International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Risk Assessment Network for Diabetic Cardiovascular Disease based on Causality and Individual Attention

Ming Zuo[1], Jingyi Deng[2], Liping Zhang[3], Qi Xu[4]

[1]Glorious Sun School of Business and Management, Donghua University 201620, Shanghai, China
zm[at]rjh.com.cn

[2]School of Computer Science and Technology, Donghua University 201620, Shanghai, China
15001859414[at]163.com

[3]School of Computer Science and Technology, Donghua University 201620, Shanghai, China
1139042518[at]qq.com

[4]Glorious Sun School of Business and Management, Donghua University 201620, Shanghai, China
xuqi[at]dhu.edu.cn

Abstract: ***With the increasing number of diabetic patients worldwide, the prevention and treatment of diabetic cardiovascular disease, a major complication, has become a major social challenge. At present, most of the research on diabetic cardiovascular disease is based on statistical methods, focusing on the correlation analysis between the risk characteristics of patients, such as age and cholesterol, and the disease risk. This approach, which considers the individual characteristics of patients and the characteristics of metabolic indicators as the same risk characteristics, ignores the causal relationship between risk characteristics and disease risk, ignores the important information carried by the individual characteristics of patients and the background of diabetes, and further ignores the impact of differences in disease background. In order to fill this gap, we proposed a new deep learning model, namely, a risk assessment model for diabetic cardiovascular disease based on Causal stability and interaction of individual characteristics (causal-NET). The causally stable and time-aware Long short-term Memory network (Causal and time-aware TLSTM) was used to learn disease risk information in the metabolic characteristics of patients and enhance the stability of the model. Secondly, our model also designed an individual feature interaction layer, which used individual features to modify the disease information hidden information obtained by learning the Causal and time-aware TLSTM unit, so as to obtain a more accurate and comprehensive disease information representation for the risk assessment task of diabetic cardiovascular disease. Our experimental results demonstrate that the model presented here performs better in the diabetic CVD risk assessment task, and consistently outperforms the contrast model. The experimental evaluation indexes reached the model accuracy, recall, F1 score and 94.33%, 89.84%, 93.33% and 93.90% under the receiver operation feature curve, respectively.***

Keywords: Diabetic cardiovascular disease; Metabolic characteristics selection; Individual characteristic interaction; Causal stable learning; Disease risk assessment

1. Introduction

Diabetes mellitus (dm) is a chronic metabolic disease that causes a variety of serious health complications, including kidney failure, blindness and cardiovascular diseases, and has become one of the leading disease burdens in China and globally [1, 2, 3]. The international diabetes federation estimates that 415 million people worldwide, or 8.8% of the world's population, are living with diabetes, and death from diabetic cardiovascular disease is one of the leading causes of death in this population [5, 6, 7]. Therefore, the search for an effective risk assessment method for diabetic cardiovascular disease for early prevention and treatment of the disease could greatly improve the survival rate of people with diabetes.

Most of the existing studies related to the risk of diabetic cardiovascular disease are based on statistical methods to calculate the correlation between risk characteristics and disease risk or disease risk score. For example, Domanski et al. [11] used statistical methods to evaluate the relationship between low-density lipoprotein and disease risk. D'Agostino et al. [12] constructed the American Framingham cardiovascular disease prediction model based on the general population, and used Cox proportional hazards regression model to evaluate the risk scores of related factors such as patient age, high-density lipoprotein and diabetes status on cardiovascular disease events. These methods have made some progress in their study cohorts, but most of them treat diabetes, an important disease background, individual characteristics and metabolic characteristics of patients as risk characteristics indiscriminately for disease risk analysis, emphasizing the statistical correlation between risk characteristics and disease risk, but ignoring the causal relationship between them [13]. At the same time, they ignore the important information carried by the individual characteristics of patients and the background of diabetes, and further ignore the influence of the difference in the distribution of data sets caused by the difference in the background of disease on the stability of the model.

Licensed Under Creative Commons Attribution CC BY

To address the above problems, we propose a risk assessment model for diabetic cardiovascular disease based on the interaction between causal attention and individual characteristics. This model can effectively consider the characteristics of diabetes, emphasize the causal relationship between risk characteristics and target tasks, and reduce the impact of differences in disease background. At the same time, the model can effectively combine the individual characteristics of patients, improve the accuracy of disease risk assessment task, help clinicians to make disease risk diagnosis, and improve the probability of early detection and treatment of disease.

In summary, the main contributions of this paper are as follows:

Considering the characteristics of chronic metabolic diseases in diabetes, the long-term medical visit data of patients were regarded as temporal information as the input of the model in the task of risk assessment of diabetic cardiovascular disease, and the modeling idea of TLSTM was adopted.

We redesigned the unit update process of TLSTM to better focus on the information carried by patients' current medical data, reduce the impact of data distribution differences caused by different background of diabetes complications on the model, and increase the accuracy and stability of the model prediction. To the best of our knowledge, this is the first time that causal correlation has been applied to the risk prediction task of diabetes cardiovascular disease.

(3) The individual feature interaction network was designed, and the individual characteristics of patients were incorporated into the model learning to further learn and modify the disease risk feature information obtained in the previous stage, so as to obtain a more comprehensive disease information feature representation.

In order to prove the effectiveness and superiority of our model, we evaluated and compared our model with traditional machine learning methods (LR, RF and GBDT) and deep learning methods (RNN, GRU, LSTM and T-LSTM) on this task. Experimental results show that our proposed model performs better in real tasks and outperforms the baseline model in AUC and other indicators.

2. Related Work

Cardiovascular disease, as the leading cause of death worldwide, is an important public health problem [10]. Over the years, the study of its related disease risk has been a hot issue, attracting the attention of many scholars and experts at home and abroad. Most of the existing research methods are based on statistics and use the correlation between risk factors (risk characteristics) and diseases to carry out disease risk regression modeling. For example, Shen Meifeng et al. [16] used Pearson and variance statistical methods to study the effects of patient's age, gender, history of diabetic complications and glycated hemoglobin index on cardiovascular complications of type 2 diabetes. Scholes et al. [18] studied the prevalence and management trends of CVD risk factors in the United Kingdom from the perspective of BMI category based on statistical analysis, and confirmed the significance of blood pressure and lipid changes and glycemic control. Similarly, Bode et al. [19] also combined statistical methods and used Wald test and Logistic regression model to study the relationship between cardiovascular disease risk factors and BMI and age with the prevalence of risk factors in American firefighters by BMI category.

These works have contributed to our study of risk factors for cardiovascular disease, but most of them regard diabetes as an important disease background as a simple risk characteristic, such as low-density lipoprotein and other risk indicators, ignoring the important information carried by the patient's diabetes disease background. Based on the correlation between risk factors and disease, some studies further used Cox hazard regression model to model and obtained the risk score of related factors on cardiovascular disease events. For example, Elley et al. [21] built the cardiovascular disease prediction model of the New Zealand Diabetes Cohort Study (DCS) based on patients with type 2 diabetes mellitus. Cox proportional hazards regression model was used to model cardiovascular events. Multiple risk factors such as age, duration of diabetes, sex, systolic blood pressure, smoking status, total cholesterol and glycosylated hemoglobin were evaluated. Conroy et al. [22] used the Weibull proportional hazards model to develop a risk scoring system for the clinical management of cardiovascular risk in European clinical practice, based on cohort study datasets from 12 European countries. Hippisley-cox J et al. [23] developed a model to estimate lifetime risk of CVD by fitting two independent Cox models, taking into account factors such as race, total cholesterol ratio, and age.

