

医疗大数据驱动的疾病预测系统架构与应用研究

杨冰倩^{1,2*}, 张艳敏^{1,2}, 崔海江¹, 孙凤杰¹, 赵占坤¹

¹河北软件职业技术学院, 河北 保定

²保定市人工智能 + 智慧交通领域重点实验室, 河北 保定

收稿日期: 2026年9月6日; 录用日期: 2026年10月7日; 发布日期: 2026年10月15日

摘要

本文针对医学影像识别与临床风险预测两类典型场景, 设计并构建了一套基于医疗大数据的疾病预测原型系统, 给出了由数据处理层、AI模型层、应用服务层和用户交互层组成的四层系统架构。在影像端, 系统以预训练的ResNet50为骨干网络, 通过迁移学习完成胸部X光片的肺炎识别, 并使用Softmax输出类别概率; 在临床风险预测方面, 系统集成逻辑回归、随机森林和XGBoost等机器学习模型, 实现冠心病风险概率及风险概率预测。同时引入SHAP方法对模型决策进行解释, 增强了结果的可信度与可用性。

关键词

医疗大数据, 疾病预测, 深度学习, 机器学习, 模型可解释性

Research on the Architecture and Applications of Medical Big Data-Driven Disease Prediction Systems

Bingqian Yang^{1,2*}, Yanmin Zhang^{1,2}, Haijiang Cui¹, Fengjie Sun¹, Zhankun Zhao¹

¹Department of Software Engineering, Hebei Software Institute, Baoding Hebei

²Baoding Key Laboratory of Artificial Intelligence and Intelligent Transportation, Baoding Hebei

Received: September 6, 2026; accepted: October 7, 2026; published: October 15, 2026

Abstract

This paper focuses on two typical scenarios—medical image recognition and clinical risk prediction—

*通讯作者。

文章引用: 杨冰倩, 张艳敏, 崔海江, 孙凤杰, 赵占坤. 医疗大数据驱动的疾病预测系统架构与应用研究[J]. 计算机科学与应用, 2026, 16(10): 1-9. DOI: 10.12677/csa.2026.1610305

and designs and implements a disease prediction prototype system based on healthcare big data. A four-layer system architecture is proposed, consisting of a data processing layer, an AI model layer, an application service layer, and a user interaction layer. For the imaging component, the system employs a pre-trained ResNet50 as the backbone network, performs pneumonia recognition on chest X-ray images through transfer learning, and outputs class probabilities using Softmax. For clinical risk prediction, the system integrates machine learning models including Logistic Regression, Random Forest, and XGBoost to predict coronary heart disease risk and associated probability. Additionally, the SHAP method is introduced to interpret model decisions, thereby enhancing the credibility and usability of the results.

Keywords

Healthcare Big Data, Disease Prediction, Deep Learning, Machine Learning, Model Interpretability

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着电子病历、医学影像和远程监测等信息化系统的广泛应用，医疗数据规模不断增长，并呈现出影像、文本、信号等多模态特征，为疾病预测和临床决策提供了重要的数据基础[1]。与此同时，医疗数据具有较强的隐私敏感性，在采集、存储和分析等环节面临一定的安全风险[2]。近年来，深度学习在医学影像分析领域得到广泛应用，CNN、ResNet 以及迁移学习等方法为医学影像识别提供了有效技术手段[3]。在临床风险预测领域，机器学习方法也逐渐应用于疾病风险评估，为传统预测方法提供了新的技术路径[4]。

基于此，本文围绕肺炎影像识别和冠心病风险预测两个场景，设计了一套疾病预测系统，构建数据处理层、AI 模型层、应用服务层和用户交互层四层架构，并结合 ResNet50、机器学习模型及 SHAP 模型解释方法，实现疾病预测与结果解释，为医疗数据驱动的疾病预测系统应用提供参考。

2. 医疗大数据与疾病预测相关技术

2.1. 医疗大数据概述

医疗大数据可以理解为在医疗与健康活动过程中产生的、能够被信息系统采集和处理的全部数据集，其来源包括电子病历、医学影像、检验检查报告、可穿戴设备监测数据以及公共卫生数据等。这些数据具有多模态、海量、时序性强和隐私敏感等典型特征，对数据的存储、管理与计算处理能力提出了更高要求[5]。

