Nursing Science 护理学, 2025, 14(6), 956-965
Published Online June 2025 in Hans. https://www.hanspub.org/journal/ns
https://doi.org/10.12677/ns.2025.146127

Hans汉斯

数据挖掘技术在代谢相关脂肪性肝病预测中应用的研究进展

邹慧慧，潘梦圆，陈正英*

吉首大学医学院，湖南 吉首

收稿日期：2025年5月10日；录用日期：2025年6月5日；发布日期：2025年6月16日

摘　要

代谢相关脂肪性肝病(MAFLD)作为全球最常见的慢性肝病，其早期预测和精准干预对预防和延缓疾病进展至关重要。近年来，随着医疗大数据和人工智能技术的快速发展，数据挖掘技术在代谢相关脂肪性肝病风险预测方面展现出巨大潜力。数据挖掘技术为代谢相关脂肪性肝病的个体化预测和精准干预提供了新工具，但其临床应用仍需跨学科协作与标准化数据生态支持。本文从数据挖掘及代谢相关脂肪性肝病的概念入手，对数据挖掘技术在代谢相关脂肪性肝病预测特征选择和模型构建中的应用现状加以总结和讨论，梳理现有预测模型的局限，以期为更好发挥数据挖掘技术在代谢相关脂肪性肝病预测的实际效用提供参考。

关键词

代谢相关脂肪性肝病(MAFLD)，数据挖掘，预测模型

Research Progress in the Application of Data Mining Techniques for Metabolic-Associated Fatty Liver Disease Risk Prediction

Huihui Zou, Mengyuan Pan, Zhengying Chen*

School of Medicine, Jishou University, Jishou Hunan

Received: May 10th, 2025; accepted: Jun. 5th, 2025; published: Jun. 16th, 2025

Abstract

Metabolic associated fatty liver disease (MAFLD), as the most prevalent chronic liver disease globally,

*通讯作者。

文章引用: 邹慧慧, 潘梦圆, 陈正英. 数据挖掘技术在代谢相关脂肪性肝病预测中应用的研究进展[J]. 护理学, 2025, 14(6): 956-965. DOI: 10.12677/ns.2025.146127

requires early prediction and precise intervention to effectively prevent and delay disease progression. In recent years, driven by advancements in medical big data and artificial intelligence technologies, data mining has demonstrated significant potential for risk prediction of MAFLD. Data mining offers novel tools for individualized prediction and precision intervention in MAFLD; however, its clinical application necessitates interdisciplinary collaboration and the establishment of a standardized data ecosystem. This article begins with an overview of the concepts of data mining and MAFLD, systematically summarizing and analyzing the current status of data mining technology in feature selection and model construction for MAFLD prediction. Furthermore, it critically evaluates the limitations of existing prediction models, aiming to provide insights and guidance for maximizing the practical utility of data mining in MAFLD prediction.

Keywords

Metabolic-Associated Fatty Liver Disease (MAFLD), Data Mining, Prediction Model

1. 引言

近年来，代谢相关脂肪性肝病(Metabolic-associated fatty liver disease, MAFLD)患病率逐年上升，逐渐成为全球范围内影响最广的慢性肝病，影响着世界超过三分之一的成年人[1]。MAFLD 虽然并非致命性疾病，但因其慢性病程长和疾病预后差，从脂肪性肝炎、肝纤维化、肝硬化、肝癌，患者多死于心脑血管疾病，疾病预测能够在早期聚焦于高风险人群，及时捕捉疾病动态进展方面具有强大优势，对疾病的早期防控具有重要意义。

随着大数据时代的到来，信息技术得到高速发展，数据挖掘技术所具备的决策、优化等强大能力，使之成为信息时代发展的基础[2]。随着预测模型的不断开展，近年来研究者们开始关注对健康人群未来慢性病的发病风险预测，以实现对慢性病的早期防控。目前，数据挖掘技术已被研究者们应用于对 MAFLD 的预测，有望在早期筛查出疾病早期高危人群以及需要生活方式干预的目标人群，本研究将对数据挖掘技术在 MAFLD 预测领域中的应用现状、现有 MAFLD 预测模型的局限和未来展望进行综述。

2. 数据挖掘的概念

数据挖掘技术是一种处理海量数据的技术，通过在海量数据中挖掘获取提炼隐含规律，用以实现分类、回归、降维、聚类等相关任务[3]。目前数据挖掘技术已得到长足发展，并且在医疗领域中也取得了一定的成果，比如构建疾病预测模型辅助诊断和预警、对医学影像中的图像数据挖掘提升医生决策诊断效率、对医疗就诊信息的挖掘获得成本效益实现医院管理的优化等。其中，数据挖掘技术在医学信息的事件预测中应用最为广泛[3]，通过对相关数据的深入挖掘与分析，运用科学知识、手段和方法[4]，对事件状态或未来发展趋势做出科学的估计和评价，实现对事件发生前的预警和精准调控。在数据挖掘技术的辅助下，预测研究已从既往对传染性疾病的短期事件发生预测转移至对慢性病的早期诊断[5]、未来发病倾向和预后研究上来。

