Heliyon 9 (2023) e18758

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

A dynamic machine learning model for prediction of NAFLD in a health checkup population: A longitudinal study

Yuhan Deng [a,b], Yuan Ma [c], Jingzhu Fu [d,e,f], Xiaona Wang [g], Canqing Yu [d,e,f,h], Jun Lv [d,e,f,h], Sailimai Man [b,d,e,f,***], Bo Wang [b,e,h,*], Liming Li [d,e,f,h,**]

[a] Chongqing Research Institute of Big Data, Peking University, Chongqing, China
[b] Meinian Institute of Health, Beijing, China
[c] School of Population Medicine and Public Health, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
[d] Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China
[e] Peking University Health Science Center Meinian Public Health Institute, Beijing, China
[f] Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China
[g] MJ Health Screening Center, Beijing, China
[h] Peking University Center for Public Health and Epidemic Preparedness & Response, Beijing, China

ARTICLE INFO

Keywords:
Machine learning
Dynamic prediction
Time-series data
Checkup records
NAFLD

ABSTRACT

Background: Non-alcoholic fatty liver disease (NAFLD) is one of the most common liver diseases worldwide. Currently, most NAFLD prediction models are diagnostic models based on cross-sectional data, which failed to provide early identification or clarify causal relationships. We aimed to use time-series deep learning models with longitudinal health checkup records to predict the onset of NAFLD in the future, and update the model stepwise by incorporating new checkup records to achieve dynamic prediction.

Methods: 10,493 participants with over 6 health checkup records from Beijing MJ Health Screening Center were included to conduct a retrospective cohort study, in which the constantly updated initial 5 checkup data were incorporated stepwise to predict the risk of NAFLD at and after their sixth health checkups. A total of 33 variables were considered, consisting of demographic characteristics, medical history, lifestyle, physical examinations, and laboratory tests. L1-penalized logistic regression (LR) was used for feature selection. The long short-term memory (LSTM) algorithm was introduced for model development, and five-fold cross-validation was conducted to tune and choose optimal hyperparameters. Both internal validation and external validation were conducted, using the 20% randomly divided holdout test dataset and previously unseen data from Shanghai MJ Health Screening Center, respectively, to evaluate model performance. The evaluation metrics included area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, Brier score, and decision curve. Bootstrap sampling was implemented to generate 95% confidence intervals of all the metrics. Finally, the Shapley additive explanations (SHAP) algorithm was applied in the holdout test dataset for model interpretability to obtain time-specific and sample-specific contributions of each feature.

Results: Among the 10,493 participants, 1662 (15.84%) were diagnosed with NAFLD at and after their sixth health checkups. The predictive performance of the deep learning model in the internal

* Corresponding author. Meinian Institute of Health, Beijing, China.
** Corresponding author. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China
*** Corresponding author. Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China
E-mail addresses: sailimai.man@meinianresearch.com (S. Man), paul@meinianresearch.com (B. Wang), lmlee@bjmu.edu.cn (L. Li).

https://doi.org/10.1016/j.heliyon.2023.e18758
Received 13 July 2023; Received in revised form 25 July 2023; Accepted 26 July 2023
Available online 27 July 2023

validation dataset improved over the incorporation of the checkups, with AUROC increasing from 0.729 (95% CI: 0.698,0.760) at baseline to 0.818 (95% CI: 0.798,0.844) when consecutive 5 checkups were included. The external validation dataset, containing 1728 participants, was used to verify the results, in which AUROC increased from 0.700 (95% CI: 0.657,0.740) with only the first checkups to 0.792 (95% CI: 0.758,0.825) with all five. The results of feature significance showed that body fat percentage, alanine transaminase (ALT), and uric acid owned the greatest impact on the outcome, time-specific, individual-specific and dynamic feature contributions were also produced for model interpretability.

Conclusion: A dynamic prediction model was successfully established in our study, and the prediction capability kept improving with the renewal of the latest checkup records. In addition, we identified key features associated with the onset of NAFLD, making it possible to optimize the prevention and control strategies of the disease in the general population.

1. Introduction

As one of the most common chronic hepatic diseases worldwide, non-alcoholic fatty liver disease (NAFLD) has affected 29.6% of the Asian population and the prevalence rate is increasing significantly over time [1]. NAFLD can transit to hepatic inflammation and fibrosis, and is highly associated with other non-liver-specific diseases, especially cardiovascular and metabolic disorders [2], presenting severe healthcare burdens globally [3]. However, given the multifactorial and intricate etiology of the disease, it's still difficult to determine a specific prevention strategy and achieve early identification of high-risk groups to reduce the prevalence of the disease. Therefore, improved prediction of the risk of NAFLD may be of great value in the prevention and control of the disease in the general population.

Artificial intelligence methods, together with massive data collected in the medical field, make it possible to precisely predict the risk of NAFLD. As expected, more and more research has been conducted to solve such issues. However, previous related studies were mostly cross-sectional and mainly focused on the development of diagnostic prediction models, which failed to determine causal relationships or provide early risk probabilities sometime before the confirmed diagnosis of NAFLD [4–6]. Besides, most existing studies were based on conventional machine learning models with a single measurement of variables [5,7,8], ignoring the variation tendency contained in multiple measurements of data.

To date, the valuable information contained in health checkup data, characterized by annually repeat-measured items, is underutilized, while the commonly used machine learning models, such as random forest and XGBoost, may be unable to handle these kinds of time-series data well [9]. Presently, deep learning-based models [10], including the recurrent neural network (RNN) and its derived models, can make full use of the constantly updated records and have been proven to show good performance, but almost in ICU settings [11]. However, despite the promising prediction performance, the complex structures and large amounts of parameters in deep learning models restrict their acceptance and practical application [12,13], while the introduction of model interpretability algorithms largely solved this problem and facilitate the understanding of the black boxes. Among all the algorithms, Shapley additive explanations (SHAP) algorithm [14], a widely used model interpretability method for any machine learning models, seems both effective and prospective in quantifying the impact of each variable on the outcomes [15,16].

To the best of our knowledge, no studies have ever used constantly updated time-series health checkup data to develop a dynamic prediction model based on deep learning algorithms to predict NAFLD. In this study, we use the long short-term memory (LSTM) model with SHAP interpretability algorithms to predict the risk of NAFLD one year after health checkups and update the prediction model

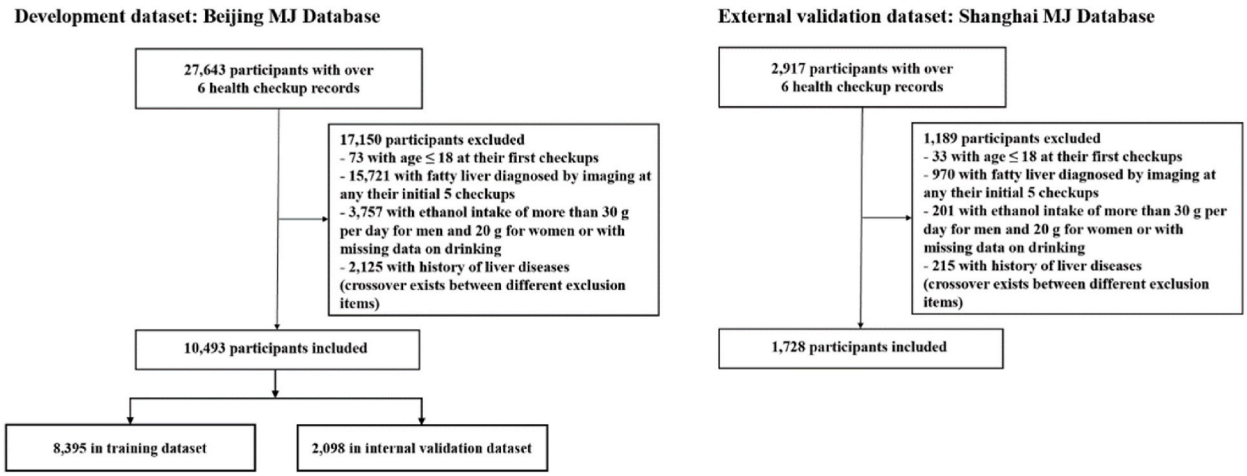

Fig. 1. Inclusion flow chart of study participants.

over checkups, and further verify the results on an external validation dataset.

2. Method

2.1. Data source and study participants

A retrospective cohort study was conducted based on data from Beijing MJ Health Screening Center. Beijing MJ Health Checkup Database contains routinely annual checkup data from 2003 to 2021, with more than 50,000 checkup records collected every year. During each checkup, a series of standardized protocols including physical examinations, laboratory tests, imaging diagnoses and a questionnaire survey about demographic characteristics, dietary habits, physical activity and other comprehensive lifestyle and health conditions were applied and inquired to all the participants. The external validation dataset was obtained from Shanghai MJ Health Screening Center, a health checkup center independent of that in Beijing, containing checkup records from 2003 to 2021.

