



## OPEN ACCESS

## EDITED BY

Chunyu Wang,  
Harbin Institute of Technology, China

## REVIEWED BY

Ran Su,  
Tianjin University, China  
Yongqing Zhang,  
Chengdu University of Information Technology,  
China

## \*CORRESPONDENCE

Zhen Li,  
✉ lizhen5000@szit.edu.cn

RECEIVED 24 November 2025

REVISED 18 December 2025

ACCEPTED 26 December 2025

PUBLISHED 21 January 2026

## CITATION

Zhao S, Nan B, Guo J, Xu W and Li Z (2026)  
Coronary heart disease risk prediction based on  
GAIN imputation and interpretable  
machine learning.  
*Front. Genet.* 16:1752811.  
doi: 10.3389/fgene.2025.1752811

## COPYRIGHT

© 2026 Zhao, Nan, Guo, Xu and Li. This is an  
open-access article distributed under the terms  
of the [Creative Commons Attribution License](#)  
(CC BY). The use, distribution or reproduction in  
other forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in this  
journal is cited, in accordance with accepted  
academic practice. No use, distribution or  
reproduction is permitted which does not  
comply with these terms.

# Coronary heart disease risk prediction based on GAIN imputation and interpretable machine learning

Shulin Zhao<sup>1</sup>, Baoyun Nan<sup>1</sup>, Jun Guo<sup>1</sup>, Wenkai Xu<sup>1</sup> and Zhen Li<sup>2\*</sup>

<sup>1</sup>The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou, China, <sup>2</sup>School of Artificial Intelligence, Shenzhen University of Information Technology, Shenzhen, China

**Introduction:** Coronary atherosclerotic heart disease (CHD) is a leading cause of morbidity and mortality worldwide, making timely identification critical for improving patient prognosis. However, traditional imaging examinations are limited by high costs and patient selection bias, while existing prediction models often lack interpretability and generalization ability. This study aimed to develop a robust, interpretable machine learning approach to address these challenges.

**Methods:** This retrospective study analyzed hospitalized patients at Quzhou People's Hospital from July 2021 to March 2025. Patients diagnosed with CHD were categorized as positive samples, while those without cardiovascular disease served as negative controls. The dataset integrated demographic data, blood biomarkers, and vital signs. A Generative Adversarial Imputation Network (GAIN) was utilized to handle missing values, and multiple machine learning models were constructed and compared for prediction performance.

**Results:** Among the evaluated algorithms, the XGBoost model achieved superior performance on the test set with an Area Under the Curve (AUC) of 0.9053. To enhance clinical utility, the integration of SHAP (SHapley Additive exPlanations) values enabled both global and local interpretation of model decisions. Key predictive factors identified included mean respiratory rate during hospitalization, age, high-sensitivity troponin I (hs-cTnI), and hypertension.

**Discussion:** The developed model demonstrates robust prediction performance combined with high clinical interpretability. Unlike traditional "black box" models, this approach clarifies the contribution of specific risk factors. Crucially, the tool is well-suited for dual deployment: serving as an automated screening tool integrated into hospital electronic health records (EHRs) and as a self-monitoring aid for individuals with underlying health conditions via mobile health applications.

## KEYWORDS

coronary heart disease, disease prediction, explainable machine learning, feature fusion, GAIN

# 1 Introduction

Coronary atherosclerotic heart disease (CHD) is one of the most prevalent and deadliest diseases worldwide (Yang et al., 2023). It is characterized by the narrowing or occlusion of the coronary artery lumen. The deleterious effects of CHD are progressive and potentially lethal, manifesting as a spectrum from arrhythmias and angina pectoris to myocardial infarction and heart failure. CHD significantly compromises patients' life expectancy and quality of life while imposing a substantial economic burden on families and society (Colantonio et al., 2017; Ladak et al., 2020; Pickles and Keller, 2025).

Beyond therapeutic management, effective risk prediction is crucial, enabling timely intervention and preventative measures. Disease prediction is a continuous spectrum that includes both anticipation of future patients and screening of patients who are currently ill but have not been detected. For chronic and often insidious conditions like CHD, the latter is particularly important (Koloi et al., 2024). In hospitalized populations, undetected occult CHD significantly elevates perioperative risks—especially during non-cardiac surgeries—thereby severely impacting prognosis and exacerbating medical burdens. Additionally, for the general population with underlying conditions such as hypertension and diabetes, the occult nature of CHD makes it difficult to detect through routine self-examinations, often leading to delayed diagnosis until severe cardiovascular events occur, causing patients to miss the critical window for early intervention (Sawaf et al., 2024; Zaninotto et al., 2024).

Although imaging techniques such as computed tomography angiography (CTA) and invasive coronary angiography (ICA) can assess the degree of coronary artery stenosis and plaque burden, their widespread clinical application is constrained by high costs, operator dependency, and selection bias (Min et al., 2022; Xiong et al., 2024). Usually, only patients with a high clinical suspicion of disease undergo these expensive or radiation-intensive procedures. This means that there is a severe lack of healthy but slightly abnormal samples and atypical symptoms cases in the imaging patient dataset. Conversely, biomarkers derived from routine blood tests offer a non-invasive, cost-effective, and scalable evaluation method accessible at all levels of healthcare (Sanchez-Morillo et al., 2024). Combining personal basic information (gender, age, etc.) with easily accessible vital sign information (blood pressure, blood oxygen saturation SpO<sub>2</sub>, body temperature, etc.) of smart wearable devices can identify high-risk individuals for diseases earlier and more widely (Kundrick et al., 2025; Nenova and Shang, 2022).

Although machine learning or deep learning driven models can improve prediction performance, they often lack interpretability due to their “black box” nature, which cannot clearly reveal the correlation mechanism between risk factors and disease probability (Topranin et al., 2025; Li et al., 2021; Liu et al., 2019), limiting clinical doctors' trust in prediction results and the development of personalized intervention strategies. Although traditional models such as Framingham risk score have some interpretability, they have problems such as insufficient prediction accuracy and weak generalization ability, making it difficult to meet the current needs of precision medicine (Rehman et al., 2025).

Therefore, this research constructed a specific group of non-cardiovascular disease hospitalized patients as negative samples, combined with their personal basic information, blood biomarkers, and vital sign information, to construct an efficient and stable interpretable model for predicting CHD risk. This model not only predicted the probability of individual disease risk, but also clearly explained the specific contributions of various risk factors to the prediction results. The framework flowchart shown in Figure 1 illustrates the comprehensive process from data collection to clinical interpretation. This approach aims to provide intuitive basis for clinical doctors to understand the mechanism of disease association and formulate personalized intervention strategies, and to provide low-cost and easy to promote practical tools for independent heart health monitoring in populations with underlying conditions.

