Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

Construction and evaluation of a predictive model for the degree of coronary artery occlusion based on adaptive weighted multi-modal fusion of traditional Chinese and western medicine data

Jiyu ZHANG, Jiatuo XU*, Liping TU, Hongyuan FU

College of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 200120, China

ARTICLE INFO

Article history
Received 17 December 2024
Accepted 24 April 2025
Available online 25 June 2025

Keywords
Coronary artery disease
Deep learning
Multi-modal
Clinical prediction
Traditional Chinese medicine diagnosis

ABSTRACT

Objective To develop a non-invasive predictive model for coronary artery stenosis severity based on adaptive multi-modal integration of traditional Chinese and western medicine data.

Methods Clinical indicators, echocardiographic data, traditional Chinese medicine (TCM) tongue manifestations, and facial features were collected from patients who underwent coronary computed tomography angiography (CTA) in the Cardiac Care Unit (CCU) of Shanghai Tenth People's Hospital between May 1, 2023 and May 1, 2024. An adaptive weighted multi-modal data fusion (AWMDF) model based on deep learning was constructed to predict the severity of coronary artery stenosis. The model was evaluated using metrics including accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic (ROC) curve (AUC). Further performance assessment was conducted through comparisons with six ensemble machine learning methods, data ablation, model component ablation, and various decision-level fusion strategies.

Results A total of 158 patients were included in the study. The AWMDF model achieved excellent predictive performance (AUC = 0.973, accuracy = 0.937, precision = 0.937, recall = 0.929, and F1 score = 0.933). Compared with model ablation, data ablation experiments, and various traditional machine learning models, the AWMDF model demonstrated superior performance. Moreover, the adaptive weighting strategy outperformed alternative approaches, including simple weighting, averaging, voting, and fixed-weight schemes.

Conclusion The AWMDF model demonstrates potential clinical value in the non-invasive prediction of coronary artery disease and could serve as a tool for clinical decision support.

1 Introduction

In recent years, the incidence of coronary artery disease (CAD) has been steadily rising, with morbidity and mortality rates remaining persistently high, and the number of affected individuals is estimated to be approximately 330 million [1]. The current "gold standard" for diagnosing coronary artery stenosis is coronary computed tomography angiography (CTA). However, due to its invasive nature, non-invasive diagnostic approaches for CAD have become a research focus recently. With advances in computational medicine, the value of non-invasive diagnostics leveraging data and model development has gained growing recognition [2-4]. Traditional Chinese medicine

*Corresponding author: Jiatuo XU, E-mail: xjt@fudan.edu.cn.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2025.05.005

Citation: ZHANG JY, XU JT, TU LP, et al. Construction and evaluation of a predictive model for the degree of coronary artery occlusion based on adaptive weighted multi-modal fusion of traditional Chinese and western medicine data. Digital Chinese Medicine, 2025, 8(2): 163-173.

(TCM) diagnostic methods maintain an inherent advantage in non-invasive diagnostics for CAD. Recent studies have shown that the objectification of TCM's four diagnostic techniques contributes to the diagnosis of CAD [5]. TCM diagnostic approaches, including facial diagnosis [6], tongue diagnosis [7], and pulse diagnosis [8], offer unique advantages in identifying CAD.

TCM diagnostics emphasize the integration of the four diagnostics (inspection, ausculation and olfaction, inquiry, and palpation) to form a comprehensive judgment of diseases, which bear certain similarities to the multimodal data fusion approach employed in contemporary artificial intelligence (AI) research. By integrating data from multiple sources, multimodal fusion in medicine facilitates complementary information exchange among diverse datasets, promoting prediction accuracy and model robustness [9].

This study combined tongue and facial images from TCM diagnostics with laboratory test results and echocardiographic data from modern medicine to construct an integrated non-invasive diagnostic model. Additionally, an adaptive weighting module was incorporated into the model, enabling dynamic adjustment of the contribution of data from different modalities to optimize predictive performance. This approach aims to provide accurate probability predictions for high-risk coronary artery stenosis, which strengthens risk stratification for CAD.

2 Data and methods

2.1 Study participants

The patients included in this study were recruited from the Cardiac Care Unit (CCU) of Shanghai Tenth People's Hospital. The data, which were collected from May 1, 2023 to May 1, 2024, consist of laboratory parameters, echocardiographic findings, and TCM diagnostic images of the tongue and face. This study was reviewed and approved by the Ethics Committee of Shanghai University of Traditional Chinese Medicine (2020-916-125) and registered in the Chinese Clinical Trial Registry (ChiCTR2100043546). The study was conducted in strict accordance with the ethical principles outlined in the Declaration of Helsinki by the World Medical Association. Written informed consent was obtained from all participants prior to their enrollment, and comprehensive information regarding the study's purpose, procedures, potential risks, and expected benefits was provided to each participant.

2.1.1 Diagnostic criteria The diagnostic criteria for CAD were based on the *Internal Medicine (9th Edition)*, which outlines acute and chronic CAD diagnostic standards [10].

2.1.2 Inclusion criteria Patients were included if they met the following criteria: (i) fulfill the diagnostic criteria for CAD; (ii) age between 20 and 85 years; (iii) presence of typical chest pain symptoms (e.g., episodic angina or oppressive pain); (iv) presence of weakened heart sounds on auscultation; (v) documentation of ST-segment abnormalities on electrocardiography; (vi) coronary angiography-confirmed stenosis $\geqslant$ 1 vessel with severity > level 1 according to the 2019 European Society of Cardiology (ESC) Guidelines for the Diagnosis and Management of Chronic Coronary Syndromes [11].

2.1.3 Exclusion criteria The exclusion criteria were as follows: (i) patients who did not meet CAD diagnostic criteria; (ii) patients diagnosed with malignant tumors or critical illnesses; (iii) women who were pregnant or lactating; (iv) individuals with incomplete data.

