文章编号：2095-2163(2023)11-0172-08 中图分类号：TP391 文献标志码：A

融合随机森林与 SHAP 的心脏病预测及其特征分析研究

程祉元，张博良，蔡雨晨，马雨生，邵泽国，刘巧红
（上海健康医学院 医疗器械学院，上海 201318）

摘 要：心脏病是一种常见的心血管疾病，对人类生命健康有极大的威胁，准确预测是否患有心脏病能够帮助心脏病的早发现、早治疗，提升心脏病患者的生活质量和寿命。本文以克利夫兰心脏病数据集为研究对象，首先对原始数据集进行数据变换、标准化处理等工作，将处理后的数据作为随机森林模型的输入进行训练，将预测结果与线性逻辑回归、K-最近邻、决策树等多种机器学习模型进行比较，结果表明本文模型在准确率、查准率、查全率、*F*1 值、*AUC* 值等 5 种性能评价指标上均优于对比的模型。最后，引入了 SHAP 模型加强预测模型的可解释性，并进行特征分析识别出影响心脏病的主要因素，为临床决策提供可参考的依据。

关键词：心脏病；随机森林；预测模型；SHAP；特征分析

Combination of Random Forest and SHAP for heart disease prediction and feature analysis research

CHENG Zhiyuan, ZHANG Boliang, CAI Yuchen, MA Yusheng, SHAO Zeguo, LIU Qiaohong
(School of Medical Instrumentation, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China)

Abstract: Heart disease is a common cardiovascular disease, which further poses threats to human health. Accurate prediction of heart disease can foster the early detection and treatment of heart disease, and furthermore improve the life quality and longevity of patients with heart disease. This study is based on the Cleveland heart disease dataset. In the research, on the basis of data transformation and normalization of the raw data set, the processed data are trained as the input of random forest model. The prediction results are compared with LR, KNN, decision tree and other machine learning models. The results show that the model is superior to the comparison model in five performance evaluation indexes, such as accuracy, precision, recall, *F*1 - score and *AUC*. Therefore, the SHAP model is introduced to enhance the interpretability of prediction model, and the main factors affecting heart disease are identified by feature analysis, providing a reference basis for clinical decision making.

Key words: heart disease; Random Forest; prediction model; SHAP; feature analysis

0 引 言

人的循环系统包括心脏、血管以及调节血液循环的神经体液组织，而循环系统疾病（心血管病）包括了上述所有组织器官的疾病，而心脏病在其中最为多见，也常见于内科疾病，会导致患者的劳动力严重丧失。随着生活水平的提高，人们对自己的生活质量，尤其是身体健康有着更高的要求。然而，根据《中国心血管健康与疾病报告 2020》，心血管疾病约有 3.3 亿人，包括 1 300 万脑卒中，1 139 万冠心病，500 万肺源性心脏病，4 530 万下肢动脉疾病以及 2.45亿高血压[1]。心血管病给社会带来的经济负担日益加重，已成为重大的公共卫生问题。

研究可知，心脏病因其多样复杂的发病类型、极高的死亡率，成为了医学上多年来想要攻克的难题[2]。现阶段心脏疾病的诊断更多依赖于医生对各类检查生成的医学影像的阅片以及患者的生活环境、家族病史、生理指标等因素的综合诊断。最终的

基金项目：国家自然科学基金（61801288）；上海市科委科技创新行动计划项目（22DZ2305300）；国家社会科学基金（20BTQ073）。

作者简介：程祉元（2001-），女，本科生，主要研究方向：医学大数据分析；张博良（2002-），男，本科生，主要研究方向：医学大数据分析；蔡雨晨（2003-），男，本科生，主要研究方向：医学大数据分析；马雨生（2001-），男，本科生，主要研究方向：医学大数据分析；邵泽国（1978-），男，博士，副教授，主要研究方向：人工智能与计算医疗。

通讯作者：刘巧红（1979-），女，博士，副教授，硕士生导师，主要研究方向：医学图像处理。Email：hqllqh@163.com

收稿日期：2022-11-09

诊断结果易受到医生经验和诊断方式等主观因素影响，不同医生的诊断结果常常不一致，甚至出现误诊和漏诊等现象[3]。近年来，随着人工智能在医疗领域逐步深入的应用，人们发现利用机器学习算法针对医疗健康数据建立模型，辅助医生对于疾病的诊断，增强评估的客观性，可以大大提高诊断准确率。同时，还可降低医生由于自身临床经验不足及疲劳工作而导致的误判风险，提高诊断效率，以及解决现阶段普遍存在的医疗诊断滞后性的问题，做到早发现、早干预。例如，林志远[2]采用了决策树算法构建了心脏病预测模型，分析了 ID3 和 CART 的区别。李岭海[4]对比 SIFT、SURF、KAZE，发现深度学习可以提高分类超声心电图的准确率，对心脏病的分类效果更好。石胜源等学者[5]的实验结果表明，随机森林算法在心血管疾病预测中准确率为 73.55%，具有较大的优势，并且性能优于其他算法，对心血管疾病的预测研究和早期病人的及时有效治疗具有重要意义。陈洞天等学者[6]利用 Xgboost 模型预测心脏病，准确率为 76.5%，且利用了指标分析法对预测模型的进行特征分析。Krithiga 等学者[7]利用贝叶斯分类器应用于冠心病的早期预测，取得了不错的效果。王健等学者[8]提出了一种基于特征组合和卷积神经网络的方法预测心脏病，准确率为 89.9%，但缺少该预测方法的可解释性，即不能说明该算法的内部预测过程及其是否与临床诊断方法吻合。

