ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Risk assessment method for progression of metabolic dysfunction-associated steatosis liver disease based on multi-dimensional health data

Jing Xia[1,2,3], *Wei Wang*[4] *and Wenjing Fu*[1,2,*]

[1] School of Medicine, Hainan Vocational University of Science and Technology, Haikou 571126, Hainan, China
[2] School of Pharmacy, China Medical University, Shenyang 110122, Liaoning, China
[3] School of Basic Medical Sciences, Peking University, Beijing 100191, Beijing, China
[4] School of Basic Medicine, Xinjiang University, Urumqi 830017, Xinjiang, China

SUMMARY: *In response to the heavy burden of metabolic dysfunction-associated steatosis liver disease (NAFLD) in China, limited existing screening methods, unsuitability of foreign models for Chinese people, small sample size of domestic models, and insufficient predictive performance, the present work constructs a high-precision NAFLD progression risk assessment model for employed laborers and farmers. Firstly, based on the multi-source health check-up data of 68573 employed laborers, 14 key indicators including age, gender, blood pressure, blood glucose, blood lipids, and hepatic biochemical function were selected as features to construct a CNN-LSTM hybrid model fusion predictive framework. The performance of the model was comprehensively evaluated using AUC, classification accuracy, precision, recall, F1 value, and calibration curve. The experimental results show that the proposed CNN-LSTM hybrid model intelligent learning algorithm based NAFLD progression risk assessment model has an AUC index value of 0.9861 on the training dataset and 0.9336 on the validation dataset, with an classification accuracy index value of 0.8505, an classification accuracy index value of 0.8363, and an F1 index value of 0.8570. All indicators are superior to the comparison model, indicating that the proposed CNN-LSTM hybrid model has high classification accuracy, strong generalization and stability in predicting NAFLD progression risk, and can provide technical support for clinical precision intervention and population health management prevention and control.*

KEYWORDS: *metabolic dysfunction-associated steatosis liver disease; Multi source monitoring data; CNN; LSTM; Risk assessment; Intelligent learning algorithm; hepatic biochemical function indicators*

1 Introduction

Nonalcoholic fatty liver disease (NAFLD) can lead to impaired hepatic biochemical function, and then develop into liver late-stage liver fibrosis, hepatocellular carcinoma (HCC) and other end-stage liver diseases [1]. It is also closely related to metabolic syndrome, cardiovascular disease, type 2 diabetes and other chronic diseases. The global prevalence rate exceeds 34.2%, which is an important problem threatening population health management. A dataset of over 6.2 million adults from 30 provinces in China from 2020 to 2025 showed that after adjusting for gender and age, the prevalence of hepatic steatosis and severe hepatic steatosis were 45.18%

*ccaabb202@163.com
https://doi.org/10.65102/is20261246 Published: 10 June, 2026

and 11.09%, respectively. The prevalence of advanced liver fibrosis and late-stage liver fibrosis was 2.91% and 0.93%, respectively. China has a heavy burden of fatty liver disease. According to the 2022 Global Burden of Disease Study, as the main type of FLD, the prevalence of NAFLD related liver complications (LC-NAFLD) in China has reached 21.53%, resulting in 123800 disability adjusted life years (DALYs) [2]. Early diagnosis and intervention are of great significance for improving the prognosis of NAFLD patients and reducing medical burden. Diagnostic methods include ultrasound imaging, magnetic resonance imaging (MRI), and liver biopsy. Among them, the high cost, low accessibility, and radiation hazards of MRI testing limit its promotion and application. In addition, due to its invasiveness and potential risk of complications, the application of liver biopsy is even more limited. Although ultrasound imaging is the main method for diagnosing NAFLD in China, it is not suitable as a universal screening tool for NAFLD due to the insufficient allocation of medical resources and high expenditure in the country. Therefore, it is necessary to construct a model for predicting and evaluating the risk of fatty liver progression as a preliminary screening method. Famous models for predicting the risk of fatty liver progression in foreign countries include the Fatty Liver Index (FLI) and the United States Fatty Liver Index (USFLI), which are not suitable for Chinese people due to their race. The current domestic models for predicting the risk of fatty liver progression are mostly based on small samples, and their predictive performance is not yet ideal [3].

The application of intelligent learning algorithm algorithms in the medical field is becoming increasingly widespread, as they can automatically learn complex nonlinear relationships from massive amounts of data and construct high-precision predictive frameworks. In the field of disease prediction, classic statistical methods such as logistic regression have weak data utilization capabilities and difficulty in modeling nonlinear relationships. Compared with it, intelligent learning algorithm models can construct models with excellent predictive performance through automatic key indicator selection, high-dimensional data processing, and capturing nonlinear interaction effects between risk factors. In recent years, several models for predicting the risk of fatty liver progression have been constructed using intelligent learning algorithm algorithms in China. However, there are significant differences in the number of research samples, types and quantities of introduced predictive variables, optimal model algorithms, and predictive performance. There is still a lack of algorithm model construction supported by large sample data. The present work is based on multi-dimensional health data such as large sample physical examination, and uses convolutional neural networks (CNN) and long short-term memory networks (LSTM) to construct a risk assessment model for fatty liver progression. Combining current mainstream intelligent learning algorithm algorithms such as neural networks, decision trees, and ensemble algorithms, the optimal model is screened and deployed, aiming to provide a high-performance initial screening tool for fatty liver progression risk screening.

2 Related research

2.1 Current situation of fatty liver

Fatty liver disease (FLD) is a common pathological change in the liver, which can be divided into metabolic dysfunction-associated steatosis liver disease (NAFLD), alcoholic fatty liver disease (ANAFLD), and various other special types of fatty liver [4]. It is also a disease caused by excessive accumulation of fat in liver cells due to various reasons, and is closely related to genetic factors, environmental factors, and metabolic stress. If the lipid accumulation in the normal liver exceeds 5% of the wet weight of the liver, it is identified as fatty liver. The disease

is characterized by high incidence rate, high mortality and low control rate. In recent years, the prevalence of fatty liver is growing rapidly. Studies have shown that fatty liver has become the most common chronic liver disease in both developing and developed countries. Between 1981 and 1991, the prevalence of fatty liver in developed countries such as Europe and America were only 12%, but after 2004, the prevalence increased to 21%. Now, the incidence rate of fatty liver among adults in Britain and other developed European countries has reached 47%, higher than 25% in Japan. Nonalcoholic fatty liver disease is an increasingly serious cause of chronic liver injury, especially in Western countries where it has become the most common indication for liver transplantation. Previous studies have shown that the prevalence of metabolic dysfunction-associated steatosis liver disease (NAFLD) in Asia may be as common as in developed regions such as Europe and America, with a prevalence of>25% in most Asian countries. In recent years, research has shown that there are nearly 260 million patients with fatty liver in China, and it is expected that the domestic fatty liver incidence will continue to increase by 52% from 2018 to 2032. With the rising prevalence of fatty liver worldwide, it will cause a heavy economic burden on families and society [5].

