刘明明, 王广静, 赵子涵, 等. 基于机器学习乳腺癌预测及 SHAP 特征分析[J]. 智能计算机与应用,2024,14(10):194-200.
DOI:10.20169/j.issn.2095-2163.241028

基于机器学习乳腺癌预测及 SHAP 特征分析

刘明明[1], 王广静[1], 赵子涵[2], 骆谋萸[1], 谢 静[1]
(1 蚌埠医科大学 公共基础学院, 安徽 蚌埠 233030; 2 蚌埠医科大学 公共卫生学院, 安徽 蚌埠 233030)

摘 要: 乳腺癌作为全球新发病例最多的癌症,严重影响和损伤人们生命质量,乳腺癌的预测与了解其发病机制是目前仍需更多研究的问题。针对乳腺癌诊断的准确性需求,本文旨在通过应用机器学习算法提升乳腺癌预测模型的精确度,为医生的决策制定提供支持,有效实现“三早”预防,并为疾病病因的深入研究提供新的线索。以美国威斯康星州在 Kaggle 平台发布的乳腺癌公开数据集为研究对象,首先在数据预处理后,借助随机森林的递归特征消除法进行变量的重要性排序和特征选择。其次,利用网格搜索法优化超参数,运用 LightGBM 算法构建预测模型,并引入 SHAP 值增强模型的可解释性,进一步揭示乳腺癌相关的危险因素及其作用机制。最后,通过 *AUC* 值等评价指标对模型的预测性能进行评估。结果表明,模型的表现优于传统模型,预测准确率达到 97%,且 *AUC* 值为 0.97,有效提升了乳腺癌的正确识别能力。

关键词: 机器学习; 乳腺癌预测; LightGBM 算法; SHAP 值
中图分类号: TP181 **文献标志码**: A **文章编号**: 2095-2163(2024)10-0194-07

Machine learning based breast cancer prediction and SHAP feature analysis

LIU Mingming[1], WANG Guangjing[1], ZHAO Zihan[2], LUO Mouyu[1], XIE Jing[1]
(1 School of Public Base, Bengbu Medical University, Bengbu 233030, Anhui, China;
2 School of Public Health, Bengbu Medical University, Bengbu 233030, Anhui, China)

Abstract: Breast cancer is the most prevalent cancer in the world and has a serious impact on the quality of life. The aim of this paper is to improve the accuracy of breast cancer prediction models by applying machine learning algorithms to support doctors´ decision making, to achieve "three early" prevention and to provide new clues for further research on the cause of the disease. Using a public dataset of breast cancer from the state of Wisconsin published on the Kaggle platform, firstly the data is pre-processed and then the recursive feature elimination method of random forests is used to rank the importance of variables and select features. Secondly, a grid search method is used to optimize the hyperparameters, and the LightGBM algorithm is applied to construct a prediction model, and SHAP values are introduced to enhance the interpretability of the model to further reveal the risk factors associated with breast cancer and their mechanisms of action. Finally, the predictive performance of the model is assessed by evaluation indicators such as *AUC* values. The results shows that the model outperforms the traditional model, with a prediction accuracy of 97% and an *AUC* value of 0.97, effectively improving the correct identification of breast cancer.

Key words: machine learning; breast cancer prediction; LightGBM algorithm; SHAP values

0 引 言

近年来,乳腺癌已成为全球女性发病率最高的恶性肿瘤[1]。2021 年世卫组织最新数据显示乳腺癌已经成为了全球新发病例最多的癌症,中国乳腺癌的发病数位居全球第四(WHO,2020)[2]。早期发现及有效治疗,可以改善癌症分期,并降低乳腺癌的死亡率[3]。因此,研究乳腺癌发病的相关危险因素,寻找科学有效的防治措施对维护女性身心健康具有重要意义[4-5]。基于此,本文以美国威斯康星州乳腺癌公开数据集[6]作为研究对象,将机器学习与诊断手段结合,帮助临床医生进行快速有效的乳腺癌诊断,大量节省人工成本。同时运用 SHAP 值可解释性分析乳腺癌相关危险因素和其内部机制。