3. 代谢相关脂肪性肝病的定义与概念

代谢相关脂肪性肝病(MAFLD)是指遗传易感个体由于营养过剩和胰岛素抵抗引起的慢性代谢应激性

肝病，且与代谢功能障碍密切相关的慢性肝病，既往称为非酒精性脂肪性肝病(NAFLD)。近年来，随着各项研究对非酒精性脂肪肝与代谢之间关联的证实，于 2020 年国际肝病小组建议将其更名为代谢相关脂肪性肝病(MAFLD)，并进一步更新了疾病诊断标准[6]。又于 2023 年考虑到“脂肪”一词带来的污名化，重新进行了更名建议[7]，建议使用脂肪变性肝病代替原有脂肪肝一词。MAFLD 的疾病谱广泛：疾病进展包括单纯性脂肪肝、代谢相关脂肪性肝炎、肝纤维化，甚至肝细胞癌。MAFLD 的重新定义标志着从“排除性诊断”到“病因明确性诊断”的转变，强调代谢异常的核心作用，推动临床实践和研究的革新。其概念演变反映了对疾病机制理解的深化，也为个体化治疗和跨学科管理提供了新方向。

4. 数据挖掘技术在 MAFLD 预测特征选择中的应用现状

在 MAFLD 预测问题中，特征选择是模型构建的关键步骤，受到研究对象的复杂性和预测因素众多等因素影响，研究者们首先需在高维数据中获得能够有效描述预测问题且能让研究结果充满说服力的特征子集，同时不能降低后续预测任务的精度。特征选择[8]，就是从原始特征空间中选择与目标任务紧密相关的特征，剔除冗余特征，通过在原始特征空间中获得拥有最佳特征子集，辅助科研人员完成后续的分类、回归等预测任务，实现数据降维、提升后续任务的精确度和效率。

因 MALFD 与代谢相关因素的关联得以证实，既往研究中研究者们常聚焦于代谢相关指标、生活方式行为、是否有共病(如高血压、糖尿病等)以及家族史、性别、年龄等人口学资料，将其作为建立 MAFLD 预测模型中的预测因素，但在预测因素的特征选择方式上存在较大的差异。在数据挖掘技术应用于预测前，研究者们往往采用传统的统计学方法，如 Logistic 回归来选择与 MAFLD 相关的疾病因素。随着数据挖掘技术在疾病预测领域的开展，在预测因素的选择上也开始有了新的创新，以下借助基于特征子集评价策略的三种特征选择方法[9]对数据挖掘技术在 MAFLD 预测因素选择中的实际应用进行综述：

4.1. 过滤式(Filter Method)

该方法独立于模型构建的学习算法，先进行原始数据的特征选择，再将特征选择结果用于学习算法的输入，具有快速剔除噪声特征、计算效率高、通用性强的特点。如特征评估、特征筛选、特征排序等。在 MAFLD 预测问题中，过滤式特征选择帮助研究者从众多特征中筛选出与 MAFLD 预测高度相关的特征，进而提高预测模型的性能。

已有大量研究将该方法应用到 MAFLD 风险因素筛选及预测的模型构建中。例如，李绒[10]等人采用过滤式方法筛选出重要的风险因素(特征变量)，通过综合 Relief (relevant features)相关统计量和随机森林的特征重要性评分的平均值，对所有特征进行重要性排序，随后采用逐渐增加相关特征的贪心策略，筛选出最佳特征子集用于后续预测模型的构建。此外，部分研究对特征选择方法进行了改进，提出了基于随机森林(RF)的过滤式(Filter)特征选择算法[11]。该方法结合了过滤式特征选择、随机森林算法以及 K 折交叉验证技术。与传统的过滤式特征选择方法相比，该方法的优势主要体现在两个方面：一是引入了特征重要性之间的相关性，二是在分类器构建之前预先筛选无关变量和无信息变量，从而显著提升了特征子集的质量和分类模型的预测精度。

4.2. 包裹式(Wrapper Method)

该方法依赖后续的学习算法，将要使用的模型性能作为评价特征子集的评价标准，给定学习器选择最优性能“量身定做”的特征子集。通过多次训练模型提高学习模型性能。基于学习算法在特定子集上的性能进行评估，通常应用于小规模数据集和简单的学习算法中，如支持向量机(SVM)、极端梯度提升(XGBoost)、遗传算法等。

在 ZHAO 等人研究中[12]，研究者采用包裹式特征选择方法中的遗传算法结合 Spearman 相关系数筛

选最佳特征子集。遗传算法通过模拟自然进化过程，将问题的求解转化为类似于生物进化中染色体基因的交叉和突变的过程。在每一代进化过程中，适应度较高的个体被保留，而低适应度的个体则被淘汰。与传统的优化算法相比，遗传算法在处理复杂组合优化问题时，通常能够更快地获得更优的结果。在一些研究中，研究者会联合使用多种特征选择的方法来确定最佳特征子集。例如，在 Huang [13]等人的研究中，采用了极端梯度提升(XGBoost) - 递归特征消除(RFE)与最小绝对收缩和选择运算符(LASSO)相结合的方法筛选特征预测因子。其中，XGBoost-RFE 属于包裹式特征选择方法，通过不断减少特征数量并验证模型性能，最终筛选出最优特征子集；另一项关于 NAFLD 和氧化应激之间 31 个交集基因的研究中[14]，研究者同样使用支持向量机 - 递归特征消除(SVM-RFE)包裹式特征选择方法提取了 20 个基因。包裹式特征选择方法的优点在于能够捕捉特征之间的相互关系，提供更准确的特征子集。