Simple fatty liver can worsen and develop into metabolic dysfunction-associated steatosis liver disease, liver fibrosis, and even severe liver diseases such as Late stage liver fibrosis and hepatocellular carcinoma (HCC); In addition, it can also induce and aggravate cardiovascular and cerebrovascular diseases, which is closely related to the incidence rate of diabetes and chronic kidney disease and the impairment of immune capacity of the body, and poses a certain threat to human health and social development. However, due to the insidious and slow progression of early-stage fatty liver disease, most people have unclear clinical symptoms and unrestricted daily life. Currently, most people still lack sufficient understanding of the potential harm and prevention strategies of fatty liver [6]. However, as the early progression of fatty liver is a reversible disease, it is important to identify potential patients early and avoid exposure to high-risk factors to reduce the incidence rate of fatty liver progression. Liver biopsy is the "gold standard" for diagnosing the progression of fatty liver, but it is an invasive examination that has disadvantages such as limited sampling, large errors, and inability to intuitively reflect fat infiltration and liver cell damage. Although B-ultrasound has high classification accuracy as a functional tool for diagnosing the progression of fatty liver, its classification accuracy is highly dependent on subjective judgment.

2.2 Intelligent learning algorithm Modeling

With the continuous development of big data and artificial intelligence technologies, intelligent learning algorithm (ML) has been widely applied in many different fields as a representative and has shown great potential in assisting clinical diagnosis [7]. For example, intelligent learning algorithm can conduct deep mining and analysis of big data, and has been applied in disease prediction such as liver, diabetes and tumor, with good application potential. A good predictive model can accurately predict the progression of fatty liver progression, thereby effectively monitoring and intervening in high-risk populations in a timely manner. The present work mainly uses Convolutional Neural Networks (CNN) and Long Short Memory Networks (LSTM) to construct a risk assessment model for fatty liver progression, and combines current mainstream intelligent learning algorithm algorithms such as neural networks, decision trees, and ensemble algorithms to verify and screen the optimal fatty liver progression risk assessment model and deploy the model. The specific description is as follows [8]: (1) KNN is a non-parametric method first proposed by Fix and Hodges, which does not require a specific training stage and is therefore known as lazy learning algorithm. For a given labeled sample dataset, when new data without K labels is input, the similarity between the new input data and the data in the sample is compared, and the K data with the highest similarity is selected. The majority

voting result of the K most similar data is used as the category label of the current new data. In the KNN algorithm, the smaller the distance, the more similar the degree between two samples. (2) XGBoost is a well-known integrated machine algorithm based on gradient enhanced decision tree framework, which is more efficient in terms of speed and performance compared to most machine methods. XGBoost reduces the complexity of the model by adding a regularization term to the objective function to prevent overfitting. (3) SVM is a binary classification model that uses supervised learning to classify data. The basic idea is to find a hyperplane in the sample space based on the training dataset to divide samples of different classes, so that the optimal interval of the hyperplane is maximized and the minimum classification error rate is ensured. SVM has no prior assumptions and does not involve probability measures, which can effectively avoid the traditional process between induction and deduction [9]. It can achieve efficient "transducing inference" from training samples to prediction set samples, greatly simplifying commonly used classification and regression related problems. (4) The logistic regression algorithm is used in intelligent learning algorithm methods to handle binary classification problems. Its principle is to add the regression coefficients of all variable feature values themselves, and then input the sum into the sigmoid function to obtain a value from 0 to 1. Then, based on the threshold, the classifier is used to convert it to 0 or 1. (5) The basic idea of Naive Bayes (NB) method is to assume that the feature conditions are independent of each other. For a given item to be classified, the probability of each category under these conditions is solved. The highest probability can be used to determine which category the item belongs to. NB is suitable for scenarios where features are independent of each other, and the connotation of "simplicity" can be understood as the ability of features to be independent of each other.

2.3 Risk assessment model for fatty liver progression

There is relevant literature both domestically and internationally that have developed predictive models for the risk of worsening metabolic dysfunction-associated steatosis liver disease [10]. As Eren et al. [11] used basic information to construct a prediction of the risk of metabolic dysfunction-associated steatosis liver disease progression, the AUC was only 0.744. De Vincentis et al. [12] established a risk assessment model for fatty liver progression based on health checkups, and compared three intelligent learning algorithm methods with traditional logistic regression. The results showed that the random forest predictive framework had the best performance in predicting the risk of fatty liver progression. Tamaki et al. [13] evaluated the performance of six intelligent learning algorithm algorithms in predicting the risk of fatty liver progression, and the results showed that XGBoost algorithm (classification accuracy of 0.873) had more advantages than logistic regression (classification accuracy of 0.769). Unlike the above, Roh et al. [14] used decision tree and logistic models to construct a NAFLD disease risk assessment model, and the results showed that traditional logistic regression was significantly better than decision tree models. In the research on the risk assessment model of fatty liver progression, the optimal predictive framework algorithm is not consistent. The focus of risk assessment is on improving the classification accuracy of the model, while there is less attention paid to the interpretation of the model. This hinders the practical application of intelligent learning algorithm in clinical practice and makes it difficult to promote the model [15]. In the early stage of reform and opening up, employed laborers in China became an important part of construction workers and a great contributor to the socialist country. However, employed laborers have a high risk of illness and low health literacy due to irregular diet, which requires special attention. Chronic metabolic related diseases, mainly characterized by the progression of fatty liver, seriously threaten the health level of employed laborers. Early

identification of high-risk factors is a prerequisite for effective prevention. At present, there is a lack of relevant research on the risk assessment model for the progression of fatty liver among employed laborers in foreign trade both domestically and internationally. Therefore, the present work constructs a risk assessment model for fatty liver progression among employed laborers based on six intelligent learning algorithm algorithms, providing a basis for early prevention of fatty liver progression among employed laborers [16].

3 Research subjects and methods

3.1 Research Object

Statistical analysis was conducted using the 2024 health check-up database of employed laborers in a certain city in China. The examination items included gender, age, smoking history (current smoking status), alcohol consumption history (current alcohol consumption status), height, weight, systolic blood pressure (SBP), diastolic blood pressure (DBP), fasting blood glucose (FPG), alanine aminotransferase (ALT), aspartate aminotransferase (AST), triglycerides (TG), cholesterol (TC), low-density lipoprotein cholesterol ester (LDL-C), high-density lipoprotein cholesterol ester (HDL-C), liver ultrasound imaging examination, etc [17]. The data was cleaned, and those with missing key items, incomplete input, or suspicious values were excluded. At the same time, patients with other liver diseases were excluded from liver ultrasound examination. Finally, a total of 68573 people were enrolled in the present work.

3.2 Research Methods

Calculate body mass index (BMI, kg/m^2) =weight/height ^{2}. Use age, gender, alcohol consumption history, smoking history, SBP, DBP, LDL-C, HDL-C, FPG, BMI, TG, TC, ALT, AST as predictive variables, and non-fatty liver progression/fatty liver progression as target variables to establish a CNN-LSTM hybrid model for predicting the risk of non-alcoholic fatty liver progression. Combined with various intelligent learning algorithm models constructed (KNN, decision tree, Bayesian network, support vector machine, random forest, neural network, logistic regression, etc.), evaluate the predictive performance of each model for the risk of non-alcoholic fatty liver progression to screen for models with excellent performance. The evaluation indicators include F1 score (F1 Score), area under the ROC curve (AUC), calibration curve, precision, recall, receiver operating characteristic (ROC), decision curve analysis (DCA), and the optimal model is selected for research deployment [18].

