Comprehensive Monitoring and Evaluation of Cardiovascular Disease Risk

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. In this study, eleven 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 in preventive healthcare.

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

Introduction

Hazards of cardiovascular disease (CVD)

Cardiovascular diseases (Cardiovascular Disease, CVD) are diseases affecting the heart or blood vessels, typically associated with fat deposition within arteries (atherosclerosis) and thrombosis (Table 1.1). 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].

Table 1.1 Four main types of cardiovascular disease (CVD). Types of CVD are defined by the location of fat deposition[1].

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	Poor circulation, slow healing wounds, pain in limbs
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, causing approximately 19.8 million deaths annually and accounting for 32% of all global deaths. Among these, 85% of deaths are due to heart disease or stroke[2]. 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[3].

The cardiovascular disease mortality rate in rural areas was 364.16 cases per 100,000 people in 2021[4]. 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.

Complexity of CVD and 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 global population has at least one risk factor for CVD, and only 2-7% has no risk factors at all[5]. The complex interactions among these risk factors, such as the synergistic interaction between hypertension and diabetes that exacerbates atherosclerosis, further amplify the probability of disease occurrence and the importance of improved risk assessment.

Since there is no minimum disease threshold that would trigger CVD, most events occur not in individuals with a single extreme CVD risk factor, but in people with several moderately elevated ones. 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 that may have been previously categorized as low-risk. Previous studies have also shown that comprehensive multi-factor intervention is more effective in reducing the incidence of CVD than the control of single risk factors[6]. Early intervention is considered the best strategy to prevent CVD, but there are many bottlenecks in the existing medical system (Table 1.2), illustrating the necessity of early identification and intervention.

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

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

AI in Healthcare

Artificial Intelligence (AI), especially machine learning, is emerging as a transformative tool in healthcare. By recognizing relationships in complex datasets and using them to predict future trends, machine learning becomes an effective tool for helping doctors efficiently analyze large amounts of patient data. Integrating AI into healthcare offers promising solutions for early detection, risk assessment, and intervention.

AI has been shown to predict the mortality of heart disease patients for the next 5 years with 80% accuracy[7], while the accuracy rate of clinical doctors' predictions is only 60%[8]. With the accumulation of open-source cardiac assessment data, 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).

Technical bottlenecks and opportunities in research

While AI has shown great potential in enhancing CVD diagnoses, the studies also revealed the algorithm's 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 model's generalization capability. Especially in scenarios involving complex nonlinear relationships, some linear models (like logistic regression) struggle to adequately capture the interactions between variables[9].

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. The study aims to address the gap in early diagnosis by reducing the high cost and inconvenience of traditional diagnostic methods through non-invasive means. It would also improve the efficiency of medical resource utilization by reducing the burden of delayed treatment and by promoting the development of personalized medicine through transparent, explanatory AI predictions using SHapley Additive exPlanations.

Methods

Artificial intelligence

Data processing

The data used in this study were obtained from the UCI Machine Learning Repository, collected by David Aha, and included participants ranging from age 29 to 71[10]. The dataset contains 303 patient heart assessment instances covering 14 attributes, including age, gender, type of chest pain, cholesterol levels, and resting ECG results (Table 2.1). After excluding 6 records with missing values, I was left with a dataset of 297 complete entries.

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 clinical testing and use shorthand labels, which reduces the model's transparency. Therefore, to ensure the reliability of the research results, the scope was narrowed down to these 10 attributes because they are most compatible with hardware and relevant for cardiac assessment. 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[9]

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

Feature processing was only applied to some models that benefit from it, such as linear models and some tree-based models. Categorical data are encoded using OneHot to enable the model to identify differences in discrete features; numerical data are standardized using StandardScaler to set each feature's mean to 0 and its 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 (Equation 1).

Equation 1

$$z=(x-\mu)$$

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.

Model development and evaluation

To verify the performance of different machine learning algorithms, ten commonly used classification models were selected, including logistic regression, support vector machine (SVM), KNN, random forest, decision tree, XGBoost, CatBoost, and LightGBM (Table 2.2). These models were chosen for their diverse algorithmic principles, including linear, non-linear, distance-based, and ensemble learning approaches. All models were trained on the same dataset and evaluated using the same metrics to ensure fairness and comparability.

