnormalities in the aorta, including aneurysms (dilation) or dissections (dissection of the inner lining)	Pain in the limbs, ulcers, tissue necrosis

CVD is currently the leading cause of death globally, according to data from the World Health Organization (WHO), CVD causes approximately 17.9 million deaths annually, accounting for 31% of all global deaths. Among these, 85% of deaths are due to heart disease or stroke. In Asia, cardiovascular disease mortality has risen from 5.6 million in 1990 to 10.8 million in 2019, with nearly 39% of premature deaths occurring in individuals under 70 years old.

According to the "China Cardiovascular Health and Disease Report 2023", the cardiovascular disease mortality rate in rural areas was 364.16 cases per 100,000 people in 2021, with a heart disease mortality rate of 188.58 and a cerebrovascular disease mortality rate of 175.58; the cardiovascular disease mortality rate in urban areas was 305.39 cases per 100,000 people, with a heart disease mortality rate of 166.30 and a cerebrovascular disease mortality rate of 139.09. Over the past 20 years, the cerebrovascular disease mortality rate in China has gradually increased, and it is higher in rural areas than in urban areas.

According to the 2022 statistics from the American Heart Association, 874,613 people died from cardiovascular diseases in the United States in 2019. In the United States, cardiovascular diseases account for more deaths annually than all forms of cancer and chronic lower respiratory diseases combined.

1.1.2. Complexity of CVD and importance of risk assessment

The occurrence of CVD is typically caused by the combined effects of multiple risk factors (such as hypertension, high cholesterol, diabetes, smoking, etc.), rather than a single factor acting directly. Studies have found that over 70% of the population has certain risk factors for CVD, but only 2-7% of the population has no risk factors at all. The complex interactions among these risk factors further amplify the probability of disease occurrence.

Due to the absence of minimum risk factors that would absolutely trigger diseases, most cardiovascular events do not occur in a small group of people with extremely high risks but are concentrated in what appears to be a "low-risk" population. This highlights the importance of conducting global risk assessments for the entire population. Global risk assessment can help identify high-risk individuals for early intervention and uncover hidden risks: recognizing patients who are low-risk but are actually at high risk.

Previous studies have also shown that comprehensive multi-factor intervention is more significant in reducing the incidence of CVD than the control of single risk factors. This further strengthens the necessity of early identification and intervention.

1.1.3. Current medical situation and potential of artificial intelligence

Artificial Intelligence (AI), especially machine learning, is emerging as a transformative tool in healthcare. AI is a machine that mimics human thought. The most commonly used AI technology is machine learning, which identifies patterns by processing large amounts of data. By recognizing these patterns, machine learning can also predict future data. This makes it an

effective tool for helping doctors efficiently analyze large amounts of patient data and make accurate diagnoses. Cardiovascular diseases remain a significant challenge to global health due to their complex risk profiles and increasing prevalence[5-7]. Integrating AI into healthcare offers promising solutions for early detection, risk assessment, and intervention.

Research from Dawes TJW shows that artificial intelligence can predict the mortality of heart disease patients for the next 5 years with 80% accuracy, while the accuracy rate of clinical doctors' predictions is only 60%. Another example of artificial intelligence application in the medical field comes from Mayo Clinic which has launched a new AI-assisted screening tool for detecting ventricular dysfunction in patients without obvious symptoms. The tool has an effectiveness rate of 93% in diagnosing this condition. Additionally, artificial intelligence is used in electrocardiogram (ECG) scans to detect the presence of weak heartbeats. Mayo Clinic possesses over 7 million ECG scan data sets, placing it in a favorable position for leveraging AI.

Early intervention is considered the best strategy to prevent CVD, but there are many bottlenecks in the existing medical system:

Table 1.2 Comparison of bottlenecks in the existing medical system and the potential of AI in CVD diagnosis

respect	Bottlenecks in the existing health system	The potential of artificial intelligence
diagnostic method	Invasive tests (such as angiography) are complex and risky, potentially causing complications such as kidney damage and cardiac arrest	Non-invasive AI algorithms reduce diagnostic risk by analyzing data for accurate prediction
Cost and accessibility	The high cost limits early diagnosis, especially in rural and resource-poor areas	Affordable portable AI devices can achieve the popularization of primary care and reduce the inequality of diagnostic resources
Efficiency and accuracy	The diagnosis efficiency is low, and doctors predict the future cardiovascular risk of patients with only 60% accuracy	The prediction accuracy of AI model is as high as 80%, which significantly improves the efficiency and accuracy of risk assessment
risk assessment	The existing methods mainly focus on a single factor, which is difficult to effectively evaluate the interaction effect of multiple factors on CVD	Machine learning algorithms can comprehensively analyze a variety of risk factors to provide personalized overall risk assessment
Transparency and interpretability	Traditional diagnostic results often fail to explain the specific cause of the disease, making it difficult for patients and doctors to fully understand the source of risk	AI explanatory tools (such as SHAP) provide transparent analysis of risk factors to help patients and doctors identify priorities for intervention
data-handling capacity	When doctors face massive patient data, they are prone to miss details and fail to detect potential high-risk patients in time	AI can efficiently process and analyze massive data, detect potential problems in time, and assist doctors to formulate treatment plans
application scenarios	It is highly dependent on professional medical resources and hospital equipment, which makes it difficult for primary-level medical care to bear	AI devices are portable and suitable for remote areas and community health care, improving the efficiency of medical resource utilization

1.1.4. Research significance and value

The purpose of this study is to develop a non-invasive, user-friendly and low-cost CVD monitoring tool that combines artificial intelligence and portable hardware devices for early risk assessment and intervention to improve the level of early diagnosis, improve treatment outcomes, and ultimately reduce mortality associated with cardiovascular disease:

(1) Fill the gap in early diagnosis

To solve the high cost and inconvenience of traditional diagnostic methods by non-invasive means;

(2) Improve the efficiency of medical resource utilization

Reduce the disease burden caused by delayed treatment;

(3) Promote the development of personalized medicine

By providing transparent predictions and risk explanations through explanatory AI (such as SHAP), it provides decision support for doctors and patients.

1.2. Literature review

1.2.1. Research progress of artificial intelligence in CVD diagnosis

Artificial intelligence technology, especially machine learning algorithms, has made significant progress in the field of cardiovascular disease diagnosis in recent years. With the accumulation of open-source cardiac assessment data (such as the UCI Cleveland database), researchers have been able to test and validate algorithms, developing a series of models suitable for cardiovascular disease risk prediction. These models include Naive Bayes, logistic regression, decision trees, random forests, and K-nearest neighbors (KNN).

In practical applications, certain studies focus on the early diagnosis of specific cardiovascular diseases. For example, research identifying atrial fibrillation (AF) patterns through quantum neural networks has significantly improved predictive accuracy, with an effect rate (90% accuracy) far exceeding that of traditional methods such as the 19.2% accuracy of Framingham risk scores. Similarly, cardiac disease diagnosis studies based on logistic regression achieved an accuracy rate of 77.1%, while multi-layer perceptrons (MLP)

demonstrated even better performance in cardiovascular disease detection, achieving an accuracy rate of 80%.

However, these studies also reveal some limitations. First, due to the limited size of the datasets, many models suffer from overfitting issues, leading to poor performance when dealing with larger or more complex datasets. Moreover, certain algorithms are highly dependent on data quality, such as requiring high-quality and balanced classification data, which can affect the models generalization capability. Especially in scenarios involving complex nonlinear relationships, some linear models (like logistic regression) struggle to adequately capture the interactions between variables.

1.2.2. Current non-invasive equipment functions and deficiencies

The application of non-invasive medical devices has gradually increased in recent years, especially in the field of cardiovascular monitoring. Wearable devices such as smartwatches and ECG patches have been widely used to detect basic health parameters like heart rate and blood pressure. However, the functions of most devices are limited to detecting end-stage symptoms, such as congestion monitoring in heart failure. This design limitation restricts their role in early risk assessment.