4.3. 嵌入式(Embedding Method)

该方法是前两者的组合算法，将特征选择过程作为学习算法中的一部分，由于嵌入式特征选择与学习算法紧密结合，因此可以获得更好的准确度和速度，比如正则化线性回归(LASSO、Ridge 回归)、决策树验证、CART 算法等。

在实际应用中嵌入式特征选择方法中的正则化线性回归多与其他方法结合使用。一项前瞻性队列研究[13]运用极端梯度提升(XGBoost) - 递归特征消除(RFE)算法结合最小绝对收缩和选择运算符(LASSO)进行特征预测因子的筛选。其中，LASSO 作为一种经典的数据降维方法，通过构建 L1 正则化惩罚函数，能够将不显著变量的回归系数压缩为 0，从而实现特征变量的有效筛选，但其未充分考虑变量间的多重共线性问题。Peng 等人[15]也在其研究中采用了 LASSO 回归结合随机森林的方法，成功筛选出 NAFLD 风险的重要预测因子。然而，不同数据条件下的特征选择方法表现不同。黄娅[16]探讨了四种回归模型在正确选择影响因素的平均数量与剔除影响因素方面的表现。研究结果显示：随机森林-Lasso Logistic 回归模型和 Lasso Logistic 回归模型在正确选择影响因素的平均数量上高于最优子集回归模型和逐步 Logistic 回归模型；当阳性率为 50%时，四种回归模型的筛选效果达到最佳。此外，当样本量达到自变量个数的 10 倍以上时，样本量的变化对四种回归模型的影响较小。将随机森林-Lasso Logistic 回归模型应用于脂肪肝健康风险因素筛选的实例中，结果表明，该模型在拟合效果和预测性能上均显著优于另外三种模型。

综上，数据挖掘技术在 MAFLD 预测中的特征选择方法具有多样性，各种方法可以根据数据的特点和具体需求进行选择。未来研究可进一步整合多维度指标，以提升预测模型的全面性和实用性。通过有效的特征选择，可以提高 MAFLD 预测模型的精度和可解释性，进而为临床诊断与风险评估提供更可靠的支持。

5. 数据挖掘方法在 MAFLD 预测模型构建中的应用现状

随着医学数据的积累和计算机技术的发展，数据挖掘技术被广泛应用于 MAFLD 的预测模型构建中。MAFLD 的早期诊断通常依赖于医学影像、实验室检查、临床数据等多种数据源，这为数据挖掘提供了丰富的原材料。通过运用数据挖掘技术，可以整合、分析这些多源数据，发现 MAFLD 的潜在预测因子，并建立准确的预测模型，从而为临床提供早期诊断工具。在 MAFLD 预测模型的研究中，常用的数据挖掘方法包括以下三大类：

5.1. 传统统计学习

在 MAFLD 预测中，传统统计学习方法如 Logistics 回归、Cox 回归、列线图等，提供了强有力的工具来分析和预测疾病风险。这些方法因其简洁性、可解释性和高效性，广泛应用于临床医学数据分析。

尽管随着数据规模的增大，深度学习等更复杂的算法展现出强大的潜力，但传统统计方法仍在一定场景下占据着重要地位。

在瘦型代谢功能障碍相关脂肪性肝病(MASLD)预测领域[17]，研究者开发了列线图预测模型，该模型在预测准确性、辨别力和临床实用性方面均优于传统指标脂肪肝指数(FLI)和肝脂肪变性指数(HSI)，为瘦型人群 MASLD 风险的早期识别提供了有力工具。李楠[18]等研究者基于多因素 Cox 回归分析构建的中青年人群 NAFLD 发生风险的列线图预测模型，可为临床医师提供 1~3 年 NAFLD 的发病风险的个体化评估。然而，该研究仅纳入常规体检指标，未考虑饮食、体力活动及心理等生活方式因素，这在一定程度上限制了模型对 NAFLD 风险因素的全面评估能力。张家丽[19]利用 CRT 分类树、列线图两种方法构建 NAFLD 风险预测模型，这两个风险预测模型均具有良好的鉴别能力、临床实用性以及较高的准确性。高福来[20]等人开发了基于血清 Betatrophin 水平的 NAFLD 列线图预测模型，其预测效能为 85.71%，特异性为 88.68%，敏感性为 82.98%，该模型显示出良好的模型拟合度和临床价值，但研究样本量较小可能影响结果的可靠性。此外，在特定人群研究方面，一项针对护士职业人群的前瞻性队列研究采用 logistic 回归构建了 NAFLD 风险预测模型[21]。该模型预测性能较好，外部验证结果显示其灵敏度达 90%，特异度为 63.2%，总体预测正确率为 65.4%。另外一项病例回顾研究构建了绝经前 HR+乳腺癌患者内分泌治疗相关脂肪肝的列线图预测模型[22]。该模型基于 Cox 多因素分析结果构建，经内部验证显示出良好的分辨度，其预测值与实际观察值具有较好的一致性，模型的准确性和符合度均较理想，为 HR+乳腺癌患者内分泌治疗相关脂肪肝的风险预测提供了可靠工具。