Model training employs 10-fold cross-validation, and hyperparameters are optimized through grid search to ensure model robustness. The dataset was divided into training and testing sets in an 8:2 ratio. The training set, comprising 80% of the UCI Cleveland dataset, was used to train the model to recognize patterns in the data, while the remaining data, labelled as the testing set, was used to evaluate the model's ability to handle unseen 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 neural network)	non-linear model	High-dimensional complex data such as images, voice and text	Can deal with complex high-dimensional data, good at nonlinear, multi-dimensional samples
decision tree	Tree-based approach	Classification and regression are suitable for	Easy to explain, by recursively dividing the data

		handling various types of features	
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

Prediction models performance

To evaluate the performance of different machine learning models in cardiovascular disease (CVD) risk prediction, this study systematically compared the 10 algorithms' performance through AUC, precision, accuracy, recall, and F-1 score (Table 3.1). The study prioritized AUC because it assesses the model's consistency in consistently ranking patients with CVD at higher risk than those without, regardless of the disease threshold. This is important because different hospitals may use different definitions of a cardiovascular disease threshold, depending on clinical priorities, such as minimizing false negatives to avoid missed diagnoses. Thus, the AUC of a model can still reflect its performance comprehensively even when the disease threshold is changed.

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.	Accuracy=TP+TNTotal
Accuracy	Reflect the reliability of positive class prediction, that is, the proportion of true positive in positive prediction.	Precision=TPTP+FP
Recall	It reflects the coverage ability of positive samples, that is, the proportion of correct prediction of all positive samples.	Recall=TPTP+PN

F1-score	The harmonic mean of accuracy and recall rate comprehensively reflects the performance of the model.	F1=2·Precision·RecallPrecision+Recal

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.

Hardware equipment design and integration

Microcontroller 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, allowing 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. 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.

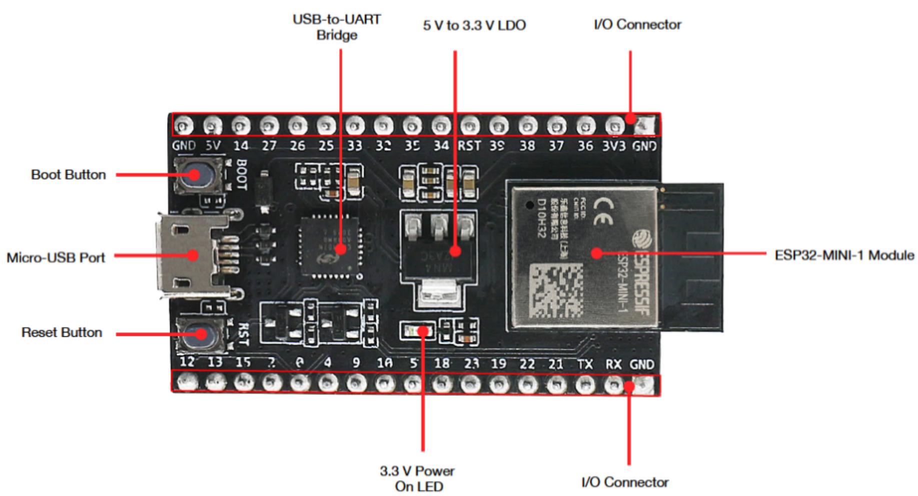

Figure 2.1.1 ESP32 image

Electrocardiogram Sensor 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. As an operational amplifier, it amplifies clear signals from the PR interval and QT interval. It filters out unwanted small bioelectric potential signals, extracting diagnostic signals effectively in a noisy environment.

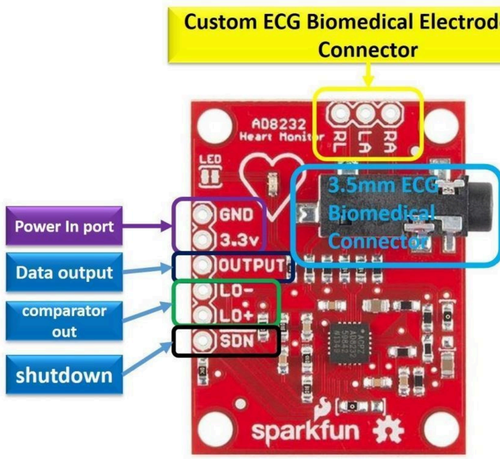

Figure 2.1.2 Schematic diagram of the labeling of the SparkFun AD8232, heart rate monitoring module

In this study, the AD8232 device collects ST-segment data from patients and identifies potential cardiac abnormalities by detecting increases or decreases in the ST segment. The ST-segment reflects ventricular recovery on an ECG, with abnormal elevation or depression signalling myocardial ischemia or damage. The sensor transmits the collected data to the ESP32 for further processing, ensuring the accuracy and real-time nature of the ECG signals.