These risk prediction algorithms are usually developed using multivariate regression models and usually assume that all of these factors are linearly related to CVD outcome. The limitations of modeling assumptions and the limited number of predictors make existing algorithms usually show moderate predictive performance [24]. Therefore, some scholars proposed data-driven techniques based on Machine Learning (ML) to unknowingly identify new risk predictors and their more complex interactions, so as to improve the performance of risk prediction. For example, Mohan et al. [12] combined Random Forest (RF) and Linear Method (LM) modeling and proposed a Linear hybrid Random Forest model to improve the accuracy of cardiovascular disease prediction. Dinh et al. [25] used Logistic Regression (LR) and Support Vector Machine (SVM) to improve the accuracy of cardiovascular disease prediction. Multiple supervised learning models, such as SVM and Integration Model, were used to classify high-risk patients to achieve better performance than a single algorithm. Alaa et al. [23] An ML-based model was developed to predict cardiovascular disease risk based on 473 available variables.

To some extent, these machine learning models make up for the shortcomings of previous models based on multivariate regression. However, these models treat all risk factors

Licensed Under Creative Commons Attribution CC BY

related to the disease equally as the same, lack of attention to the heterogeneity of individual characteristics of patients, and lack of learning of important disease background information. A number of studies have been conducted [26-30]It has been shown that people with diabetes have an increased risk of cardiovascular disease, and the correlation is not negligible. Diabetes mellitus is considered to be an independent risk factor for cardiovascular disease, and cardiovascular disease is the most common cause of death in patients with diabetes [7, 31, 32]Therefore, it is very important to further explore and utilize diabetes information for the task of cardiovascular disease risk prediction. In this paper, we propose a new deep learning model for the risk assessment task of cardiovascular disease. This model considering the diabetes chronic metabolic disease characteristics, using the patients medical clinic data as input for a long time, and get the weights of causal factors based on the balance of covariate in patients with metabolic characteristics of stability study, study to interact with the characteristics of individual patients, in addition to the patients with metabolic characteristics of individual characteristic information, improve the reliability and accuracy of the model of task.

3. Problem Elaboration

3.1 Patient dataset description

In the problem we defined, each patient data consists of two parts: temporal metabolic characteristics and individual patient characteristics. In metabolic characteristics on the choice of the risk factors, and the most risk factors based on clinical experience or relevant statistical study of the traditional feature selection methods, in order to further study characteristics and target the potential relationship between disease risk, combined with the characteristics of target population data set, lower unrelated or low correlation characteristic of the model, the influence of In order to improve the performance of the model, SHAP based on game theory and random forest algorithm, a commonly used method in machine learning, were used to rank the importance of features on the relevant risk indicators, and the high-importance features were selected as the input data of the model.

As shown in Figure 3-1. According to the ranking results of characteristics, we finally selected seven indicators of glycosylated hemoglobin (HbA1c), two hours postprandial blood glucose (GLU4), cholesterol (CHOL), high density lipoprotein (UHDL), low density lipoprotein (ULDL), triglyceride (TG-B) and apolipoprotein B (APOB) as our metabolic characteristics input. This is also consistent with the direction chosen in most clinical studies.

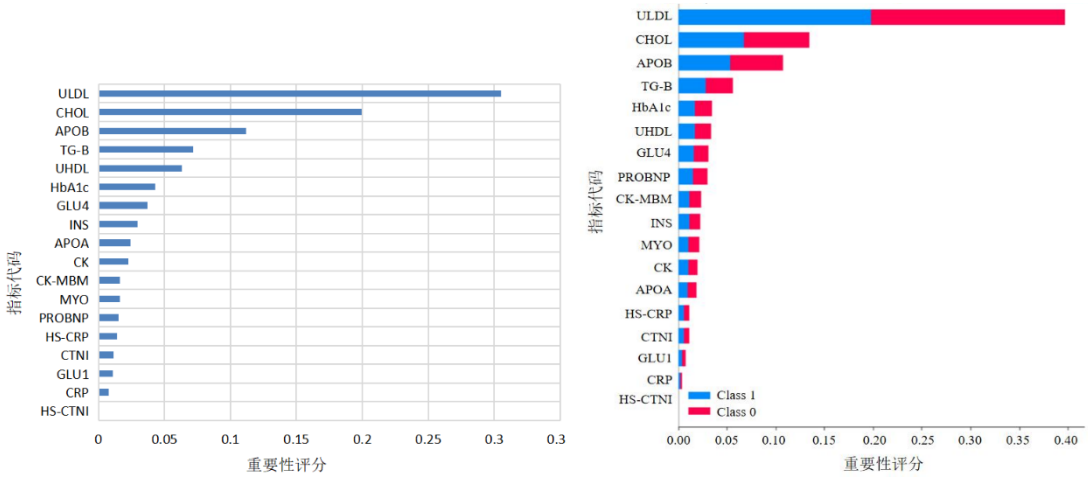


Figure 3: 1 Left: Importance ranking of indicator features based on Random Forest; Figure 3-1 right: Ranking of importance of indicator features based on SHAP.

3.2 Problem Elaboration

In the definition of the problem in this paper, each patient's data is composed of metabolic characteristics and individual characteristics. In order to better reflect the real patient visit situation, the patient visit time is also included as input information in the task learning of the model.

Thus, a patient data can be described as $P, P = \{[v_1, v_2, \ldots v_T], d, \tau\} = \{V, d, \tau\}$, where T represents the number of examinations and v_t represents the t visits, including glycated hemoglobin (HbA1c), two hours postprandial blood glucose (GLU4), HDL lipoprotein (UHDL), low density lipoprotein (ULDL), cholesterol (CHOL), triglycerides (TGB) and apolipoprotein B (APOB). For unified expression, the number of medical features is recorded here as N_v, so the patient t visit record data is described as $v_t = [v_t^1, v_t^2, \ldots, v_t^{N_v}], v_t \in \mathbb{R}^{N_v}$, here $N_v = 7. d$

Licensed Under Creative Commons Attribution CC BY

represents six individual characteristics including patient gender, age and the other four common complications of diabetes (here, diabetic foot disease, diabetic nephropathy, diabetic peripheral neuropathy and diabetic eye disease), as described by N_d, $d = \{d^1, d^2, \ldots, d^{N_d}\}, d \in \mathbb{R}^{N_d}, \tau$ represents the time interval of patient visits, $\tau = [\Delta_1, \Delta_2, \ldots \Delta_T], \tau \in \mathbb{R}^{N_t}$. Specifically, the Δ_i represents the time interval between the patient's t visit, v_i, and the last visit, v_{i-1}, which should be noted as $\Delta_1 = 0$.

The problem in this paper can be described as, given the l sample size dataset
$D, D = \{ (P_1, y_1), (P_2, y_2), \ldots, (P_l, y_l) \} = \{ (P_i, y_i) \}_{i=1}^{l}$, in which the input P_i for each sample consists of patient visit medical feature sequence V, visit interval τ and individual feature τ, namely $P_i = \{V, d, \tau\} = \{ [v_1, v_2, \ldots v_T], d, \tau\}$. The goal of this paper is to learn a non-linear mapping function to assess the diabetic CVD risk of patients based on the medical visit data D, while minimizing the error of the target function and the sample output, as shown in formulas (4-1) and (4-2):

$$\hat{y}_i = f(P_i; \omega) \quad (4\text{-}1)$$

$$\omega^* = \underset{\omega}{argmin} \sum_{i=1}^{l} L(y_i, \hat{y}_i) \quad (4\text{-}2)$$

Specifically, f (·) is the nonlinear mapping function, L (·) is the loss function for defining the model task, and ω is the parameter of the model network. $\hat{y}_i$ represents the output of the target model prediction function, and y_i represents the true value of the risk of the cardiovascular disease occurrence in the i^{th} patient, where, $\hat{y}_i, y_i \in \{0, 1\}$. It should be pointed out here that in order to better express the size of the patient's current cardiovascular disease risk, $\hat{y}_i$, the binary label, and the $\tilde{y}_i$ ($\tilde{y}_i \in [0, 1]$), are also used as the output.

