Neural Computing and Applications (2020) 32:9683–9698
https://doi.org/10.1007/s00521-019-04489-y

ORIGINAL ARTICLE

Risk prediction of type 2 diabetes in steel workers based on convolutional neural network

Jian-Hui Wu[1,2] · Jing Li[1] · Jie Wang[1] · Lu Zhang[1] · Hai-Dong Wang[1] · Guo-Li Wang[1,2] · Xiao-lin Li[3] · Ju-Xiang Yuan[1,2]

Received: 26 May 2019 / Accepted: 29 August 2019 / Published online: 16 September 2019
© Springer-Verlag London Ltd., part of Springer Nature 2019

Abstract
With the change in environment and lifestyle, the number of diabetic patients is increasing rapidly. Diabetes has become one of the most important chronic diseases affecting the health of the Chinese people, and complications, disability, death and treatment costs caused by diabetes have placed a heavy burden on families and society. If the high-risk population of diabetes can be identified and the adverse lifestyle can be changed as soon as possible, the incidence of diabetes can be reduced or the onset of diabetes can be slowed down. Risk prediction model can accurately predict the risk of disease and has been widely used in the field of health management and medical care. This study was based on the special occupational group of steel workers, the risk prediction model of type 2 diabetes was established by using convolutional neural network, and the feasibility of the model was discussed. The results showed that the prediction accuracy of the established model in the training set, verification set, and test set is relatively high, which is 94.5%, 91.0%, and 89.0%, respectively. The area under the ROC curve was 0.950 (95 CI 0.938–0.962), 0.916 (95 CI 0.888–0.945), and 0.899 (95 CI 0.899–0.939), respectively, indicating that the model can accurately predict the risk of type 2 diabetes among steel workers, provide a basis for self-health management of steel workers, facilitate the rational allocation of medical and health resources and the development of health services, and provide a basis for government departments to make decisions.

Keywords Convolutional neural network · Steel worker · Type 2 diabetes mellitus (T2DM) · Prediction of morbidity risk

1 Introduction

Diabetes mellitus (DM) is a metabolic disorder syndrome caused by many factors such as genetic factors, environmental factors, and their interactions. Clinically, high blood sugar is the main feature. Its typical cases can show more urine, more drink, more eating, weight loss, etc., which is the "three more and one less" symptoms. DM is generally divided into type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (T2DM) and gestational diabetes mellitus (GDM). Among them, type 1 diabetes patients are young, mostly younger than 30 years old, with sudden onset, polyuria, polydipsia, and high blood sugar levels. Many patients with ketoacidosis as the first symptom, and serum insulin and C-peptide levels are low. ICA, IAA, or GAD antibodies are positive, and oral drugs alone are ineffective and require insulin treatment [1]. Type 2 diabetes is common in middle-aged and elderly people. Obesity has a high incidence and can often be associated with hypertension, dyslipidemia, and arteriosclerosis. The onset of this disease is not easy to detect, there is no symptom in the early stage of the disease, or only mild fatigue and thirst. People with unimportant blood sugar increase need to do a glucose tolerance test to confirm the diagnosis. The patient's serum insulin level is normal or high in the early stage, and the late stage is low. GDM is

✉ Ju-Xiang Yuan
gwxyjxb@ncst.edu.cn

[1] School of Public Health, North China University of Science and Technology, Tangshan 063210, Hebei, People's Republic of China

[2] Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, North China University of Science and Technology, Tangshan 063210, Hebei, People's Republic of China

[3] INSA Centre Val de Loire, Campus de Blois, 3 Rue de la Chocolaterie, CS 23410, 41034 Blois Cedex, France

insulin resistance from cells, but it is caused by hormones secreted by pregnant women, who usually heal after delivery.

In patients with DM, the proportion of type II DM is about 95% [2]. DM can cause macrovascular complications (lesions of cerebrovascular, cardiovascular and lower extremity blood vessels, etc.), microvascular complications (kidney lesions and fundus lesions), and neuropathy (sensory nerves, motor nerves, autonomic neuropathy, etc.). These chronic complications are the main causes of DM disability and death. Since the early stages of DM are generally not easy to find, many type II DM patients have chronic complications before diagnosis.

In recent years, with the aging of the population and changes in lifestyles and living environment, the number of people with diabetes has grown rapidly, and diabetes has become one of the most important chronic diseases affecting the health of our people. According to reports, there were 114 million diabetic patients in China of 2013, which has become the country with the largest number of diabetes patients [3], and complications, disability, death, and treatment caused by diabetes put a heavy burden on families and society. If you can identify high-risk groups of diabetes and make lifestyle changes as soon as possible, it can effectively reduce the number of people with diabetes or reduce their incidence. This is a prevention strategy based on primary prevention of chronic diseases. At present, countries all over the world have recognized the role of risk prediction models in the prevention and treatment of diabetes and established diabetes risk prediction models based on epidemiological survey data from various countries, such as the American Framingham risk score diabetes model [4], Mexican descendants of the USA and non-Spanish Caucasian diabetes prediction model [5], Japanese American individual diabetes risk prediction model [6], Finnish population diabetes risk score model [7], Thailand population risk score model [8], and the British population individual diabetes risk scoring model [9]. Domestic scholars built the first diabetes risk assessment model for the Chinese population in 2007 [10]. In 2009, Chien [11] et al. based on the Taiwanese community population and the Framingham cardiovascular prediction model, constructed a diabetes individual risk scoring model for the Chinese population. In addition, Mi Sheng-Quan [2] used a synthetic analysis method to use the 2002 survey of Chinese residents' nutrition and health status to construct a prediction model for the risk of diabetes in Chinese adult individuals. Due to regional and population differences, the established model is less universal and has localization advantages.

The steel industry is one of the main economic pillars of China. The population of this industry is large, and the health of steel workers is directly related to the development of China's steel industry. For steel workers, because of their long-term exposure to occupational factors such as shift, high temperature, and noise, the risk factors for type 2 diabetes are different from those of the general population. Therefore, the current models are not suitable for the risk prediction of type 2 diabetes in steel workers. In order to protect the health of workers in the steel industry and ensure the stable development of the steel industry in China, it is necessary to establish a type 2 diabetes risk prediction model for workers in the steel industry.

In the current disease risk prediction, the commonly used methods are logistic regression [12], Cox regression [13], support vector machine [14], artificial neural network, etc. [15]. Logistic regression and Cox regression are traditional statistical modeling methods. They are highly interpretable and can obtain the links between odds ratio (OR), risk ratio (RR), and hazard ratio (HR) indicators and diseases. Therefore, they are the most widely used, but these methods have more data. It is required that the prediction accuracy will be lower when the data used do not meet the conditions. Both support vector machine and artificial neural network are common machine learning methods. Support vector machines show unique advantages in small sample learning, nonlinear, high-dimensional problems and generalization ability, but the choice of kernel function will directly affect model effect, and backpropagation (BP) neural network is more widely used in artificial neural network, which has no requirement for data, and establishes the optimal model through self-adjustment and self-adaptation. However, it is prone to fall into local minimum values and overfitting, which limits the generalization ability. In recent years, with the development of computer technology, artificial neural networks have also developed rapidly. CNNs have become one of the research hot spots in many scientific fields due to their high accuracy and good generalization ability. The study intends to use the CNN to establish a model for predicting the risk of type 2 diabetes in workers in the steel industry, in order to provide a basis for self-health management.

2 Design of convolution depth neural network algorithm

2.1 Algorithm principle

Convolution neural network is a feedforward neural network which is different from the common neural network structure [16, 17]. The CNN is mainly composed of convolution layer and lower sampling layer, forming the feature extractor. In the convolution layer of CNN, there are usually several feature planes, each feature plane is

composed of several rectangular neurons, and neurons on the same feature plane share weights, among which the shared weights are called convolution kernel. The shared weights can reduce the number of connections between neurons and the risk of overfitting. Downsampling, also known as pooling, can be regarded as a special convolution process, which can reduce the computational dimension and improve the generalization ability of the model. Convolution and subsampling greatly simplify model complexity and reduce model parameters. The training process of CNN mainly includes forward propagation and backward propagation. The convolutional layer and the lower sampling layer appear alternately in the training process. The basic structure of CNN is shown in Fig. 1.

The purpose of forward propagation is to extract features, which is mainly realized by convolution operation and downsampling operation. The image goes from the input layer to the convolution layer, and then the output value is obtained by activating the function.

$$x^l = f\left(W^l x^l + b^l\right) \tag{1}$$

where l represents the number of layers, ω represents the weight, b represents the bias, and f represents the activation function.

