

基于电子病历的慢性病辅助诊断模型与特征边际效应分析

王颖帅¹ 王智飞² 万艳丽¹ 谢莉琴¹ 秦盼盼¹ 雷行云¹ 陈庆锟¹ 胡红濮^{1,3}

(¹ 中国医学科学院/北京协和医学院医学信息研究所 北京 100020

² 人力资源和社会保障部信息中心 北京 100716

³ 中国医学科学院/北京协和医学院马克思主义学院人文和社会科学学院 北京 100730)

【摘要】 目的/意义 基于电子病历数据中的主诉、生命体征、辅助检查、现病史等特征，预测 6 种常见慢性病患概率，为基层医生辅助诊断提供参考。**方法/过程** 基于公开发布的电子病历数据，采用数据清洗、自然语言处理等技术，建设精细化特征工程，采用逻辑回归、朴素贝叶斯、深度神经网络、决策树、支持向量机和轻量级梯度提升机 6 种机器学习算法建模。采用精确度、召回率、AUC 等指标评估模型性能，使用 SHAP 值增强模型可解释性。**结果/结论** 轻量级梯度提升树效果较好，能够有效预测慢性病风险。特征分析显示主诉、专科检查、现病史等对模型预测的边际效应影响较大，为个性化预防提供了科学依据。

【关键词】 人工智能；基层服务；慢性病；辅助诊断

【中图分类号】 R-058 **【文献标识码】** A **【DOI】** 10.3969/j.issn.1673-6036.2025.10.003

The Auxiliary Diagnosis Model of Chronic Diseases Based on Electronic Medical Records and Analysis of Feature Marginal Effects

WANG Yingshuai¹, WANG Zhifei², WAN Yanli¹, XIE Liqin¹, QIN Panpan¹, LEI Xingyun¹, CHEN Qingkun¹, HU Hongpu^{1,3}

¹Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China;

²Information Center of the Ministry of Human Resources and Social Security, Beijing 100716, China; ³School of Marxism, School of Humanities and Social Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China

【Abstract】 Purpose/Significance To predict the probability of six common chronic diseases based on features from electronic medical record (EMR) data, including chief complaints, vital signs, ancillary examinations, and history of present illness, so as to provide references for auxiliary diagnosis for primary care physicians. **Method/Process** Using publicly available EMR data, techniques such as data cleaning and natural language processing (NLP) are employed to create a detailed feature. Models are built by using six machine learning algorithms: logistic regression, naive Bayes, deep neural networks, decision trees, support vector machine, and light gradient boosting machine (LightGBM). The performance of models is evaluated using multiple metrics, including accuracy, recall, and AUC. SHAP values are used to enhance the interpret ability of the models. **Result/Conclusion** LightGBM shows the best performance and can effectively predict chronic disease risk. Feature analysis reveals that chief complaints, specialty examinations, and history of present illness have a significant marginal effect on model predictions, providing a scientific basis for personalized prevention strategies.

【修回日期】 2025-09-03

【作者简介】 王颖帅，博士，发表论文 20 余篇；通信作者：胡红濮。

【基金项目】 中国卫生经济学会医药卫生体制改革重点研究课题；中国医学科学院医学与健康科技创新工程项目（项目编号：2022-I2M-1-019）；中央级公益性科研院所基本科研业务费专项（项目编号：2024-ZHCH630-01）。

[**Keywords**] artificial intelligence (AI); primary care services; chronic diseases; assisted diagnosis