4. Model Approach

4.1 Overall Architecture

As shown in figure 4-1, Causal cardiovascular disease risk assessment model causal-net mainly consists of three parts: (1) LSTM module based on Causal stability and time awareness; (2) individual feature interaction module based on attention mechanism; (3) Output module based on fully connected network.

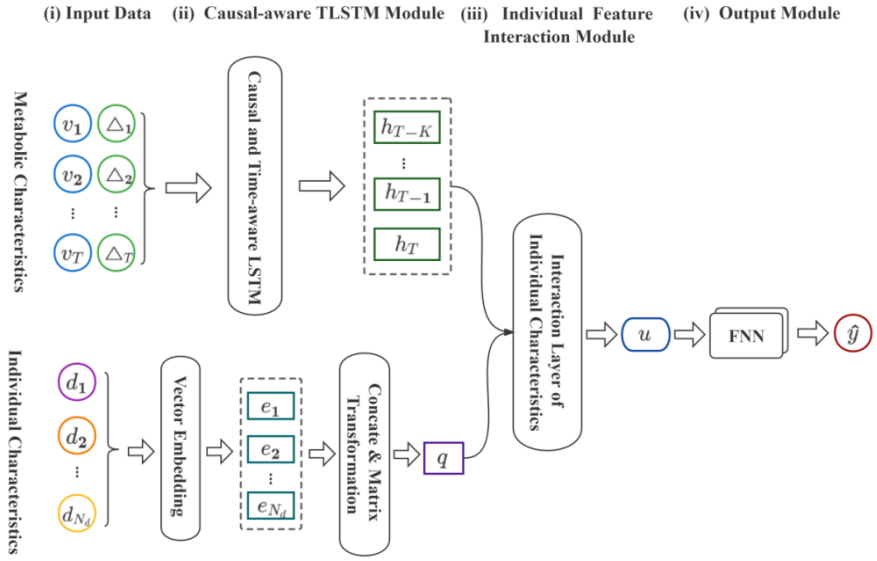


Figure 4-1: Overall architecture of Causal-aware TLSTM model

The model Causal-Net takes the feature sequenceV in the patient's medical visit record and the visit time interval τ as the input of the Causal and Time-aware LSTM module (Causal-aware TLSTM) based on causal stability and time-perception, which first obtains the weight factor α_i of each step by balancing covariates, reducing the influence of different complication disease background on the index characteristics of diabetes patients. Subsequently, the Causal-aware TLSTM module performed feature learning of patient metabolic metrics based on the weight factor $\alpha_i, v_i$and τ_i at the current step, yielding the hidden state h_i and c_i at the time of i. In order to further study disease risk in the hidden state information, this chapter in the second stage of individual feature interaction module, set K size observation window, the previous stage output $h_{T-K:T}$ ($h_{T-K:T} = [h_{T-K}, \ldots, h_{T-1}, h_T]$) and patient individual characteristics d as input to the stage, based on the attention mechanism of disease information hidden state sequence $h_{T-K:T}$ reweighted weight β ($\beta = [\beta_{T-K}, \ldots, \beta_{T-1}, \beta_T]$). Use these modified reweighted hidden state sequence $\tilde{h}_{T-K:T}$ ($\tilde{h}_{T-K:T} = [\tilde{h}_{T-K}, \ldots, \tilde{h}_{T-1}, \tilde{h}_T]$) and individual characteristics d to obtain a more accurate and comprehensive disease risk representation u. Finally, the model uses a fully connected network for disease risk assessment to obtain a risk assessment value of$\tilde{y}$ between 0 and 1, and then it is mapped

Licensed Under Creative Commons Attribution CC BY

to {0, 1} via the softmax function to obtain the $\tilde{y}$.

4.2 LSTM based on causal stability and time awareness

This section describes the first module in Causal-Net, the Causal stability and time-aware long short-term Memory Unit (Causal-Aware TLSTM). This module is mainly composed of two parts: weight factor calculation based on covariate balance and causal stability learning based on time awareness.

4.2.1 Weight calculation based on covariate balance

There are a large number of diabetic complications, and their prevalence varies greatly, leading to unbalanced distribution of patients' disease background, which brings some distribution differences to the dataset, resulting in inaccurate parameter estimation and unstable prediction of unknown test data. Intuitively speaking, since the incidence of diabetic nephropathy is higher than that of other complications, the distribution of patients' disease backgrounds in the data sets is unbalanced. In this condition, in order to better improve the performance of the model, such as accuracy, the model will pay more attention to and learn the data characteristics of patients with nephropathy complications during training, while ignoring the information of patients with other disease backgrounds.

Therefore, this section proposes a feature de-correlation weighting algorithm based in literature [10] to calculate the weight factor α_i for each step of visit data, by adjusting the weight size to achieve covariate balance, thus removing the impact of the difference in disease background distribution, and increasing the accuracy and stability of the model. The main calculation process is as follows:

First, One characteristic variable in all clinic medical characteristics is Z_i ($i \in N_v, Z_i \in \mathbb{R}^{N_n*1}$), N_n represents the total number of visits of all patients in the dataset, Initialized causal weight is W ($W \in \mathbb{R}^{N_n*1}, \alpha \in W$); To achieve a covariate equilibrium, That is, to make the difference between $E[Z_i^T \Sigma_W Z_{-i}]$ and $E[Z_i^T W]E[Z_{-i}^T W]$ as small as possible, Where $\Sigma_W = \text{diag}(W_1, \ldots, W_{N_n})$, $\sum_{j=1}^{N_n} W_j = N_n$, Z_{-i} represents the other feature variables other than Z_i, The objective function can therefore be described as shown in the formula (4-3):

$$W^{\alpha} = \underset{W}{argmin} \sum_{i=1}^{N_v} ||E[Z_i^T \Sigma_W Z_{-i}] - E[Z_i^T W]E[Z_{-i}^T W]||_2^2 \quad (4\text{-}3)$$

4.2.2 TLSTM unit based on causal stability

Traditional TLSTM considers that if the time span between two consecutive records is large, the current dependence on the previous record should be differentiated, that is, the short-term memory should be adjusted according to the time span between records with time steps t and t−1 without dismissing the long-term effects. Therefore, compared with LSTM, the main improvement of TLSTM architecture is the adjustment of the amount of information contained in the previous time step. TLSTM focuses on the information dependency between v_{t-1} and v_t and proposed the heuristic decay function $Y(\cdot)$, applied to short-term memory information C_{t-1}^S to obtain the adjusted short-term memory $\widehat{C}_{t-1}^S$.

Similarly, the contribution of patient records to the current information should vary differently, and to reduce the data-agnostic distribution difference effects of the disease background, emphasize the attention learning of the current visit information v_t during the TLSTM unit updating process. This subsection presents the TLSTM-based on causal attention, namely Causal-aware TLSTM, as shown in Figures 4-3. Causal-aware TLSTM uses the feature causal weight W^{α} calculated in the previous subsection to adjust the current candidate memory $\widetilde{C}$ during the neural network unit update, allowing the model to pay attention to more complete information on the clinic feature.