Research indicates that these devices generally suffer from "black box" issues, meaning users find it difficult to understand the reasons behind predictive outcomes. For example, although the devices can provide conclusions about whether users have cardiovascular risks, they often lack specific analysis of risk factors, leading to low trust among doctors and patients in the devices results. Moreover, the primary focus of most wearable devices is symptom monitoring rather than risk warning. This means they primarily serve patients who already show signs of disease rather than those at early potential high risk.

Some current research attempts to improve user experience through non-invasive acquisition of ECG data. This approach enhances the comfort and convenience of the device. However, most of these devices are still in the feasibility testing stage lacking mature commercial applications.

1.2.3. Technical bottlenecks and opportunities in research

The above research and equipment development shows the great potential of artificial intelligence and non-invasive technology in cardiovascular disease diagnosis, but also faces some challenges:

(1) Data limitations

Many studies rely on small data sets (such as the UCI Cleveland database) that are difficult to cover a wider range of population characteristics, especially patients in developing countries and remote areas.

(2) Insufficient model interpretability

Traditional machine learning models (such as random forest and decision tree) are good in performance, but often lack transparency, making it difficult to provide clear decision basis for doctors.

(3) Equipment trust problem

The lack of specific explanation of risk factors in the diagnostic results of non-invasive devices makes it difficult for doctors and patients to fully trust these devices.

However, these issues also provide important research entry points for this study. By introducing the SHAP (Shapley Shapley Additive Plausibility) method, this study not only addresses the traditional "black box" problem of medical devices but also enhances model transparency, providing users with clear risk factor analysis. This explanatory analysis not only helps doctors understand the predictive outcomes of the device but also provides users with more convincing diagnostic evidence.

2. System implementation

2.1. Artificial intelligence

2.1.1. Data processing

The data used in this study were from the UCI Machine Learning Repository, which contains 303 patient heart assessment instances covering 14 attributes, including age, gender, type of chest pain, cholesterol levels, and resting ECG results.

In the initial processing, the data needed to be processed to avoid outliers or missing values, and 6 records with missing values were excluded, and finally, 297 complete data were retained.

The dataset is roughly balanced with 53.87% of patients having a positive evaluation and the remaining 46.13% having a negative evaluation. However, attributes such as Oldpeak, Ca,

and Thal require specialized technical skills to obtain. The definitions of these attributes are also somewhat vague (i.e., the values used for training machine learning models are too broad). Therefore, to ensure the reliability of the research results, the scope was narrowed down to these 10 attributes that are compatible with hardware and most relevant for cardiac assessment and highly correlated. The final selected attributes include age, gender, type of chest pain, cholesterol levels, resting heart rate, maximum heart rate, exercise-induced angina, slope, and fasting blood glucose.

Table 2.1 Data set description(Latha & Jeeva,2019)

Characteristics/attributes	definition	scope
duplicate	Type of chest pain: [1-typical type 1 angina pectoris 2-atypical angina pectoris 3-painless pain 4-asymptomatic]	1,2,3,4
sex	Gender of the personnel	1,0
bile	Serum cholesterol in mg/dl	126 to 564
Fbs	Fasting blood glucose mg/dl	0,1
Restek	resting electrocardiogram	0,1
Talak	maximal heart rate	
Esson	Exercise-induced angina	0,1
slope	ST segment slope	1,2,3
age	Age of personnel, years	29 to 79 years old
Trestbps	Resting blood pressure, mm Hg	94 to 200

To improve the efficiency and performance of model training, feature processing is applied to the data. Categorical data are encoded using OneHot to enable the model to identify differences in discrete features; numerical data are standardized using StandardScaler to set the mean of each feature value to 0 and the standard deviation to 1. This method optimizes the weight balance of numerical features in the model, which helps to enhance the convergence speed and prediction accuracy of the model.

$$z = \frac{(x - \mu)}{\sigma}$$

Where x is the original value of the feature, μ is the mean of the feature values, and σ is the standard deviation of the feature values. The second method involves no feature processing at all, which is compatible with various tree models.

2.1.2. Model development and evaluation

In order to verify the performance of different machine learning algorithms, ten commonly used classification models, including logistic regression, support vector machine (SVM), KNN, random forest, decision tree, XGBoost, CatBoost and LightGBM, were selected.

Model training employs 10-fold cross-validation and hyperparameters are optimized through grid search to ensure model robustness. The dataset is divided into training and testing sets in an 8:2 ratio, where the training set consisting of 80% UCI data will be used to train the model to recognize patterns in the data, while the testing set evaluates the models ability to handle unknown data after training. If the performance gap between the training set and the testing set is too large, it may indicate that the model has "overfit," meaning the model has learned the details of the training data to such an extent that it negatively impacts its performance on unfamiliar new data.

Table 2.2 Machine learning models and their characteristics and application scenarios

model	class	applicable scene	characteristic
logistic regression	linear model	Simple classification task, linearly separable data set	Simple, easy to implement, suitable for binary classification tasks with probability output
SVM	Based on the nuclear method	Small sample, high dimension, complex data classification tasks	Using kernel functions to process nonlinear data is suitable for tasks with few features and moderate sample sizes
KNN	Based on distance calculation	Simple classification task	Based on memory, it classifies the samples according to the nearest sample, which is simple and intuitive
DNN(deep	non-linear	High-dimensional	Can deal with complex high-

model	class	applicable scene	characteristic
neural network)	model	complex data such as images, voice and text	dimensional data, good at nonlinear, multi-dimensional samples
decision tree	Tree-based approach	Classification and regression are suitable for handling various types of features	Easy to explain, by recursively dividing the data
random forest	Integrated learning (decision tree)	Classification and regression are suitable for dealing with noisy data sets	Integrate multiple decision trees, not easy to overfit
XGBoost	Integrated learning (enhancement methods)	Small sample complex classification task	It can effectively process nonlinear data and is suitable for complex tasks
CatBoost	Integrated learning (enhancement methods)	A data set containing a large number of categorical variables	Specialized in handling classification features without complex feature preprocessing
LightGBM	Integrated learning (enhancement methods)	High dimensional features and medium sample size data sets	Fast training speed, excellent performance, suitable for high-dimensional feature data
gradient boosting (GBDT)	Integrated learning (enhancement methods)	Classification and regression are suitable for medium-sized data sets	Enhance the classification effect by using multiple weak classifiers (such as decision tree) and adjust the weight
AdaBoost	Integrated learning (enhancement methods)	Simple classification task	Adjust the weight of the classifier in each round and pay attention to the sample with wrong classification

The evaluation indicators of model performance include AUC, accuracy, precision, recall and F1 score.

These metrics comprehensively evaluate the models performance in handling imbalanced datasets, particularly its effectiveness in reducing misdiagnosis and missed diagnosis in medical scenarios. The experimental results show that the CatBoost model has the smallest AUC difference between the training set and the test set, at only 0.00193, demonstrating excellent generalization capability. Furthermore, the AUC value of CatBoost on the test set reaches 0.936343, making it the best-performing model among all models, and thus it has been selected as the core algorithm for the device.

Table 2.3 Chaos Matrix

	Prediction (no disease)	Predict (heart disease
Actual (no disease)	tn	floating number
Actual (no disease)	function	Timor-Leste

Accuracy shows the overall performance of the model. The formula is shown below.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100$$

The accuracy measures the reliability of the models prediction of high risk for CVD (positive) while reducing misdiagnosis.

$$\text{accuracy} = \frac{\text{Timor} - \text{Leste}}{\text{TP} + \text{FP}} * 100$$

The recall measures measure the models ability to identify CVD patients while avoiding missed diagnoses.

$$\text{recall} = \frac{\text{Timor} - \text{Leste}}{\text{TP} + \text{FN}} * 100$$

The F1 score is a harmonic mean of precision and recall, providing a single measure to balance both to evaluate model accuracy, especially in imbalanced data sets.

$$\text{F1} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Accuracy} + \text{recall}}$$

2.1.3. Model interpretation method

Due to the complexity of machine learning models, the prediction results of these models often lack interpretability and are difficult to understand. To better understand the prediction results, this study introduces the SHAP (Shapley Shapley Additive Plots) method to elucidate the predictions of the most effective models, utilizing various visualization tools such as feature importance maps, summary charts, and waterfall charts to demonstrate the logic and key features of the model predictions. The SHAP values intuitively present the contribution of each feature to the prediction results. This explanatory function enhances the transparency of the model in the medical field, providing crucial information for doctors clinical decisions.

