ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Construction of type 2 diabetes risk early warning model based on multi-modal data fusion

Wenjing Fu[1], *Wei Wang*[3] *and Jing Xia*[1,2,*]

[1] School of Medicine, Hainan Vocational University of Science and Technology, Haikou 571126, Hainan, China
[2] School of Pharmacy, China Medical University, Shenyang 110122, Liaoning, China
[3] School of Basic Medicine, Xinjiang Medical University, Urumqi 830017, Xinjiang, China

SUMMARY: *In order to improve the early identification ability of high-risk people with type 2 diabetes mellitus, a risk warning model based on multi-modal data fusion was constructed. In this study, the clinical indicators, continuous monitoring signals, lifestyle information and fundus image data of 3126 subjects were integrated. After preprocessing, single-modal feature extraction, dynamic weighted fusion and gated restructuring, the risk warning framework of type 2 diabetes mellitus was established. The results show that the accuracy, F1 value and AUC of the proposed model on the test set reach 0.934, 0.906 and 0.978, respectively, which is better than that of the traditional machine learning model, the single-modal deep model and the conventional early fusion model. This method can more fully characterize the individual metabolic risk characteristics and provide a feasible computational support for the early screening and early warning of type 2 diabetes.*

KEYWORDS: *Type 2 diabetes mellitus; Multi-modal data fusion; Risk early warning model; Deep learning*

1 Introduction

Type 2 diabetes mellitus (T2DM) is a chronic metabolic disease with glucose metabolism disorder as the core manifestation. Its pathogenesis is usually related to insufficient insulin secretion, enhanced insulin resistance and imbalance of glucose and lipid metabolism. Mohsen et al. pointed out in their review that the current research on diabetes risk prediction has gradually shifted from single clinical indicator discrimination to a comprehensive risk identification framework supported by artificial intelligence [1]. Different from type 1 diabetes mellitus with more acute onset and more prominent immune damage characteristics, type 2 diabetes mellitus is often in a state of insidious progression, and patients often lack sufficient clinical perception in the early stage of the disease, which makes its risk identification and early warning have higher practical necessity. Aoki et al. believe that there is usually a window period for intervention between abnormal glucose metabolism and clear diagnosis, and if risk stratification and behavioral intervention can be implemented in this stage, it will help to delay the progression of the disease [2].

From the perspective of clinical manifestations, type 2 diabetes is not only a single point problem of abnormal blood glucose indicators, but also a complex disease related to obesity, lipid disorders, abnormal blood pressure, inflammatory response, lifestyle imbalance and

*abcjx133@126.com
https://doi.org/10.65102/is2026794
Published: 30 April, 2026

genetic susceptibility. Based on the analysis of cohort data, Talebi Moghaddam et al. found that variables such as age, body mass index, fasting blood glucose, blood lipid levels and family history had continuous and stable identification value in the prediction of diabetes [3]. Hoyos et al. further pointed out that the formation of diabetes risk is not the result of linear accumulation of a single factor, but is closer to the nonlinear evolution process under the joint action of multiple physiological indicators, behavioral characteristics and environmental exposure [4]. This means that if coarse-grained judgments still rely on a small number of structured variables, it is often difficult to fully present the true differences of individual risk status.

Once type 2 diabetes enters the continuous progression stage, it often causes cardiovascular damage, kidney damage, retinopathy and neurological complications, and the related health burden will be further extended in the direction of long-term and complex. Duckworth et al. pointed out in their study on real-time abnormal blood glucose prediction that the risk associated with diabetes is not limited to the diagnosis of the disease itself, but also reflected in the continuous clinical consequences of subsequent hyperglycemia, hypoglycemia and individualized control imbalance [5]. Han et al. also proposed that the formation of complications in diabetes patients has significant stage accumulation characteristics, and if a more sensitive early warning model can be constructed in the front-end, it will help to improve the pertinence of subsequent diagnosis and treatment decisions [6]. Therefore, compared with "post-diagnosis management", "pre-diagnosis warning" around high-risk groups is more in line with the development direction of moving the focus of chronic disease prevention and control.

In this context, machine learning and deep learning methods provide a new technical path for type 2 diabetes risk identification. Tuppad et al. believe that machine learning can extract complex patterns from clinical data that are difficult to be revealed by traditional statistical methods, thus improving the sensitivity and adaptability of risk assessment [7]. Talukder et al. pointed out that the improvement of diabetes prediction models in recent years is no longer at the level of classifier replacement, but increasingly relies on the collaborative optimization of data engineering, feature construction and model fusion strategies [8]. Tanabe et al. showed that the prediction of diabetes subtypes based on machine learning has obvious advantages in revealing individual heterogeneity, which also indicates that the risk identification model should pay attention to the complex structure within the sample instead of only pursuing the overall classification accuracy [9].

However, there are still obvious limitations in the existing research. One class of studies mainly relies on structured indicators in electronic medical records, which are convenient for modeling, but underutilize heterogeneous data such as imaging information, behavioral trajectories and text descriptions. Ding et al. pointed out when using large language multimodal models to carry out diabetes prediction, that a single data source is difficult to fully cover multi-level signals in the process of disease formation, and multimodal joint modeling is more likely to improve the ability to identify new cases [10]. The other type of research has introduced interpretable methods or hybrid deep networks, but mostly focuses on single-modal feature enhancement, and has not fully solved the problems of cross-modal feature scale inconsistency, large differences in semantic expression, and information redundancy superposition. Nguyen et al. proposed that explainable hybrid deep learning has a good performance in pre-diabetes prediction, but the model performance is still affected by feature organization and input dimension structure [11]. Talari et al. also believe that the improvement of early prediction effect depends not only on the classifier itself, but also on the design quality of feature screening and fusion mechanism [12].

It is worth noting that multimodal data fusion provides a more potential computer

modeling basis for type 2 diabetes risk early warning. The so-called multimodal data includes not only phenotypic and laboratory indicators such as age, body mass index, blood glucose, blood lipid, and blood pressure, but also physical examination images, physiological monitoring signals, and lifestyle records. In the study of joint modeling of fundus images and traditional risk factors, Lee et al. confirmed that there was a complementary relationship between image modalities and conventional clinical modalities, and reasonable fusion could enhance the robustness of risk prediction [13]. Baharoon et al. also found in the hypertension multimodal recognition system that the model's ability to capture complex risk representations was significantly enhanced after the joint input of visual information and cardiometabolar risk factors [14]. This kind of research provides a reference for the construction of type 2 diabetes risk early warning model, which uses multi-source data to three-dimensional characterize individual metabolic risk, and then uses deep learning to complete cross-modal feature alignment, weight allocation and joint discrimination.

Based on the above understanding, this paper focuses on the main line of "multimodal data fusion - risk representation learning - early warning model construction", and intends to construct an intelligent early warning model for type 2 diabetes risk recognition. At the data level, the model comprehensively incorporated structured clinical indicators, continuous monitoring information, lifestyle information and fundus image features. At the method level, high-dimensional risk representation was realized through single-modal feature extraction and cross-modal fusion mechanism. Wang et al. pointed out that the real value of artificial intelligence in diabetes care does not lie in replacing clinical judgment, but in providing risk support tools with earlier, finer and more individual differences perception ability [15]. Based on this, this paper attempts to further answer two questions on the basis of existing research. First, whether multimodal information can improve the performance of type 2 diabetes risk warning more effectively than single-modal input. Secondly, whether the fusion modeling framework can enhance the stability and application feasibility of the model while ensuring the prediction accuracy. Focusing on these two problems will not only help to improve the technical path of early identification of type 2 diabetes, but also provide a transferable method reference for intelligent early warning research of chronic metabolic diseases.

2 Materials and Methods

2.1 Multi-modal data acquisition

In order to improve the ability of type 2 diabetes risk early warning model to describe individual metabolic status, this paper constructed a multimodal data acquisition system including structured clinical indicators, continuous monitoring signals, lifestyle information and fundus imaging features before modeling (as shown in Figure 1). The data were collected from the physical examination center, endocrine clinic follow-up database and health management platform records of a tertiary hospital, and the collection time span was from January 2021 to June 2024. Participants aged 18 years and older were included in the study. The inclusion criteria included having a complete basic physical examination record, completing at least one fasting blood glucose or glycated hemoglobin test within the past 12 months, and keeping lifestyle questionnaire or wearable device monitoring data. To ensure sample comparability, patients with gestational diabetes mellitus, type 1 diabetes mellitus, severe liver and kidney dysfunction, and individuals with more than 30% missing key modality information were not included in the study.