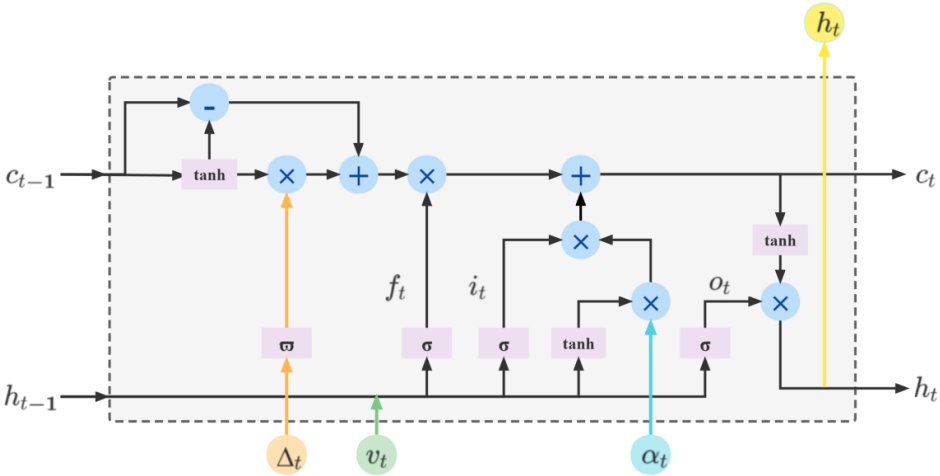


Figure 4-3: TLSTM unit based on causal stability1
Figure 4-3: The cell of Causal-aware TLSTM

In Causal-Aware TLSTM, the cell status update process at time step t is as follows:

In the cause-Aware TLSTM unit update process, the calculation principle of long short-term memory at time step t-1 is shown in equations (4-4) to (4-7):

$$C_{t-1}^S = tanh(W_s C_{t-1} + b_s) \quad (4\text{-}4)$$

Licensed Under Creative Commons Attribution CC BY

$$\hat{C}_{t-1}^{S} = C_{t-1}^{S} * \Upsilon(\Delta_t) \quad (4\text{-}5)$$
$$C_{t-1}^{L} = C_{t-1} - C_{t-1}^{S} \quad (4\text{-}6)$$
$$C_{t-1}^{*} = C_{t-1}^{L} + \hat{C}_{t-1}^{S} \quad (4\text{-}7)$$

The calculation principle of the candidate memory $\widetilde{C}^{*}$ based on the causal weights is shown in Equations (4-8) and (4-9):

$$\tilde{C} = tanh(W_c v_t + U_c h_{t-1} + b_c) \quad (4\text{-}8)$$
$$\tilde{C}^{*} = \tilde{C} * \alpha_t \quad (4\text{-}9)$$

The calculation process of forgetting gate f_t, input gate i_t and output gate o_t are as shown in equation (4-10), (4-11) and (4-12) respectively:

$$f_t = \sigma\,(W_f v_t + U_f h_{t-1} + b_f) \quad (4\text{-}10)$$
$$i_t = \sigma\,(W_i v_t + U_i h_{t-1} + b_i) \quad (4\text{-}11)$$
$$o_t = \sigma\,(W_o v_t + U_o h_{t-1} + b_o) \quad (4\text{-}12)$$

The memory unit C_t and the hidden state h_t at time step t in Causal-aware TLSTM are calculated as shown in formulas (4-13) and (4-14):

$$C_t = f_t * C_{t-1}^{*} + i_t * \widetilde{C}^{*} \quad (4\text{-}13)$$
$$h_t = o_t * \tanh\,(C_t) \quad (4\text{-}14)$$

where v_t represents the current input, h_{t-1} and h_t are the hidden states of the previous and current steps, respectively, and C_{t-1} and C_t are the unit memories of the previous and current steps, respectively. Δ_t is the access time interval between v_{t-1} and v_t, and $\Upsilon(\cdot)$ is a heuristic decay function based on Δ_t values, which have less effect on short-term memory. C_{t-1}^{S} represents short-term memory in the previous step, $\widehat{C}_{t-1}^{S}$ is short-term memory after adjustment of the time inspiration function, C_{t-1}^{L} represents long-term memory in the previous step, and C_{t-1}^{*} represents long and short-term memory after adjustment. As with the standard LSTM unit update process, $\widetilde{C}$ is the current candidate memory, $\widetilde{C}^{*}$ is the candidate memory based on the causal weight α_t adjustment, and the current unit memory C_t is obtained based on these two parts of the unit memory. In addition, W, U and b are all network parameters to be trained. α_t represents the weight size of the patient's current visit record, which is used to solve the problem of inaccurate parameter estimation and unknown prediction caused by differences in disease background, and enhance model stability learning.

4.3 Attention-based interaction of individual features

In addition to the metabolic data in medical visit characteristics, the information carried by individual characteristics of patients, such as history of other complications under diabetes, age and gender, also plays an important role in improving the performance of the target task in this paper, which should not be ignored. In addition, the disease risk information concerned by the various hidden state h_i obtained in the Causal-aware TLSTM in the previous stage is not the same, and its contribution size to the assessment task should not be regarded as undifferentiated. Therefore, in order to obtain more accurate feature information that can represent the current disease risk of patients, this paper designed an individual feature interaction module based on attention mechanism in Causal-NET, as shown in Figure 4-4 below.

As shown in Figure 4-4, This module uses the vector embedding and the individual patient feature vector q obtained from feature extraction and the disease information hidden feature sequence obtained in the previous module Causal-aware TLSTM $h_{T-K:T}$ ($h_{T-K:T} = [h_{T-K}, \ldots, h_{T-1}, h_T]$) are used, as the input, And use individual features to correct the $h_{T-K:T}$ output in the learning Causal-aware TLSTM: T, To get more accurate disease information representing the u, For the diabetic cardiovascular disease risk assessment task used in this paper.

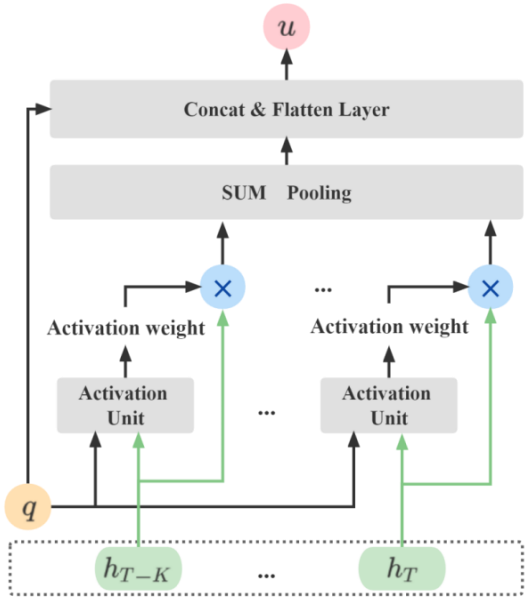


Figure 4-4: Interaction module of individual characteristics based on attention 2

4.3.1 Individual characteristics representation layer

First, this section counts the discrete number of the discrete individual features as the word list size, and the word vector dimension size is set to be N_s according to the size of the available number of values of the discrete features. Subsequently, the discrete individual features $[d_1, \ldots, d_{N_d}]$ were input to the embedding layer to obtain the embedding vector of individual features based on Word2Vec, $[e_1, \ldots, e_{N_d}]$, where $e_i \in \mathbb{R}^{N_s}$. The matrix representation $\hat{q}$ ($\hat{q} \in \mathbb{R}^{1\times(N_d*N_s)}$) is finally multiplied by the parameter matrix W_q ($W_q \in \mathbb{R}^{(N_d*N_s)\times N_h}$) to obtain the latest representation of individual characteristics q ($q \in \mathbb{R}^{1\times N_h}$).