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

The MKS-5V45-HRV-FP (MKS-FP) is a high-precision sensor for accurate measurement of physiological parameters such as pulse waveform, heart rate value, blood oxygen, etc. This sensor has high sensitivity and noise resistance, allowing it to distinguish the actual signal from noise with high accuracy. In this research, the sensor will be used to measure heart rate and blood pressure. The physiological signals from MKS-FP will be converted into electrical signals and transmitted to the ESP32, enabling real-time data processing.

Figure 2.1.3 MKS-5V45-HRV-FP image

Communication protocol (MQTT)

This system adopts MQTT as the core communication protocol due to 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.

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.

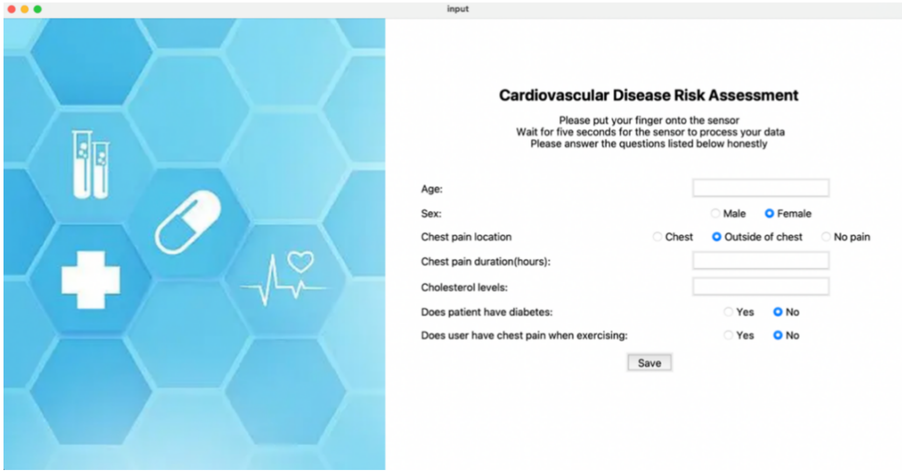

Figure 2.1.4 User Interface Image

The patient will receive an email informing them of their cardiovascular risk the waterfall chart generated by SHAP that interprets the algorithm's prediction. The email provides guidance on how to read 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.

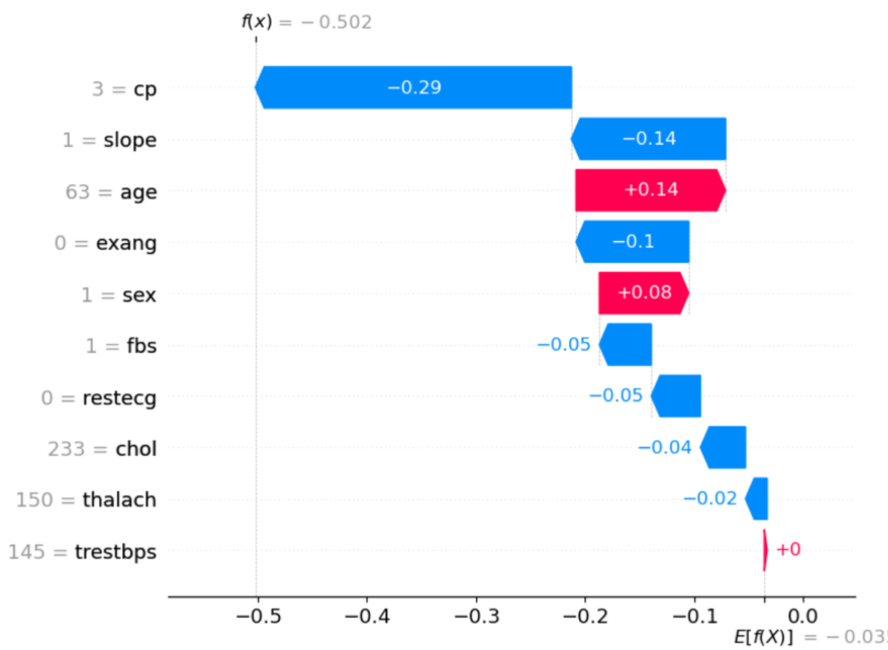

Figure 2.1.5 SHAP (Example Shapley Additive waterfall plot chart)

Individualized interpretation

The impact of features on outcomes is ranked from important to unimportant, with features on the top being most significant. The SHAP values demonstrate the effect of features (risk factors) on the final outcome. A waterfall chart is generated to illustrate the model's decision-making process.