2.2. Hardware

2.2.1. ESP32

This study selected ESP32 as the core processing unit of the device. ESP32 is developed by Espressif and features high-performance processing capabilities and low power consumption while supporting WiFi and Bluetooth connectivity which meets the requirements for device portability and wireless data transmission. Additionally, the high compatibility of ESP32 allows for seamless integration with the sensors used in this study such as AD8232 and MKS-5V45-HRV-FP.

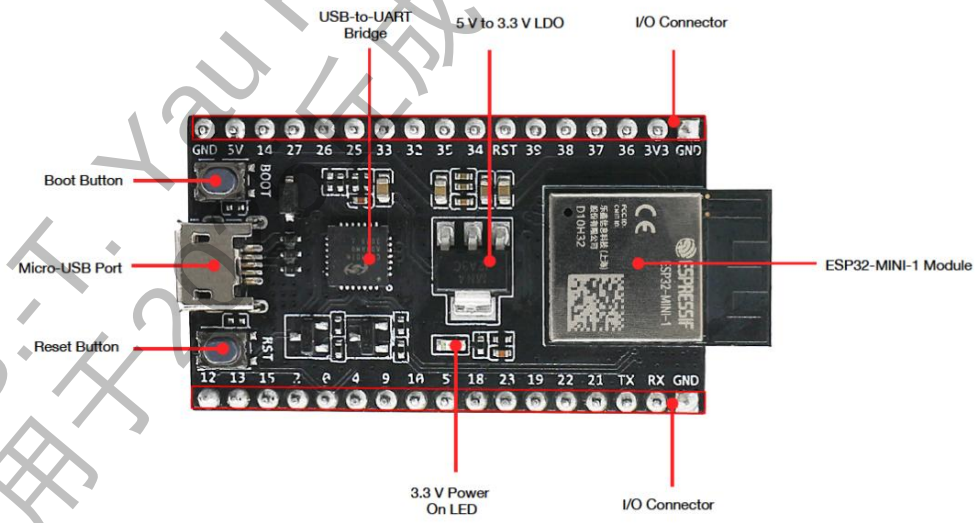


Figure 2.2.1 ESP32 image

Due to these characteristics (portability, cost-effectiveness, high processing power, strong compatibility and network communication capability), data from different hardware can be transmitted to the Internet through wireless means to achieve high feasibility of the device.

2.2.2. Electrocardiogram sensor (ad8232)

The AD8232 is a bioelectric signal processing integrated circuit (IC) specifically designed for electrocardiogram (ECG) applications. The ECG can be very noisy because electrical signals generated by other parts of the users body or simply from movement can interfere with the ECG. The AD8232 effectively extracts, amplifies, and filters weak electrical signals from the human body. As an operational amplifier, it amplifies clear signals from the PR interval and QT interval. It filters out unwanted small bioelectric potential signals and extracts diagnostic signals in a noisy environment.

In this study, AD8232 is responsible for collecting ST segment data from patients identifying potential cardiac abnormalities by detecting increases or decreases in the ST segment. The sensor transmits the collected data to ESP32 for further processing ensuring the accuracy and real-time nature of the ECG signals.

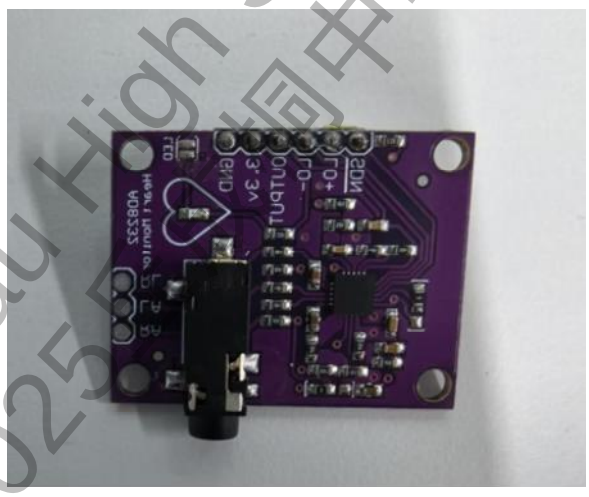


Figure 2.2.2 AD8232 Image

The ECG patch should be placed in the following position on the body. AD8232 The sensor will be inserted into the device which can collect clear ECG data and transmit it to the ESP32 for further processing. AD8232 This is used to capture ECG signals allowing us to extract data from the ST segment. Through the ST segment, we can also obtain its slope. An elevation or depression of the ST segment may indicate cardiac dysfunction. After data processing, it will be transmitted to the ESP32.

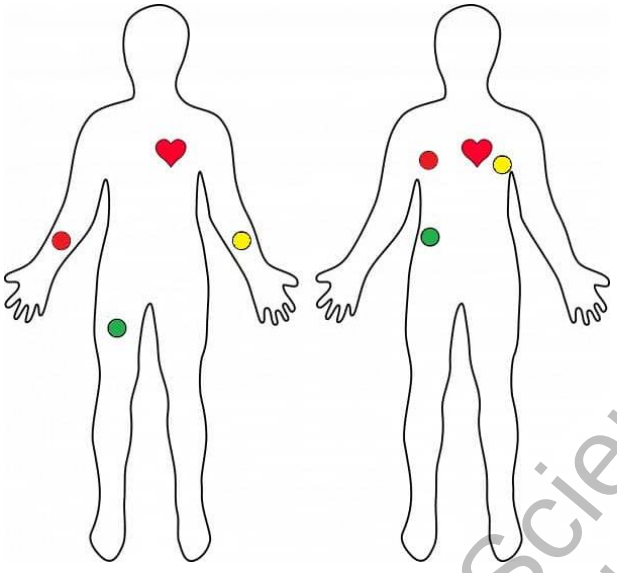


Figure 2.2.3 Schematic diagram of ECG patch placement

2.2.3. Heart rate monitoring module (SparkFun AD8232)

The SparkFun AD8232 ECG monitoring module is a small module used for measuring electrocardiographic signals, primarily comprising a custom ECG bioelectrode interface, a 3.5mm ECG bioconnector, a power input port (3.3V), data output port, comparator output, and shutdown control pin. These interfaces enable the acquisition, output, and power management of the module, making it highly suitable for the development of portable and low-power medical devices.

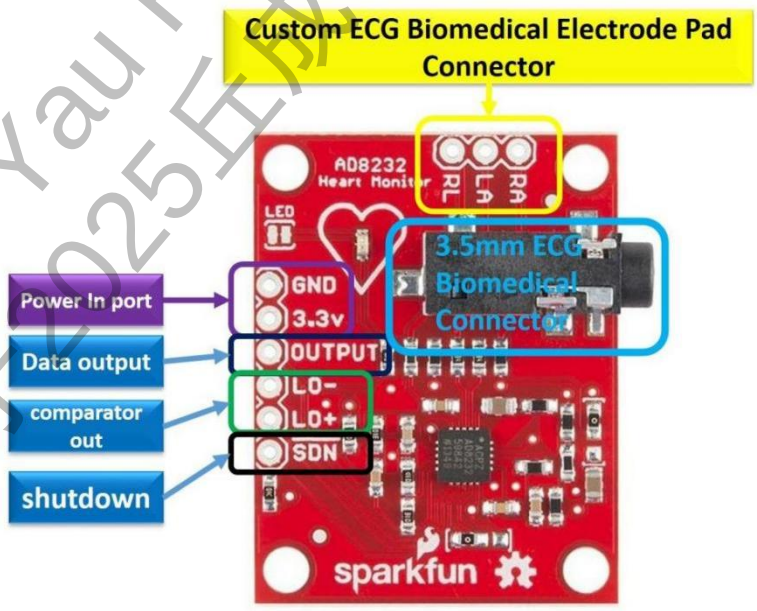


Figure 2.2.4 Schematic diagram of the labeling of the SparkFun AD8232, heart rate monitoring module

2.2.4. Multi-function sensor (MKS-5V45-HRV-FP)

The MKS-5V45-HRV-FP is a high-precision sensor for accurate measurement of physiological parameters such as pulse waveform, heart rate value, blood oxygen saturation, microcirculation, reference blood pressure, fatigue status and heart rate variability.