4.3.2 Calculation of feature weights

Take the individual feature q and the hidden state sequence $h_{T-K:T}$ as input to calculate the external product p of the two features, and then combine the feature external product p into Concat splicing together with the individual feature q and the hidden state sequence $h_{T-K:T}$ to get a new feature representation. Then the new feature representation is input into the multi-fully connected network, ReLu is selected as the activation function, and finally the reweighted weight of

Licensed Under Creative Commons Attribution CC BY

the hidden state sequence is obtained by the linear layer output, β ($\beta = [\beta_{T-K}, \ldots, \beta_{T-1}, \beta_T]$).

Here, individual h_i and q are taken as examples, and the computational principle is shown in Figure 4-5.

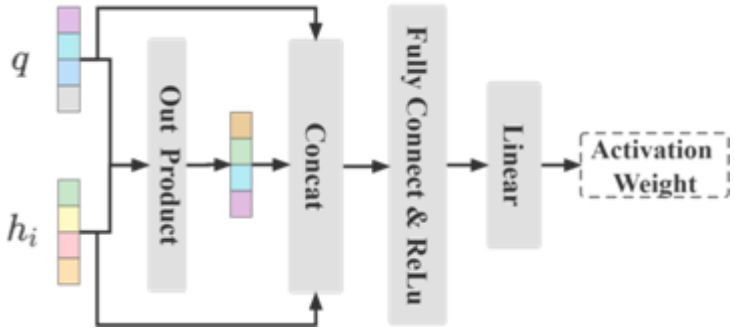


Figure 4- 5: Activation Unit calculation process3

The calculation process is shown in Equations (4-15) to (4-18):

$p = q * h_i$ (4-15)

$R_1 = \text{ReLu}\{W_{r1}(q \oplus p \oplus h_i) + b_{r1}\}$ (4-16)

$R_2 = \text{ReLu}(W_{r2}R_1 + b_{r2})$ (4-17)

$\beta_i = \text{SoftMax}(W_{r3}R_2 + b_{r3})$ (4-18)

4.3.3 Feature Interaction layer

Based on the attention weight β obtained in the previous step and the hidden state sequence $h_{T-K:T}$, get the corrected disease hidden information $\tilde{h}_{T-K:T}$ ($\tilde{h}_{T-K:T} = [h'_{T-K}, \ldots, h'_{T-1}, h'_T]$). The $\tilde{h}_{T-K:T}$ is input into the Sum Pooling layer for summing, and then stitched together with the individual feature information q to obtain the final disease information feature u representation.

The calculation process is shown in formula (4-19):

$u = q \oplus \sum_{j=0}^{K} \beta_{T-j} h_{T-j}$ (4-19)

4.4 Risk assessment of diabetic cardiovascular disease

The disease risk assessment module takes the output u of the individual feature interaction module as the input. Through a fully connected network, a binary label $\hat{y}$ (model training process) output indicates the risk of the patient's diabetic cardiovascular disease. It should be noted that in order to better display the size of the disease risk, in addition to the binary label $\hat{y}$ output during the model training, the evaluation results before softmax operation $\tilde{y}_i$ is used as the output.

Furthermore, this chapter selects the cross-entropy function to calculate the losses, with a mathematical representation as shown in formulas (4-20), (4-21), and (4-22):

$\tilde{y} = W_y u + b_y$ (4-20)

$\hat{y} = \text{softmax}(\tilde{y})$ (4-21)

$\mathcal{L}(y, \hat{y}) = -(\hat{y}\log(y) + (1 - \hat{y})\log(1 - y))$ (4-22)

Specifically, W_y and b_y are the network parameters, y represents the true value of the patient's diabetic CVD risk, $\hat{y}$ is the output value of the model disease risk assessment function, and $\mathcal{L}(\cdot)$ is the loss function of the model task.

5. Experiments

5.1 Experimental data description

The study was approved by the ethics Committee of Ruijin Hospital, and written informed consent was obtained from each participating patient in accordance with the Declaration of Helsinki. The patient information is shown in Table 5-1. Our dataset consists of biochemical index data and personal information of diabetic patients in Shanghai Ruijin Hospital from August 1, 2009 to July 30, 2021. A total of 33048 patients and 61646 medical records were included, including 19899 men and 13149 women. Based on domestic and foreign literature and feature selection, Seven metabolic indicators, including glycosylated hemoglobin (HbA1c), 2-hour postprandial blood glucose (GLU4), cholesterol (CHOL), triglyceride (TG-B), high-density lipoprotein (UHDL), low-density lipoprotein (ULDL) and apolipoprotein B (APOB), were selected as metabolic characteristics in this dataset. At the same time, six individual characteristics, including gender, age, history of diabetic foot disease, history of diabetic eye disease, history of diabetic nephropathy and history of diabetic peripheral neuropathy, were selected as the experimental data together with the above seven metabolic characteristics.

Table 5- 1: Details of Patient Information

	Statistic	Value
DataSet	# patients	33048
	# visit	61646
	# positive label	12680
	# negative label	20368
	% female	60.21%

5.2 Experimental Model

We evaluated our proposed causal-NET model on baseline models, including three traditional machine learning methods (LR, RF, GBDT) and four deep learning methods (RNN, GRU, LSTM, and TLSTM). In order to demonstrate the effectiveness of Causal stability and individual feature

Licensed Under Creative Commons Attribution CC BY

interaction in causal-NET, We also implemented five versions of causal-NE and TLSTM, respectively. It is worth noting that there are many advanced clinical prediction models that utilize attention mechanisms to extract long-term dependencies in patient history visits [41, 42, 43, 19], and they are orthogonal to our contribution. Cause-net focuses on incorporating the heterogeneity of individual patient characteristics and the difference of disease background into model learning, which can be easily combined with attention mechanisms.

5.3 Experimental environment and evaluation indicators

We implement our proposed baseline and target models on TensorFlow 2.2.0 and Scikit-Learn 1.0.2, and use Adam optimizer for training. Through parameter tuning, in this section, the Epoch and Batch Size parameters during model training are set to 100 and 125, the learning rate is set to 0.001, the dimension of the individual feature embedding vector used in the deep learning baseline model and the Causal-Net model is set to 64, and the dimension of the hidden vector is set to 128. In addition, the dataset is randomly divided into 10, and all experimental results are averaged by ten-fold cross-validation. Seven groups of training are used each time, one group of validation and two groups of testing, and the validation set is used to determine the best value of parameters in the training iteration.

Finally, we use the four most commonly used evaluation indexes in dichotomous classification problems as experimental evaluation criteria to compare the performance of all methods. Namely, Accuracy, Recall, F1-score and Area Under the Receiver Operating Characteristic curve (AUC).

5.3 Comparative Experiment

In order to better discuss and analyze the performance of causal-net model on the risk assessment task of diabetic cardiovascular disease, and evaluate the effectiveness of TLSTM unit and individual feature interaction layer based on Causal stability learning in the target model, a number of comparative experiments were set up in this section. Four experiments were conducted, including the comparison experiment of important parameters of the experiment, the comparison experiment of individual feature fusion method and the whole model comparison experiment.