目前，医疗大数据分析的主要技术路径大致可归纳为三类：基于传统统计分析的方法、基于经典机器学习的方法和基于深度学习的方法。这三类方法在医疗风险预测、疾病辅助诊断、健康管理等场景中均有应用，并常常组合使用以兼顾准确性与可解释性。

随着医疗信息化程度的提高和数据集规模不断扩大，基于机器学习的疾病预测模型在准确率和泛化能力上通常优于传统统计模型，这也是本文选择以机器学习和深度学习为主要技术手段的原因。

2.2. 医学影像疾病识别技术

卷积神经网络(CNN)是医学影像识别领域常用的深度学习模型，能够通过局部感受野、权值共享和池

化等机制自动提取图像特征[6]。近年来, CNN 已广泛应用于胸片、CT 和病理切片等医学影像的分类、分割与检测任务,并在部分任务中取得了接近甚至超过经验丰富临床医生的识别效果[7]。

随着网络深度不断增加,深度神经网络容易出现梯度消失或梯度爆炸等问题,导致模型训练困难。He 等提出的深度残差网络(ResNet)通过引入跨层残差连接,使网络由直接学习目标映射转变为学习残差映射,从而有效改善深层网络的训练问题。ResNet50 作为 ResNet 系列的典型模型,在网络深度、计算复杂度和特征提取能力之间取得了较好的平衡,因此被广泛应用于医学影像分析及迁移学习任务。

在肺炎影像识别研究方面, Rajpurkar 等[8]提出的 CheXNet 采用 DenseNet-121 对胸部 X 光片进行肺炎检测,其研究结果显示模型具有较好的肺炎识别能力。随后, Irvin [9]等构建了 CheXpert 数据集,包含超过 20 万张胸部 X 光片,并提供多种疾病标签及不确定性标注,为肺炎等胸部疾病的智能识别和算法研究提供了重要的数据资源。这些研究为基于深度学习和迁移学习开展肺炎 X 光影像识别提供了较好的理论与实践基础。

2.3. 临床数据疾病预测技术

基于临床指标的疾病风险预测是医疗大数据应用的重要方向之一。现有研究通常利用监督学习方法构建疾病风险预测模型,并结合特征重要性分析和可解释性方法探讨影响疾病发生的关键因素。辛瑞昊等采用 T 检验进行特征筛选,结合多种机器学习模型开展心脏病风险预测,并利用 SHAP 方法对模型预测结果进行解释,为结构化临床数据的风险预测提供了较好的研究范式。

在传统风险评估方法中, Framingham 风险评分通过多个临床指标的线性组合评估心血管事件发生风险,是经典的心血管风险预测工具。但由于其主要基于线性关系进行建模,对于临床指标之间复杂的非线性关系和高阶交互作用刻画能力有限。相比之下,随机森林通过集成多棵决策树进行预测,能够较好地处理特征之间的非线性关系,并可通过特征重要性评价变量的相对贡献。XGBoost 则采用梯度提升框架,并结合正则化等策略进一步提升模型的预测性能,在结构化临床数据的风险预测任务中具有较好的应用效果,因此可作为本文临床风险预测模块的候选模型。

对于具有明显时间依赖性的临床监测数据,长短期记忆网络(LSTM)通过门控机制能够学习序列数据中的长期依赖关系,已应用于 ICU 监测等场景。考虑到本文系统主要针对患者单次就诊时输入的静态临床指标进行风险预测,暂不涉及连续监测数据,因此采用适用于结构化数据的机器学习模型开展预测,时序深度学习模型可作为后续功能扩展的研究方向。

2.4. 模型可解释技术

医疗场景对模型可解释性有较强需求,原因在于医生和患者通常希望了解模型“为什么会给出这样的预测”。SHAP (SHapley Additive exPlanations) [10]基于合作博弈论中的 Shapley 值,将每个特征对预测结果的贡献进行量化分配,能够同时给出全局和个体层面的解释。

LIME (Local Interpretable Model-agnostic Explanations)通过在样本局部构建线性代理模型来解释复杂模型的预测结果,实现简单且与模型无关,适合作为 SHAP 之外的补充解释手段。