5.2. 机器学习算法

近年来，机器学习技术在 MAFLD 预测领域应用取得了显著进展。相比传统统计方法，机器学习算法可以处理更大规模的数据集、更复杂的模式和关系，模型更加灵活，但不如深度学习对大数据的依赖强。通过运用典型机器学习算法来构建 MAFLD 的风险预测模型，是目前整合医学资源和数学模型的一种较新的数据挖掘方法。在众多算法中，XGBoost 算法显现出较大优势，如 Zhao [12]等人基于电子体检记录，采用 XGBoost 算法构建脂肪肝预测模型，显著提升了训练速度，有效降低了方差并防止过拟合。同时，其支持特征粒度上的并行计算，进一步提高了计算效率。该模型在测试集上表现出了优异的预测能力，AUC 达到 0.89，与神经网络等其他算法相比，XGBoost 基于决策树模型的特性使其具有更好的可解释性，有助于深入理解体检数据在模型中的作用机制。此外，该模型不仅在横断面数据上表现优异，在纵向数据中依然具有较高的预测效能。雷丽[23]等研究者系统评估了六种机器学习算法包括(决策树、XGBoost、Bagging、随机森林、人工神经网络和支持向量机)在脂肪肝预测建模中的表现，通过 10,000 次重复模拟试验发现，XGBoost 算法在纵向亚模型构建中展现出最优性能(AUC = 0.958，召回率 = 0.790，精确率 = 0.761，准确率 = 0.898)，其综合预测效能显著优于传统 Logistic 回归模型(AUC = 0.732)。进一步结合时依 Cox 生存函数构建的 XGBoost-Joint 联合模型，在 24,106 人次的 11 年纵向数据验证中，表现出良好的模型稳定性和预测一致性。另一项基于 22,140 人队列的纵向数据构建的 XGBoost 预测模型[24]，在平衡数据集上展现出最优分类性能在平衡数据集上展现出最优分类性能(准确率 = 0.835，灵敏度为 = 0.835，特异性 = 0.834，AUC = 0.914)。其综合预测效能(约登指数 = 0.669，F-1 值 = 0.833)显著优于同期比较的其他三类机器学习模型，进一步验证了 XGBoost 算法在脂肪肝风险预测中的优势地位。

除 XGBoost 算法外，随机森林算法在 MAFLD 预测领域中得到了广泛应用，它结合了决策树的预测能力和随机性的优点，被广泛应用于分类和回归问题中。有研究者采用随机森林法构建了 2 型糖尿病患者罹患 NAFLD 的多维度预测模型[25]，不同模型的预测精度介于 81.5%至 83.6%之间，显著优于传统 ZJU 指数(评估 NAFLD 风险的生物化学评分系统) 70.9%的平均精度和 72.3%的最高精度，该模型在精确度和

适用性方面均优于传统统计方法所设计的指数预测模型，为 2 型糖尿病患者并发 NAFLD 风险提供了更可靠的工具。在涉及 k 最近邻(kNN)、支持向量机(含径向基函数)、高斯过程(GP)、随机森林、神经网络(NN)、Adaboost 和朴素贝叶斯的比较研究中[26]，随机森林算法在 NAFLD 分期预测中展现出独特优势，其对脂肪肝存在期、脂肪变期和纤维化期的预测准确率分别达到 82%、52%和 57%。特别是 82%的总体准确率，为基于人体测量指标的 NAFLD 一级预防筛查提供了可靠的技术路径。此外，随机森林和支持向量机模型在 NAFLD 筛查中也表现出良好的性能[27]。

值得注意的是，数据挖掘技术在中医领域的脂肪肝预测中也展现出独特的价值。例如，吕航[28]等人运用决策树模型，探讨了中医人格体质类型对 2 型糖尿病患者伴发 NAFLD 的预测作用。所构建的 NAFLD 患病风险模型达到了 87.1%的预测准确度，为中医疾病预测的客观化和量化提供了新的研究思路。此外，机器学习算法在非肥胖人群的脂肪肝风险预测中也有应用，如王菊芳[29]等人采用人工神经网络(ANN)方法构建的非肥胖人群的脂肪肝风险预测模型，也具有较高的预测价值。这些研究结果表明，机器学习方法在 MAFLD 风险预测中具有广阔的应用前景，特别是在复杂疾病和个体化医疗领域。

5.3. 深度学习算法

深度学习是机器学习的一个分支，与既往基于横断面数据的预测模型不同，基于纵向数据的时间序列预测在疾病早期风险预测方面呈现出强大优势，而深度学习在时间序列预测中具有强劲性能[30]。深度学习[31] (Deep Learning, DL)算法能够借助多个处理层完成对复杂样本数据的特征提取和挖掘，在面对高维医学数据时比经典的机器学习算法能够拥有更优的处理方式，并通过自动学习维持高维特征提高预测性能。常见的三大类深度学习算法为卷积神经网络(CNN)、循环神经网络(RNN)以及长短期记忆神经网络(Long short term memory, LSTM)。其中，LSTM 以输入门、遗忘门和输出门选择关键信息，同时能够有效处理梯度消失和梯度爆炸的相关问题，因此近年来研究者多选择 LSTM 及其变体完成对健康人群的未来脂肪肝发病风险预测的初步探索。深度学习算法在 MAFLD 预测模型构建中有别于机器学习算法，在增加预测准确率的同时也实现了对健康人群未来 MAFLD 风险预测。