3.3 Distribution of Progression of Fatty Liver

There is a total of 68573 people in the sample, including 36764 males, accounting for 53.61%; 31809 women, accounting for 46.39%. The youngest age is 19 years old, the oldest is 86 years old, and the median (interquartile range) is 38.74 (11.68) years old. There were 18539 patients with worsening alcoholic fatty liver, accounting for 27.04%, and 50034 patients with worsening non-alcoholic fatty liver, accounting for 72.96%. The distribution of relevant indicators for different populations with worsening fatty liver is shown in Table 1.

Table 1: Distribution of relevant indicators in the population with worsening alcoholic fatty liver disease/metabolic dysfunction-associated steatosis liver disease

Related indicators		Alcoholic fatty liver (percentage)	Nonalcoholic fatty liver (percentage)	x^2/Z/t value	P value
Gender	Man	22349(60.79%)	14415(39.21%)	5023.65	<0.001
	Woman	26213(82.41%)	5596(17.59%)		
Smoking history	Smoking	10368(61.15%)	6586(38.85%)	1308.42	<0.001
	No smoking	38221(74.04%)	13398(25.96%)		
Drinking history	Drinking Wine	15108(60.08%)	10037(39.92%)	2924.75	<0.001
	Not drinking alcohol	33658(77.50%)	9770(22.50%)		
Age (years)		38.65(11.84)	43.82(15.37)	27.17	<0.001
BMI(kg/m²)		23.24±2.79	28.12±2.86	197.12	<0.001
DBP(mmHg)		78.32±11.37	88.73±12.16	101.69	<0.001
SBP(mmHg)		119.65±17.60	136.15±19.43	100.53	<0.001
FPG(mmol/L)		5.12±0.83	5.73±1.65	62.36	<0.001
TG(mmol/L)		1.18±0.86	2.32±1.75	88.81	<0.001
TC(mmol/L)		4.48±0.75	5.08±1.13	66.53	<0.001
HDL-C(mmol/L)		1.36±0.31	1.17±0.25	49.49	<0.001
LDL-C(mmol/L)		2.57±0.72	2.81±0.76	38.12	<0.001
ALT(U/L)		19.08±14.83	33.15±22.38	79.58	<0.001
AST(U/L)		23.11±9.40	28.12±13.19	46.48	<0.001

Divide the samples into a training dataset (47361) and a testing set (20299) in a 7:3 ratio. The composition ratios of metabolic dysfunction-associated steatosis liver disease progression/metabolic dysfunction-associated steatosis liver disease progression in the two regions are 71%:29% and 73%:27%, respectively, x^2=1.08, P=0.27. In univariate analysis, 14 variables including gender (categorical variable, 0=male, 1=female), smoking history (categorical variable, 0=non-smoking, 1=smoking), alcohol consumption history (categorical variable, 0=non drinking, 1=drinking), age, TC, TG, SBP, BMI, LDL-C, HDL-C, DBP, FPG, ALT, AST (all continuous variables) all showed P<0.001. The 14 variables were defined as "input" roles, and metabolic dysfunction-associated steatosis liver disease progression (0=metabolic dysfunction-associated steatosis liver disease progression, 1=metabolic dysfunction-associated steatosis liver disease progression) was used as the "target" role. The "feature selection" in modeling was applied to predict the 14 inputs. The importance of variables ("important", "marginal", "unimportant") was further screened for predictive variables, and all 14 variables showed "important".

4 CNN-LSTM hybrid model intelligent learning algorithm model

4.1 CNN Learning Model

Convolutional Neural Network (CNN) is a special type of multi-layer perceptron that belongs to the feedforward neural network architecture. Its standard structure typically includes an input

layer, alternately stacked convolution layers and pooling layers, fully connected layers, and output layers, as shown in Figure 1.

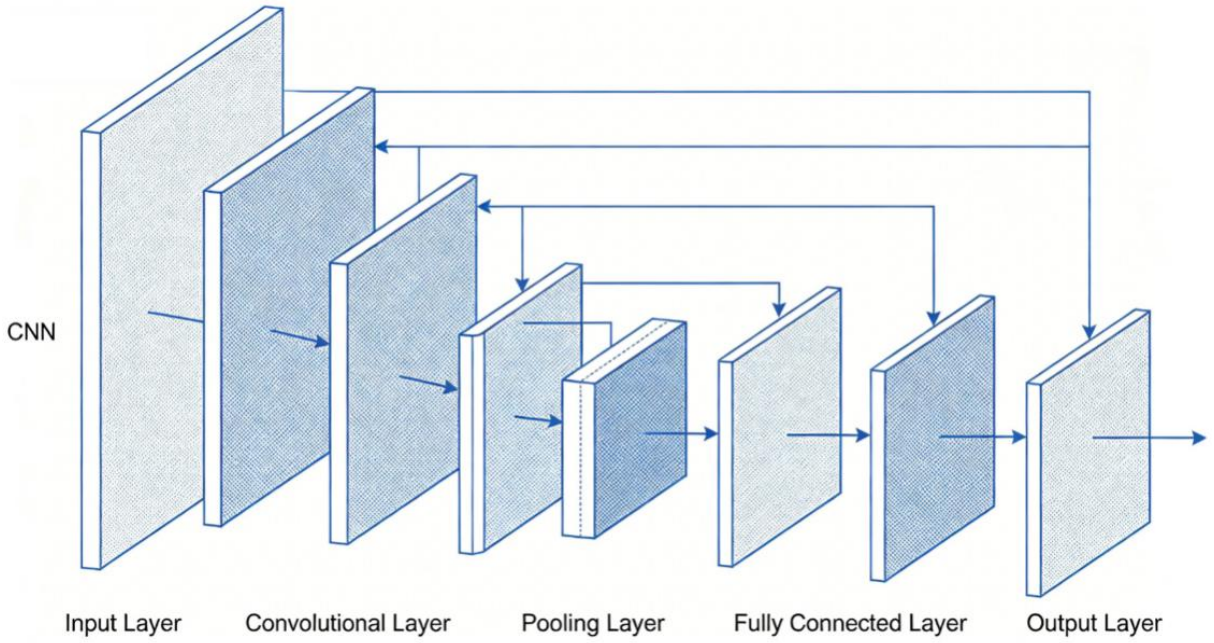


Figure 1: CNN Structure Diagram

The core advantage of CNN lies in its ability to efficiently process high-dimensional input data (such as images and time series data segments) through convolution and pooling operations. While significantly reducing computational complexity, it effectively extracts key features of the input data, laying the foundation for subsequent recognition and prediction tasks. Specifically:

Convolutional layer: responsible for extracting local spatial (or temporal) features of input variables, scanning the input data through sliding convolution kernels, and capturing local patterns in the data.

Pooling layer: The main function is to achieve data dimensionality reduction by significantly reducing the parameter size of feature maps through operations such as MaxPooling and AveragePooling, enhancing the spatial invariance and computational efficiency of the model.

