Progress in Medical Devices
ISSN: 2957-5478

Zentime

Research Article https://doi.org/10.61189/568091unpkqk

Construction and comparative analysis of an early screening prediction model for fatty liver in elderly patients based on machine learning

Xiaolei Cai[1*], Qi Sun[2*], Cen Qiu[2*], Zhenyu Xie[1], Jiahao He[2], Mengting Tu[3], Xinran Zhang[2], Yang Liu[2], Zhaojun Tan[2], Yutong Xie[2], Xixuan He[1], Yujing Ren[1], Chunhong Xue[1], Siqi Wang[2], Linrong Yuan[2], Miao Yu[2], Xuelin Cheng[4], Xiaopan Li[4], Sunfang Jiang[4], Huirong Zhu[1]

[1]Tangqiao Community Health Service Center, Shanghai 200127, China. [2]Shanghai University of Medicine and Health Sciences, Shanghai 201318, China. [3]Shanghai DianJi University, Shanghai 201306, China. [4]Health Management Center, Zhongshan Hospital Affiliated to Fudan University, Shanghai 200032, China.
*The authors contribute equally.

Corresponding authors: Sunfang Jiang and Huirong Zhu.

Acknowledgement: None.

Declaration of conflict of interest: None.

Declaration of Ethics: This article does not contain any studies with human participants or animals performed by any of the authors. Ethical approval for this study was obtained from Tangqiao Community Health Service Center.

Received May 11, 2024; Accepted July 16, 2024; Published September 30, 2024

Highlights

- This study collected three years of physical examination data from older adults in the Tangqiao community of Shanghai, which is more regionally representative.
- The most suitable model for this study was selected from six machine learning models to construct a fatty liver risk prediction model for the elderly.
- This study combines six feature selection algorithms with varying performance to screen the features most rele vant to fatty liver.

Abstract

Objective: To construct a prediction model for fatty liver disease (FLD) among elderly residents in community using machine learning (ML) algorithms and evaluate its effectiveness. **Methods:** The physical examination data of 4989 elderly people (aged over 60 years) in a street of Shanghai from 2019 to 2023 were collected. The subjects were divided into a training set and a testing set in a 7:3 ratio. Using feature selection and importance sorting methods, eight indicators were selected, including high-density lipoprotein cholesterol, body mass index, uric acid, triglycerides, albumin, red blood cell, white blood cell, and alanine aminotransferase. Six ML models, including Categorical Features Gradient Boosting, eXtreme Gradient Boosting, Light Gradient Boosting Machine, Random Forest, Decision Tree, and Logistic Regression, were constricted, and their predictive performances were compared via accuracy, precision, recall, F1 score, and Area Under Receiver Operating Characteristic Curve. **Results:** Among the six ML models, the Categorical Features Gradient Boosting model demonstrated the highest prediction accuracy of 0.74 for FLD in elderly community population, along with a precision of 0.70, a recall of 0.73, a F1 score of 0.71, and an area under the curve of 0.74. **Conclusions:** In the context of rapid development of artificial intelligence, a community-based elderly FLD prediction model constructed using ML algorithms aid family general practitioners in the early diagnosis, early treatment, and health management of local FLD patients.

Keywords: Fatty liver, machine learning models, disease screening, health management, community diagnosis

Address correspondence to: Sunfang Jiang, Health Management Center, Zhongshan Hospital Affiliated to Fudan University, Gate 5 East Campus, No. 179 Fenglin Road, Xuhui District, Shanghai 200032, China. Email: jiang.sunfang@zs-hospital.sh.cn. **Huirong Zhu,** Tangqiao Community Health Service Center, No.131 Pujian Road, Pudong New District, Shanghai 200127, China. Email: rachel1022@126.com.