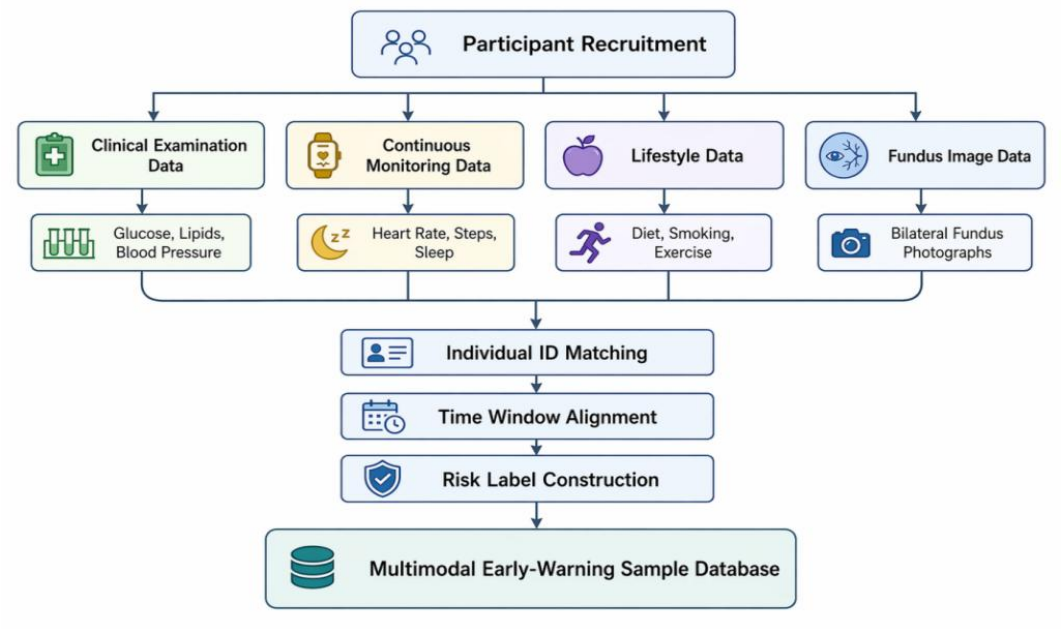

Figure 1: Framework of multimodal data acquisition

In terms of specific collection content, the structured clinical modality mainly included age, gender, body mass index, waist circumference, systolic blood pressure, diastolic blood pressure, fasting blood glucose, 2 h postprandial blood glucose, glycosylated hemoglobin, triglyceride, high density lipoprotein cholesterol, low density lipoprotein cholesterol and family history of diabetes. The continuous monitoring mode mainly recorded resting heart rate, average daily steps, sleep duration, sleep interruption times and ambulatory blood pressure fluctuation information to reflect the individual's daily physiological load and behavioral rhythm. The lifestyle mode focused on collecting dietary preference, frequency of sugar-sweetened beverage intake, smoking and drinking, frequency of exercise and sedentary time. In the imaging modality, the fundus color photos of both eyes acquired synchronously during the physical examination were selected to capture the morphology of microvessels, the texture around the optic disc and the subtle lesion signals. The collection strategy takes into account the biochemical characteristics of metabolic abnormalities, life exposure, behavioral patterns and visual pathological cues, which is closer to the real path of type 2 diabetes risk than a single test index.

In order to achieve consistent management of cross-modal samples, this paper uses the unique number of the subject as the index, performs time alignment and individual matching of data from different sources, and defines the observation sample of the ith subject as:

$$X_i = \{C_i, S_i, L_i, I_i\} \tag{1}$$

Here, C_i represents structured clinical features, S_i represents continuous monitoring signal features, L_i represents lifestyle features, and I_i represents fundus imaging features. Based on the clinical diagnostic criteria and follow-up results, the risk label is denoted as y_i in this paper. When the subjects meet the fasting blood glucose ≥7.0mmol/L, glycated hemoglobin ≥6.5%, or have been clinically diagnosed as type 2 diabetes mellitus, they are denoted as high-risk samples, namely:

$$y_i = \begin{cases} 1, \text{High risk} \\ 0, \text{Low risk} \end{cases} \tag{2}$$

This annotation method makes the model training not only retain the medical judgment basis, but also provide a clear label boundary for subsequent supervised learning. After sample screening and number integration, 3126 valid subject data were finally formed, including 1128 high-risk samples and 1998 low-risk samples. After the completion of the collection, the data of each modality were entered into the unified database management platform, the textual questionnaire information was stored in structured fields, the monitoring sequence was summarized according to the daily granularity, and the image data was uniformly converted into PNG format and the original resolution was retained. The purpose of this processing is not to simply expand the input dimension, but to incorporate the complementary information that can reflect the metabolic risk in different modalities into the same computational framework, so as to provide a stable data basis for subsequent single-modal feature extraction and multi-modal fusion modeling.

2.2 Data preprocessing

There are obvious differences in the source, scale, sampling frequency and information density of multi-modal data. If the original data is directly input into the risk early warning model, it is easy to cause the imbalance of feature distribution, the superposition of noise between modalities and the bias of the training process. Therefore, after sample collection and number matching are completed, this paper implements unified preprocessing of structured clinical data, continuous monitoring signals, lifestyle records and fundus imaging data, and the overall process is shown in Figure 2. This process is not just a mechanical repair of outliers and missing values, but a standardization of data representation of different modalities around the modeling requirements of "computable, aligned, and comparable".

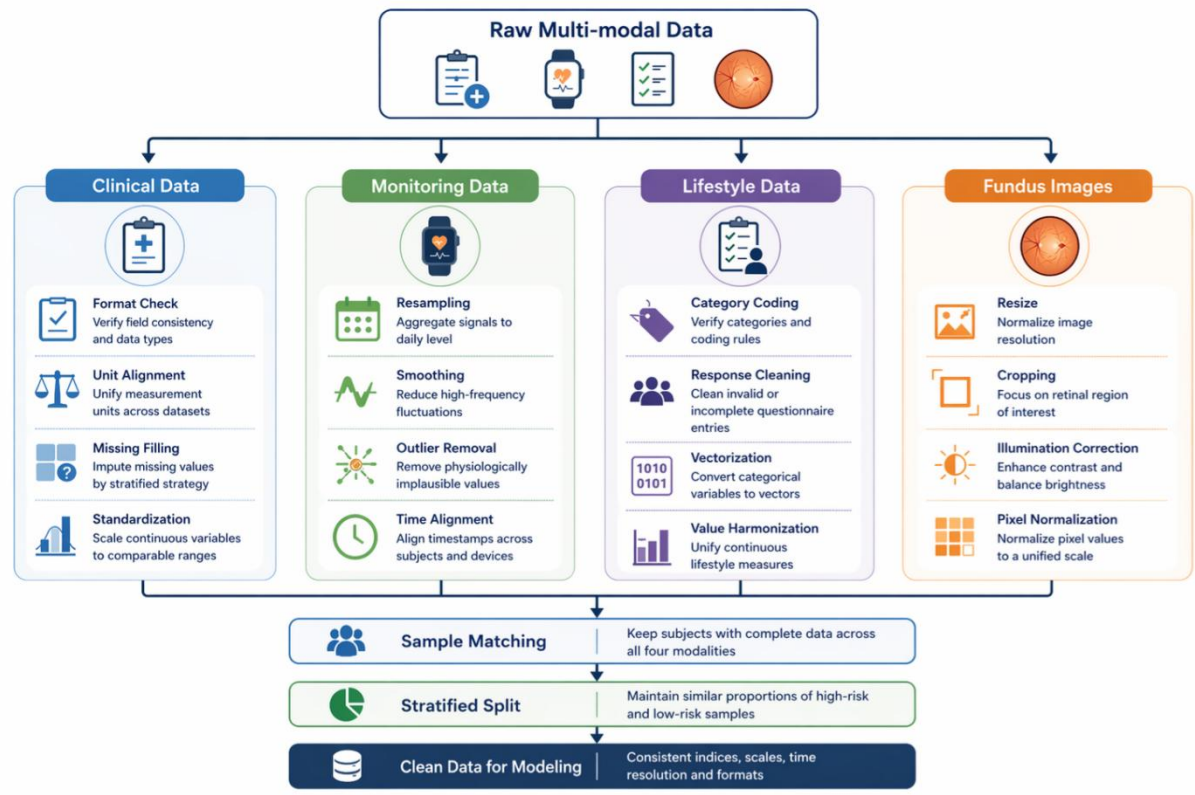

Figure 2: Data preprocessing process

For structured clinical variables and lifestyle variables, this paper first carried out field

consistency verification and dimension unification processing, and transformed continuous variables such as blood glucose, blood lipid, blood pressure, and body mass index into comparable standardized forms. Let the original value of the JTH feature in the ith sample be x_{ij}, and the standardized result is denoted as follows.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{3}$$

