·医学工程·

基于机器学习构建脂肪肝预测模型

宋莹莹[1] 张展召[2]

（1.新乡医学院，新乡453003；2.新乡医学院三全学院，新乡453003）

【摘要】 目的 本研究旨在利用机器学习技术构建一种高效的脂肪肝预测模型，以帮助医务人员准确识别和分类脂肪肝高危人群，从而为早期干预和个性化治疗提供支持。**方法** 研究数据来自2017—2020年的美国国家和健康调查（National Health and Nutrition Examination Survey，NHANES）数据库。最终纳入11个变量进行分析。采用6种机器学习算法决策树（Decision Tree，DT）、极端梯度提算（Extreme Gradient Boosting，XGBoost）、自适应提升算法（Adaptive Boosting，Adaboost）、K最邻近（K–Nearest Neighbor，KNN）、Logistic回归、人工神经网络（Artificial neural network，ANN）模型构建脂肪肝预测模型，采用随机生成的70%作为训练集，剩下的30%作为测试集。模型评估指标包括精确率、准确率、召回率、F1值、受试者工作特征曲线（Receiver Operating Characteristic curve，ROC）及曲线下面积（Area Under Curve，AUC）。**结果** 共纳入3383名受试者，其中脂肪肝患者1506例，DT、XGBoost、AdaBoost、KNN、Logistic回归、ANN模型的AUC分别为0.98838、0.94624、0.88935、0.99110、0.90047、0.98990。除了KNN模型外，其余所有模型的校准曲线均表现良好，其中AdaBoost模型在六个模型中展现出最佳的预测性能。**结论** AdaBoost模型被认为是最优的预测模型，适用于脂肪肝的预测，并能够解释特征变量之间的交互关系。体质指数、甘油三酯和低密度脂蛋白指标能够有效识别脂肪肝的高危人群。

【关键词】 脂肪肝；机器学习；模型预测；自适应提升算法；受试者工作特征曲线

【中图分类号】 TH77; R575.5 **【文献标识码】** A **【DOI】** 10.11876/mimt202501008

Constructing a fatty liver prediction model based on machine learning

Song Yingying[1], Zhang Zhanzhao[2]

(1. Xinxiang Medical University, Xinxiang 453003, China; 2. Sanquan College of Xinxiang Medical University, Xinxiang 453003, China)

【Abstract】 Objective This study aims to use machine learning technology to construct an efficient fatty liver prediction model to help medical personnel accurately identify and classify high–risk populations for fatty liver, thereby providing support for early intervention and personalized treatment. **Methods** The research data was sourced from the National Health and Nutrition Examination Survey (NHANES) database in the United States from 2017 to 2020. Finally, 11 variables were included for analysis. Using 6 machine learning algorithms, Decision Tree (DT) Extreme Gradient Boosting (XGBoost), Adaptive Boosting (Adaboost), K–Nearest Neighbor (KNN), Logistic Regression, and Artificial Neural Network (ANN) models are used to construct a fatty liver prediction model. The randomly generated 70% is used as the training set, and the remaining 30% is used as the testing set. The model evaluation metrics include accuracy, precision, recall, F1 score, Receiver Operating Characteristic curve (ROC), and Area Under Curve (AUC). **Results** A total of 3383 participants were included, including 1506 patients with fatty liver. The decision tree The AUC of XGBoost, AdaBoost, KNN, logistic regression, and ANN models are 0.98838, 0.94624, 0.88935, 0.99110, 0.90047, and 0.98990, respectively. Except for the KNN model, the calibration curves of all other models performed well, with the AdaBoost model showing the best predictive performance among the six models. **Conclusion** The AdaBoost model is considered the optimal

基金项目：河南省2022年省级专创融合特色示范课课程（195）

第一作者：宋莹莹，女，本科，助教，研究方向：机器学习与图像处理

predictive model for predicting fatty liver and can explain the interaction between feature variables. Body mass index, triglycerides, and low-density lipoprotein indicators can effectively identify high-risk populations for fatty liver.

【Key words】 Fatty liver; Machine learning; Model prediction; Adaptive boosting; Receiver operating characteristic curve