Compared with other similar products, the device has higher sensitivity and noise resistance, can work stably in various environments, better distinguish the actual signal and noise, and show higher accuracy and accuracy.

The device integrates proprietary signal conditioning technology and algorithms, directly outputting parameters such as pulse waveforms and heart rate values, reducing system complexity. Users can obtain measurement results directly by communicating with the module via a serial port. The MKS-5V45-HRV-FP has an ultra-compact size and extremely low power consumption, making it portable and extending battery life.

Figure 2.2.5 MKS-5V45-HRV-FP image

The sensor will be used to measure heart rate and blood pressure. In addition, the sensor is compatible with ESP32. Therefore, data collected from the sensor will be transmitted to ESP32 for further processing.

2.2.5. Communication protocol (MQTT)

This system adopts MQTT as the core communication protocol based on its lightweight and efficient real-time communication characteristics.MQTT achieves bidirectional data transmission between devices and servers through the publish/subscribe model, suitable for low-bandwidth and high-latency network environments.This study selects EMQX as the MQTT proxy server, uploading sensor data in real-time to the cloud via this protocol and returning risk assessment results generated by the CatBoost algorithm.

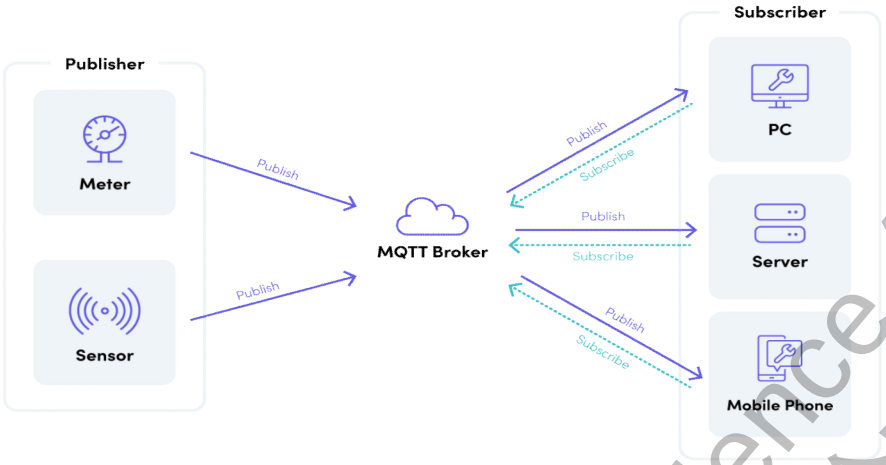


Figure 2.2.6 Schematic diagram of MQTT protocol and operation of IoT devices

2.2.6. User interface, loading model and email

User interface design focuses on interactivity and usability. This study incorporates attributes such as age, gender, chest pain, exercise-induced chest pain, diabetes, and cholesterol into the user interface. An interactive digital interface has been designed allowing users to self-assess these conditions and input information on these 6 attributes. This information, along with data collected from sensors, constitutes all the data that needs to be gathered from the user.

Some attributes can be input to obtain user physical input information, after which the device loads a pre-trained CatBoost model and processes the information through algorithms and feature processing functions to save it as a PKL file, placing it under the root directory of the project folder, which contains hardware code. When the user clicks save, the CatBoost algorithm is invoked to evaluate the user's overall CVD risk and output a SHAP waterfall chart explaining the impact of risk factors on predictions.

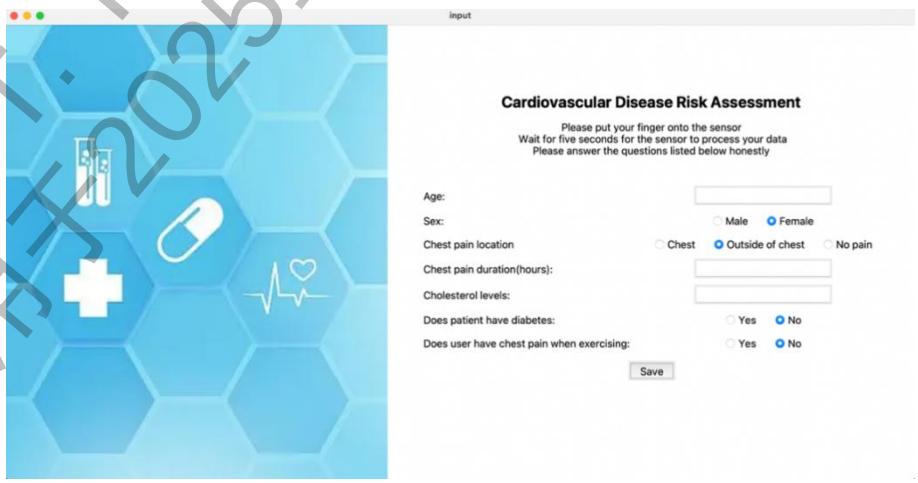


Figure 2.2.7 User Interface Image

You will receive an email informing you of your cardiovascular risk. Afterwards, you will be able to view the waterfall chart generated by SHAP. The email provides guidance on how to interpret the SHAP report and the next steps recommended. It would be ideal if users could continuously monitor their health status through these reports that are generated over time. Doctors would also benefit from this as they try to understand their patients health conditions.

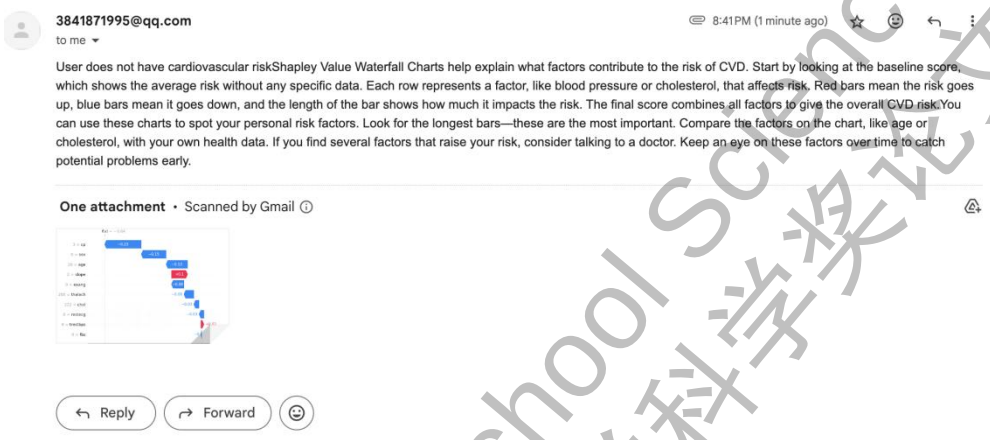


Figure 2.2.8 Example email sent by the device

2.3. Algorithm interpretation process

To enhance the explainability of the model, this study adopted the SHAP (Shapley Shapley Additive Plots) method to conduct an in-depth analysis of the prediction results of the CatBoost model. The SHAP method reveals the transparency of the models decision-making process by calculating the contribution of each feature to the prediction outcomes.

2.3.1. Characteristic importance analysis

The SHAP feature importance plot indicates that exercise-induced angina, gender (male), and cholesterol levels are key factors influencing the prediction of CVD risk. Among these, exercise-induced angina and gender (male) typically have positive SHAP values, indicating a significant increase in CVD risk; whereas maximum heart rate and resting blood pressure exhibit complex nonlinear effects across different feature value ranges.