基金项目: 安徽省高校人文社会科学重点项目(SK2020A0357); 蚌埠医学院自然科学重点项目(KYBY1704ZD)。
作者简介: 刘明明(2001-),女,本科生,主要研究方向:机器学习,物联网工程。
通讯作者: 谢 静(1985-),女,讲师,主要研究方向:物联网,计算机教学。Email:xiejingbbmc@163.com
收稿日期: 2023-06-12

由于机器学习在疾病风险预测方面可以提高预测准确率，因此应用机器学习算法预测疾病的发病风险已经成为当今研究的重要课题。例如，高媛媛[7]基于多特征融合和机器学习的疾病基因检测大数据分类模型构建方法，采用主成分分析融合数据特征，达到了效率高、病情反映能力强的疾病基因大数据分类。陈静雯等学者[8]采用支持向量机、K 近邻、决策树和随机森林多种机器学习方法建立呼吸道疾病预测模型，并提供自动和手动数据特征选择方式，以数据可视化方式展示给用户。黄光成等学者[9]则比较了多种机器学习算法在不同疾病预测中的应用，叙述了不同类型的机器学习算法的适用条件，为预测特定疾病选取合理的机器学习算法提供了支持。而关于乳腺癌的预测，国内外很多学者采用了多种机器学习算法进行了深入探究。Anisha 等学者[10]使用机器学习中的随机森林算法预测乳腺癌，并比较了决策树、逻辑回归等算法的预测准确率，证明了随机森林算法的预测准确率高于其他。Zorgani 等学者[11]使用 K-近邻，SVM 算法对乳腺癌进行分类，结果表明，SVM 和 KNN 的分类准确率分别达到了 93.75%和 88.75%。吴泽琪等学者[12]采用随机森林(Random Forest, RF)、极端梯度提升(Extreme Gradient Boosting, XGBoost)、逻辑回归(Logistics Regression, LR)和支持向量机(Support VectorMachine, SVM)算法构建乳腺癌腋窝淋巴结转移预测模型，并对比了这些模型的性能。

本文以美国威斯康星州乳腺癌公开数据集作为研究对象，将机器学习与诊断手段结合，帮助临床医生进行快速有效的乳腺癌诊断，大量节省人工成本。文章首先通过随机森林的递归特征消除法对特征进行筛选，利用初步筛选出的特征使用 LightGBM 算法构建乳腺癌风险预测模型，使用网格调参和 K 折交叉验证的方法得到模型的最佳参数。最后，通过分析 SHAP 值，本文识别与预测出了为恶性肿瘤相关的危险因素及其对预测结果的影响。

1　乳腺癌分析预测方法

1.1　基于随机森林的递归特征消除法

RF-RFE 算法采用随机森林分析变量重要性，并根据变量重要性排序，进而通过 RFE 方法选择重要变量[13]。其原理是通过构建随机森林模型来评估每一个特征的重要性，并将最不重要的特征逐步消除，重新训练随机森林模型，直到剩余的特征数量达到预设的阈值或者模型性能不再提升，从而提高模型的预测性能和鲁棒性。

1.2　LightGBM 算法

LightGBM[14]是轻量级的梯度提升机器，是 GBDT 模型的另一个进化版本[15]，包括基于直方图的决策树算法、直方图差加速，限制最大深度的 Leaf-wise 叶子节点生长策略等。