2.2 Collection and processing of multi-modal data

2.2.1 Collection and analysis of clinical data Clinical data are comprised of information on comorbidities, echocardiographic vascular data, laboratory biochemical indicators, and thromboelastography results. Data were categorized based on the presence or absence of $\geqslant$ 75% stenosis of the coronary artery following CTA examination. Cases with stenosis < 75% were classified as negative samples, while cases with stenosis $\geqslant$ 75% as positive samples.

2.2.2 Collection of TCM tongue and facial data Tongue and facial images were collected and analyzed using the TFDA-1 digital tongue and facial diagnostic device (registration No. 20212200604). The image acquisition parameters were standardized as follows: shutter speed of 1/125 s, aperture value of F6.3, and international organization for standardization (ISO) sensitivity of 200. The diagnostic device is shown in Figure 1. Image acquisition followed a standardized protocol between 8:30 am and 11:30 am. Patients were seated upright, facing the imaging device, with the built-in standardized light source activated. Each patient's head was positioned at the chin alignment mark on the device. During tongue image collection, patients were instructed to close their eyes and protrude their tongue. During facial image collection, patients maintained both mouth and eyes closed.

2.3 Construction of the multi-modal algorithm

The model was built in a Python environment using the PyTorch framework. The architecture of the adaptive weighted multi-modal data fusion (AWMDF) model based on deep learning is illustrated in Figure 2, and the model parameters are listed in Table 1. Due to the presence of residual modules and attention blocks, the model's computational complexity rises, which may affect computational accuracy and lead to overfitting. To

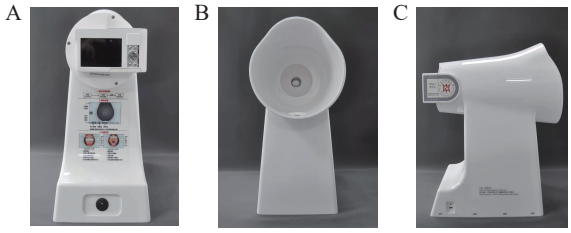

Figure 1 TFDA-1 digital tongue diagnostic instrument
A, rear view. B, front view. C, side view.

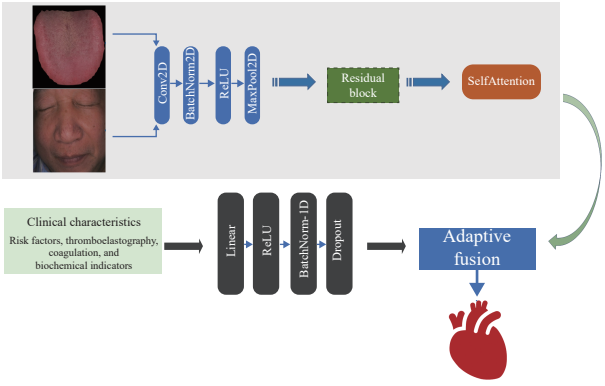

Figure 2 The schematic diagram of the AWMDF model

address this, a dropout layer was added after the fully connected layer in the model, and L2 regularization was employed to constrain the model parameters. These measures help reduce parameter redundancy and promote the model's generalization capability.

Facial and tongue images were preprocessed through standardization and subsequently passed through residual and self-attention modules to extract data features. These features were then put into the adaptive fusion module. After feature selection, clinical indicators were integrated with the tongue and facial data within the fusion module. The model dynamically adjusted the output weights of the tongue, facial, and clinical data following the relative importance of each data type in the prediction process.

The model integrated weighted multimodal features derived from tongue, facial, and clinical data to perform binary classification for identifying CAD patients with significant stenosis (⩾ 75% luminal narrowing).

2.4 Processing of image data

Tongue and facial image data were normalized to a resolution of 256 × 256 pixels. The ResNet50 architecture was used by the residual module to boost the depth of image feature extraction, facilitating the identification of differential features. Additionally, a self-attention module was integrated to strengthen the model's learning of critical feature information. After processing using the residual and self-attention modules, the extracted features were concatenated and passed into the adaptive decision fusion module for further processing and integration.

2.5 Processing of clinical data

The included variables comprised risk factors of CAD, echocardiographic data, and laboratory biochemical parameters. Variables with over 30% missing data were excluded. For variables with under 10% missing data, mean imputation was applied, whereas those with 10% to 30%

Table 1 Parameters of the AWMDF model

Modal	Layer name	Type	Output shape
Face	Conv2D-1	Convolutional Layer	[batch, 64, 128, 128]
	BatchNorm-1	Batch normalization	[batch, 64, 128, 128]
	ReLU-1	Activation function	[batch, 64, 128, 128]
	MaxPool2D-1	Max pooling	[batch, 64, 64, 64]
	ResidualBlock-1	Residual block	[batch, 64, 64, 64]
	SelfAttention-1	Self-attention mechanism	[batch, 64, 64, 64]
	AdaptiveAvgPool2D-1	Adaptive average pooling	[batch, 64, 1, 1]
Tongue	Conv2D-1	Convolutional layer	[batch, 64, 128, 128]
	BatchNorm-1	Batch normalization	[batch, 64, 128, 128]
	ReLU-1	Activation function	[batch, 64, 128, 128]
	MaxPool2D-1	Max pooling	[batch, 64, 64, 64]
	ResidualBlock-1	Residual block	[batch, 64, 64, 64]
	SelfAttention-1	Self-attention mechanism	[batch, 64, 64, 64]
	AdaptiveAvgPool2D-1	Adaptive average pooling	[batch, 64, 1, 1]
Lab	Linear-1	Fully connected layer	[batch, 128]
	ReLU-2	Activation function	[batch, 128]
	BatchNorm1D-1	Batch normalization	[batch, 128]
	Dropout-1	Dropout	[batch, 128]

missing data were handled using placeholder values for subsequent estimation.