## 2 Methods

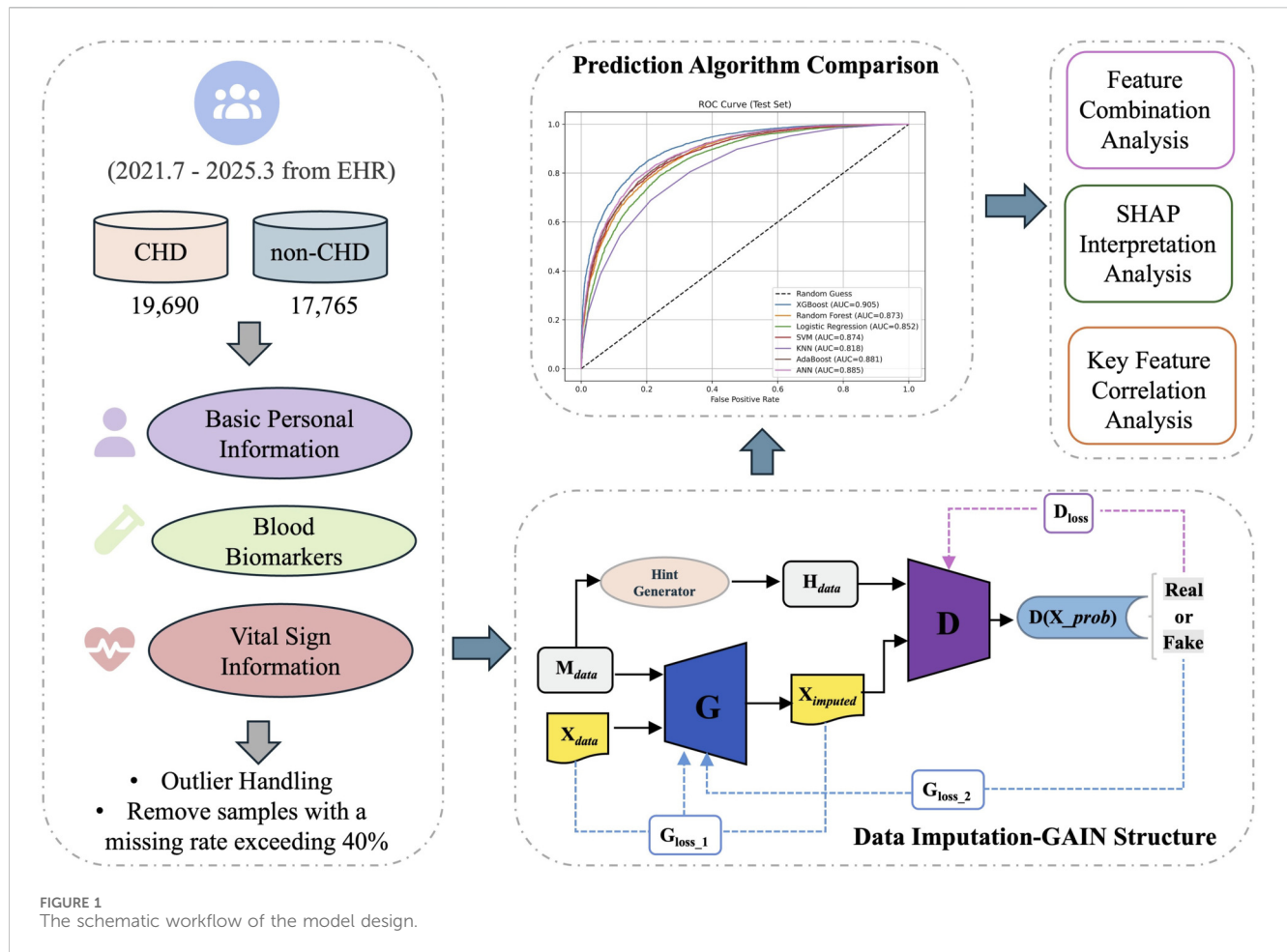
### 2.1 Research population design

We retrospectively enrolled hospitalized patients at Quzhou People's Hospital from July 2021 to March 2025. The condition for positive sample collection is based on patients diagnosed as coronary atherosclerotic heart disease after discharge and whose length of stay is  $\geq 2$ , excluding patients with cancer/tumor. A total of 19,690 eligible patients with CHD were included. The negative sample set comprised patients without a discharge diagnosis of cardiovascular-related diseases, hospitalized for  $\geq 2$  days, excluding patients with cancer/tumors. Ultimately, 17,765 eligible non-cardiovascular disease (non-CHD) patients were included.

Utilizing a healthy population as a control often causes the model to learn merely the generalized differences between ill patients and healthy individuals, rather than the specific pathological features that distinguish CHD from other diseases. Consequently, applying such a model to patients with existing comorbidities often results in unacceptable false positive rates, diminishing the clinical utility of the predicted results. To achieve the goal of disease screening within medical institutions and self-screening among individuals with underlying conditions, this research innovatively used other hospitalized patients with non-cardiovascular diseases as negative controls.

### 2.2 Data variables and preprocessing

The dataset comprises three variable categories: demographic characteristics, blood biomarkers, and vital signs, all extracted from electronic health records (EHR). The basic personal information includes the patient's gender, blood type, and age; lifestyle factors included smoking and drinking status; and comorbidities included diabetes and hypertension. Age was recorded at the time of treatment; smoking and drinking status were obtained from medical history records; and diabetes and hypertension status were derived from discharge diagnoses. Blood biomarkers, derived from initial admission tests, included complete blood



counts (CBC), biochemical indicators (e.g., liver and kidney function), and D-dimer levels, among others. The vital sign information included the patient's initial admission temperature, heart rate, respiratory rate, systolic blood pressure (SBP), and diastolic blood pressure (DBP), and SpO<sub>2</sub>. Additionally, the maximum, minimum, and mean values of SBP, DBP, body temperature, respiratory rate and SpO<sub>2</sub> measured during hospitalization were recorded.

In laboratory testing, sample quality issues caused by hemolysis, instrument errors, or other factors can produce extreme outliers. As these outliers do not reflect true physiological or pathological states, rigorous detection and cleaning were performed on the blood biomarker data. We adopted a modified Z-score method to identify outliers, which is more robust to outliers (Kuo et al., 2024).

$$MAD = \text{median}(|X_i - X_m|)$$

$$Z\text{-score} = 0.6745 \times (X_i - X_m) / MAD$$

Among them,  $X_i$  is the sample feature value,  $X_m$  is the median of the sample feature value, and  $MAD$  is the median absolute deviation. Values with a Z-score > 3.5 were identified as outliers and replaced with null values (NaN). Further screening was conducted on samples with missing values below 40%, while retaining samples with more valid data. Consequently, 12 CHD and 7 non-CHD samples were removed.

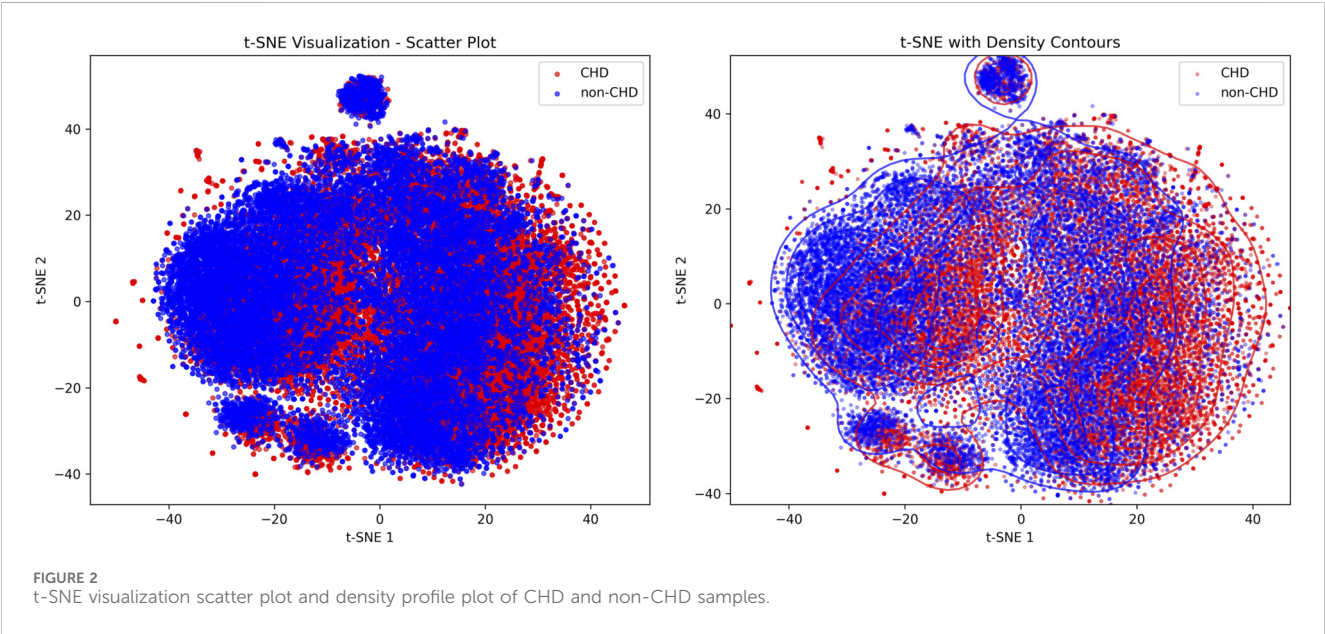
The partial continuous value feature names, abbreviations, units, distribution descriptions (mean, standard deviation), and missing rates on the positive and negative sample sets are shown in the Table 1. Given the high dimensionality of the dataset, the complete feature table is included in the Supplementary Material.