Introduction

Fatty Liver Disease (FLD) is a globally prevalent disease. With changes in lifestyle and population demographics, as well as factors such as obesity and excessive alcohol consumption, FLD has become the leading chronic liver disease in China, and its prevalence is on the rise, posing a significant threat to people's health, particularly that of the elderly [1]. FLD can be diagnosed through liver biopsy or imaging techniques. Liver biopsy is the gold standard for diagnosing steatosis but associated with invasiveness and bleeding risk [2]. Meanwhile, imaging methods such as computed tomography and magnetic resonance imaging are time-consuming and expensive and are not always available in remote areas. Early diagnosis of FLD based on risk factors helps clinicians identify adverse events related to FLD and take more lifestyle intervention measures for its prevention [3]. Machine learning (ML) is a promising technique for medical data analysis and disease prediction, making significant contributions to the field of healthcare. For instance, Jiang et al. used ML to comprehensively describe the overall decline in intrinsic capabilities among elderly people in communities [4]; Islam et al. developed an efficient, interpretable ML framework for predicting the risk of hypertension in Ethiopian patients [5]; Abnoosian et al. developed a novel ML framework to enhance the performance of diabetes prediction models and address various challenges [6]. It is evident that ML is playing an increasingly important role in various fields, particularly in healthcare. By analyzing big data, building models, and making automated decisions, ML offers more efficient and accurate solutions for the healthcare industry. This study aims to propose a ML model to predict FLD, helping doctors to preliminarily screen high-risk populations and make early predictions for FLD.

Materials and methods

Data preparation

The overall workflow of this study is shown in **Figure 1**. The dataset used was provided by a community health service center on a street in Shanghai, containing health examination data for the elderly over three years (2019, 2022, 2023), with a total of 4,989 entries. All individuals were aged 60 and above, and the data was anonymized. The dataset includes 24 variables such as gender, blood biochemical indicators, and age, with the distribution of 23 continuous variables shown in **Figure 2**. To preserve the authenticity of the data, 1,144 entries with missing values were removed. In this study, FLD diagnosis was based on ultrasound results. Ultimately, out of the 3,845 participants in the study, 2,074 were potentially diagnosed with FLD through ultrasound.

Statistical analysis

Data analysis was conducted using Python 3.11.4 software. The continuous variables were expressed as mean ± SD, and the comparison between the FLD group and non-FLD group was conducted using an independent sample t-test. A p-value less than 0.05 were considered statistically significant.

Data processing

In ML, most algorithms, such as Logistic regression (LR), Support Vector Machines, and K-nearest neighbors, can only process numerical data and are unable to handle text. In such cases, to make data compatible with algorithms and libraries, encoding is performed, which involves converting textual data into numerical data.

In this study, the dataset consisted of 25 columns, with data types including only numerical and categorical variables. For numerical variables, since the collected data came with units, these units were removed, and the data type was converted to float. For categorical variables, one hot encoding was used for gender, transforming the feature into dummy variables, with "female" marked as 1 and "male" marked as 0; for ultrasound results, label encoding was used, with "fatty liver" marked as 1 and "non-fatty liver" marked as 0.

Feature selection

The dataset included 24 clinical features: age, sex, body mass index (BMI), white blood cell (WBC), red blood cell (RBC), hemoglobin (Hb), serum creatinine (SCR), mean corpuscular volume (MCV), platelets (PLT), triglycerides (TG), mean corpuscular hemoglobin concentration (MCHC), direct bilirubin (DBIL), aspartate aminotransferase (AST), alanine aminotransferase (ALT), AST/ALT, albumin-globulin ratio (A/G), total cholesterol (TC), uric acid (UA), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), albumin (ALB), globulin (GLB), total protein (TP), and total bilirubin (TBIL). Feature selection aims to remove irrelevant or redundant features while retaining other original features to obtain a subset that better describes the given problem with minimal performance loss. This study employed