Fully connected layer: receives and integrates feature information processed by convolutional and pooling layers, maps these features to the final output space through nonlinear transformation, and generates the expected prediction results.

Given the outstanding performance of CNN in key indicator selection, the present work uses CNN to perform preliminary key indicator selection on multi-source detection data of fatty liver, providing a reliable basis for subsequent time series modeling and control strategy design.

4.2 Long Short Term Memory Network (LSTM)

Long Short Term Memory (LSTM) was first proposed by Hochreiter and Schmidhuber, and is an important variant of Recurrent Neural Network (RNN). The core design goal is to solve the common problem of vanishing gradient or exploding gradient in traditional RNNs when processing long sequence data, in order to accurately and effectively capture long-distance dependencies in sequence data.

Recurrent neural networks (RNNs) are naturally suitable for processing sequential data

(such as time series, speech, text, etc.) due to their inherent cyclic connectivity structure. Its uniqueness lies in the introduction of a state transfer mechanism in the time dimension, allowing information to flow between different time steps and utilizing historical temporal information to assist current decisions. However, standard RNNs face many challenges when dealing with actual long sequence tasks:

(1) Gradient problem: When updating parameters through Back Propagation Through Time (BPTT) algorithm, the gradient values are repeatedly multiplied in the time dimension, which can easily lead to rapid decay of the gradient values to near zero (gradient disappearance) or rapid growth to infinity (gradient explosion), seriously weakening the model's ability to learn long-term dependencies.

(2) Computational burden: As the sequence length increases, RNNs need to store and process state information for all time steps, resulting in a significant increase in computational and memory requirements.

To effectively overcome the above-mentioned shortcomings of standard RNNs, the present work selected LSTM network to accurately model the temporal characteristics of multi-source detection data of fatty liver, and utilized its powerful long-term memory ability to capture the inherent laws of system state changes.

The key innovation of LSTM networks lies in the introduction of a sophisticated "gating" mechanism, mainly composed of forget gates, input gates, and output gates. These gate structures adaptively control the flow of information in the cell state through learnable parameters: retaining important information, forgetting residual information, and adding new information, thereby achieving efficient modeling of long-term dependencies. The basic structure of the LSTM unit is shown in Figure 2.

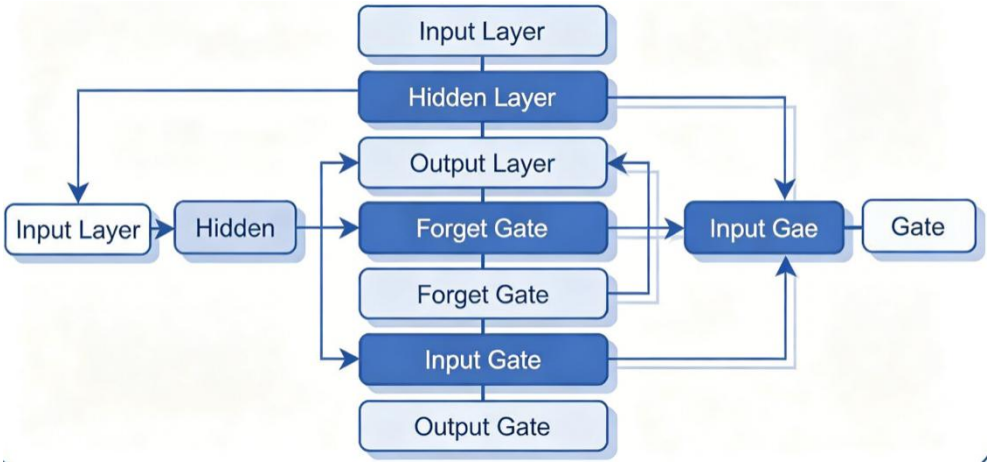


Figure 2: LSTM Structure Diagram

The forget gate determines the forgetting ratio of cell state information in the previous time step based on the hidden state of the previous time step and the current input data; The input gate is responsible for incorporating the weights of new information into the cell state matrix of the previous time step, and generating new candidate cell states based on the hidden state of the previous time step and the current input data; By the synergistic effect of the forget gate and input gate, the cell state at the current time step is updated. The various gate control update mechanisms are shown in equation (1):

$$\begin{cases} f_t = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \\ i_t = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \\ \tilde{C}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \\ C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ O_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \\ h_t = o_t \odot \tanh\left(C_t\right) \end{cases} \tag{1}$$

where, b is the bias generated during network training, W is the weight matrix, $\sigma(\cdot)$ and $\tanh(\cdot)$ are the activation functions; W_f, W_i, W_c, W_o is the weight matrix corresponding to gate control and state calculation; b_f, b_i, b_c, b_o is the corresponding bias vector; f_t, i_t, o_t represents the activation vectors (range [0,1]) of the forget gate, input gate, and output gate at time step t; C_{t-1}, C_t represents the cell state of the previous time step and the current time step respectively (long-term memory carrier). Mainly updating the unit state by combining the results of input and output gates; h_{t-1}, h_t represents the hidden states (short-term memory/unit output) of the previous time step and the current time step, respectively; $\sigma(\cdot)$ stands for Sigmoid activation function, with an output range of (0,1), used for gating signals; The o_t output gate determines the hidden state of the output terminal, $\tanh(\cdot)$ represents the hyperbolic tangent activation function, and the output range is (-1,1), which is used for candidate state calculation and output scaling; $\tilde{C}_t$ represents the candidate cell state for the current time step.

With the significant advantage of LSTM network in modeling long-term temporal dependencies, the present work inputs the spatial features extracted by the CNN key indicator selection module into the LSTM network, enabling the fusion model to have both powerful spatial key indicator selection and long-term dynamic sequence modeling capabilities, ultimately forming an intelligent model suitable for fatty liver control.

4.3 Algorithm Architecture Design

The CNN-LSTM hybrid model fusion model constructed in the present work consists of two core modules (as shown in Figure 3), which work together to process time-series data of fatty liver:

(1) CNN key indicator selection module: As a front-end feature extractor, it includes several alternately stacked convolutional layers and pooling layers. Convolutional layers are responsible for capturing local feature patterns (such as short-term fluctuations and specific morphological features) in input time series segments (which can be viewed as one-dimensional signals); The pooling layer follows closely behind for feature selection and dimensionality reduction, preserving key information and reducing subsequent computational complexity.

(2) LSTM temporal modeling module: As a backend temporal modeler, it receives low dimensional feature sequences output by the CNN key indicator selection module. The LSTM layer utilizes its internal gating mechanism and cell state to learn the long-term dynamic evolution patterns and state dependencies contained in the feature sequence. The overall process is as follows: the input multidimensional time series data is first subjected to local spatial key indicator selection through the CNN key indicator selection module, and the obtained feature sequence is then input into the LSTM temporal modeling module for temporal dependency modeling. Finally, the output of LSTM is integrated and mapped through one or more Fully Connected Layers to generate the multi-dimensional prediction output of the model.