## 2.3 GAIN architecture

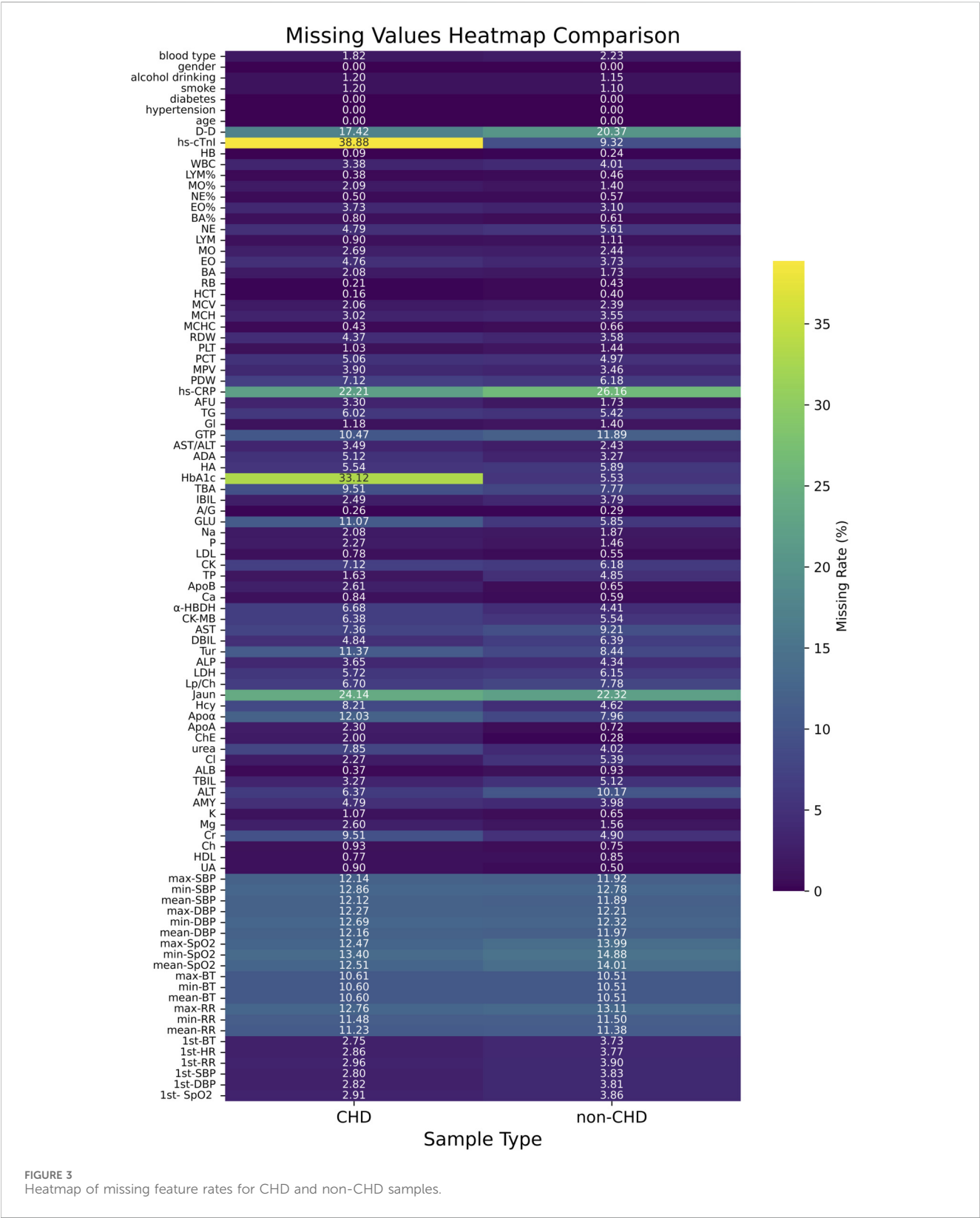
Generative adversarial imputation Nets (GAIN) represent a data imputation method based on generative adversarial networks (GANs) (Xu et al., 2025; Nayak et al., 2024). By leveraging adversarial training between a generator and a discriminator, GAIN learns the underlying data distribution to generate plausible imputed values. The architecture is illustrated in Figure 1. The generator, which serves as the core component for imputation, utilizes a three-layer fully connected neural network structure. The input consists of the original data tensor  $X_{data}$  containing missing values and the mask tensor  $M_{data}$ , where 1 denotes observed data and 0 denotes missing data. The discriminator is tasked with distinguishing between observed true values and imputed values produced by the generator; its network structure mirrors that of the generator. The discriminator input comprises a concatenated tensor of the imputed data  $X_{imputed}$  and

TABLE 1 Description of partial features.

Feature	Abbreviation	Unit	CHD			Non-CHD		
			Mean	std	Missing_rate (%)	Mean	std	Missing_rate (%)
Age	Age	years	70.81	10.95	0	60.96	14.33	0
D-dimer	D-D	mg/L FEU	0.5	0.38	17.42	0.47	0.37	20.37
High-sensitivity troponin I	Hs-cTnI	μg/L	0.005	0.0045	38.88	0.003	0.0037	9.32
Hemoglobin	HB	g/L	120.73	21.76	0.09	122.07	21.27	0.24
White blood cell count	WBC	*10 <sup>9</sup> /L	6.25	2.1	3.38	6.25	2.22	4.01
Lymphocyte percentage	LYM%	%	22.2	9.64	0.38	23.63	10.89	0.46
Monocyte percentage	MO%	%	8.22	2.63	2.09	7.84	2.64	1.4
Neutrophil percentage	NE%	%	66.57	11.46	0.5	65.77	12.78	0.57
Eosinophil percentage	EO%	%	1.99	1.6	3.73	1.8	1.56	3.1
Basophil percentage	BA%	%	0.41	0.26	0.8	0.4	0.27	0.61
Neutrophil count	NE	*10 <sup>9</sup> /L	4.12	1.72	4.79	4.06	1.86	5.61
Lymphocyte count	LYM	*10 <sup>9</sup> /L	1.32	0.57	0.9	1.4	0.6	1.11
Monocyte count	MO	*10 <sup>9</sup> /L	0.51	0.2	2.69	0.48	0.2	2.44
Eosinophil count	EO	*10 <sup>9</sup> /L	0.11	0.09	4.76	0.1	0.09	3.73
Basophil count	BA	*10 <sup>9</sup> /L	0.02	0.02	2.08	0.02	0.02	1.73



the hint vector  $H_{data}$ . The hint vector, derived from a randomly generated probability matrix and a mask tensor, provides auxiliary information to assist the discriminator in identifying missingness patterns. The generator loss is composed of a weighted adversarial loss  $G_{loss\_1}$  and an MSE loss  $G_{loss\_2}$ , where the adversarial loss is achieved by minimizing the discriminator's recognition accuracy of the generated values, and the MSE loss constrains the generator to not destroy the original information at known data positions. Optimize the discriminative ability of the discriminator by calculating the classification loss  $D_{loss}$  between the original real data and the generated imputed data. The entire model gradually learns the inherent distribution pattern of the data through a continuous adversarial game between the generator and discriminator, ultimately generating missing values that are close to the true distribution.





## 2.4 Evaluation metrics

To comprehensively evaluate the model’s predictive capability and clinical utility, we employed multiple complementary metrics, including accuracy, precision, recall, F1-score and Area Under the Receiver Operating Characteristic Curve (AUC) (Chen et al., 2025a; Kumar et al., 2024; Rimal and Sharma, 2023; Chen et al., 2025b; Qiao et al., 2024). These are the core metrics for evaluating the performance of binary classification models, calculated based on four fundamental values in the confusion matrix: true positive cases (TP), false positive cases (FP), true negative cases (TN), and false negative cases (FN) (Zeng et al., 2025; Zulfiqar et al., 2024; Qiao et al., 2025; Xie et al., 2025; Wang et al., 2025; Wang et al., 2024).

$$ACC = (TP + TN) / (TP + TN + FP + FN)$$
$$Precision = TP / (TP + FP)$$
$$Recall = TP / (TP + FN)$$
$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

## 3 Discussion and results

### 3.1 Analysis of data distribution and feature missing

Figure 2 showed the visualization results of positive and negative samples using t-distributed stochastic neighbor embedding (t-SNE), which was used to display the distribution patterns of CHD and non-CHD samples in a high-dimensional feature space, intuitively presenting the feature differences and clustering patterns of the two groups (Nollmann et al., 2024).