LightGBM 运用了直方图算法。其思想是：首先，将连续的浮点特征值离散化为多个离散值，用来构筑宽度直方图(bin)。接着，再循环训练数据，以离散后的特征值为索引，统计直方图中的 所有离散值的累计统计量，最后根据索引遍历寻找最佳的分割点后，进行特征选择。对于直方图算法，LightGBM 还进行了进一步的优化。使用了具有最大深度限制的叶子节点生长(Leaf-wise)策略，从当前所有叶子节点中找到分裂所计算的增益最大，即数据量最大的叶子节点，让其分裂，再循环重复这一步骤，进行多次分裂。并且，该算法对 Leaf-wise 设置了最大深度的限制，即设置超参数叶子数和最大深度(max depth)，在保持高效率的同时防止过拟合。

1.3　模型评价方法

在机器学习中，评价指标用于评估模型的性能和预测能力。本文使用准确率 (*Accuracy*)、精确率(*Precision*)、召回率(*Recall*)、*F*1 - Score、*AUC* 作为评价指标，具体介绍如下。

表 1　混淆矩阵

Table 1　Confusion matrix

	预测为恶性	预测为良性
真实为恶性	*TP*	*FN*
真实为良性	*FP*	*TN*

表 1 中，*TP* 表示正确预测为恶性样本数量，*FN* 表示错误预测为恶性样本数量，*FP* 表示错误预测为良性的样本数量，*TN* 表示正确预测为良性的样本数量。接下来，将对各指标给出阐释分述如下。

(1)准确率 (*Accuracy*)：即模型预测正确样本占总样本数的比例。数学计算公式具体如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

(2)精确率 (*Precision*)：即模型预测为正例(Positive)的样本中，实际为正例的样本所占的比例。数学计算公式具体如下：

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

(3)召回率 (*Recall*)：即实际为正例的样本中，

被模型正确预测为正例的样本所占的比例。数学计算公式具体如下：

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

(4) $F1$ 得分($F1$ - Score)：即精确率和召回率的加权平均值,用于综合评估模型的性能。数学计算公式具体如下：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

(5) AUC 值(Area Under the Curve)：ROC 曲线下的面积,用于综合评估模型的性能和鲁棒性。AUC 的取值范围在 0.5~1.0 之间,越接近 1 表示模型性能越好,越接近 0.5 表示模型性能越差。

2 SHAP 值方法

考虑到大多数集成学习算法虽然预测能力优异,但却缺乏可解释性且难以确定影响疾病进程的关键因素,而这对协助医生诊断和治疗疾病至关重要。因此,本文采取 SHAP 值对冠心病预测模型进行可解释性分析。

SHAP 值是衡量特征的边际贡献度,是当前模型解释的最佳方法之一,对于模型进行可视化的全局解释、局部解释,可以在一定程度上满足业务对于模型解释性的要求。其全局解释(特征对于整体模型的影响)可以作为特征重要性帮助筛选变量;局部解释(对单个样本的预测结果进行解释)可以直观地看到单个样本预测结果的主要影响因素,即特征有哪些、以及相应的影响程度。

SHAP 值的计算方法源于博弈论,表示在一个有限的合作游戏中,一个玩家对整个游戏价值的贡献。在 SHAP 模型中,将特征视为玩家,模型预测值视为游戏价值。

SHAP 值满足以下 4 个公平性原则,分别是:效率、对称性、空值玩家和线性。SHAP 值的计算公式如下：

$$\phi_i(f) = \sum_{S \subset N \setminus i} [f(S \cup \{i\}) - f(S))] \frac{[|S|!\ (|N| - |S| - 1)!\ ]}{|N|!} \tag{5}$$

其中,$\phi_i(f)$ 表示特征 i 的 SHAP 值;f 表示模型;N 表示特征集合;S 表示不包含特征 i 的特征子集。式(5)表示特征 i 的 SHAP 值为所有可能的特征子集 S 对模型预测结果的贡献的加权平均值。权重是根据特征子集的大小计算的,确保了公平性原则得到满足。