随着现代社会经济的快速发展以及人们生活方式和饮食结构的持续变化，脂肪肝（Fatty Liver Disease，FLD）的发病率和检出率已超越病毒性肝炎，成为肝脏疾病中最常见的[1,2]，并可导致肝硬化和肝细胞癌的非胆汁淤积性[3,4]。FLD在过去几十年中变得越来越普遍，目前世界上超过33%的人口受到影响[3]。我国FLD的患病率达到29.2%（95% CI: 27.7% ~ 30.7%）[5]，并且这一比例正逐年上升，且存在明显的地区差异。FLD一种非传染性的肝脏代谢性疾病，其诱发因素多种多样，包括肥胖、营养不良以及酒精摄入等[6]。此外，随着病情的进展，FLD患者极易并发心血管疾病和消化系统疾病[7]，导致预后不良。FLD对身体健康造成严重威胁，影响患者的日常生活，给患者及其家庭带来了沉重的生活负担，因此，预防和控制FLD的形势十分严峻。FLD作为一种常见的肝脏疾病，其患病率的上升不仅对患者的健康构成威胁，也带来了显著的经济负担。因此，准确识别高风险个体和早期诊断FLD对于降低这一经济负担至关重要。早期识别FLD患者可以为诊断、预防和适当治疗提供巨大的好处。通过及时的干预，患者可以采取生活方式的改变，降低疾病进展的风险，避免更严重的健康问题（如肝硬化）及其相关的高昂治疗费用。此外，早期识别能给医生在制定个性化的治疗方案中提供帮助，可能会使患者的生活质量提高。在过去的十年中，肝活检被广泛用于对FLD患者进行分层，并被认为是评估肝脏脂肪浸润的“金标准”[8,9]。但是，肝活检是一种可能会对机体产生损伤的检测方式，如出血、感染和疼痛等结果的出现[8]。此外，活检的结果可能受到采样误差的影响，特别是在肝脏脂肪分布不均的情况下，这可能导致误诊或漏诊[9]。所以肝活检虽然被视为诊断的“金标准”，但超声检查作为一种非侵入性的方法，已被广泛应用于FLD诊断。超声检查具有较高的准确性和安全性，能够有效评估肝脏脂肪的积聚程度。但是，超声检查对操作员有较高的要求，例如需要操作员有丰富的经验等。不同的操作员可能会对同一患者的超声结果产生不同的解读，这可能影响诊断的一致性。

机器学习（Machine Learning，ML）是计算机科学的一个重要的分支，是通过使用各种机器学习模型从海量数据中挖掘出各种结果的预测。随着数据量的激增，机器学习技术已逐渐成为多个领域中预测和决策的强大工具，尤其是在医疗健康领域，其潜力尤为显著。ML通过模拟训练不同算法，提取重要特征，例如临床数据、影像学资料等[10,11]。通过分析，能够识别出潜在的趋势，从而为未来的决策提供帮助。ML在肝脏疾病[12,13]、糖尿病[14]和肿瘤[15]等领域的预测已经得到了广泛应用前景[16]。FLD可以通过早期发现，改变一些因素从而回归健康的一种疾病，通过ML构建FLD风险预测模型，定期筛查与其相关的特征因素，对高风险人群提供重要依据，有助于降低相关医疗费用。因此，本研究的目标是基于机器学习算法，构建NHANES数据库中2017—2020年体检人群的FLD预测模型，对其早期识别、诊断和治疗提供依据。

1 资料与方法

1.1 资料来源

研究数据来自2017—2020年的美国国家和健康调查（National Health and Nutrition Examination Survey，NHANES）数据库。NHANES是由美国疾病控制与预防中心（Centers for Disease Control and Prevention，CDC）每年开展的一项具有全国代表性的横断面研究。该调查是对美国普通人群的健康状况和营养水平进行评估。数据是通过问卷调查、家庭访谈、体检和实验室测试收集的。本研究选

取了2017—2020年的数据，共有15560名参与者，共排除了12177名参与者。本研究使用肝脂肪指数（Fatty Liver Index，FLI）[17]对FLD进行诊断，计算公式为：$FLI = (e^{(0.953 \times \log e(\text{甘油三酯}) + 0.139 \times BMI + 0.718 \times \log e(\gamma\text{-谷氨酰转肽酶}) + 0.053 \times \text{腰围} - 15.745)} / (1 + e^{(0.953 \times \log e(\text{甘油三酯}) + 0.139 \times BMI + 0.718 \times \log e(\gamma\text{-谷氨酰转肽酶}) + 0.053 \times \text{腰围} - 15.745)}) \times 100$，结果FLI≥60，则认为该患者患有FLD，FLI<30，则不患FLD。排除标准：FLI数据不全（如身体质量指数（Body Mass Index，BMI）和γ-谷氨酰转肽酶（γ-glutamyl transpeptidase，GGT）等变量缺失，共计n=11146），以及协变量数据不全（如性别、年龄等变量缺失，共计n=1031）。最终，共纳入3383名研究对象，将其分为FLD组和非FLD组。

NHANES获得了美国国家卫生统计中心伦理审查委员会（审批编号：2018-01）的批准，并获得了所有参与者的知情同意[18]。

1.2 临床资料

本研究对参与者的基本资料、身体测量数据以及生理和生化指标信息等数据进行了分析。其中信息包括年龄（years）、性别（sex）、BMI、白蛋白、肌酸酐、肌酸激酶（Creatine Pphosphokinase，CPK）、球蛋白、葡萄糖、GGT等39个数据指标。