The Y-axis shows the features (chest pain, ST-segment slope, age, exercise angina, sex, fasting blood sugar, resting ECG, cholesterol, maximum heart rate, resting blood pressure) in the dataset ranked by their average contribution to the final outcome. The X-axis shows the predicted risk value, which indicate the extent to which each feature influences changes in risk predictions. 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). A value below 50% indicates no risk of heart disease, and a value above 50% indicates a risk of heart disease. Such visualizations can help doctors understand the specific sources of patient risk, thereby providing targeted intervention recommendations.

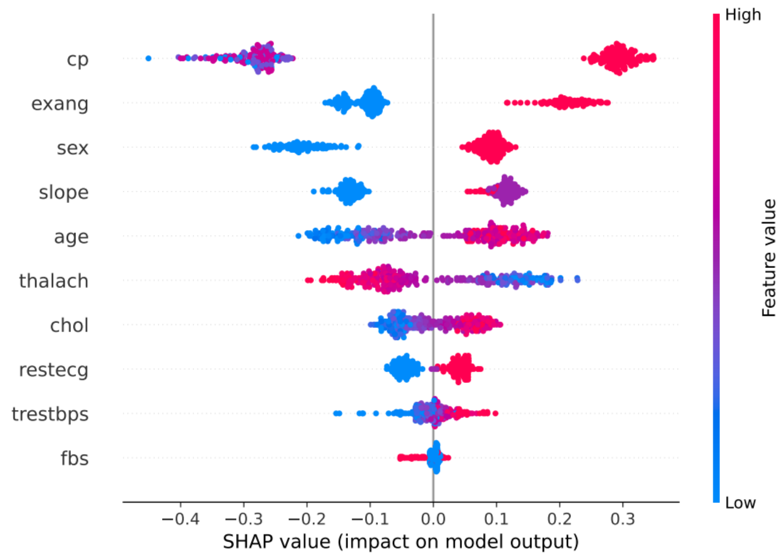

Figure 2.2.1 Shape feature importance diagram

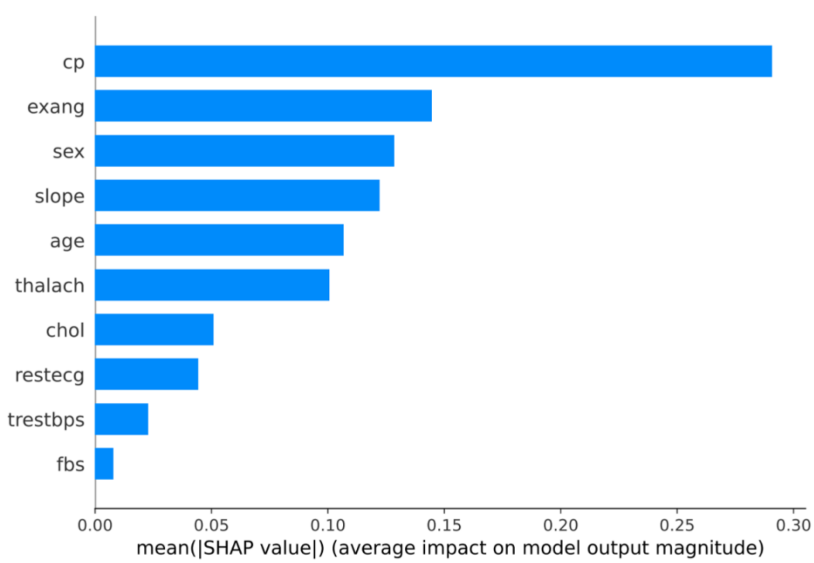

Figure 2.2.2 Average values of the influence of each feature in model prediction (|Shap value |).

Fig. 2.2.2 shows the average influence of each feature on model predictions. 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.