一项回顾性队列数据收集 7 年内台北某医学门诊的体检人群数据，完成了对健康体检人群当前就诊(CVP)和下次就诊(NVP)的脂肪肝风险预测模型的初步构建[32]，该研究使用 K-邻近分类(KNNC)、Adaboost、SVM、逻辑回归(LR)、随机森林(RF)、高斯朴素贝叶斯(GNB)、决策树 C4.5 以及决策树 CART 构建了当前就诊脂肪肝风险的预测模型，同时使用 LSTM 等时间序列模型构建了可预测下一次就诊时的脂肪肝风险。在固定区间特征条件下，各 LSTM 变体模型的预测准确率在 78.36%~79.32%，这一发现为深度学习模型在纵向健康数据预测中的稳健性提供了实证支持。基于传统机器学习算法的 CVP 预测模型与基于深度学习的 NVP 预测模型相结合，可为脂肪肝风险的动态监测提供可靠的技术支持。此外，随着时间序列数据的不断更新采集，有研究实现了 NAFLD 风险的动态预测[33]。该研究中采用 LSTM 架构结合 SHAP 可解释性算法，通过持续纳入新健康体检记录进行模型迭代更新，显著提升了预测系统的时序适应能力，并使用重复测量方差分析(ANOVA)，在五个连续时间点上对比 LSTM 与 L1 惩罚逻辑回归(LR)模型的整体性能差异，研究结果显示，LSTM 模型展现出显著的时间序列学习优势。在内部验证集中，LSTM 的平均 AUC 值达 0.770，高于 LR 模型(0.752)。随着体检记录的时间跨度增加，LTSM 模型的性能改善幅度(最终就诊与初次就诊的 AUC 差值 0.089)显著高于 LR 模型(最终就诊与初次就诊的 AUC 差值 0.060)，这一趋势在外部验证中得到进一步证实。模型动态更新过程中，深度学习模型的优势在纵向数据分析中尤为突出，LSTM 模型能够有效捕捉时序特征间的非线性交互作用，其动态权重调整机制使模型对新纳入体检数据的适应速度较传统模型有所提升。基于 LSTM 的模型的性能随着新的健康体检记录的增加而进一步增强，这意味着 LSTM 模型可以在更长的数据时间跨度上能够提供更准确的预测，这一发

现为长期健康监测场景下的模型选择提供了重要依据。

5.4. 多种数据挖掘方法建立 MAFLD 预测模型的对比

不同数据挖掘方法在 MAFLD 风险预测中的效能差异主要源于算法特性与数据特征的适配性。有研究将不同的数据挖掘方法进行了比较：

一项对比研究发现[34]，Ridge 回归在临床预测中表现最佳，其高阴性预测值高达 96%，相比之下，AdaBoost 与决策树均存在过拟合问题。虽然 Logistic 回归与 Ridge 回归表现相似，但 Ridge 回归因计算简便且高阴性预测值更受推荐。随机森林算法在部分研究中展现出突出的预测性能。如在 RF、朴素贝叶斯(NB)、人工神经网络(ANN)和 LR 四种模型的对比中，随机森林预测性能最优[35]，此外，在区分 NAFLD 分子簇方面，其表现也优于 SVM、XGBoost 和 GLM 模型[36]。然而，在另一项 NAFLD 发病风险模型比较研究中[13]，CatBoost 表现最佳，随机森林次之。尽管 XGBoost 具有较强的校正能力，但 Logistic 回归因在验证集中预测价值最高，因此被选为最优模型。在脂肪肝分类预测研究中，将决策树、神经网络、支持向量机、贝叶斯网络和随机森林算法 5 种常见的机器学习算法在脂肪肝分类预测研究中的应用进行了实现和比较[37]，并将其模型运用到 2337 例体检数据中，提取了重要的指标作为参数进行了分析比较，观察数据发现决策树模型的预测准确率最高，达到了 70%以上，支持向量机和神经网络模型次之，处于 68%左右的水平，而贝叶斯网络模型的预测性能最低，仅有 62.17%，由此可见，决策树模型的分类预测效果最优，应用在小样本数据上有优势。

不同的数据挖掘方法由于采用的算法、模型不一致，因此，模型中纳入的特征子集也不尽相同。李绒[10]等基于女性体检数据构建了 6 种机器学习预测模型，通过不断迭代增加特征分析最佳特征子集，结果可见模型性能随特征值个数增加而逐步提高，随机森林的 AUC 值最高，其他 5 种算法模型的性能比较接近，XGBoost、AdaBoost 和 MLPC 达到最高性能时的特征个数较少(分别是 16、20 和 20 个)。这种方法与传统方法相比，具有较强的可解释性，能减轻传统机器学习的“黑盒”模式影响。对于某些算法来说，使用最佳特征子集与全部特征集合的性能基本相同，因此在临床研究时可减少所收集的特征数量(如只收集最佳特征子集)，从而降低模型训练的数据成本，适用于临床疾病预警和辅助决策支持。