In the scatter plot, blue (non-CHD) samples form a core cluster and two independent small clusters, indicating that the characteristics of the non-CHD population have strong concentration. The red (CHD) samples are interspersed in the form of scattered dots within and at the edges of the blue clusters, with only mild clustering in local areas and a relatively scattered overall distribution. In the density contour map, the density contour of non-CHD samples covers most of the areas in the map, and the core area has a high density, further verifying the concentration of non-CHD population characteristics and the stability of subgroup structure. The density profile of CHD samples highly overlaps with non-CHD, with only weak independent trends in local areas, indicating that the characteristic boundaries between CHD and non-CHD are blurred and overlap is high. At the same time, the heterogeneity of CHD features, such as different disease courses, subtypes, and comorbidities leads to their scattered distribution.

Figure 3 showed the heatmap of feature missing rates for the CHD and non-CHD groups. The missing rates for most variables were similar in both groups, but significant differences existed in key clinical indicators. The CHD group had significantly higher missing rates for hs-cTnI (38.88%) and HbA1c (33.12%) than the non-CHD group (9.32% and 5.53%, respectively). This difference may reflect insufficient detection of these important diagnostic and monitoring indicators in patients with CHD in clinical practice. In contrast, the non-CHD group had slightly higher missing rates for indicators such as D-dimer (20.37%) and hs-CRP (26.16%). The missing rates

TABLE 2 Comparison of 5-cv performance of imputation algorithms.

Method	Accuracy	Precision	Recall	F1	AUC
KNN	0.8208	0.8206	0.8198	0.8201	0.9016
MICE	0.8265	0.8263	0.8256	0.8259	0.9078
AE	0.8322	0.8321	0.8313	0.8316	0.9146
DAE	0.8350	0.8349	0.8339	0.8343	0.9157
VAE	0.8388	0.8385	0.8382	0.8383	0.9196
GAIN	0.8343	0.8342	0.8333	0.8336	0.9154

TABLE 3 Comparison of independent testing performance of imputation algorithms.

Method	Accuracy	Precision	Recall	F1	AUC
KNN	0.8096	0.8145	0.8261	0.8202	0.8946
MICE	0.8234	0.8256	0.8420	0.8337	0.9043
AE	0.8285	0.8281	0.8503	0.8390	0.9095
DAE	0.8317	0.8342	0.8486	0.8413	0.9135
VAE	0.8350	0.8381	0.8504	0.8442	0.9165
GAIN	0.8354	0.8405	0.8477	0.8441	0.9156

for most routine laboratory indicators and vital signs remained below 10% in both groups, indicating relatively complete basic clinical data collection. However, key indicators with high missing rates require appropriate missing data processing strategies in subsequent analyses.

### 3.2 Performance analysis of imputation models

Although incomplete recorded data may be reasonable in clinical practice, the performance of machine learning algorithms is often affected by biased and incomplete data. Medical record data is extremely valuable for disease research. If partially missing samples are directly removed and models are constructed using non missing samples, although this approach is simple, it wastes a lot of available information.

This research comprehensively compared and analyzed the performance of traditional imputation algorithms (KNN, MICE) (Varol et al., 2025), deep learning autoencoder series (AE, DAE, VAE) (Gautier et al., 2024; Shi et al., 2024), and GAIN in medical record data. Due to the lack of real data references, the performance of the downstream tasks is generally taken as the standard. Table 2 compared the 5-fold cross validation (5-cv) performance of the imputation algorithms, and Table 3 compared its performance on the independent test set. The performance metrics of traditional imputation algorithms were significantly lower than those of deep learning methods, and they were limited to being unable to adapt to the complex nonlinear correlations between features in the data, resulting in insufficient expression ability in high-dimensional medical data scenarios.

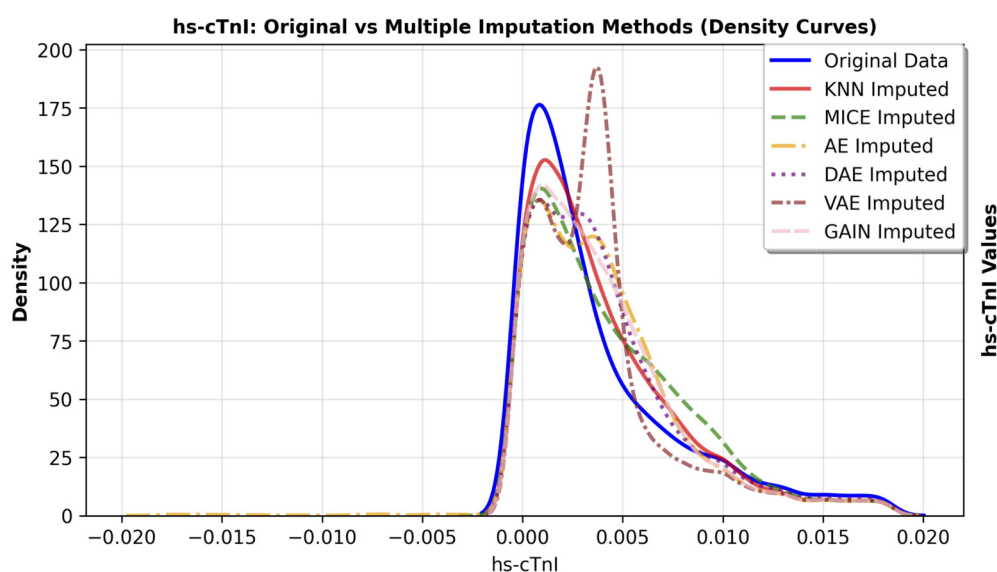


FIGURE 4  
Density curve of feature hs-cTnI distribution before and after imputation.

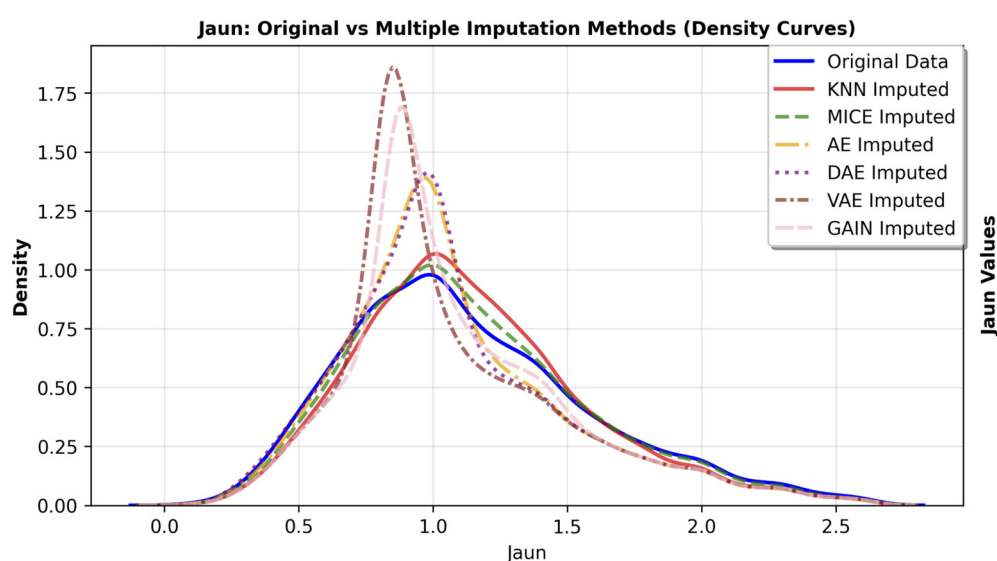


FIGURE 5  
Density curve of feature Jaun distribution before and after imputation.

As a basic autoencoder, AE achieved a 5-cv AUC of 0.9146 and an independent test AUC of 0.9095, which preliminarily demonstrated the modeling ability of deep learning for complex data. DAE enhanced robustness through noise reduction mechanism, further improving performance (5-cv AUC 0.9157, independent test AUC 0.9135), and had a higher tolerance for data noise. After introducing variational inference, VAE had better flexibility in distribution modeling, with a 5-cv AUC 0.9196 and independent testing AUC 0.9165, ranking among the top in multiple performance metrics. As a generative model, GAIN not only considered the distribution of individual features in adversarial learning, but also

comprehensively considers the complex correlations between all other features. Its 5-cv AUC 0.9154 and independent test AUC 0.9156 were slightly lower than VAE.