在医疗领域,已有研究将 SHAP 等可解释方法用于心脏病、房颤等疾病的风险预测,验证了其在揭示模型决策依据方面的有效性。

3. 疾病预测系统总体架构设计

3.1. 系统需求分析

系统需求包括功能需求、数据需求与用户需求三部分。功能需求涵盖医学影像上传、临床指标输入、

疾病预测、风险评估、结果展示与模型解释六类操作；数据需求涉及医学影像数据、临床结构化数据、模型预测结果以及用户反馈；用户群体主要包括医疗研究人员、医疗辅助人员和教学科研人员。

3.2. 系统总体架构

基于上述需求，本文将系统划分为四层结构：用户交互层、应用服务层、AI 模型层与数据处理层，自上而下形成清晰的调用关系。系统总体架构如图 1 所示。



Figure 1. Overall architecture of the medical big data-driven disease prediction system
图 1. 医疗大数据驱动的疾病预测系统总体架构

用户交互层负责与用户直接交互，提供图像上传、数据输入与结果展示三类界面；应用服务层对业务进行封装，包含肺炎识别、冠心病预测与结果解释三个服务模块；AI 模型层集中存放模型资产，包括用于影像识别的 ResNet50、用于临床预测的机器学习模型以及 SHAP 模型解释方法；数据处理层则承担数据读取、清洗、预处理与特征提取的职责，为上层提供可用的数据输入。各层之间通过明确定义的接口传递数据，便于后续扩展与替换。

3.3. 系统业务流程

系统的业务流程覆盖用户从登录到获取预测结果的全过程。其业务流程如图 2 所示。



Figure 2. Business process of the disease prediction system
图 2. 疾病预测系统业务流程

用户首先选择预测任务类型，随后上传医学影像或输入临床指标；系统调用数据处理层对原始数据进行预处理与特征提取，再交由 AI 模型层完成预测；预测结果与解释信息通过应用服务层返回用户交互

层进行展示。整个流程在结构上保持一致，仅在数据输入与模型调用环节存在差异，便于用户理解与系统维护。

3.4. 系统功能模块设计

系统按业务职责划分为六个功能模块，详见表 1。

Table 1. Overview of system functional modules

表 1. 系统功能模块一览

模块名称	主要功能	所属层次
影像上传模块	接收用户上传的胸部 X 光片，并做尺寸校验	用户交互层
数据输入模块	接收结构化临床指标，支持单条与批量两种方式	用户交互层
肺炎识别模块	对上传影像进行预处理并调用 ResNet50 给出肺炎概率	应用服务层
冠心病预测模块	对临床指标进行预处理并调用机器学习模型给出风险概率	应用服务层
结果解释模块	基于 SHAP 方法对模型预测进行特征级解释	应用服务层
结果展示模块	将预测结果与解释信息以图表形式呈现给用户	用户交互层

4. 疾病预测系统关键技术实现

4.1. 肺炎影像预测模型

肺炎影像预测模块以胸部 X 光片为输入，目标是判断图像中存在肺炎的可能性。整体处理流程如图 3 所示。

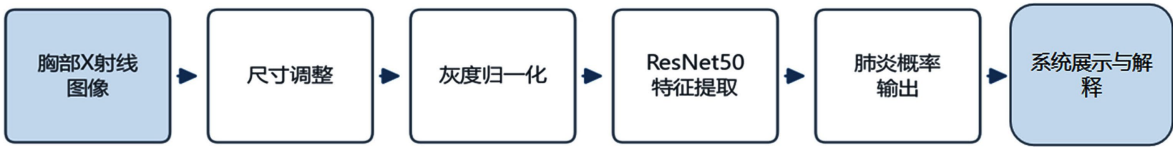


Figure 3. Pneumonia image prediction processing workflow

图 3. 肺炎影像预测处理流程

为缓解深层网络的训练困难，本文在特征提取阶段使用 ResNet50 作为骨干网络。残差块通过公式(1)学习残差映射：

$$y = F(x, \{W_i\}) + x \tag{1}$$

式中： x 为残差块的输入， $F(x, \{W_i\})$ 为需要学习的残差映射， $\{W_i\}$ 为对应卷积层的权重集合。当 $F(x, \{W_i\})$ 趋近于 0 时，残差块退化为恒等映射，从而避免梯度消失。