In our study, participants who attended more than 6 checkups were included. The exclusion criteria were as follows: (1) age under 18 years at the first checkup; (2) having been diagnosed with fatty liver on liver ultrasound at any of the initial five checkups; (3) having ethanol intake of more than 30 g per day for men and 20 g for women or with missing data on drinking at any of the initial five checkups; (4) having a self-reported history of hepatic diseases at any of the initial five checkups. After excluding those who meet the exclusion criteria, a total number of 10,493 participants were included for model development and 1728 for model validation in our study (Fig. 1). The median time intervals between two consecutive checkups of the participants were presented in Supplemental F. 1.

2.2. Predictors and outcomes

The following variables at each of the initial 5 visits were extracted: (i) demographic characteristics: sex, age, education, income; (ii) physical examinations: body mass index (BMI), waist circumference (WC), body fat percentage (BFP), systolic blood pressure (SBP), diastolic blood pressure (DBP); (iii) laboratory tests: fasting blood glucose (FBG), total cholesterol (TC), triglyceride (TG), high-density lipoprotein cholesterol (HDLC), low-density lipoprotein cholesterol (LDLC), alanine transaminase (ALT), aspartate aminotransferase (AST), direct bilirubin (DB), total bilirubin (TB), total protein (TP), albumin (ALB), globulin (GLB), glutamyl transferase (GGT), alkaline phosphatase (ALP), lactate dehydrogenase (LD), uric acid (UA); (iv) smoking status; (v) dietary habits: dietary regularity, fruit intake, dairy intake, fried food intake; (vi) medical history: hypertension, diabetes and hyperlipidemia.

The consumption of food was classified into three tiers of low, moderate, and high, delineated by the following specifics: the low fruit intake is defined as either no consumption or less than 150 g per day, moderate intake is defined as consuming 150 g–300 g fruit per day, and high intake is defined as consuming more than 300 g fruit per day; the low intake of dairy products is defined as no consumption or less than one cup of milk (240 ml) or one serving of dairy products (30 g of cheese or 1 slice of cheese) per week, moderate intake is defined as consuming 1–3 cups of milk or 1–3 servings of dairy products per week, and high intake is defined as consuming more than four cups of milk or more than four servings of dairy products per week; a low level of intake of fried food indicates either no consumption or consumption of less than one serving per week, a moderate level of intake indicates consumption of 1–3 servings per week, and a high level of intake indicates consumption of four or more servings per week, with one serving defined as half a bowl.

The outcome was defined as whether the participant was diagnosed with NAFLD on ultrasonography from their sixth health checkups until the end of their follow-ups.

2.3. Data preprocessing and statistical analysis

Participants in the development dataset from Beijing MJ database were randomly divided into a training set (80%) and an internal validation dataset (20%), and imputations of missing values were conducted in the two datasets, together with the external validation dataset, respectively. Except for certain variables (sex for example) that were constant at each checkup, most of the variables were time-series forms, which means that they were measured repetitively at each checkup and not independent of each other, so the last observation carried forward (LOCF) method was conducted to impute missing values for repeated measurements. After the first imputation, variables with missing rates over 30% were excluded from our analysis, including TP and ALB. For continuous variables, their means were used to impute the rest missingness. Outliers were identified as values distributed more or less than three standard deviations from the mean and were handled the same way as missing values. All continuous variables were maintained as their original forms in case of information loss, and categorical variables with multiple levels were re-classified, then the unordered were one-hot encoded, while the ordered were kept as their reclassification forms. Continuous variables were presented as means with standard deviations (SDs) and utilized Student's t-test or Wilcoxon rank-sum test for statistical analysis. Categorical variables were presented as counts and percentages, and the Chi-square test was employed for comparison.

After the preprocessing step, L1-penalized logistic regression (LR) was conducted based on the newest records in the training set to select a subset of the total features generated from the preprocessing procedure, then the LSTM algorithm was introduced to develop the prediction model, and the model was updated as the number of visit times increased and new measurements of predictors entered. The hyperparameters were selected through a five-fold cross-validation process, in which all combinations of hyperparameters were tested and the optimal combination was chosen based on the mean area under the receiver operating curve (AUROC) across five validation sets derived from the training dataset. Besides, considering the imbalanced issue of the dataset [17], in which participants with NAFLD were less than those without, we kept all the samples in the minority class in each epoch and randomly selected the same

number of samples from the majority class to train the model, and both balanced datasets and imbalanced datasets were tested through five-fold cross-validation. In addition, to establish a benchmark for the prediction of NAFLD, LR was conducted using data based on each health checkup.

In the internal validation dataset and external validation dataset, discrimination and calibration of the prediction models were assessed, in which AUROC, sensitivity, and specificity were used as metrics of discrimination ability, while Brier score was used to evaluate calibration capability. Besides, Platt scaling method [18] was used to transform outputs into risk probabilities by fitting LR to improve calibration, and the Brier score was calculated after recalibration. Bootstrap sampling was performed to generate 95% confidence intervals (CIs) of the metrics above. Specifically, the same size of samples as the validation dataset was randomly selected with replacement 1000 times, and the distribution of the results was used to obtain 95% CIs. Meanwhile, decision-curve characteristic on the internal validation dataset was evaluated. The Delong test was utilized to compare the consecutive AUCs resulting from model updates. Furthermore, as predictive performance with model updates was not independent, we employed repeated measures analysis of variance (ANOVA) to compare the overall performance between models based on LSTM algorithms and LR across all five time points. After the comparison of the updated models, the best-performed model was selected to show feature significance. The SHAP algorithm was applied in the internal validation dataset to enhance models' interpretability by obtaining patient-specific contributions of each feature at a specific time point.

All the data preprocessing was conducted using SAS 9.4 and statistical analyses using Python 3.7.

2.4. LSTM for model development

LSTM [19] is introduced on the basis of the RNN, which can tackle time-series problems and selectively retain important information through its gating systems: input gate (I_t), forget gate (F_t) and output gate (O_t). Specifically, the input gate determines how much information to input at the present step, the forget gate controls how much information to throw away from the previous steps, and the output gate decided how much current information to output. After the disposal conducted by the gate systems, redundant information is filtered out, largely releasing the memory of the hidden layer. The equations and diagram of an LSTM neuron (Fig. 2) are as below:

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i)$$

$$F_t = \sigma\left(X_t W_{xf} + H_{t-1} W_{hf} + b_f\right)$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o)$$

$$\widetilde{C}_t = \tanh\ (X_t W_{xc} + H_{t-1} W_{hc} + b_c)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \widetilde{C}_t$$

$$H_t = O_t \odot tanh(C_t)$$

In our study, the NAFLD risk prediction was updated timely by incorporating new checkup records to construct dynamic models, and all these longitudinal data integrated into the models were used for the prediction of NAFLD in future checkups.

2.5. SHAP for model interpretability

In the SHAP algorithm [14], feature importance is presented according to Shapley values. For a single sample, a probability is a certain output from a prediction model, indicating the outcome risk of this sample, and in the same way, we can get a mean probability of all samples. Through the SHAP algorithm, each of the current features of this sample corresponds to a Shapley value, which indicates the contribution of this feature (considering its interactions with other features) to the outcome. The sum of all Shapley values is the difference between the actual probability of a single sample and the mean probability of whole samples, which can be represented as:

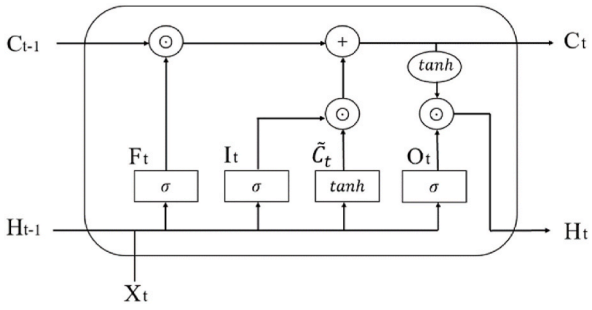

Fig. 2. Diagram of an LSTM neuron.

Table 1
Characteristics of study participants at baseline stratified by outcome.