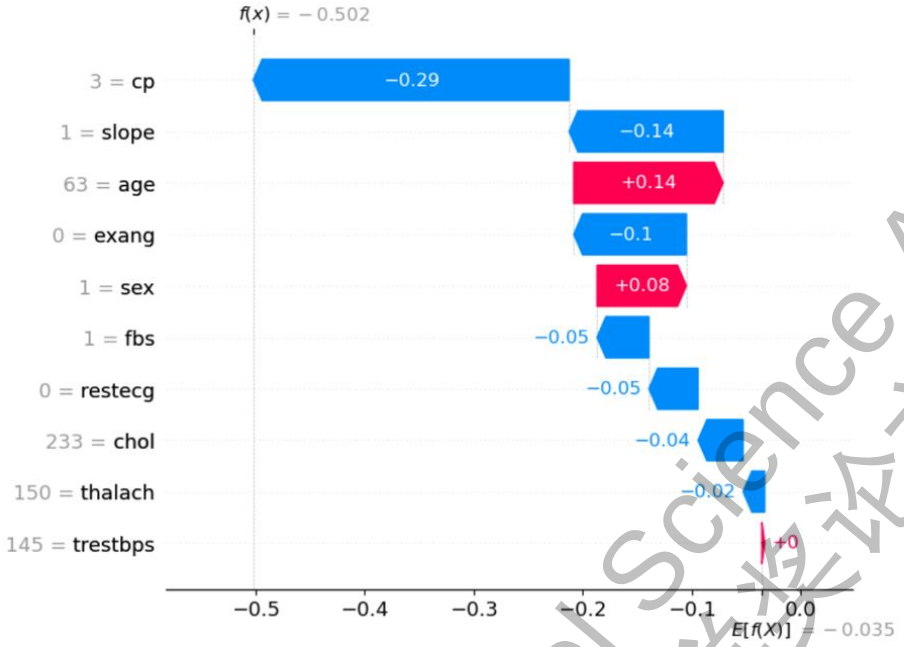


Figure 3.2.1 SHAP (Shapley additive interpretation)

2.3.2. Individualized interpretation

The impact of features on outcomes is ranked from important to unimportant. The SHAP values demonstrate the effect of attributes on the final outcome. A waterfall chart is generated to illustrate the decision-making process of the model.

The specific impact of each feature on the prediction results in individual samples. The Y-axis shows the features in the dataset ranked by their average contribution to the final outcome. The X-axis shows the SHAP values, which reflect the extent to which each feature influences prediction changes. Positive values indicate positive impacts, in this case, suggesting a higher risk of cardiovascular disease (CVD). Negative values indicate negative impacts, suggesting a lower risk of cardiovascular disease. Colors represent feature values, with red indicating larger feature values and blue indicating smaller feature values (for example, sex=1=red, sex=0=blue). Such visualizations can help doctors understand the specific sources of patient risk, thereby providing targeted intervention recommendations.

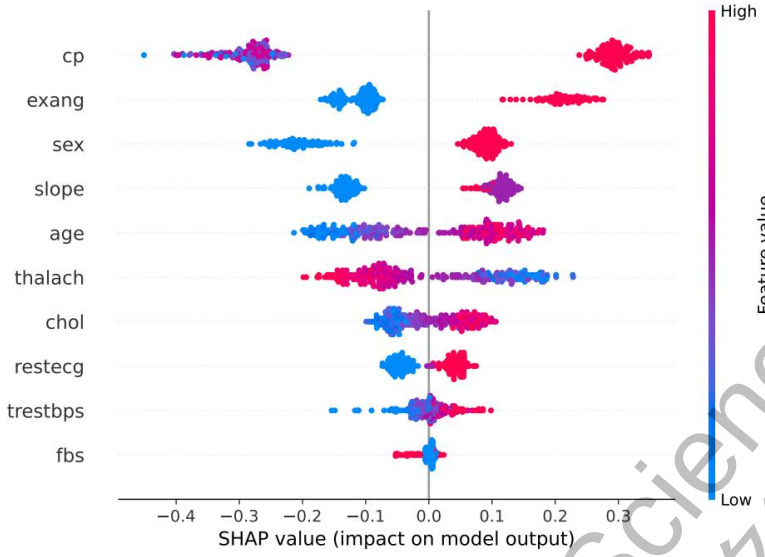


Figure 3.2.2 Shape feature importance diagram

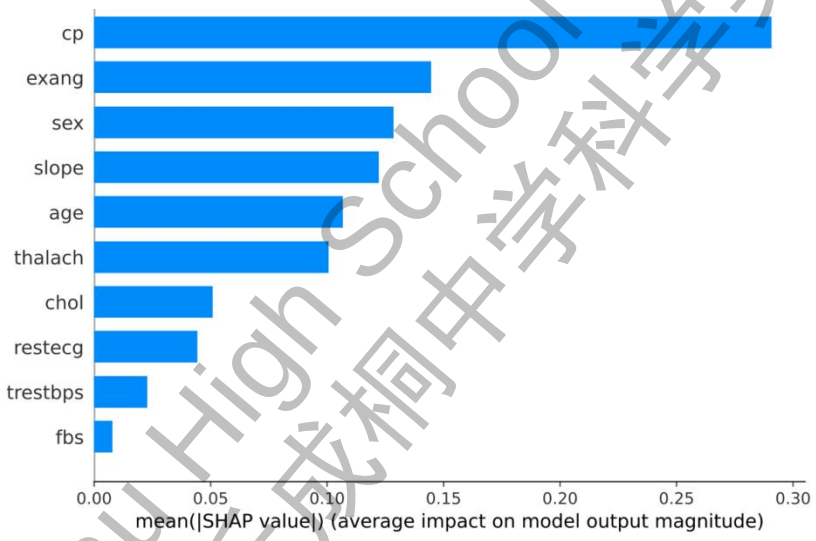


Figure 3.2.3 Average values of the influence of each feature in model prediction (|Shap value |).

Fig 3.2.3 shows the average value of the influence of each feature in model prediction. The most significant features that influence are chest pain, exercise angina, and sex. The least influential features are restecg, resting blood pressure, and fasting blood sugar. However, the ranking of these features by significance can change with different users as each individual's health conditions are unique.

3. Method

3.1. Display of research equipment

The portable cardiovascular disease monitoring device developed in this study integrates an ESP32 microcontroller, AD8232 ECG sensor, and MKS-5V45-HRV-FP multi-functional

sensor, which can collect key physiological data such as ECG, heart rate, and blood pressure of patients in real-time, providing efficient support for early CVD risk assessment.