In the process of forward propagation, several feature graphs of the upper layer are convolved by a learnable convolution kernel, and then a new feature graph can be obtained by activating the function.

$$x_j^l = f\left(\sum_{i \in M_j} X_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{2}$$

where l represents the current layer, $l-1$ represents the previous layer, X_j^l represents the j-th feature graph of the current layer, k_{ij}^l represents the convolution kernel corresponding to the i-th feature graph of the current layer and the j-th feature graph of the previous layer, and b_j^l is the offset value.

After the lower sampling layer is immediately connected with the convolution layer, the relative position changes of target tilt and rotation can be ignored, so as to improve the performance and robustness of the algorithm, reduce the dimension of feature graph, and avoid overfitting to some extent. The calculation method of the lower sampling layer is as follows:

$$x_j^l = f\left(\beta_j^l \operatorname{down}\left(x_j^{l-1}\right) + b_j^l\right) \tag{3}$$

where down () represents the lower sampling function.

The purpose of backpropagation is to update the weight of convolution kernel, that is, to perform error operation on the output value of forward propagation and the given sample label, and then obtain gradient through convolution and pooling to update the convolution kernel [18–20]. Loss function of error operation generally adopts loss function of difference of two squares. For multi-classification problems with c number of categories and N samples, the loss function table of difference of two squares:

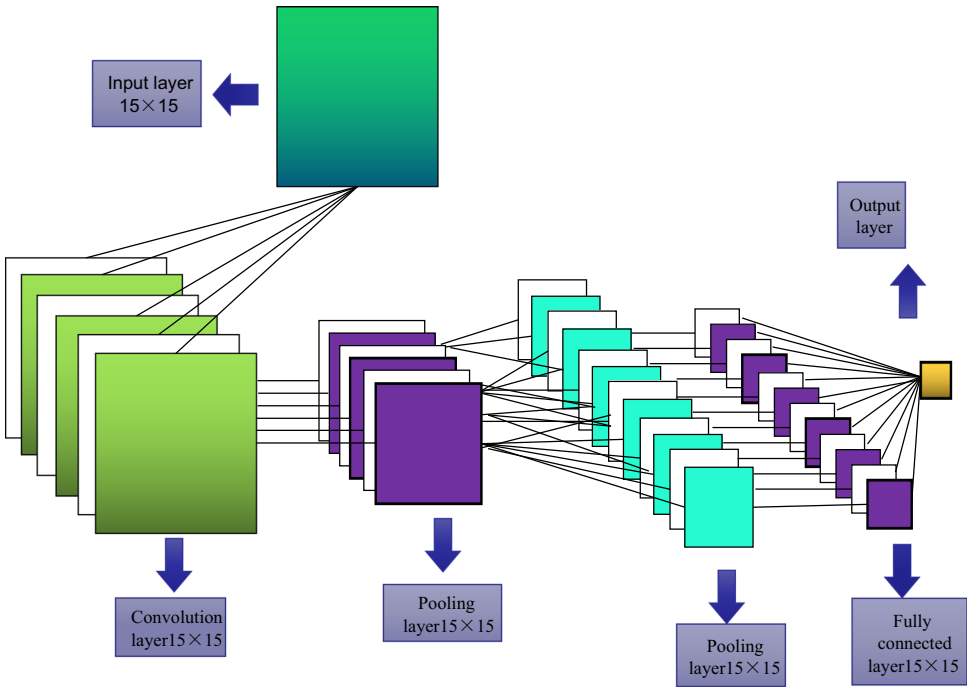

Fig. 1 The basic structure of CNNs

$$E^N = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}\left(t_k^n - y_k^n\right)^2 \quad (4)$$

where E^N represents the total error of N samples, t_k^n represents the k-dimension label corresponding to the n-th sample, and y_k^n represents the k-dimensional output corresponding to the n-th sample.

After obtaining the features through convolution, the features were used for classification. In order to solve the overfitting problem, the features at different positions were aggregated and statistically analyzed. This aggregation operation is called pooling. Pooling layer has the function of feature invariance, feature dimensionality reduction, and preventing overfitting to a certain extent to facilitate optimization. The common method for pooling layer is Max pooling [21]. The idea is very simple. For each 2 * 2 window, select the maximum number as the value of the corresponding element of the output matrix; for example, if the largest number in the first 2 * 2 window of the input matrix is 6, then the first element of the output matrix is 6, and so on, as shown in Fig. 2.

The fully connected layer after pooling plays the role of "classifier" in the entire CNN; that is, after the deep network of convolution, activation function and pooling, the results are recognized and classified through the fully connected layer. First, the results of convoluted, activated, and pooled deep network are strung together, as shown in Fig. 3.

2.2 Algorithm implementation flow

(1) Basic principle

The feature set output from the convolutional layer is obtained by convolving data using a number of learnable filters and data from the previous layer. The role of convolution is:

(a) Convolving the input image and the filter to reduce the noise interference while enhancing the feature information of the original image;

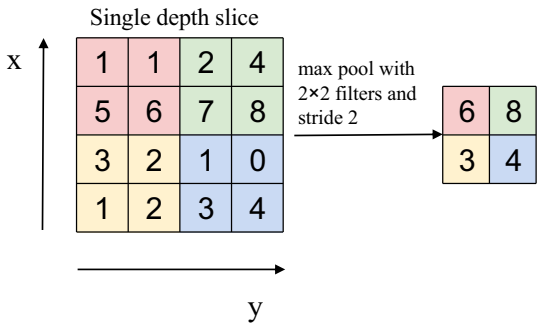

Fig. 2 Schematic diagram of max pooling

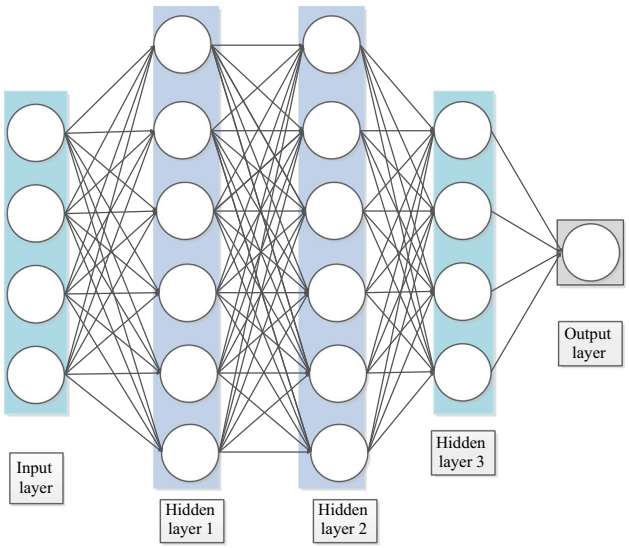

Fig. 3 Full connection layer structure

(b) Reflecting the local receptive field and weight sharing;

(c) Automatically learn image features by convolution, saving time and effort.

The CNN training process is shown in Fig. 4.

In the convolutional layer, the feature map of the upper layer is convoluted with a learnable convolution kernel, and the obtained result is input into an activation function,

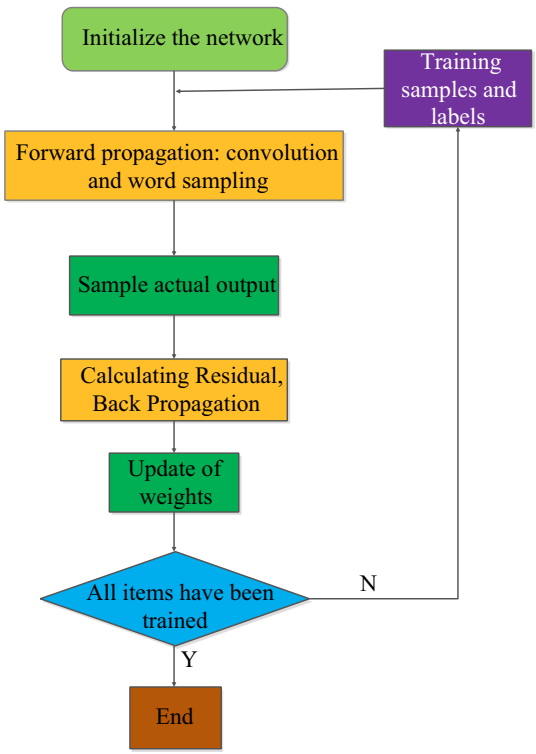

Fig. 4 CNN training flowchart

and finally the output is combined to form a new feature map, wherein different volumes are included. The kernel is obtained with different feature maps, and each output feature map may be obtained by a combined convolution of a plurality of feature maps. The calculation of the convolutional layer is as follows:

$$X_j^l = f\left(\sum_{i\in M_j} X_i^{l-1} \times K_{ij}^l + b_j^l\right) \tag{5}$$

X_j^l represents the *jth* feature map of the *lth* layer, K_{ij}^l is the convolution kernel function, and $f()$ is the activation function. b_j^l is the bias parameter. M_j represents a selected set of input feature maps from which a combination is selected as an input feature map, each output feature map having an offset coefficient, but for each particular output graph. Convolution kernels of each input feature map are simultaneously convolved, for example, two feature maps a and b are output, and their input feature maps all include c, but they are summed from c through different convolution kernels.