Due to the imbalance in the sample size within this model, this study introduces the Synthetic Minority Over-Sampling Technique (SMOTE) to promote the model's ability to recognize minority class samples, reduce the data bias caused by imbalanced samples, and avoid overfitting owing to random oversampling [12]. Using this method, minority class samples were selected, and for each minority sample, its nearest neighbors were identified. The minority samples were expanded with the use of a weighted combination of the original sample and its neighbors. This approach effectively addresses the recognition problems of minority class samples in binary classification.

2.6 Adaptive module architecture

The adaptive decision module dynamically updated the output weights of the model based on the individual accuracy performance of each data modality (tongue, facial, and clinical data). The updated weights were then integrated into the model to strengthen the final classification outcome. Unlike fixed-weight fusion methods, the adaptive decision fusion mechanism allowed the model to learn weights during training, enabling the contribution of input features from different modalities to be dynamically adjusted in line with task-specific requirements. Under this mechanism, the final fused feature vector is represented as:

$$f_{\text{combined}} = \omega_{\text{tongue}} \times f_{\text{tongue}} + \omega_{\text{face}} \times f_{\text{face}} + \omega_{\text{lab}} \times f_{\text{lab}} \quad (1)$$

Here, $f_{(\text{tongue/face/lab})}$ represents the function outputs of tongue data, facial data, and clinical data after feature extraction by the model, respectively. Where ω and b denote the classification weight matrix and bias term, respectively. The weighted fusion results were passed through a fully connected layer, and the final binary classification output is expressed as:

$$\text{output} = \sigma(\omega \times \text{f}_{\text{combined}} + \text{b})$$

Softmax was used as the activation function. The detailed process for calculating model parameters is as follows: initial weight data were provided for different modalities, and classification probabilities were obtained through the activation function (softmax). The model loss was calculated using the cross-entropy loss function. In the adaptive mechanism, the model dynamically adjusted the weights during training based on the contribution of each modality to classification performance. During each backpropagation step, weight updates were performed based on the gradient and learning rate. The weights were updated using the stochastic gradient descent (SGD) optimizer.

The gradient and learning rate calculation processes are represented as follows:

$$w_{\text{tongue}} \leftarrow w_{\text{tongue}} - \eta \times \frac{\partial L}{\partial w_{\text{tongue}}} \quad (2)$$

$$w_{\text{face}} \leftarrow w_{\text{face}} - \eta \times \frac{\partial L}{\partial w_{\text{face}}} \quad (3)$$

$$w_{\text{lab}} \leftarrow w_{\text{lab}} - \eta \times \frac{\partial L}{\partial w_{\text{lab}}} \quad (4)$$

In contrast to dynamic weight updating, fixed-weight fusion methods use predetermined, constant weight values. In this context, η represents the learning rate, ∂ denotes the partial derivative, and L stands for the loss function. Although such methods can perform effectively in certain simple tasks, they exhibit notable limitations, for instance, the inability to adapt to the varying requirements of different tasks. In this study, the selected modalities (tongue, facial, and pulse features) do not demonstrate strong complementarity in identifying coronary artery stenosis. Therefore, the adaptive mechanism, through dynamic weight adjustment, overcomes these limitations by raising the model's flexibility and accuracy, allowing it to better accommodate the demands of diverse tasks. In this study, several alternative decision-level fusion methods were also compared, including SimpleWeight, AverageWeight, Voting, and FixedWeight, to evaluate their performance against the proposed dynamic weight fusion approach.

2.7 Model hyperparameters

The model was implemented using the PyTorch framework (v2.0.1) with the following computational resources: GPU (Tesla V100), CPU (4 cores), and RAM (32 GB). The dataset was divided into training and testing sets with an 8 : 2 ratio. The hyperparameter settings were: learning rate = 0.001, batch_size = 16, optimizer = Adam, training_epochs = 50. Data augmentation techniques, such as scaling and rotation, were applied to improve model robustness during training.

2.8 Gradient-weighted class activation mapping (Grad-CAM) heatmap visualization

Grad-CAM was employed to visualize the model's focus during decision-making. Grad-CAM extends conventional CAM methods by overcoming the limitation of requiring global average pooling layers in network architectures [13]. It calculates gradient information from the convolutional layer outputs to weight each feature map, capturing the relationship between class predictions and feature maps. Grad-CAM allows intuitive visualization of the key areas the model focuses on during training. The calculation is expressed as:

$$\text{Grad-CAM}_c = \text{ReLU}\left(\sum_k a_k^c A^k\right)$$

where a_k^c denotes the gradient-weighted coefficient for the class concerning the feature map A^k.

2.9 Comparison with other machine learning frameworks

To evaluate the model's feature selection performance, tongue and facial images were processed using the TCM Tongue Diagnosis Analysis System (TDAS) v2.0, developed by Shanghai University of Traditional Chinese Medicine. In CAD diagnostics, the tongue and facial images were important diagnostic references [14, 15].

Tongue image features: evaluation of tongue body indicators includes color space values from different color domains. Texture indicators consist of contrast (CON), angular second moment (ASM), entropy (ENT), and mean (MEAN). Texture indicators reflect the fineness and depth of texture in the image. A higher ASM value corresponds to lower CON, ENT, and MEAN values, indicating finer texture. PerAll and PerPart represent tongue coating indices, where PerAll is the ratio of tongue coating area to the total tongue area, and PerPart is the ratio of tongue coating area to the area without coating. Facial color indicators include facial hue, saturation, and value (HSV) color space indicators, facial red, green, and blue (RGB) color space indicators, and facial Lab color space indicators.

Six commonly used machine learning models were implemented in Python to compare prediction performance: logistic regression, decision tree, random forest, gradient boosting, support vector machine (SVM), and k-nearest neighbors (KNN). The performance of tongue, facial, and clinical data after adaptive fusion was compared with that of the AWMDF model to evaluate its superiority in feature integration and accuracy prediction.