Here, $\bar{x}_j$ is the mean value of the feature in all samples, and s_j is the standard deviation. Through this transformation, different indicators can be compressed to a unified scale range, thereby reducing the dominant effect of a single high-dimensional variable in model training. For discrete variables such as smoking history, drinking frequency, and family history, one-hot encoding is used to convert them into numerical vectors, so as to retain category differences and avoid misdirection caused by artificial ranking. For missing data, this paper implements hierarchical processing according to missing mechanism and variable attribute. If too high a percentage of a field is missing, it will weaken its statistical stability, so the missing rate of the JTH feature is defined as:

$$r_j = \frac{n_j^{miss}}{N} \tag{4}$$

where n_j^{miss} is the number of missing samples for feature j and N is the total sample size. When $r_j > 0.30$, this field will not be included in the subsequent modeling. When the missing rate was in the acceptable interval, the median of the same age group and the same sex group were used to fill in the continuous variables, and the mode was used to fill in the discrete variables. This process preserves the overall distribution shape while minimizing the shift introduced by simple mean substitution.

Continuous monitoring signals have time series characteristics, which are susceptible to wear interruption, device drift and incidental noise. To this end, this paper resampled and summarized the data such as heart rate, steps, and sleep duration in a daily unit, and used a sliding window to smooth high-frequency fluctuations, and then eliminated the abnormal records with obvious distortion according to the reasonable range of medicine. The resolution adjustment, brightness correction, background cropping and pixel normalization of fundus image data are unified to enhance the identification of optic disc area, vascular texture and small lesion signal. Considering the synchronization requirement of image modality and phenotype modality in the number of samples, this paper only retained the records of subjects who had completed matching in all four types of modalities into the final sample database.

2.3 Single-modal feature extraction

After data preprocessing, although different modal information has unified input conditions, its internal structure differences are still obvious. Structured clinical variables emphasize numerical relationships and risk thresholds, continuous monitoring signals contain time dependence and fluctuation patterns, and fundus images are more dependent on spatial texture and local lesion expression. If the same extraction strategy is used for all types of inputs, it is easy to weaken the discriminative advantage of the mode itself. Therefore, this paper constructs corresponding single-modal feature extraction modules for different data types, and converts the original input into a low-dimensional representation suitable for subsequent fusion modeling.

For clinical physical examination indicators and lifestyle variables, this paper uses the fully connected mapping method to extract their nonlinear combination features. Let the structured input vector of the ith sample be ui. After linear mapping and activation function transformation, its unimodal representation is denoted as follows.

$$h_i^{(u)} = \phi(W_u u_i + b_u) \tag{5}$$

Here, W_u and b_u represent the weight matrix and bias term, respectively, and $\phi(\cdot)$ represents the nonlinear activation function. The expression does not only retain the original amplitude of a single variable, but describes the interaction between age, body mass index, blood glucose, blood lipid, blood pressure and behavioral variables through parameter learning, so as to enhance the representation ability of structured modalities for risk status. Since the formation of type 2 diabetes mellitus is usually related to the coordinated changes of multiple metabolic indicators, this mapping method is more beneficial to retain the implicit joint risk information than traditional manual screening. For continuous monitoring signals, in this paper, a gated recurrent unit is used to extract time-dependent features. Let the ith sample form a sequence $S_i = \{s_{i1}, s_{i2}, \ldots, s_{iT}\}$, whose hidden state update process is expressed as follows.

$$g_{it} = \mathrm{GRU}(s_{it}, g_{i,t-1}) \tag{6}$$

where g_{it} is the hidden state vector at time t. After the time dimension recursion, the hidden state g_{iT} at the last time is used as the temporal feature representation of the mode, namely:

$$h_i^{(s)} = g_{iT} \tag{7}$$

This process can preserve the long-term dependence of step change, sleep fluctuation, resting heart rate fluctuation and ambulatory blood pressure rhythm, so that the model is not limited to static average values, but can identify dynamic trends with early warning significance. For the risk of glucose metabolism, irregular fluctuations in continuous signals are often more indicative of underlying abnormalities than single measurement results, which is why the timing extraction module is necessary.

For fundus image modality, this paper uses convolutional neural network to extract visual features. Let the input fundus image be I_i and the convolutional extractor be denoted as $F_{cnn}(\cdot)$, then its feature vector is expressed as follows.

$$h_i^{(v)} = F_{cnn}(I_i) \tag{8}$$

The process gradually aggregates local texture, blood vessel distribution, morphology around the optic disc and small lesion area information through multi-layer convolution and pooling operation, so as to obtain a deep visual representation related to diabetic microvascular changes. Compared with the direct use of manual measurement indicators, the convolutional extraction method can automatically learn the risk-related image modes with fewer prior constraints, and improve the representation integrity of the image modes.

In order to facilitate different modalities to enter the subsequent fusion stage, the output of each single mode is further unified into an embedding vector of the same dimension, and the sample-level one-to-one correspondence is maintained. After the above processing, structural variables, temporal signals and fundus images form feature representations with clear semantic emphasis respectively. The former highlights metabolic states and behavioral

exposures, the middle reflects daily physiological rhythms and fluctuations, and the latter complements microscopic visual pathological cues. Single-modal feature extraction is not the final discriminative goal, but provides complementary input for multi-modal fusion.

2.4 Multi-modal feature fusion method

After the single-modal feature extraction, structured clinical features, continuous monitoring features and fundus image features have formed discriminative representation vectors respectively, but the three types of vectors are not consistent in dimension scale, statistical distribution and semantic level. If they are directly concatenated into the classifier, although the implementation is simple, it is prone to the problems of high-dimensional redundancy accumulation, excessive dominance of strong modes and the cover of weak association information. The risk of type 2 diabetes is not determined by a single indicator, but the result of metabolic state, behavioral rhythm and microvascular abnormalities. Therefore, the fusion process should not stop at the level of data superposition, but should turn to cross-modal semantic alignment and complementary information aggregation. Based on this understanding, this paper constructs a multi-modal feature fusion method of "unified mapping, weight assignment, gated recombination, and joint output", and its overall process is shown in Figure 3.