(2) Gradient calculation of convolutional layer

The general convolutional layer will be followed by a downsampling layer $l + 1$. It can be known from the BP neural network model [22] that to obtain the weight update of the convolutional layer l, the error signal δ of each neuron in the l layer is first obtained. In order to obtain δ, it is necessary to first sum the error signals of the neurons of the next layer to obtain δ^{l+1}, and then multiply the weight W which corresponding to the connections, and multiply by the activation function f of the input of the neurons in the layer. Derivative, from which the error signal δ^i of each neuron in layer l can be obtained. However, the convolutional layer is followed by a downsampling layer, and the error signal of the neuron node of the sampling layer corresponds to the area of the sampling window in the convolutional layer output feature map, so each neuron of the feature map in the layer is only connected one of the neurons in the $l + 1$ layer corresponds to the feature map. Therefore, the feature map corresponding to the downsampling layer is required to obtain the error signal of the layer, so that the size is consistent with the feature map size of the convolution layer. The weights of the downsampling layer feature maps are all the same constant β, so the result obtained by the previous step is multiplied by β, and the error signal δ of the layer l can be obtained, matching the characteristic map j of the corresponding downsampling layer. This process can be repeated to obtain the error signal b_j^l of a feature map in the convolutional layer:

$$\delta_j^l = \beta_j^{l+1}\left(f'\left(u_j^l\right) \bullet up\left(\delta_j^{l+1}\right)\right) \tag{6}$$

Among them, for an up-sampling operation, we can pass the above formula; the error signals in the layers are summed to obtain a bias based on the bias:

$$\frac{\partial E}{\partial b_j} = \sum_{u,v}\left(\delta_j^l\right)_{uv} \tag{7}$$

Finally, the weight gradient for the convolution kernel can be obtained using the conventional bp algorithm [23]. Since many connections in a convolutional neural network are weight-shared, for a given weight, all connections associated with that weight need to be graded for that point, and finally these gradients are summed:

$$\frac{\partial E}{\partial K_{ij}^l} = \sum_{u,v}\left(\partial_j^l\right)_{uv}\left(p_i^{l-1}\right)_{uv} \tag{8}$$

Here, p_i^{l-1} is a small block in X_i^{l-1} that is multiplied by K_{ij}^l element by element at the time of convolution. The value of the (u, v) position of the output convolutional feature map is the result of multiplying the small block of the upper (u, v) position by the convolution kernel K_{ij}^l by element. The above formula can be implemented in a convolution function in Python without the need to laboriously remember which of the small pieces of the input feature map each neuron corresponds to the output feature map:

$$\frac{\partial E}{\partial K_{ij}^l} = \mathrm{rot}180\left(\mathrm{conv}2\left(X_i^{l-1}, \mathrm{rot}180\left(\delta_j^l\right), ``\mathrm{valid}''\right)\right). \tag{9}$$

Here, rotate δ feature map is for cross-correlation rather than convolution, and the output is reversed back. When the sample is convolved in forward propagation, the convolution kernel will follow the predetermined direction.

(3) Gradient calculation of the downsampling layer

The principle of the downsampling layer is relatively simple, and each output feature map size is a reduced version of the size of the input feature map:

$$X_j^l = f\left(\beta_j^l \mathrm{down}\left(X_j^{l-1}\right) + b_j^l\right) \tag{10}$$

Among them, down() is the downsampling function, and the sampling window size is $n \times n$. Thus, the output feature map is reduced by n times, and the purpose is to reduce the resolution to obtain the scale invariance. Each output feature map has its own multiplicative bias parameter β and additive bias parameter b.

To calculate the error signal of the downsampling layer feature map, first find which small block in the sensitivity feature map of the current layer corresponds to which pixel of the sensitivity feature map of the next layer, and then use the backpropagation algorithm to calculate. The error signal of the current subsampling layer is obtained by recursing the error signal of the next layer:

$$\delta_j^l = f'\left(u_j^l\right) \bullet \text{conv}\,2\left(\delta_j^{l+1}, \text{rot}\,180\left(K_j^{l+1}\right), \text{“full”}\right) \quad (11)$$

Before the calculation, the convolution kernel needs to be rotated 180°, so that the convolution function can perform correlation calculation, and full is a full convolution function, which can process the convolution boundary and fill the missing pixels by 0. The next step can obtain the gradient of β and b:

$$\frac{\partial E}{\partial b_j} = \sum_{u,v}\left(\delta_j^l\right)_{uv} \quad (12)$$

$$\frac{\partial E}{\partial \beta_j} = \sum_{u,v}\left(\delta_j^l \bullet \text{down}\left(X_j^{l-1}\right)\right)_{uv} \quad (13)$$

Similar to the BP algorithm, the weight update from the time t to $t + 1$ in the CNN can be obtained as follows:

$$W(t+1) = W(t) + \eta\delta(t)x(t) \quad (14)$$

Among them, η is the learning rate, $x(t)$ is the input of neurons, $\delta(t)$ is error.

2.3 Parameter setting and hardware and software platform

TensorFlow [24] is a system that transmits complex data structures to an artificial intelligence neural network for analysis and processing. TensorFlow can be used in many machine depth learning areas such as speech recognition or image recognition. It supports CNN, RNN, and LSTM algorithms, which are the most popular deep neural network models currently available in image, speech and NLP.

Since TensorFlow can support CNN, RNN and LSTM algorithms, there must be certain data processing rules within it [25]. The following figure will show the specific process of TensorFlow processing data more clearly:

It can be seen from the figure x that the input data will pass through one or more hidden layers. Inside the hidden layer, each data will be given a certain weight and offset, so the input data will be given a new value to be output.

$$x_{i\_\text{new}} = \omega_i x_i + b_i \quad (15)$$

These new values are not exactly in line with the ideal expectations, so they will be subjected to a comprehensive data processing process. Through the data processing process, these new data will get a new weight and offset. Then loop through the above process to get the final output.

In this paper, TensorFlow will be used to support the characteristics of CNN algorithm, and based on the results of thousands of questionnaires, the analysis and prediction of whether individuals have diabetes.

Figure 5 reveals that running the CNN algorithm on TensorFlow is a process of constant cyclic assignment. Weight when performing this process And the offset b is an indispensable parameter. Although the relevant parameters will be updated continuously with the data processing process, since the related data update of the CNN algorithm is based on the gradient descent algorithm, if the original parameter setting is not appropriate, it will lead to similar the problem of gradient disappearing occurs too early. In many TensorFlow instances, the parameters are initialized to 0, but doing so will cause all the neurons to start doing the same calculations, and eventually each neuron in the same layer gets the same parameters. Therefore, it is necessary to set a good initialization parameter for the purpose of accelerating convergence and getting an optimal result.

Current mainstream weight initialization methods include all-zero initialization, random initialization, Xavier initialization, He initialization, and Pre-train initialization. The all-zero initialization, random initialization method only considers the variance of each layer of input, usually there are some problems, such as small gradient in the back propagation calculation, the gradient “signal” is weakened, the variance of the intermediate result $\sum_i \omega_i x_i + b$ is small, the neurons are not saturated Learning speed will not slow down, etc. Therefore, the above method is only applicable to small neural networks [26].

Combined with the above problems, this paper uses the Xavier initialization method to set the initialization weight of the input data. With output layer Y and input layer according to TensorFlow's data processing principle x, there must be the following relationship:

$$Y = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n \quad (16)$$

According to the variance formula of the independent variable multiplication, we can calculate:

$$\begin{aligned}\text{Var}(\omega_i x_i) = {} & E[x_i]^2\text{Var}(\omega_i) + E[\omega_i]^2\text{Var}(x_i) \\ & + \text{Var}(\omega_i)\text{Var}(x_i)\end{aligned} \quad (17)$$

In the Xavier initialization method, the input X and the weight W are both zero mean, so the above equation is simplified to $\text{Var}(\omega_i x_i) = \text{Var}(\omega_i)\text{Var}(x_i)$.