Characteristic, n (%)	Total (N = 10,493)	NAFLD		*P* value
		Yes (N = 1662)	No (N = 8831)	
Demographic characteristics				
Sex				<0.001
Male	3162 (30.13)	714 (42.96)	2448 (27.72)	.
Female	7331 (69.87)	948 (57.04)	6383 (72.28)	.
Age/year	35.40 ± 8.73	37.13 ± 9.02	35.08 ± 8.63	<0.001
Education				<0.001
University degree or below	1781 (16.97)	390 (23.47)	1391 (15.75)	
Undergraduate	6018 (57.35)	961 (57.82)	5057 (57.26)	.
University degree and above	2694 (25.67)	311 (18.71)	2383 (26.98)	.
Income				<0.001
Low	7867 (74.97)	1326 (79.78)	6541 (74.07)	
Moderate	1456 (13.88)	191 (11.49)	1265 (14.32)	.
High	1170 (11.15)	145 (8.72)	1025 (11.61)	.
Body measurement				
BMI/(kg/m^2)	21.30 ± 2.27	22.45 ± 2.15	21.08 ± 2.23	<0.001
Waist circumference/cm	72.35 ± 7.25	75.93 ± 7.11	71.68 ± 7.08	<0.001
Body fat percentage/%	24.66 ± 5.54	26.33 ± 5.80	24.35 ± 5.44	<0.001
SBP/mmHg	105.86 ± 11.65	108.00 ± 12.34	105.46 ± 11.47	<0.001
DBP/mmHg	64.88 ± 8.72	65.91 ± 9.32	64.69 ± 8.59	<0.001
Laboratory tests				
FBG/(mmol/L)	5.25 ± 0.36	5.33 ± 0.35	5.24 ± 0.36	<0.001
TC/(mmol/L)	4.37 ± 0.70	4.47 ± 0.73	4.35 ± 0.70	<0.001
TG/(mmol/L)	0.84 ± 0.34	0.95 ± 0.37	0.82 ± 0.33	<0.001
HDLC/(mmol/L)	1.57 ± 0.29	1.50 ± 0.28	1.59 ± 0.29	<0.001
LDLC/(mmol/L)	2.58 ± 0.59	2.67 ± 0.59	2.56 ± 0.59	<0.001
ALT/(U/L)	15.93 ± 7.36	17.70 ± 7.84	15.59 ± 7.22	<0.001
AST/(U/L)	17.94 ± 4.51	18.44 ± 4.61	17.84 ± 4.48	<0.001
Direct bilirubin/(μmol/L)	3.00 ± 2.15	2.42 ± 2.15	3.11 ± 2.13	<0.001
Total bilirubin/(μmol/L)	8.63 ± 6.25	7.29 ± 6.41	8.88 ± 6.18	<0.001
Total protein/(g/L)	72.80 ± 2.44	72.81 ± 2.08	72.80 ± 2.50	0.838
Albumin/(g/L)	46.43 ± 1.53	46.48 ± 1.34	46.42 ± 1.56	0.087
Globulin/(g/L)	26.38 ± 2.15	26.34 ± 1.83	26.38 ± 2.21	0.351
GGT/(U/L)	15.24 ± 5.95	16.36 ± 5.84	15.03 ± 5.95	<0.001
ALP/(U/L)	57.65 ± 10.59	59.28 ± 9.59	57.34 ± 10.74	<0.001
LD/(U/L)	148.94 ± 15.65	150.34 ± 14.61	148.68 ± 15.83	<0.001
Uric acid/(μmol/L)	277.90 ± 58.55	292.31 ± 58.32	275.19 ± 58.19	<0.001
Living habits				
Smoking status				<0.001
Current	971 (9.25)	237 (14.26)	734 (8.31)	.
Former	200 (1.91)	25 (1.50)	175 (1.98)	.
Never	9322 (88.84)	1400 (84.24)	7922 (89.71)	.
Dietary regularity				0.161
Regular	8563 (81.61)	1336 (80.39)	7227 (81.84)	
Irregular	1930 (18.39)	326 (19.61)	1604 (18.16)	.
Fruit intake				0.061
Low	2117 (20.18)	357 (21.48)	1760 (19.93)	.
Moderate	7468 (71.17)	1183 (71.18)	6285 (71.17)	.
High	908 (8.65)	122 (7.34)	786 (8.90)	.
Dairy intake				0.003
Low	4905 (46.75)	840 (50.54)	4065 (46.03)	.
Moderate	2707 (25.80)	401 (24.13)	2306 (26.11)	.
High	2881 (27.46)	421 (25.33)	2460 (27.86)	.
Fried food intake				0.132
Low	6177 (58.87)	951 (57.22)	5226 (59.18)	.
Moderate	3734 (35.59)	604 (36.34)	3130 (35.44)	.
High	582 (5.55)	107 (6.44)	475 (5.38)	.
Personal history				
Hypertension				<0.001
Yes	404 (3.85)	106 (6.38)	298 (3.37)	
No	10,089 (96.15)	1556 (93.62)	8533 (96.63)	.
Diabetes				0.483
Yes	133 (1.27)	24 (1.44)	109 (1.23)	
No	10,360 (98.73)	1638 (98.56)	8722 (98.77)	.
Hyperlipidemic				<0.001
Yes	2625 (25.02)	516 (31.05)	2109 (23.88)	
No	7868 (74.98)	1146 (68.95)	6722 (76.12)	.

BMI = body mass index. SBP = systolic blood pressure. DBP = diastolic blood pressure. FBG = fasting blood glucose. TC = total cholesterol. TG = triglyceride. HDL-C = high-density lipoprotein cholesterol. LDL-C = low-density lipoprotein cholesterol. ALT = alanine transaminase. AST = aspartate aminotransferase. GGT = glutamyl transferase. ALP = alkaline phosphatase. LD = lactate dehydrogenase.

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{ij})$$

In our study, the global contribution, together with the time-specific and sample-specific contributions of each feature can be generated by this algorithm.

3. Results

After excluding individuals who did not meet the inclusion criteria, a total of 10,493 participants were included for model development in our study. Among them, 1662 (15.84%) were diagnosed with NAFLD at and after their sixth checkup. Characteristics of study participants at baseline stratified by the outcome are presented in Table 1. People with the onset of NAFLD were older ($P < 0.001$), comprised more males ($P < 0.001$), with lower educational degrees ($P < 0.001$) and lower income ($P < 0.001$). Moreover, they had higher values of anthropometric measures, including BMI ($P < 0.001$), WC ($P < 0.001$), BFP ($P < 0.001$), SBP ($P < 0.001$), and DBP ($P < 0.001$). The differences in most of the laboratory tests between the two groups were also statistically significant ($P < 0.05$). Besides, the NAFLD group comprised more current smokers ($P < 0.001$), and had lower dairy intake ($P = 0.003$), and the prevalence of hypertension ($P < 0.001$) and hyperlipidemia ($P < 0.001$) was also significantly higher in the NAFLD group. To comprehensively

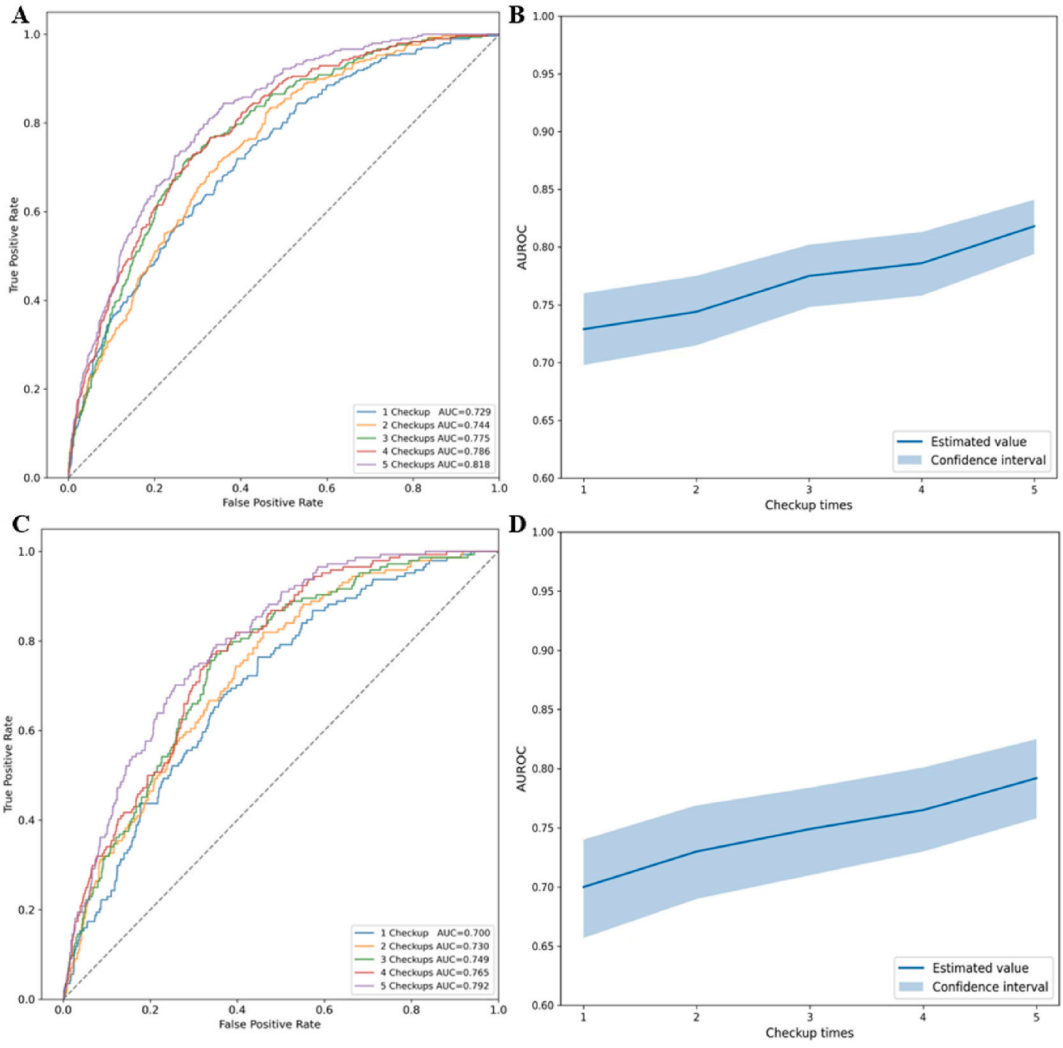

Fig. 3. Comparison of AUCs as the number of checkups increased. ROC curves of internal validation (A) and external validation (C); AUC as a function of checkup times in the internal validation dataset (B) and external validation dataset (D).

demonstrate the longitudinal characteristics of the study participants, we also presented the characteristics of their fifth health checkup and the checkup prior to their last follow-up in Supplemental Tab. 1 and Supplemental Tab. 2, respectively. Following the random partitioning of the dataset, we investigated differences in the baseline characteristics of the study participants between the training and internal validation sets. Supplemental Tab. 3 displays the results of this assessment, which demonstrated no statistically significant differences in any of the analyzed features.