1 引言

数智化技术赋能基层慢性病辅助诊断可以分为 3 个方向,一是基于统计学分析方法的辅助诊断,二是基于机器学习的辅助诊断,三是基于大模型的辅助诊断。统计学方法在基层慢性病管理中可提供多方面参考。例如,利用上下文学习和合成数据提升诊断性能^[1],通过风险模型分析睡眠与慢性病的关系^[2],结合调查数据探讨卫生服务利用关键因素^[3],以及评估患者对健康管理服务偏好等^[4],为优化基层医疗体系提供科学依据。机器学习在基层慢性病诊断中发挥重要作用。例如,通过融合朴素贝叶斯、随机森林等算法,基于美国国家癌症研究所的癌症监测、流行病学和结果数据库构建肿瘤辅助诊断模型,提升预测精准性^[5];基于多中心数据,使用 9 种方法建立疾病风险预测模型,使用 SHAP 值解释并开发在线工具支持临床决策^[6];有研究^[7]发现基于韩国、日本、英国队列数据的集成模型对 2 型糖尿病风险预测效果良好;还有研究^[8]系统分析疾病共病预测流程,涵盖数据提取、聚类、网络分析与风险识别,提出慢性病防治新思路。大模型在基层慢性病诊断领域展现出较大潜力。例如,通过多模态影像与认知评分结构化数据的融合,构建轻量级辅助诊断网络,提升阿尔茨海默病诊断效率^[9];基于大语言模型与深度学习的图像-语言集成系统,为基层糖尿病护理和视网膜病变筛查提供解决方案,提升患者自我管理与慢性病服务效率^[9-10]。此外,有研究^[11-12]采用引导思维链微调策略,验证了用于临床决策的医学大模型具有较强的实用性。

电子病历涵盖主诉、现病史、生命体征等丰富的临床信息,为慢性病诊断提供了宝贵的数据基础。本研究基于电子病历,通过精细化特征工程,在有限算力下构建辅助决策模型,有助于优化基层医疗资源配置和提升服务水平。

2 材料与方法

2.1 数据来源

基于中文电子病历数据集 CEMRs^[13](<https://github.com/YangzlTHU/C-EMRs>),共采集 18 624 条样本,17 个特征。特征包括主诉、手术史、生命体征、专科情况、一般情况、过敏史、营养状态、自杀倾向、专科检查、手术外伤史、并发症、现病史、生育史、辅助检查、个人史、既往史和家族史,将入院初诊作为标签。选择糖尿病、高血压、慢性阻塞性肺疾病、心律失常、哮喘、痛风 6 类发病率高、具有代表性的慢性病作为采集对象^[14]。

2.2 实验环境及模型参数

实验环境为 NVIDIA Tesla V100 的 GPU 服务器,显存 32GB,内存 512GB,32 核 CPU,代码基于 Python 3.11 实现。轻量级梯度提升机(light gradient boosting machine, LightGBM)最佳参数设置为:叶子节点数量=21(调试范围 10~30),树最大深度=5(调试范围 3~7),迭代轮数=30(调试范围 10~100),叶子节点最小样本数=20(调试范围 1~30),学习率=0.05(调试范围 0.01~0.1),L1 正则化参数=0.1(调试范围 0.05~0.5),L2 正则化参数=0.1(调试范围 0.05~0.3)。基于网格搜索自动调试得出最佳参数。特征重要性使用 LightGBM 的信息分裂增益指标,计算在所有树中的累积信息增益。此外,通过博弈论 SHAP 方法量化特征边际贡献。

2.3 数据处理与统计分析

数据预处理包括文本编码统一、分词、去停用词等,使用 jieba 分词工具,制作医疗术语表。按照 8:1:1 的比例将数据集划分为训练集、验证集和测试集。为确保训练集、验证集、测试集中包含所有类别的疾病,采用分层抽样。实验结果基于对数据集的 3 次随机划分取其平均值。

统计分析方法包括词云图、数据统计分布、特征解释等。词云图通过对词语大小、颜色和位置的编码,呈现文本中词语出现频率或重要性的图形表示,能够快速识别文本中的主要概念和主题关联,是医学信息挖掘和疾病研究的有效工具。统计分布是指数据集中变量在其取值范围内出现的频率或概率模式,描述数据值分布状况,揭示数据的集中趋势、离散程度等。特征解释是对数据集中变量的含义、重要性及其与目标变量之间关系进行阐释的过程。

2.4 特征设计

2.4.1 统计特征提取器 对结构化表格中的类别变量、数值变量和文本变量进行处理,类别变量表示为 $x_{cat} \in R^{d_1}$, d_1 是类别数量,每个类别对应查找表中的 embedding 词嵌入向量;数值变量表示为 $x_{num} \in R^{d_2}$, d_2 是数值变量的数量,归一化缩放到 $[0, 1]$ 之间;文本变量表示为 $x_{text} \in R^n$, n 是关键词数量,然后利用语言模型词向量进行编码。

$$X = \text{Embed}(t) \quad (1)$$

$X \in R^{n \times d}$, d 表示词向量维度,编码表示如下。

$$X = \begin{bmatrix} \text{word}_1 \\ \text{word}_2 \\ \dots \\ \text{word}_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix} \quad (2)$$