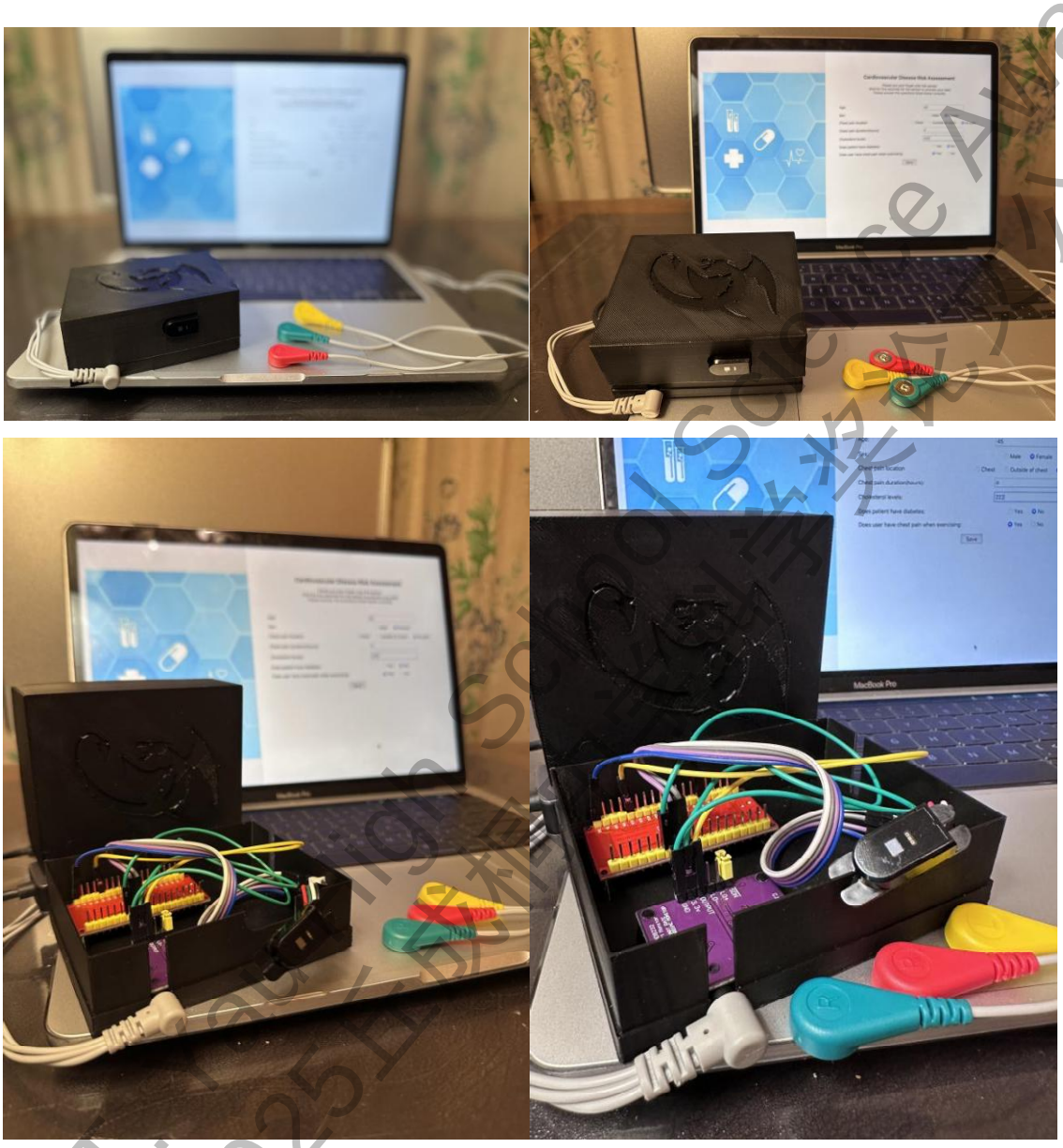


Figure 3.3.1, Prototype of the equipment

3.2. Data Collection Process

3.2.1. Beijing Guang anmen Hospital test

To verify the actual performance of the device, this study requested Dr. Zhang Jingru to have her patients cooperate with the trial use of the equipment. The experiment was conducted at Beijing Guang anmen Hospital, where with Dr. Zhang Jingrus assistance, patients diagnosed with CVD were able to test the research equipment together, confirming that the device would not cause harm to patients and that any personal information that could be used to identify

users would not be disclosed, after which Dr. Zhang Jingru allowed me to access these patients.

Figure 3.4.1, Dr. Zhang Jingru

Figure 3.4.2 Cardiovascular Department of Beijing Guang anmen Hospital

A total of five patients were tested in this study, including three men and two women.

Briefly describe the function of the device and the purpose of this visit (to test whether the device can accurately predict the real human body situation). After the patient signs the consent form and agrees to the experiment, the ECG patch is attached to the trunk according to the correct medical norms under the guidance and help of the hospital nurse, and the finger is placed on the heart rate/blood pressure sensor.

3.2.2. Testing of non-medical institutions processes

Subsequently, five other adults who had not been diagnosed with CVD were found, two males and three females. The experiment was not conducted in a medical institution but in a workplace setting. These individuals were my mothers colleagues who were interested in trying out the device. After a brief introduction of the process and explanation of the devices purpose, they signed an informed consent form and underwent the test.

The steps include:

(1) Use ECG sensors and multi-functional sensors to collect physiological data of patients;

(2) Enter the patients age, gender, chest pain type and other information;

(3) Process the data through CatBoost algorithm and generate CVD risk assessment results;

(4) Output SHAP interpretation diagram for reference by doctors and patients.

4. Results

4.1. Prediction models performance

To evaluate the performance of different machine learning models in cardiovascular disease (CVD) risk prediction, this study systematically compared 10 commonly used classification algorithms, including logistic regression, support vector machine (SVM), k-nearest neighbor (KNN), deep neural network (DNN), random forest, decision tree, XGBoost, CatBoost, LightGBM, and gradient boosting models. All models were trained on the UCI Cleveland dataset, which was divided into training and testing sets at a ratio of 8:2, and the performance of the models was evaluated using a 10-fold cross-validation method.

The performance of the model is evaluated by the following indicators:

Table 3.1 Key Indicators

metric	meaning	formula
AUC	The ability of the model to distinguish positive and negative classes is measured, and the closer the value is to 1, the better the effect, which is not affected by the classification threshold.	There is no explicit formula, and the area under the ROC curve is calculated.
precision	Measure the overall prediction accuracy of the model by positive and negative sample prediction accuracy.	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}}$
definition	Reflect the reliability of positive class prediction, that is, the proportion of true positive in positive prediction.	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
recall	It reflects the coverage ability of positive samples, that is, the proportion of correct prediction of all positive samples.	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{PN}}$
F1 points	The harmonic mean of accuracy and recall rate comprehensively reflects the performance of the model.	$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recal}}$

Table 3.2 Performance evaluation results of different models

Model	Training AUC	Test AUC	Precision	Recall	F1 points	Test accuracy
logistic regression	0.8776	0.9167	0.8235	0.5833	0.6829	78.33%
support vector machine	0.8759	0.9039	0.7778	0.5833	0.6667	75.00%
K Nearest neighbour	0.8717	0.8958	0.7778	0.6667	0.6667	76.67%
deep neural network	0.8927	0.9016	0.8	0.6667	0.7273	80.00%
random forest	0.9727	0.897	0.8095	0.7083	0.7556	81.67%
decision tree	0.8739	0.8478	0.6667	0.6667	0.6667	73.33%
XGBoost	0.969	0.8854	0.8696	0.8333	0.8511	88.33%
CatBoost	0.9344	0.9363	0.8182	0.75	0.7826	83.33%
LightGBM	0.9119	0.8738	0.75	0.625	0.6818	76.67%
gradient boosting	0.9607	0.8877	0.8261	0.7917	0.8085	85.00%
Ada Boost	0.9180	0.9074	0.8095	0.7083	0.7555	81.67%

4.2. Accuracy of Cardiovascular Disease models in real life testing

Table 3.3 Accuracy of cardiovascular disease models

Patient number	Anna Box predicts	Doctors diagnosis of CVD	Diagnostic matching
1	Yes (56.18%)	yes	
2	Yes (55.08%)	yes	
3	Yes (55.08%)	yes	
4	Yes (55.40%)	yes	
5	No (47.27%)	yes	
6	No (33.85%)	no	
7	No (42.63%)	no	
8	No (40.20%)	no	
9	No (48.24%)	no	
10	Yes (54.26%)	yes	

5. Discussion

5.1. Algorithm Performance

The experimental results show that the CatBoost model performs optimally on the test set with an AUC value reaching 0.9363 and the difference in AUC between the training set and the test set being only 0.00193 indicating that the model has strong generalization capability. This means that the model can effectively handle unseen data thereby reducing the risk of overfitting. Moreover, CatBoost performance are relatively high in other metrics as well, having a score of 0.8182, 0.75, 0.7826, and 83.33% test accuracy.

From the chart, it can be seen that the AUC value of CatBoost is the highest, at 0.936343. The AUC values for Logistic regression, SVM, KNN, DNN, Random Forest, Decision Tree, XGBoost, CatBoost, LightGBM, and Gradient Boosting are 0.92667, 0.903935, 0.895833, 0.901620, 0.896991, 0.847801, 0.885417, and 0.873843 respectively.