2.10 Model evaluation

In deep learning, five commonly used evaluation metrics are accuracy, the area under the receiver operating characteristic (ROC) curve (AUC), precision, recall, and F1 score. Accuracy represents the proportion of correctly predicted samples to the total number of samples, providing an intuitive measure of the model's overall performance. A higher accuracy value indicates better predictive performance.

AUC evaluates the model's classification performance at various thresholds. An AUC value approaching 1 indicates superior model performance. This metric reflects the model's ability to discriminate between positive and negative samples and is particularly robust to class imbalance issues. Precision is defined as the proportion of predicted positive samples that are positive. It measures the reliability of the model's positive predictions. A higher precision value indicates fewer false positives when predicting positive samples. Recall (also called sensitivity) is the proportion of actual positive samples correctly identified as positive. It assesses the model's ability to capture positive samples. A higher recall value indicates that the model successfully identifies more positive instances. F1 score is the harmonic mean of precision and recall, balancing the trade-off between the two metrics. The F1 score ranges from 0 to 1, with a value closer to 1 demonstrating superior model performance.

Meanwhile, ablation studies on the model itself from both the data and module perspectives were conducted. The data-level ablation includes tongue images, facial images, and laboratory data, while the module-level ablation involves the residual, self-attention, and adaptive fusion modules. The model was evaluated under each ablation setting.

2.11 Statistical analysis

Statistical analyses were performed using SPSS 27.0. Continuous variables were expressed as mean ± standard deviation (SD). For between-group comparisons, the independent samples t test was applied to data meeting assumptions of normality and homogeneity of variance, while the Mann-Whitney U test was used for data that did not meet these assumptions. The rank-sum test was employed for categorical data. $P < 0.05$ was considered statistically significant.

3 Results

3.1 Comparison of clinical data

A total of 158 patients with CAD were included, comprising 84 males and 74 females. Significant differences were observed in aortic sinus diameter, left ventricular ejection fraction (LVEF), creatine kinase-MB (CK-MB), and arachidonic acid (AA) inhibition rate when coronary artery occlusion above 75% was treated as a binary variable ($P < 0.05$). Moreover, 45 clinical indicators were initially included, and indicators with missing data of more than 30% were excluded, leaving 34 valid indicators (Table 2).

3.2 Model data presentation

As shown in Figure 3, the AWMDF model achieved an AUC of 0.973, accuracy of 0.937, precision of 0.937, and recall of 0.929. During the 50 training epochs, accuracy exhibited fluctuations around the 20th epoch, but gradually increased from the 20th to the 50th epoch. Both AUC and precision improved progressively with the number of training epochs. Recall decreased at the 20th epoch but gradually increased from the 20th to the 50th epoch as training progressed.

Table 2 Differences in clinical data of two categories of vascular obstruction degree

Item	Vascular occlusion < 75% (n = 60)	Vascular occlusion ⩾ 75% (n = 98)	$\chi^2/t/U$ value	P value
Coronary heart disease (yes/no)	31/29	70/32	0.92	0.862 6
Hypertension (yes/no)	36/24	29/69	1.88	0.078 8
Type 2 diabetes (yes/no)	40/20	33/65	1.58	0.221 0
Gender (male/female)	38/22	46/52	0.55	0.122 7
Aortic sinus diameter (mm)	32.83 ± 4.08	34.12 ± 3.18	1 106.50	0.033 2
Left atrial diameter (mm)	36.85 ± 5.53	36.66 ± 4.44	1 469.00	0.959 2
Left ventricular end-diastolic diameter (mm)	45.90 ± 5.62	46.44 ± 4.92	1 358.00	0.541 0
Left ventricular end-systolic diameter (mm)	31.95 ± 7.15	32.05 ± 5.61	1 388.50	0.669 2
Interventricular septal thickness (mm)	10.75 ± 1.97	10.48 ± 1.03	1 446.00	0.932 0
Left ventricular posterior wall thickness (mm)	10.18 ± 1.55	9.97 ± 0.87	1 416.50	0.780 1
LVEF (%)	52.45 ± 10.19	47.25 ± 11.23	1 900.00	0.008 1
S wave peak value	0.07 ± 0.02	0.07 ± 0.02	1 635.00	0.283 3
E/E'	11.03 ± 2.60	11.16 ± 2.30	1 424.50	0.997 4
Myoglobin (ng/mL)	86.29 ± 94.67	125.77 ± 260.07	2 322.00	0.902 2
CK-MB (ng/mL)	10.67 ± 26.14	30.66 ± 66.43	1 594.00	0.001 6
NT-proBNP (pg/L)	5 453.18 ± 9 798.11	2 625.69 ± 4 024.53	2 297.50	0.899 7
Total cholesterol (mmol/L)	4.33 ± 1.30	4.34 ± 1.15	1 700.50	0.667 4
Triglycerides (mmol/L)	1.44 ± 1.21	3.24 ± 15.35	1 583.00	0.256 9
HDL-C (mmol/L)	1.27 ± 0.29	2.52 ± 11.58	1 978.50	0.591 5
LDL-C (mmol/L)	2.50 ± 1.17	2.61 ± 0.96	1 433.00	0.303 3
Glycated hemoglobin (%)	6.82 ± 1.70	7.22 ± 1.80	1 217.00	0.118 1
Glucose (mmol/L)	6.64 ± 3.0	6.74 ± 2.68	1 675.50	0.648 8
INR	1.09 ± 0.32	1.00 ± 0.10	2 418.50	0.409 9
Fibrinogen (g/L)	3.96 ± 1.55	4.03 ± 1.30	2 177.00	0.600 8
D-Dimer (mg/L)	1.56 ± 2.34	1.29 ± 2.86	2 686.50	0.285 4
FDP (μg/mL)	5.76 ± 6.44	6.64 ± 18.26	1 404.50	0.826 7
Thromboelastography R (min)	5.58 ± 2.51	5.17 ± 1.00	693.00	0.979 7
Thromboelastography K (min)	1.65 ± 0.76	1.62 ± 0.47	631.00	0.549 9
Thromboelastography angle (deg)	67.65 ± 8.15	67.34 ± 5.37	741.00	0.607 3
Thromboelastography MA (mm)	66.26 ± 9.86	68.34 ± 5.87	− 0.95	0.350 8
Thromboelastography EPL (%)	0.14 ± 0.21	0.52 ± 2.19	684.00	0.880 1
Thromboelastography CI	1.19 ± 3.19	1.73 ± 1.41	625.50	0.448 5
AA inhibition rate (%)	67.67 ± 39.33	89.16 ± 22.35	514.00	0.013 2
ADP inhibition rate (%)	61.13 ± 27.54	72.35 ± 25.72	541.50	0.051 8