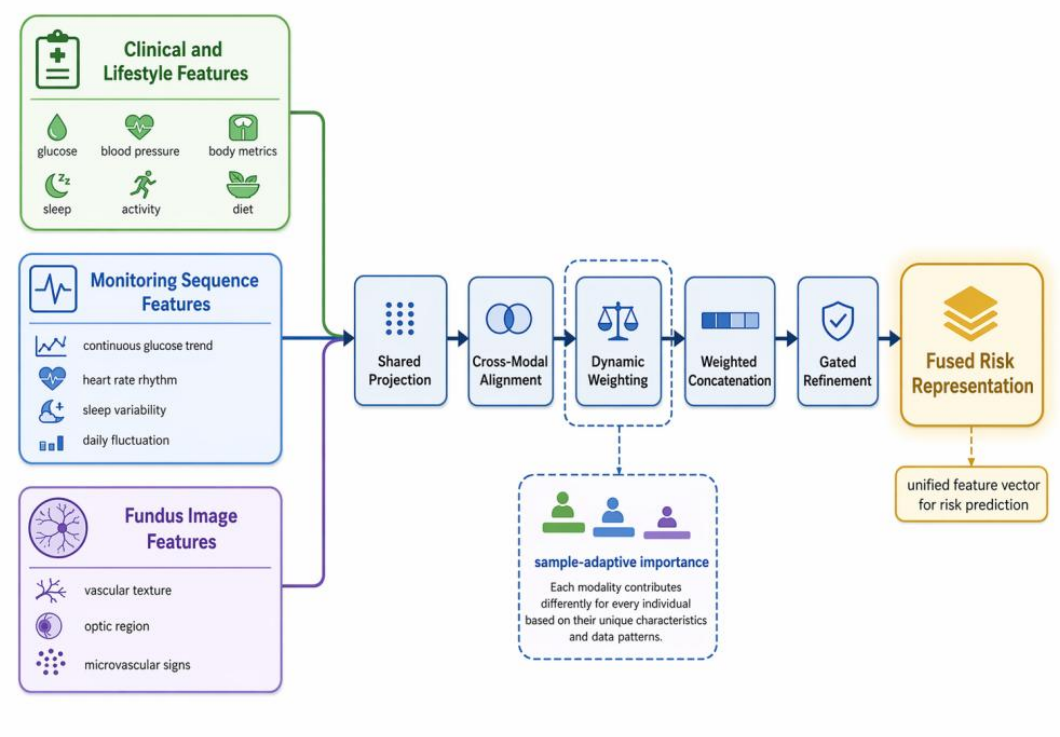

Figure 3: Process of multimodal feature fusion

2.4.1 Cross-modal alignment mechanism

Let the ith sample be extracted by single modality to obtain three types of feature vectors, denoted as h_i^c、h_i^s and h_i^y, respectively, corresponding to clinical and lifestyle modality, continuous monitoring modality and visual imaging modality. Since the feature dimensions of different modalities are not the same, this paper first maps them to the unified latent space. The projected representation of the m-th mode is written as follows.

$$q_i^m = P_m h_i^m + b_m, m \in \{c, s, v\} \tag{9}$$

Here, P_m represents the modal projection matrix and b_m represents the bias term. After this transformation, the risk information from different sources is compressed to the latent semantic space of the same scale, which provides conditions for the subsequent fusion operation. The unified mapping here is not a simple dimension reduction, but retains the core expression most related to type 2 diabetes risk through parameter learning, so that heterogeneous cues such as blood glucose, blood pressure, sleep fluctuations and fundus vascular texture can enter the same discriminant framework.

2.4.2 Integrating mathematical processes

After the alignment is completed, the attention weight mechanism is introduced to measure the contribution strength of each modality to the risk identification of the current sample. For the ith sample, the relevance score for the m-th modality is defined as follows.

$$e_i^m = u^\top \tanh(W_m q_i^m + d_m) \tag{10}$$

where u is the shared rating vector, W_m and d_m are learnable parameters. Then, the Softmax function is used to normalize the scores of each modality into weight coefficients:

$$\alpha_i^m = \frac{\exp(e_i^m)}{\sum_{k\in\{c,s,v\}} \exp(e_i^k)} \tag{11}$$

This weight reflects the difference in effectiveness of different modalities on the current sample. When the fundus image of a subject shows more obvious microvascular abnormalities, the visual modality weight will be increased accordingly. When the behavioral rhythm fluctuation is more prominent, the continuous monitoring mode will occupy a higher proportion in the fusion. The weight obtained in this way is not a preset constant, but dynamically adjusted with the state of the sample, so it is more in line with the actual characteristics of individual risk heterogeneity.

Based on the weight allocation, the three types of modes are combined by weighting, and the redundant noise is further suppressed by the gating unit. The weighted stitching result is denoted as follows.

$$r_i = [\alpha_i^c q_i^c \,\|\, \alpha_i^s q_i^s \,\|\, \alpha_i^v q_i^v] \tag{12}$$

Here, $\|$ denotes the vector concatenation operation. In order to avoid information accumulation caused by simple splicing, this paper constructs the gated vector:

$$g_i = \sigma\left(W_g r_i + b_g\right) \tag{13}$$

where $\sigma(\cdot)$ represents the Sigmoid function. The final fused representation is defined as follows.

$$z_i = g_i \odot r_i \tag{14}$$

Here, $\odot$ denotes the Hadamard element-wise product. The role of the gating mechanism is to retain the feature components with discriminative value, and weaken the repeated expression and unstable noise, so that the fusion result not only retains the multi-modal

complementary information, but also avoids the excessive expansion of invalid dimensions.

2.5 Construction of type 2 diabetes risk early warning model

After the multi-modal feature extraction and fusion, the focus of research is no longer the information compression within a single modality, but how to transform the joint representation into the risk warning results with medical interpretation significance. Type 2 diabetes risk identification is not a simple binary classification decision problem, its essence is to continuously estimate the individual metabolic abnormality propensity, and then form an early warning output according to the risk intensity. If only a single-layer linear mapping is used, the computational cost is low, but it is difficult to describe the nonlinear interaction in the fusion feature. If the model structure is too complex, it is easy to overfit on medium scale medical samples. Based on this, this paper constructs a type 2 diabetes risk early warning model composed of "fusion input-hidden layer representation-risk mapping-threshold discrimination", so that it can better identify the potential risk patterns in multimodal data while maintaining computational control. Let the joint feature vector z_i be obtained from the ith sample after multi-modal fusion, and its risk hidden representation is defined as follows.

$$m_i = \psi(W_1 z_i + b_1) \tag{15}$$

Here, W_1 is the hidden layer weight matrix, b_1 is the bias vector, and $\psi(\cdot)$ is the nonlinear activation function. The role of this step is to further compress the redundant components in the fused features while retaining the high-order combined information that is most relevant to type 2 diabetes risk. Since different modalities may still have local correlation overlap after fusion, a new intermediate expression is formed through hidden layer mapping, which is helpful to transform the linkage features between "blood glucose, behavior and image" into a compact representation that can be used for early warning discrimination.

Based on this, in this paper, the hidden representation is mapped to the probability value of the occurrence of high-risk events for the individual. The risk output function is written as:

$$p_i = \frac{1}{1 + \exp[-(W_2 m_i + b_2)]} \tag{16}$$

where $p_i \in (0,1)$ represents the predicted probability that the ith sample is judged to be a high-risk individual of type 2 diabetes, and W_2 and b_2 represent the output layer parameters respectively. This form makes the output of the model no longer stay at the rigid category label, but can give the continuous risk intensity, which provides a direct basis for hierarchical early warning. When p_i is closer to 1, it means that the high-risk features of individual metabolic abnormalities, behavioral exposure and imaging pathological cues in the joint space are more obvious. In order to transform continuous probabilities into early warning results that can be used in medical management, this paper sets up a risk discriminant function:

$$\hat{y}_i = \begin{cases} 1, p_i \geq \tau, \\ 0, p_i < \tau, \end{cases} \tag{17}$$

Here, τ represents the risk threshold, $\hat{y}_i = 1$ represents the high risk warning, and $\hat{y}_i = 0$ represents the low risk state. Considering that the screening of type 2 diabetes emphasizes the early identification of high-risk individuals, this paper does not simply fix the empirical threshold, but selects the optimal threshold based on the sensitivity, specificity and F1 value on the validation set. This processing can avoid the discrimination shift caused by the uneven class distribution, and make the model output more suitable for early warning scenarios. In

addition to binary early warning, the risk stratification results are further constructed to enhance the application readability of the model. Let the risk level function be G_i, then according to the predicted probability interval, it can be defined as follows.

$$G_i = \begin{cases} \text{Low}, 0 \le p_i < 0.35, \\ \text{Medium}, 0.35 \le p_i < 0.65 \\ \text{High}, 0.65 \le p_i \le 1. \end{cases} \tag{18}$$

This hierarchical way enables the model to not only identify "high risk or not", but also distinguish the difference in risk degree, which facilitates the implementation of differentiated intervention in subsequent health management. Follow-up observation is the main method for low-risk individuals, lifestyle adjustment and index review should be strengthened for medium-risk individuals, and high-risk individuals should enter the key screening and further diagnosis process.

From the perspective of the calculation process, the type 2 diabetes risk early warning model constructed in this paper takes the fusion feature as the unified input, obtains the risk representation through the hidden layer nonlinear mapping, and then outputs the individual high-risk probability through the Sigmoid function. Finally, the early warning judgment and risk stratification are completed according to the threshold. Compared with the traditional single classifier, this model does not rely on the absolute advantage of a certain modality, but uses the joint representation formed by different data sources as the basis of discrimination, so as to improve the recognition ability of complex metabolic abnormal states. Figure 4 presents the overall structure of the type 2 diabetes risk early warning model.