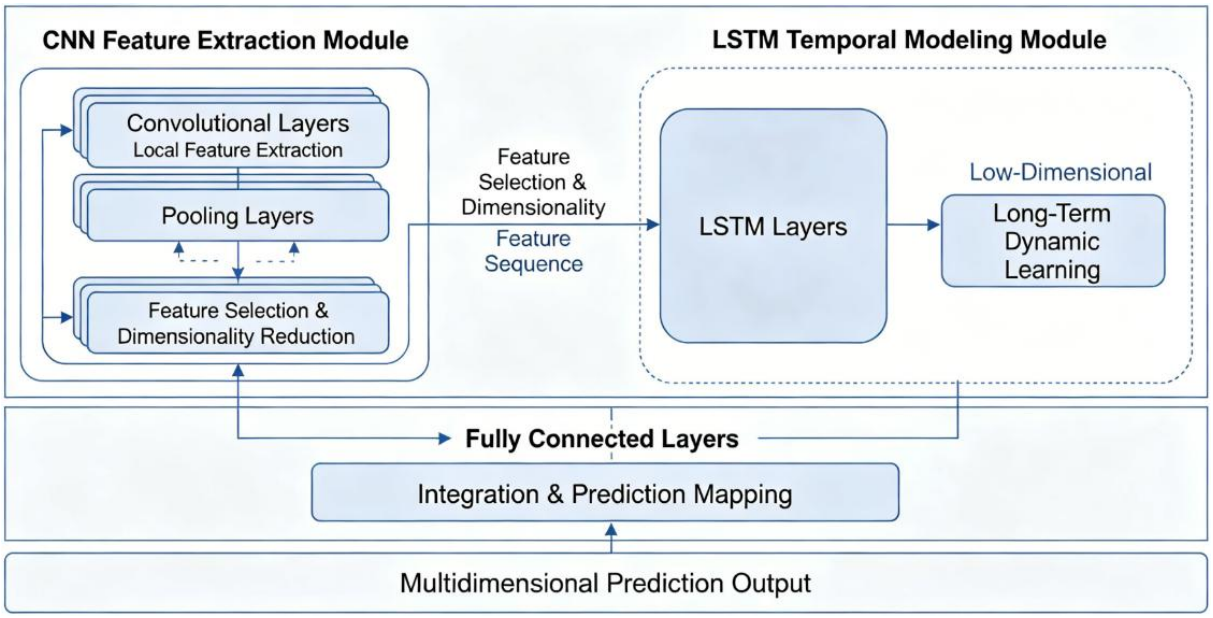


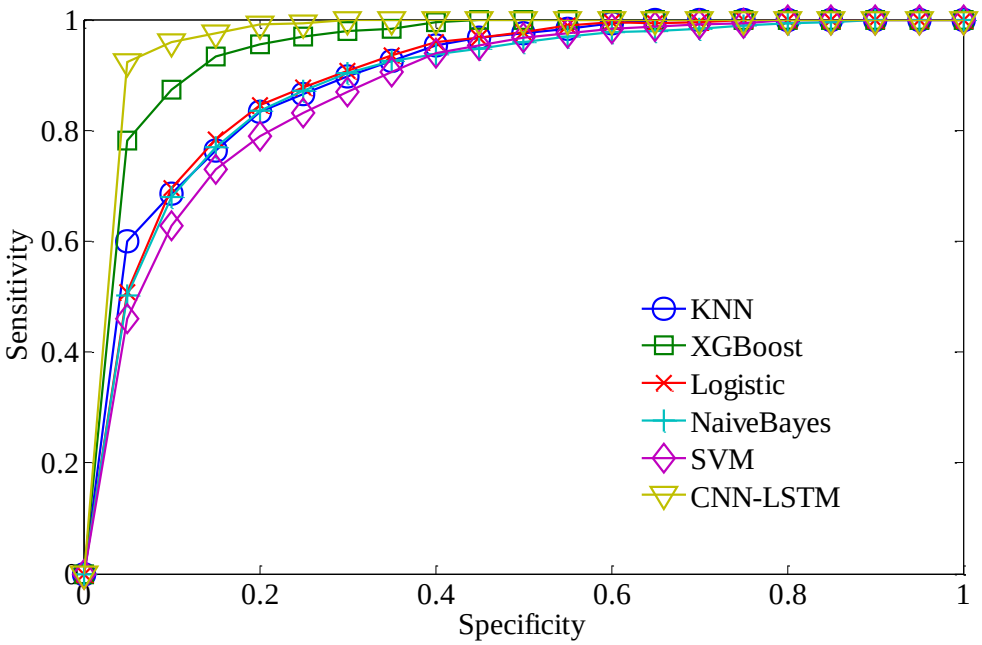
Figure 3: CNN-LSTM hybrid model Fusion Model

5 Experimental results

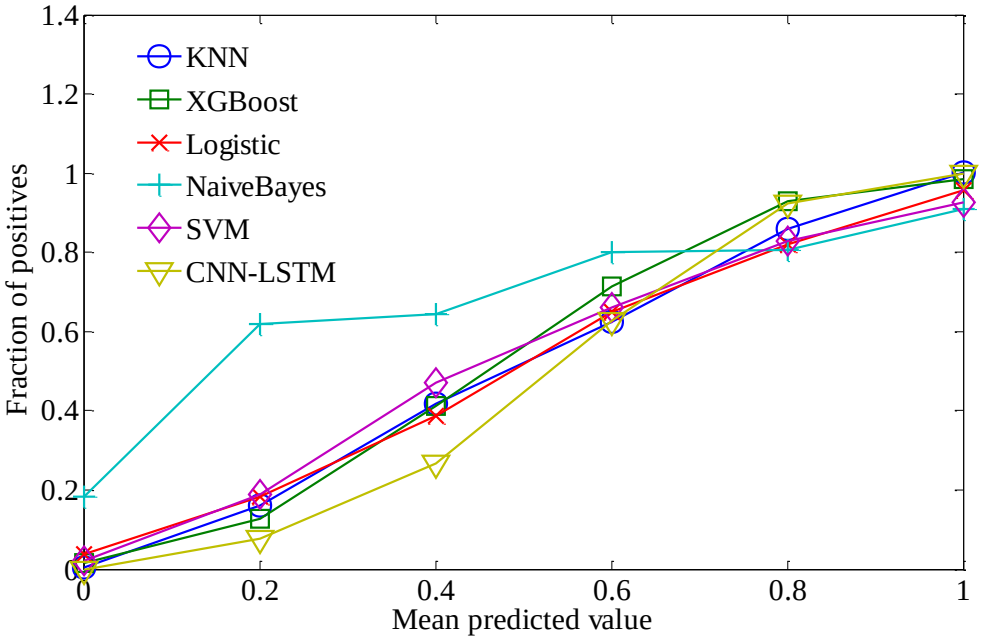
Table 2 shows the experimental results of performance evaluation indicators based on the validation dataset for each model. Figure 4 and Figure 5 respectively compare the predictive performance of intelligent learning algorithm models in the training and testing sets.