Because all x_i and ω_i are independent, there are:

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(\omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n) \\ &= n\,\text{Var}(\omega_i)\text{Var}(x_i)\end{aligned} \quad (18)$$

That is, the variance of the output is related to the input. To make the variance of the output the same as the input, it means $n\text{Var}(\omega_i) = 1$, therefore $\text{Var}(\omega_i) = \frac{1}{n} = \frac{1}{n_{\text{in}}}$, $\text{Var}(\omega_i) = \frac{1}{n} = \frac{1}{n_{\text{out}}}$.

Since n_{in} and n_{out} usually are not equal, these two variances cannot be met at the same time; use $\text{Var}(\omega_i) = \frac{2}{n_{\text{in}}+n_{\text{out}}}$ to replace. According to the evenly distributed variance, the uniform distribution of W is reversed. Since the variance of the uniform distribution of the $[a, b]$

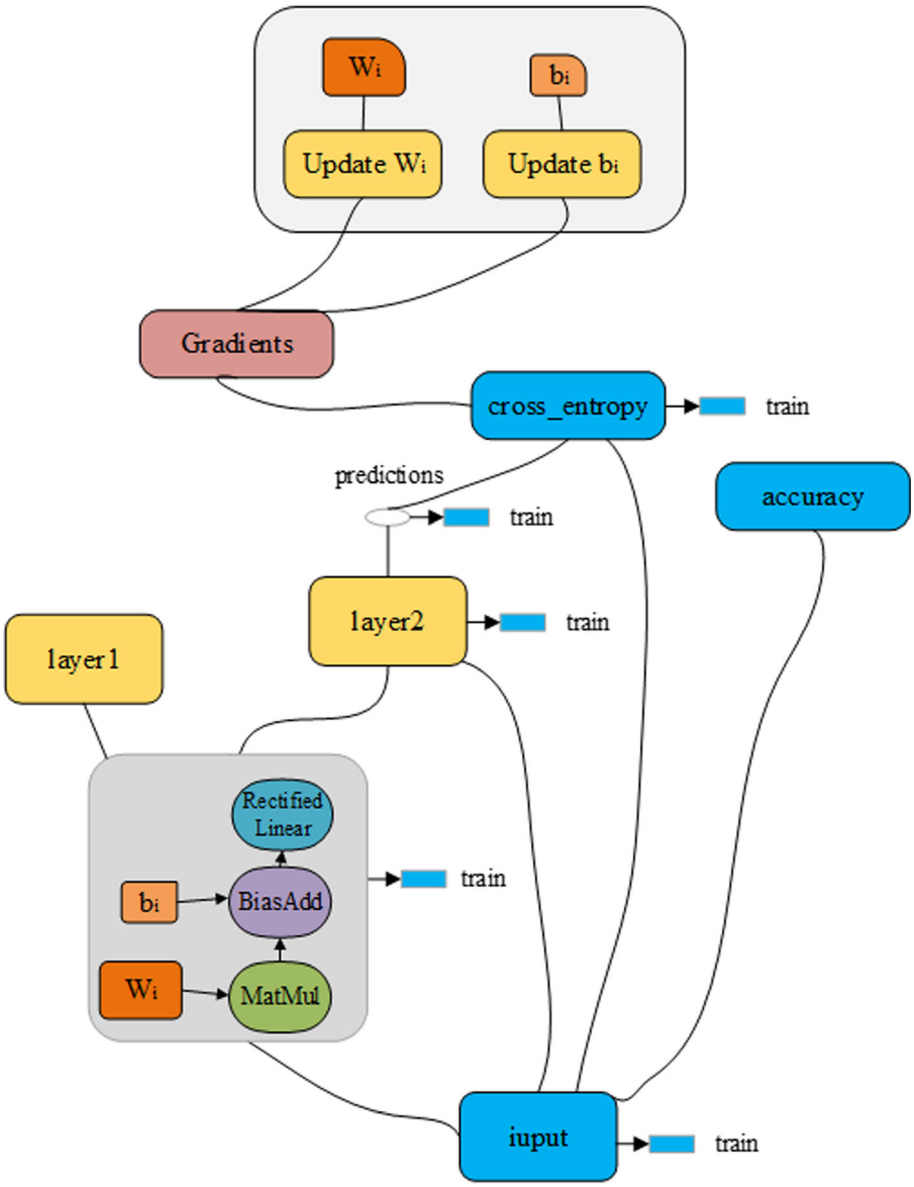

Fig. 5 Flowchart of TensorFlow processing data

interval is Var $= \frac{(b-a)^2}{12}$ with an average of 0, you can get $b = \frac{\sqrt{6}}{\sqrt{n_{\text{in}}+n_{\text{out}}}}$.

In summary, the initialization weight should be set according to the input dimension n, the output latitude m, and the specific value should be in uniform distribution $\left[-\sqrt{\frac{6}{m+n}}, \sqrt{\frac{6}{m+n}}\right]$.

Since the real problems are not all capable of being linearly separable through the origin. Without bias, all of our dividing lines pass through the origin, and the offset is close to the true value [27]. However, since the offset will change continuously during the calculation process, the initial value setting will not have an excessive effect on the final result. Therefore, for the initialization of the bias, the all-zero initialization method is adopted, that is, the offset of all input values is 0.

The learning rate represents the speed at which information accumulates over time in a neural network. If the learning rate is set too low, the training will progress very

Table 1 The flowchart of research object selection

Information of research subjects	Number of cases
Total cases	8648
Working hours less than 1 year	938
Age over 60	357
Incomplete information	1399
Final choice	5954

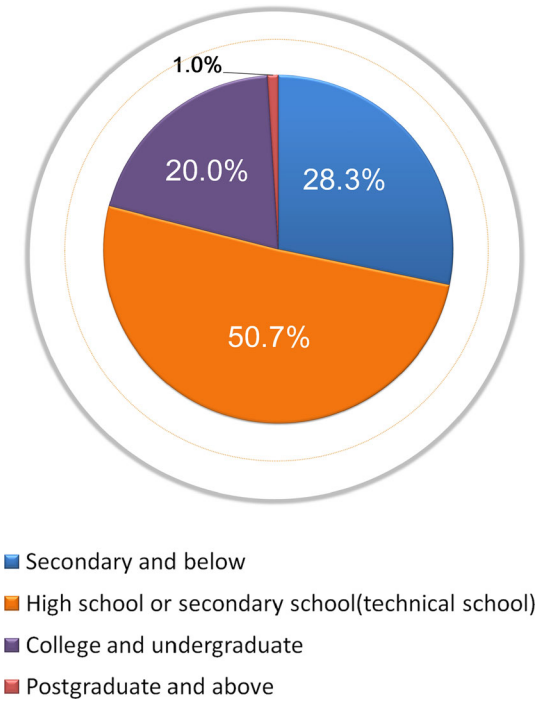

Fig. 6 Education of the study subjects

slowly: however, if the learning rate is set too high, it may have undesired consequences on the loss function. Ideally, it will start at a large learning rate and gradually reduce the speed until the loss value no longer diverges. The choice of learning rate is usually determined at 0.01–0.9 [28].

Platform requirements: TensorFlow can automatically run models from a variety of platforms, from mobile phones, from a single CPU/GPU to a distributed system of hundreds of GPU cards. For ordinary users, a personal computer with a 16G can also operate accordingly.

3 Research objects and research methods

3.1 Research objects

A total of 5954 on-the-job steel workers who performed occupational health examinations and health checkups at a steel group company hospital from February 2017 to June 2017 were selected for the study and were required to be on duty and employed for at least 1 year and aged $\leq$ 60 years. See Table 1.