本文基于集成学习随机森林算法，以克利夫兰心脏病数据集作为研究对象，在对其进行数据预处理、模型训练、超参数优化、模型性能分析、可解释性等工作的基础上，建立了性能优越的预测模型。本文的主要工作体现在以下 2 个方面：

(1)提出使用随机森林模型预测心脏病，并通过网格搜索技术进行参数优化提高模型性能，采用准确率、查准率、查全率、$F1$ 值、AUC 值等 5 种指标评价预测效果，混淆矩阵、AUC 可视化分析预测效果，与线性逻辑回归、K - 最近邻、决策树等模型对比，验证了本文模型性能的优越性。

(2)在保证随机森林模型预测性能的基础上，引入 SHAP 可解释性模型来增强随机森林模型的可解释性，对影响心脏病的关键因素进行了特征分析，为心脏病的临床诊断和决策提供了可参考的依据。

1　方法及原理

1.1　随机森林算法

随机森林算法的本质是利用集成理论将多个弱分类器(决策树)通过训练之后生成多棵独立分布的决策树并将决策树集成一体，形成强分类器(随机森林)。算法有效地解决了单棵决策树存在的不稳定性、无法保证全局最优及过度拟合等问题。这是 Bootstrap 与决策树算法的结合，方法是先从原始数据集 D 中采用 Bootstrap 重采样技术，采用放回式取样抽取一定数量的训练样本集，生成对应数量的决策树；决策树训练过程中，每个节点的特征都是从该决策树数据集特征中按照特定比例地无放回随机抽取新的特征子集[9]；最后，从新特征子集中选出能使信息增益率最大化的特征，并以其为分割点。信息增益公式如下：

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{v} \frac{|D^v|}{|D|} Ent(D^v) \quad (1)$$

其中，$Gain(\,)$ 表示信息增益；$Ent(\,)$ 表示信息熵；D 表示原始数据集；a 表示新特征子集中某个特征；v 表示使用特征 a 有 v 个可能的分支节点。最终分类结果，由所有独立决策树的结果投票决定，公式如下：

$$H(x) = arg \max_{Y} \sum_{i=1}^{k} I\,(h_i(x) = Y) \quad (2)$$

其中，$H(x)$ 表示对样本 x 的包外预测；k 表示弱分类器的迭代次数；$h(\,)$ 表示基学习器；Y 表示某个样本特征的标签；I 表示示性函数。这种方式保证了输入每棵决策树的训练集的随机性以及每个划分节点的随机性。优势在于其能够处理高维度数据集，实现比较简单，训练速度快，还可以将不平衡数据集的误差缩小，并对于存在大量缺失值的数据样本也能较好地处理。

1.2　SHAP 模型解释

随机森林预测模型虽然可以得到较高的准确率，但其“黑盒”性质决定了对结果的解释力很弱，例如很难解释为什么算法可以准确预测患者是否罹患特定的疾病。

SHAP (SHapley Additive exPlanation)能够观察到某一个样本的预测中各个特征对预测结果产生的影响，对随机森林模型的单个预测做出解释。SHAP 模型的原理是给每个单独的预测样本都生成一个预测值，而单个样本中对应其特征分配的数值表现为 SHAP value。假设第 i 个样本的第 j 个特征为 x_{ij}，模型对该样本的预测值为 y_i，模型的基线(默认所有样本目标变量的均值为基线)为 y_{base}，那么 SHAP value 服从以下公式：

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{im}) \quad (3)$$

其中，$f(x_{ij})$ 表示第 i 个样本的第 j 个特征对样本预测值 y_i 的贡献度。当 $f(x_{ij}) > 0$，表示该特征使得预测值升高，有积极的影响；反之，则说明该特征使得预测值降低，有消极的影响[6]。SHAP value 的优势在于 SHAP 能反映出每一个样本中各特征的影响力以及影响力的正负性，并且特征本身在模型内部还有交互作用。本文利用 SHAP 来解释随机森林算法内部是如何预测结果的。

2 分类模型构建

2.1 模型构建

心脏病分类预测模型的设计思路主要包含数据探索，对数据集的统计分布进行可视化展示，观察数据的分布情况；特征工程，完成数据预处理，如数据变换、数据标准化等，保证数据的质量；模型构建，构建随机森林的心脏病预测模型；超参数优化，采用网格搜索技术对随机森林算法的超参数进行优化调参，提高模型的预测能力；模型训练，利用十折交叉验证将数据集随机地划分为训练集和测试集进行验证，提高模型的泛化能力；可解释性分析，采用 SHAP 对模型中的心脏病的影响因素进行解释分析，增强模型的可解释性。整个基于随机森林的心脏病风险预测及特征分析模型的构建流程如图 1 所示。