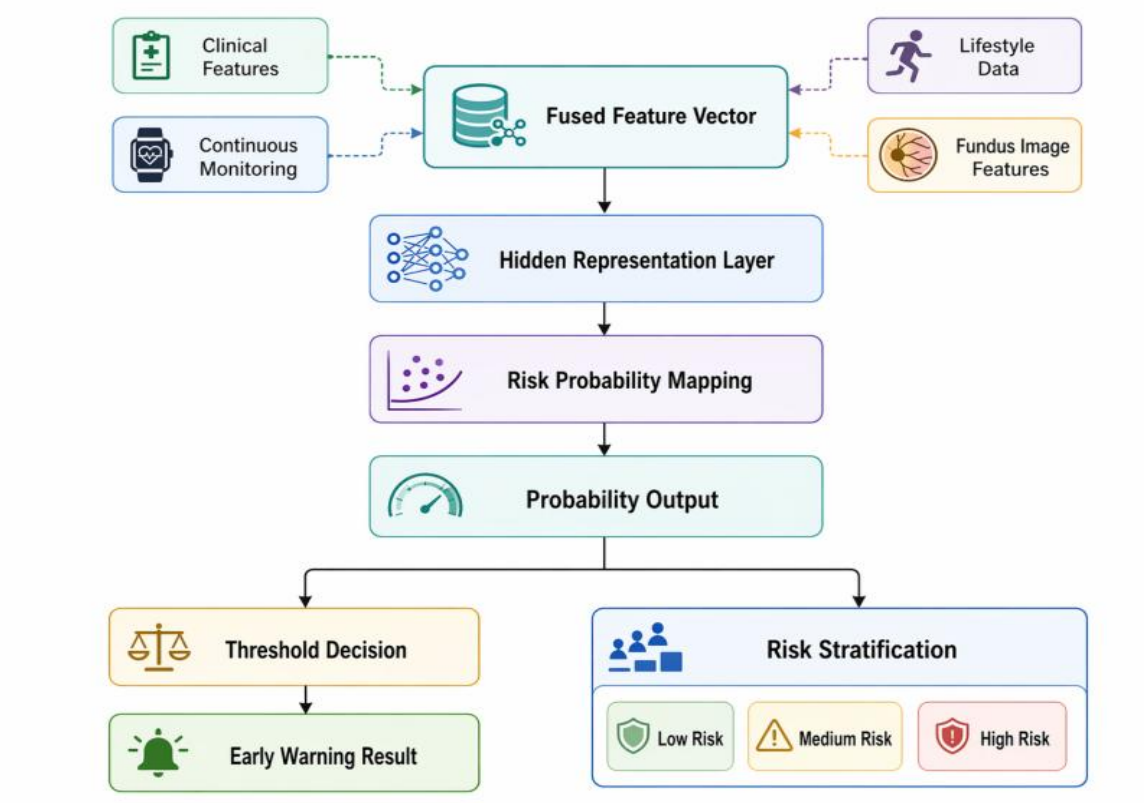

Figure 4: Structure of type 2 diabetes risk warning model

2.6 Model Training and parameter optimization

After the multi-modal feature fusion and risk early warning model construction is completed, the formation of model performance not only depends on the network structure itself, but also is affected by parameter initialization, loss function design, optimizer update method and termination condition setting. Type 2 diabetes risk warning is a typical medical binary classification task, and there is still a certain degree of unbalanced distribution between high-risk samples and low-risk samples. If the training process only aims at the overall classification accuracy, it is easy to lead to a bias towards the majority class of samples. Based on this, the model training process is designed uniformly around the goal of "stable convergence, suppression of overfitting, and improvement of high-risk identification ability", and the parameter optimization strategy is used to improve the generalization performance of early warning results.

If the number of training set samples is N, the true label of the ith sample is y_i, and the high risk probability output by the model is p_i, then this paper uses the weighted binary cross entropy loss function as the basic training objective:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[\omega_1 y_i \ln p_i + \omega_0(1-y_i)\ln(1-p_i)] \tag{19}$$

Here, ω_1 and ω_0 represent the class weights of high-risk and low-risk samples, respectively. This design can improve the sensitivity of the model to high-risk individuals when the number of classes is different, and avoid the training process from excessively following the dominant distribution of low-risk samples. In order to reduce the parameter oscillation and overfitting caused by complex models under the condition of limited medical sample size, this paper further adds L_2 regularization term to the objective function to form the final optimization objective:

$$\mathcal{J} = \mathcal{L}_{cls} + \lambda\|\Theta\|_2^2 \tag{20}$$

Here, Θ represents all the trainable parameters of the model and λ represents the regularization coefficient. In this way, the model will not only fit the discrimination boundary of the training sample, but also constrain the too large parameter amplitude, so as to enhance the training stability and the generalization ability of the test phase.

In the parameter update stage, this paper uses AdamW optimizer to complete the gradient iteration. For the gradient g_t in the TTH training iteration, the first moment and second moment estimates are written as follows.

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t, v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2 \tag{21}$$

Here, β_1 and β_2 are the momentum attenuation coefficients. After bias correction, the model parameters are updated as follows.

$$\Theta_{t+1} = \Theta_t - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{22}$$

Here, η_t represents the learning rate at round t and ε is the smoothing constant that prevents the denominator from being zero. Compared with the common stochastic gradient descent method, this update method can maintain a better convergence speed in the multi-modal high-dimensional parameter space, and reduce the interference of local

fluctuations on the training process. Considering the slow decline of loss in the later stage of training, cosine annealing strategy is used to dynamically adjust the learning rate, which is expressed as follows.

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\frac{\pi t}{T}\right) \tag{23}$$

Here, η_{max} and η_{min} represent the maximum and minimum learning rate, respectively, and T represents the total number of training rounds. This strategy can maintain a strong search ability in the early stage of training, and gradually reduce the step size in the middle and late stage, so that the model can approach the optimal parameter region more stably. At the same time, we set an early stopping mechanism: if the validation set loss does not continue to decrease in eight consecutive rounds of training, the training is terminated and the parameter combination with the best performance in the validation set is retained to avoid invalid iterations and overfitting accumulation. Combined with the sample size and model complexity mentioned above, the final training parameters determined in this paper are shown in Table 1. The relevant parameters are selected based on the comparison of multiple groups of experiments, which can better balance the training efficiency, convergence speed and early warning performance, and provide unified experimental conditions for subsequent results analysis.

Table 1: Model training and parameter optimization Settings

Parameter Category	Parameter Name	Setting Value
Data Split	Training Set: Validation Set: Test Set	7:1:2
Batch Processing Setting	Batch size	32
Number of Training Epochs	Max epochs	100
Optimizer	Optimizer	AdamW
Initial Learning Rate	η_{max}	2×10^{-4}
Minimum Learning Rate	η_{min}	1×10^{-6}
Momentum Parameters	β_1, β_2	0.9, 0.999
Regularization Coefficient	λ	1×10^{-4}
Early Stopping Patience	Patience	8
Risk Threshold Determination Method	Threshold selection	Threshold corresponding to the best F1 score on the validation set

2.7 Model evaluation metrics

In order to objectively evaluate the performance of the type 2 diabetes risk early warning model based on multi-modal data fusion, this paper constructs an evaluation index system from three levels: overall discrimination ability, positive class recognition ability and comprehensive balance performance. Since medical early warning tasks pay more attention to the early detection of high-risk individuals, the model evaluation cannot be judged only by a single accuracy rate, but also needs to investigate the detection ability of high-risk samples, the level of false alarm control, and the recognition balance between different categories. Based on this, this paper takes high-risk samples as the positive class, selects precision, recall, F1 value, specificity, AUC and MCC as the core indicators, and conducts a unified analysis combined with confusion matrix.

Let the true, false positive, true negative and false negative examples in the test set be denoted as TP, FP, TN and FN, respectively. Precision is a measure of the proportion of samples that the model classifies as high-risk that are truly high-risk, and is expressed as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{24}$$

The higher this metric is, the less false positives the model has. Recall is used to reflect the ability of the model to identify the true high-risk individuals and is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{25}$$

In type 2 diabetes early warning research, the recall rate is of high importance, because the missed diagnosis of high-risk individuals will directly weaken the clinical use value of the early warning system.