The characteristics of participants at baseline and at fifth checkup in the external validation dataset after feature selection were presented in Supplemental Tab. 4 and Supplemental Tab. 5.

After feature selection, a total of 16 features, including sex, education, income, BMI, WC, BFP, ALT, UA, DB, LD, TG, AST, HDLC, ALP, GGT and GLB were included in our model. Model performance in predicting NAFLD improved as checkup records were renewed. In the internal validation dataset, the AUC increased significantly from 0.729 (95% CI: 0.698,0.760) at baseline to 0.818 (95% CI: 0.798,0.844) when consecutive 5 checkups were included (Fig. 3A–B). The recalibrated Brier scores decreased when more checkup records were included, indicating improved calibration. The decision curve analysis also showed that the net benefits of the model incorporated all five checkups significantly outperformed those of the other throughout all threshold probabilities (Fig. 4). The same growth trend in AUC can be seen in the external validation set, from 0.700 (95% CI: 0.657,0.740) at baseline to 0.792 (95% CI: 0.758,0.825) with all 5 checkups (Fig. 3C–D), although less accurate compared with those in the internal validation dataset. Brier scores also showed an improved calibration with the update of the checkups. The results of Delong test indicated that the improvement of AUC was statistically significant both in internal and external validation datasets (Supplemental Tab. 6). All metrics are presented in Table 2.

For comparison, LR was also used to develop prediction models at each health checkup, the model performance was presented in Supplemental Tab. 7. Although AUC increased with updated health checkup records, from 0.724 (95% CI: 0.695,0.753) to 0.784 (95% CI: 0.758,0.810) in internal validation dataset, and from 0.713 (0.673,0.753) to 0.787 (0.750,0.823) in external validation dataset, the overall performance of the model was inferior to that of LSTM models ($P < 0.0001$) based on the result of repeated measures ANOVA, and the improvement trend was not as strong as the latter ($P < 0.0001$), as evidenced by the results presented in Supplemental Tab. 8 and Supplemental Tab. 9. Specifically, significant differences ($P < 0.001$) were observed in the prediction performance of the models in the internal validation set as the number of health checkups varied. Additionally, the LSTM models outperformed the LR models in terms of overall performance over five time points, with mean AUC values of 0.770 and 0.752, respectively. Moreover, the trend in AUC improvement with the update of checkup records differed significantly between the two models, with the LSTM model showing a stronger improvement trend than the LR model (difference in AUC between final and initial visits: 0.089 vs. 0.060, respectively). These results were further supported by external validation. Besides, to more clearly compare the predictive performance differences between LR and LSTM models, we also used Wilcoxon signed rank tests to compare the AUC differences between the two models at or before each time point. The results are presented in Supplemental Tab. 10.

To address the concern of potential improvement in model performance due to the decreasing time interval between the measurement time of the variables and the occurrence of the outcome with the incorporation of new health checkup records, we further developed models relative to the time of the fifth checkup, incorporating the fourth, the third, and until the first gradually, and evaluating its performance. The results are presented in Fig. 5A and B and Table 3, which also showed an increased trend in model performance.

The contributions of each feature in the model related to the risk of NAFLD are shown in Fig. 6A and B. Among all these features, body fat percentage, ALT, and uric acid own the biggest impact on the outcome, with higher values above indicating greater risks of the development of NAFLD. Besides, some demographic characteristics, like income and education, are also at the priority level.

The dynamic prediction model can also be interpreted at a specific checkup time, the summary plots of which can be seen in Supplemental F. 2–6. The results of the mean importance rank of the features over time are presented in Fig. 7, which illustrates the changing contributions of a certain feature. We note that body fat percentage and ATL occupied the top ranks in almost all checkup times, HDLC gained importance with the extension of time steps, while the ranks of DB decreased, meaning losing importance over

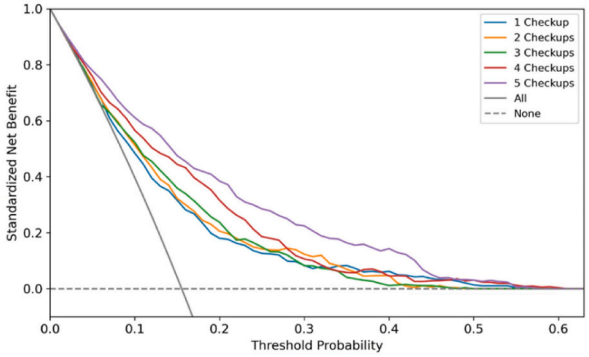

Fig. 4. Decision curve analysis with the update of checkups in the internal validation dataset. The gray solid line and the gray dashed line represent the net benefit generated by intervention, and non-intervention to all participants, respectively. Other colored lines present intervention to all potential patients identified by each model throughout threshold probabilities.

Table 2
Model performance with the update of health checkups based on LSTM.

Model	AUC	Sensitivity	Specificity	Brier score
Internal validation dataset				
1 Checkup	0.729 (0.698,0.760)	0.700 (0.548,0.870)	0.641 (0.459,0.780)	0.131 (0.119,0.142)
2 Checkups	0.744 (0.715,0.775)	0.774 (0.649,0.884)	0.605 (0.481,0.721)	0.112 (0.101,0.123)
3 Checkups	0.775 (0.748,0.802)	0.735 (0.653,0.808)	0.715 (0.652,0.784)	0.087 (0.076,0.094)
4 Checkups	0.786 (0.758,0.813)	0.748 (0.660,0.872)	0.700 (0.558,0.775)	0.067 (0.058,0.077)
5 Checkups	0.818 (0.798,0.844)	0.804 (0.713,0.873)	0.690 (0.627,0.772)	0.052 (0.041,0.066)
External validation dataset				
1 Checkup	0.700 (0.657,0.740)	0.741 (0.532,0.897)	0.590 (0.422,0.788)	0.076 (0.066,0.087)
2 Checkups	0.730 (0.690,0.769)	0.765 (0.646,0.925)	0.579 (0.446,0.745)	0.077 (0.066,0.088)
3 Checkups	0.749 (0.710,0.784)	0.786 (0.702,0.877)	0.642 (0.524,0.686)	0.076 (0.066,0.088)
4 Checkups	0.765 (0.730,0.801)	0.792 (0.702,0.904)	0.646 (0.516,0.710)	0.067 (0.057,0.078)
5 Checkups	0.792 (0.758,0.825)	0.791 (0.611,0.911)	0.695 (0.503,0.790)	0.066 (0.055,0.077)

LSTM = long short-term memory. AUC = area under the curve.

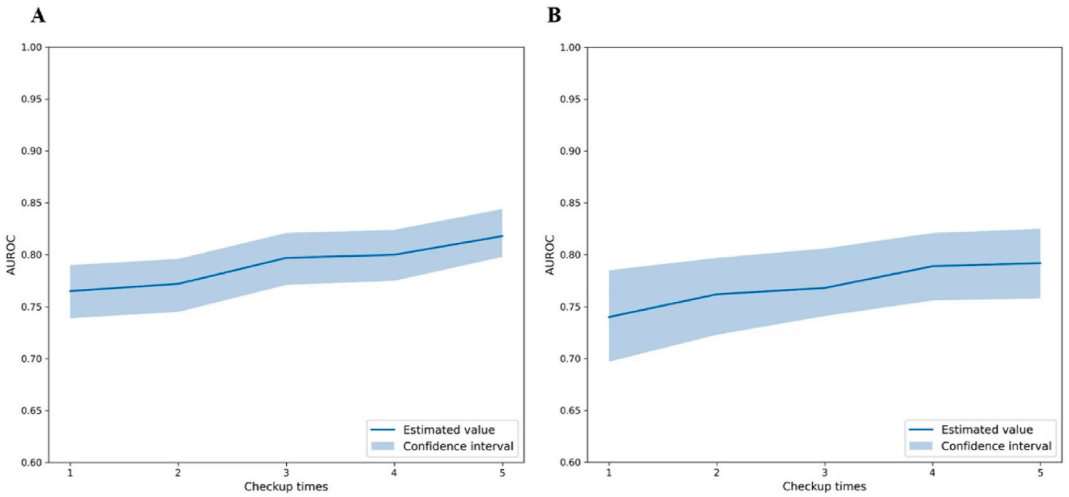

Fig. 5. AUC as a function of checkup times incorporated from the fifth checkup till the first in the internal validation dataset (A) and external validation dataset (B).