1.3 统计学分析

本研究使用SPSS 27.0和R 4.4.2软件进行统计分析。第一步先剔除缺失比例超过10%的变量，最终保留40个自变量。采用随机森林算法对缺失值少于10%的变量进行数据插补。连续型变量以（$\bar{x} \pm s$）表示，采用t检验，而计数资料以例数（百分比）和χ^2检验进行分析。通过6种机器学习算法决策树（Decision Tree，DT）、XGBoost（Extreme Gradientboosting）、AdaBoost（Adaptive Boosting）、K最邻近（K-Nearest Neighbor，KNN）、Logistic回归、人工神经网络（Artificial Neural Network，ANN）模型构建FLD预测模型。将数据随机分成7：3比例分为训练集和测试集。最终结果使用精确率、准确率、召回率、F1值、受试者工作特征曲线（Receiver Operating Characteristic curve，ROC）及曲线下面积（Area Under Curve，AUC）、校准曲线等对模型进行评估比较。这些评估指标数值越接近1，说明模型的预测能力越强。为了更清晰地阐述机器学习模型的输出结果。分析最佳模型中的风险因素的方法是SHAP（Shapley Additive Explanation）。是基于Shapley值的计算方法，每个模型预测都会有一个对应的预测值。SHAP值表示风险因素对样本预测的影响程度是多少，并能说明影响的增强还是减弱。SHAP值大于0，表示特征对预测值有正向影响；小于0时则相反。并且还使用了皮尔逊相关系数（Pearson correlation coefficient）来计算两个特征之间的相关性。检验水准设定为α=0.05。

对于一个给定的模型预测$f(x)$，SHAP值的计算公式为：

$$\phi\ i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \times (|N| - |S| - 1)!}{|N|!} \left(f(S \cup \{i\}) - f(S) \right)$$

ϕ i是特征i的SHAP值。

N是所有特征的集合。

S是特征的一个子集，且不包含特征i。

f(S)是在特征集合S下模型的预测值。

|S|是集合S的大小，|N|是集合N的大小。

2 结果

2.1 两组一般资料比较

本研究共纳入3383名受试者，未发生FLD组1877例（55.48%），发生FLD组1506例（44.52%）。两组有核红细胞、肌酸激酶、总蛋白水平比较，差异无统计学意义（P>0.05）。两组比较结果见表1。

2.2 Logistic回归模型筛选变量

基于2017—2020年横断面数据，以是否检出FLD为因变量，以年龄、性别、总胆红素、肌酐等指标作为自变量。单因素分析结果见补充表1，筛选出36个具有潜在预测能力的因素（P<0.05）。多因素分析结果见补充表2，预测因素从36个减少为11个，包括ALT、AST、GGT等（P<0.05），最终纳入机器学习算法的FLD变量。

表1 两组一般资料比较［n（%），$\bar{x} \pm s$］

自变量	总例数（n=3383）	脂肪肝（n=1506）	非脂肪肝（n=1877）	χ^2/t	P
年龄（岁）	49.04 ± 17.80	51.43 ± 16.22	47.12 ± 18.75	7.063	<0.001
性别					
男（%）	1696（50.10）	801（53.20）	895（47.70）	10.128	<0.001
女（%）	1687（49.90）	705（46.80）	982（52.30）		
白蛋白（μg/ml）	51.68 ± 324.25	67.41 ± 394.20	39.06 ± 254.01	2.530	0.011
肌酸酐（mg/dl）	138.99 ± 85.40	147.69 ± 87.71	132.01 ± 82.86	5.332	<0.001
白蛋白比肌酐（mg/g）	49.01 ± 370.92	65.17 ± 458.13	36.05 ± 281.55	2.270	0.023
ALT（U/L）	22.45 ± 20.78	26.74 ± 19.27	19.00 ± 21.29	10.963	<0.001
ALP（IU/L）	77.90 ± 25.75	83.11 ± 27.59	73.73 ± 23.35	10.714	<0.001
AST（U/L）	22.00 ± 15.84	23.00 ± 14.61	21.20 ± 16.71	3.279	<0.001
血液尿素氮（mg/dl）	14.61 ± 5.47	15.20 ± 6.29	14.14 ± 4.66	5.648	<0.001
CPK（IU/L）	176.48 ± 452.21	168.97 ± 217.88	182.50 ± 574.88	–0.864	0.387
球蛋白（g/dl）	3.11 ± 0.44	3.18 ± 0.45	3.05 ± 0.41	8.782	<0.001
葡萄糖（mg/dl）	104.98 ± 35.20	115.19 ± 44.31	96.78 ± 22.54	15.647	<0.001
GGT（IU/L）	31.93 ± 56.00	43.63 ± 77.38	22.54 ± 25.54	11.079	<0.001
LDH（IU/L）	157.44 ± 36.40	160.17 ± 35.77	155.25 ± 36.76	3.916	<0.001
渗透压（mmol/kg）	281.81 ± 5.31	282.34 ± 5.57	281.38 ± 5.05	5.275	<0.001
总胆红素（mg/dl）	0.49 ± 0.29	0.47 ± 0.28	0.51 ± 0.30	–3.623	<0.001
胆固醇（mg/dl）	183.31 ± 40.80	186.56 ± 41.01	180.70 ± 40.45	4.161	<0.001
总蛋白质（g/dl）	7.15 ± 0.44	7.13 ± 0.44	7.16 ± 0.44	–1.776	0.076
尿酸（mg/dl）	5.47 ± 1.45	5.95 ± 1.46	5.09 ± 1.32	17.967	<0.001
甘油三酯（mg/dl）	106.52 ± 71.24	138.03 ± 84.34	81.24 ± 44.79	25.093	<0.001
LDL（mg/dl）	110.15 ± 36.02	114.13 ± 35.76	106.96 ± 35.93	5.784	<0.001
HDL （mg/dl）	53.42 ± 15.69	47.49 ± 12.35	58.17 ± 16.44	–20.926	<0.001