分类输出层使用 Softmax 函数将特征向量映射为类别概率分布，如式(2)所示：

$$P(y = j | x) = \frac{\exp(z_j)}{\sum_{k=1}^K \exp(z_k)} \tag{2}$$

式中： z_j 为第 j 个类别的输出得分， K 为类别总数， $K = 2$ 时对应“正常”与“肺炎”两类。

针对医学影像标注数据有限的问题，本文采用迁移学习策略，使用 ImageNet 预训练的 ResNet50 权重初始化模型，并在公开胸部 X 光片数据集上进行微调。已有研究表明，该策略能够有效利用预训练模

型的特征表征能力，在有限标注数据条件下获得较好的识别效果。

4.2. 临床风险预测方法

临床风险预测模块以就诊时采集的若干临床指标如年龄、血压、胆固醇、心率等为输入，输出疾病发生的风险概率。整体处理流程如图 4 所示。



Figure 4. Clinical risk prediction processing workflow

图 4. 临床风险预测处理流程

本文在基线模型中选择逻辑回归作为参考模型，其形式简洁且可解释性较强。逻辑回归的预测函数如式(3)所示：

$$P(z) = \frac{1}{1 + e^{-z}} \quad (3)$$

$$z = \beta_0 + \sum_{i=1}^n \beta_i x_i$$

式中： β_0 为截距项， β_i 为特征 x_i 对应的回归系数， n 为参与建模的特征数。当 $z > 0$ 时预测概率大于 0.5，模型倾向于输出阳性类别。

4.3. 模型服务化与系统实现

本文采用 Python + FastAPI 实现模型的服务化部署，将训练好的模型权重以文件形式加载至服务端，通过 REST 接口对外提供预测能力。系统启动时一次性完成模型加载，单次推理仅需调用已加载的权重，从而获得较稳定的接口响应时延。同时，服务端在并发请求到来时通过线程池调度预测函数，避免重复加载带来的额外开销。

在结果展示层面，系统除了给出预测概率，还引入 SHAP 方法给出特征级解释，帮助使用者理解模型关注了哪些因素。SHAP 值的定义如式(4)所示：

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f_x(S \cup \{i\}) - f_x(S)] \quad (4)$$

式中： ϕ_i 为特征 i 的 SHAP 值， N 为全体特征集合， S 为不含特征 i 的子集， f_x 为模型预测函数。该式将特征贡献转化为 Shapley 值形式，能够公平地反映每个特征对预测结果的影响。实际使用中，系统会同时输出全局特征重要性图和单条预测的特征贡献图，使用户既能把握模型的整体行为，也能针对具体病例追溯决策依据。

5. 系统应用与实验分析

5.1. 系统运行环境

本文系统部署在常规的 x86 服务器或工作站上，操作系统采用 Ubuntu 22.04 LTS，深度学习框架使用 PyTorch。系统运行所需的软硬件环境见表 2。

Table 2. System operating environment
表 2. 系统运行环境

类别	配置/版本
处理器	Intel Xeon Gold 6248R 或同档 CPU
GPU	NVIDIA RTX 3090 (可选, 用于模型训练)
内存	32 GB
硬盘	1 TB SSD
操作系统	Ubuntu 22.04 LTS
深度学习框架	PyTorch 2.x
服务端框架	Python 3.10 + FastAPI
前端框架	Vue 3 + Element Plus