Table 3
Model performance incorporated checkup data from the fifth till the first.

Model	AUC	Sensitivity	Specificity	Brier score
Internal validation dataset				
5th Checkup	0.765 (0.739,0.790)	0.770 (0.679,0.858)	0.655 (0.564,0.729)	0.118 (0.110,0.127)
4th-5th Checkups	0.772 (0.745,0.796)	0.774 (0.677,0.855)	0.666 (0.566,0.750)	0.115 (0.106,0.123)
3rd-5th Checkups	0.797 (0.771,0.821)	0.789 (0.699,0.865)	0.684 (0.608,0.763)	0.092 (0.083,0.101)
2nd-5th Checkups	0.800 (0.775,0.824)	0.773 (0.686,0.844)	0.712 (0.649,0.807)	0.071 (0.060,0.084)
1st-5th Checkups	0.818 (0.798,0.844)	0.804 (0.713,0.873)	0.690 (0.627,0.772)	0.052 (0.041,0.066)
External validation dataset				
5th Checkup	0.740 (0.697,0.785)	0.629 (0.494,0.748)	0.770 (0.684,0.879)	0.088 (0.082,0.095)
4th-5th Checkups	0.762 (0.723,0.797)	0.717 (0.494,0.888)	0.682 (0.449,0.888)	0.091 (0.084,0.098)
3rd-5th Checkups	0.768 (0.741,0.806)	0.784 (0.696,0.892)	0.647 (0.533,0.723)	0.081 (0.073,0.088)
2nd-5th Checkups	0.789 (0.756,0.821)	0.804 (0.607,0.950)	0.640 (0.476,0.832)	0.079 (0.073,0.086)
1st-5th Checkups	0.792 (0.758,0.825)	0.791 (0.611,0.911)	0.695 (0.503,0.790)	0.066 (0.055,0.077)

AUC = area under the curve.

time.

We presented the fourth and fifth checkups of a representative case from the internal validation dataset in Fig. 8A and B, respectively. The participant was a 45-year-old male, with an undergraduate degree and a moderate level of income. At his fourth checkup, his BMI was 26.93 kg/m^2, waist circumference 82 cm, and body fat percentage 39.0%, with none of the three chronic diseases. Laboratory tests showed his direct bilirubin 2.67 μmol/L, ALT 16.0U/L, TG 1.95 mmol/L, and HDLC 1.01 mmol/L, which

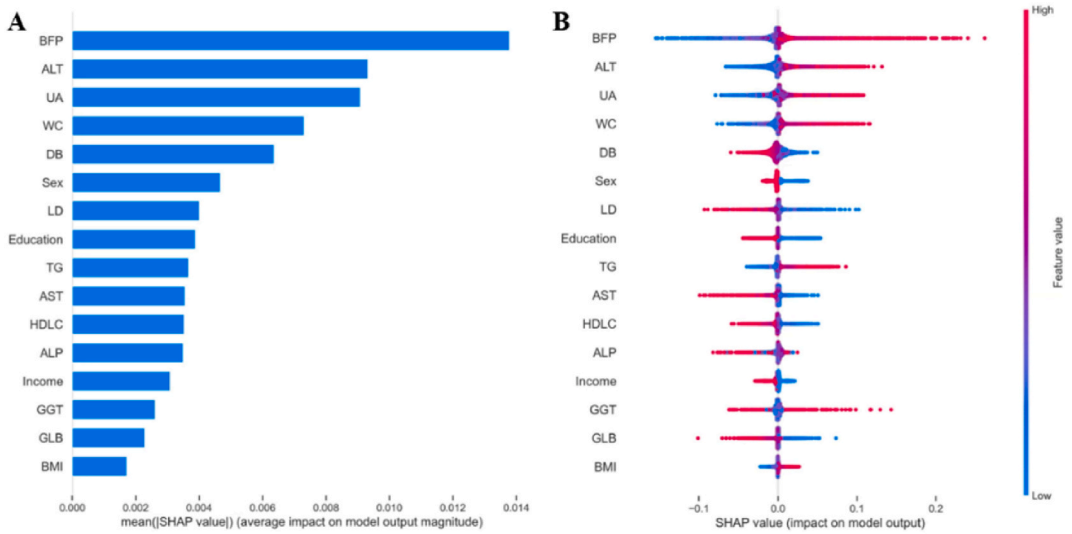

Fig. 6. Summary plot of the impact of features on the risk of NAFLD. (A) The blue bars represent the sum of the absolute SHAP value on features. The higher SHAP values indicate more contributions to prediction. (B) Feature significance decreased from top to bottom. Each point represents the impact of a feature on NAFLD prediction for one participant at a given checkup time. Red points with positive SHAP values, which are the same as blue points with negative SHAP values, represent positive associations with the outcome, while red points with negative SHAP values, similarly to blue points with positive SHAP values, mean negative relationships. BFP = body fat percentage. ALT = alanine transaminase. UA = uric acid. WC = waist circumference. DB = direct bilirubin. LD = lactate dehydrogenase. TG = triglyceride. AST = aspartate aminotransferase. HDLC = high-density lipoprotein cholesterol. ALP = alkaline phosphatase. GGT = glutamyl transferase. GLB = globulin. BMI = body mass index.

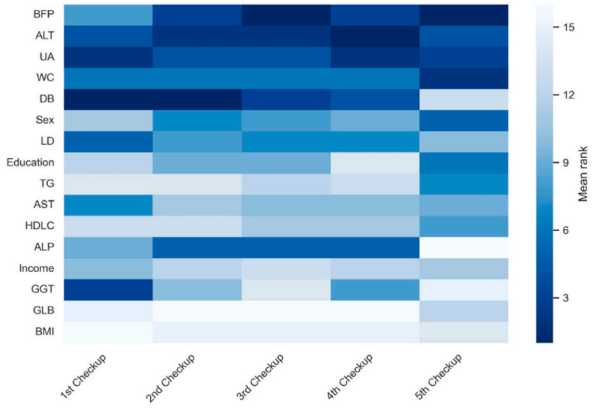

Fig. 7. The dynamic impact of features on the risk of NAFLD. BFP = body fat percentage. ALT = alanine transaminase. UA = uric acid. WC = waist circumference. DB = direct bilirubin. LD = lactate dehydrogenase. TG = triglyceride. AST = aspartate aminotransferase. HDLC = high-density lipoprotein cholesterol. ALP = alkaline phosphatase. GGT = glutamyl transferase. GLB = globulin. BMI = body mass index.

were all risk factors identified by the model. One year later, he tested a 10U/L increase in GGT, which was out of the normal range and pulled the prediction towards the onset of NAFLD. The values of waist circumstance, TG and direct bilirubin decreased but were still identified as risk factors. Meanwhile, the measurements of BMI, body fat percentage, HDLC, ALT and LD, increased a little compared with the fourth checkup, although most of which, especially the laboratory test results, were still within their normal limits, identified as strong risk factors to the prediction. This participant was diagnosed with NAFLD at his sixth checkup. More information about this individual can be seen in Supplemental Tab. 11.

BFP = body fat percentage. GGT = glutamyl transferase. DB = direct bilirubin. ALT = alanine transaminase. TG = triglyceride. HDLC = high-density lipoprotein cholesterol. UA = uric acid. WC = waist circumference. LD = lactate dehydrogenase. BMI = body mass index. GLB = globulin.

We further explored the relationships between features and the outcome, and detailed how some features interact (Fig. 9). A non-linear relationship exists between UA measured at the third checkup, ALP measured at the fourth checkup and the outcome (Fig. 9A and B). The normal ranges of UA and ALP are 180–420 mmol/L and 40–150U/L, respectively, meaning that the results of these two laboratory tests in most participants are within the normal ranges. Besides, the relationship between the measurement value and the

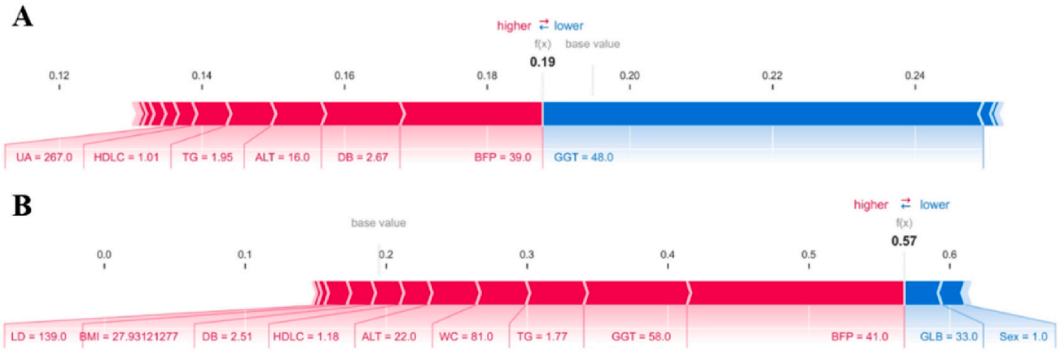

Fig. 8. Time-specific impact of features for a single participant (A) the 4th checkup (B) the 5th checkup. The features in red color drive the outcome towards the side of the development of the NAFLD, while the features in blue color pull down the risk probability.