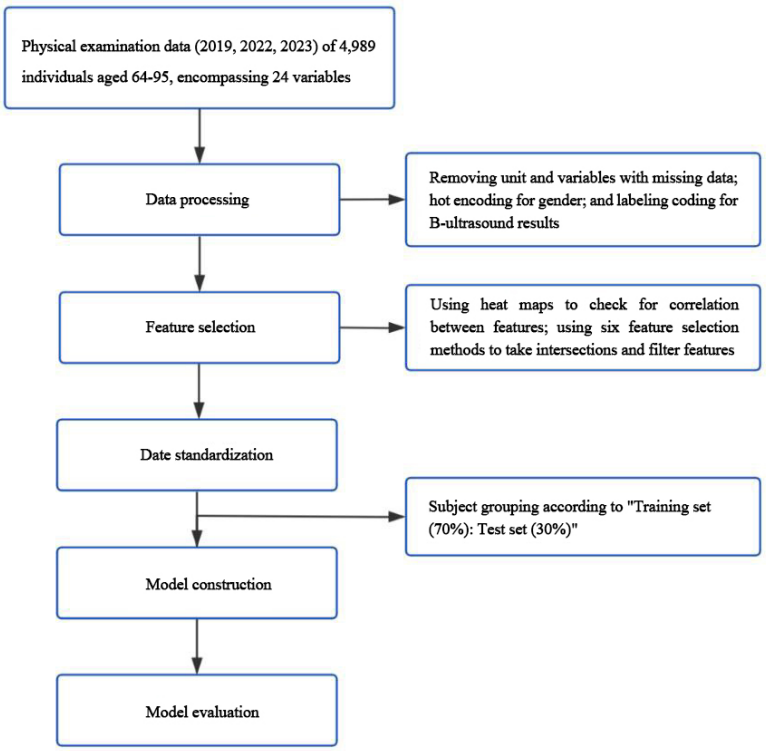

Figure 1. Workflow of the ML-based FLD prediction method proposed in this study. ML, machine learning; FLD, fatty liver disease.