Table 2 Comparison of general characteristics of patients with type 2 diabetes and non-patients in steel workers

Factors	T2DM n (%)/M (P_{25}, P_{75})		χ^2/Z	P
	No (n = 5251)	Yes (n = 703)		
Age	46 (38, 50)	50 (45, 53)	13.42	< 0.001
Income (yuan)	5500 (4000, 7000)	5000 (4000, 7000)	4.04	< 0.001
Sex			20.30	< 0.001
Male	4768 (90.8)	674 (95.9)		
Female	483 (9.2)	29 (4.1)		
Education level			86.26	< 0.001
Secondary and below	1400 (26.7)	285 (40.5)		
High school or secondary school	2675 (50.9)	345 (49.1)		
College and undergraduate	1119 (21.3)	72 (10.2)		
Graduate and above	57 (1.1)	1 (0.1)		
Marital status			23.21	< 0.001
Unmarried	196 (3.7)	5 (0.7)		
Married	4838 (92.1)	654 (93.0)		
Other	217 (4.1)	44 (6.3)		
Ethnicity				
Han nationality	5123 (97.6)	689 (98.0)	0.53	0.47
Other	128 (2.4)	14 (2.0)		
Family history of DM			60.58	< 0.001
No	4755 (90.6)	569 (80.9)		
Yes	496 (9.4)	134 (19.1)		
Hypertension			167.55	< 0.001
No	4542 (86.5)	475 (67.6)		
Yes	709 (13.5)	228 (32.4)		

Table 3 Comparison of lifestyles between type 2 diabetes patients and non-patients in steel workers

Factors	T2DM n (%)/M (P_{25}, P_{75})		χ^2/Z	P
	No (n = 5251)	Yes (n = 703)		
Physical activity			12.06	0.002
Mild	446 (8.5)	85 (12.1)		
Moderate	1939 (36.9)	230 (32.7)		
Severe	2866 (54.6)	388 (55.2)		
Smoking status			13.93	0.001
No smoking	2282 (43.5)	257 (36.6)		
Quit smoking	269 (5.1)	49 (7.0)		
smoking	2700 (51.4)	397 (56.5)		
Passive smoking			0.35	0.560
No	1890 (36.0)	3361 (64.0)		
Yes	261 (337.1)	442 (62.9)		
Drinking status			32.78	< 0.001
No drinking	3253 (62.0)	359 (51.1)		
Alcohol withdrawal	100 (1.9)	23 (3.3)		
Drinking	1898 (36.1)	321 (45.7)		
Vegetable intake			4.68	0.200
Never	41 (0.8)	9 (1.3)		
Occasionally	418 (8.0)	66 (9.4)		
Regularly	2087 (39.7)	260 (37.0)		
Every day	2705 (51.5)	368 (52.3)		
Fruit intake			72.86	< 0.001
Never	80 (1.5)	37 (5.3)		
Occasionally	1847 (35.2)	307 (43.7)		
Regularly	1987 (37.8)	227 (32.3)		
Every day	1337 (25.5)	132 (18.8)		

Table 4 Comparison of occupational factors between patients with type 2 diabetes and non-patients in steel workers

Factors	T2DM n (%)/M (P_{25}, P_{75})		χ^2/Z	P
	No (n = 5251)	Yes (n = 703)		
Shift situation			7.50	0.02
Never	866 (16.5)	115 (16.4)		
Once	924 (17.6)	153 (21.8)		
Now	3461 (65.9)	435 (61.9)		
Noise			0.52	0.47
No	4243 (80.8)	560 (79.7)		
Yes	1008 (19.2)	143 (20.3)		
Occupational heat			1.54	0.21
No	4335 (82.6)	567 (80.7)		
Yes	916 (17.4)	136 (19.3)		

3.2 Research content

The face-to-face survey of the subjects was conducted by a specially trained investigator using a self-developed questionnaire to investigate the general characteristics of the subjects, including gender, age, ethnicity, education level, marital status, income, disease history, family history, health status, etc. Lifestyle includes smoking status, drinking status, diet, physical activity, etc. Physical examination includes height, weight, abdominal circumference, waist circumference, hip circumference, blood pressure, body mass index (weight/height^2 (kg/m^2)), and waist-to-hip ratio (waist/hip circumference). Occupational factors include shifting, contact noise, exposure to high temperature, etc. Venous blood test in the morning detects fasting blood glucose, uric acid, total cholesterol, triglyceride, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, white blood cell count, high-sensitivity C-reactive protein and homocysteine, and other indicators. Hypertension is defined as self-reported hypertension, diagnosed by a doctor, or blood pressure $\geq$ 140/90 mmHg, or taking antihypertensive drugs; diagnosis of type 2 diabetes is based on WHO diagnostic criteria for diabetes [29], FPG $\geq$ 7.0 mmol/L, or reported diabetes diagnosed by a doctor, or taking insulin or hypoglycemic agents; dyslipidemia is defined as TG $\geq$ 71.70 mmol/L, or TC $\geq$ 75.20 mmol/L, or HDL-C $\leq$ 0.91 mmol/L.

Table 5 Comparison of physical examination and laboratory examination results between patients with type 2 diabetes and non-patients in steel workers

Factors	T2DM *n* (%)/*M* (P_{25}, P_{75})		χ^2/Z	*P*
	No (*n* = 5251)	Yes (*n* = 703)		
BMI (kg/m^2)	24.6 (24.0, 25.9)	24.6 (24.6, 27.0)	6.73	< 0.001
WHR	0.88 (0.84, 0.92)	0.90 (0.87, 0.94)	10.63	< 0.001
Abdominal circumference (cm)	90.0 (84.0, 96.6)	93.7 (88.4100.3)	10.58	< 0.001
Uric acid (μmol/L)	382 (322, 442)	359 (308, 438)	3.85	< 0.001
White blood cell count (*10^9/L)	6.42 (5.49, 7.55)	6.66 (5.82, 7.89)	4.80	< 0.001
C-reactive protein (mg/L)	0.01 (0, 0.06)	0.03 (0.01, 0.12)	8.41	< 0.001
Homocysteine (μmol/L)	12.1 (10.1,16.7)	12.2 (10.1, 16.6)	0.09	0.93
Dyslipidemia			60.55	< 0.001
No	2217 (42.2)	189 (26.9)		
Yes	3034 (57.8)	514 (73.1)		

Table 6 Logistic regression analysis results of influencing factors of type 2 diabetes patients in steel workers

Factors	β	Wald χ^2	*P*	OR	95% CI
Age	0.069	109.48	< 0.001	1.07	1.06, 1.09
Gender (male)	− 0.705	11.24	0.001	0.50	0.33, 0.75
Physical activity (mild)					
Moderate	− 0.49	11.24	0.001	0.61	0.46, 0.82
Severe	− 0.505	13.07	< 0.001	0.60	0.46, 0.79
Fruit intake (never)					
Occasionally	− 0.69	9.49	0.002	0.50	0.32, 0.78
Regularly	− 0.937	17.05	< 0.001	0.39	0.25, 0.61
Every day	− 1.086	21.34	< 0.001	0.34	0.21, 0.54
Family history of DM (no)	0.947	65.97	< 0.001	2.58	2.05, 3.24
Hypertension (no)	0.721	53.98	< 0.001	2.06	1.70, 2.49
BMI	0.113	50.32	< 0.001	1.12	1.09, 1.16
Abdominal circumference	0.003	5.02	0.025	1.003	1.00, 1.01
Uric acid	− 0.003	40.43	< 0.001	0.997	0.996, 0.998
White blood cell count	0.063	6.45	0.011	1.07	1.01, 1.12
Dyslipidemia (no)	0.530	30.28	< 0.001	1.70	1.41, 2.05

Table 7 Variable assignment table

Variable name	Variable meaning	Assignment method
y	T2DM	0 = no, 1 = yes
*x*1	Age	Continuous variable (year)
*x*2	Gender	1 = male, 2 = female
*x*3	Physical activity	1 = mild, 2 = moderate, 3 = severe
*x*4	Fruit intake	1 = never, 2 = occasionally, 3 = regularly, 4 = every day
*x*5	Family history of diabetes	0 = no, 1 = yes
*x*6	Hypertension	0 = no, 1 = yes
*x*7	BMI	Continuous variable (kg/m^2)
*x*8	Abdominal circumference	Continuous variable (cm)
*x*9	Uric acid	Continuous variable (μmol/L)
*x*10	White blood cell count	Continuous variable (10^9/L)
*x*11	Dyslipidemia	0 = no, 1 = yes