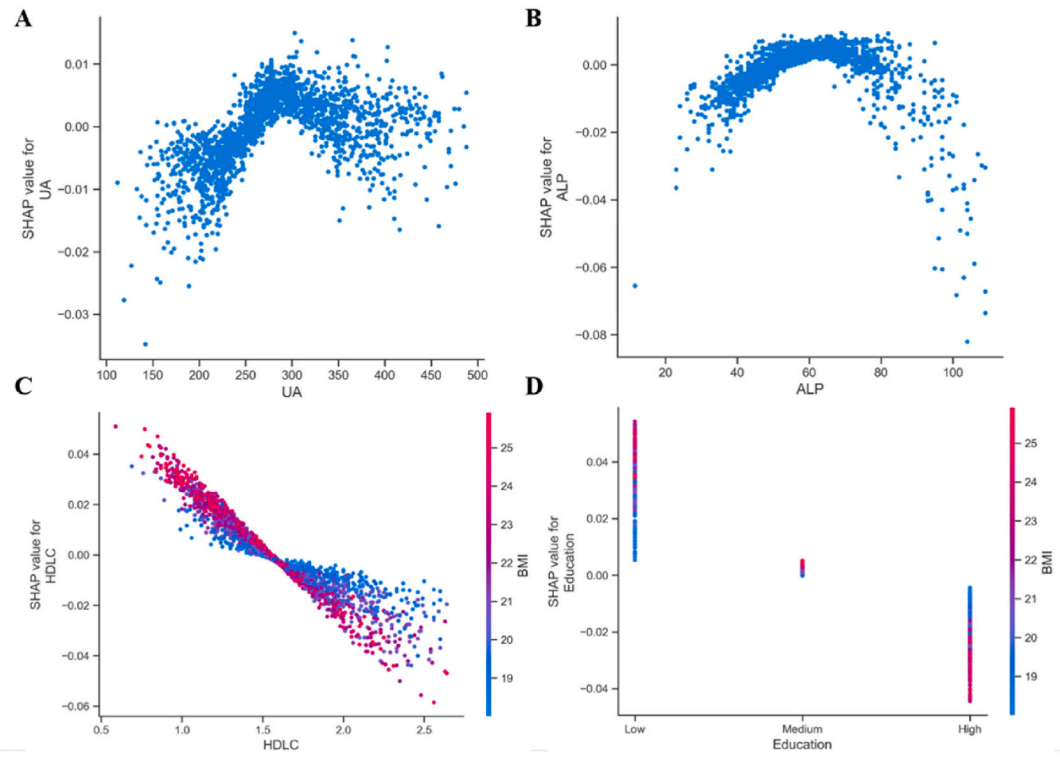

Fig. 9. Impact of feature interactions on NAFLD prediction. (A) the relationship between UA and the outcome; (B) the relationship between ALP and the outcome; (C) the interaction of HDLC and BMI; (D) the interaction of education and BMI. The SHAP value corresponding to each dot represents the impact of a feature and its interaction with other features for one participant. The positive SHAP value pushes the prediction towards the onset of NAFLD, while the negative SHAP value reduces the risk of the disease. UA = uric acid. ALP = alkaline phosphatase. HDLC = high-density lipoprotein cholesterol. BMI = body mass index.

risk of NAFLD was found to be non-linear, contrary to the expected linear relationship. Specifically, the risk of NAFLD increased initially with the measurement values, followed by a decrease, indicating a complex relationship between the predictors and the outcome. Furthermore, it is obvious that a higher value of HDLC is associated with a lower risk of NAFLD, but when considering BMI, we can see that most of the dots in red color, indicating high BMI, are distributed far away from the position of 0 SHAP value, which means that the interaction of high BMI and HDLC has a stronger impact on the prediction (Fig. 9C). Similarly, a lower educational degree forebodes a higher risk of NAFLD, and a higher BMI also strengthens the impact (Fig. 9D).

4. Discussion

In this study, a time-series deep learning model was developed to predict the risk of NAFLD within the health checkup population,

and the model was updated when new checkup records were incorporated to achieve dynamic prediction. SHAP algorithm was used to enhance model interpretability and identify features that had a great impact on outcomes.

As expected, with the extension of the time step, the latest records were incorporated and more information about temporal trends was contained, so the model performance improved significantly. However, when incorporating records from the first checkups to the fifth checkups stepwise, the time of predictors measurement will approach the time of prediction, so the model performance may improve on this account. To deal with this problem, we further developed a model relative to the time of the fifth checkup, incorporating the fourth, the third, and until the first gradually, and evaluating its performance. The results also showed an increased trend in model performance, although not as significant as the forward incorporation. This finding indicates that the features measured years before the development of NAFLD are still of great value and need to be taken into account to better predict the outcome.

While cross-sectional data-based diagnostic prediction models have shown promising results, they fall short in predicting the future incidence of NAFLD and identifying high-risk populations before the onset of the disease [5,7,8,20]. Prognosis prediction models, on the other hand, excel in early risk estimation for diseases. Compared to existing studies on prognosis prediction models, the predictive performance of the model developed in our study demonstrates certain advantages. In a study by Zhu et al., multivariate LR model was constructed to predict the risk of NAFLD among lean pre-diabetics with normal blood lipid levels, achieving an AUC of 0.796 in the longitudinal internal validation with a 5-year follow-up period [21]. Abeysekera et al. utilized variables measured at the participants' adolescence to predict the incidence risk of NAFLD at the age of 24 using LR, yielding the highest AUROC of 0.79 [22]. Wang et al. conducted a four-year cohort study to predict the incidence risk of NAFLD among 13,240 baseline NAFLD-free individuals using cox proportional hazards regression analyses. In the internal validation cohort, the AUC for 1-year, 2-year, 3-year, and 4-year risk predictions were 0.817, 0.820, 0.814, and 0.813, respectively [23]. In contrast, our study focused on developing a dynamic prediction model with constantly updated health checkup records using advanced deep learning algorithms that estimated the incidence risk of NAFLD from the sixth health checkup onwards until the end of the follow-up period. The internal validation set yielded an impressive highest AUC of 0.818 (0.798, 0.844), showcasing the innovative and competitive performance of our model.

Our research reveals that the prediction model we established surpasses the LR model used as the benchmark. We found that our LSTM-based model's performance further enhanced as more recent health checkup records become available compared to the LR model. This implies that the LSTM model could offer more accurate predictions over a longer time span of data. Additionally, as both models employ the same predictive variables, the cost of using either model remains identical. Our validation set results demonstrate that while the LR model's discriminative ability improves with updated health checkup records, its calibration ability decreases. Conversely, our LSTM-based prediction model exhibits greater robustness, stability, and generalizability. Our findings suggest that the LSTM model could serve as a more dependable and precise tool for time-series data prediction, offering superior performance to the LR model.

Although in many studies, BMI was regarded as an important predictor of NAFLD [24–26], we identified the more valid body measurement indicator, body fat percentage. This indicator could help recognize those with normal weight obesity, whose BMI cannot tell [27]. Besides, this feature had a roughly increased importance rank over time, showing more value than other body measurement indicators [28]. As a practical, easily measured, and non-invasive indicator, we recommend using body fat percentage to supplement other body measurement indicators. However, further research needs to be done to test whether there exists a causal relationship between BFP and NAFLD. ALT, a standard and commonly used indicator of liver function, was found to have a great impact on the prediction in our study. As an enzyme in the cytosol of hepatocytes, only a low level of ALT can be detected in the serum in a healthy population, elevated ALT level usually indicates the injury or apoptosis of hepatocytes [29]. Previous studies have also concluded that the increased ALT value was associated with a higher risk of NAFLD [30–32]. In a cross-sectional study on elderly Chinese man and woman, ALT was observed to have a significant joint association with serum uric acid to NAFLD prevalence [33], although we haven't considered the joint contribution of the two features in our study, serum uric acid was also found to have a strong impact on the outcome. The global contribution of UA in all five checkups showed that a high level of the feature driving the outcome towards the side of NAFLD. Meanwhile, we also found a non-linear relationship between UA and the outcome that existed in a single checkup. Some cross-sectional studies have demonstrated that serum uric acid was positively associated with the prevalence of NAFLD in Chinses population [34]. Besides, a case-control study also indicated that uric acid was an independent predictor of NAFLD [35]. As the end product of purine metabolism, high levels of serum uric acid could induce fat accumulation and then lead to the development of fatty liver [36]. Waist circumference was also a significant predictor of NAFLD, consistent finding has been revealed in previous studies [7, 37]. Furthermore, our results also showed that direct bilirubin levels from five or more years ago can have a significant and continuous effect on the development of NAFLD, which may hint at its long-term effect on the onset of the disease. Nevertheless, lifestyle habits collected from the questionnaire survey, including dietary condition and smoking status, although proved to be associated with the development of NAFLD by other studies [38–40], were not regarded as important factors in our study. It makes sense when considering the long-time interval between the survey time of these habits and the prediction, along with the possible changeability of these features. From the SHAP algorithms, we also illustrated the interactions between some features. We found that a high BMI could strengthen the impact of certain features on the outcome, including HDLC and education. Specifically, the interactions between high BMI and the two features tended to drive the outcome towards two extremes. Thus, minor changes in laboratory test results for the high BMI population may have a strong impact on the development of NAFLD, which deserves more attention.