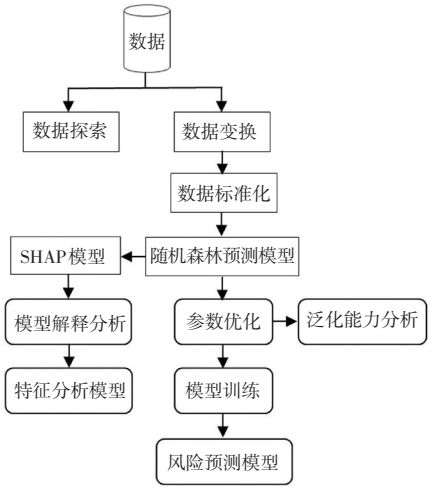


图 1 心脏病风险预测及特征分析模型流程图

Fig. 1 Flow chart of heart disease risk prediction and characteristic analysis model

2.2 数据探索

本研究采用 kaggle 平台提供的数据集，其来源于 University of California, Irvine (UCI) 机器学习数据库中的 the Cleveland database 数据集，此数据库包含 76 个属性，但所有已发布的实验都引用并使用其中 14 个属性的子集，即克利夫兰心脏病数据集。

该数据集中一共有 303 个样本，每个样本有 14 个特征，其中 13 个特征为自变量，描述样本的基本患病信息，最后 1 个特征“Target”为因变量，表示患者是否患有心脏病，所有的特征及其含义见表 1。

表 1 克利夫兰心脏病数据集的基本特征

Tab. 1 Basic characteristics of the Cleveland heart disease data set

编号	特征属性	特征属性含义	特征类型
1	*age*	年龄，1~100	数值型
2	*sex*	性别，1=male，0=female	分类型
3	*cp*	胸痛类型，0=典型心绞痛，1=非典型心绞痛，2=非心绞痛，3=没有症状	分类型
4	*trestbps*	静息血压，90~200	数值型
5	*chol*	血清胆固醇，100~600	数值型
6	*restecg*	静息心电图，0=正常，1=患有 ST-T 波异常，2=根据 Estes 的标准显示可能或确定的左心室大	分类型
7	*fbs*	空腹血糖，>120mg/dl；1=true，0=false	分类型
8	*thalach*	达到的最大心率，90~200	数值型
9	*exang*	运动诱发的心绞痛，1=yes，0=no	分类型
10	*oldpeak*	相对于休息的运动引起的 ST 值，0~5	数值型
11	*slope*	运动高峰 ST 段的坡度，1= up sloping、向上倾斜，2=flat、持平，3= down sloping、向下倾斜	分类型
12	*ca*	大血管数量，0~3	数值型
13	*thal*	地中海贫血，1=正常，2=固定缺陷，3=可逆转缺陷	分类型
14	*target*	生病有无，1=yes，0=no	分类型

通过对数据质量的探索和数据特征的分析，观察数据样本和特征的数量、数据类型及数据概率分布等信息，用于指导预测模型建立。根据对心脏病原始数据的描述性统计分析发现，未患病人群中男性所占比例远超女性，而患病人群中男性占比仍多于女性。将年龄对患病情况的影响绘制出的柱状统计分布如图 2 所示。由图 2 可知，中年患病几率较大。