The primary goal of imputation in medical research is to faithfully preserve the original data distribution and minimize bias. We further observed the fitting degree of each imputation algorithm on the data distribution, and selected hs-cTnI and Jaun features with high missing rates. The feature density curves before and after imputation were shown in Figures 4, 5. Compared to VAE, the data distribution after GAIN imputation has a higher degree of fit with the original data in terms of morphology, which can avoid

TABLE 4 Comparison of performance metrics of different prediction algorithms on the training set.

Model	Accuracy	Precision	Recall	F1	AUC
XGBoost	0.8345	0.8346	0.8345	0.8344	0.9184
Random forest	0.7999	0.8019	0.7999	0.7989	0.8874
Logistic regression	0.7789	0.7788	0.7789	0.7787	0.8582
SVM	0.8108	0.8113	0.8108	0.8104	0.8923
KNN	0.7963	0.8004	0.7963	0.7963	0.8844
AdaBoost	0.8079	0.8078	0.8079	0.8078	0.8911
ANN	0.8381	0.8382	0.8381	0.8379	0.9182

TABLE 5 Comparison of performance metrics of different prediction algorithms on the test set.

Model	Accuracy	Precision	Recall	F1	AUC
XGBoost	0.8246	0.8247	0.8246	0.8246	0.9053
Random forest	0.7868	0.7888	0.7868	0.7868	0.8732
Logistic regression	0.7750	0.7749	0.7750	0.7750	0.8517
SVM	0.7919	0.7922	0.7919	0.7919	0.8744
KNN	0.7356	0.7411	0.7356	0.7356	0.8177
AdaBoost	0.7976	0.7976	0.7976	0.7976	0.8811
ANN	0.8035	0.8034	0.8035	0.8035	0.8854

additional bias caused by distribution offset due to imputation and ensure the authenticity and effectiveness of feature information in subsequent analysis. To strictly quantify this observation, we used Kullback-Leibler (KL) divergence and Kolmogorov-Smirnov (KS) test statistics for the imputed data

of VAE and GAIN on independent test sets. A lower KL or KS value indicates a distribution closer to the ground truth., GAIN achieved the lowest mean KL divergence (0.158) and mean KS statistic (0.067) outperforming VAE (KL: 0.210; KS: 0.068). This statistical evidence demonstrates that GAIN is superior in capturing the complex underlying probability distribution of the real data, avoiding the distributional shifts often introduced by variational inference in VAEs. Consequently, considering both the robust downstream performance and the superior data fidelity, GAIN was selected as the optimal imputation algorithm for this research.

### 3.3 Comparison of prediction algorithm performance

In this section, we conducted a performance comparison analysis of XGBoost, random forest, logistic regression, SVM, KNN, AdaBoost, and ANN algorithms. The prediction metrics of training set in Table 4 showed that XGBoost (AUC = 0.9184, Accuracy = 0.8345) and ANN (AUC = 0.9182, Accuracy = 0.8381) have the most outstanding comprehensive performance. In the test set prediction performance in Table 5, XGBoost exhibited excellent generalization stability, while ANN’s generalization ability is significantly insufficient. Figure 6 showed the comparison of ROC curves of different prediction algorithms on the training and testing sets, which intuitively proves that the XGBoost model had the strongest ability to distinguish positive and negative samples and excellent generalization.

### 3.4 Feature combination analysis

In many research, derived indicators based on blood biomarkers have shown excellent performance. We have established 8 derived indicators according to the obtained blood

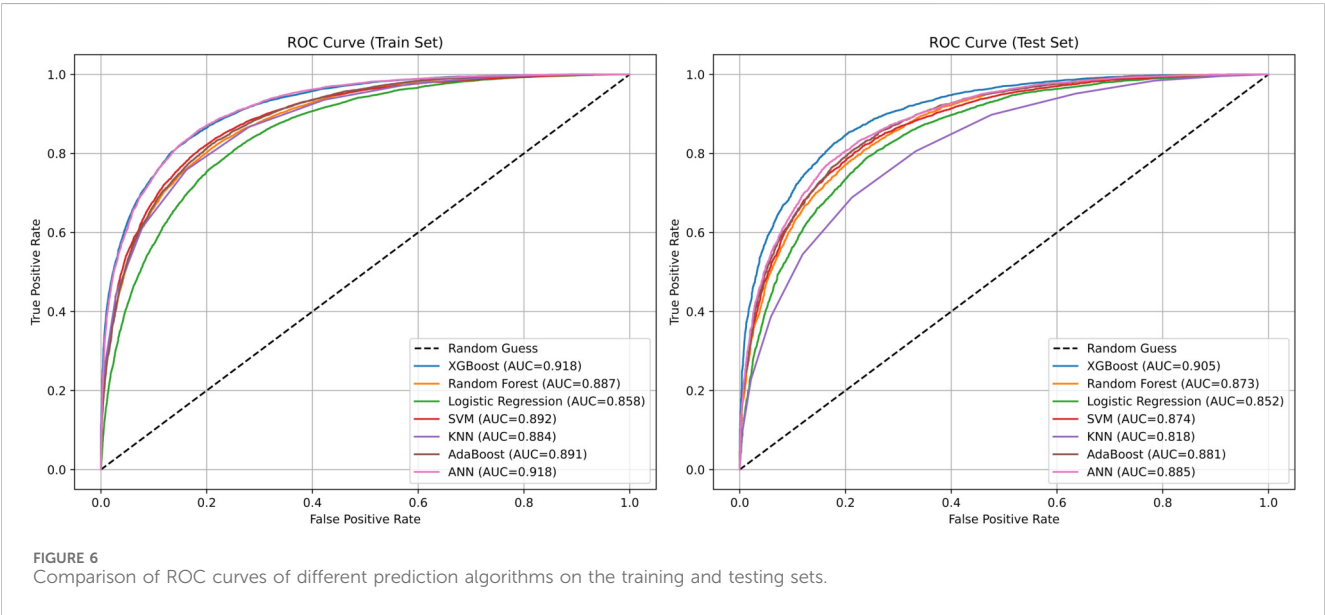




TABLE 6 Performance comparison of different feature combinations on the test set.

Feature set	Accuracy	Precision	Recall	F1	AUC
BI	0.7078	0.7081	0.7078	0.7065	0.7776
BB	0.7510	0.7510	0.7510	0.7505	0.8312
VSI	0.6988	0.7008	0.6988	0.6990	0.7786
DI	0.5922	0.5922	0.5922	0.5844	0.6212
BI+BB	0.7823	0.7829	0.7823	0.7817	0.8634
BI+VSI	0.7868	0.7869	0.7868	0.7865	0.8704
BI+DI	0.7160	0.7168	0.716	0.7144	0.7849
BB+VSI	0.7979	0.7978	0.7979	0.7979	0.8812
BB+DI	0.7532	0.7532	0.7532	0.7527	0.8308
VSI+DI	0.7105	0.7116	0.7105	0.7107	0.7945
BI+BB+VSI	<b>0.8218</b>	<b>0.8219</b>	<b>0.8218</b>	<b>0.8216</b>	<b>0.9047</b>
BI+BB+DI	0.7786	0.7792	0.7786	0.7780	0.8607
BI+VSI+DI	0.7866	0.7865	0.7866	0.7863	0.8709
BB+VSI+DI	0.7984	0.7984	0.7984	0.7984	0.8813
BI+BB+VSI+DI	0.8208	0.8208	0.8208	0.8205	0.9044

Bold text denoted the best performance among different feature combinations.

markers, namely, plasma atherogenic index ( $AIP = \log_{10} [TG/HDL]$ ), uric acid/high-density lipoprotein ratio ( $UHR = UA/(18 * HDL)$ ), neutrophil/lymphocyte ratio ( $NLR = NE/LYM$ ), platelet/lymphocyte ratio ( $PLR = PLT/LYM$ ), monocyte/lymphocyte ratio ( $MLR = MO/LYM$ ), and systemic immune inflammation index ( $SII = PLT \times NE/LYM$ ), systemic inflammatory response index ( $SIRI = NE \times MO/LYM$ ), and systemic inflammatory composite index ( $AISI = NE \times MO \times PLT/LYM$ ) (Wu et al., 2023; E et al., 2025).