We selected a representative sample from the internal validation dataset to show the quantitative and individual-specific impact of features and provide the changing process over time. This sample was predicted to develop NAFLD in the future based on his consecutive 5 checkups. But when regardless of his initial 4 checkups and only his last checkup record was included, he would be predicted as exempted from NAFLD, which appears to be a wrong prediction and is contrary to reality. Specifically, almost all features in his fifth checkup seemed not very abnormal and most were within normal limits, but when compared with his fourth checkup, we

can find that there exists a slight increase in BMI, body fat percentage, HDLC, ALT, and many other features. Thus, we can conclude that it was the variation trend contained in the time-series checkup data that prompted the correct prediction. Although in some studies, time-series data were transformed into summary statistics, including mean, minimum and maximum values, to fully harness the information contained in the repeated measurement data [41,42], the temporal trends in detail cannot be captured adequately [43, 44]. Therefore, instead of focusing on snapshot measurements or the transformed statistics of multiple measurements, the original measurement history and long-term trends of some features also need to be taken into account.

It should be noted that our study was subject to the presence of right censoring as well as potential bias due to loss to follow-up, which are inherent limitations associated with utilizing health checkup data. As presented in Supplementary Tab. 12, we report the number and proportion of participants lost to follow-up at each subsequent health checkup after the sixth one. However, we performed a thorough analysis of participant characteristics between those who were lost to follow-up and those who were not at their sixth health checkup, as presented in Supplementary Tab. 13. Our comparison revealed that there were negligible differences between the two groups, as most indicators did not exhibit statistically significant differences ($P > 0.05$), and the standardized mean difference (SMD) results indicated that almost all indicators had an SMD of less than 0.1 [45]. In addition, we also conducted a separate comparison of baseline characteristics between participants lost to follow-up and those who were not, taking into account the outcome (Supplementary Tab. 14). This allowed for a better comparison with Table 1. The analysis revealed that the distribution of most characteristic differences between different outcomes was similar across both the lost to follow-up and not lost to follow-up groups. These findings partly mitigate the potential impact of loss to follow-up on our results.

While our health check-up database contained data spanning 2003–2012, we chose to analyze only those individuals with six or more health checkups. This decision aimed to balance the trade-off between selecting too few or too many years of data, as the former would not fully reflect changes in the time-series data while the latter would result in an unacceptably small sample size. Supplementary Tab. 15 illustrates the relationship between the number of health checkups included and corresponding sample sizes and NAFLD incidence rates. Although selecting patients with seven or more health checkup records would provide richer data, it would reduce the sample size to below 10,000, compromising the statistical power of our deep learning model. Additionally, if the selected time span is too long, the corresponding population may be relatively healthy, with a low incidence rate of NAFLD, and thus the practical value of the model would be limited. Conversely, choosing five or fewer health check-up records would produce a relatively short time-series that might not capture the full spectrum of changes in the underlying data. Therefore, we opted to focus on patients with at least six recorded health checkups to maintain an optimal balance between sample size and data richness.

We own some strengths in this study. Firstly, we made full use of the routine health checkup data and kept their time-series forms to extract the valuable information contained, and used the SHAP algorithm to visualize changing contributions of top predictors and provided the individual-level model explanation. Secondly, we explored the dynamic prediction by incorporating newly updated checkups to adapt to the evolving health checkup picture, which is more useful than a static one and can provide timely guidance to participants. Thirdly, we focused on developing a prognostic prediction model with a time interval of over one year to help identify high-risk groups at early stages, which may own more value compared with most of the current diagnostic prediction models. Fourthly, we verified the predictive performance of our model through external validation on a completely independent dataset, so the generalization and the practical value of the prediction model are in prospect.

There are still several limitations in our study. Firstly, as the high prevalence rate of NAFLD, we filtered out those diagnosed with NAFLD in their initial five checkups and kept a relatively healthier population, which may limit the application of the model. Secondly, we trained the LSTM model with consecutively updated 5 checkups of each participant. Even though our longitudinal data spanned over 5 years, the checkup records hadn't been renewed at a very high frequency. While the strength of the LSTM algorithm lies in its learning about long-term dependencies, which may partially affect our results [46]. Thirdly, our study did not comprehensively investigate the potential differences in missingness patterns between outcome groups and over time during the imputation of missing values, and the use of the LOCF method may not be optimal for our checkup data. Therefore, a more in-depth exploration of missingness imputation techniques is warranted to enhance the robustness of our results. Fourthly, although the MJ Health Screening Center performs calibration of its examination instruments and equipment regularly, the lengthy time frame of our study may still result in natural year-to-year differences in biochemical values. Fifthly, our feature selection process did not consider the longitudinal nature of the data, but instead relied solely on variables obtained during a single checkup, which may partly bias our results. Lastly, despite the relatively high sensitivity exhibited by our prediction model, there are still individuals who are incorrectly predicted as negative, resulting in false negatives. This may be due to the right censoring of data, which could lead to some study participants being diagnosed with NAFLD long after the fifth check-up. Nevertheless, it is essential to recognize that various factors may contribute to the manifestation of false negatives, and their quantification poses a significant challenge.

In conclusion, we developed a dynamic and continuously updated deep learning model based on time-series health checkup data for the prediction of NAFLD in the future checkup and achieved good performance both in calibration and discrimination. Predictors of top significance were identified and their impact on the outcome was visualized both in global settings and individual settings.

Ethics statement

Ethical approval from the Institutional Review Board of Peking University Health Science Center (ID of the approval: IRB00001052-19077) was received by this study. And as all data were de-identified in this study, individual informed consent was waived.

Author contribution statement

Yuhan Deng, Yuan Ma and Jingzhu Fu: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Jianchun Yin, Canqing Yu, Jun Lv: Conceived and designed the experiments; Analyzed and interpreted the data.

Sailimai Man, Bo Wang, Liming Li: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was funded by the Ministry of Science and Technology of the People's Republic of China (grant number 2020YFC2004703); the National Natural Science Foundation of China (grant number 82192901, 82192904, 82192900); and the Ministry of Science and Technology of the People's Republic of China (grant number 2020YFC2003405).

Data availability statement

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e18758.

References

[1] J. Li, B. Zou, Y.H. Yeo, Y. Feng, X. Xie, D.H. Lee, et al., Prevalence, incidence, and outcome of non-alcoholic fatty liver disease in Asia, 1999-2019: a systematic review and meta-analysis, Lancet Gastroenterol Hepatol 4 (5) (2019) 389–398, https://doi.org/10.1016/S2468-1253(19)30039-1.

[2] T. Marjot, A. Moolla, J.F. Cobbold, L. Hodson, J.W. Tomlinson, Nonalcoholic fatty liver disease in adults: current concepts in etiology, outcomes, and management, Endocr. Rev. 41 (1) (2020) bnz009, https://doi.org/10.1210/endrev/bnz009.

[3] J.M. Paik, K. Kabbara, K.E. Eberly, Y. Younossi, L. Henry, Z.M. Younossi, Global burden of NAFLD and chronic liver disease among adolescents and young adults, Hepatology 75 (5) (2022) 1204–1217, https://doi.org/10.1002/hep.32228.

[4] M. Noureddin, F. Ntanios, D. Malhotra, K. Hoover, B. Emir, E. McLeod, 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, Hepatol Commun 6 (7) (2022) 1537–1548, https://doi.org/10.1002/hep4.1935.

[5] X. Ma, C. Yang, K. Liang, B. Sun, W. Jin, L. Chen, et al., A predictive model for the diagnosis of non-alcoholic fatty liver disease based on an integrated machine learning method, Am J Transl Res 13 (11) (2021) 12704–12713.

[6] A. Atsawarungruangkit, P. Laoveeravat, K. Promrat, Machine learning models for predicting non-alcoholic fatty liver disease in the general United States population: NHANES database, World J. Hepatol. 13 (10) (2021) 1417–1427, https://doi.org/10.4254/wjh.v13.i10.1417.

[7] W. Ji, M. Xue, Y. Zhang, H. Yao, Y. Wang, A machine learning based framework to identify and classify non-alcoholic fatty liver disease in a large-scale population, Front. Public Health 10 (2022), 846118, https://doi.org/10.3389/fpubh.2022.846118.

[8] Y.X. Liu, X. Liu, C. Cen, X. Li, J.M. Liu, Z.Y. Ming, et al., Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: an extended study, Hepatobiliary Pancreat. Dis. Int. 20 (5) (2021) 409–415, https://doi.org/10.1016/j.hbpd.2021.08.004.