LVEF, left ventricular ejection fraction. S wave, systolic wave peak velocity. E/E', early mitral inflow velocity to early diastolic mitral annular velocity ratio. CK-MB, creatine kinase-MB. NT-proBNP, N-terminal pro-B-type natriuretic peptide. HDL-C, high-density lipoprotein cholesterol. LDL-C, low-density lipoprotein Cholesterol. INR, international normalized ratio. FDP, fibrin degradation products. AA, arachidonic acid. ADP, adenosine diphosphate.

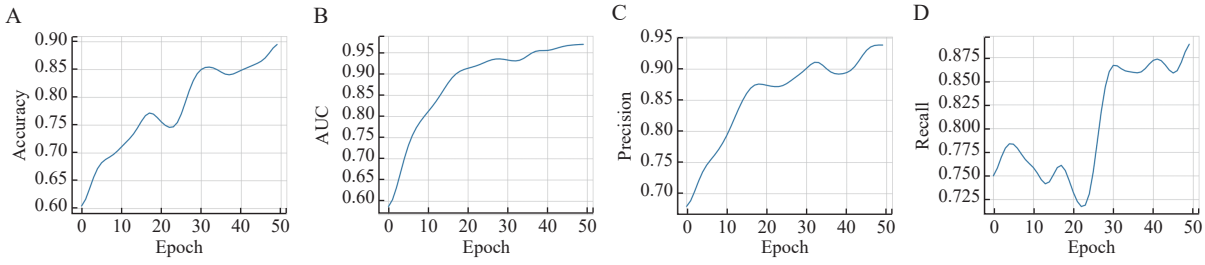

Figure 3 Visual presentation of different training parameters in the AWMDF model

A, accuracy. B, AUC. C, precision. D, recall.

3.3 Model ablation experiments

The use of only fully connected layers to predict tongue, facial, and clinical data resulted in poor performance (accuracy = 0.552, AUC = 0.557, precision = 0.805, and F1 score = 0.561). The full model demonstrated significantly better performance, with accuracy values exceeding 0.800. However, the individual performance of the residual, self-attention, and adaptive modules alone was suboptimal (Table 3). These findings highlight the critical role of integrating the adaptive and image convolution modules in the model construction process.

Table 3 Ablation experiment of the model module

Model	Accuracy	AUC	Precision	Recall	F1 score
None	0.552	0.557	0.805	0.421	0.561
Res	0.689	0.863	0.916	0.578	0.705
SA	0.758	0.857	0.928	0.684	0.789
Ada	0.598	0.615	0.815	0.503	0.621
Res + SA	0.793	0.936	0.933	0.736	0.826
Res + Ada	0.804	0.913	0.901	0.817	0.857
SA + Ada	0.835	0.887	0.804	0.834	0.819
Res + SA + Ada	0.937	0.973	0.937	0.929	0.933

Res is the residual module, SA is the self-attention module, and Ada is the adaptive module.

3.4 Model data ablation experiments

The combination of tongue, facial, and laboratory data achieved the best predictive performance. Only clinical data resulted in moderate performance (precision = 0.851, AUC = 0.784, recall = 0.894, F1 score = 0.827, and accuracy = 0.827). Combining clinical data with tongue image data yielded a more remarkable accuracy increase than clinical data in combination with facial image data (Table 4).

Table 4 Ablation experimental results of model data

Model	Precision	AUC	Accuracy	Recall	F1 score
Lab + tongue + face	0.937	0.973	0.937	0.929	0.875
Lab + tongue	0.875	0.831	0.758	0.736	0.758
Lab + face	0.853	0.947	0.827	0.894	0.827
Lab only	0.851	0.784	0.827	0.894	0.827

3.5 Comparison of the AWMDF model parameters with other machine learning models

The SVM and KNN models demonstrated relatively strong performance, with prediction accuracy exceeding 0.8 as compared with the results of machine learning models trained on feature-selected data. However, the AWMDF model outperformed machine learning classification models, achieving superior predictive accuracy and robustness (Figure 4).

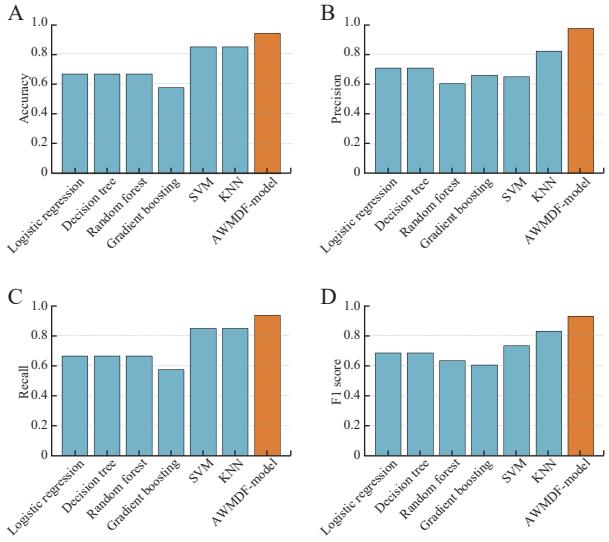

Figure 4 Comparison of the AWMDF model parameters with different machine learning models

A, accuracy. B, precision. C, recall. D, F1 score.