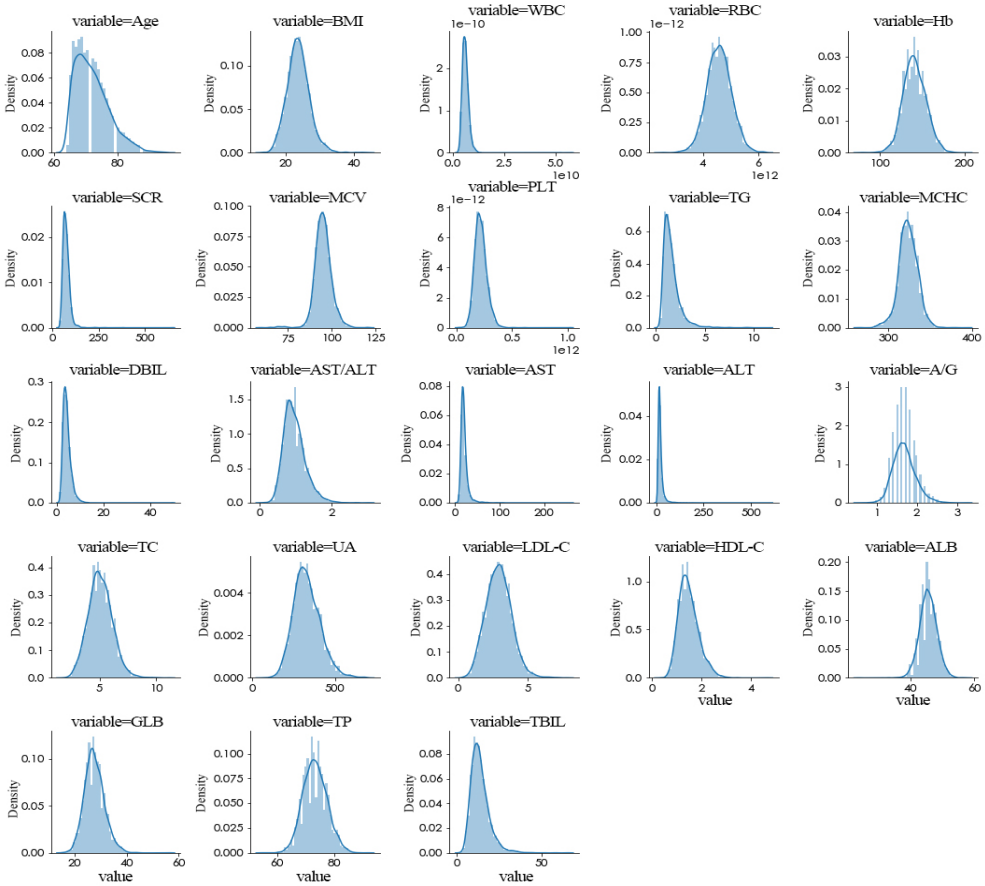

Figure 2. Distribution graphs of various continuous variables. BMI, body mass index; WBC, white blood cell; RBC, red blood cell; Hb, hemoglobin; SCR, serum creatinine; MCV, mean corpuscular volume; PLT, platelets; TG, triglycerides; MCHC, mean corpuscular hemoglobin concentration; DBIL, direct bilirubin; AST, aspartate aminotransferase; ALT, alanine aminotransferase; A/G, albumin-globulin ratio; TC, total cholesterol; UA, uric acid; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; ALB, albumin; GLB, globulins; TP, total protein; TBIL, total bilirubin.

six typical feature selection methods: Variance Threshold, Correlation Coefficient, Chi-square test, Mutual Information, Tree-based Feature Selection, and Recursive Feature Elimination.

Variance Threshold selects features based on their variance, by default removing all zero-variance features. The Correlation Coefficient method (Pearson correlation coefficient) calculates the correlation coefficient between features and the target variable, selecting features with a significant correlation with the target. The Chi-Square test evaluates the correlation between features and the target variable based on the chi-square statistic, which can reduce dimensionality and improve model performance. Mutual Information selects features with high mutual information, suitable for both continuous and discrete variables. Tree-based Feature Selection considers interactions among features and can handle nonlinear features and high-dimensional data, appropriate for datasets with complex relationships among features. Recursive Feature Elimination progressively removes features based on model coefficients or feature importance, selecting the most significant features. After comparing six feature subsets, this study chose the eight most commonly used features in clinical practice to construct the prediction model.

Model construction and validation

All processing steps were performed in Python 3.11.4. After feature selection, data normalization was conducted using the StandardScaler from the sklearn.preprocessing module. The selected 8 features were used as independent variables, with the FLD label column as the dependent variable y. The dataset was split into training and testing sets in a 7:3 ratio using train test split, followed by the construction of ML models. This study primarily built six widely used ML models and optimized their performance. For the Categorical Features Gradient Boosting (CatBoost), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) models, the study employed RandomizedSearchCV for random search, which selects a subset of parameter combinations from a preset parameter space and a preset number of trials for training and validation [7-9]. This method returns the best-performing parameter combination from the trials, significantly reducing search time and computational resources [10].

For the Decision Tree (DT), Random Forest (RF), and LR models, the study used GridSearchCV for parameter tuning. GridSearchCV is a hyperparameter optimization tool that automatically adjusts parameters of ML models. It performs grid search by accepting a dictionary of all possible parameter combinations passed to the 'param_grid' parameter. It also requires an integer to be passed to the 'cv' parameter, indicating the number of cross-validation iterations [11].

Performance metrics

The performance of the six ML models used was evaluated using five widely used metrics: accuracy, precision, recall, F1-score, and the area under the curve (AUC). True Positive (TP) indicates that the patient has FLD and the algorithm correctly predicts it as FLD. True Negative (TN) indicates that the patient does not have FLD and the algorithm correctly predicts it as non-FLD. False Negative (FN) indicates that the patient has FLD, but the algorithm predicts it as non-FLD, mistakenly predicting an actual positive as negative. False Positive (FP) indicates that the patient does not have FLD, but the algorithm predicts it as FLD, mistakenly predicting an actual negative as positive.

Accuracy: The proportion of correctly predicted samples out of the total number of samples.