Table 2: Performance evaluation indicators of each model based on the validation dataset

Dataset splitting	Model	AUC	Classification accuracy	Precision	Recall	F1 value
Training dataset	CNN-LSTM hybrid model	0.9861	0.9343	0.9317	0.9428	0.9375
	XGBoost	0.9635	0.8942	0.8865	0.9117	0.8982
	KNN	0.9043	0.8176	0.8234	0.8223	0.8235
	logistic	0.8986	0.8220	0.8260	0.8289	0.8276
	NaiveBayes	0.8865	0.7875	0.8804	0.6810	0.7658
	SVM	0.8808	0.7903	0.8226	0.7573	0.7884
Validation dataset	CNN-LSTM hybrid model	0.9336	0.8505	0.8363	0.8809	0.8570
	XGBoost	0.9327	0.8474	0.8318	0.8835	0.8563
	KNN	0.8018	0.7518	0.7543	0.7676	0.7601
	logistic	0.8946	0.8136	0.8075	0.8412	0.8210
	NaiveBayes	0.8953	0.7950	0.8710	0.7065	0.7808
	SVM	0.8847	0.8102	0.8276	0.7968	0.8113

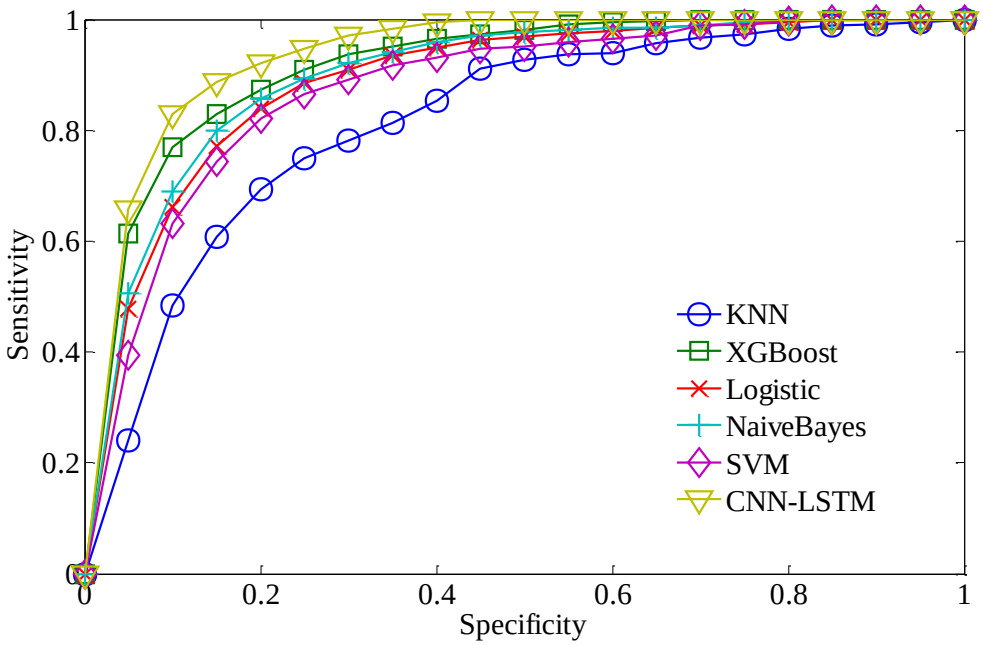


(a) ROC curve evaluates the predictive performance of each model

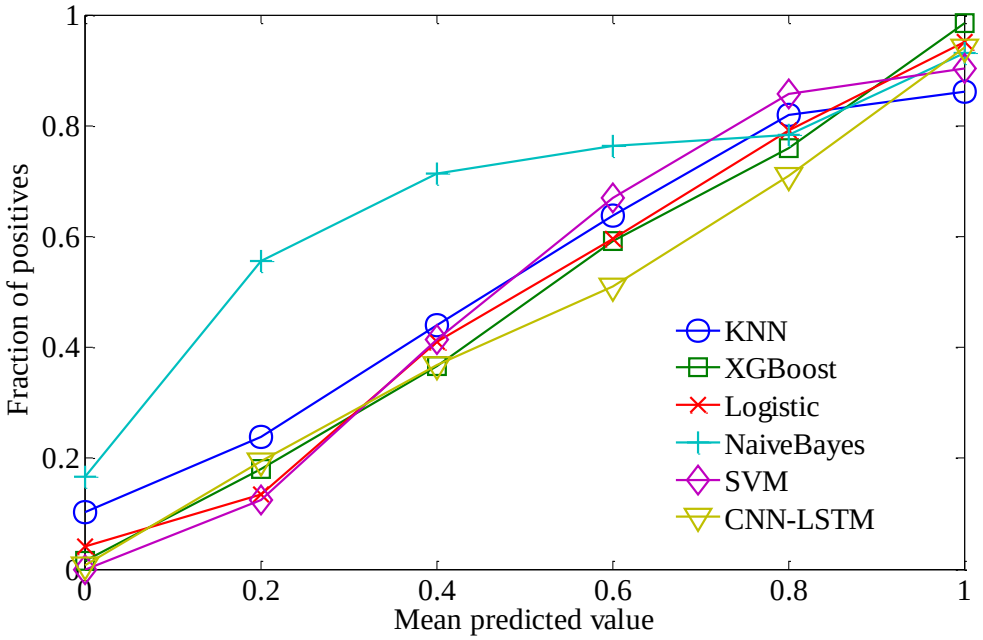


(b) Calibration curve analysis evaluates the predictive performance of each model

Figure 4: Comparison of predictive performance of intelligent learning algorithm models in the training dataset



(a) ROC curve evaluates the predictive performance of each model



(b) Calibration curve analysis evaluates the predictive performance of each model

Figure 5: Comparison of predictive performance of intelligent learning algorithm models in the validation dataset

According to Table 2 and the experimental results in Figure 4 and Figure 5, it can be seen that in the training dataset, the AUC index of the CNN-LSTM hybrid model metabolic dysfunction-associated steatosis liver disease progression predictive framework reached 0.9861, the classification accuracy index reached 0.9343, the classification accuracy index reached 0.9317, the recall index reached 0.9428, and the F1 value index reached 0.9375, which were

significantly higher than models such as XGBoost, KNN, and logistic regression. On the validation dataset, the AUC index of the CNN-LSTM hybrid model metabolic dysfunction-associated steatosis liver disease progression predictive framework reached 0.9336, the classification accuracy index reached 0.8505, the classification accuracy index reached 0.8363, the recall index reached 0.8809, and the F1 value index reached 0.8570, still leading all comparison algorithms. For the XGBoost metabolic dysfunction-associated steatosis liver disease progression predictive framework, the AUC index on the validation dataset reached 0.9327, and all indicators were close to CNN-LSTM hybrid model, ranking second in overall performance. The AUC index of the logistic regression validation dataset using traditional statistical methods is only 0.8946, and the classification accuracy and F1 value are lower than those of ensemble models. The KNN model performed the worst, with an AUC index of only 0.8018 on the validation dataset, and all indicators were significantly lower. The overall performance of SVM and Naive Bayes models is moderate, with a recall index of only 0.7065 for the Naive Bayes algorithm model, indicating significant shortcomings. The ROC curve and calibration curve were further validated, and the curve of CNN-LSTM hybrid model was closest to the upper left corner, with the best calibration degree and the best consistency between predicted probability and actual observed values. It had the best prediction classification accuracy, generalization ability, and stability for the risk of metabolic dysfunction-associated steatosis liver disease progression, and has good clinical implementation potential.