注：ALT：丙氨酸氨基转移酶；ALP：碱性磷酸酶；AST：天冬氨酸转氨酶；BMI：体质指数；CPK：肌酸激酶；DBP：舒张压；FLI：肝脂肪指数；GGT：γ–谷氨酰转移酶；HDL：高密度脂蛋白胆固醇；LDH：乳酸脱氢酶；LDL：低密度胆固醇；SBP：收缩压。

表2 Logistic回归分析结果（$\bar{x} \pm s$）

自变量	总例数（n=3383）	脂肪肝（n=1506）	非脂肪肝（n=1877）	t	P
总胆固醇（mg/dL）	183.07 ± 40.79	186.41 ± 41.05	180.39 ± 40.38	4.280	<0.001
淋巴细胞百分比（%）	31.36 ± 8.76	30.71 ± 8.81	31.89 ± 8.69	–3.881	<0.001
单核细胞百分比（%）	8.27 ± 2.15	8.12 ± 2.16	8.39 ± 2.14	–3.565	<0.001
红细胞计数（million cells/μl）	4.77 ± 0.50	4.84 ± 0.51	4.72 ± 0.49	6.616	<0.001
血红蛋白（g/dL）	14.12 ± 1.54	14.22 ± 1.62	14.03 ± 1.46	3.454	<0.001
血小板计数（1000 cells/μl）	241.98 ± 63.89	248.28 ± 67.95	236.94 ± 59.97	5.151	<0.001
有核红细胞	0.08 ± 0.09	0.08 ± 0.08	0.08 ± 0.11	0.631	0.528
糖化血红蛋白（%）	5.84 ± 1.14	6.17 ± 1.40	5.58 ± 0.79	15.547	<0.001
铁蛋白（ng/ml）	159.03 ± 192.96	184.27 ± 199.12	138.78 ± 185.46	6.861	<0.001
空腹血糖（mg/dl）	112.52 ± 36.88	122.93 ± 46.37	104.16 ± 23.88	15.203	<0.001
体重（kg）	83.33 ± 22.37	100.00 ± 20.55	69.96 ± 12.69	52.143	<0.001
身高（cm）	167.25 ± 9.96	168.00 ± 9.80	166.65 ± 10.04	3.934	<0.001
BMI（kg/m^2）	29.71 ± 7.31	35.42 ± 6.66	25.14 ± 3.69	56.900	<0.001
腰围（cm）	100.20 ± 17.40	114.61 ± 13.35	88.64 ± 10.12	64.353	<0.001
臀围（cm）	107.15 ± 14.58	117.45 ± 14.32	98.89 ± 8.07	47.548	<0.001
SBP	123.61 ± 18.68	126.29 ± 18.37	121.47 ± 18.65	7.531	<0.001
DBP	74.31 ± 11.31	77.45 ± 11.25	71.78 ± 10.72	14.951	<0.001
FLI	52.09 ± 33.18	84.79 ± 11.79	25.85 ± 18.05	109.379	<0.001

注：ALT：丙氨酸氨基转移酶；ALP：碱性磷酸酶；AST：天冬氨酸转氨酶；BMI：体质指数；CPK：肌酸激酶；DBP：舒张压；FLI：肝脂肪指数；GGT：γ–谷氨酰转移酶；HDL：高密度脂蛋白胆固醇；LDH：乳酸脱氢酶；LDL：低密度胆固醇；SBP：收缩压。

2.3 机器学习算法模型性能

针对FLD危险因素构建的六种机器学习模型中，AdaBoost模型在精确度、准确率、召回率和F1值等指标上表现优异，分别达到了0.96014、0.94778、0.94863和0.95435，因此被选定为最佳模型，见表3。根据AUC、精确度、准确率、召回率和F1值可知最佳模型AdaBoost模型，AUC是0.99045，最差模型DT模型，AUC是0.88887，见图1。

表3 训练集预测模型的性能比较

模型	AUC	精确率	准确率	召回率	F1值
Xgboost	0.99018	0.92955	0.94778	0.94896	0.93915
DT	0.88887	0.85146	0.90640	0.94432	0.89549
KNN	0.92245	0.87793	0.89261	0.86775	0.87281
Logistic回归	0.99110	0.93151	0.94778	0.94664	0.93901
ANN	0.91627	0.86031	0.84138	0.86473	0.86251
AdaBoost	0.99045	0.96014	0.94778	0.94863	0.95435