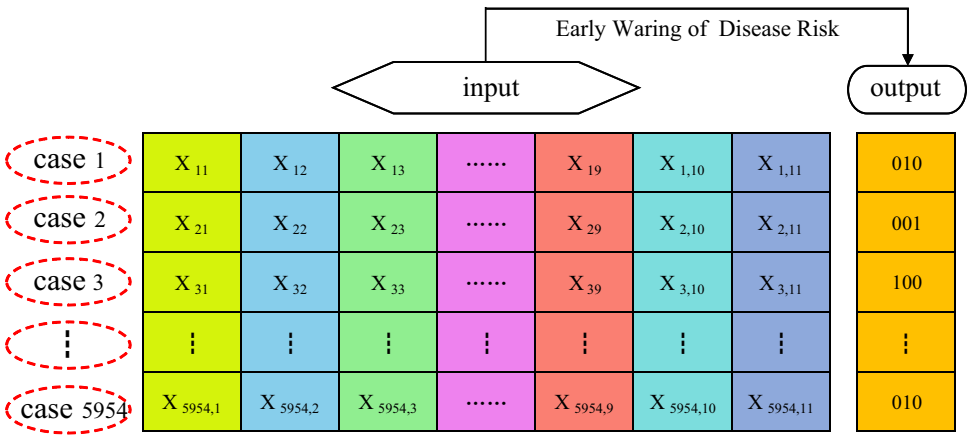

Fig. 7 Sample set structure

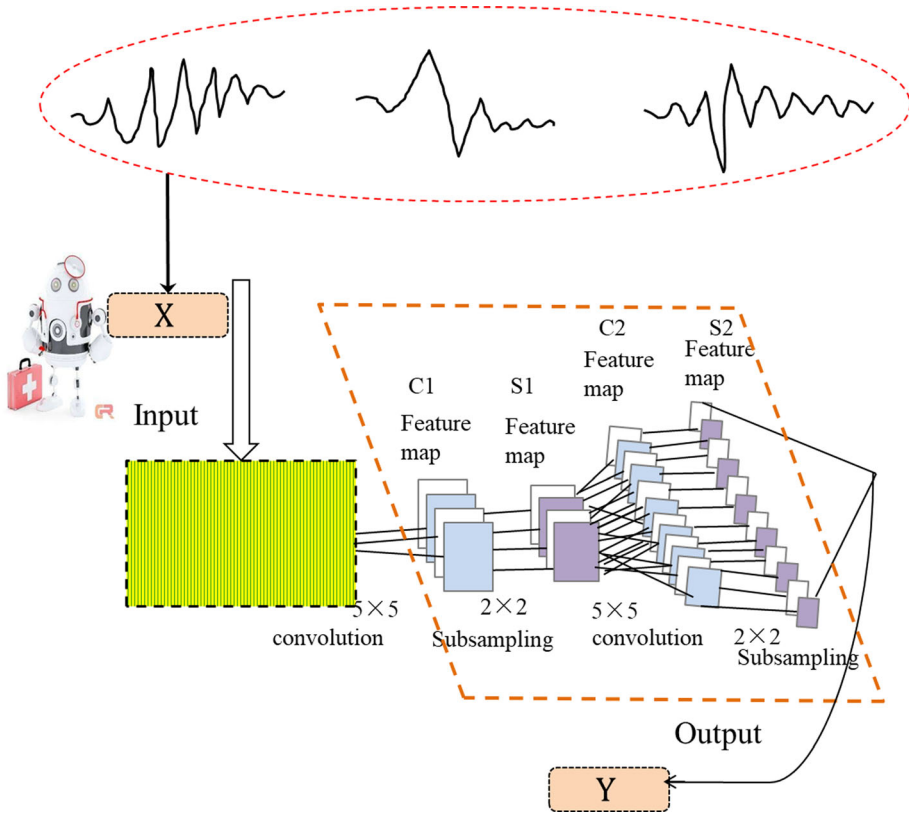

Fig. 8 Sample learning framework for early warning models of diabetes risk

3.3 Statistical method

Use Excel2016 to establish a database, IBM SPSS24.0 for statistical analysis, obey the normal distribution of measurement data, use the independent samples t test for comparison between groups, and non-normal distribution of measurement data with M (P25, P75), between groups. The comparison was performed using the rank sum test; the count data were used as the ratio, and the Chi-square test was used for comparison between groups; the rank data were expressed by the composition ratio, the rank sum test was used for comparison between groups, and the multivariate analysis of the influencing factors was performed using the unconditional logistic regression analysis. The test level is 0.05.

4 Results

4.1 General situation of the subject

Among the 5954 respondents, 5442 were males, accounting for 91.4%; the age range was 22–60 years, the median age was 46 (39, 50) years, the Han nationality was 5812,

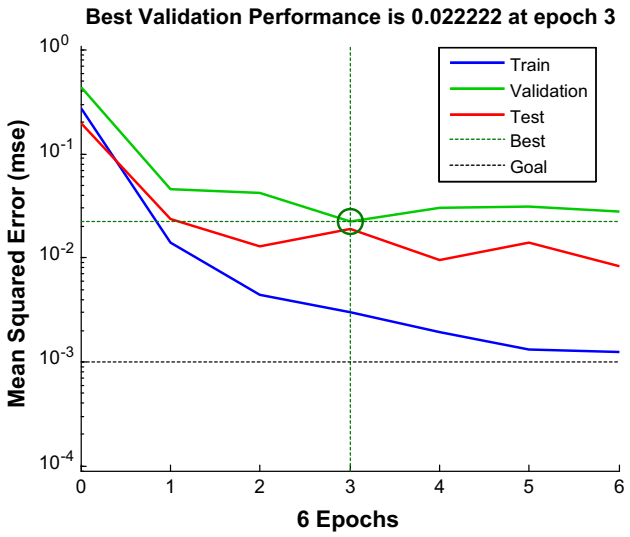

Fig. 9 Error diagram of algorithm learning effect

accounting for 97.6%, and the married persons were 5492, accounting for 92.2%. The average income is 5000 (4000, 7000) yuan, and the number of people in high school or secondary school (technical school) is 3020, accounting for

Table 8 Prediction results of training set of type 2 diabetes risk prediction model [*n* (%)]

Predictive value	Actual value		Total
	Yes	No	
Yes	463 (95.5)	207 (5.6)	670
No	22 (4.5)	3477 (94.4)	3499
Total	485	3684	4169

50.7%, as shown in Fig. 6; the prevalence rate of 703 people with type 2 diabetes is 11.8%, and that of the non-patient 5251, is 88.2%.

4.2 Independent variable screening

Univariate analysis of type 2 diabetes patients and non-patients in steel workers showed results in age, economic income, gender, education, marital status, physical activity, smoking status, drinking status, fruit intake, and family history of diabetes. There were statistical differences in hypertension, shift, BMI, waist-to-hip ratio, abdominal

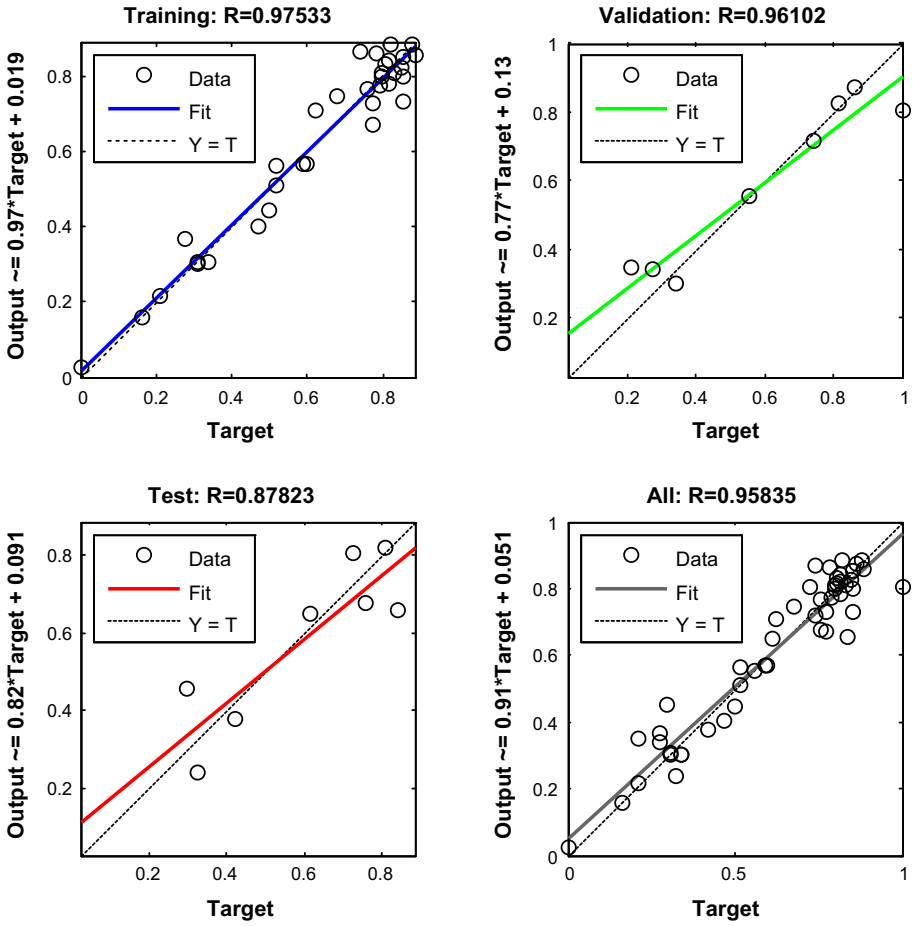

Fig. 10 Results of good fit of each data set model

Table 9 Type 2 diabetes risk prediction model validation set prediction results [*n* (%)]