3 结果分析

3.1 数据来源及预处理

本文选取 Kaggle 中威斯康星乳腺癌数据集[6]。该数据集共有 569 个样本,共有 31 列特征。其中,357 例为良性样本,212 例为恶性样本。

首先,对样本进行预处理,使用 3σ 原则,检测数据集中是否存在异常值,对于存在的异常值数据使用中位数进行填充,并对数据集中连续性变量进行标准化处理,消除量纲。

由于数据集中存在较多的特征,本文基于上述数据的处理,选择基于随机森林的递归消除特征法,对特征进行筛选,选取 15 个重要特征,详见表 2。

表 2 特征含义表

Table 2 Feature meaning

变量名	特征含义
radius_mean	半径(点中心到边缘的距离)平均值
concave_points_worst	凹缝(轮廓的凹部分)最大值
concavity_worst	凹度(轮廓凹部的严重程度)最大值
smoothness_worst	平滑程度(半径内的局部变化)最大值
perimeter_worst	周长最大值
area_worst	面积最大值
radius_worst	半径(点中心到边缘的距离)最大值
texture_worst	纹理(灰度值的标准值)最大值
perimeter_se	周长标准差
concavity_mean	凹度(轮廓凹部的严重程度)平均值
texture_mean	纹理(灰度值的标准值)平均值
concave points_mean	凹缝(轮廓的凹部分)平均值
perimeter_mean	周长平均值
area_mean	面积平均值
area_se	面积标准差

在以上数据处理的基础上,将数据集以 7:3 的比例划分为训练集和测试集。

3.2 结果分析

为了保证模型能有更好的效果,通常需要采取对模型调参的操作,因此采取网格搜索的方法来解决机器学习过程中的超参数搜索问题。目前,网格搜索优化法是最简单、应用最广泛的超参数搜索算法,通过对超参数组合列表中的每一个组合,实例化给出模型,做 cv 次交叉验证,这里选择 5 次交叉验证,将平均得分最高的超参数组合作为最佳的选择,再返回模型对象。因此,使用网格搜索优化法来进行超参数调优。LightGBM 集成学习模型基于网格搜索优化的优化结果的参数的详细设置情况见

表 3。

表 3　重要参数值

Table 3　Important parameter values

参数	*learning_rate*	*max_depth*	*min_child_weight*	*n_estimators*
值	0.1	6	6	100

由表 3 可知，使用网格寻参和五折交叉验证的方法得出 *LightGBM* 算法的重要参数分别是：*learning_rate* 为 0.1，*max_depth* 为 6，*min_child_weight* 为 6，*n_estimators* 为 100。

基于上述设置来构建模型在测试集上进行验证，仿真结果见表 4。实验仿真得到的 ROC 曲线如图 1 所示。

表 4　评价结果

Table 4　Evaluation results

评价指标	准确率	精确率	召回率	*F*1 − Score
结果	0.970	0.953	0.968	0.960

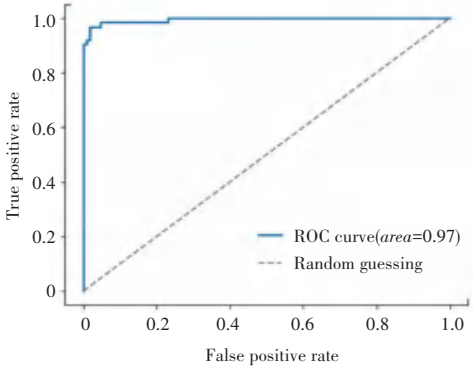


图 1　ROC 曲线

Fig. 1　ROC curve

结合表 4 和图 1 可以看出，该模型性能表现良好，准确率达到 97%，且 *AUC* 值为 0.97。

3.3　SHAP 分析

为了进一步研究乳腺癌的主要影响因素，提升分类模型的可解释性，本文引入 SHAP 方法对乳腺癌数据集进行特征分析。图 2 为乳腺癌数据集的特征重要性分析结果。由图 2 可知，肿瘤凹缝（轮廓的凹部分）最大值、纹理（灰度值的标准差）平均值和面积标准差是影响乳腺癌的重要因素。