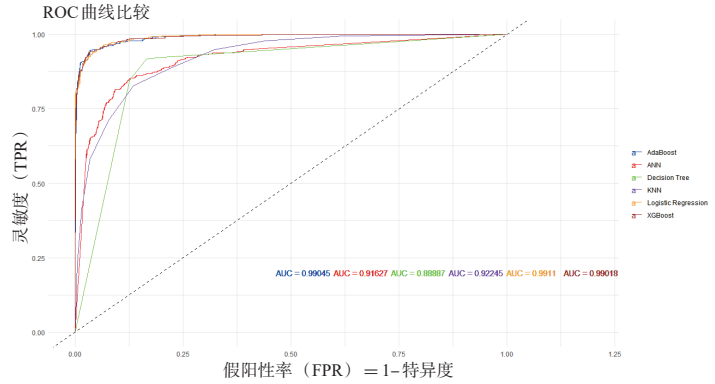


图1 ROC曲线评估结果

2.4 变量在模型中的可视化实现

以AdaBoost作为最佳模型，重要变量的SHAP特征见图2，表示影响FLD的因素重要性的排序结果。影响最显著的因素是TG，并且所有因素对FLD有正向作用。多个预测因素与预测结果的关系如图3。图3表示某些特征（如ALT、AST、BMI、GGT、TG和SF）与FLD的关系呈现出明显的非线性趋势，而其他特征（如HDL、LDL、单核细胞、TC1）则显示出较为平稳或负相关的趋势。可能表明在某个范围内ALT升高与FLD风险相关，但过高的ALT可能与其他因素有关，可能表明在某个BMI范围内，肥胖与FLD风险相关。单核细胞、HDL水平和SF水平的变化与FLD风险呈现出负相关趋势。TG水平的增加与FLD风险呈现出正相关，说明TG水平越高，FLD的风险越高。AST的变化与FLD的关系类似，显示出一种非线性的关系。LDL的变化对FLD的影响较小，整体趋势相对平稳。可视化特征之间的相关性如图4所示，BMI、Monocytes、TG、AST和ALT之间存在较强的正相关关系（深红色），这表明这些特征的值随着彼此的增加而增加。