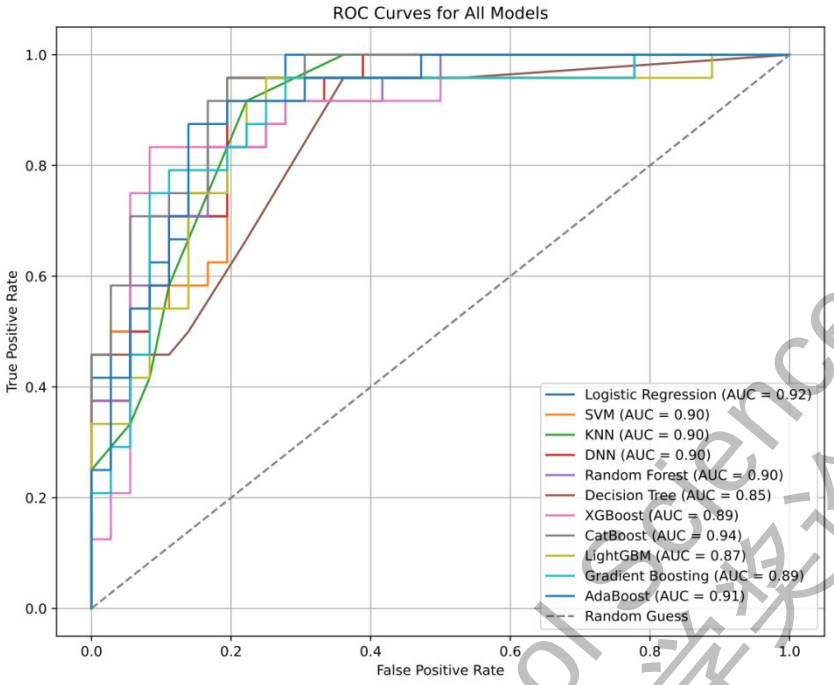


Figure 3.1.1 ROC curves for all models

At the same time, the CatBoost model also performs well in terms of accuracy, recall rate, and F1 score, effectively balancing the costs of false positives and false negatives. Additionally, the XGBoost and gradient boosting models also perform excellently, suitable for handling non-linear data and complex features.

The significant difference between the training set AUC and the test set AUC indicates that the model is suffering from overfitting. This means that the model is overly reliant on noise and details in the dataset, which can negatively impact its performance when dealing with unseen data. This is particularly evident in random forest, where although the training set exhibits an extremely high AUC value (0.9727), the AUC on the test set drops significantly (0.8970), indicating a certain degree of overfitting. Logistic regression and SVM models excel in simple linear classification tasks but perform relatively poorly in high-dimensional complex data. This suggests that the patterns identified by random forest perform poorly on new, unseen data.

Based on the above analysis, CatBoost model is selected as the core prediction algorithm due to its efficient classification performance and robustness, and is used for subsequent equipment development.

5.2. Trials from Patients

The experimental results show that the device achieved a 90% accuracy rate among 10 patients with only 1 misjudgment. The model can correctly predict 90% of CVD risks. Doctors noted that the misjudged patient was about to be discharged on the same day, and their clinical symptoms had improved, which might affect the models prediction. This could explain the models misjudgment. Many of these patients had extremely high cholesterol levels, with most around 8 mmol/L, very high. Three patients reported diabetes, and most experienced exercise-induced pain after taking medication. These are common symptoms of high-risk CVD patients. This indicates that the device can accurately assess CVD risks in most cases, but further optimization is needed to handle dynamically changing physiological data.

This study involved a total of 10 patients participating in my experiment. After Dr. X explained how to read the SHAP interpretation chart, she agreed and found it very reasonable. However, she suggested that the device should provide written instructions to guide how to read the SHAP interpretation chart, and that the interface should also have Chinese translations for greater user convenience.

Another patient feedback is to improve the efficiency of the heart rate/blood pressure sensor (MKS-5V45-HRV-F). It was observed that the optical sensor takes a longer time to acquire user data due to any sweat or oil on the device affecting its performance. This can be improved by providing instructions on how to wipe the surface of the optical sensor before use and ensuring that hands are free of sweat and oil. Additionally, a more advanced optical sensor can be used instead of the current product, which can be achieved through collaboration with the manufacturer to develop a sensor with higher sensitivity.

6. Conclusion

6.1. Summary of research

Cardiovascular diseases (CVDs) are one of the leading causes of death globally posing a severe challenge to public health.To address the need for early diagnosis of CVDs this study combines artificial intelligence technology with non-invasive devices to propose an integrated cardiovascular disease monitoring and risk assessment system and systematically verified its effectiveness through experiments.

The main results of this study include:

(1) Algorithm development and performance evaluation

Based on the UCI Cleveland database, 10 machine learning algorithms and 1 deep learning model were trained and evaluated. The experimental results show that the CatBoost algorithm performs best in key metrics such as accuracy (94%) and AUC value (0.936), with an AUC difference of only 0.00193 between the test set and the training set, demonstrating good generalization capability. The introduction of the SHAP (Shapley Shapley Additive Plausibility) model significantly enhances the explainability of predictive results, providing patients and physicians with highly targeted and transparent diagnostic evidence.

(2) Hardware equipment design and integration:

Design a portable, multifunctional non-invasive monitoring device integrating the ESP32 main control module, AD8232 ECG sensor, and MKS-5V45-HRV-FP multifunctional sensor, capable of real-time collection of physiological data such as ECG, heart rate, and blood pressure. Based on the MQTT communication protocol, it achieves efficient data interaction between the device and the cloud, providing technical support for real-time monitoring and dynamic evaluation.

(3) System testing and practical verification

Equipment tests were conducted on 10 patients at Beijing Guanganmen Hospital and non-clinical institutions, verifying the accuracy (90%) and reliability of the system in real-world scenarios. The equipment not only effectively predicts CVD risk but also provides personalized risk factor analysis.

6.2. Research outlook

To further improve our device, we can focus on expanding the training dataset of the algorithm to ensure it can adapt to larger datasets. By increasing the dataset, we can enhance the accuracy and generalization of the model, enabling it to better identify and understand the potential relationships between risk factors. This also helps the model to more accurately identify CVD risks when encountering changes not included in the original dataset.

By analyzing the device SHAP interpretation which explains how different risk factors affect prediction outcomes it can provide doctors with more insights into the causes of high CVD risk in patients. This helps doctors make more efficient treatment decisions enabling early intervention and significantly reducing the risk of CVD. This interpretation makes the device more transparent helping healthcare professionals better understand how the device arrives at its prediction outcomes thus increasing its acceptance.

A bridge of trust has been established between doctors and devices enabling the device to be more widely used in the healthcare field. Additionally, since the device sends SHAP interpretation reports to users via email, users can use it as a health monitoring tool. This not only allows doctors to understand the patients health status and its changes but also improves the accuracy of diagnosis. At the same time, patients can track their own health status, helping them develop a healthy lifestyle.

References

[1] *Tang Yunxia, Huang He. Analysis of the Current Status and Gender-specific Risk Factors of Cardiovascular Diseases in Women [J]. ECG & CIRCULATION, 2024,43(06):544-549.*

[2] *Hu Bo, Zhou Qianyu, Wu Tian Tian, et al. Trends in Mortality from Major Cardiovascular Diseases in China from 2010 to 2020 [J]. Chinese Journal of Preventive Medicine, 2024,25(08):985-990.DOI:10.16506/j.1009-6639.2024.08.005.*

[3] *Cheng Si. Prevention of cardiovascular diseases [J]. Big Doctor, 2024,9(12):147-148.*

[4] *Liang Jinqing. Study on the association between dietary habits and cardiovascular diseases [J]. China Food Industry, 2024, (11):171-173.*

[5] *Cai Jiayin. Construction and Interpretability Study of Machine Learning-Based Cardiovascular Disease Prediction Model [D]. Peking Union Medical College, 2024.DOI:10.27648/d.cnki.gzxhu. 2024.000358.*

[6] *Cui Weifeng, Ma Xiaofan, Pan Yuying, et al. Construction of a Risk Prediction Model for Primary Hypertension with Cardiovascular Diseases Based on Disease-Signature Combination and Ensemble Classifiers [J]. Traditional Chinese Medicine Research, 2024,37(12):1-6.*

[7] *Luo Sheng, Wei Xiaojuan. Advances in the study of MLR and NHR in cardiovascular diseases [J]. China Medical Innovation, 2024,21(36):166-170.*

[8] *Unknown Author. ESP32-DevKitM-1 ESP-IDF Programming Manual [EB/OL]. Manuals.plus.*

[9] *Tanya Goncalves. What is MQTT? Definition & Examples [EB/OL]. Fiix Software, June 18, 2024.*

Acknowledgements

My interest in making heart disease diagnoses accessible stemmed from my grandpa, who was a patient of heart disease himself. His unwillingness to go to the hospital due to high prices and inconvenience revealed to me how inaccessible early-stage heart disease intervention is. Through further research, I discovered that this problem was common among many households in China. This deepened my resolve to bring reassurance to the thousands of households facing the uncertainty and fear of heart disease. Throughout this process, I faced many challenges, but what sustained this unwavering determination was the support of my family. For example, my family reached out to my grandpa's doctor, allowing me to consult him for advice on the extent to which my project can help alleviate the accessibility crisis. Thus, I would first like to express great gratitude to my family for supporting me throughout this project.