3.6 Comparison of different fusion methods

This study compared four different fusion methods with the adaptive fusion approach. The first method was SimpleWeight, a linear fusion based on the size of each data modality's dimensions. Three alternative fusion strategies were employed for comparison: (i) Average Weight, in which equal weights were assigned to each data modality (lab = 0.34, face = 0.33, tongue = 0.33); (ii) Voting, where the output weights were adjusted using a voting mechanism; and (iii) FixedWeight, which assigned fixed fusion weights to each modality (lab = 0.5, face = 0.25, tongue = 0.25).

The adaptive fusion module demonstrated better performance across accuracy, precision, recall, and F1 score. The FixedWeight model outperformed the SimpleWeight, AverageWeight, and Voting strategies. However, methods relying solely on linear weighting (SimpleWeight), equal weighting (AverageWeight), or voting mechanisms (Voting) exhibited suboptimal performance (Table 5).

Table 5 Parameters of different fusion methods

Method	Accuracy	AUC	Precision	Recall	F1 score
SimpleWeight	0.733	0.861	0.808	0.801	0.804
AverageWeight	0.733	0.980	0.923	0.687	0.791
Voting	0.739	0.689	0.753	0.901	0.819
FixedWeight	0.809	0.949	0.875	0.796	0.834
AdaptiveWeight	0.937	0.973	0.937	0.929	0.933

3.7 Visualization of deep learning results of tongue and face images

The CAM heatmap indicates that the patient's central facial region was the AWMDF model's primary focus during the model's training and prediction process (Figure 5). Key areas of attention included the nasal region, cheekbones, and parts of the forehead. This visualization highlights the model's ability to identify critical facial data regions that contribute significantly to its predictions. Additionally, the CAM heatmap demonstrates that the model effectively learned features across most areas of the tongue body during training. The model concentrated on the tongue coating, highlighting its significance in prediction.

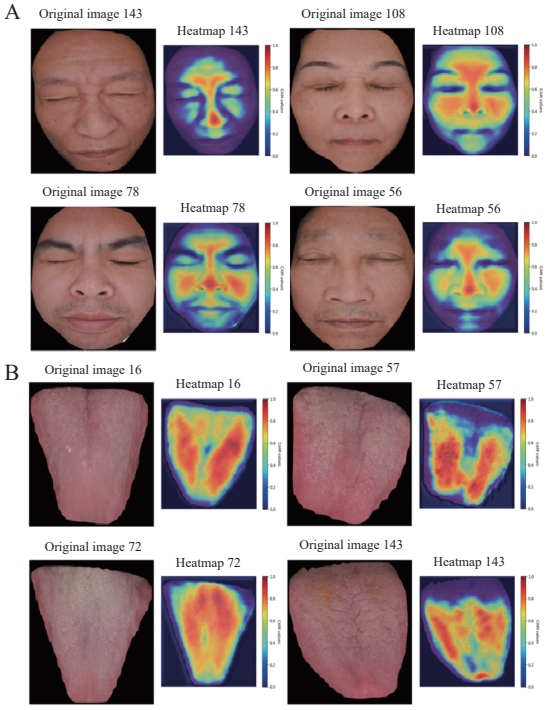

Figure 5 Multi-modal feature visualization based on Grad-CAM

A, heatmap of facial features. B, heatmap of tongue features.

4 Discussion

4.1 Model assessment and investigation

From a diagnostic perspective, facial diagnosis—part of TCM's inspectional methods—provides valuable information related to CAD by assessing changes in facial complexion and luster—pulse diagnosis. Evaluating the pulse characteristics reflects the cardiovascular system's condition and offers unique advantages in assessing vascular stenosis risk—tongue diagnosis. Observing changes in the tongue coating and body provides insights into intrinsic conditions associated with CAD. Research has found that the tongue features of patients with CAD exhibit certain specificity, particularly in the differences in the texture of tongue coating and parameters in various color spaces [16]. Furthermore, recent studies have shown that facial complexion indices can serve as objective clinical indicators for the differentiation of TCM syndromes in CAD patients [6, 17, 18]. Integrating TCM diagnostic approaches enhances CAD risk assessment by providing multidimensional and comprehensive insights that compensate for the limitations of single-method western diagnostic paradigms. The growing intelligence and modernization of TCM diagnostics further underscore the advancements in this field [19].

This study predicts high-risk coronary artery stenosis (⩾ 75% blockage) by integrating tongue and facial images with clinical data. The developed model demonstrates excellent binary classification performance for vascular obstruction, achieving an AUC of 0.973, accuracy of 0.937, precision of 0.937, and recall of 0.929 on the test set. These results highlight the effectiveness of data fusion in classifying coronary artery stenosis. Ablation experiments revealed that the individual residual module, self-attention module, or adaptive module demonstrated poor performance. However, combining residual or self-attention modules with the adaptive module at the decision layer improved prediction accuracy substantially. The adaptive tongue texture module enables the model to identify critical features during multimodal data fusion autonomously, avoiding information loss caused by single-weight connections [20, 21].

4.2 Multi-modal fusion and data ablation

In the multi-modal fusion of models, marked progress has been made in recent years. For instance, transformer-based attention mechanisms can adjust the weight relationships among input data elements. Research has shown that the dynamic transformer fusion mechanism achieved sensitivity and specificity both above 95% [22]. Additionally, the dynamic routing fusion method is a technique used in deep learning to strengthen the relationships among features and improve model performance. It aims to address the issues arising from the loss of translational invariance in convolutional neural networks (CNNs). The capsule mechanism establishes dynamic inter-feature connections, enabling the model to learn feature relationships. Research has shown that this method has achieved favorable results in diagnosing mild cognitive impairment in Alzheimer's disease patients [23]. However, in this study, facial and tongue diagnostic features exhibit inherent independence in TCM practice. Different modal data have certain independence, and the attention to the interaction in the data extraction stage is weak. Therefore, different from the focus on the content between data using the existing fusion mode, the back-end fusion mode of this model is more in line with the clinical dialectical reality. In the experimental study, the adaptive fusion achieved good results.