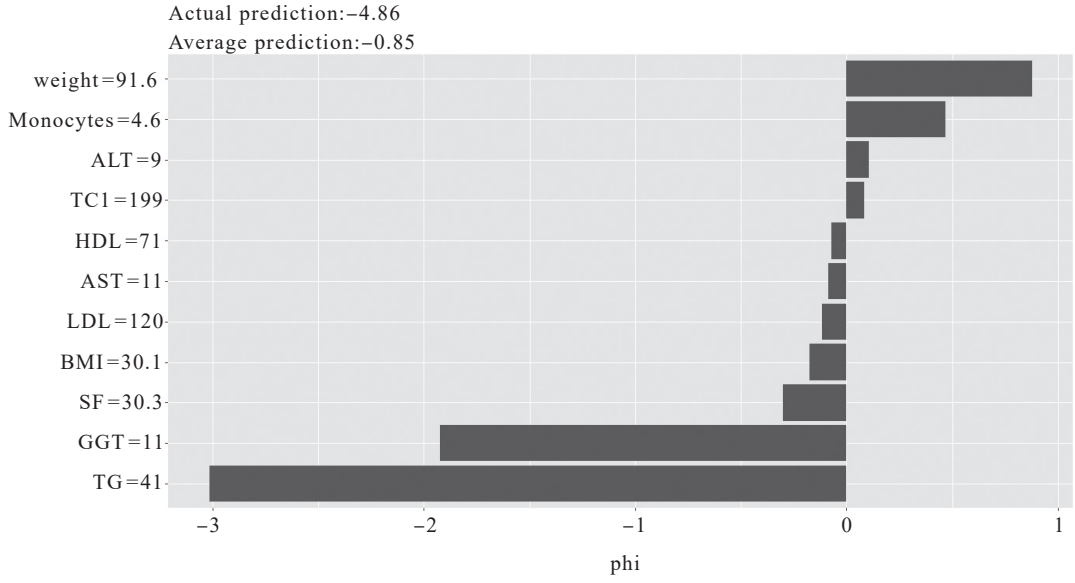


图2 SHAP特征解释图

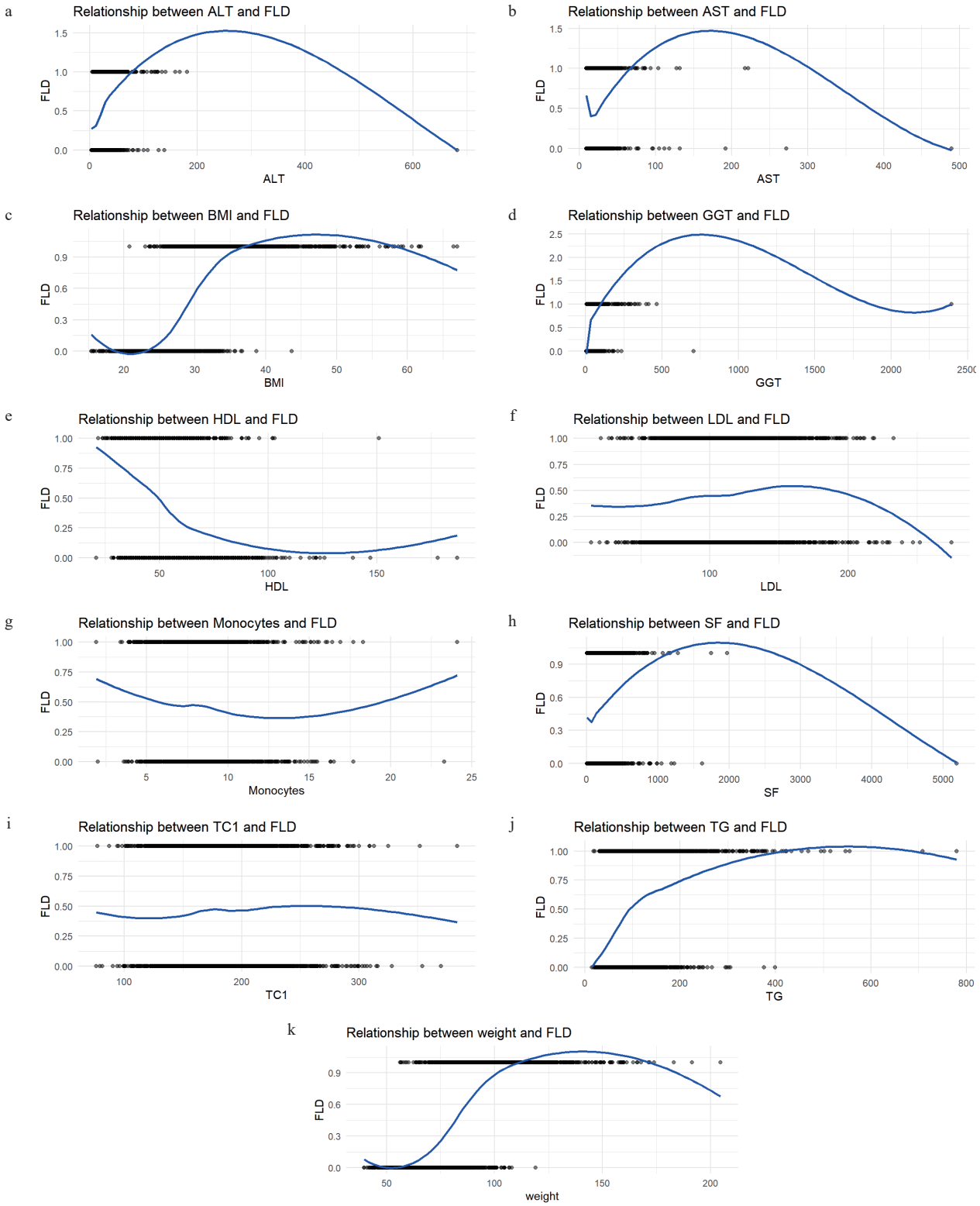


图 3 11 个特征和预测脂肪肝之间关系

注：a. ALT水平在200U/L左右时，FLD的预测值达到最高，之后逐渐下降。b. AST水平在200U/L左右时，FLD的预测值也达到最高，之后呈现下降趋势。c. BMI在30m/kg^{2}左右时，FLD的预测值达到最高，之后逐渐平稳。d. GGT水平在300IU/L左右时，FLD的预测值达到最高，之后逐渐下降。e. HDL水平增加时，FLD的预测值逐渐下降，表明HDL与FLD呈负相关。f. LDL水平与FLD之间的关系较为平稳，几乎没有显著变化。g. 单核细胞数量增加时，FLD的预测值略有上升，但变化不大。h. SF水平在1000ng/mL左右时，FLD的预测值达到最高，之后逐渐下降。i. TC1与FLD之间的关系较为平稳，几乎没有显著变化。j. TG水平在200mg/dL左右时，FLD的预测值逐渐上升，之后趋于平稳。k. 体重在100 ~ 150kg时，FLD的预测值逐渐上升，之后趋于平稳。



图 4 可视化特征之间的相关性

3 讨论

在本研究中，我们采用了六种不同的算法，包括 DT、XGBoost、AdaBoost、KNN、Logistic 回归和 ANN，以构建 FLD 的预测模型。通过对比各模型的预测效能，我们发现 AdaBoost 模型在训练集中的 AUC 值达到了 0.99045，表现出最佳的预测能力。表明 AdaBoost 在进行 FLD 预测时，能够有效地提高模型的准确性和稳定性。与其他算法相比，尽管 XGBoost 也展现了较高的 AUC 值（0.99018），但在精确率和 F1 值等指标上，AdaBoost 仍然占据优势。这一发现为 FLD 的早期诊断和干预提供了重要的理论依据，提示我们在临床应用中可以优先考虑使用 AdaBoost 算法进行 FLD 的风险评估。