此外，多数研究采用 SHAP 方法来增强模型的可解释性，例如，有研究结合了基于集成的机器学习 XGBoost 模型和可解释的人工智能 SHAP 方法来检测高风险的非酒精性脂肪性肝炎，该模型优于常用的临床风险指数，并且可以增加在资源有限环境中对高危非酒精性脂肪性肝炎患者的识别[38]。ML 与 SHAP 可解释性的整合为 NAFLD 提供了强大的预测工具，增强了疾病的早期识别和潜在管理[39]。综上，各数据挖掘方法在不同场景下各有优劣，在脂肪肝预测问题上，应根据不同的使用场景、数据量大小、变量间的关系选择最佳的数据挖掘方法。同时，模型的解释性、计算资源和训练时间等也是数据挖掘方法选择时需要考虑的因素。

6. 现有 MAFLD 风险预测模型的局限及未来展望

6.1. 诊断标准以及结局定义存在差异

受到近年来代谢相关脂肪性肝病的更名影响，诊断标准的不断更新，造成了在此类疾病预测模型构建中的较大差异。此外，由于肝活检的金标准在健康人群和 MAFLD 人群中的使用受限，目前较多研究使用 B 超作为代替影像学判断方式，也有研究使用 FLI、HSI 等计算公式代替脂肪肝的诊断。因此，更名、诊断金标准的差异造成了目前所构建的预测模型对 MAFLD 结局的定义不一致，造成类似预测模型之间的可比性较低，实用性较差。

6.2. 预测特征选择结果差异大且存在实用性问题

现有 MAFLD 预测模型的特征选择存在显著异质性。对预测因子的纳入个数，同时在纳入因素中获得临床实际易获得的数据特征从而推进临床实际应用也存在很大差异。在预测特征选择上过度依赖传统指标且忽视临床实际需求，导致模型“学术价值高，应用落地难”。未来需建立标准化特征筛选框架，优先保障易获取指标的核心地位，同时探索动态、多模态数据的整合路径，以弥合科研与临床的鸿沟。

6.3. 缺失数据处理文内未报告易产生偏倚

缺失数据处理方式的选择会显著影响研究结果的可靠性。然而，在多数相关研究中，研究者未能充分报告具体的缺失数据处理方法，这可能会引入选择偏倚或信息偏倚，最终影响模型的可重复性和泛化能力。目前研究中缺失数据处理方法呈现明显的异质性。从方法学复杂程度来看，这些处理方式大致可分为两类：一是基于简单统计量的快速插补法(如采用均值或众数进行单变量插补)，二是基于统计建模的复杂方法(如通过链式方程实现的多变量插补技术)。这种处理方法的多样性虽然提供了灵活的选择空间，但也给研究间的可比性带来了挑战。

6.4. 研究设计类型差异限制了预测模型的用途

在当前所构建的 MAFLD 相关预测模型中，以横断面数据所开展的诊断模型较多。尽管基于横断面数据的诊断预测模型已得到可观的进展，但以队列数据为基础构建的预后模型较为有限，在预测 MAFLD 的未来发病风险方面存在不足。在既往研究中，对 NAFLD 人群开展的肝纤维化预后模型开展较多，但针对健康体检人群的 MAFLD 早期预测的风险模型极为有限。

6.5. 模型实际应用与可视化呈现存在障碍

目前，在模型的实际应用中存在较大局限，虽然已有较多研究针对该疾病构建了科学的风险预测模型，但所构建的预测模型无法在临床进行实际应用。而在模型推广至临床应用前，模型的预测效能验证也存在一定的问题，大多数研究仅在开发模型的数据集内部完成了验证，但少有研究报道了所构建的预测模型在外部数据集中的验证情况。除此之外，目前已有较多研究者为了推动预测模型结果在临床的实际应用，借助可视化方式进行呈现，比如列线图、网页制作等方式，但较多研究仅止步于模型构建，缺少对结果的可视化呈现。

7. 结论

综上所述，数据挖掘技术在特征选择和模型构建中均有应用，与传统统计算法相比数据挖掘技术有更好的表现效果，预测模型取得更好的预测效能，预测效率得到了显著提升。数据挖掘技术在脂肪肝领域的应用，有助于发现疾病关联的关键特征，提早发现 MAFLD 高风险人群，辅助疾病的早期防控。预测个体未来发生疾病的风险是健康管理中的关键一步[40]，MAFLD 风险预测应从“静态单一”向“动态多元”范式转变，未来模型需突破传统机器学习框架，构建“数据–算法–临床”三位一体的智能系统。通过跨学科协作攻克可解释性、动态适应性和临床实用性三大核心挑战，最终实现从风险预测到精准干预的闭环管理。未来可考虑纳入生活方式行为因素作为预测因素，在队列数据的基础综合考虑时间和高危因素，借助数据挖掘技术构建临床实用性强的预测模型，为人群健康保驾护航。

致　谢

感谢本文全体作者的大力支持。

基金项目

吉首大学校级课题(Jdy24082)。

参考文献

[1] Riazi, K., Azhari, H., Charette, J.H., Underwood, F.E., King, J.A., Afshar, E.E., *et al*. (2022) The Prevalence and Incidence of NAFLD Worldwide: A Systematic Review and Meta-Analysis. *The Lancet Gastroenterology & Hepatology*, **7**, 851-861. https://doi.org/10.1016/s2468-1253(22)00165-0

[2] 张君秋, 赵建光, 孟凡明, 等. 基于数据挖掘技术的相关模型与算法研究综述[J]. 中国新通信, 2023, 25(2): 45-48.