5.2. 肺炎影像预测应用

肺炎影像预测应用以公开胸部 X 光片数据集为测试对象，测试流程为：上传 X 射线图像→系统处理→模型预测→返回结果。系统的肺炎识别界面如图 5 所示。

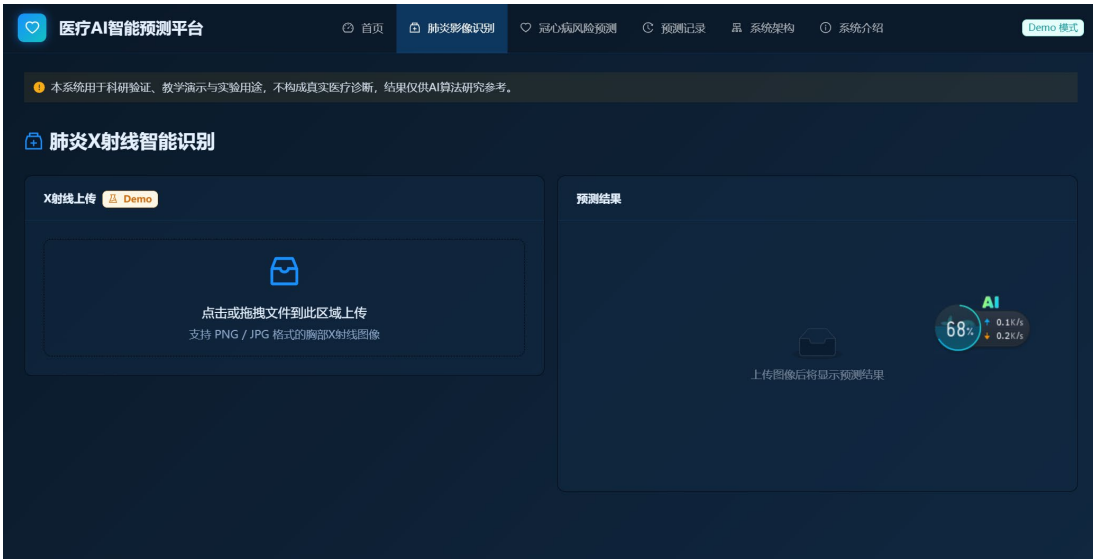


Figure 5. Pneumonia image recognition system interface
图 5. 肺炎影像识别系统界面

5.3. 冠心病风险预测应用

冠心病风险预测应用的输入包括年龄、血压、胆固醇、心率及其他相关临床指标，输出为风险概率。其界面如图 6 所示。

5.4. 模型解释应用

针对最终选定的临床风险预测模型，采用 SHAP 方法开展全局特征重要性分析和个体样本预测解释。模型解释应用对单条预测给出特征贡献图，对全体样本给出全局特征重要性排序。结果既可单独查看，也可与预测概率一起呈现。用户在使用过程中可以根据特征贡献判断预测是否与已有的临床经验一致，

识别潜在的异常病例。

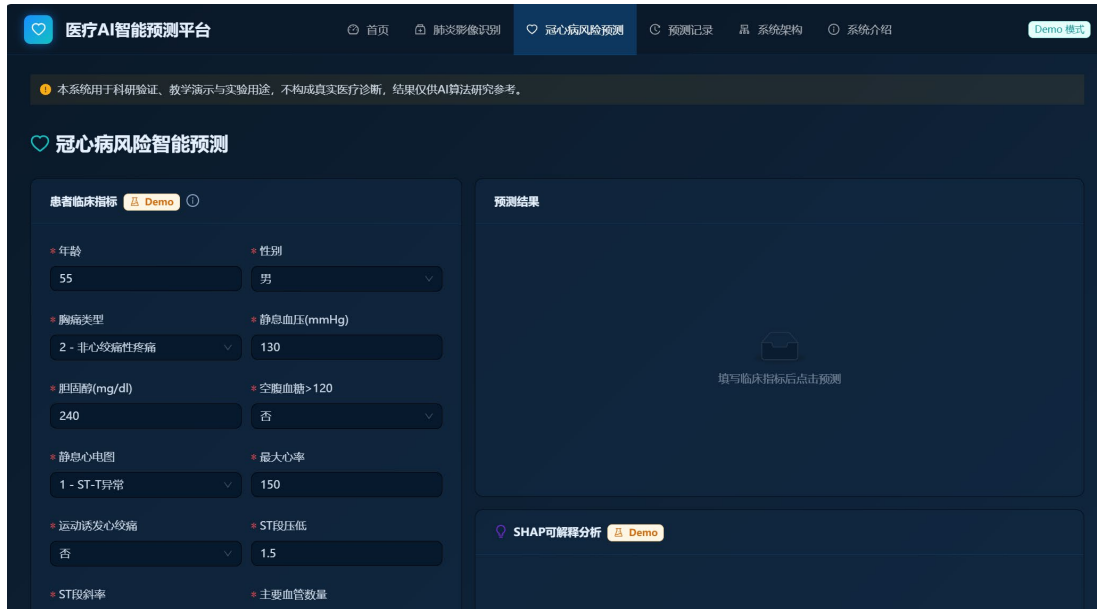


Figure 6. Coronary heart disease risk prediction interface
图 6. 冠心病风险预测界面

5.5. 模型预测性能分析

系统性能分析主要从预测性能方面展开。预测性能采用医学领域常用的四项指标：准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1 值(F1-score)，其定义分别如式(5)~式(8)所示：