2.4.2 局部特征提取器 利用卷积神经网络对电子病历进行局部特征学习,给定一个大小为 $s \times d$ 的卷积核。

$$W = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1d} \\ w_{21} & w_{22} & \dots & w_{2d} \\ \dots & \dots & \dots & \dots \\ w_{s1} & w_{s2} & \dots & w_{sd} \end{bmatrix} \quad (3)$$

通过将 W 在 X 上滑动并进行卷积运算,得到粒度为 s 的局部特征 v 。

$$v = \text{Conv}(X, W) \quad (4)$$

2.4.3 全局特征提取器 类别特征和文本特征提取模块由一个向量嵌入层和多个叠加的 Transformer 层组成,Transformer 层由多头注意力层和位置前馈层组成。类别变量特征提取可表示为 $E_{cat} = \text{Cat_Encoding}(x_{cat})$, 文本变量特征提取可表示为

$E_{text} = \text{Text_Encoding}(x_{text})$, 数值变量特征提取可表示为 $E_{num} = \text{Num_Encoding}(x_{num})$, 见图 1。

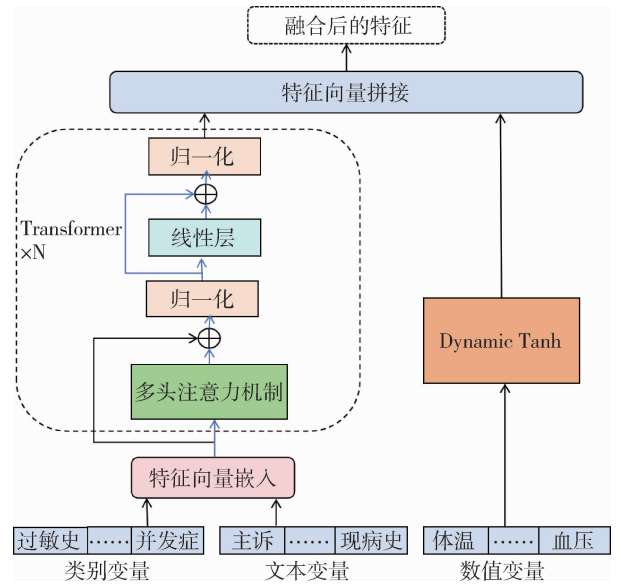


图 1 基于 Transformer 的全局特征提取

Transformer 是一种强大的特征提取工具,可以有效捕捉序列之间的上下文依赖关系。在对类别变量和文本变量进行嵌入向量表示后,利用 Transformer 进行特征提取,自注意力机制如下。

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

其中, Q, K, V 分别代表查询矩阵、键矩阵和值矩阵, d_k 表示键值向量的维度, Transformer 模块可以堆叠 N 层。Dynamic Tanh^[15] 是一种新颖的层归一化方法,数学表达如下。其中, α 是可学习的标量参数,负责动态调整输入的缩放程度; γ 和 β 是可学习的向量参数,用于将输出缩放到合适的范围,移除了传统层归一化操作,减少了计算量,保留了模型对特征尺度的自适应调整能力。

$$\text{DyT}(x) = \gamma \times \tanh(\alpha x) + \beta \quad (6)$$

2.5 模型选择和实施

本研究目标是构建输入与输出的映射关系。通过对模型参数的优化,推动预测值 $\hat{y}_i = f(x_i)$ 向真实标签 y_i 逼近,从电子病历数据中提取有价值的信息,以提高医生的临床辅助决策水平。选取 6 种具有代表性的机器学习方法进行慢性病患者病概率预

测，分别是逻辑回归、朴素贝叶斯、深度神经网络、决策树、支持向量机以及 LightGBM。其中逻辑回归作为基础且可解释的线性模型，是评估基线性能的良好选择；朴素贝叶斯是高效的概率分类器，在数据量有限的情况下表现稳定；深度神经网络由多层神经元组成，通过非线性变换和特征学习，能捕捉复杂模式；决策树通过对特征的递归划分，可实现快速决策且易于解释；支持向量机擅长在高维空间中找到最优分类超平面，对于处理结构化数据具有较强的泛化能力；而 LightGBM 是一种先进的集成学习方法，可以捕捉非线性关系。通过对比不同模型表现，评估不同范式利用电子病历数据进行慢性病风险预测的有效性，确定适合该任务的模型。