建立有效的预测模型对预防和治疗 FLD 等慢性疾病具有重要意义，有助于提高人们的生活质量。本研究使用六种不同的机器学习算法模型来预测 FLD。DT[19]是一种用于分类和回归的机器学习模型，它通过树形结构对数据进行划分。其构建过程基于最大化信息增益或最小化基尼不纯度等标准[20]，优点在于易于理解和解释，适合非专业人士使用。决策树对数据预处理要求低，能处理缺失值和非线性关系，但存在过拟合风险，通常需结合剪枝或集成方法（如随机森林）来提高泛化能力[21]。XGBoost[22]是一种高效的梯度提升树算法，广泛用于机器学习竞赛中。它通过构建多个弱学习器（通常是决策树）组合成强学习器，具备高效性和灵活性，支持并行计算，缩短训练时间。XGBoost 采用正则化技术防止过拟合，并引入缺失值处理机制，增强模型稳固性，适用于多种任务[23]。AdaBoost[24]是一种集成学习算法，通过整合多个弱学习器来提升分类的准确性。该算法通过迭代调整样本的权重，使得后续的弱学习器更加关注那些分类错误的样本，从而最终构建出一个强学习器[25]。AdaBoost 简单易用，适合二分类问题，但对噪声敏感[26,27]。KNN[28]是一种简单而有效的监督学习算法，它通过计算样本之间的距离来进行分类。K 值选择对模型性能影响显著，KNN 实现简单，但计算复杂度随数据集增大而上升[29]。Logistic 回归[30]是是一种用于二分类问题的统计模型，它利用逻辑函数来预测事件发生的概率，模型简单易解释，但对线性可分性要求较高[31]。ANN[32]模仿生物神经网络，由多个相互连接的节点组成，适用于复杂非线性关系。ANN 具备强大拟合能力，但训练成本高，超参数敏感[33]。近年来，在图像识别和自然语言处理等领域取得了显著的进展[34]。本研究结果显示，所构建的 AdaBoost 模型在 FLD 预测中展现出最佳的预测性能。

本研究的优势在于多个方面，首先，采用了多种先进的机器学习算法（如 AdaBoost、XGBoost、DT 等）进行 FLD 预测模型的构建，充分发挥了各种算法的优势，从而提升了模型的预测准确性和稳定性。特别是 AdaBoost 模型在训练集中的 AUC 值高达 0.99045，显示出其在处理复杂数据时的优越性。其次，本研究在数据预处理阶段进行了细致的缺失值填补和特征选择，通过单因素和多因素 Logistic 回归分析，确保了所选特征的统计学意义。这种严谨的数据处理方法为后续模型的构建奠定了坚实的基础，减少了混杂因素的干扰，提高了模型的可信度。此外，本研究还强调了模型的可视化应用，未来可以将构建的预测模型整合到可视化平台，以便医务人员能够迅速、方便地评估个体发生 FLD 的风险。这种用户友好的界面设计不仅提高了临床应用的可行性，也为医务人员提供了有效的决策支持工具，有助于识别高危人群并制定个性化的干预措施。最后，本研究的结果为 FLD 的早期诊断和干预提供了重要的理论依据，促进了公众对 FLD 的认识和重视，具有重

要的社会意义。通过提高人们对FLD风险的认知，能够有效推动健康教育和预防措施的实施，从而降低FLD的发病率，提升整体公共健康水平。

综上所述，本研究发现AdaBoost集成算法在综合预测性能方面表现优异，适合用于FLD等慢性疾病的风险评估。未未来的研究将继续探索基于机器学习算法构建针对中国人群的其他常见慢性病风险预测模型，以为控制慢性病风险因素和推动精准医疗提供科学依据。

[参 考 文 献]

[1] Lazo M, Clark JM. The epidemiology of nonalcoholic fatty liver disease: a global perspective [J]. Semin Liver Dis, 2008, 28(4):339–350.

[2] 洪森,张廷光,徐慧.非酒精性脂肪肝患者血清中同型半胱氨酸的水平及影响非酒精性脂肪肝发生的高危因素[J].中华养生保健,2023,41(20):60–62.

[3] Le MH, Yeo YH, Li X, et al. 2019 Global NAFLD Prevalence: A systematic review and meta–analysis [J]. Clin Gastroenterol Hepatol, 2022, 20(12):2809–2817.e28.

[4] Adams LA, Lymp JF, St Sauver J, et al. The natural history of nonalcoholic fatty liver disease: a population–based cohort study [J]. Gastroenterology, 2005, 129(1):113–121.

[5] Zhou F, Zhou J, Wang W, et al. Unexpected rapid increase in the burden of NAFLD in China from 2008 to 2018: A systematic review and meta–analysis [J]. Hepatology, 2019, 70(4):1119–1133.

[6] 卫霞.血脂,血糖,肝功能和血尿酸联合检测在老年脂肪肝诊断中的应用研究[J].临床医学,2022,42(5):76–78.