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FN} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (8)$$

式中：TP 为真阳性、TN 为真阴性、FP 为假阳性、FN 为假阴性。F1 值是精确率和召回率的调和均值，能够在数据类别不均衡时对模型给出较为综合的评价。

肺炎识别和冠心病预测两类任务上的模型性能对比结果见表 3 和表 4。

Table 3. Pneumonia image recognition model performance
表 3. 肺炎影像识别模型性能

任务	模型	准确率	精确率	召回率	F1 值
肺炎识别	ResNet50 (迁移学习)	0.8910	0.8833	0.9513	0.9160
MobileNetV2	0.8157	0.7996	0.9410	0.8645	0.9242
DenseNet121	0.8542	0.8188	0.9846	0.8941	0.9567

Table 4. Comparison of performance of coronary heart disease risk prediction models
表 4. 冠心病风险预测模型性能对比

任务	模型	准确率	精确率	召回率	F1 值
冠心病预测	逻辑回归	0.8891	0.8871	0.8571	0.8721
冠心病预测	随机森林	0.8242	0.7826	0.8571	0.8182
冠心病预测	XGBoost	0.8242	0.8250	0.7857	0.8049

6. 结论

本文围绕医疗大数据驱动的疾病预测系统开展研究,构建了由数据处理层、AI 模型层、应用服务层和用户交互层组成的四层系统架构,并以肺炎影像识别和冠心病风险预测为典型应用场景,实现了相关功能模块。影像识别模块采用 ResNet50 作为骨干网络,并结合迁移学习策略缓解医学影像标注数据不足的问题;临床风险预测模块以逻辑回归为基线,引入随机森林和 XGBoost 等机器学习模型,并结合 SHAP 方法增强预测结果的可解释性。在此基础上,系统实现了医学影像上传、临床指标输入、疾病预测、风险评估、结果展示及模型解释等主要功能,为医疗数据驱动的辅助诊疗提供了初步的系统实现方案。

尽管本文完成了系统架构设计与主要功能模块的实现,但仍存在进一步完善的空间。首先,医疗数据具有较高的隐私敏感性,后续需要进一步完善数据脱敏、访问控制和隐私保护机制,以提高系统的数据安全性。其次,随着基础模型在医疗人工智能领域的发展,如何将大模型的通用能力与本文面向具体疾病任务的专科模型相结合,实现模型协同与能力扩展,是后续值得研究的问题。最后,当前系统主要基于公开数据集开展实验,尚缺乏大规模真实医疗数据和独立外部数据集的验证,模型的临床适用性和泛化能力仍需通过更严格的实验进一步评估。

参考文献

- [1] 杜志成, 张志杰, 姜庆五. 医疗大数据分析技术研究进展[J]. 中国血吸虫病防治杂志, 2022, 34(5): 465-468+492.
- [2] 郭子菁, 罗玉川, 蔡志平, 等. 医疗健康大数据隐私保护综述[J]. 计算机科学与探索, 2021, 15(3): 389-402.
- [3] Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., *et al.* (2021) Deep Learning-Enabled Medical Computer Vision. *npj Digital Medicine*, 4, Article No. 5. <https://doi.org/10.1038/s41746-020-00376-2>
- [4] 王霄, 朱恩照, 艾自胜. 卷积神经网络的原理及其在医学影像诊断中的应用[J]. 中国医学物理学杂志, 2022, 39(12): 1485-1489.
- [5] 辛瑞昊, 董哲原, 苗冯博, 等. 基于机器学习的心脏病预测模型研究[J]. 吉林化工学院学报, 2022, 39(9): 27-32.
- [6] D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., *et al.* (2008) General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation*, 117, 743-753. <https://doi.org/10.1161/circulationaha.107.699579>
- [7] Breiman, L. (2001) Random Forests. *Machine Learning*, 45, 5-32. <https://doi.org/10.1023/a:1010933404324>
- [8] Rajpurkar, P., Irvin, J., Zhu, K., *et al.* (2026) CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225.
- [9] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., *et al.* (2019) CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 590-597. <https://doi.org/10.1609/aaai.v33i01.3301590>
- [10] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. <https://doi.org/10.1145/2939672.2939785>