Predictive value	Actual value		Total
	Yes	No	
Yes	129 (91.5)	95 (9.1)	224
No	12 (8.5)	948 (90.9)	960
Total	141	1043	1184

Table 10 Type 2 diabetes risk prediction model test set prediction results [*n* (%)]

Predictive value	Actual value		Total
	Yes	No	
Yes	69 (89.6)	58 (11.1)	127
No	8 (10.4)	466 (88.9)	474
Total	77	524	601

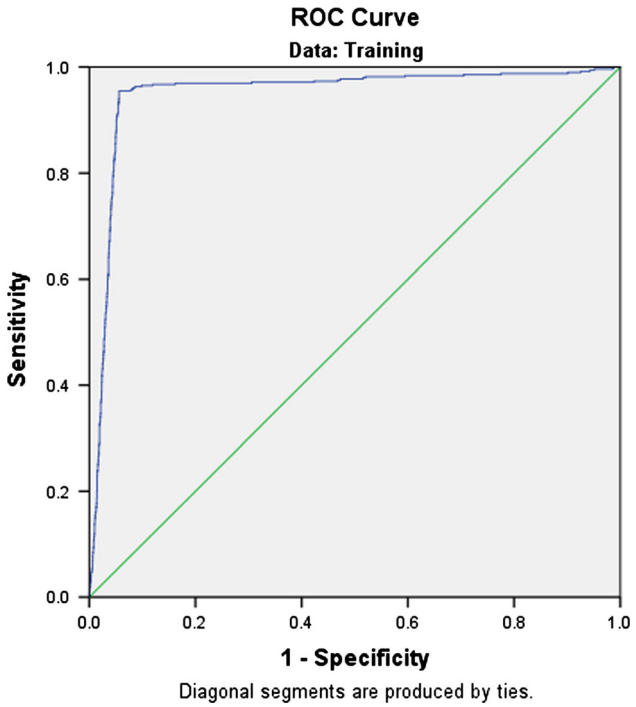

Fig. 11 ROC curves of CNN models in the training set

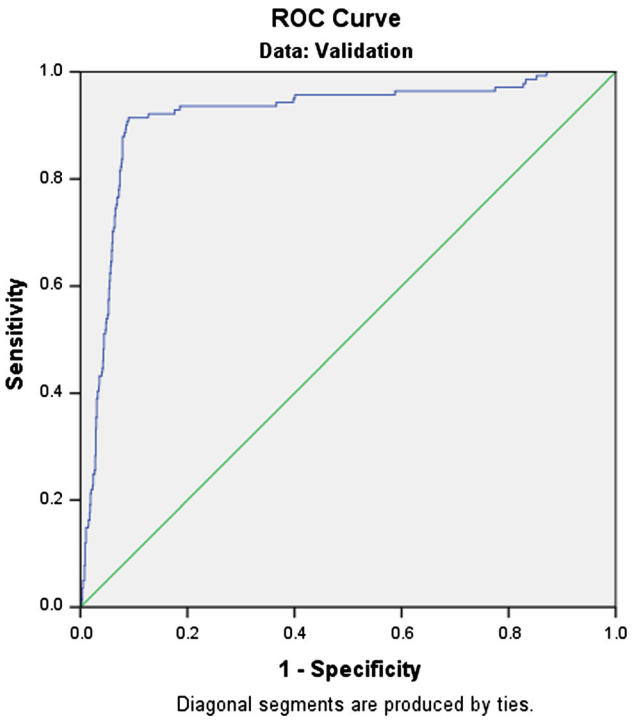

Fig. 12 ROC curves of CNN models in the validation set

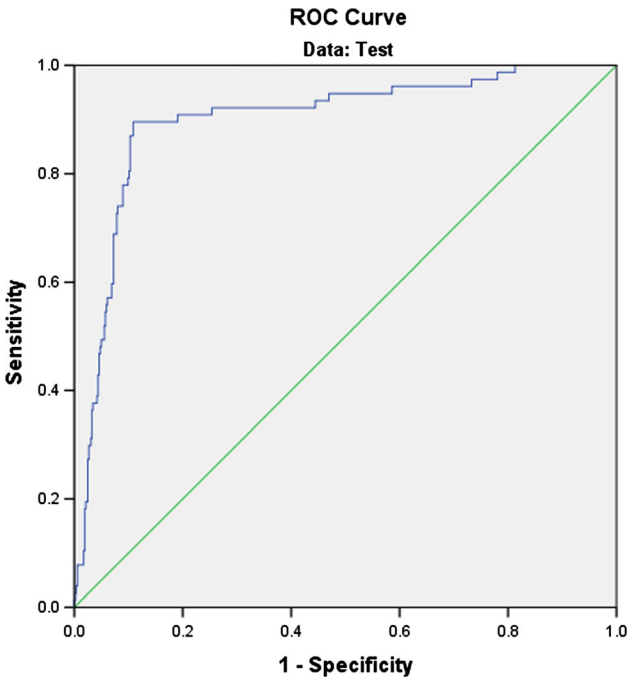

Fig. 13 ROC curves of CNN models in the test set

circumference, uric acid, white blood cell count, C-reactive protein, and dyslipidemia ($P < 0.05$), while the two groups were ethnic, passive smoking, vegetable intake. There were no significant differences in noise, high temperature and homocysteine levels ($P > 0.05$). The results are shown in Tables 2, 3, 4, and 5.

The statistically significant factors of univariate analysis were introduced into unconditional logistic regression for multivariate analysis. The results showed that the higher the age, the higher the risk of diabetes, and the risk of moderate and heavy physical activity was lower than that of light physical activity. The risk of fruit is reduced.

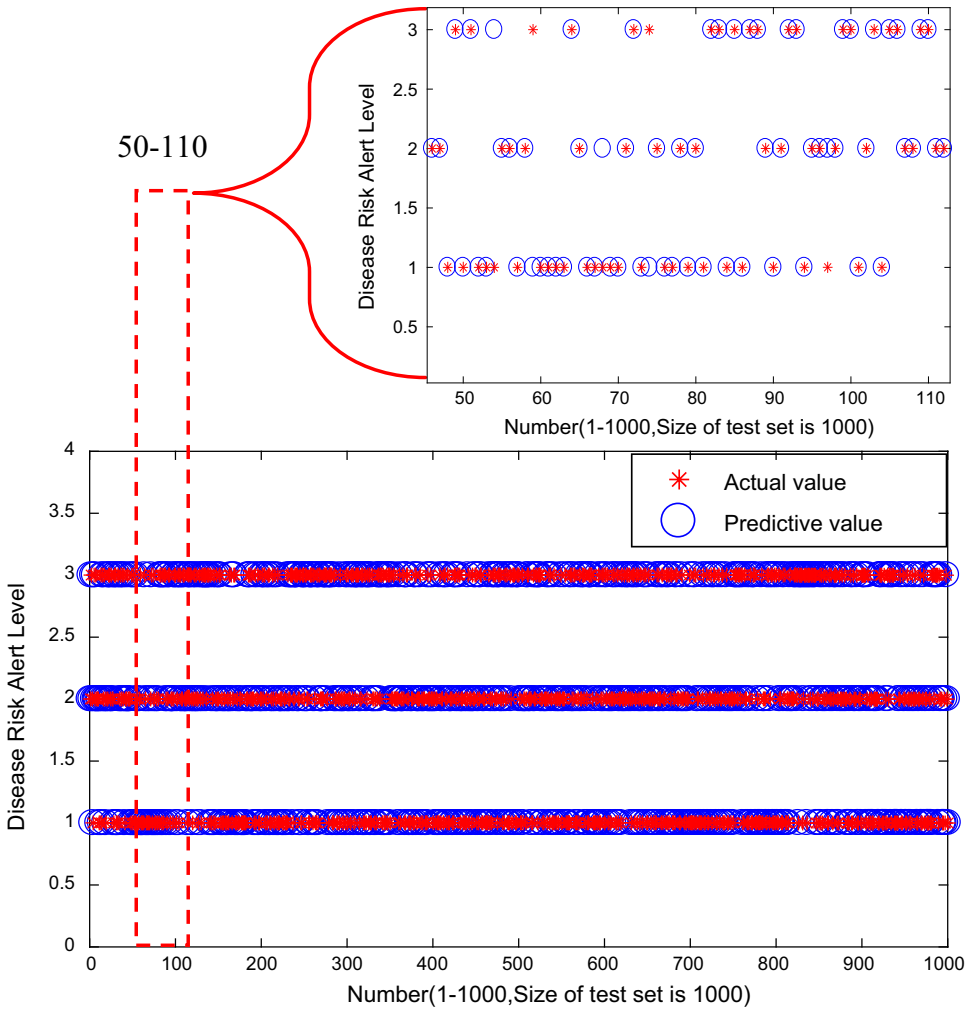

Fig. 14 Comparison of actual risk alert and forecast risk alert under sequence

People with diabetes and family history of hypertension are more likely to develop type 2 diabetes. High BMI, large abdominal circumference, high white blood cell count, and dyslipidemia are risk factors for type 2 diabetes, while uric acid is type 2. The protective factors of diabetes are shown in Table 6.