2.6 模型评估与解释

采用以下指标和方法进行模型评估和解释。一是精确度，衡量模型预测为正类的样本中，实际为正类的比例。设 TP 为真正例，FN 为假反例，FP 为假正例，TN 为真反例。

$$\text{精确度} = \frac{TP}{TP + FP} \tag{7}$$

二是召回率，表示模型正确预测为正例的样本占有所有实际正例样本的比例。

$$\text{召回率} = \frac{TP}{TP + FN} \tag{8}$$

三是 F1 值，是精确度和召回率的调和平均数，旨在平衡二者之间的关系。

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{9}$$

四是准确度，表示预测正确的样本数占总样本数的比例，反映模型在所有预测中正确分类的能力。

$$\text{准确度} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

五是受试者工作特征（receiver operating characteristic, ROC）曲线，是评估模型性能的图形化工具，ROC 曲线下面积（area under curve, AUC）表示随机选择一个正例和一个反例，模型将正例排在反例前面的概率。其中，TPR 表示召回率，FPR 表示假正例率，真正例率 AUC 越大，表示预测效果越好。

$$AUC = \int_0^1 TPR(FPR) d(FPR) \tag{11}$$

六是 SHAP 值，是一种基于博弈论的模型可解释方法，为模型的预测提供特征贡献度量。通过计算特征对模型输出的边际贡献来量化其影响程度。对于预测函数 f 和输入实例 x ，特征 i 的 SHAP 值 φ_i 的计算方式如下。其中 N 是全部特征的集合， S 是不包含特征 i 的子集， $f_x(S)$ 是只使用特征子集的模型预测， $|S|$ 和 $|N|$ 分别表示子集 S 和全部特征集 N 的基数。

$$\varphi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f_x(S \cup \{i\}) - f_x(S)] \tag{12}$$

3 结果

3.1 统计描述

在慢性病辅助诊断任务中，结构化表格数据是基本表现形式，部分样例，见表 1。数据类型包括非结构化文本和结构化表格，表格特征分为数值型和类别型，如主诉、手术史、现病史等是文本特征，体温、脉搏、血压等生命体征是数值特征，一般情况、营养状况、过敏史等是类别特征。数据集中有 23 条数据的 17 个特征均为空，将其过滤，得到有效数据 18 601 条，不同疾病的分布，见图 2。

表 1 数据样例

字段	样例
入院初诊（标签）	糖尿病
主诉	蛋白尿 10 余年，恶心 1 月，腹泻伴纳差 1 周
手术史	2016 年前因冠心病行“心脏支架置入术”
生命体征	体温：36.8℃；脉搏：60 次/分，规则；呼吸：19 次/分，规则；血压：130/60 mmHg
专科检查	神清，颈软，颜面部轻度浮肿，贫血貌，双肺呼吸音清晰，无干湿啰音，心率 60bpm，律齐，无杂音，腹软，无压痛及反跳痛，双肾区无叩痛，腰骶部及双下肢轻度水肿
手术外病史	2016 年前因冠心病行“心脏支架置入术”。无外病史
并发症	无
现病史	患者 10 多年前行常规检查时发现尿蛋白 2+，当时血肌酐正常，无颜面及双下肢水肿、四肢关节疼痛、发热、皮疹、肉眼血尿等不适，未行特殊治疗。于 4 年前血肌酐开始出现增高，当时血肌酐约 100umol/L 左右，在湖北省人民医院行肾穿刺活检确诊为“糖尿病肾病”，并给予对症治疗
……	……