In order to measure the coordination level between precision and recall at the same time, this paper uses F1 value as the comprehensive discriminant index, which is calculated as follows.

$$\text{F}_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

When the model can maintain a high hit rate and avoid too many false positives in positive class recognition, the F1 value will be improved accordingly. Compared with observing precision or recall alone, this metric is more suitable for evaluating binary classification models with imperfectly balanced class distribution in medical scenarios. In addition to the investigation of positive samples, this paper also introduces the specificity evaluation model's ability to exclude low-risk individuals, namely:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{27}$$

This metric reflects the ability of the model to avoid misclassifying healthy or low-risk samples as high-risk. If the specificity is too low, the model may have a high recall rate, but it will bring too many unnecessary warning tips and reduce the screening efficiency. In order to further evaluate the robustness of the model on the overall classification, Matthews Correlation Coefficient (MCC) is used as a supplementary measure, which is expressed as follows.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{28}$$

MCC utilizes four elements of confusion matrix simultaneously, which can reflect the consistency of prediction results more comprehensively under the condition of class imbalance. In terms of threshold sensitivity evaluation, this paper uses the area under the receiver operating characteristic curve AUC as the overall ranking ability index. The larger the AUC, the more the model is able to rank the high-risk samples before the low-risk samples, indicating that its probability output has better discrimination. In summary, precision, recall and F1 value focus on measuring the effect of high risk warning, specificity reflects the

ability to exclude low risk, AUC and MCC are used to evaluate the overall discrimination quality of the model. The above indicators together constitute the model evaluation framework of this paper, which can comprehensively reflect the performance of the multimodal fusion early warning model in type 2 diabetes risk recognition. Table 2 shows the meaning of each evaluation index and its main role.

Table 2: Model evaluation indicators and their functional descriptions

Metric Name	Mathematical Basis	Main Function
Precision	Proportion of true positives among predicted positive samples	Evaluates false positive control capability
Recall	Proportion of correctly identified positives among all actual positive samples	Evaluates high-risk detection capability
F1-score	Harmonic mean of Precision and Recall	Evaluates overall performance of positive class identification
Specificity	Proportion of correctly identified negatives among all actual negative samples	Evaluates low-risk exclusion capability
AUC	Area under the ROC curve	Evaluates overall discrimination capability
MCC	Joint measure based on the four components of the confusion matrix	Evaluates consistency under imbalanced conditions

3 Results and discussion

3.1 Analysis of model convergence characteristics

In order to test the learning efficiency and stability of the type 2 diabetes risk early warning model based on multimodal data fusion in the training process, this paper analyzes the change trend of the loss function and the convergence trajectory of AUC, and compares the fusion model with the single modal model. The convergence characteristics not only reflect the fitting speed of the model to the sample distribution, but also reveal whether there are problems such as oscillation amplification, local stagnation or over-fitting accumulation in the later stage of training, so it is an important basis for judging the availability of the early warning model.

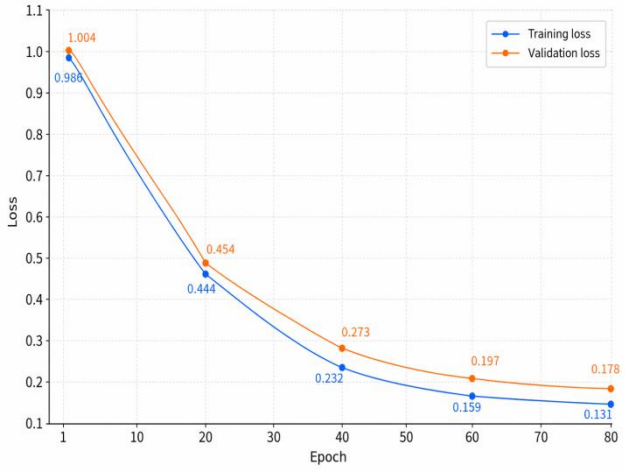

Figure 5: Model training and validation loss convergence curves

As shown in Figure 5, the training loss and validation loss of the fusion model continue to decrease with the increase of iteration rounds, showing a relatively smooth convergence process as a whole. In the first round of training, the training loss and validation loss are 0.986 and 1.004, respectively. After the 20th iteration, the two have dropped to 0.444 and 0.454 respectively, indicating that the model can quickly complete the learning of the main risk patterns in the early stage. After entering the 40th round, the loss decline began to narrow, the training loss was 0.232, the validation loss was 0.273, and the slope of the curve slowed down significantly, indicating that the model had transferred from the fast fitting stage to the refinement adjustment stage. In the 60th round, the losses of the two categories further decreased to 0.159 and 0.197, and then the changes tended to be flat. By the end of the 80th round, the training loss is stable at 0.131, the validation loss is stable at 0.178, and the difference between them is controlled within 0.047, without obvious divergence. This shows that the training strategy constructed in this paper can better suppress parameter oscillation, and the model maintains a relative balance between convergence speed and generalization performance.

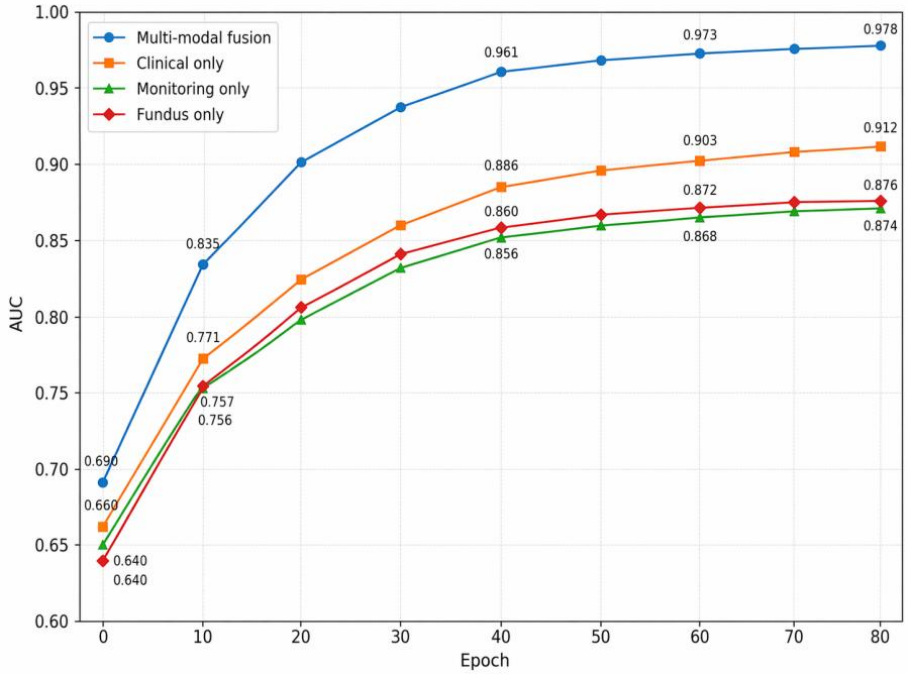

Figure 6: Comparison curves of AUC convergence between multimodal fusion model and single-modal model

Figure 6 further shows the AUC changes of different models during training. It can be seen that the improvement range of the multimodal fusion model is significantly higher than that of the three types of single-modal models. At the 10th round, the AUC has reached 0.835, which is 0.064, 0.078 and 0.079 higher than that of the clinical single-modal model, the monitoring single-modal model and the fundus imaging single-modal model, respectively. This advantage grows as the training progresses. At the 40th round, the AUC of the fusion model increased to 0.961, while the clinical, surveillance and imaging unimodal models were 0.886, 0.856 and 0.860, respectively. In the 60th round, the AUC of the fusion model reached 0.973, which was close to the optimal state. At the end of training, the AUC of the proposed model stabilized at 0.978, which was significantly higher than 0.912 of the clinical unimodal model, 0.874 of the monitoring unimodal model and 0.876 of the imaging unimodal model. The results show that the multi-modal fusion strategy can more effectively integrate the

complementary information in blood glucose, life behavior, continuous monitoring and fundus images, so that the model can obtain higher discrimination ability under the same training rounds.

Combining the loss convergence curve and AUC change trajectory, it can be seen that the proposed model has basically completed effective convergence after about 60 rounds, and the performance gain brought by subsequent iterations is small, but no obvious overfitting is produced. This shows that the proposed multi-modal fusion early warning framework has good training stability and parameter adaptation ability, and also provides a more reliable experimental basis for the comparison of early warning effects between different methods.