In this section, we focused on different types of features, such as basic information (BI), blood biomarkers (BB), vital signs information (VSI), derivative indicators (DI). Feature combinations analysis was conducted, and its performance on the test set is shown in Table 6. From the view of single category features, blood biomarkers (BB) demonstrated core prediction value, reflecting the direct correlation of blood biomarkers in the pathological mechanisms of CHD such as lipid metabolism and inflammatory response. In multi class feature combinations, the performance of three class feature fusion (BI+BB+VSI) reached its peak, with accuracy 0.8218 and AUC 0.9047 being the best among all combinations. The integration of basic information, blood biomarkers, and vital signs has constructed a complete CHD risk profile from three dimensions, clinical phenotype, biochemical mechanisms, and physiological status, maximizing the complementarity between features. Although the feature combination performance of DI is theoretically guaranteed, it does not exceed BI+BB+VSI. This may be because DI introduced redundant information, which slightly interferes with the model's generalization. This also proved the prediction algorithm's ability to mine the cross-complementarity of feature.

### 3.5 SHAP based model interpretation and key feature correlation analysis

The highest mean absolute SHAP value of mean-RR in Figure 6 indicated that it has the most significant global influence on CHD prediction in the model, followed by age, hs-cTnI, and hypertension, which collectively constitute the core drivers of model decision-making. We further analyzed the SHAP dependency plots for key features in Figure 7. The mean-RR dependency plot showed that a low respiratory rate is weighted as a positive contributor to CHD risk. In hospitalized patients or underlying disease populations, shortness of breath is an extremely common non-specific symptom with various causes, such as pain, anemia, anxiety, etc. Through data-driven analysis, the model identified high RR as strongly correlated with non-CHD hospitalization causes. Consequently, a relatively lower RR served as a distinguishing signal for occult CHD within this specific patient population. Regarding age, SHAP values increased monotonically, confirming age as a robust risk factor. The observed plateauing effect in the elderly suggested a deceleration in risk accumulation, consistent with established clinical knowledge regarding the progression of coronary atherosclerosis. For hs-cTnI, when hs-cTnI exceeded 0.0024, the SHAP value rapidly turned positive and remained at a high level, even within the clinical normal reference range, indicating that an increase in hs-cTnI has significantly increased the risk of CHD. This highlighted the sensitivity of high-sensitivity troponin in early myocardial injury and risk prediction. The role of SHAP analysis is limited to reflecting specific behavioral patterns of the model and cannot be used to infer causal relationships. At the same time, the numerical values of its results will also vary with the distribution of data and the structure of the model.

However, the control group may include individuals with acute respiratory conditions, potentially introducing a confounding bias where elevated respiratory rates reflect the pathology of the control group rather than a direct risk factor for CHD. To address this concern and verify the model's robustness, we conducted a sensitivity analysis by excluding all respiratory-related features (mean-RR, max-RR, min-RR, and 1st-RR) and retraining the XGBoost model. The results showed that while the AUC on the test set experienced a moderate decline from 0.9053 to 0.8693, it remained within a clinically excellent range. This performance retention confirmed that although respiratory rate significantly contributes to discrimination, the model's prediction power is fundamentally driven by the comprehensive integration of multiparametric features, rather than solely relying on distinguishing respiratory-related anomalies in the control group.

To further explore feature interactions, we calculated the Spearman rank correlation coefficients for the top 20 features ranked by SHAP importance, as visualized in Figure 8. The absolute value of the correlation coefficient was represented by the radius of a circle. The larger the radius, the stronger the correlation. The color represented the direction of correlation, with red indicating positive correlation and blue indicating negative correlation. The "X" in the upper triangle indicated that