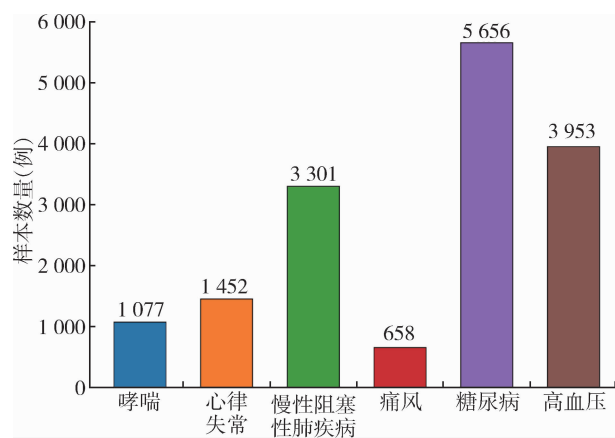


图2 常见慢性病样本本分布

3.2 慢性病症状分析

糖尿病和高血压主诉词云图，见图3。



图3 慢性病词云图

糖尿病词云图显示“血糖高”“多饮”和“多尿”为显著词汇，高血压词云图显示“反复”“头晕”“头昏”和“加重”为显著词汇，反映了核心症状与患者关注点。同时，“发现”“年”“月”等词表明了病情发展、诊断及管理的时间维度与过程。

3.3 特征重要性

6 种常见慢性病的特征重要性，见图4。

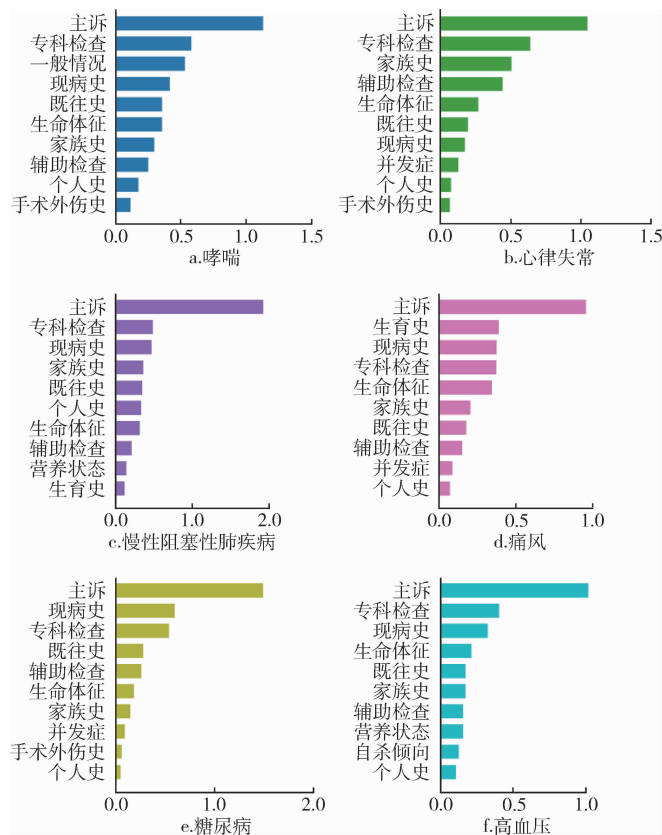


图4 慢性病特征重要性

3.4 模型评估

为了筛选最佳模型，对逻辑回归、朴素贝叶斯、深度神经网络、决策树、支持向量机和 LightGBM 这 6 种机器学习算法进行 6 折交叉验证，计算平均 AUC 和 F1 值，见表 2。可以看出，LightGBM 在 AUC 和 F1 值上均优于其他基线模型。

表 2 不同模型预测不同慢性病的 AUC 和 F1 值

模型名称	慢性病名称	AUC	F1
逻辑回归	哮喘	0.866 4	0.826 7
	心律失常	0.766 7	0.701 3
	慢性阻塞性肺疾病	0.860 6	0.856 4
	痛风	0.733 9	0.681 7
	糖尿病	0.773 0	0.813 9
	高血压	0.728 8	0.776 5