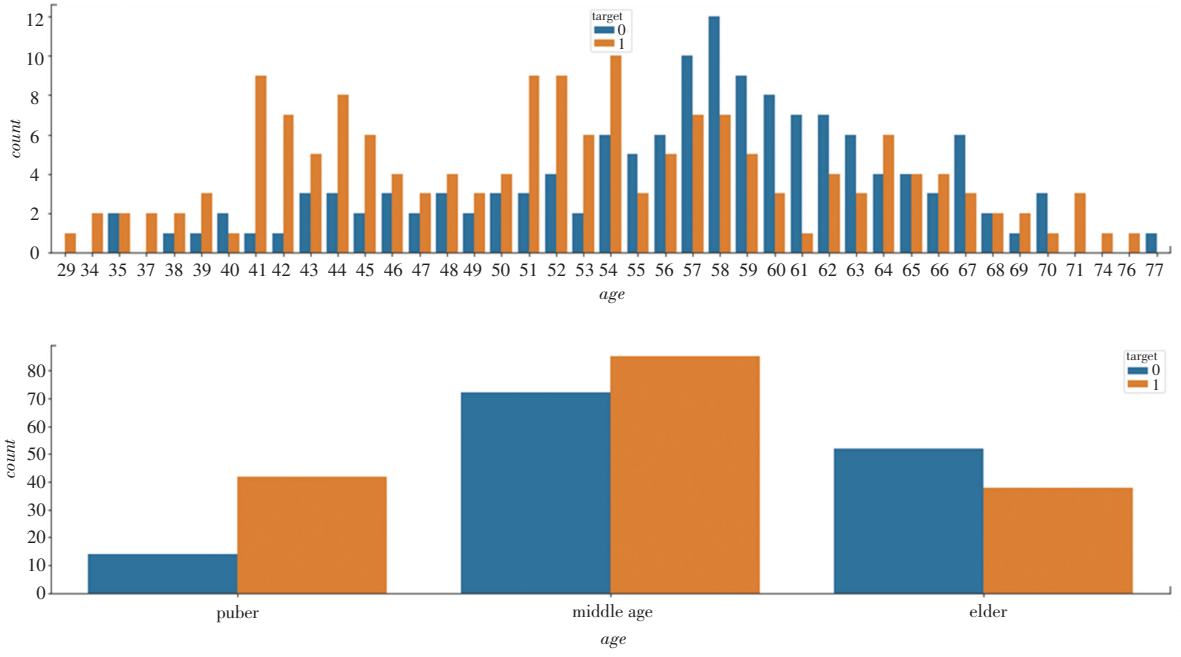
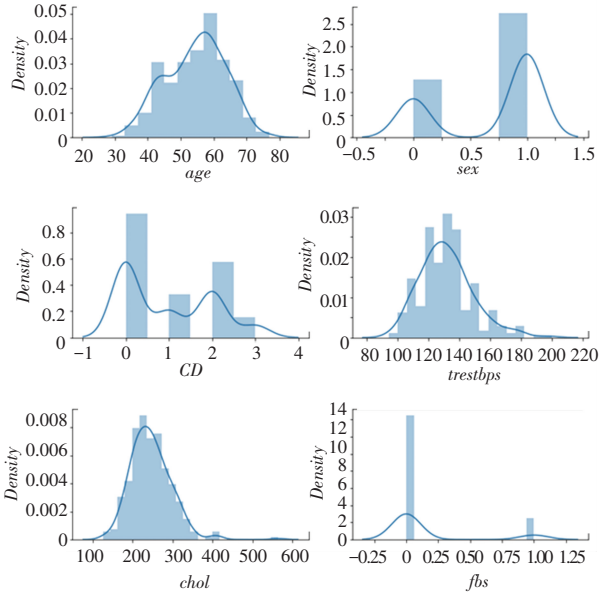


图 2　根据年龄分析患病情况

Fig. 2　Analysis of prevalence by age

图 3 是心脏病数据集中 14 个特征的单变量分布密度图，从图 3 中可以看出每个特征的数据类型及取值分布，其中 *age*、*trestbps*、*chol*、*thalach* 和 *oldpeak* 五个特征为连续型特征，*sex*、*cp*、*fbs*、*restecg*、*exang*、*slope*、*ca*、*thal* 和 *target* 九个特征为非连续型特征，需要进行数据预处理操作。



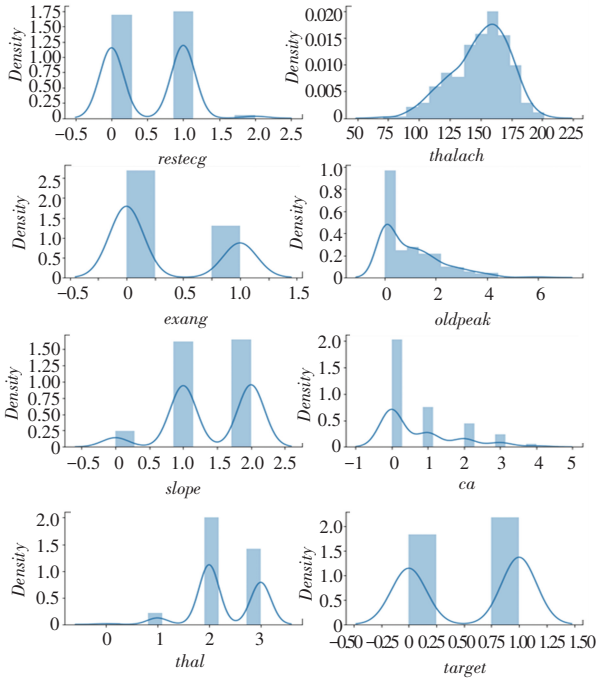


图 3　单变量特征统计分布

Fig. 3　Statistical distribution of univariate characteristics

2.3　特征工程

2.3.1　特征相关性

图 4 给出了能够反映特征之间关系的热力图，通过热力图来发掘特征之间的关系。热力图表示了

2 个数据之间的相关性,数值范围是-1 到 1 之间,大于 0 表示 2 个数据之间是正相关的,小于 0 表示 2 个数据之间是负相关的,等于 0 就是不相关。由图 4 可知,*cp*、*thalach* 和 *slope* 这 3 个特征与 *target* 之间正相关且系数大,表明其与是否患病的关系较为密切。