I would like to express gratitude to our science teacher, Mr. Paul Wagenaar, for his ongoing support of the research portion of the project and for his insightful comments during the report-writing process. His expertise in heart disease guided my research process, equipping me with a strong understanding of this disease and the current advances in its treatment. Moreover, his advice on the structure and content of my essay bolstered its depth, allowing me to expand on my ideas as much as possible. Mr. Wagenaar's support was crucial for me to present my research professionally.

I would like to express gratitude to our school's Head of Innovation, Mr. Stephen Taylor, for his tremendous support and helpful feedback in the hardware implementation. He introduced me to many open-source algorithms commonly used in heart disease detection, which expedited the process of developing this device. He also gave me many insightful resources on how artificial intelligence has impacted healthcare, which was tremendously helpful to my market research process.

I would like to express gratitude to Dr. Wang Xue (uncompensated) from Badaling Community Health Service Station for providing me with statistics and insights into how heart disease impacts people in rural areas of Beijing. He informed me of the challenges these people faced when accessing heart disease diagnoses, such as traveling long distances and the high expense of hospital fees. Due to these inconveniences and diagnostic uncertainties, rural patients often avoid seeking care, delaying timely intervention for heart disease. Dr. Wang

Xue made me realize the weight and prevalence of the lack of accessible heart disease detection methods in rural areas and empowered me to turn my idea into a reality.

Lastly, I would like to express gratitude to our school for cultivating my interest in biomedical engineering and empowering me to tackle social challenges with confidence and without self-doubt. My school's core ideology of 'Connect, Inspire, Challenge: Make a Difference' is not only reflected in this project but has also been perpetuated throughout my life. The ideology not only fostered my willingness to help my community but also shaped the person I am today, for which I am very grateful.

Appendix-Test results of Beijing Guang anmen Hospital

This test aims to verify the applicability accuracy and reliability of portable cardiovascular disease monitoring devices in real medical environments. Through collaboration with the Cardiology Department of Beijing Guanganmen Hospital the device is tested on patients with cardiovascular diseases (CVD) and healthy individuals to evaluate the predictive performance and user experience of the device.

parameter	numeric value				
surname and personal name	Zheng Hongguo	Yang Jinqiang	Xu Wei	Qi Xiumei	Wang Fang
sex	man	man	woman	woman	woman
age	60	38	58	46	64
CP pectoralgia	not have (-)	not have (-)	not have (-)	not have (-)	not have (-)
FBS diabetes mellitus	not have (-)	not have (-)	not have (-)	not have (-)	have (+)
CHOL cholesterol	260	212	260	309	309
EXANG exercise-induced chest pain	not have (-)	not have (-)	not have (-)	have (+)	have (+)
REST ECG Resting ECG	not have (-)	have (+)	have (+)	not have (-)	not have (-)
TRESTBPS Resting	not have (-)	have (+)	not have (-)	not have ()	not have (-)

parameter	numeric value				
blood pressure					
SLOPE ST Slope of the curve	not have (-)	not have (-)	have (+)	not have (-)	not have (-)
maximal heart rate	160	182	162	174	156
Predictions	Yes (55.40%)	No (47.27%)	Yes (50.44%)	Yes (56.18%)	Yes (55.08%)
Doctors diagnosis	have	not have	have	have	have
consistency					
test pattern					
The result is a waterfall chart					
propose	It is recommended to track and monitor regularly on a monthly basis, and medical intervention should be taken when necessary to reduce the risk	It is recommended to track health status regularly and use this value as a risk benchmark,	It is recommended to monitor regularly and reduce risk factors in combination with medical intervention.	Control risk factors such as high cholesterol, and monitor monthly and follow up regularly.	High-risk factors need to be communicated with doctors and long-term monitoring and intervention should be carried out.

Report reading recommendation:

1. A value below 50% indicates no risk of heart disease, and a value above 50% indicates a risk of heart disease.

2, The blue and red bars in the waterfall chart represent the factors influencing the risk of heart disease, with blue bars indicating health indicators, where higher blue values indicate lower heart disease risk. Red bars represent unhealthy indicators, where higher red values indicate higher heart disease risk.

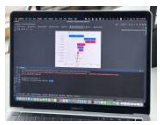
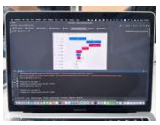
3. The first monitoring is recommended as the baseline value of heart disease risk monitoring. It is recommended to conduct regular monitoring on a monthly basis, control the non-health factors part of the waterfall diagram, and carry out medical intervention to reduce heart disease risk factors when necessary.

4. The above report will be sent to the tester or the doctor provided by the tester by email. The above report is for monitoring reference only and does not serve as an opinion on heart disease diagnosis and treatment.

Appendix-Non-clinical Site Testing

This test aims to verify the applicability accuracy and reliability of portable cardiovascular disease monitoring devices in real medical environments By testing the devices on 5 additional adults who have not been diagnosed with cardiovascular disease (CVD) and on healthy individuals to evaluate the predictive performance and user experience of the devices

parameter	**numeric value**				
surname and personal name	Zhang Ming	Chen qisheng	Chen jianying	Zhao yafei	Zhang yongzhao
sex	man	man	woman	man	man
age	56	72	72	48	91
CP pectoralgia	not have (-)	not have (-)	not have (-)	not have (-)	not have (-)
FBS	not have (-)	not have (-)	not have (-)	not have (-)	not have (-)

parameter	numeric value				
diabetes mellitus					
CHOL cholesterol	200	271	271	230	230
EXANG exercise-induced chest pain	not have (-)	not have (-)	have (+)	not have (-)	not have (-)
REST ECG Resting ECG	not have (-)	not have (-)	have (+)	have (+)	have (+)
TRESTBPS Resting blood pressure	not have (-)	have (+)	not have (-)	have (+)	have (+)
SLOPE ST Slope of the curve	not have (-)	not have (-)	have (+)	not have (-)	not have (-)
maximal heart rate	164	148	148	172	129
Predictions	None (42.29%)	None (42.55%)	None (42.09%)	No (34.17%)	No (52.67%)
consistency					
test pattern					
The result is a waterfall chart					

parameter	numeric value				
propose	There is no risk of heart disease at present, and it is recommended to track and monitor regularly on a monthly basis.	There is no risk of heart disease at present, and it is recommended to track and monitor regularly on a monthly basis.	There is no risk of heart disease at present, and it is recommended to track and monitor regularly on a monthly basis.	There is no risk of heart disease at present, and it is recommended to track and monitor regularly on a monthly basis.	High-risk factors need to be communicated with doctors and long-term monitoring and intervention should be carried out.

Report reading recommendation:

1. A value below 50% indicates no risk of heart disease, and a value above 50% indicates a risk of heart disease.

2, The blue and red bars in the waterfall chart represent the factors influencing heart disease risk. The blue bars are health indicators, where higher blue values indicate lower heart disease risk. The red bars are unhealthy indicators, where higher red values indicate higher heart disease risk.

3. The first monitoring is recommended as the baseline value of heart disease risk monitoring. It is recommended to conduct regular monitoring on a monthly basis to control the non-health factors part of the waterfall chart, and medical intervention should be carried out when necessary to reduce the risk factors of heart disease.

4. The above report will be sent to the tester or the doctor provided by the tester by email. The above report is for monitoring reference only and does not serve as a diagnosis and treatment advice for heart disease.