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)*100\% \quad (1)$$

Precision: The proportion of true positive samples among all samples predicted as positive (TP and FP), reflecting the model's ability to identify positive samples.

$$\text{Precision} = TP/(TP+FP)*100\% \quad (2)$$

Recall: The proportion of true positive samples among all actual positive samples (TP and FN), reflecting the model's ability to detect positive samples.

$$\text{Recall} = TP/(TP+FN)*100\% \quad (3)$$

F1-score: The harmonic mean of precision and recall.

$$\text{F1-score} = 2TP/(2TP+FP+FN)*100\% \quad (4)$$

Receiver Operating Characteristic Curve (ROC): the area under the ROC is an important metric for assessing model performance. The ROC curve plots the True Positive Rate on the y-axis and the False Positive Rate on the x-axis, showing the model's performance at different thresholds. The AUC is obtained by calculating the area enclosed by the ROC curve, the x-axis, and the False Positive Rate =1. The closer the

Table 1. Feature selection outcomes of six feature selection methods

	Variance threshold	Correlation coefficient	Chi-square test	Mutual information	Tree-based feature selection	Recursive feature elimination
Age	√	√	√	√		
Sex	√			√		
BMI	√	√	√	√	√	√
WBC	√	√	√	√	√	√
RBC	√	√	√	√	√	√
Hb	√	√	√	√	√	
SCR	√	√	√			√
MCV	√			√		√
PLT	√	√	√	√		√
TG	√	√	√	√	√	√
MCHC	√			√		√
DBIL	√					
AST/ALT		√	√	√	√	√
AST	√	√	√		√	
ALT	√	√	√	√	√	√
A/G						
TC	√					
UA	√	√	√	√	√	√
LDL-C	√					√
HDL-C	√	√	√	√	√	√
ALB	√	√	√	√	√	√
GLB	√			√		√
TP	√	√	√		√	
TBIL	√	√	√			√

Note: BMI, body mass index; WBC, white blood cell; RBC, red blood cell; Hb, hemoglobin; SCR, serum creatinine; MCV, mean corpuscular volume; PLT, platelets; TG, triglycerides; MCHC, mean corpuscular hemoglobin concentration; DBIL, direct bilirubin; AST, aspartate aminotransferase; ALT, alanine aminotransferase; A/G, albumin-globulin ratio; TC, total cholesterol; UA, uric acid; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; ALB, albumin; GLB, globulins; TP, total protein; TBIL, total bilirubin.

Table 2. Evaluation metrics for six models

ML model	Accuracy	Precision	Recall	F1-score	AUC
RF	0.73	0.68	0.73	0.70	0.73
DT	0.70	0.68	0.60	0.64	0.69
LR	0.74	0.70	0.71	0.70	0.73
CatBoost	0.74	0.70	0.73	0.71	0.74
XGBoost	0.73	0.67	0.72	0.70	0.72
LightGBM	0.73	0.68	0.74	0.71	0.73

Note: ML, machine learning; AUC, area under the curve; RF, random forest; DT, decision tree; LR, logistic regression; CatBoost, categorical features gradient boosting; XGBoost, extreme gradient boosting; LightGBM, light gradient boosting machine.

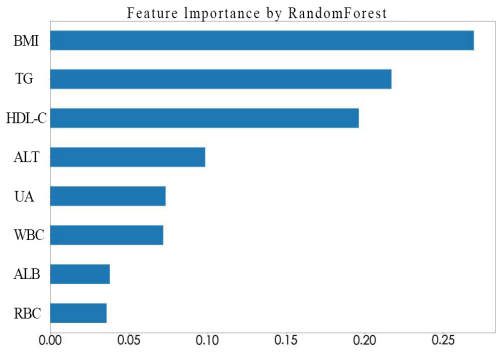

Figure 3. Feature importance ranking based on RF. RF, Random Forest; BMI, body mass index; TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; ALT, alanine aminotransferase; UA, uric acid; WBC, white blood cell; ALB, albumin; RBC, red blood cell.

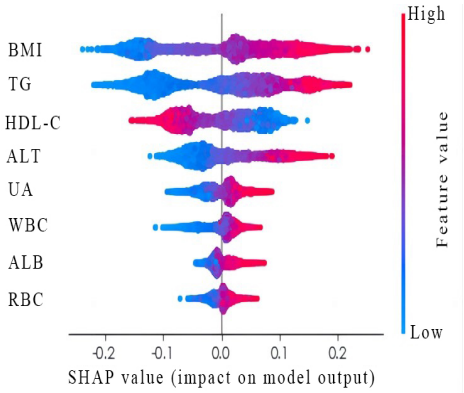

Figure 4. SHAP summary plot. SHAP, shapley additive explanations; BMI, body mass index; TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; ALT, alanine aminotransferase; UA, uric acid; WBC, white blood cell; ALB, albumin; RBC, red blood cell.