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.

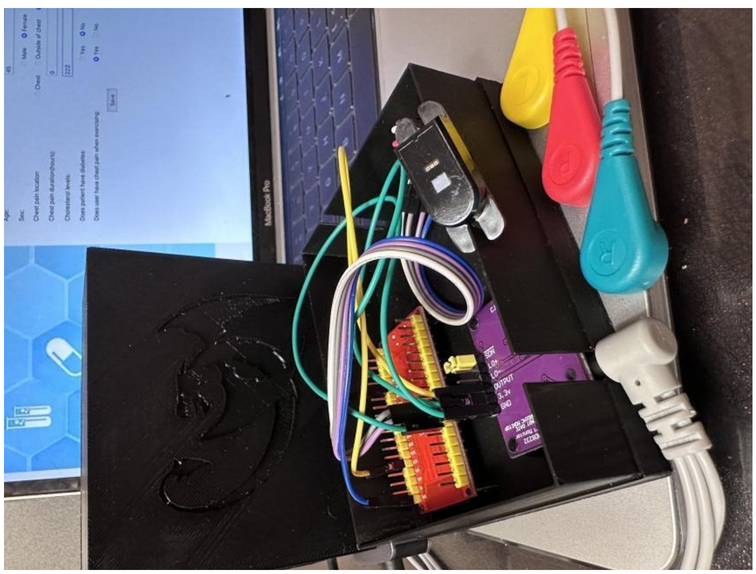

Figure 2.2.3, Prototype of the equipment

System testing and practical verification

Beijing Guang anmen Hospital test

For performance validation 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 2.3.1, Dr. Zhang Jingru

Figure 2.3.2 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.

Testing of non-medical institutions' processes

Subsequently, five other adults （2 male and 3 female） who were not CVD patients were used as a control variable. After a brief introduction of the process and explanation of the device's purpose, they signed an informed consent form and underwent the test.

The users are first asked to attach the ECG patches to the top left, top right, and bottom right of their torso (shown in the three-lead ECG diagrams). Simultaneously, the user would place their finger on the MKS-FP sensor on the device, which collects heart rate and blood pressure data. The user would enter the rest of their personal information (age, gender, chest pain type, etc.) through the user interface. After they click submit, the algorithm would receive the processed data and generate CVD risk assessment results. This risk prediction and a SHAP interpretation diagram will be sent to the user's email.

Results

Algorithm Performance

AUC-ROC (AUC) value of CatBoost is the highest, at 0.936 (Figure 3.1.1). The AUC values for Logistic regression, SVM, KNN, DNN, Random Forest, Decision Tree, XGBoost, CatBoost, LightGBM, and Gradient Boosting are 0.92667, 0.904, 0.900,0.902,0.897,0.848, 0.885, and 0.874 respectively (Table 3.2).

Overall, the CatBoost model's performance was the most optimal with the AUC score's difference between training and testing to be only 0.002, 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 is relatively high in other metrics as well, having a score of 0.813, 0.750, 0.783, and 83.33% in recall, precision, F1, and test accuracy (Table 3.2).