续表 2

模型名称	慢性病名称	AUC	F1
朴素贝叶斯	哮喘	0.829 6	0.815 7
	心律失常	0.757 7	0.712 8
	慢性阻塞性肺疾病	0.828 6	0.816 9
	痛风	0.653 2	0.673 8
	糖尿病	0.760 7	0.810 2
支持向量机	高血压	0.735 6	0.765 3
	哮喘	0.890 4	0.841 2
	心律失常	0.887 4	0.734 1
	慢性阻塞性肺疾病	0.917 6	0.875 2
	痛风	0.812 0	0.687 9
决策树	糖尿病	0.893 2	0.836 9
	高血压	0.850 7	0.745 8
	哮喘	0.880 1	0.852 7
	心律失常	0.829 2	0.721 6
	慢性阻塞性肺疾病	0.934 2	0.892 5
深度神经网络	痛风	0.804 5	0.690 2
	糖尿病	0.887 2	0.821 5
	高血压	0.803 4	0.720 1
	哮喘	0.880 5	0.835 9
	心律失常	0.863 5	0.741 1
LightGBM	慢性阻塞性肺疾病	0.912 5	0.887 3
	痛风	0.764 6	0.683 7
	糖尿病	0.893 6	0.843 6
	高血压	0.838 6	0.746 9
	哮喘	0.919 5	0.902 1
	心律失常	0.897 0	0.761 1
	慢性阻塞性肺疾病	0.969 6	0.900 3
	痛风	0.889 7	0.705 2
	糖尿病	0.923 4	0.883 6
	高血压	0.879 7	0.782 7

初筛选出的最佳模型 LightGBM，在不同慢性病的精确度、召回率、F1 值、准确率，见表 3。基于 LightGBM 模型，对引入的局部特征和全局特征进行消融实验，验证其精细化特征工程的价值。

表 3 LightGBM 不同慢性病消融实验结果

消融实验	慢性病	精确度	召回率	F1	准确率
LightGBM	哮喘	0.910 3	0.888 9	0.902 1	0.901 5
	心律失常	0.800 3	0.732 6	0.761 1	0.868 9
	慢性阻塞性肺疾病	0.890 4	0.919 2	0.900 3	0.908 7
	痛风	0.785 7	0.636 2	0.705 2	0.861 7
	糖尿病	0.889 2	0.876 9	0.883 6	0.908 9
	高血压	0.749 8	0.820 3	0.782 7	0.879 2

续表 3

消融实验	慢性病	精确度	召回率	F1	准确率
LightGBM + 精细化特征工程	哮喘	0.918 8	0.891 4	0.911 6	0.909 3
	心律失常	0.811 7	0.745 2	0.768 3	0.872 2
	慢性阻塞性肺疾病	0.901 2	0.923 5	0.910 7	0.913 8
	痛风	0.791 2	0.658 2	0.712 5	0.870 8
	糖尿病	0.891 3	0.879 2	0.889 7	0.910 8
	高血压	0.750 6	0.825 6	0.788 9	0.880 3