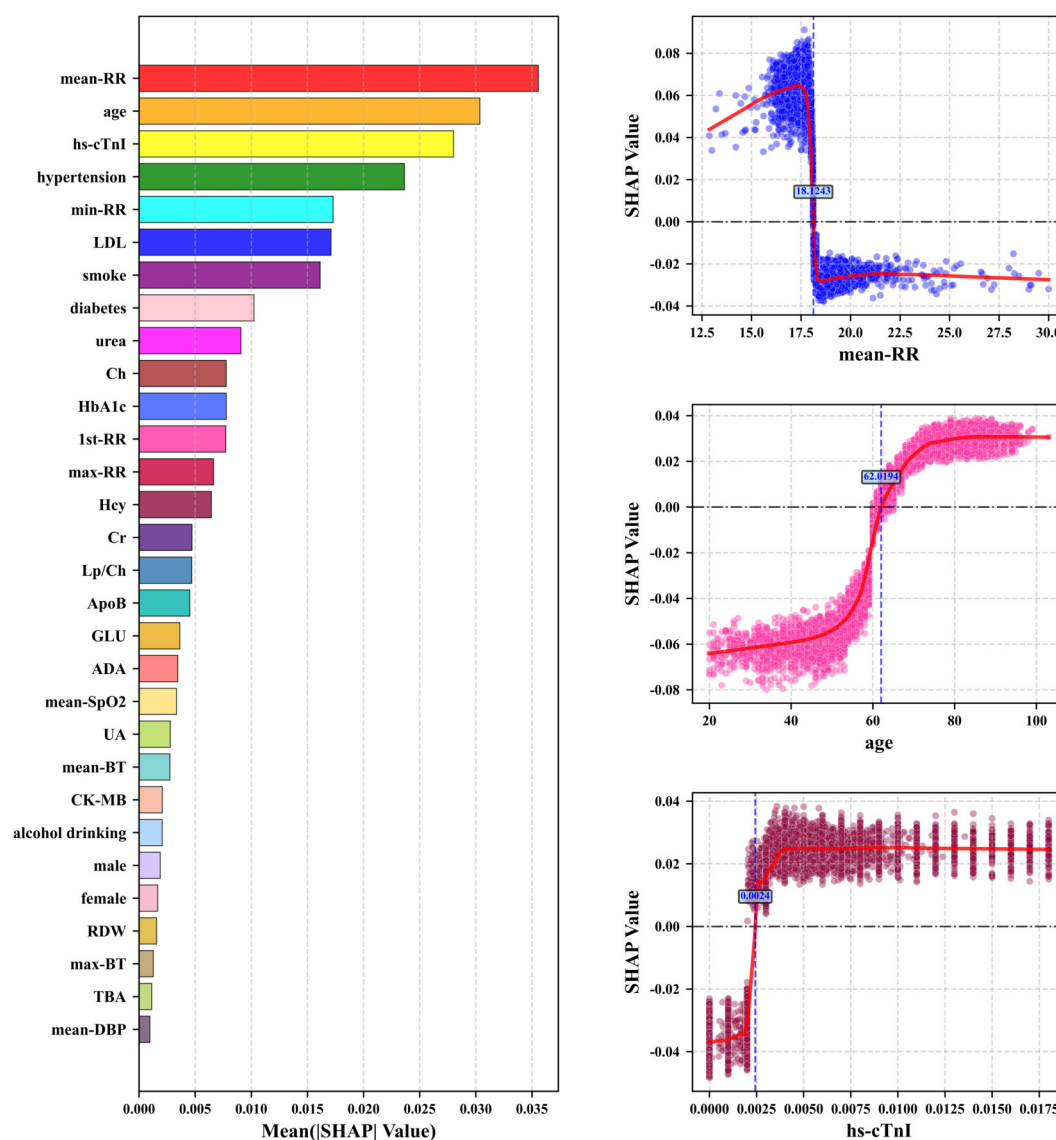


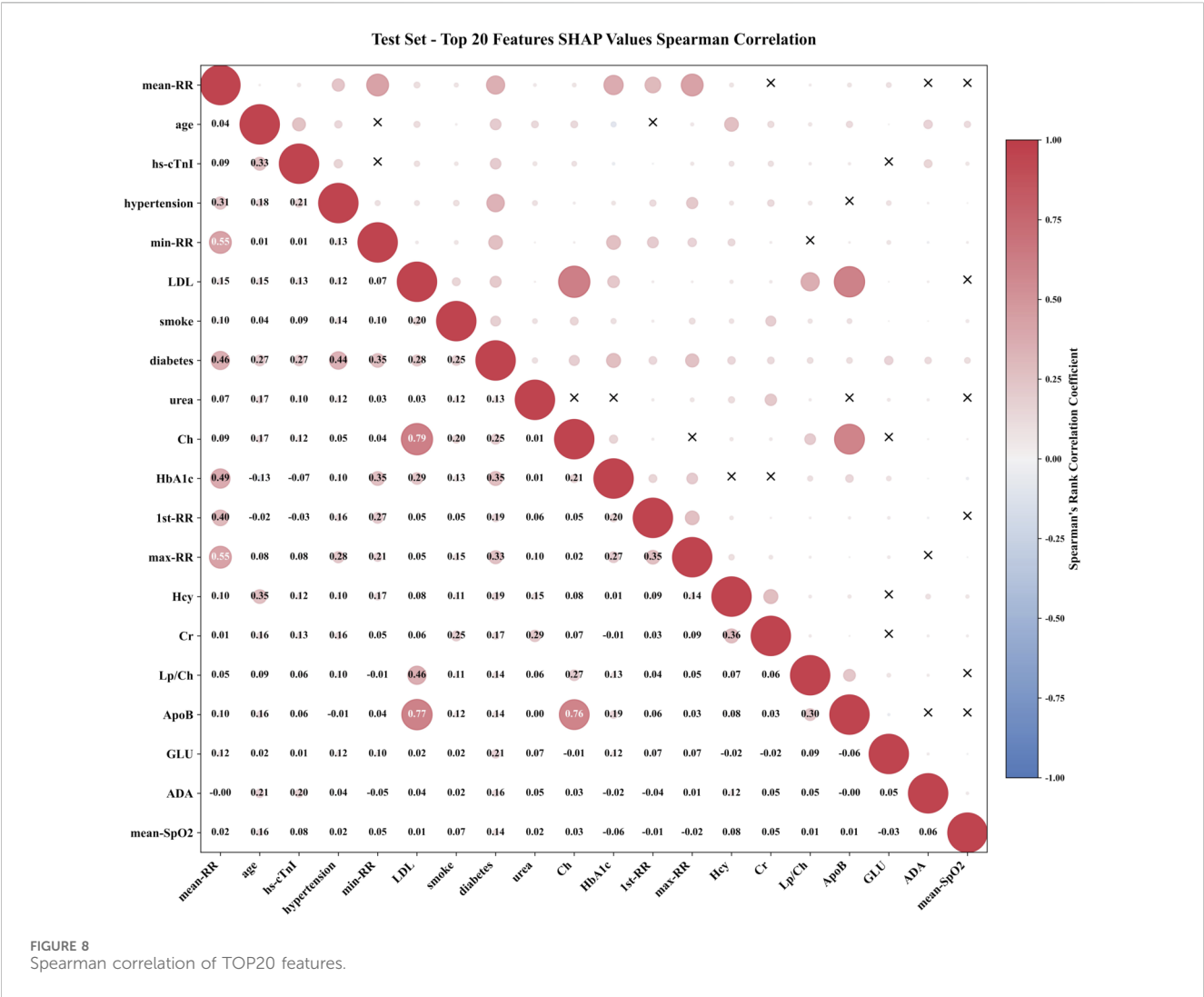
FIGURE 7

The left sub figure: the top 20 features with mean absolute SHAP values. The right three sub figures: SHAP dependency plots for the top 3 key features including SHAP value fitting curves and feature boundary values when SHAP = 0.

the feature pair is not statistically significant, while the specific correlation coefficient values were labeled in the lower triangle. The diagonal represented the autocorrelation of the feature. In addition to RR series features, the correlation coefficients of lipid metabolism features Ch, ApoB, and LDL are 0.79, 0.77, and 0.76, respectively, reflecting their synergistic effects in the process of lipid transport and atherosclerosis. This type of strong correlation prompt required attention to the joint effect of feature groups when interpreting model decisions, rather than the independent contribution of a single feature. However, features without statistically significant associations, such as age and some metabolic indicators, smoke and LDL, had no statistically significant association at the SHAP level ( $P > 0.05$ ), indicating that the weight allocation of these features by the model was relatively independent and can reduce the interference of multicollinearity on model stability.

### 3.6 Clinical implementation and limitations

To facilitate the clinical translation of this low-cost screening tool, a dual approach can be implemented: integration into in-hospital EHR systems to generate automated risk alerts for non-cardiology departments, and the development of mobile health applications to enable self-monitoring for individuals with underlying conditions. However, widespread deployment necessitates addressing critical barriers, including strict adherence to data privacy regulations, the necessity of dynamic model updating to counter concept drift, and the challenge of fostering clinician trust—which is partially mitigated by the SHAP interpretability framework employed in this research. Furthermore, the interpretation of findings must be tempered by the limitations of a retrospective, single-center design. While the current model incorporates age and comorbidities as features, rigorous



validation in future multi-center prospective studies is required to ensure fairness, robustness, and equitable healthcare outcomes across diverse subpopulations.

4 Conclusion

This research successfully developed a machine learning based CHD risk prediction model, effectively improving its generalization ability and practicality in complex clinical backgrounds. By using GAIN imputation method to process missing data, combined with XGBoost algorithm to achieve high-precision prediction, and utilizing SHAP method to reveal the contribution of key features to the prediction results, the interpretability of the model is enhanced. The research results indicate that the model has significant potential in identifying hidden CHD, which can assist clinical doctors in early intervention and personalized management, and provide a low-cost and easy to promote self-monitoring method for the population with underlying diseases. Although the model cannot replace the gold-standard diagnosis, it has important practical value in the connection between public health

prevention and clinical diagnosis and treatment, which helps to reduce medical burden.

This approach aims to provide intuitive basis for clinical doctors to understand the mechanism of disease association and formulate personalized intervention strategies, and to provide low-cost and easy to promote practical tools for independent heart health monitoring in populations with underlying conditions.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by Ethics Committee of Quzhou People’s Hospital. The studies were conducted in accordance with the local legislation and institutional requirements.