[9] Æ.Ö. Kristinsson, Y. Gu, S.M. Rasmussen, J. Mølgaard, C. Haahr-Raunkjær, C.S. Meyhoff, E.K. Aasvang, H.B.D. Sørensen, Prediction of serious outcomes based on continuous vital sign monitoring of high-risk patients, Comput. Biol. Med. 147 (2022), 105559, https://doi.org/10.1016/j.compbiomed.2022.105559.

[10] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[11] S. Nemati, A. Holder, F. Razmi, M.D. Stanley, G.D. Clifford, T.G. Buchman, An interpretable machine learning model for accurate prediction of sepsis in the ICU, Crit. Care Med. 46 (4) (2018) 547–553, https://doi.org/10.1097/CCM.0000000000002936.

[12] A.J. London, Artificial intelligence and black-box medical decisions: accuracy versus explainability, Hastings Cent. Rep. 49 (1) (2019) 15–21, https://doi.org/10.1002/hast.973.

[13] W.X. Lim, Z. Chen, A. Ahmed, The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: a review, Med. Biol. Eng. Comput. 60 (3) (2022) 633–642, https://doi.org/10.1007/s11517-021-02487-8.

[14] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Proceedings of the 31st International Conference on Neural Information Processing Systems (2017) 4768–4777.

[15] M.D. Nemesure, M.V. Heinz, R. Huang, N.C. Jacobson, Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence, Sci. Rep. 11 (1) (2021) 1980.

[16] Y. Zoabi, S. Deri-Rozov, N. Shomron, Machine learning-based prediction of COVID-19 diagnosis based on symptoms, NPJ Digit Med 4 (1) (2021) 3.

[17] E. Tasci, Y. Zhuge, K. Camphausen, A.V. Krauze, Bias and class imbalance in oncologic data-towards inclusive and transferrable AI in large scale oncology data sets, Cancers 14 (12) (2022) 2897.

[18] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in Large Margin Classifiers 10 (1999) 61–74.

[19] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[20] S. Qin, X. Hou, Y. Wen, C. Wang, X. Tan, H. Tian, Q. Ao, J. Li, S. Chu, Machine learning classifiers for screening nonalcoholic fatty liver disease in general adults, Sci. Rep. 13 (1) (2023) 3638.

[21] W. Zhu, P. Shi, J. Fu, A. Liang, T. Zheng, X. Wu, S. Yuan, Development and application of a novel model to predict the risk of non-alcoholic fatty liver disease among lean pre-diabetics with normal blood lipid levels, Lipids Health Dis. 21 (1) (2022) 149.

[22] K.W.M. Abeysekera, J.G. Orr, F.H. Gordon, L.D. Howe, J. Hamilton-Shield, J. Heron, M. Hickman, Evaluating future risk of NAFLD in adolescents: a prediction and decision curve analysis, BMC Gastroenterol. 22 (1) (2022) 323.
[23] J. Wang, Y. Tang, K. Peng, H. Liu, J. Xu, Development and validation of a nomogram for predicting nonalcoholic fatty liver disease in the non-obese Chinese population, Am. J. Tourism Res. 12 (10) (2020) 6149–6159.
[24] S. Saokaew, S. Kanchanasuwan, P. Apisarnthanarak, A. Charoensak, P. Charatcharoenwitthaya, P. Phisalprapa, et al., Clinical risk scoring for predicting non-alcoholic fatty liver disease in metabolic syndrome patients (NAFLD-MS score), Liver Int. 37 (10) (2017) 1535–1543.
[25] A.E. Rigamonti, A. Bondesan, E. Rondinelli, S.G. Cella, A. Sartorio, The role of aspartate transaminase to platelet ratio index (APRI) for the prediction of non-alcoholic fatty liver disease (NAFLD) in severely obese children and adolescents, Metabolites 12 (2) (2022) 155.
[26] K.W.M. Abeysekera, J.G. Orr, F.H. Gordon, L.D. Howe, J. Hamilton-Shield, J. Heron, et al., Evaluating future risk of NAFLD in adolescents: a prediction and decision curve analysis, BMC Gastroenterol. 22 (1) (2022) 323.
[27] E. Oliveros, V.K. Somers, O. Sochor, K. Goel, F. Lopez-Jimenez, The concept of normal weight obesity, Prog. Cardiovasc. Dis. 56 (4) (2014) 426–433.
[28] A.G. Mainous 3rd, B.J. Rooks, J.F. Medley, S.B. Dickmann, Body composition among adults at a healthy body mass index and association with undetected non-alcoholic fatty liver, Int. J. Obes. 46 (7) (2022) 1403–1405.
[29] W.R. Kim, S.L. Flamm, A.M. Di Bisceglie, et al., Serum activity of alanine aminotransferase (ALT) as an indicator of health and disease, Hepatology 47 (2008) 1363–1370.
[30] X. Ma, S. Liu, J. Zhang, M. Dong, Y. Wang, M. Wang, et al., Proportion of NAFLD patients with normal ALT value in overall NAFLD patients: a systematic review and meta-analysis, BMC Gastroenterol. 20 (2020) 10.
[31] D.N. Amarapurka, A.D. Amarapurkar, N.D. Patel, et al., Nonalcoholic steatohepatitis (NASH) with diabetes: predictors of liver fibrosis, Ann. Hepatol. 5 (2006) 30–33.
[32] M.T. Long, A. Pedley, L.D. Colantonio, J.M. Massaro, U. Hoffmann, P. Muntner, et al., Development and validation of the framingham steatosis index to identify persons with hepatic steatosis, Clin. Gastroenterol. Hepatol. 14 (2016) 1172–1180.
[33] H. Yang, D. Li, X. Song, F. Liu, X. Wang, Q. Ma, et al., Joint associations of serum uric acid and ALT with NAFLD in elderly men and women: a Chinese cross-sectional study, J. Transl. Med. 16 (1) (2018) 285.
[34] Y. Li, C. Xu, C. Yu, L. Xu, M. Miao, Association of serum uric acid level with non-alcoholic fatty liver disease: a cross-sectional study, J. Hepatol. 50 (5) (2009) 1029–1034.
[35] A. Lonardo, P. Loria, F. Leonardi, A. Borsatti, P. Neri, M. Pulvirenti, et al., Fasting insulin and uric acid levels but not indices of iron metabolism are independent predictors of non-alcoholic fatty liver disease. A case-control study, Dig. Liver Dis. 34 (3) (2002) 204–211.
[36] Y.J. Choi, H.S. Shin, H.S. Choi, J.W. Park, I. Jo, et al., Uric acid induces fat accumulation via generation of endoplasmic reticulum stress and SREBP-1c activation in hepatocytes, Lab. Invest. 94 (10) (2014) 1114–1125.
[37] A. Atsawarungruangkit, P. Laoveeravat, K. Promrat, Machine learning models for predicting non-alcoholic fatty liver disease in the general United States population: NHANES database, World J. Hepatol. 13 (10) (2021) 1417–1427.
[38] E. Molina-Molina, G.E. Furtado, J.G. Jones, P. Portincasa, A. Vieira-Pedrosa, A.M. Teixeira, et al., The advantages of physical exercise as a preventive strategy against NAFLD in postmenopausal women, Eur. J. Clin. Invest. 52 (3) (2022), e13731.
[39] S. Ebrahimi Mousavi, N. Dehghanseresht, F. Dashti, Y. Khazaei, S. Salamat, et al., The association between Dietary Diversity Score and odds of nonalcoholic fatty liver disease: a case-control study, Eur. J. Gastroenterol. Hepatol. 34 (6) (2022) 678–685.
[40] J.H. Lee, H.S. Lee, S.B. Ahn, Y.J. Kwon, Dairy protein intake is inversely related to development of non-alcoholic fatty liver disease, Clin Nutr 40 (10) (2021) 5252–5260.
[41] Z. Zeng, S. Yao, J. Zheng, X. Gong, Development and validation of a novel blending machine learning model for hospital mortality prediction in ICU patients with Sepsis, BioData Min. 14 (1) (2021 Aug 16) 40.
[42] Y. Zhu, J. Zhang, G. Wang, R. Yao, C. Ren, G. Chen, et al., Machine learning prediction models for mechanically ventilated patients: analyses of the MIMIC-III database, Front. Med. 8 (2021), 662340.
[43] J. Lopez Bernal, S. Soumerai, A. Gasparrini, A methodological framework for model selection in interrupted time series studies, J. Clin. Epidemiol. 103 (2018) 82–91.
[44] Y. Xue, D. Klabjan, Y. Luo, Predicting ICU readmission using grouped physiological and medication trends, Artif. Intell. Med. 95 (2019) 27–37.
[45] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, second ed., Lawrence Erlbaum Associates, Hillsdale, *NJ*, 1988, pp. 1–17.
[46] I. Gandin, A. Scagnetto, S. Romani, G. Barbati, Interpretability of time-series deep learning models: a study in cardiovascular patients admitted to Intensive care unit, J. Biomed. Inf. 121 (2021), 103876.