Ablation studies also showed that combining clinical data with tongue images achieved higher prediction accuracy than clinical data in combination with facial images, indicating that tongue images play a more critical role in predicting coronary artery stenosis. The importance of tongue image information in diagnosing CAD has also been manifested in recent clinical studies [7, 24, 25].

To test the model's ability to learn from raw images, feature extraction of tongue and facial images using machine learning was compared. The AWMDF model significantly outperformed conventional machine learning approaches. In earlier studies, important tongue and facial features were extracted mainly by analyzing color components in different color spaces. The superior performance of the proposed model is likely to stem from its ability to extract more crucial predictive features via residual and self-attention modules.

4.3 Interpretation mechanism of CAM-based model

A heatmap mechanism was employed to promote interpretability. The extracted tongue features covered most of the tongue body and coating, which centers more on the coating. The extracted tongue features suggest that coating thickness may be related to the progression of CAD. Studies have shown that gut microbiota varies among CAD patients with greasy and non-greasy tongue coatings, suggesting that these bacterial differences could serve as potential biological markers or factors influencing greasy tongue coating formation in these patients [26]. Furthermore, there is a correlation between gastrointestinal health and coronary artery lesions [27]. Facial feature extraction revealed that the model concentrated primarily on the nose, cheekbones, and forehead—regions with more prosperous blood supply and microcirculation. This finding is consistent with previous research reporting differences in these regions for CAD diagnosis [28].

Integrating non-invasive diagnostic data from TCM and modern clinical data considerably enhances the accuracy of CAD risk assessment, which holds promise for clinical applications. Research on TCM diagnostics modernization, mostly through multi-modal data, is rapidly expanding [29]. This study provides new perspectives for integrating TCM diagnostic methods into modern medicine and introduces innovative, non-invasive techniques for diagnosing CAD. This study has some limitations, such as a small patient sample size. Expanding the dataset, particularly incorporating incomplete data samples (e.g., missing tongue or facial image data), will improve real-world reliability and validity. In the model optimization, this study applied regularization techniques to constrain the model and prevent overfitting. Despite the implementation of regularization to mitigate overfitting, the SMOTE-based oversampling technique yielded good F1 score performance, indicating the model's strong learning ability for minority class samples. However, the issue of increasing datasets remains an area that needs to be explored in future research. During model training, some fluctuations were observed, which suggests that the sample size needs to be increased in subsequent studies to achieve better model fitting and more accurate predictions in real-world scenarios.

Additionally, this study focused on model fusion strategies and multi-modal data integration but lacked external validation. Future work will warrant multi-center, multi-hospital datasets to strengthen generalizability. The research team will further integrate additional TCM diagnostic dimensions, such as pulse diagnosis, to optimize and validate this method, aiming to provide clinicians with a more precise and effective diagnostic tool.

5 Conclusion

This study developed an AWMDF model based on multi-modal data in a non-invasive manner to predict the degree of vascular stenosis in patients with CAD by integrating data and image information, achieving high predictive performance. The adaptive weighting method demonstrated effective performance in backend fusion of multi-modal data. This model can serve as a non-invasive clinical auxiliary diagnostic tool.

Fundings

Construction Program of the Key Discipline of State Administration of Traditional Chinese Medicine of China (ZYYZDXK-2023069), Research Project of Shanghai Municipal Health Commission (2024QN018), and Shanghai University of Traditional Chinese Medicine Science and Technology Development Program (23KFL005).

Acknowledgements

This study was funded by the College of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, and the Center for Information Science and Technology of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine. Thanks to the clinical data support by Shanghai Tenth People's Hospital.

Competing interests

Jiatuo XU is an editorial board member for *Digital Chinese Medicine* and was not involved in the editorialreview or the decision to publish this article. All authorsdeclare that there are no competing interests.

References

[1] LIU MB, HE XY, YANG XH, et al. Interpretation of report on cardiovascular health and diseases in China 2023. Journal of