对于划分好的数据，本文借助 Python 中的 SHAP 库计算了 LightGBM 模型的 SHAP 值，如图 3 所示。每个点对应数据集的一个实例，即一个样本。x 轴上的位置、即实际的 SHAP 值，表示该特征对特定样本的模型输出、即对特定样本的相对患病风险的影响[16]。换言之，较高 SHAP 值的样本相对于较低 SHAP 值的样本具有较高的患恶性肿瘤风险。此外，各个特征按其重要性沿 y 轴排列，其重要性由其绝对 SHAP 值的平均值给出。特征位置越高，说明重要性越高。

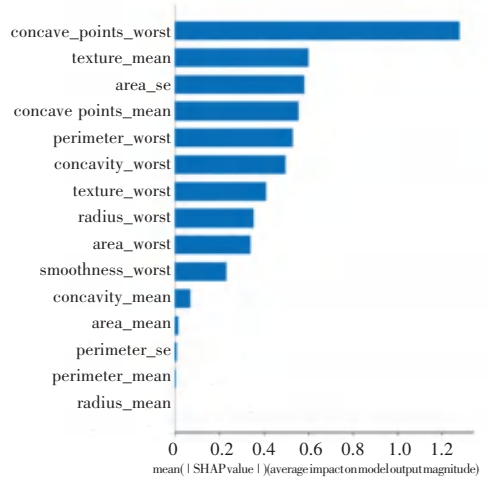


图 2　特征重要度

Fig. 2　Feature importance

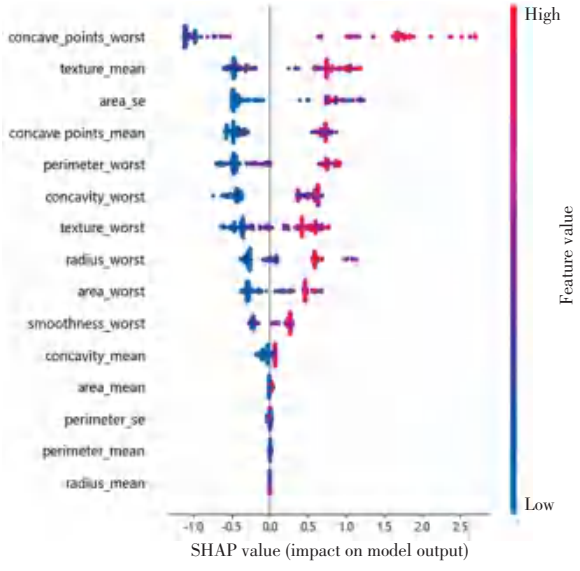


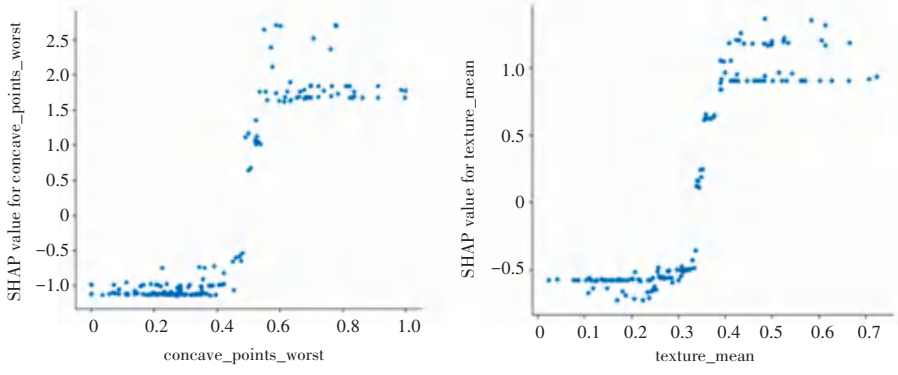
图 3　SHAP 蜂窝图

Fig. 3　SHAP honeycomb diagram

SHAP 值反映特征对患乳腺癌种类的相对风险的影响，SHAP 值越高，对预测患恶性肿瘤率的贡献越大。图 3 显示在 LightGBM 模型中，除了肿瘤凹度（轮廓凹度的严重程度）平均值、面积平均值、周长标准差、周长平均值和半径（点中心到边缘的距离）平均值的 SHAP 值趋近于 0，表明这些特征对乳腺癌预测结果贡献基本为 0 之外。其余特征均倾向于有很长的右尾。这表明，该特征在较多样本的预测值中均产生了较高的恶性肿瘤风险分数，这就意味着，这些样本患乳腺癌恶性肿瘤率会更高。从直观而言，这个结果是有一定道理的。