5.3.1 Selection of important parameters

Parameter T: Diabetic disease is a chronic metabolic disease. To accurately assess the risk of diabetic cardiovascular disease, it is important to track and learn the long-term health status of patients. According to the data analysis in Section 3.1.3, 65.37% of the patients had only one visit record, and the sample of patients with long-term regular visits (T > 10 here) was small, accounting for only 1.17% of the data set. In order to reduce the impact of the difference in data volume, this section only discusses the case when the step size T is less than or equal to 7. At the same time, in order to increase the research scope of data and reduce data loss, this section discusses the data records of the last T times of patients, that is, all patients have at least one visit record. The data statistics are shown in Figure 4-6.

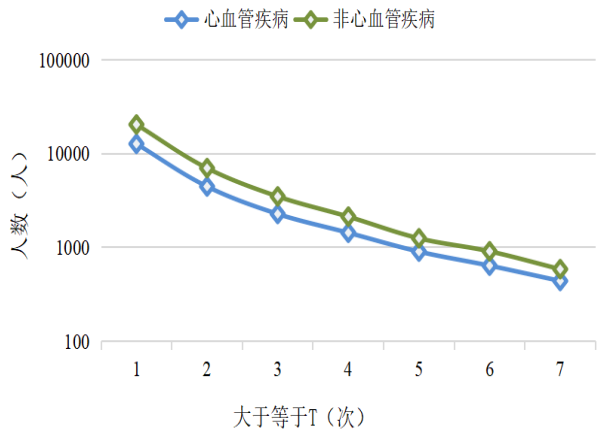


Figure 4-6: Statistics of the number of visits not less than T4

where, K=1 was set, and the step size T was compared on the model Causal-Net. The best experimental results were obtained by adjusting the experimental parameters such as Epoch and Batch, as shown in Table 4-3.

Table 4-3 Parameter selection of step size T1

Step T	Accuracy	Recall	F1-Score	AUC
T=1	0.8972	0.7893	0.8549	0.8768
T=2	0.9033	0.8054	0.8672	0.8859
T=3	0.9148	0.8687	0.8911	0.9072
T=4	0.9251	0.8786	0.9028	0.9171
T=5	**0.9314**	**0.8817**	**0.9188**	**0.9261**
T=6	0.9260	0.8317	0.9008	0.9108
T=7	0.9211	0.8731	0.9007	0.9150

The experimental results show that, with the increase of T=1 to T=5, the evaluation indexes of the model are getting better, and the long-term information of patients can be obtained by tracking and learning, which can effectively improve the accuracy of disease risk assessment. It is considered that this is caused by diabetes itself as a chronic metabolic disease. Therefore, when the amount of data allows, it is necessary to collect as much information as possible to improve the accuracy of the disease risk assessment task. In addition, it can be seen from Table 4-3 that when T is 5, the model index reaches the best level and then begins to show a downward trend. Combined with the data statistics in Figure 4-6, the influence of data volume is considered here. Therefore, in combination with the experimental results in Table 4-3 and to reduce the impact of too small data volume on other model experiments, the patient visit step T=5 is selected as the parameter of the subsequent experiments in this chapter.

Parameter K: T=5 is set here to discuss the influence of parameter K in the interaction module of individual characteristics on the model. The experimental results are shown in Table 4-4.

Licensed Under Creative Commons Attribution CC BY

Table 4-4: Parameter selection of observation window K2

Watch window K	Accuracy	Recall	F1-Score	AUC
K=1	0.9314	0.8817	0.9188	0.9261
K=2	**0.9433**	**0.8984**	**0.9333**	**0.9390**
K=3	0.9267	0.8441	0.9109	0.9178
K=4	0.9173	0.8978	0.9051	0.9152
K=5	0.9102	0.8495	0.8978	0.9000

As shown in Table 4-4, when K=2, the model achieves the best performance; When K is greater than 2, the performance of the model decreases, which is because the long-term dependence of information has been modeled and learned in Causal-Net. When T=5 and the step size is small, too much value of K will cause the model to pay too much attention to the redundant part of feature information and reduce the performance of the model, which proves the advantages of LSTM unit in long-term information dependence learning to a certain extent. Therefore, K=2 is chosen as the observation window size of the interaction layer of individual features in subsequent experiments.

5.3.2 Comparison of individual feature fusion methods

The interaction of individual features based on attention mechanism is the focus of the design of Causal Net model in this chapter. In order to prove its effectiveness, four versions of Causal Net and TLSTM are respectively implemented based on traditional feature fusion methods (Concat and Add here). Table 4-5 describes the evaluation results of different feature fusion methods on model accuracy, recall and F1 score. "Metabo" in Table 4-5 indicates that the model only uses the metabolic characteristics of patients as input data, and "Concat" and "Add" indicate the individual feature fusion methods adopted by the model. Here, the metabolic medical characteristic data of patients is defined as "Metabo", the individual characteristic data of patients is defined as "Indifac", and the attention-based individual feature interaction method proposed in this chapter is defined as "Interfus".

It should be noted that, in order to further observe and compare the performance of different feature fusion methods on the task of risk assessment of diabetic cardiovascular disease, the performance of AUC on each model was separately presented in this experiment, as shown in Figure 4-7.

Table 4-5: Comparison of feature fusion methods

Model	Metabo	Indifac	Interfus	Accuracy	Recall	F1-Score
TLSTM_Metabo	Square root			0.8983	0.7688	0.8693
TLSTM_Add	Square root	Square root		0.9243	0.8656	0.9069
TLSTM_Concat	Square root	Square root		0.9209	0.875	0.8974
TLSTM_At	**Square root**	**Square root**	**Square root**	**0.9312**	**0.8762**	**0.9117**
Causal-Net_Metabo	Square root			0.9008	0.8281	0.8724
Causal-Net_Add	Square root	Square root		0.9267	0.8817	0.9136
Causal-Net_Concat	Square root	Square root		0.9338	0.8602	0.9195
Causal-Net	**Square root**	**Square root**	**Square root**	**0.9433**	**0.8984**	**0.9333**

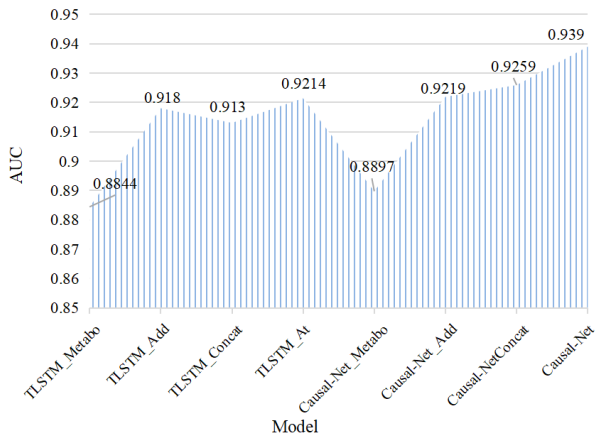


Figure 4-7: Comparison of feature fusion methods on AUC5

As can be seen from Table 4-5 and Figure 4-7, the traditional TLSTM model and the model Causal-Net proposed in this chapter can also achieve good results when only metabolic indexes are used as model input data, with Accuracy indexes reaching 89.83% and 90.08%, respectively. The AUC index was 88.44% and 88.97%, respectively. Then, Concat and Add, the traditional feature fusion methods, were used to Add the individual feature data into the model learning, and the F1-score and Recall indexes of the above two basic models were significantly improved. Compared with TLSTM_Metabo, TLSTM_Concat increased the index Recall by 13.81%, and Causal Net_Add increased the index F1-score by 4.72% compared with Causal Net_Metabo. This shows the importance of individual feature learning in disease risk task.