Clinical Cardiology, 2024, 40(8): 599–616.
[2] MA C, XU S, LIU YD. Non-invasive diagnosis of coronary heart disease based on integrated optimized kernel extreme learning machine. Computer Applications and Research, 2017, 34(6): 1671–1676.
[3] LI F, CHEN Y, XU HZ. Coronary heart disease prediction based on hybrid deep learning. The Review of Scientific Instruments, 2024, 95(1): 015115.
[4] NISHI T, YAMASHITA R, IMURA S, et al. Deep learning-based intravascular ultrasound segmentation for the assessment of coronary artery disease. International Journal of Cardiology, 2021, 333: 55–59.
[5] WANG YQ, GUO R, XU CX, et al. Application of objective research in the four diagnostic methods of traditional Chinese medicine in the diagnosis of coronary heart disease. Chinese Medicine Journal, 2016, 57(3): 199–203.
[6] DU B, HU YH, JIANG YC, et al. Ideas and discussions on the clinical application of facial recognition in diagnosing coronary heart disease with blood stasis syndrome. World Journal of Traditional Chinese and Western Medicine, 2020, 15(2): 377–380.
[7] WANG J, JIANG LH, CHANG TY, et al. Progress in the application of traditional Chinese medicine tongue diagnosis in coronary heart disease. Practical Journal of Heart, Brain, Lung, and Vascular Diseases, 2024, 32(3): 99–102.
[8] ZHANG HF, LU XZ, YU ZF, et al. Discussion on the pulse changes in 289 patients with coronary heart disease. Western Journal of Traditional Chinese Medicine, 2017, 30(6): 1–3.
[9] SUN YY, JIA ZT, ZHU HY, A review of multimodal deep learning. Computer Engineering and Applications, 2020, 56(21): 1-10.
[10] CHEN HZ, ZHONG NS, LU ZY, et al. Internal Medicine, 9th Edition. Beijing: People's Medical Publishing House, 2018: 213.
[11] European Society of Cardiology (ESC). 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. European Heart Journal, 2020, 41(3): 407–477.
[12] CHAWLA NV, BOWYER KW, HALL LO, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357.
[13] SELVARAJU RR, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 2020, 128(2): 336–359.
[14] GUAN Q, XU Y, YANG S, et al. Study on the relationship between traditional Chinese medicine facial diagnosis features and diseases. Chinese Journal of Traditional Chinese Medicine, 2022, 37(2): 902–905.
[15] XU JT, JIANG T, LIU S. Current status and outlook on tongue diagnosis analysis based on machine learning. Digital Chinese Medicine, 2024, 7(1): 3–12.
[16] DUAN MY, WANG CH, TAN YQ, et al. Objective study of tongue feature characteristics in 315 patients with coronary heart disease. Chinese Journal of Traditional Chinese Medicine, 2024, 65(9): 921–927.
[17] CHEN C, HONG J, DING XD, et al. Analysis of facial diagnosis image feature parameters in coronary heart disease with phlegm and blood stasis syndrome. Shizhen Journal of Traditional Chinese Medicine and Materia Medica, 2019, 30(7): 1768–1770.
[18] CAO YY, LI FF, WANG YQ, et al. Study on clinical differentiation of facial diagnosis color features in coronary heart disease patients. Chinese Journal of Traditional Chinese Medicine, 2013, 31(9): 1867–1869.
[19] ZHANG SQ, SUN YH, XIAN NX, et al. Progress in the objective and intelligent research of the four diagnostic methods of traditional Chinese medicine. Chinese Medicine Herald, 2023, 29(6): 170–174.
[20] YIN TZ, PENG YK, LI YY, et al. A multilevel cooperative attention network of precise quantitative analysis for predicting ractopamine concentration via adaptive weighted feature selection and multichannel feature fusion. Food Chemistry, 2025, 464: 141884.
[21] WEN RH, LIU CY, LIU S, et al. Adaptive weight detail-preserving multi-exposure image fusion. Laser and Optoelectronics Progress, 2024, 61(18): 326–335.
[22] DINESHKUMAR R, AMEELIA A, KANTH T, et al. Adaptive transformer-based multi-modal image fusion for real-time medical diagnosis and object detection. International Journal of Computational and Experimental Science and Engineering, 2024, 10(4): 890-897.
[23] JIAO ZC, HUANG P, KAM TE, et al. Dynamic routing capsule networks for mild cognitive impairment diagnosis. Medical Image Computing and Computer-Assisted Intervention, 2019, 2019: 620–628.
[24] DUAN MY. Research on machine learning-assisted diagnosis of coronary heart disease in traditional Chinese and western medicine based on tongue feature. Beijing: Beijing University of Chinese Medicine, 2023.
[25] YANG JH, XIAN NX, ZHANG SQ, et al. Objective study on tongue feature characteristics in age-related coronary heart disease with phlegm and blood stasis syndrome. Chinese Journal of Traditional Chinese Medicine, 2025, 43(3): 56-62I0010.
[26] LI X, CAO XT, SHI Y, et al. Study the correlation between gut microbiota structure and tongue coating in coronary heart disease patients. Lishizhen Medicine and Materia Medica Research, 2022, 33(1): 151–154.
[27] XU JW, SUN W. Discussion on the correlation between the degree of coronary artery lesions and gut microbiota in coronary heart disease patients. Modern Medicine and Health Research Electronic Journal, 2024, 8(20): 117-119.
[28] REN Q, TANG FF, ZHOU X, et al. Preliminary study on the objective analysis of facial diagnosis in coronary heart disease with phlegm and blood stasis syndrome. Chinese Journal of Basic Medicine in Traditional Chinese Medicine, 2020, 26(9): 1280–1283.
[29] LI HZ, YANG WW, QU H, et al. Thoughts on the integrated Chinese and western medicine multimodal diagnostic model. Journal of Integrated Traditional Chinese and Western Medicine for Cardiovascular Diseases, 2023, 21(21): 4020–4024.

基于自适应权重多模态中西医数据融合方法的冠心病血管阻塞程度预测模型的构建与评价

张冀豫, 许家佗*, 屠立平, 付洪媛

上海中医药大学中医学院, 上海 200120, 中国

【摘要】目的 基于中西医多模态数据的自适应融合，构建一种用于预测冠状动脉狭窄严重程度的无创模型。**方法** 收集 2023 年 5 月 1 日至 2024 年 5 月 1 日期间，在上海市第十人民医院心脏重症监护病房（CCU）接受冠状动脉计算机断层扫描血管造影（CTA）检查患者的临床指标、超声心动图数据、中医舌象特征及面部特征信息。基于深度学习构建了一个自适应加权多模态数据融合（AWMDF）模型，以预测冠状动脉狭窄的严重程度。采用准确率、精确率、召回率、F1 值及受试者工作特征曲线下面积（AUC）等指标对模型进行评估。通过与六种集成机器学习方法的比较、数据消融实验、模型组件消融实验及多种决策层融合策略进一步评估模型性能。**结果** 研究共纳入 158 例患者。AWMDF 模型具有优异的预测性能（AUC = 0.973，准确率 = 0.937，精确率 = 0.937，召回率 = 0.929，F1 值 = 0.933）。与模型消融、数据消融实验及多种传统机器学习模型比较结果显示，AWMDF 模型性能更出色。此外，自适应加权策略优于简单加权、平均法、投票法及固定权重等替代方案。**结论** AWMDF 模型对冠心病无创化预测有一定价值，可作为临床辅助诊断。

【关键词】冠心病；深度学习；多模态；临床预测；中医诊断