3.2 Comparison of early warning effects of different methods

In order to further verify the effectiveness of the type 2 diabetes risk early warning model based on multi-modal data fusion in actual discrimination, this paper compares the proposed method with the traditional machine learning model, the single-modal deep model and the conventional early fusion model. The comparison objects included logistic regression (LR), Random forest (RF), XGBoost, Clinical-DNN based only on structured Clinical data, Monitoring-BigrU based only on continuous Monitoring sequences, Base-CNN based only on Fundus images, and direct stitching Early Fusion. And the attention-gated multimodal fusion model proposed in this paper. Each model was trained under the same partition of training set, validation set and test set, and measured by the unified evaluation index set in the previous section.

From the overall results of the test set, the proposed model shows obvious advantages in all key indicators. Table 3 shows that the accuracy, precision, recall, F1 value, specificity, AUC and MCC of the proposed model reach 0.934, 0.915, 0.903, 0.906, 0.952, 0.978 and 0.856, respectively, which are better than those of other comparison methods. Compared with the traditional statistical learning model, logistic regression only achieved an accuracy of 0.821 and an AUC of 0.861, indicating that linear discrimination alone was difficult to fully capture the nonlinear interaction in the formation of type 2 diabetes risk. The AUC of XGBoost is increased to 0.921, but the recall rate and F1 value are still lower than that of the multimodal fusion model, indicating that although the tree model can identify some complex patterns, it is still limited by the expression ability of deep semantic relationships between modalities.

Table 3: Comparison of warning effects of different methods on the test set

Model	Accuracy	Precision	Recall	F1-score	Specificity	AUC	MCC
LR	0.821	0.786	0.742	0.763	0.866	0.861	0.598
RF	0.859	0.828	0.801	0.814	0.892	0.905	0.688
XGBoost	0.881	0.852	0.832	0.842	0.911	0.921	0.734
Clinical-DNN	0.889	0.861	0.841	0.851	0.918	0.912	0.751
Monitoring-BiGRU	0.842	0.812	0.789	0.800	0.878	0.874	0.662
Fundus-CNN	0.846	0.816	0.793	0.804	0.881	0.876	0.668
Early Fusion	0.907	0.883	0.873	0.868	0.931	0.949	0.786
Proposed	0.934	0.915	0.903	0.906	0.952	0.978	0.856

Single-modal deep models perform better than traditional machine learning models, but there are still clear boundaries in general. The AUC of Clinical-DNN was 0.912, and the AUC of Monitood-bigrU and Basin-CNN were 0.874 and 0.876, respectively, which was consistent with the trend obtained by convergence analysis in Section 3.1. Structured clinical modality is

still a strong input source when used alone, but it is difficult to cover the supplementary signals in behavioral fluctuations and microvascular abnormalities by relying only on biochemical indicators and signs information. Although the continuous monitoring modality and fundus image modality can provide dynamic behavioral clues and visual pathological information, the model is still not sufficient to grasp the overall risk state without clinical basic variables. The performance of the direct concatenation Early Fusion is significantly improved after the joint input of multi-source information, with the AUC reaching 0.949, the accuracy and F1 value increasing to 0.907 and 0.868, respectively, indicating that the multimodal joint modeling direction itself has strong effectiveness. Nevertheless, Early Fusion is still lower than the AUC of the proposed model by 2.9 percentage points and the F1 value by 3.8 percentage points, which indicates that simple concatenation cannot adequately solve the problem of weight adaptation and redundancy suppression between different modalities.

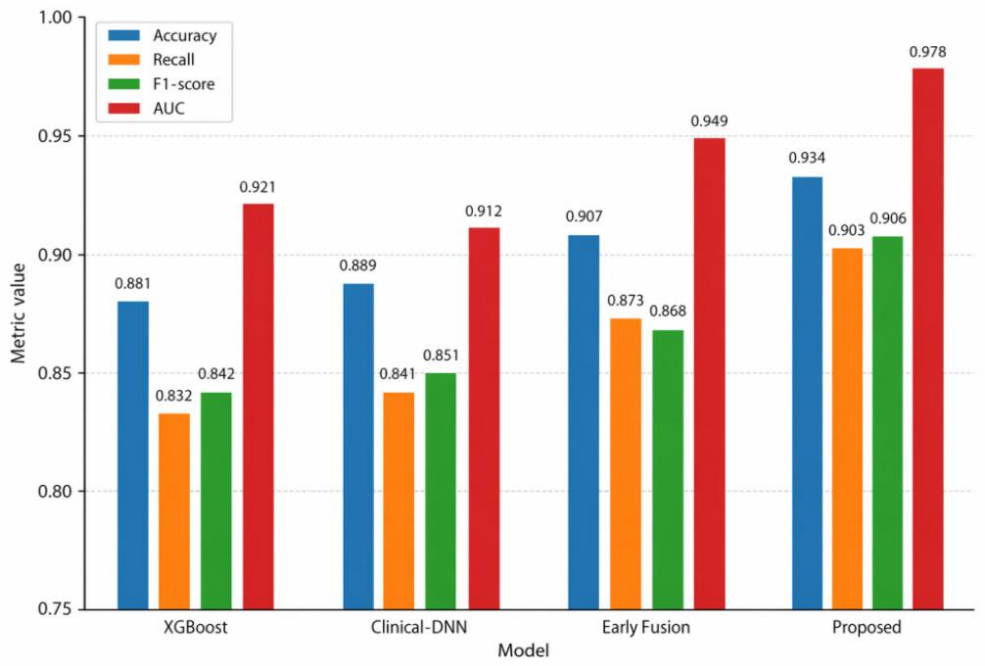

Figure 7: Comparison of key indicators of different methods

Compared with Early Fusion, the proposed model improves the accuracy by 2.7 percentage points, the recall rate by 3.0 percentage points, and the MCC by 0.070. This result shows that the dynamic weight allocation and gated recombination mechanism introduced in the fusion stage not only improves the detection ability of high-risk individuals, but also improves the overall consistency under the condition of class imbalance. Figure 7 visually displays the key indicators of XGBoost, Clinical-DNN, Early Fusion and the proposed model. It can be seen that the proposed model maintains the highest level in the four indicators of Accuracy, Recall, F1-score and AUC. And the lead is more obvious in AUC and F1 value. This shows that multi-modal fusion does not only bring local performance gain, but also achieves improvement in the overall discrimination quality and the balance of high-risk identification.

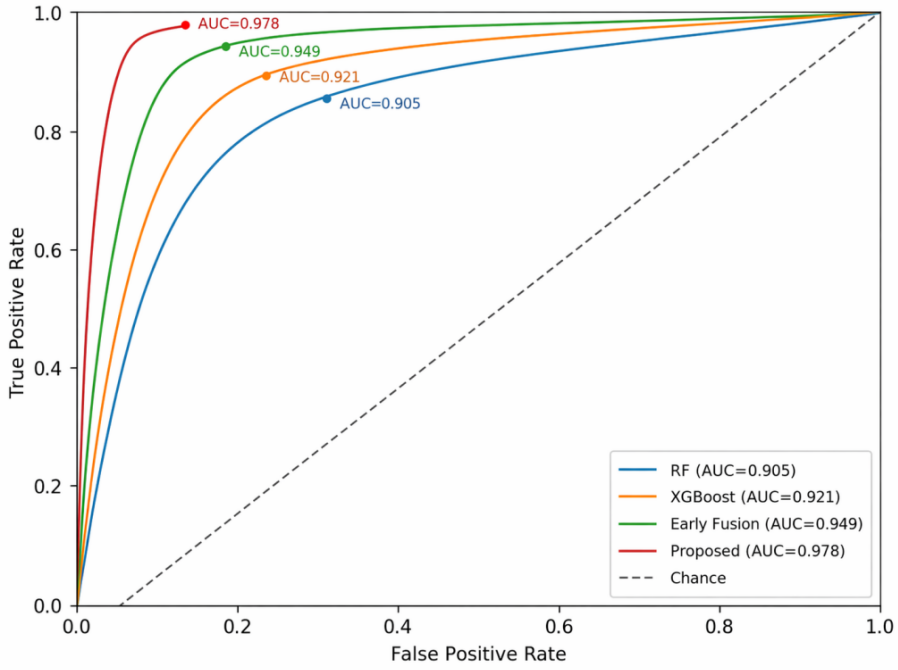

Figure 8: Comparison of ROC curves of key models

From the ROC curve comparison results, the proposed model shows a faster true positive rate increase in the low false positive rate interval, the curve is closer to the upper left corner, and the AUC reaches 0.978. Figure 8 shows that although Early Fusion is better than RF and XGBoost, its curve still has a certain climbing lag in the middle and high threshold region. In contrast, the proposed model can maintain high sensitivity at a low false alarm level, which is more suitable for clinical front-end screening and health management scenarios. This advantage shows that the attention-weighted multimodal joint representation can more accurately distinguish the two types of samples of "metabolic abnormalities but not yet diagnosed" and "low risk normal fluctuations", thereby improving the flexibility of early warning threshold selection.