3.5 模型解释

最佳模型 LightGBM 中特征的解释分为全局解释和个体样本解释，慢性病预测全局特征边际效应，见图 5。

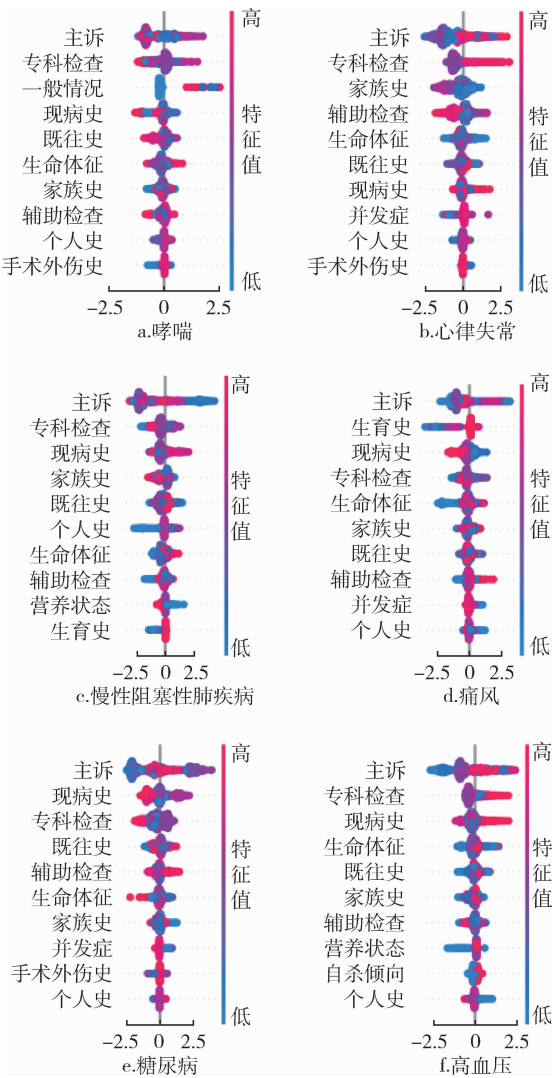


图 5 集成模型 LightGBM 特征边际效应

个体样本特征边际效应，见图 6，该样本中并发症、既往史、主诉、专科检查、生命体征和一般

情况均影响患病概率。

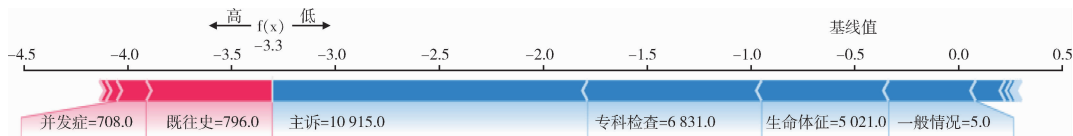


图 6 单样本特征解释

4 讨论

本研究提出基层慢性病辅助诊断机器学习模型，评估不同慢性病风险因子，为医生决策提供参考^[16-18]。对于糖尿病和高血压，主诉、现病史、专科检查是排名前 3 位的特征。主诉中如果出现“多饮、多尿”或“头晕、头痛”等高频词，医生应给予高度关注，并结合现病史询问，安排相关专科检查，以实现早期筛查和干预。对于慢性阻塞性肺疾病，主诉、现病史、专科检查是关键特征。当患者出现持续性咳嗽、咳痰、呼吸困难等主诉时，医生应了解其吸烟史和职业暴露史，尽快安排肺功能检查，以明确诊断，指导后续治疗。对于心律失常，主诉、专科检查和家族史是排名前 3 位的特征。在问诊过程中，医生除了关注心慌、心悸等主诉和心电图等专科检查，还应询问患者家族中是否有心血管病史。对于哮喘，主诉、专科检查和一般情况是关键特征。医生除了要关注典型的哮喘症状和肺功能检查，还应重视患者的一般情况，包括过敏史、生活环境、精神状态等。哮喘的诊疗需要药物结合患者生活状态进行综合管理。对于痛风，主诉、生育史和现病史是排名前 3 位的特征。问诊时，医生应注意询问女性患者的生育史，并结合关节红肿热痛等主诉和痛风发作的病史进行综合判断，以便进行早期识别和风险评估。

既往研究^[19-20]表明机器学习模型可以较好地预测常见慢性病患者患病概率。本研究从诊断准确性、患者管理和可解释性 3 个方面分析在基层使用机器学习方法的可行性。一是机器学习利用电子病历数据挖掘隐藏模式，可以提升早期或非典型病例识别的准确性。二是机器学习基于多维特征，可进行更

精细的边际效应分析、患者风险分层等，在提升患者管理精细度上具有可行性。三是相较于深度学习“黑箱”模型，LightGBM 在提供预测结果的同时还可向基层医生解释原因，增强医生的信任度，引导其进行更深入的检查和判断。

本研究模型具有较高精准度和较强可解释性，在基层慢性病防控场景具有较大应用潜力。通过将卷积神经网络的局部特征提取与 Transformer 的全局注意力机制结合，并借助消融实验，验证了精细化特征工程的价值。本研究尚存在局限，例如，数据集尽管包括常见慢性病，但主要是面向中国人群，可能导致模型的泛化能力受限。模型可能对未纳入的特征（如环境因素等）预测能力有限，有待进一步优化。

5 结语

本研究通过深度分析电子病历数据，应用机器学习方法，构建常见慢性病患者患病概率预测模型，可以辅助基层医生提升慢性病早筛准确率，增强预防效果。使用逻辑回归、朴素贝叶斯、深度神经网络、决策树、支持向量机、LightGBM，识别影响常见慢性病发病风险的关键因素，如主诉、专科检查、现病史等，尽管存在数据来源种群单一和泛化能力弱等局限，但模型在测试集上显示出较好的评估性能和特征可解释性。未来研究应探索更多样化的数据集和风险因子，以进一步赋能基层慢性病管理。

作者贡献：王颖帅负责实验实施、论文撰写；王智飞负责文献调研；万艳丽负责提供技术指导；谢莉琴负责图表制作；秦盼盼负责数据标准化；雷行云负责代码审核；陈庆锐负责方法设计；胡红濮负责