4.3 Construction of the sample set

Logistic regression analysis was used as the independent variable of the CNN model, and the type 2 diabetes risk prediction model was established based on whether type 2 diabetes occurred. The variable assignment table is shown in Table 7, and the sample set structure is shown in Fig. 7.

4.4 Model effect

All the data were randomly divided into training set, verification set and test set according to the ratio of 7:2:1. The data learning was carried out according to the sample learning framework shown in Fig. 8. The error result of the model is shown in Fig. 9. After three steps of training, the test set error is minimized.

The goodness of fit of the type 2 diabetes risk prediction model in the training set, validation set and test set was 0.975, 0.961, and 0.878, respectively, indicating that the model fitting ability is very good. The selected CNN is suitable for the risk prediction of type 2 diabetes in steel workers. The results are shown in Fig. 10.

The CNN model predicts 4169 training set samples, 1184 proof set samples and 601 test set samples. The accuracy rates are 94.5%, 91.0%, and 89.0%, respectively. The training set sensitivity is 95.5%, and the specificity is 94.4%; the area under the ROC curve (AUC) is 0.950 (95 CI 0.938–0.962); the sensitivity of the verification set is 91.5%, the specificity is 90.9%, and the area under the ROC curve (AUC) is 0.916 (95 CI 0.888–0.945); the test set sensitivity was 89.6%, the specificity was 88.9%, and the area under the ROC curve (AUC) was 0.899 (95 CI 0.859–0.939). The results are shown in Tables 8, 9, 10 and Figs. 11, 12, 13.

The congested depth neural network algorithm is used to obtain the diabetes risk alert prediction model. The empirical results are shown in the following figure. The figure shows that the accuracy of the police forecast is high, and the test results are provided for the accurate treatment of related patients. It is further shown that the convolution depth neural network algorithm is an algorithm suitable for complex and variable case data learning, and it is worthy of promotion in related fields (Fig. 14).

5 Conclusion

Using the convolution neural network to establish the risk prediction model of the steel worker T2DM, the fitting advantages of the model in the training set, the verification set and the test set were 0.975, 0.961, and 0.878, respectively, indicating that the selected model is suitable for the steel worker's medical examination data; the prediction accuracy of the model in the three data sets was 94.5%, 91.0%, and 89.0%, respectively, and the AUC was 0.950 (95 CI 0.938–0.962), 0.916 (95 CI 0.888–0.945), and 0.899 (95 CI 0.859–0.939). It shows that the established model can accurately predict the risk of type 2 diabetes in steel workers, which can provide a basis for steel workers' self-health management, and also facilitate the rational allocation of medical and health resources and the development of health services, and provide evidence for government departments to make decisions.

Acknowledgements This work was supported by the National Key R&D Program of China (No. 2016YFC0900605).

Compliance with ethical standards

Conflict of interest The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Colman PG, Mcnair PD, Gellert S (2010) Development of autoantibodies to islet antigens during childhood: implications for preclinical type 1 diabetes screening. Pediatric Diabetes 3(3):144–148
2. Mi S-Q (2011) Establishing and validating of type II diabetes incident prediction model of Chinese adult at individual level. Ph.D. dissertation, Chronic center, Chinese Center for Disease Control and Prevention, Beijing, China
3. Hong Y (2016) Research on diabetes prediction models based on machine learning algorithm. M.S. thesis, School of Management, Harbin Institute of Technology, Harbin, China
4. Nichols GA, Brown JB (2008) Validating the framingham offspring study equations for predicting incident diabetes mellitus. Am J Managed Care 14(9):574–580
5. Stern Michael P (2002) Identification of persons at high risk for type 2 diabetes mellitus: Do we need the oral glucose tolerance test? Ann Intern Med 136(8):575–581
6. McNeely MJ, Boyko EJ, Leonetti DL, Kahn SE, Fujimoto WY (2003) Comparison of a clinical model, the oral glucose tolerance test, and fasting glucose for prediction of type 2 diabetes risk in Japanese Americans. Diabetes Care 26(3):758–763
7. Han Y, Li J, Yang XL, Liu WX, Zhang YZ (2018) Dynamic prediction research of silicon content in hot metal driven by big data in blast furnace smelting process under hadoop cloud platform. Complexity 2018:1–16
8. Liu FC, Liu YL, Jin DH, Jia XY, Wang TT (2018) Research on workshop-based positioning technology based on internet of things in big data background. Complexity 2018:1–11
9. Yang AM, Yang XL, Chang JC, Bai B, Kong FB, Ran QB (2018) Research on a fusion scheme of cellular network and wireless sensor networks for cyber physical social systems. IEEE Access 6(99):18786–18794
10. Hai-Yun Wu, Pan Ping, He Yao, Liu Xiao-Ling, Zeng Qiang, Wei Cheng-Jun, Huang Shi-Jian (2007) A risk model for prediction of diabetes in Chinese adults. Chin J Health Manag 1(2):95–98
11. Chien K, Cai T, Hsu H, Su T, Chang W, Chen M (2009) A prediction model for type 2 diabetes risk among Chinese people. Diabetologia 52(3):443–450
12. Siyouneh B, Hsuan-Ping L, Richard P, Melanie M, Martha S, Bradlee M (2018) Dietary cholesterol intake is not associated with risk of type 2 diabetes in the Framingham offspring study. Nutrients 10(6):1–9
13. Hippisley-Cox J, Coupland C (2010) Individualizing the risks of statins in men and women in England and Wales: population-based cohort study. Heart 96(12):939–947
14. Wu JH, Wei W, Zhang L (2019) Risk assessment of hypertension in steel workers based on LVQ and fisher-SVM deep excavation. IEEE Access 7(1):23109–23119
15. Zhang Y, Zhao CK, Li XL, He YP, Ren WW, Zou CP (2017) Virtual touch tissue imaging and quantification: value in malignancy prediction for complex cystic and solid breast lesions. Sci Rep 7(1):1–10
16. Oyedotun OK, Khashman A (2016) Document segmentation using textural features summarization and feedforward neural network. Appl Intell 45(1):1–15
17. Konomi M, Sacha GM (2017) Feedforward neural network methodology to characterize thin films by electrostatic force microscopy. Ultramicroscopy 182:243–248
18. Kalchbrenner N, Grefenstette E, Blunsom P (2014) A convolutional neural network for modelling sentences. Eprint Arxiv, 655–665
19. Jeon J, Paik J (2015) Single image super-resolution based on subpixel shifting model. Optik Int J Light Electron Optics 126(24):4954–4959
20. Herrera C, Nava FA (2009) Earthquake hazard assessment in seismogenic systems through Markovian artificial neural network estimation: an application to the Japan area. Earth Planets Space 61(11):1223–1232
21. Xu R, Chen T, Xia Y, Lu Q, Liu B, Wang X (2015) Word embedding composition for data imbalances in sentiment and emotion classification. Cognit Comput 7(2):226–240
22. He F, Zhang L (2018) Prediction model of end-point phosphorus content in BOF steelmaking process based on PCA and BP neural network. J Process Control 66:51–58
23. Feng Yu, Xiaozhong Xu (2014) A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network. Appl Energy 134:102–113
24. Adler J, Oktem O (2018) Learned Primal-dual Reconstruction. IEEE Trans Med Imaging 2018:1–10

25. Agostinelli F, Hoffman M, Sadowski P, Baldi P (2014) Learning activation functions to improve deep neural networks. Comput Sci 1–9. ArXiv preprint arXiv:1412.6830
26. Wang Y, Wang X, Song G, Liu H, Zhu D, Ding J (2016) Genetic connection between mud shale lithofacies and shale oil enrichment in Jiyang depression, Bohai bay basin. Pet Explor Dev 43(5):759–768
27. Yang AM, Li SS, Lin HL, Jin DH (2018) Edge extraction of mineralogical phase based on fractal theory. Chaos Solitions Fractals 117:215–221
28. Kim J, Kim J, Jang GJ, Lee M (2017) Fast learning method for convolutional neural networks using extreme learning machine and its application to lane detection. Neural Netw 87:109–121
29. World Health Organization (2006) Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. Geneva World Health Organization

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.