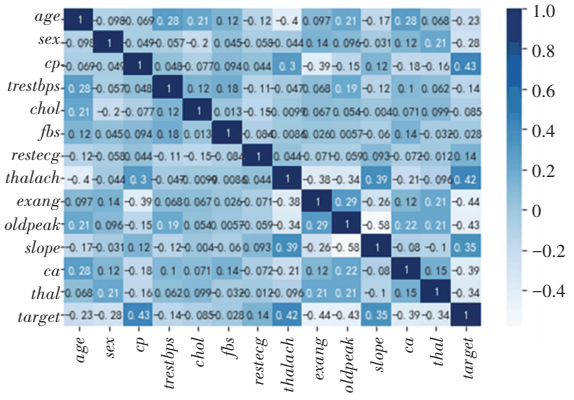


图 4 各项特征之间的相关性热力图

Fig. 4 Thermodynamic diagram of correlation between features

2.3.2 非连续型数值转换

经过数据探索和特征相关性分析发现,*cp*、*thal* 和 *slope* 为不连续的多分类特征,该类型的数据不适合作为分类器输入,因此,首先将*cp*、*thal*和*slop*三个特性转换成独热编码的形式参与模型训练。原始特征 *cp* 转换为 4 个代表不同取值的特征 *cp*_0、*cp*_1、*cp*_2 和 *cp*_3,原始特征 *thal* 转换为 4 个代表不同取值的特征 *thal*_0、*thal*_1、*thal*_2 和 *thal*_3,原始特征 *slope* 转换为 3 个代表不同取值的特征 *slope*_0、*slope*_1 和 *slope*_2,并将原始特征删除。经过数据转换处理后的特征维度由原始数据的 14 增加到了 22。

2.3.3 数据归一化

为了消除数据之间的量纲影响,减小数据集中数据的差异性,对数据进行了归一化处理,将数据统一归一化到[-1,1]之间。原始数据经过数据标准化处理后,处于同一数量级,能够有效地提升模型精度和收敛速度。

2.4 参数优化

随机森林模型涉及到多个参数选择,参数值的选择影响到模型的性能。具体的参数取值见表 2。对于表 2 中的 6 个核心参数,本文采用了网格搜索技术进行调参。网格搜索在规定的参数取值范围内逐步调整参数,用调整后的参数对随机森林模型进行训练,使得模型性能最优的参数确定为最佳参数。

表 2 随机森林算法参数意义及取值

Tab. 2 Meaning and value of random forest algorithm parameters

编号	参数	参数意义	参数取值范围	参数取值
1	*max_features*	构建决策树最优模型的最大特征数	1,3,5	3
2	*max_leaf_nodes*	最大叶子节点数	/	16
3	*n_estimators*	对原始数据集进行有放回抽样生成的子数据集个数,即决策树的个数	400,420,440	420
4	*n_jobs*	设定工作的 core 数量-1 表示 cpu 里的所有 core 进行工作	/	-1
5	*oob_score*	是否采用袋外样本来评估模型的好坏	/	True
6	*random_state*	随机模式的设置,指定随机数生成器的种子	/	666

3 实验分析

3.1 模型性能度量

为了客观评价该算法的有效性,采用了 *F*1 值、准确率、查准率、查全率和 *AUC* 值这 5 种评价指标对模型性能进行度量。

(1)准确率(*Accuracy*)。表示所有样本中被预测正确的样本的比率。可由如下公式计算求值:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

(2)查准率(*Precision*)。表示预测样本中预测为真阳性的概率。可由如下公式计算求值:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

(3)查全率(*Recall*),真阳性率(True Positive Rate,*TPR*),灵敏度(*Sensitivity*)。表示阳性样本被预测为真阳性的概率。可由如下公式计算求值:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

(4) *F*1 值(*F*1 - score)。用来衡量二分类模型

精确度的一种指标，可以看作是模型查准率和查全率的一种加权平均。该指标同时兼顾了分类模型的查准率和查全率，最大值是 1，最小值是 0。可由如下公式计算求值：

$$F1-score=\frac{2\times TP}{2\times TP+FN+FP}=2\times\frac{precision\times recall}{precision+recall} \tag{7}$$

其中，真阳性（True Positive，*TP*）表示样本中正确识别的数量；假阳性（False Positive，*FP*）表示样本中错误识别的数量；真阴性（True Negative，*TN*）表示正确识别为错误的样本数；假阴性（False Negative，*FN*）表示错误识别为正确的样本数。除了上述指标之外，还使用了 *ROC* 曲线和 *AUC* 值。

3.2　模型性能评估

3.2.1　模型对比

为验证本文的随机森林模型的有效性，与逻辑回归、K-最近邻、决策树等常用模型进行比较分析。为了提高模型之间对比的公平性及可靠性，实验中采用了十折交叉验证方法进行性能评估。各种模型在准确率、查准率、查全率、*F*1 值和 *AUC* 值这 5 项指标上的对比结果见表 3，各种模型的 *ROC* 曲线对比如图 5 所示。从表 3 和图 5 的实验结果可以看出，本文的集成学习模型随机森林的预测准确率为 86%，查准率为 85%，查全率为 83%，*F*1 值为 84%，*AUC* 值为 0.89，均高于其它对比的方法。随机森林模型的 *ROC* 曲线（红色）下方面积比逻辑回归模型、K-最近邻模型、决策树模型的面积大，由 *ROC* 曲线的性质可知，曲线下方面积（*AUC*）越大、准确率越高，体现了本文模型的优越性。

表 3　不同分类模型对阳性样本的预测能力

Tab. 3　The predictive ability of different classification models for positive samples