6 Discussion

6.1 Comparative analysis of predictions

Intelligent learning algorithm algorithms have shown good predictive performance and can assist medical personnel in making medical decisions [19]. Their high classification accuracy and strong operability greatly improve clinical work efficiency, and their combination with clinical medicine has gradually become a hot topic. The progression of metabolic dysfunction-associated steatosis liver disease is the most common chronic liver disease, and how to use data mining methods to assist in prevention and treatment will be a valuable research field. The present work constructs a risk assessment model for metabolic dysfunction-associated steatosis liver disease in construction workers based on ensemble algorithms (CNN-LSTM hybrid model, XGBoost) and traditional intelligent learning algorithm (SVM, Naive Bayes, KNN, logistic regression). CNN-LSTM hybrid model, XGBoost, and NaiveBayes models are significantly better than traditional logistic regression in intelligent learning algorithm. Due to logistic regression being a traditional statistical modeling method, its parameter meanings are clear and the result indicators are easy to understand. However, it has high requirements for the data in the sample. When the data used does not meet the conditions, the prediction classification accuracy is low [20]. KNN is commonly used for disease risk assessment. The KNN algorithm is easy to implement, simple to understand, and particularly suitable for handling multi classification problems. In the validation dataset, KNN showed the weakest performance in terms of AUC and classification accuracy compared to the other five models. This may be because a single distance in the KNN model makes it difficult to accurately calculate the sample spacing that contains both discrete and continuous features, resulting in a decrease in classification accuracy. SVM, as a commonly used intelligent learning algorithm algorithm, has unique advantages in dealing with high-dimensional problems and generalization ability. The lower classification accuracy of SVM in the training dataset compared to the other four types of intelligent learning algorithm may be due to its lower predictive performance for problems with large sample sizes. Naive Bayes, as a commonly used intelligent learning algorithm

algorithm, has the advantages of being less sensitive to missing data and having stable classification efficiency? But in the training dataset, the classification accuracy of Naive Bayes is lower than the other five intelligent learning algorithm methods [21].

Boosting, as an ensemble learning algorithm, has a modeling phase and a voting phase. In modeling, surrogate modeling is often used to establish the model, and the optimal model can be selected in the voting phase. Finally, a set of weak classifiers is combined to create a strong classifier. The CNN-LSTM hybrid model and XGBoost models are Boosting algorithms, which have better model performance than traditional machine models. The XGBoost model has the characteristics of flexibility and efficiency, but it has defects such as high spatial complexity and unsuitability for processing high-dimensional feature data. The XGBoost model's more conservative fitting strategy compared to the CNN-LSTM hybrid model has caused slight underfitting issues, resulting in lower classification accuracy compared to the CNN-LSTM hybrid model. The present work constructs a metabolic dysfunction-associated steatosis liver disease (NAFLD) progression risk assessment model based on six intelligent learning algorithm models [22]: CNN-LSTM hybrid model, XGBoost, KNN, logistic, SVM, and NaiveBayes. The prediction performance of the model is evaluated and compared in terms of classification accuracy, precision, recall, F1 score, and AUC. The comparison results show that the CNN-LSTM hybrid model outperforms the other five models in terms of performance evaluation in both the training and testing sets.

6.2 Clinical implementation Value

Apply the constructed predictive framework to practical situations, fully leverage the role of the predictive framework, and facilitate its application value. There are many prognostic predictive frameworks widely used in clinical practice, such as scoring or column charts. For example, Wang Jiao et al. used Framingham risk assessment to evaluate the risk of cardiovascular disease; Boursier et al. [23] established a pelvic floor ultrasound score for postpartum stress urinary incontinence based on logistic regression model; Noureddin et al. [24] established a nomogram and used multivariate Cox regression to predict the postoperative survival of patients with lymph node positive pancreatic cancer. The above are all predictive frameworks constructed based on generalized linear models, and the established predictive frameworks are applied to clinical practice in the form of scores or column charts [25]. With the widespread application of artificial intelligence technology, intelligent learning algorithm algorithms often have better predictive performance than logistic regression in clinical practice. However, due to the complexity of intelligent learning algorithm algorithms, they cannot be applied to clinical practice in a scoring or column chart manner like logistic regression, which may encounter certain obstacles in their application process. The online calculator-based method provides a new choice, which can embed complex intelligent learning algorithm algorithms into the back end of the online computing system, input the corresponding feature information in the computing system, and then call the predictive framework of the back end to get the corresponding results. For example, Sanyal et al. [26] used intelligent learning algorithm algorithm to predict the risk of gestational diabetes, and applied the intelligent learning algorithm predictive framework to practice in the form of online web pages. An online evaluation tool based on predictive models, with a user-friendly interface that is convenient for patients and doctors to use at any time, greatly reducing the work pressure of medical staff and the strain on medical resources.

The present work is based on the medical examination data of employed laborers from a certain medical institution nationwide. After single factor and multi factor analysis, 16 risk factors were selected and included in six intelligent learning algorithm models, including CNN-LSTM hybrid model, XGBoost, KNN, logistic regression, SVM, and NaiveBayes, to construct

a nonalcoholic fatty liver progression risk assessment model [27]. Based on classification accuracy, precision, recall, F1 value, area under the curve (AUC), and calibration curve, the predictive performance of the model was evaluated and compared. Finally, the optimal model constructed by CNN-LSTM hybrid model was determined. The CNN-LSTM hybrid model algorithm ensures the prediction performance and provides decision-making reference for the medical diagnosis of metabolic dysfunction-associated steatosis liver disease progression [28]. It is conducive to providing a basis for early prevention of metabolic dysfunction-associated steatosis liver disease progression in employed laborers, thereby improving the health status and quality of life of employed laborers.

7 Conclusion

The present work focuses on the high prevalence and heavy disease burden of metabolic dysfunction-associated steatosis liver disease (NAFLD) in China, limited early screening methods, unsuitability of foreign predictive frameworks for Chinese physique, small sample size and insufficient predictive performance of domestic models. Based on multi-dimensional health data, a CNN-LSTM hybrid model metabolic dysfunction-associated steatosis liver disease (NAFLD) progression risk assessment method was developed, and six mainstream intelligent learning algorithm algorithms including XGBoost, KNN, SVM, logistic regression, and Naive Bayes were compared. AUC, classification accuracy, precision, recall, F1 value, ROC curve, and calibration curve were used as evaluation indicators to complete model training, validation, and performance optimization. The experimental results show that all indicators of CNN-LSTM hybrid model are superior to traditional statistical models and conventional intelligent learning algorithm models, with higher prediction classification accuracy, generalization ability, and stability. It achieves large sample, multi-dimensional, and highly accurate prediction of NAFLD progression risk, providing a lightweight and highly reliable intelligent tool for early screening of NAFLD. It has important practical significance in reducing the incidence of end-stage liver diseases such as late-stage liver fibrosis and hepatocellular carcinoma (HCC), reducing the burden of population health management and medical care, and improving the health level of employed laborers.