[3] 任芳, 刘硕. 数据挖掘技术在医学信息中的广泛应用[J]. 中国多媒体与网络教学学报(上旬刊), 2019(6): 9-10.

[4] 李志鹏, 杨阳朝, 廖勇, 等. 数据驱动的事件预测技术最新研究进展[J]. 信息安全学报, 2022, 7(1): 40-55.

[5] 李金金. 天津市慢性病队列研究及风险预测模型的建立[D]: [博士学位论文]. 天津: 天津医科大学, 2018.

[6] Eslam, M., Newsome, P.N., Sarin, S.K., Anstee, Q.M., Targher, G., Romero-Gomez, M., *et al*. (2020) A New Definition for Metabolic Dysfunction-Associated Fatty Liver Disease: An International Expert Consensus Statement. *Journal of Hepatology*, **73**, 202-209. https://doi.org/10.1016/j.jhep.2020.03.039

[7] Lazarus, J.V., Newsome, P.N., Francque, S.M., Kanwal, F., Terrault, N.A. and Rinella, M.E. (2023) Reply: A Multi-Society Delphi Consensus Statement on New Fatty Liver Disease Nomenclature. *Hepatology*, **79**, E93-E94. https://doi.org/10.1097/hep.0000000000000696

[8] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 161-166.

[9] 施启军, 潘峰, 龙福海, 等. 特征选择方法研究综述[J]. 微电子学与计算机, 2022, 39(3): 1-8.

[10] 李绒, 李敏, 杨鹏程, 等. 基于女性体检数据的非酒精性脂肪肝风险因素筛选及预测模型[J]. 军事医学, 2022, 46(2): 140-144, 149.

[11] 张培文. 基于 stacking 融合模型的脂肪肝致病影响因素的筛选分析[D]: [硕士学位论文]. 重庆: 重庆大学, 2022.

[12] Zhao, M., Song, C., Luo, T., Huang, T. and Lin, S. (2021) Fatty Liver Disease Prediction Model Based on Big Data of Electronic Physical Examination Records. *Frontiers in Public Health*, **9**, Article 668351. https://doi.org/10.3389/fpubh.2021.668351

[13] Huang, G., Jin, Q. and Mao, Y. (2023) Predicting the 5-Year Risk of Nonalcoholic Fatty Liver Disease Using Machine Learning Models: Prospective Cohort Study. *Journal of Medical Internet Research*, **25**, e46891. https://doi.org/10.2196/46891

[14] Wang, H., Cheng, W., Hu, P., Ling, T., Hu, C., Chen, Y., *et al*. (2024) Integrative Analysis Identifies Oxidative Stress Biomarkers in Non-Alcoholic Fatty Liver Disease via Machine Learning and Weighted Gene Co-Expression Network Analysis. *Frontiers in Immunology*, **15**, Article 1335112. https://doi.org/10.3389/fimmu.2024.1335112

[15] Peng, H., Duan, S., Pan, L., Wang, M., Chen, J., Wang, Y., *et al*. (2023) Development and Validation of Machine Learning Models for Nonalcoholic Fatty Liver Disease. *Hepatobiliary & Pancreatic Diseases International*, **22**, 615-621. https://doi.org/10.1016/j.hbpd.2023.03.009

[16] 黄娅. 随机森林-Lasso Logistic 回归模型筛选脂肪肝健康风险因素效果研究[D]: [硕士学位论文]. 成都: 西南交通大学, 2020.

[17] 秦嘉怡, 周学谦, 孟祥勇, 等. 瘦型代谢功能障碍相关脂肪性肝病风险预测模型的构建与验证[J]. 陆军军医大学学报, 2025, 47(9): 969-979.

[18] 李楠, 王雪莹, 郭佳桐, 等. 中青年人群非酒精性脂肪肝发生风险预测模型的建立[J]. 中国慢性病预防与控制, 2021, 29(3): 167-171.

[19] 张家丽. 两种机器学习方法构建非酒精性脂肪性肝病患病风险预测模型的研究[D]: [硕士学位论文]. 桂林: 桂林医学院, 2022.

[20] 高福来, 谢长顺, 张利利. 基于血清 Betatrophin 水平的非酒精性脂肪肝列线图预测模型的建立与分析[J]. 中国医药导报, 2019, 16(10): 103-106.

[21] 王林, 夏春玲, 范玲. 基于职业护士健康队列研究建立护士非酒精性脂肪肝风险预测模型[J]. 护理研究, 2020, 34(14): 2445-2451.

[22] 徐红平, 李旭, 王浩, 等. 绝经前 HR+乳腺癌患者内分泌治疗相关脂肪肝预测模型的研究[J]. 现代预防医学, 2020, 47(7): 1332-1335.