In addition, when the model adopts the attentional mechanism based individual feature interactive fusion method proposed in this chapter (namely TLSTM_At and causal-Net), the Accuracy, Recall and F1-score of other evaluation indexes are significantly better than other models and fusion methods. The accuracy of TLSTM_At and Causes-NET reached 93.12% and 94.33%, the recall rate was 87.62% and 89.84%, the F1 score was 91.17% and 93.33%, and the AUC index was 92.14% and 93.90%, respectively. The results showed that TLSTM_At and Causes-NET had the best performance in their comparison models. The experimental results strongly prove that the individual feature interaction layer is effective in improving the performance of the target task.

Licensed Under Creative Commons Attribution CC BY

5.3.3 Causal stable learning analysis

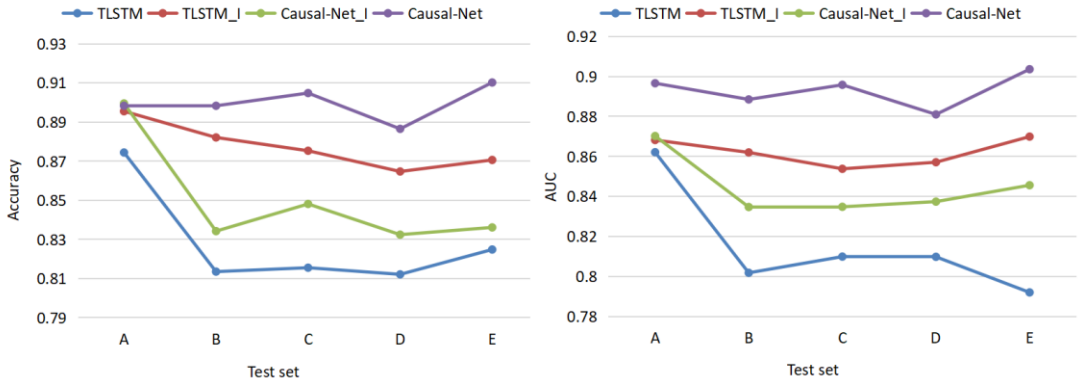
From the data analysis in Section 3.1, it can be seen that the background of diabetic complications among the patients in this dataset is different. Therefore, the experiment in this section will discuss the performance of the long short-term memory unit based on Causal stability and time perception in the causal-net model on this target task. According to the distribution statistics of individual characteristics of patients in Chapter 3, in order to reduce the error caused by the small amount of data, the background conditions of five diseases with a large number of patients were selected here. The details are shown in Table 4-6 below.

Table 4-6: Prevalence of diabetic complications3

Alpha code	Diabetes Complications
A	No other complications of diabetes occurred
B	Only diabetic nephropathy
C	Only diabetic eye disease
D	Only diabetic peripheral neuropathy
E	Concurrent diabetic nephropathy and diabetic eye disease

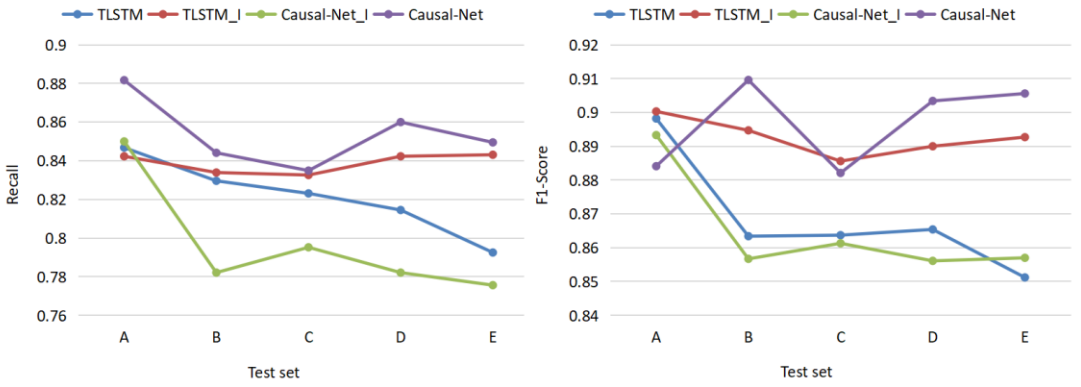
In order to facilitate the observation of the impact of differences in disease background, data of patients without other complications of diabetes were used as the training set, and data of patients in Table 4-6 were used as the test set. The ratio of training set to test set was 4: 1 for data preparation. The experimental results are shown in Figure 4-8.

It can be seen from Figure 4-8 that TLSTM and causal-NET_I models without Causal weights perform best when the patient samples in the training set and test set are of the same disease background, namely, data set A, and the evaluation indexes such as Accuracy, AUC and Recall all reach the maximum value. At the same time, it can be found that the model TLSTM and causal-NET_I without Causal weights in the rest of the test sets, that is, when the patient samples in the training set and the test set have different disease backgrounds, the model performance is significantly different from that in the test set A. When the Causal stability learning module is added to the model, which refers to TLSTM_I and causal-net, it can be seen from the figure that the evaluation indexes of the model on different test sets are relatively similar in size, which can reflect the stability of the model to a certain extent. By comparing the two groups of models, namely, model TLSTM and TLSTM_I, and model causal-net and causal-net_i, it can be seen that the model based on Causal stability and time awareness has more stable and better index performance, and can better complete the task of risk assessment of diabetic cardiovascular disease.



a) Model Accuracy on different test sets

b) Model AUC on different test sets



Model Recall on different test sets

Model F1-Score on different test sets

Figure 4- 8: Comparison of models in different disease backgrounds6

5.3.4 Overall comparison of models

In order to further observe the performance of the causes-NET model in the risk assessment task of diabetic cardiovascular disease, this section evaluates the causes-NET model proposed in this chapter on different baseline models, including LR, RF, GBDT and RNN, GRU, RNN and RNN. Four deep learning methods, LSTM and TLSTM, were proposed in this section. The experimental results are shown in Table 4-7. The performance of disease risk assessment tasks on machine learning is almost worse

Licensed Under Creative Commons Attribution CC BY

than that of the deep learning model, which is considered because the machine learning model loses the timing information of medical visits and the individual characteristics of patients. At the same time, the results in the table show that TLSTM is superior to LSTM model, which indicates the importance of irregular visit time information in patients' medical data in the risk assessment task of diabetic cardiovascular disease. In addition, it should be noted that the individual feature fusion method of TLSTM model here is "Concat" method, which has a better performance in the previous section. In this case, Causal Net, the model proposed in this chapter, also shows a significant advantage.

Table 4-7: Comparison of models on target tasks4

Model	Accuracy	Recall	F1-Score	AUC
LR	0.8497	0.7485	0.7913	0.8302
RF	0.8608	0.8010	0.8225	0.8489
GBDT	0.8603	0.8147	0.8352	0.8565
RNN	0.9078	0.8441	0.8895	0.9009
GRU	0.9031	0.8548	0.8858	0.8979
LSTM	0.9152	0.8550	0.8959	0.9075
TLSTM	0.9243	0.8656	0.9069	0.9180
Causal-Net	0.9433	0.8984	0.9333	0.9390

In conclusion, in this section, the important modules and their overall performance of the model Causal-Net are experimentally analyzed and compared. At the same time, the model parameter selection is compared, and the optimal parameter is selected. The comparative experimental results with the baseline model provide evidence for the effectiveness and superiority of the cause-NET model in the target task.