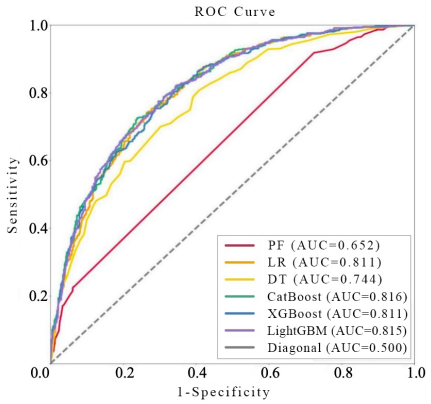

Figure 5. ROC curves of six models. ROC, receiver operating characteristic; RF, random forest; LR, logistic regression; DT, decision tree; CatBoost, categorical features gradient boosting; XGBoost, extreme gradient boosting; LightGBM, light gradient boosting machine.cell.

AUC is to 1, the better the model's classification performance.

Results

Selected features

Table 1 contains six types of feature selection algorithms, each with its unique selection mechanism and evaluation criteria. It details the specific circumstances under which these feature selection algorithms were used to select features. Ultimately, these selected features were used to construct an FLD risk assessment model, ensuring that the model could perform accurate risk assessments based on the most relevant and predictive features.

By taking the intersection of the six feature selection methods, the eight most meaningful features were identified: HDL-C, BMI, UA, TG, ALB, RBC, WBC, and ALT. Their importance was ranked (**Figure 3**), which illustrates the contribution of each data feature.

Shapley additive explanations (SHAP) is a Python package developed to explain the output of any ML model. All features are considered 'contributors' [12]. The 'summary plot' visualizes the SHAP values for each feature of each sample, providing an intuitive understanding of the overall pattern and helping to identify potential prediction outliers. Each row represents a feature, with the x-axis showing the SHAP values. Each point represents a sample, where the color indicates the feature value (red for high, blue for low), and purple near the mean. The wider the color area, the greater the impact of the feature. As shown in **Figure 4**, higher BMI values increased the predicted risk for FLD, whereas higher HDL-C values decreased the predicted risk for FLD.

Results of performance metrics across different models

Table 2 shows the performance scores for five metrics across six ML models. LR and CatBoost were comparable in terms of accuracy and precision, but CatBoost performed better in the remaining three metrics. Overall, the CatBoost model demonstrated the best performance.

The ROC curves for the six models are shown in **Figure 5**. CatBoost exhibited the highest AUC value at 0.816, followed closely by LightGBM at 0.815. XGBoost and LR were comparable, both with an AUC of 0.811, while RF and DT performed less favorably compared to the other models.

Discussion

The rising incidence of FLD, a prevalent liver disease, has attracted widespread attention from the medical community and society [13, 14]. Early identification and prediction of fatty liver risk is crucial for developing effective prevention and treatment strategies. In recent years, with the rapid development of big data and ML technologies, health risk assessment using these advanced technologies has become a new trend in research [15, 16]. In particular, by mining and analysing key indicators in physical examination data, it can provide a scientific basis for the prediction of fatty liver.

Focusing on the old three-year physical examination data in Tangqiao community, this study aims to predict the risk of fatty liver through ML algorithms, especially CatBoost model.