[23] 雷丽, 郭望, 李运明. 基于机器学习与生存模型建立脂肪肝 Joint 联合预测模型[J]. 现代预防医学, 2021, 48(17): 3259-3264.

[24] Cao, T., Zhu, Q., Tong, C., Halengbieke, A., Ni, X., Tang, J., *et al*. (2024) Establishment of a Machine Learning Predictive Model for Non-Alcoholic Fatty Liver Disease: A Longitudinal Cohort Study. *Nutrition, Metabolism and Cardiovascular Diseases*, **34**, 1456-1466. https://doi.org/10.1016/j.numecd.2024.02.004

[25] 陈霆, 蒋伏松, 朱兴敏, 等. 基于随机森林法的 2 型糖尿病合并非酒精性脂肪肝预测模型[J]. 中国数字医学, 2018, 13(11): 61-63.

[26] Razmpour, F., Daryabeygi-Khotbehsara, R., Soleimani, D., Asgharnezhad, H., Shamsi, A., Bajestani, G.S., *et al*. (2023) Application of Machine Learning in Predicting Non-Alcoholic Fatty Liver Disease Using Anthropometric and Body Composition Indices. *Scientific Reports*, **13**, Article No. 4942. https://doi.org/10.1038/s41598-023-32129-y

[27] Qin, S., Hou, X., Wen, Y., Wang, C., Tan, X., Tian, H., *et al*. (2023) Machine Learning Classifiers for Screening Non-alcoholic Fatty Liver Disease in General Adults. *Scientific Reports*, **13**, Article No. 3638. https://doi.org/10.1038/s41598-023-30750-5

[28] 吕航, 王昊, 刘媛, 等. 基于决策树的中医人格体质对 2 型糖尿病患者伴发非酒精性脂肪肝病风险的预测研究[J]. 中国中医基础医学杂志, 2017, 23(9): 1257-1259.

[29] 王菊芳, 范瑞, 杜金满. 非肥胖人群脂肪肝患病风险的神经网络分析[J]. 现代实用医学, 2020, 32(6): 668-670.

[30] 梁宏涛, 刘硕, 杜军威, 等. 深度学习应用于时序预测研究综述[J]. 计算机科学与探索, 2023, 17(6): 1285-1300.

[31] 雪峰豪, 蒋海波, 唐聃. 深度学习在健康医疗中的应用研究综述[J]. 计算机科学, 2023, 50(4): 1-15.

[32] Wu, C., Chu, T. and Jang, J.R. (2021) Current-Visit and Next-Visit Prediction for Fatty Liver Disease with a Large-Scale Dataset: Model Development and Performance Comparison. *JMIR Medical Informatics*, **9**, e26398. https://doi.org/10.2196/26398

[33] Deng, Y., Ma, Y., Fu, J., Wang, X., Yu, C., Lv, J., *et al*. (2023) A Dynamic Machine Learning Model for Prediction of NAFLD in a Health Checkup Population: A Longitudinal Study. *Heliyon*, **9**, e18758. https://doi.org/10.1016/j.heliyon.2023.e18758

[34] Yip, T.C.-F., Ma, A.J., Wong, V.W.-S., Tse, Y.-K., Chan, H.L.-Y., Yuen, P.-C., *et al*. (2017) Laboratory Parameter-based Machine Learning Model for Excluding Non-alcoholic Fatty Liver Disease (NAFLD) in the General Population. *Alimentary Pharmacology & Therapeutics*, **46**, 447-456. https://doi.org/10.1111/apt.14172

[35] Weng, S., Hu, D., Chen, J., Yang, Y. and Peng, D. (2023) Prediction of Fatty Liver Disease in a Chinese Population Using Machine-Learning Algorithms. *Diagnostics*, **13**, Article 1168. https://doi.org/10.3390/diagnostics13061168

[36] Liu, J., Li, Y., Ma, J., Wan, X., Zhao, M., Zhang, Y., *et al*. (2023) Identification and Immunological Characterization of Lipid Metabolism-Related Molecular Clusters in Nonalcoholic Fatty Liver Disease. *Lipids in Health and Disease*, **22**, Article No. 124. https://doi.org/10.1186/s12944-023-01878-0

[37] 余秋燕, 赵莹, 孙继佳, 等. 典型机器学习算法在脂肪肝分类预测研究中的实现与比较[J]. 数理医药学杂志, 2019, 32(1): 1-3.

[38] Njei, B., Osta, E., Njei, N., Al-Ajlouni, Y.A. and Lim, J.K. (2024) An Explainable Machine Learning Model for Prediction of High-Risk Nonalcoholic Steatohepatitis. *Scientific Reports*, **14**, Article No. 8589. https://doi.org/10.1038/s41598-024-59183-4

[39] Yang, B., Lu, H. and Ran, Y. (2024) Advancing Non-Alcoholic Fatty Liver Disease Prediction: A Comprehensive Machine Learning Approach Integrating SHAP Interpretability and Multi-Cohort Validation. *Frontiers in Endocrinology*, **15**, Article 1450317. https://doi.org/10.3389/fendo.2024.1450317

[40] 曾芷青, 杨淞淳, 余灿清, 等. 慢性肾脏病发病风险预测模型研究进展[J]. 中华流行病学杂志, 2023, 44(3): 498-503.