[7] 徐景城.脂肪肝远不只伤肝[J].肝博士,2024,(5):40.

[8] Berger D, Desai V, Janardhan S. Con: Liver biopsy remains the gold standard to evaluate fibrosis in patients with nonalcoholic fatty liver disease [J]. Clin Liver Dis (Hoboken), 2019, 13(4):114–116.

[9] Castéra L, Foucher J, Bernard PH, et al. Pitfalls of liver stiffness measurement: a 5–year prospective study of 13,369 examinations [J]. Hepatology, 2010, 51(3):828–835.

[10] Ngiam KY, Khor IW. Big data and machine learning algorithms for health–care delivery [J]. Lancet Oncol, 2019, 20(5):e262–e273.

[11] Handelman GS, Kok HK, Chandra R V, et al. eDoctor: machine learning and the future of medicine [J]. J Intern Med, 2018, 284(6):603–619.

[12] Huang L, Li J, Zhu S, et al. Machine learning–based prognostic prediction and surgical guidance for intrahepatic cholangiocarcinoma [J]. Bioscience trends, 2025, 18(6):545–554.

[13] Wang Z, Wang Z, Li F, et al. Construction and validation of a machine learning–based prediction model for short–term mortality in critically Ill patients with liver cirrhosis [J]. Clin Res Hepatol Gastroenterol, 2024, 49(1):102507.

[14] Zhao M, Wan J, Qin W, et al. A machine learning–based diagnosis modelling of type 2 diabetes mellitus with environmental metal exposure [J]. Comput Methods Programs Biomed, 2023, 235: 107537.

[15] 沈翊康,刘海煜,王少丽,等.基于机器学习方法预测胃癌前状态的比较研究[J/OL].中华中医药学刊,1–10[2025–02–28].http://kns.cnki.net/kcms/detail/21.1546.r.20241219.1524.002.html.

[16] 张博晖,禹天同,赵帅,等.基于机器学习的冠状动脉慢性完全闭塞患者术后不良心血管事件预测模型的构建[J].心脏杂志,2025,(3):264–270.

[17] Bedogni G, Bellentani S, Miglioli L, et al. The Fatty Liver Index: A simple and accurate predictor of hepatic steatosis in the general population [J]. BMC Gastroenterol, 2006, 6:33.

[18] National Health and Nutrition ExaminationSurvey: questionnaires, and relateddatasets [J/OL]. https ://wwwn. cdc.gov/nchs/nhanes/Default.aspx.

[19] Carrizosa E, Molero–Río C, Romero Morales D. Mathematical optimization in classification and regression trees [J]. Top (Berl), 2021, 29(1):5–33.

[20] 陈云樱,吴积钦,徐可佳.决策树中基于基尼指数的属性分裂方法[J].微机发展,2004,14(5):66–68.

[21] 袁焦,王珣,潘兆马,等.基于机器学习的列车设备故障预测模型研究[J].计算机与现代化,2020,(12):49–54.

[22] 胡雪梅,李佳丽,蒋慧凤.机器学习方法研究肝癌预测问题[J].系统科学与数学,2022,42(2):417–433.

[23] 许裕栗,杨晶,李柠,等.Xgboost算法在区域用电预测中的应用[J].自动化仪表,2018,39(7):1–5.

[24] Qin Y. AdaBoost–Theory and Application [J]. Musicmcgillca, 1970.

[25] 杨新武,马壮,袁顺.基于弱分类器调整的多分类Adaboost算法[J].电子与信息学报,2016,38(2):373–380.

[26] Schwenk H, Bengio Y. Boosting neural networks [J]. Neural Comput, 2000, 12(8):1869–1887.

[27] Zuo D, Yang L, Jin Y, et al. Machine learning–based models for the prediction of breast cancer recurrence risk [J]. BMC Med Inform Decis Mak, 2023, 23(1):276.

[28] 孙岩,吕世聘,王秀坤,等.基于结构学习的KNN分类算法[J].计算机科学,2007,34(12):184–186+237.

[29] Sun S , Huang R. An Adaptive k–Nearest Neighbor Algorithm [C] // International Conference on Fuzzy Systems and Knowledge Discovery.2010.

[30] 顾鑫,曹丹华,吴裕斌,等.基于逻辑回归的多任务域快速分类学习算法[J].计算机工程与应用,2017,53(15):47–56.

[31] Song X, Liu X, Liu F, et al. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta–analysis [J]. Int J Med Inform, 2021, 151:104484.

[32] Zou J, Han Y, So SS. Overview of artificial neural networks [J]. Methods Mol Biol, 2008, 458:15–23.

[33] Bourquin J, Schmidli H, Van Hoogevest P, et al. Basic concepts of artificial neural networks (ANN) modeling in the application to pharmaceutical development [J]. Pharm Dev Technol, 1997, 2(2):95–109.

[34] Bertolaccini L, Solli P, Pardolesi A, et al. An overview of the use of artificial neural networks in lung cancer research [J]. J Thorac Dis, 2017, 9(4):924–931.