模型	准确率	查准率	查全率	*F*1 值	*AUC* 值
逻辑回归	0.80	0.80	0.80	0.80	0.89
K-最近邻	0.82	0.83	0.83	0.83	0.86
决策树	0.72	0.73	0.73	0.72	0.79
随机森林	**0.86**	**0.85**	**0.83**	**0.84**	**0.89**

各种模型的训练时间和测试时间的对比见表 4。随机森林模型作为一种集成学习算法，模型复杂度本身高于其它几种对比的方法，同时采用网格搜索技术的参数优化较为耗时，因此在训练时间上相对较长。图 6 还给出了本文模型的混淆矩阵，可以看出预测结果中，测试集中非心脏病被预测为非心脏病有 27 例，心脏病被预测为心脏病有 36 例，非心脏病被预测为心脏病有 8 例，心脏病被预测为非心脏病有 5 例。显而易见的是，随机森林模型的真阳性和真阴性数量高，而假阳性和假阴性的值较低，因此，本文提出的模型有较好的分类性能。

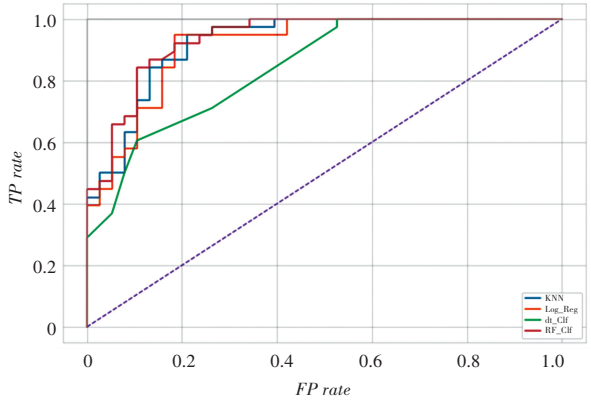


图 5　四种模型的 *ROC* 曲线

Fig. 5　*ROC* curves for the four models

表 4　各模型时间性能比较

Tab. 4　Comparison of time performance of each model

模型	Train 平均用时/s	Test 平均用时/s
逻辑回归	3.26	0.00
K-最近邻	2.91	0.02
决策树	2.69	0.00
随机森林	9.89	0.03

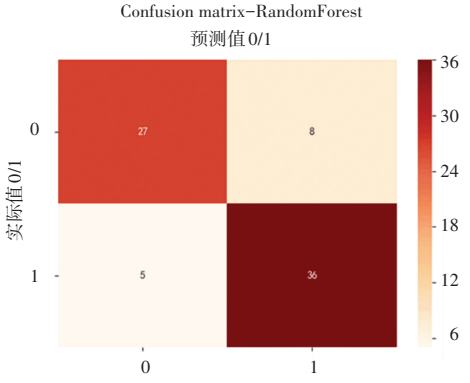


图 6　随机森林的混淆矩阵

Fig. 6　Confusion matrix of random forests

3.2.2　相关研究对比

为了进一步验证本文模型的优越性，与文献[8]、文献[10]、文献[11]和文献[12]等相关工作进行了对比实验。所有文献都针对克利夫兰心脏病数据集进行研究，文献[8]首先采用特征组合增强样本的属性关联，再利用卷积神经网络模型进行训练，在准确率上获得了高达 90%的预测精度。文献[10]与本文模型相似，但其样本量在克利夫兰心脏病数据集的基础上增加到 573 个，且在网络搜索优化参数上仅优化了 *n_estimators*、*max_depth*、*max* -

Leaf_nodes 三个参数。文献[11]使用未优化的随机森林模型训练获得了 85% 的准确度。文献[12]基于聚类和 XGBoost 算法进行预测分析，准确率达到 83%。

不同方法的准确率比较见表 5。从表 5 可以看出，本文模型的预测结果优于文献[10]、[11]和[12]，但略低于文献[8]。然而本文与其它文献的最大区别之处在于，本文在模型训练后，引入了 SHAP 可解释性模型，对模型进行可解释增强，识别出临床实际中影响心脏病的主要因素，为临床上的诊断和决策提供了有利的参考。

表 5 不同方法的准确率比较

Tab. 5 Comparison of accuracy of different methods

文献	方法	准确率
[8]	特征组合+卷积神经网络	0.90
[10]	随机森林+网格搜索	0.83
[11]	随机森林	0.85
[12]	聚类+XGboost	0.83
本文	随机森林+网格搜索+SHAP	0.86

3.3 基于 SHAP 的模型可解释性分析

图 7 是随机森林模型的特征重要性排序图。图 7 中，纵坐标是从上到下按照特征重要性排序的各个特征，横坐标是平均 *SHAP* 值。图 7 中显示特征重要性排序前六的特征分别是 *thal*_2(固定缺陷型地中海贫血症)、*cp*_0(典型心绞痛)、*ca*(大血管数量)、*thal*_3(可逆转缺陷型地中海贫血症)、*oldpeak*(运动高峰的心电图 *ST* 段)、*thalach*(最大心率)，可见这 6 个因素是影响是否患有心脏病的最关键因素。