提供指导、论文审核。

利益声明：所有作者均声明不存在利益冲突。

参考文献

- 1 HOLLMANN N, MÜLLER S, PURUCKER L, et al. Accurate predictions on small data with a tabular foundation model [J]. *Nature*, 2025, 637 (8045): 319–326.
- 2 ZHENG N S, ANNIS J, MASTER H, et al. Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the all of us research program [J]. *Nature medicine*, 2024, 30 (9): 2648–2656.
- 3 周翔, 段桂敏, 万圣洁, 等. 中老年慢病患者卫生服务利用及影响因素研究: 基于 CHARLS 数据的实证分析 [J]. *中国卫生事业管理*, 2025, 42 (2): 172–176, 185.
- 4 王萱萱, 蔡思好, 王刚, 等. 老年多重慢性病患者对社区健康管理服务模式的选择偏好研究 [J]. *中国卫生政策研究*, 2025, 18 (3): 32–40.
- 5 GUO Q H, XIE F C, ZHONG F M, et al. Application of interpretable machine learning algorithms to predict distant metastasis in ovarian clear cell carcinoma [J]. *Cancer medicine*, 2024, 13 (7): e7161.
- 6 HOU F, ZHU Y, ZHAO H, et al. Development and validation of an interpretable machine learning model for predicting the risk of distant metastasis in papillary thyroid cancer: a multicenter study [J]. *Eclinicalmedicine*, 2024, 77 (11): 102913.
- 7 LEE H, HWANG S H, PARK S, et al. Prediction model for type 2 diabetes mellitus and its association with mortality using machine learning in three independent cohorts from South Korea, Japan, and the UK: a model development and validation study [J]. *Eclinicalmedicine*, 2025, 80 (2): 103069.
- 8 DUO X U, ZESHUI X U. Machine learning applications in preventive healthcare: a systematic literature review on predictive analytics of disease comorbidity from multiple perspectives [J]. *Artificial intelligence in medicine*, 2024, 156 (10): 102950.
- 9 王光明, 柏正尧, 宋帅, 等. 阿尔茨海默病辅助诊断的多模态数据融合轻量级网络 [J]. *浙江大学学报 (工学版)*, 2025, 59 (1): 39–48.
- 10 XIAO H, ZHOU F, LIU X, et al. A comprehensive survey of large language models and multimodal large language models in medicine [J]. *Information fusion*, 2025, 117 (5): 102888.
- 11 LI J, GUAN Z, WANG J, et al. Integrated image-based deep learning and language models for primary diabetes care [J]. *Nature medicine*, 2024, 30 (10): 2886–2896.
- 12 LIU X, LIU H, YANG G, et al. A generalist medical language model for disease diagnosis assistance [J]. *Nature medicine*, 2025, 31 (3): 932–942.
- 13 YANG Z, HUANG Y, JIANG Y, et al. Clinical assistant diagnosis for electronic medical record based on convolutional neural network [J]. *Scientific reports*, 2018, 8 (1): 1–9.
- 14 周宝森. 慢性病管理的挑战、策略与进展 [J]. *健康中国观察*, 2025 (1): 78–81.
- 15 ZHU J, CHEN X, HE K, et al. Transformers without normalization [C]. *Nashville: The Computer Vision and Pattern Recognition Conference*, 2025.
- 16 张馨丹, 和锦浩, 尹悦, 等. 整合型服务体系下县域慢病管理中心建设现状及启示 [J]. *中国卫生经济*, 2025, 44 (4): 79–83.
- 17 周佳坤, 张明哲, 陈恩琳, 等. 数字健康技术在基层医疗卫生机构慢性病健康管理的应用: 以糖尿病、高血压为例 [J]. *军事护理*, 2025, 42 (4): 62–65.
- 18 李婷婷, 苏鹏, 陈晋波, 等. 基于机器学习比较 2 型糖尿病患者发生糖尿病肾脏疾病风险预测模型的研究 [J]. *中国糖尿病杂志*, 2025, 33 (4): 241–247.
- 19 蒋美艳, 张辉. 机器学习算法对心脏病预测效能的研究 [J]. *中国医学物理学杂志*, 2024, 41 (7): 905–909.
- 20 王英杰, 孙高峰. 2 型糖尿病预测模型研究进展 [J]. *预防医学*, 2025, 37 (4): 369–372, 377.