6. Summarizes

In this study, we propose a novel deep learning model for the risk assessment of diabetic cardiovascular disease. Our feasible model was divided into three stages. In the first stage, the patient's visit record and the time between visits were taken as the input, and a set of causal weights were obtained based on the covariate balance, which were used to weaken the confounding influence between variable features and the target task, and enhance the stable learning of the model. In the second stage, the causal weights obtained in the previous stage and the individual characteristics of patients were used as inputs, and through the redesigned CA-TLSTM unit, the effective information in the current patient visit data was focused on learning, and the preliminary disease information feature vector was obtained. Then, combined with the individual feature interaction layer, the individual features of patients and the current disease information features are interacted and integrated to obtain a more comprehensive and accurate disease risk feature representation of the final feature information. In the third stage, the fully connected layer is used for our final disease risk prediction. Experimental results show that our model based on causally enhanced CA-TLSTM and individual interaction design can better learn effective features, making it consistently better than the basic model. Compared with other models, our model also consistently performs better on this task, with the experimental evaluation index reaching 94.33%, 89.84%, 93.33% and 93.90% in model accuracy, recall, F1 score and receiver operation feature curve, respectively.

Our proposed model effectively takes the causal relationship between risk factors and the risk of diabetic cardiovascular disease into account in the learning of the model, and enhances the stable learning of the model. At the same time, the integrated learning of individual characteristics of patients strengthened the attention to the heterogeneity of individual characteristics of patients, emphasized the clinical significance of individual characteristics, and solved the problems of confounding association between risk factors and personalized auxiliary diagnosis. In clinical practice, we hope that our model can help physicians identify patients at high risk of diabetes cardiovascular disease to prevent or delay the occurrence of adverse outcomes. In the future, the adaptability and effectiveness of our model in cross-hospital and cross-disease problems need to be further verified on a larger scale, so as to better promote the application of artificial intelligence models in the field of diabetes complication risk prediction.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2019YFE0190500.

References

[1] Parthiban G, Srivatsa S K. Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients [J]. international journal of applied, 2012.
[2] Forbes J M, Cooper M E. Mechanisms of diabetic complications [J]. Physiological reviews, 2013, 93 (1): 137-188.
[3] Leustean A M, Ciocoiu M, Sava A, et al. Implications of the intestinal microbiota in diagnosing the progression of diabetes and the presence of cardiovascular complications [J]. Journal of diabetes research, 2018, 2018.
[4] Cho N H, Shaw J E, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045 [J]. Diabetes Research & Clinical Practice, 2018: 271.
[5] Zafar M I, Zheng J, Wen K, et al. The role of vascular endothelial growth factor-B in metabolic homoeostasis: current evidence [J]. Bioscience Reports, 2017, 37 (4): BSR20171089.
[6] Grøntved A, Hu F B. Television viewing and risk of type 2 diabetes, Cardiovascular disease, and all-cause mortality: a meta-analysis [J]. Jama, 2011, 305 (23): 2448-2455.
[7] Glovaci D, Fan W, Wong N D. Epidemiology of diabetes mellitus and cardiovascular disease [J]. Current cardiology reports, 2019, 21 (4): 1-8.
[8] Grundy S M, Benjamin I J, Burke G L, et al. Diabetes and cardiovascular disease: a statement for healthcare professionals from the American Heart Association [J]. Circulation, 1999, 100 (10): 1134-1146.

Licensed Under Creative Commons Attribution CC BY

[9] Dinesh K G, Arumugaraj K, Santhosh K D, et al. Prediction of cardiovascular disease using machine learning algorithms [C]//2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018: 1-7.
[10] Yang L, Wu H, Jin X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China [J]. Scientific reports, 2020, 10 (1): 1-8.
[11] Domanski M J, Tian X, Wu C O, et al. Time course of LDL cholesterol exposure and cardiovascular disease event risk [J]. Journal of the American College of Cardiology, 2020, 76 (13): 1507-1516.
[12] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques [J]. IEEE access, 2019, 7: 81542-81554.
[13] Gao X, Xie W, Wang Z, et al. Improved Functional Causal Likelihood-Based Causal Discovery Method for Diabetes Risk Factors [J]. Computational and mathematical methods in medicine, 2021, 2021.
[14] Baytas I M, Xiao C, Zhang X, et al. Patient subtyping via time-aware LSTM networks [C]//Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining.2017: 65-74.
[15] Kuang K, Xiong R, Cui P, et al. Stable prediction with model misspecification and agnostic distribution shift [C]//Proceedings of the AAAI Conference on Artificial Intelligence.2020, 34 (04): 4485-4492.
[16] Shen Mei-feng, Zhang Qiu-yi, Gu Shu-jun. Predictive value of HBA1C variation index for cardiovascular complications in type 2 diabetes mellitus [J]. Chin j health laboratory science, 2021, 31 (20): 2527-2530.
[17] Yang Z L, Ke T Y. Research progress of diabetes complicated with cardiovascular disease [J]. China Geriatric Healthcare Medicine, 2021, 19 (1): 4.
[18] Scholes S, Fat L N, Mindell J S. Trends in cardiovascular disease risk factors by BMI category among adults in England, 2003-2018 [J]. Obesity, 2021, 29 (8): 1347-1362.
[19] Bode E D, Mathias K C, Stewart D F, et al. Cardiovascular disease risk factors by BMI and age in United States firefighters [J]. Obesity, 2021, 29 (7): 1186-1194.
[20] D 'Agostino Sr R B, Vasan R S, Pencina M J, et al. General Cardiovascular Risk Profile for Use in Primary Care: the Framingham Heart Study [J]. Circulation, 2008, 117 (6): 743-753.
[21] Elley C R, Robinson E, Kenealy T, et al. Derivation and validation of a new cardiovascular risk score for people with type 2 diabetes: the New Zealand diabetes cohort study [J]. Diabetes care, 2010, 33 (6): 1347-1352.
[22] Conroy R M, Pyrl K, Fitzgerald A P, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project [J]. European Heart Journal, 2002, 24: 987.
[23] Hippisley-Cox J, Coupland C, Robson J, et al. Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database [J]. Bmj British Medical Journal, 2011, 342.
[24] Alaa A M, Bolton T, Di Angelantonio E, et al. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423, 604 UK Biobank participants [J]. PloS one, 2019, 14 (5): e0213653.
[25] Dinh A, Miertschin S, Young A, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning [J]. BMC medical informatics and decision making, 2019, 19 (1): 1-15.
[26] De Rosa S, Arcidiacono B, Chiefari E, et al. Type 2 diabetes mellitus and cardiovascular disease: genetic and epigenetic links [J]. Frontiers in endocrinology, 2018, 9: 2.
[27] Strain W D, Paldanius P M. Diabetes, Cardiovascular disease and the microcirculation [J]. Cardiovascular Diabetology, 2018, 9: 2.2018, 17 (1): 1-10.
[28] Einarson T R, Acs A, Ludwig C, et al. Prevalence of cardiovascular disease in type 2 diabetes: A systematic literature review of Scientific evidence from across the world from 2007 to 2017 [J]. Cardiovascular Diabetology, 2012.2018, 17 (1): 1-19.
[29] Cai X, Zhang Y, Li M, et al. Association between prediabetes and risk of all cause mortality and cardiovascular disease: updated meta-analysis [J]. Bmj, 2020, 370.
[30] Benjamin. Heart Disease and Stroke Statistics-2018 Update: A Report From the American Heart Association (vol 137, pg e67, 2018).

Licensed Under Creative Commons Attribution CC BY