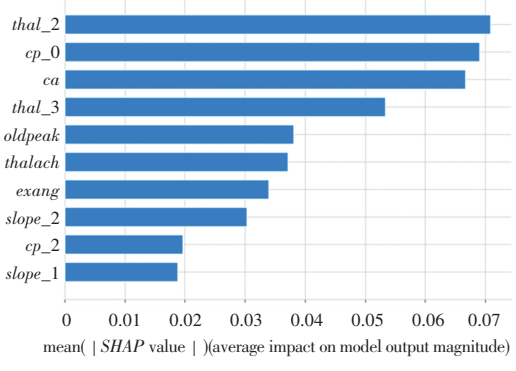


图 7 基于 *SHAP* value 的特征重要性排序

Fig. 7 Sorts by features importance based on *SHAP* value

图 8 显示了 SHAP 摘要图，该图对影响心脏病患病的因素重要性进行了排序。图 8 中的一个点表示一个样本，样本点的颜色从蓝色到红色表示样本特征值从小到大，纵坐标的各特征标签不仅显示了特征重要性排序，还显示了各个特征值与 *SHAP* 值的关系与分布。图 8 中绘制了重要性排序前 10 的特征对预测结果的影响，其中 *thal*_2(固定缺陷型地中海贫血症)、*thalach*(最大心率)对预测结果有正向贡献，*cp*_0(典型心绞痛)、*ca*(大血管数量)、*thal*_3(可逆转缺陷型地中海贫血症)、*oldpeak*(运动高峰的心电图 *ST* 段)对模型预测为心脏病的输出结果有负向贡献。

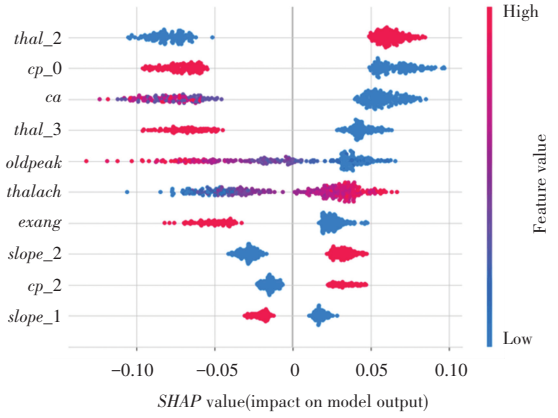


图 8 *SHAP* 特征分析

Fig. 8 *SHAP* feature analysis

4 讨论与分析

临床上，诊断心脏病的常规检查主要有常规心电图(ECG)与动态心电图(DCG)，心电图异常可提示心肌梗死、心肌缺血、心肌炎、心室肥厚等病症。相关研究对于各类心脏疾病的诊断有如下常见的标准：

(1)心电图 ST 段趋势的改变可以作为重要参考依据，指标过高可能是冠心病，指标过低则有可能是心肌缺血等病症，还用以诊断确定心室是否肥大[13-15]。

(2)心肌缺血在 ECG 的诊断标准为在同一导联上，*T* 波小于 *R* 波的十分之一，同时，*ST* 段水平下移 0.05 mV及以上；在 DCG 的诊断标准为与等电位线比较，*ST* 段下斜或压低 0.1 mV 及以上并持续下移大于1 min[16]。

(3)冠心病、肥厚型心肌病常伴有心绞痛等症状，分为典型心绞痛和非典型心绞痛，主要的病因为心肌缺血。

(4)荧光显色主要血管数目越少(数目与血糖、胆固醇相关)证明血液流动越通畅，血管腔狭窄会使患冠心病的风险大大增加[17-18]。临床常选择冠脉造影这种有创性检查，作为判断动脉狭窄程度的“金标准”。

（5）地中海贫血症是先天性贫血症影响红细胞的寿命，易导致红细胞数量不足，使得体内铁超载，从而加重心脏负担，长期的慢性贫血会诱发心绞痛，会造成心力衰竭[19-20]。

本文通过对原始数据集的预处理，构造了一个包括 22 个影响心脏病患病可能的特征，并将这些特征作为随机森林模型的输入，结合网格搜索技术的调优和十折交叉验证的模型训练，取得了高达 86% 的准确率。进一步利用 SHAP 模型对所有特征进行了事后解释分析，通过特征分析发现 *thal*（地中海贫血类型）、*ca*（主要血管数目）、*cp*（心绞痛）、*oldpeak*（心电图 *ST* 段趋势的改变）、*thalach*（最大心率）、*exang*（心绞痛型胸痛）等指标都是影响心脏病患病的重要因素。对于地中海贫血，综合观察 *thal*_2、*thal*_3，可以看出固定缺陷型地中海贫血与心脏病风险显著正相关，即会明显增加风险；而可逆转缺陷型对风险的增加不明显。对于心绞痛，综合观察 *cp*_0、*cp*_2 以及 *exang*，可以看出心绞痛、无论典型心绞痛还是非典型心绞痛，亦或是运动诱发的心绞痛对风险的增加不明显；而非心绞痛型的胸痛与心脏病风险呈正相关，会明显增加风险；究竟哪些非心绞痛型的胸痛明显增加心脏病风险还需进一步探讨。从 *ca* 指标可以观察到，大血管数量越少，心脏病风险系数越高；同样，*oldpeak* 值（即相对于休息的运动引起的 *ST* 值）越低，心脏病风险系数越高。从 *thalach* 指标可以很明显地看到最大心率值越大，心脏病风险系数越高。综合观察 *slope*_1、*slope*_2，可见运动高峰 *ST* 段的坡度持平与心脏病风险成正相关，*ST* 段的坡度向上倾斜与心脏病风险成负相关，这与心电图运动试验阳性诊断标准条件之一“运动中或运动后 *ST* 段程水平或下斜型压低 ≥ 0.10 mV”相吻合。