In future research, (1) the scope of the research subjects can be further expanded to include different occupational groups such as construction workers, manufacturing workers, and service workers, in order to enhance the universality of the model's population; (2) Continuous time-series monitoring data can be introduced, combined with wearable devices and long-term follow-up data, to enhance the modeling ability of LSTM for long-term health trends and dynamic risk evolution, achieving an upgrade from "static prediction" to "dynamic tracking". (3) We can conduct research on model light weighting and interpretability, and combine attention mechanisms, feature contribution visualization, and other methods to improve the deployment usability and clinical acceptance of the model in primary healthcare institutions.

References

[1] Beran A, Ayesh H, Mhanna M, et al. Triglyceride-glucose index for early prediction of nonalcoholic fatty liver disease: a meta-analysis of 121,975 individuals[J]. Journal of Clinical Medicine, 2022, 11(9): 2666.

[2] Mózes F E, Lee J A, Vali Y, et al. Performance of non-invasive tests and histology for the prediction of clinical outcomes in patients with non-alcoholic fatty liver disease: an

individual participant data meta-analysis[J]. The lancet Gastroenterology & hepatology, 2023, 8(8): 704-713.

[3] Asaturyan H A, Basty N, Thanaj M, et al. Improving the accuracy of fatty liver index to reflect liver fat content with predictive regression modelling[J]. PLoS One, 2022, 17(9): e0273171.

[4] Razmpour F, Daryabeygi-Khotbehsara R, Soleimani D, et al. Application of machine learning in predicting non-alcoholic fatty liver disease using anthropometric and body composition indices[J]. Scientific reports, 2023, 13(1): 4942.

[5] Crudele L, De Matteis C, Novielli F, et al. Fatty Liver Index (FLI) is the best score to predict MASLD with 50% lower cut-off value in women than in men[J]. Biology of sex Differences, 2024, 15(1): 43.

[6] Song K, Lee H W, Choi H S, et al. Comparison of the modified TyG indices and other parameters to predict non-alcoholic fatty liver disease in youth[J]. Biology, 2022, 11(5): 685.

[7] Fujiwara N, Kubota N, Crouchet E, et al. Molecular signatures of long-term hepatocellular carcinoma risk in nonalcoholic fatty liver disease[J]. Science translational medicine, 2022, 14(650): eabo4474.

[8] Song K, Park G, Lee H S, et al. Comparison of the triglyceride glucose index and modified triglyceride glucose indices to predict nonalcoholic fatty liver disease in youths[J]. The Journal of pediatrics, 2022, 242: 79-85. e1.

[9] Barbosa J V, Milligan S, Frick A, et al. Fibrosis-4 index can independently predict major adverse cardiovascular events in nonalcoholic fatty liver disease[J]. Official journal of the American College of Gastroenterology| ACG, 2022, 117(3): 453-461.

[10] Nam D, Chapiro J, Paradis V, et al. Artificial intelligence in liver diseases: Improving diagnostics, prognostics and response prediction[J]. Jhep Reports, 2022, 4(4): 100443.

[11] Eren F, Kaya E, Yilmaz Y. Accuracy of Fibrosis-4 index and non-alcoholic fatty liver disease fibrosis scores in metabolic (dysfunction) associated fatty liver disease according to body mass index: failure in the prediction of advanced fibrosis in lean and morbidly obese individuals[J]. European journal of gastroenterology & hepatology, 2022, 34(1): 98-103.

[12] De Vincentis A, Tavaglione F, Jamialahmadi O, et al. A polygenic risk score to refine risk stratification and prediction for severe liver disease by clinical fibrosis scores[J]. Clinical Gastroenterology and Hepatology, 2022, 20(3): 658-673.

[13] Tamaki N, Munaganuru N, Jung J, et al. Clinical utility of 30% relative decline in MRI-PDFF in predicting fibrosis regression in non-alcoholic fatty liver disease[J]. Gut, 2022, 71(5): 983-990.

[14] Roh E, Hwang S Y, Yoo H J, et al. Impact of non-alcoholic fatty liver disease on the risk of sarcopenia: a nationwide multicenter prospective study[J]. Hepatology International, 2022, 16(3): 545-554.

[15] Joshi S, Shamanna P, Dharmalingam M, et al. Digital twin-enabled personalized nutrition improves metabolic dysfunction-associated fatty liver disease in type 2 diabetes: results of a 1-year randomized controlled study[J]. Endocrine Practice, 2023, 29(12): 960-970.

[16] Crudele L, De Matteis C, Piccinin E, et al. Low HDL-cholesterol levels predict hepatocellular carcinoma development in individuals with liver fibrosis[J]. JHEP Reports, 2023, 5(1): 100627.

[17] Abdelhameed F, Kite C, Lagojda L, et al. Non-invasive scores and serum biomarkers for fatty liver in the era of metabolic dysfunction-associated steatotic liver disease (MASLD): a comprehensive review from NAFLD to MAFLD and MASLD[J]. Current obesity reports, 2024, 13(3): 510-531.

[18] Dritsas E, Trigka M. Supervised machine learning models for liver disease risk prediction[J]. Computers, 2023, 12(1): 19.

[19] Tincopa M A, Loomba R. Non-invasive diagnosis and monitoring of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis[J]. The Lancet Gastroenterology & Hepatology, 2023, 8(7): 660-670.

[20] Stefan N, Cusi K. A global view of the interplay between non-alcoholic fatty liver disease and diabetes[J]. The lancet Diabetes & endocrinology, 2022, 10(4): 284-296.

[21] Boursier J, Hagström H, Ekstedt M, et al. Non-invasive tests accurately stratify patients with NAFLD based on their risk of liver-related events[J]. Journal of hepatology, 2022, 76(5): 1013-1020.

[22] Kaneva A M, Bojko E R. Fatty liver index (FLI): more than a marker of hepatic steatosis: Kaneva and Bojko[J]. Journal of physiology and biochemistry, 2024, 80(1): 11-26.

[23] Boursier J, Hagström H, Ekstedt M, et al. Non-invasive tests accurately stratify patients with NAFLD based on their risk of liver-related events[J]. Journal of hepatology, 2022, 76(5): 1013-1020.

[24] Noureddin M, Ntanios F, Malhotra D, et al. Predicting NAFLD prevalence in the United States using National Health and Nutrition Examination Survey 2017–2018 transient elastography data and application of machine learning[J]. Hepatology Communications, 2022, 6(7): 1537-1548.

[25] Younossi Z M, Paik J M, Al Shabeeb R, et al. Are there outcome differences between NAFLD and metabolic-associated fatty liver disease?[J]. Hepatology, 2022, 76(5): 1423-1437.

[26] Sanyal A J, Castera L, Wong V W S. Noninvasive assessment of liver fibrosis in NAFLD[J]. Clinical Gastroenterology and Hepatology, 2023, 21(8): 2026-2039.

[27] Sanyal A J, Williams S A, Lavine J E, et al. Defining the serum proteomic signature of hepatic steatosis, inflammation, ballooning and fibrosis in non-alcoholic fatty liver disease[J]. Journal of hepatology, 2023, 78(4): 693-703.

[28] Huang D Q, Terrault N A, Tacke F, et al. Global epidemiology of cirrhosis—aetiology,

trends and predictions[J]. Nature reviews Gastroenterology & hepatology, 2023, 20(6): 388-398.