This study identified eight features through six feature selection methods: BMI, HDL-C, UA, TG, ALB, RBC, WBC, and ALT. Previous studies have highlighted various clinical features associated with FLD. For instance, age, BMI, waist circumference, ALT, fasting glucose, UA, and PLT have been reported as related to FLD [17]. Weng et al. developed an ML model to screen for high-risk populations of FLD, identifying some meaningful clinical features: BMI, ALB, ALT, GB, HDL-C, LDL-C, and TG [18]. Chen et al. used a large dataset and the XGBoost model to investigate the best overall predictive ability for FLD in the diagnosed population, additionally identifying features such as BMI, waist circumference, TG, serum γ-glutamyl transferase, glucose, age, creatinine, gender, and LDL-C as the significant factors [19]. Pei et al. conducted feature importance analysis based on the ranking of each variable, determining that serum UA, BMI, and TC are the three most likely factors, followed by HDL-C and Hb [20]. Su et al. chose waist circumference, BMI, blood pressure, WBC, Hb, PLT, glucose, serum biochemical parameters, and lipids [3]. Guarneros et al., by applying ML techniques, identified the main risk factors for FLD in four clinical datasets of liver disease patients: ALT, alkaline phosphatase, AST, alpha-fetoprotein, and γ-glutamyl transferase [21]. Hence, the features selected in this study hold significant relevance in clinical research on FLD.

CatBoost is a gradient boosting DT-based ML algorithm designed to enhance model performance through ensemble learning and feature handling [22]. Known for its robustness and generalization capabilities, the CatBoost model can automatically adjust learning rates and iteration numbers during training to prevent overfitting. Additionally, CatBoost incorporates randomization strategies to enhance model diversity, thereby improving generalization performance. This study concludes that the CatBoost model has the best overall predictive ability in the preliminary screening of FLD populations.

In the early detection of FLD, the CatBoost algorithm can be used to construct a predictive model to identify potential FLD cases. Through training on a large FLD dataset, the CatBoost model learns to predict the presence of FLD in new samples with high accuracy and sensitivity, effectively detecting early stages of the disease. The algorithm and model obtained in this study offer technical tools for community health management, particularly benefiting residents who may not routinely undergo auxiliary examinations such as ultrasound but receive blood tests during clinical visits. The community health service center can leverage this algorithm and model to intelligently assess the health of these residents, offering early warnings, predictions, and health recommendations for FLD. This supports early diagnosis and treatment by prompting at-risk residents to seek further evaluation and appropriate care.

To improve the model developed in this study, future research should collect more detailed information about patients and further refine the survey factors when conducting statistical analysis of the examined population. Parameters such as waist circumference and γ-glutamyl transferase can also be considered as features to further enhance such ML models by providing more information about the likelihood of disease occurrence and the degree of steatosis and fibrosis.

This study also has certain limitations. First, the patient grouping was based on abdominal ultrasound diagnostic results, which have a lower level of evidence compared to liver biopsy and MRI. This may affect the accuracy of the predictions. Second, the data came from only one street community, and a multicenter dataset and external validation could provide better and more reliable performance. Additionally, this study did not include some clinical parameters, such as waist circumference and γ-glutamyl transferase, which have been proven to be related to FLD. Finally, the dataset lacked patients' clinical information. The presence of diseases such as diabetes, hepatitis B, hepatitis C, and the medications used may affect the predictions of the ML model.

Conclusion

In this study, eight features closely related to

fatty liver risk were successfully screened by CatBoost algorithm and six feature selection methods. These features are not only widely recognised in clinical studies, but their validity in fatty liver prediction is further demonstrated by the validation of this study. Therefore, this study is innovative and practical in the field of fatty liver risk prediction, and provides a scientific basis for the prevention and intervention of fatty liver through the combined use of advanced ML algorithms and abundant clinical data.

Supplementary Material: None.

Availability of data and material: None.

Author Contributions: Xiaolei Cai, Qi Sun, Cen Qiu (Lead Author): Led the overall conception and design of the study. Conducted the literature review and synthesized relevant findings to inform the research questions and methodology. Drafted the initial manuscript and incorporated feedback from co-authors to refine the final version.

Qi Sun, Cen Qiu, Jiahao He, Yutong Xie, Siqi Wang, Miao Yu (Data Analyst): Contributed significantly to the data preprocessing and feature selection phases of the study. Developed and implemented the feature selection methods to identify the eight key features for the CatBoost model. Contributed to the writing of the Methods and Results sections, focusing on the technical aspects of the study.