SHAP 变化如图 4 所示。从图 4 中可以看到肿

瘤凹缝(轮廓的凹部分)最大值和纹理(灰度值的标准差)平均值都在大于某一个临界值之后开始对患恶性肿瘤风险预测产生正的影响,即相应数值越大,预测样本的患恶性肿瘤风险就越大。而其余特征对预测产生的影响是不确定的。以肿瘤凹缝(轮廓的凹部分)最大值为例,如图 4(a)所示,当肿瘤凹缝(轮廓的凹部分)最大值大于 0.5 时,随着特征值的增加 SHAP 值越来越大,即此时该特征对患恶性肿瘤风险的预测产生正影响。而在肿瘤凹缝(轮廓的凹部分)最大值小于 0.5 之前,该特征所在的 SHAP 值均为负值,且随着肿瘤凹缝(轮廓的凹部分)最大值的下降 SHAP 值也越来越小,直到维持在-1.0 左右。说明此时该特征对患恶性肿瘤风险的预测产生负影响。图 5 为肿瘤凹缝(轮廓的凹部分)最大值的部分依赖图(Partial Dependence Plot,PDP)。从图 5 中可以看出,在肿瘤凹缝(轮廓的凹部分)最大值小于 0.5 时,预测结果约为 0.4,表明该特征值在此范围内的变化对预测结果影响不大且样本预测为良性肿瘤的概率大。而在肿瘤凹缝(轮廓的凹部分)最大值大于 0.5 时,该特征值的增加对预测结果产生显著影响且增加了样本被预测为恶性肿瘤的概率。直至肿瘤凹缝(轮廓的凹部分)最大值大于 0.6,该特征对样本预测为恶性的概率维持在 0.6 左右基本不变。