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%

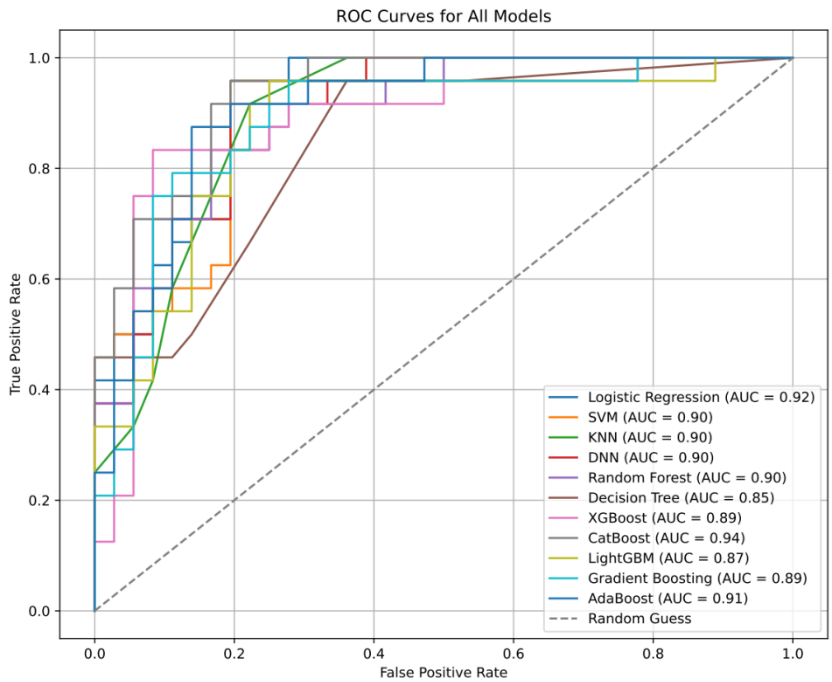

Figure 3.1.1 ROC curves for all models

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, XGBoost and gradient-boosting models perform well, are suitable for handling nonlinear data and complex features, and are widely used.

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 (0.973), the AUC on the test set drops significantly (0.897), 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 was selected as the core prediction algorithm due to its efficient classification performance and robustness, and is used for subsequent equipment development.

Patient Trials

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	

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 model' s prediction. This could explain the model’s 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.

Discussion

Summary of research

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.

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. **E. O. Lopez, B. D. Ballard, A. Jan.** "Cardiovascular Disease." *StatPearls – NCBI Bookshelf*, 22 Aug. 2023. https://www.ncbi.nlm.nih.gov/books/NBK535419/.
2. **World Health Organization.** *Cardiovascular Diseases (CVDs)*. 31 July 2025, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

3. **Zhao, Dong.** “Epidemiological Features of Cardiovascular Disease in Asia.” *JACC Asia*, vol. 1, no. 1, June 2021, pp. 1–13. https://doi.org/10.1016/j.jacasi.2021.04.007.
4. **Su, Shuyao, and Fangchao Liu.** “Cardiovascular Health and Disease Report in China: Two Decades of Progress.” *Biomedical and Environmental Sciences*, Aug. 2025. https://doi.org/10.3967/bes2025.098.
5. **Dahlöf, Björn.** “Cardiovascular Disease Risk Factors: Epidemiology and Risk Assessment.” *The American Journal of Cardiology*, vol. 105, no. 1, Dec. 2009, pp. 3A–9A. https://doi.org/10.1016/j.amjcard.2009.10.007.
6. **Levine, Jon H.** “Managing Multiple Cardiovascular Risk Factors: State of the Science.” *Journal of Clinical Hypertension*, vol. 8, no. s10, Oct. 2006, pp. 12–24. https://doi.org/10.1111/j.1524-6175.2006.05924.x.
7. **A. A. Mahayni, Z. I. Attia, J. R. Medina-Inojosa, M. F. A. Elsisy, P. A. Noseworthy, F. Lopez-Jimenez, S. Kapa, S. J. Asirvatham, P. A. Friedman, J. A. Crestenallo, M. Alkhouli.** “Electrocardiography-Based Artificial Intelligence Algorithm Aids in Prediction of Long-term Mortality After Cardiac Surgery.” *Mayo Clinic Proceedings*, vol. 96, no. 12, Dec. 2021, pp. 3062–70. https://doi.org/10.1016/j.mayocp.2021.06.024.
8. **Bösner, Stefan, J. Haasenritter, M. Abu Hani, H. Keller, A. C. Sönnichsen, K. Karatolios, J. R. Schaefer, E. Baum, N. Donner-Banzhoff.** “Accuracy of General Practitioners’ Assessment of Chest Pain Patients for Coronary Heart Disease in Primary Care: Cross-sectional Study With Follow-up.” *Croatian Medical Journal*, June 2010. https://doi.org/10.3325/cmj.2010.51.243.
9. **Y. Q. Cai, D. X. Gong, L. Y. Tang, Y. Cai, H. J. Li, T. C. Jing, M. Gong, W. Hu, Z. W. Zhang, X. Zhang, G. W. Zhang.** “Pitfalls in Developing Machine Learning Models for Predicting Cardiovascular Diseases: Challenge and Solutions.” *Journal of Medical Internet Research*, vol. 26, June 2024, p. e47645. https://doi.org/10.2196/47645.
10. **Latha, C. Beulah Christalin, and S. Carolin Jeeva.** “Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques.” *Informatics in Medicine Unlocked*, vol. 16, 2019, p. 100203. https://doi.org/10.1016/j.imu.2019.100203.