5　结束语

本文基于集成学习的随机森林算法构建了心脏病预测模型，同时引入了 SHAP 对预测模型做进一步增强解释。首先针对 Kaggle 平台提供的心脏病数据集进行数据变换、标准化等预处理后，采用网格搜索技术对模型的参数进行优化，并对处理后的数据集进行十折交叉验证训练模型；然后，采用查准率、查全率、*F*1 值、混淆矩阵、*AUC* 值等指标对模型进行评估，与逻辑回归、K-最近邻、决策树等机器学习模型的结果进行对比，验证了随机森林具有较强的泛化能力、更好的分类效果；最后，还引入 SHAP 模型对随机森林模型做进一步解释，识别出影响心脏病患病的主要因素，并解释这些特征与临床诊断的关系。模型增加了可解释说明，从而提高了模型的分类识别效率，为临床决策服务，具有重要的实用价值。

参考文献

[1] 国家心血管病中心.《中国心血管健康与疾病报告 2020》概述[J]. 中国心血管病研究，2021，19(7)：582-590.

[2] 林志远. 基于决策树算法的心脏病预测研究[J]. 电子制作，2019，370(6)：25-27.

[3] 赵梦蝶，孙九爱. 机器学习在心血管疾病诊断中的研究进展[J]. 北京生物医学工程，2020，39(2)：208-214.

[4] 李岭海. 基于深度学习的心脏病检测的研究[J]. 现代计算机（专业版），2017(9)：91-93，110.

[5] 石胜源，朱磊，叶琳，等. 基于随机森林算法的心血管疾病预测研究[J]. 智能计算机与应用，2021，11(4)：176-178，181.

[6] 陈洞天，单杰，周文丹. 基于 Xgboost 的心血管疾病预测模型和指标分析研究[J]. 现代医院，2021，21(6)：958-961.

[7] KRITHIGA B，SABARI P，JAYASRI I，et al. Early detection of Coronary Heart Disease by using Naive Bayes Algorithm[J]. Journal of Physics Conference Series，2021，1717(1)：012040.

[8] 王健，李孝虔. 一种基于特征组合和卷积神经网络的心脏病预测新方法[J]. 黑龙江大学自然科学学报，2019，36(1)：115-120.

[9] ASADI S，ROSHAN S E，KATTAN M W. Random forest swarm optimization-based for heart diseases diagnosis[J]. Journal of Biomedical Informatics，2021，115：103690.

[10] 赵金超，李仪，王冬，等. 基于优化的随机森林心脏病预测算法[J]. 青岛科技大学学报（自然科学版），2021，42(2)：112-118.

[11] 孙铁铮，于泽灏. 基于机器学习的心脏病例分类预测研究[J]. 电脑知识与技术，2021，17(26)：96-97+104.

[12] 刘宇，乔木. 基于聚类和 XGboost 算法的心脏病预测[J]. 计算机系统应用，2019，28(1)：228-232.

[13] VOGEL B，CLAESSEN B E，ARNOLD S V，et al. ST-segment elevation myocardial infarction[J]. Nature Reviews Disease Primers，2019，5(1)：39.

[14] 刘燕. 心电图检查结合临床特征在冠心病心绞痛诊断中的应用价值分析[J]. 中国实用医药，2021，16(2)：16-18.

[15] 臧传欣. 动态心电图对冠心病诊断价值的研究进展[J]. 中国医疗器械信息，2021，27(14)：27-28，122.

[16] 肖蕾，孙晓臣，罗溶. 动态心电图在冠心病心肌缺血与心律失常诊断中的价值分析[J]. 解放军医药杂志，2022，34(01)：61-64.

[17] 北京高血压防治协会，北京糖尿病防治协会，北京慢性病防治与健康教育研究会，等. 基层心血管病综合管理实践指南 2020[J]. 中国医学前沿杂志（电子版），2020，12(08)：1-73.

[18] 张博，潘晓芳，隋春兴，等. 运动负荷超声心动图诊断冠状动脉严重病变假阴性结果的影响因素分析[J]. 中国循环杂志，2021，36(08)：756-761.

[19] SALAMA K，ABDELSALAM A，ELDIN H S，et al. Iron overload parameters and early detection of cardiac disease among Egyptian children and young adults with β-thalassaemia major and sickle cell disease：a cross-sectional study[J]. F1000Research，2020，9：1108.

[20] PEPE A，PISTOIA L，GIUNTA N，et al. The strong link between pancreas and heart in thalassemia major[J]. European Heart Journal，2018，39(suppl_1)：P3706.