(a) concave_points_worst SHAP 变化图 (b) texture_mean SHAP 变化图

图 4 SHAP 变化图

Fig. 4 SHAP change chart

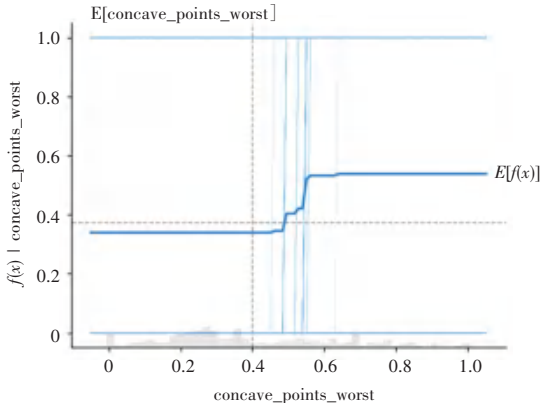


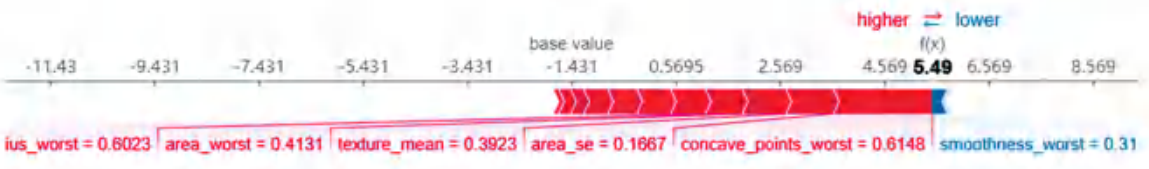
图 5 部分依赖图

Fig. 5 Partial dependency chart

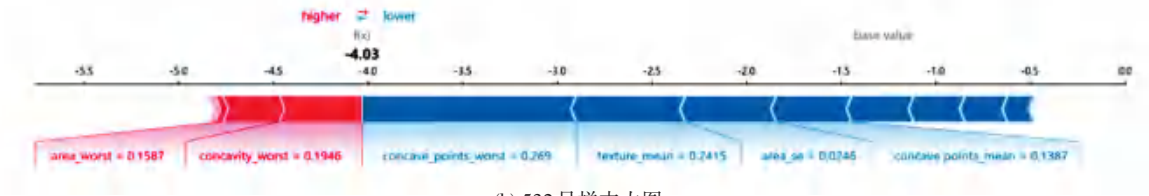
除了对患癌风险的影响预测进行总体分析之外,本文还对患病风险的影响预测进行了个体样本分析,试图从总体到个体更为全面地分析患癌风险各因素对患癌风险的影响。本文随机选择 2 个样本,分析影响因素,其中 SHAP 解释图中蓝色表示对被诊断为恶性肿瘤有负向影响,红色表示被诊断为恶性肿瘤预测有正向影响。基线(Base Value)为平均预测概率,是所有样本汇总的平均预测值[17],因此每一个样本的基线值均相同、为-1.431。将该样本 SHAP 值汇总之后得到 $f(x)$。

SHAP 值随机个例解释如图 6 所示,决策路径如图 7 所示。从图 6(a)随机选择的 70 号数据可以看出,该样本的患恶性肿瘤风险预测为 5.49,高于基线值-1.431,说明该样本患恶性肿瘤风险较高,这与该样本被诊断为恶性肿瘤的结果也是对应的。除此之外,还可看出对于样本 70 号,产生正向影响的特征对预测结果起主要作用。在产生正向影响的特征中肿瘤凹缝(轮廓的凹部分)最大值对预测值的影响最大,接下来依次为面积标准差、纹理(灰度值的标准差)平均值、面积最大值和半径(点中心到边缘的距离)平均值。同时,分析图 7(a)中该样本的决策路径可知:在平滑程度(半径内的局部变化)最大值这一特征处折线向左偏斜,说明该特征对于预测结果起负向作用,且其纵坐标的排列顺序是按照该样本的特征重要度进行排序,因此,可以看出对于 70 号样本,影响其诊断结果的重要特征依次为:凹缝(轮廓的凹部分)最大值、面积标准差、纹理(灰度值的标准值)平均值等。

从图 6(b)随机选择的 532 号数据可以看出，该样本的患恶性肿瘤风险预测为-4.03，低于基线值-1.431，说明该样本患恶性肿瘤风险较小，这与该样本被诊断为良性肿瘤的结果是对应的。除此之外，还可看出对于患者 532，肿瘤凹度(轮廓凹度的严重程度)最大值和面积最大值对其预测值产生较大正向影响，肿瘤凹缝(轮廓的凹部分)最大值特征对预测结果的负向影响最大，这就反映出模型预测该患者患恶性肿瘤风险较小的内部机制。同时，结合图 7(b)该样本的决策路径可知：在面积最大值、面积平均值等特征处折线向右偏斜，说明该特征对于预测结果起正向作用，其余特征对于预测结果起负向作用，且根据纵坐标可以看出对于 532 号样本，影响其诊断结果的重要特征依次为：凹缝(轮廓的凹部分)最大值、纹理(灰度值的标准值)平均值、面积标准差等。