## Author contributions

SZ: Writing – original draft, Methodology. BN: Validation, Writing – review and editing, Supervision, Conceptualization. JG: Methodology, Writing – original draft, Data curation. WX: Writing – review and editing. ZL: Writing – original draft, Methodology.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by Zhejiang Health Information Association Research Program (2023XHSZ-Y06), Quzhou City Competitive Science and Technology Project (2025K021) and Doctoral Initiation Projects of Shenzhen Institute of Information Technology (No. SZIIT2025KJ058). The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- Chen, Y., Wang, A., Liu, Z., Yue, J., Zhang, E., Li, F., et al. (2025a). MoSViT: a lightweight vision transformer framework for efficient disease detection via precision attention mechanism. *Front. Artif. Intell.* 8, 1498025. doi:10.3389/fgene.2025.1498025
- Chen, Y., Wang, Z., Wang, J., Chu, Y., Zhang, Q., Li, Z. A., et al. (2025b). Self-supervised learning in drug discovery. *Sci. China Inf. Sci.* 68 (7), 170103. doi:10.1007/s11432-024-4453-4
- Colantonio, L., Gamboa, C., Richman, J., Levitan, E., Soliman, E., Howard, G., et al. (2017). Black-white differences in incident fatal, nonfatal, and total coronary heart disease. *Circulation* 136 (2), 152–166. doi:10.1161/CIRCULATIONAHA.116.025848
- E, Y., Yao, Z., Ge, M., Huo, G., Huang, J., Tang, Y., et al. (2025). Development and validation of a machine learning model for predicting vulnerable carotid plaques using routine blood biomarkers and derived indicators: insights into sex-related risk patterns. *Cardiovasc. Diabetol.* 24 (1), 326. doi:10.1186/s12933-025-02867-6
- Gautier, V., Bousse, A., Sureau, F., Comtat, C., and Maxim, S. B. (2024). Bimodal PET/MRI generative reconstruction based on VAE architectures. *Phys. Med. Biol.* 69 (24), 245019. doi:10.1088/1361-6560/ad9133
- Koloi, A., Loukas, V., Hourican, C., Sakellarios, A., Quax, R., Mishra, P., et al. (2024). Predicting early-stage coronary artery disease using machine learning and routine clinical biomarkers improved by augmented virtual data. *Eur. Heart J. - Digital Health* 5 (5), 542–550. doi:10.1093/ehjdh/ztae049
- Kumar, A., Dhanka, S., Singh, J., Khan, A., and Maini, S. (2024). Hybrid machine learning techniques based on genetic algorithm for heart disease detection. *Innovation Emerg. Technol.* 11, 2450008. doi:10.1142/s2737599424500087
- Kundrick, J., Naniwadekar, A., Singla, V., Kancharla, K., Bhonsale, A., Voigt, A., et al. (2025). Machine learning applied to wearable fitness tracker data and the risk of hospitalizations and cardiovascular events. *Am. J. Prev. Cardiol.* 22, 101006. doi:10.1016/j.ajpc.2025.101006
- Kuo, H., Chen, S., Chen, I., Cheng, W., Liu, S., Guo, M., et al. (2024). Novel multiple Z-score models for detection of coronary artery dilation: application in Kawasaki disease. *Pediatr. Rheumatol.* 22 (1), 108. doi:10.1186/s12969-024-01040-9
- Ladak, L., Gallagher, R., Hasan, B., Awais, K., Abdullah, A., and Gullick, J. (2020). Health-related quality of life in adult CHD surgical patients in a low middle-income country: a mixed-methods study. *Cardiol. Young* 30 (8), 1126–1137. doi:10.1017/S1047951120001663
- Li, H., Pang, Y., and Liu, B. (2021). BioSeq-BLM: a platform for analyzing DNA, RNA, and protein sequences based on biological language models. *Nucleic Acids Res.* 49 (22), e129. doi:10.1093/nar/gkab829
- Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47 (20), e127. doi:10.1093/nar/gkz740
- Min, J., Chang, H., Andreini, D., Pontone, G., Guglielmo, M., Bax, J., et al. (2022). Coronary CTA plaque volume severity stages according to invasive coronary angiography and FFR. *J. Cardiovasc. Comput. Tomogr.* 16 (5), 415–422. doi:10.1016/j.jcct.2022.03.001
- Nayak, A., Venugopala, P., and Ashwini, B. (2024). A systematic review on generative adversarial network (GAN): challenges and future directions. *Archives Comput. Methods Eng.* 31 (8), 4739–4772. doi:10.1007/s11831-024-10119-1
- Nenova, Z., and Shang, J. (2022). Chronic disease progression prediction: leveraging case-based reasoning and big data analytics. *Prod. Operations Manag.* 31 (1), 259–280. doi:10.1111/poms.13532
- Nollmann, C., Moskorz, W., Wimmenauer, C., Jäger, P., Cadeddu, R., Timm, J., et al. (2024). Characterization of CD34+ cells from patients with acute myeloid leukemia (AML) and myelodysplastic syndromes (MDS) using a t-Distributed stochastic neighbor embedding (t-SNE) protocol. *Cancers* 16 (7), 1320. doi:10.3390/cancers16071320
- Pickles, D., and Keller, K. (2025). The economic burden of complex CHD in the United States. *Cardiol. Young* 35 (9), 1751–1758. doi:10.1017/S1047951125109256
- Qiao, J., Jin, J., Yu, H., and Wei, L. (2024). Towards retraining-free RNA modification prediction with incremental learning. *Inf. Sci.* 660, 120105. doi:10.1016/j.ins.2024.120105
- Qiao, J., Jin, J., Wang, D., Teng, S., Zhang, J., Yang, X., et al. (2025). A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. *Nat. Commun.* 16 (1), 1–16. doi:10.1038/s41467-025-59634-0
- Rehman, M., Naseem, S., Butt, A., Mahmood, T., Khan, A., Khan, I., et al. (2025). Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Sci. Rep.* 15 (1), 13361. doi:10.1038/s41598-025-96437-1
- Rimal, Y., and Sharma, N. (2023). Hyperparameter optimization: a comparative machine learning model analysis for enhanced heart disease prediction accuracy. *Multimedia Tools Appl.* 83, 55091–55107. doi:10.1007/s11042-023-17273-x
- Sanchez-Morillo, D., León-Jiménez, A., Guerrero-Chanivet, M., Jiménez-Gómez, G., Hidalgo-Molina, A., and Campos-Caro, A. (2024). Integrating routine blood biomarkers

## Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2025.1752811/full#supplementary-material>

and artificial intelligence for supporting diagnosis of silicosis in engineered stone workers. *Bioeng. and Transl. Med.* 9 (6), e10694. doi:10.1002/btm2.10694

Sawaf, B., Swed, S., Alibrahim, H., Bohsas, H., Dave, T., Nasif, M., et al. (2024). Triglyceride-glucose index as predictor for hypertension, CHD and STROKE risk among non-diabetic patients: a NHANES cross-sectional study 2001-2020. *J. Epidemiol. Glob. Health* 14 (3), 1152–1166. doi:10.1007/s44197-024-00269-7

Shi, C., Wang, C., Zhou, X., and Qin, Z. (2024). DAE-Net: dual attention mechanism and edge supervision network for image manipulation detection and localization. *Ieee Trans. Instrum. Meas.* 73, 1–17. doi:10.1109/tim.2024.3451570

Topranin, V., Wiig-Fisketjon, A., Botten, E., Dalen, H., Langaas, M., and Bye, A. (2025). Sex-specific cardiovascular disease risk prediction using statistical learning and explainable artificial intelligence: the HUNT study. *Eur. J. Prev. Cardiol.*, zwaf135. doi:10.1093/eurjpc/zwaf135

Varol, B., Omurlu, I., and Ture, M. (2025). Simulation comparison of the effects of missing data imputation methods on classification performance in high dimensional data. *Commun. Statistics-Simulation Comput.*, 1–20. doi:10.1080/03610918.2025.2524548

Wang, Y., Zhai, Y., Ding, Y., and Zou, Q. (2024). SBSM-Pro: support bio-sequence machine for proteins. *Sci. China Inf. Sci.* 67 (11), 212106. doi:10.1007/s11432-024-4171-9

Wang, L., Qian, Y., Xie, H., Ding, Y., and Guo, F. (2025). Structured sparse regularization-based deep fuzzy networks for RNA N6-Methyladenosine sites prediction. *IEEE Trans. Fuzzy Syst.* 33 (1), 131–144. doi:10.1109/tfuzz.2024.3428402

Wu, S., Yang, S., Wang, M., Song, N., Feng, J., Wu, H., et al. (2023). Quorum sensing-based interactions among drugs, microbes, and diseases. *Sci. China-Life Sci.* 66 (1), 137–151. doi:10.1007/s11427-021-2121-0

Xie, H., Wang, L., Qian, Y., Ding, Y., and Guo, F. (2025). Methyl-GP: accurate generic DNA methylation prediction based on a language model and representation learning. *Nucleic Acids Res.* 53 (6), gkaf223. doi:10.1093/nar/gkaf223

Xiong, Q., Fu, X., Ku, L., Zhou, D., Guo, S., and Zhang, W. (2024). Diagnostic performance of coronary computed tomography angiography stenosis score for coronary stenosis. *BMC Med. Imaging* 24 (1), 39. doi:10.1186/s12880-024-01213-8

Xu, Y., Fang, S., and Xing, X. (2025). Self-supervised multi-level generative adversarial network data imputation algorithm. *Int. J. Approx. Reason.* 187, 109553. doi:10.1016/j.ijar.2025.109553

Yang, H., Luo, Y. M., Ma, C. Y., Zhang, T. Y., Zhou, T., Ren, X. L., et al. (2023). A gender specific risk assessment of coronary heart disease based on physical examination data. *NPJ Digital Medicine* 6 (1), 136. doi:10.1038/s41746-023-00887-8

Zaninotto, P., Steptoe, A., and Shim, E. (2024). CVD incidence and mortality among people with diabetes and/or hypertension: results from the English longitudinal study of ageing. *Plos One* 19 (5), e0303306. doi:10.1371/journal.pone.0303306

Zeng, T., Wang, Y., Tang, B., Cui, H., Tang, D., Ding, H., et al. (2025). Colorectal liver metastasis pathomics model (CLMPM): integrating single cell and spatial transcriptome analysis with pathomics for predicting liver metastasis in colorectal cancer. *Mod. Pathology* 38, 100805. doi:10.1016/j.modpat.2025.100805

Zulfiqar, H., Guo, Z., Ahmad, R. M., Ahmed, Z., Cai, P., Chen, X., et al. (2024). Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings. *Front. Med.* 10, 1291352. doi:10.3389/fmed.2023.1291352