In order to investigate the stability of the model, this paper further carried out 5-fold cross validation, and the results are shown in Table 4. It can be seen that the AUC of the proposed model on five folds is 0.974, 0.979, 0.976, 0.971 and 0.981, respectively, with an average value of 0.976 and a standard deviation of only 0.004. In contrast, the average AUC of XGBoost and Early Fusion are 0.921 and 0.949, respectively, which also show certain stability, but the overall level is still significantly lower than that of the proposed model. This shows that the multi-modal attention-gated fusion framework not only achieves better results in a single test, but also maintains strong generalization ability and robustness under different data partitioning conditions.

Table 4: AUC results of 5-fold cross-validation for key models

Model	K1	K2	K3	K4	K5	Mean ± SD
XGBoost	0.916	0.923	0.919	0.926	0.921	0.921 ± 0.004
Early Fusion	0.944	0.951	0.947	0.954	0.949	0.949 ± 0.004
Proposed	0.974	0.979	0.976	0.971	0.981	0.976 ± 0.004

In general, the traditional model still has certain reference value in the modeling of structured variables, but it is difficult to fully reveal the cross-modal association in the

formation of type 2 diabetes risk. Single-modal deep models can improve the ability of local feature expression, but are limited by a single source of information. Although direct concatenation fusion has shown the advantages of multi-modal modeling, there is still room for improvement in feature selective utilization and noise suppression. Through unified representation, dynamic weight allocation and gated refinement, the proposed model achieves better performance in overall discrimination ability, high-risk identification ability and cross-partition stability, which indicates that the proposed method is more suitable for intelligent modeling requirements for early screening and risk warning of type 2 diabetes.

4 Conclusion

Focusing on the problems of insufficient representation of single data source, risk information dispersion and limited discriminant stability in the early identification of type 2 diabetes, this paper constructs a risk early warning model based on multi-modal data fusion, and forms a relatively complete technical link from data acquisition, preprocessing, single-modal feature extraction, cross-modal fusion, risk mapping and parameter optimization. The results show that multimodal joint modeling can more effectively integrate structured clinical indicators, continuous monitoring signals and complementary information from fundus images, thereby improving the identification ability of high-risk individuals.

The experimental results show that the accuracy, precision, recall, F1 score, AUC and MCC of the proposed model on the test set reach 0.934, 0.915, 0.903, 0.906, 0.978 and 0.856, respectively. It is better than LR, RF, XGBoost, single-modal depth model and conventional Early Fusion method. Among them, compared with Early Fusion, the AUC of the proposed model is increased by 0.029, and the F1-score is increased by 0.038, indicating that after introducing dynamic weight allocation and gated recombination, the model is enhanced in the discrimination accuracy and risk identification balance. The cross-validation results further show that the mean value of the 5-fold AUC of the model is 0.976, and the standard deviation is only 0.004, indicating that it has good generalization ability and training stability. However, this paper still has some limitations. Firstly, the sample source is still dominated by single center data, and the coverage of regional differences and population heterogeneity is insufficient. Second, although the modality composition has included clinical, behavioral and imaging information, it has not included deeper risk factors such as genetic markers, dietary details and long-term drug exposure. Third, the current research focus is still on model performance improvement, and the discussion on interpretable early warning mechanism and clinical deployment process needs to be further deepened. In the future, the multi-center sample expansion, external validation, feature interpretation and lightweight deployment can be further improved to enhance the application value of the model in real chronic disease management scenarios.

References

[1] Mohsen F, Al-Absi H R H, Yousri N A, et al. A scoping review of artificial intelligence-based methods for diabetes risk prediction[J]. npj Digital Medicine, 2023, 6(1): 197.

[2] Aoki J, Khalid O, Kaya C, et al. Progression from prediabetes to diabetes in a diverse US population: a machine learning model[J]. Diabetes Technology & Therapeutics, 2024, 26(10): 748-753.

[3] Talebi Moghaddam M, Jahani Y, Arefzadeh Z, et al. Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm[J]. BMC Medical Research Methodology, 2024, 24(1): 220.

[4] Hoyos W, Hoyos K, Ruiz R, et al. An explainable analysis of diabetes mellitus using statistical and artificial intelligence techniques[J]. BMC medical informatics and decision making, 2024, 24(1): 383.

[5] Duckworth C, Guy M J, Kumaran A, et al. Explainable machine learning for real-time hypoglycemia and hyperglycemia prediction and personalized control recommendations[J]. Journal of Diabetes Science and Technology, 2024, 18(1): 113-123.

[6] Han B C, Kim J, Choi J. Prediction of complications in diabetes mellitus using machine learning models with transplanted topic model features[J]. Biomedical engineering letters, 2024, 14(1): 163-171.

[7] Tuppad A, Patil S D. Machine learning for diabetes clinical decision support: a review[J]. Advances in Computational Intelligence, 2022, 2(2): 22.

[8] Talukder M A, Islam M M, Uddin M A, et al. Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications[J]. Digital Health, 2024, 10: 20552076241271867.

[9] Tanabe H, Sato M, Miyake A, et al. Machine learning-based reproducible prediction of type 2 diabetes subtypes[J]. Diabetologia, 2024, 67(11): 2446-2458.

[10] Ding J E, Thao P N M, Peng W C, et al. Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records[J]. Scientific reports, 2024, 14(1): 20774.

[11] Nguyen H V, Choi Y, Byeon H. An explainable hybrid deep learning model for prediabetes prediction in men aged 30 and above[J]. J Mens Health, 2024, 20(10): 52-72.

[12] Talari P, N B, Kaur G, et al. Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2[J]. Plos one, 2024, 19(1): e0292100.

[13] Lee Y C, Cha J, Shim I, et al. Multimodal deep learning of fundus abnormalities and traditional risk factors for cardiovascular risk prediction[J]. NPJ digital medicine, 2023, 6(1): 14.

[14] Baharoon M, Almatar H, Alduhayan R, et al. HyMNet: a multimodal deep learning system for hypertension classification using fundus photographs and cardiometabolic risk factors[J]. arXiv preprint arXiv:2310.01099, 2023.

[15] Wang S C Y, Nickel G, Venkatesh K P, et al. AI-based diabetes care: risk prediction models and implementation concerns[J]. NPJ digital medicine, 2024, 7(1): 36.

[16] Yagin F H, Al-Hashem F, Ahmad I, et al. Pilot-study to explore metabolic signature of type 2 diabetes: a pipeline of tree-based machine learning and bioinformatics techniques

for biomarkers discovery[J]. Nutrients, 2024, 16(10): 1537.

[17] Thanh Phuc P, Nguyen P A, Nguyen N N, et al. Early detection of dementia in populations with type 2 diabetes: Predictive analytics using machine learning approach[J]. Journal of Medical Internet Research, 2024, 26: e52107.

[18] Khan S, Mohsen F, Shah Z. Genetic biomarkers and machine learning techniques for predicting diabetes: systematic review[J]. Artificial Intelligence Review, 2024, 58(2): 41.

[19] Nderitu P, Nunez do Rio J M, Webster L, et al. Predicting 1, 2 and 3 year emergent referable diabetic retinopathy and maculopathy using deep learning[J]. Communications Medicine, 2024, 4(1): 167.

[20] El-Sofany H, El-Seoud S A, Karam O H, et al. A proposed technique using machine learning for the prediction of diabetes disease through a mobile app[J]. International Journal of Intelligent Systems, 2024, 2024(1): 6688934.