Xiaolei Cai, Sunfang Jiang, Huirong Zhu, Zhenyu Xie, Mengting Tu, Xinran Zhang, Yang Liu (Clinical Expert): Provided critical clinical input throughout the study, ensuring the research questions and methodology were grounded in clinical practice. Reviewed and interpreted the results from a clinical perspective, highlighting their relevance and implications for public health and clinical practice. Contributed to the Discussion section, drawing connections between the findings and existing clinical literature.

Yujing Ren, Chunhong Xue, Xuelin Cheng, Xiaopan Li, Zhaojun Tan, Linrong Yuan (Statistical Advisor): Offered guidance on statistical analysis and modeling techniques. Ensured the rigorous application of machine learning methods, particularly the CatBoost algorithm. Assisted in the interpretation of the results and provided recommendations for enhancing the robustness and reproducibility of the findings. Contributed to the Methods section, emphasizing the statistical rigor of the study.

References

[1] Teng ML, Ng CH, Huang DQ, et al. Global incidence and prevalence of nonalcoholic fatty liver disease. Clin Mol Hepatol 2023;29(Suppl):S32-S42.

[2] Chalasani N, Younossi Z, Lavine JE, et al. The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatology 2018;67(1):328-357.

[3] Su PY, Chen YY, Lin CY, et al. Comparison of Machine Learning Models and the Fatty Liver Index in Predicting Lean Fatty Liver. Diagnostics (Basel) 2023;13(8):1407.

[4] Jiang X, Chen F, Yang X, et al. Effects of personal and health characteristics on the intrinsic capacity of older adults in the community: a cross-sectional study using the healthy aging framework. BMC Geriatr 2023;23(1):643.

[5] Islam MM, Alam MJ, Maniruzzaman M, et al. Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia. PLoS One 2023;18(8):e0289613.

[6] Abnoosian K, Farnoosh R, Behzadi MH. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics 2023;24(1):337.

[7] Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv 2018.

[8] Yang G. Generalized XGBoost Method. arXiv 2022.

[9] Alafate J, Freund Y. Faster Boosting with Smaller Memory. arXiv 2019.

[10] Anghel A, Papandreou N, Parnell T, et al. Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms. arXiv 2019.

[11] Ahmad GN, Fatima H, Ullah S, et al. Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV. IEEE Access 2022;10:80151-80173.

[12] Amoukou SI, Brunel NJB, Salaün T. Accurate Shapley Values for explaining tree-based models. arXiv 2023.

[13] Liu K, Chen Y, Chen J, et al. Genetically determined circulating micronutrients and the risk of nonalcoholic fatty liver disease. Sci Rep 2024;14(1):1105.

[14] Younossi ZM, Golabi P, Paik JM, et al. The global epidemiology of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH): a systematic review. Hepatology 2023;77(4):1335-1347.

[15] Zare A, Shafaei Bajestani N, Khandehroo M. Machine Learning in Public Health. J Res Health 2024;14(3):207-208.
[16] Anuntakarun S, Khamjerm J, Tangkijvanich P, et al. Classification of Long Non-Coding RNAs s Between Early and Late Stage of Liver Cancers From Non-coding RNA Profiles Using Machine-Learning Approach. Bioinform Biol Insights 2024;18:11779322241258586.
[17] Gong XY, Zhang LF, Ge R, et al. Analysis of risk factors of metabolic fatty liver disease based on community physical examination data. J Community Med 2023;21(21):1106-1111.
[18] Weng S, Hu D, Chen J, et al. Prediction of Fatty Liver Disease in a Chinese Population Using Machine-Learning Algorithms. Diagnostics (Basel) 2023;13(6):1168.
[19] Chen YY, Lin CY, Yen HH, et al. Machine-Learning Algorithm for Predicting Fatty Liver Disease in a Taiwanese Population. J Pers Med 2022;12(7):1026.
[20] Pei X, Deng Q, Liu Z, et al. Machine Learning Algorithms for Predicting Fatty Liver Disease. Ann Nutr Metab 2021;77(1):38-45.
[21] Guarneros-Nolasco LR, Alor-Hernández G, Prieto-Avalos G, et al. Early Identification of Risk Factors in Non-Alcoholic Fatty Liver Disease (NAFLD) Using Machine Learning. Mathematics 2023;11(13):3026.
[22] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7(1):94.