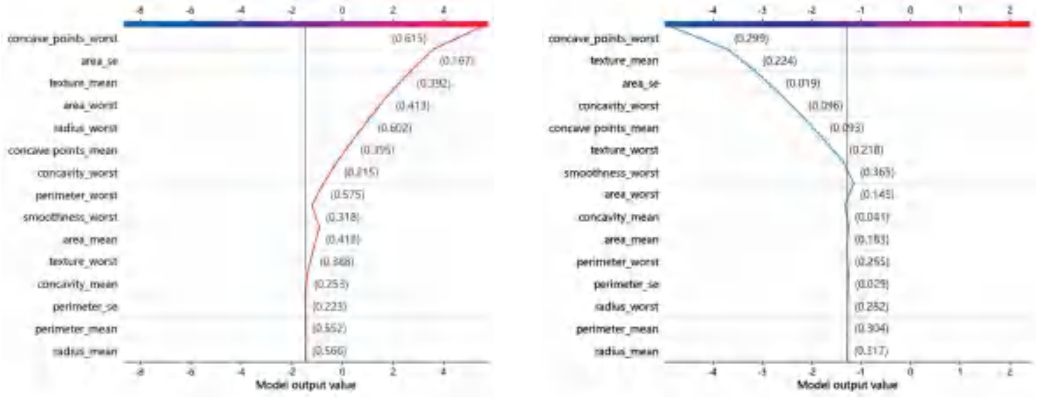
(a) 70 号样本力图



(b) 532 号样本力图

图 6　SHAP 值随机个例解释图

Fig. 6　Example of SHAP interpretation for random cases



(a) 70 号样本　　(b) 532 号样本

图 7　决策路径

Fig. 7　Decision path

4　结束语

本文模型构建的思想基于 LightGBM 算法并使用网格搜索优化法建立乳腺癌预测模型，相比于传统机器学习模型，被证明在预测乳腺癌方面具有更好的效果。在分析过程中，基于 SHAP 值对乳腺癌预测模型进行了可解释研究，提供每一特征对于预测的重要性程度的参考，最终得出诊断为恶性肿瘤主要因素有凹缝(轮廓的凹部分)最大值、纹理(灰度值的标准值)平均值、面积标准差等，为临床医生进行快速有效的乳腺癌诊断提供依据。本文的局限性在于所使用的数据集特征较为常规，如果有更丰富全面的数据，预测模型的效果可能会有进一步的提升。随着更高质量临床数据的收集和存储，在未来的研究中可以使用更多样和全面的数据集，纳入更多相关的因素于模型中，进一步提升乳腺癌预测

准确率。

参考文献

[1] 张雅聪,吕章艳,宋方方,等 . 全球及我国乳腺癌发病和死亡变化趋势[J]. 肿瘤综合治疗电子杂志,2021,7(2):14-20.

[2] 何梦. “丁香医生”微信公众号乳腺癌信息传播的框架研究[D]. 北京:北京外国语大学,2022.

[3] 张玉辉. 中疾控:2030 年全国乳腺癌患者将超 40 万例[N]. 医师报,2023-04-20(A04).

[4] 谢小红,顾锡冬,赵虹,等. 973 例乳腺癌患病相关危险因素分析[J]. 中华全科医学,2014,12(6):960-962.

[5] 罗恩,杨晓虹,王雪清. 女性乳腺癌危险因素的成组病例对照研究[J]. 成都医学院学报,2013,8(3):269-273.

[6] 阿里云. 威斯康星乳腺癌数据分析及自动诊断[EB/OL].[2021-07-21]. https://tianchi. aliyun. com/dataset/106831.

[7] 高媛媛. 基于多特征融合和机器学习的疾病基因检测大数据分类模型[J]. 微型电脑应用,2023,39(3):25-27,39.

[8] 陈静雯,张鹏鹏,徐思语,等. 基于机器学习的呼吸道疾病预测可视化系统[J]. 物联网技术,2023,13(2):68-70.

[9] 黄光成,周良,石建伟,等. 机器学习算法在疾病风险预测中的应用与比较[J]. 中国卫生资源,2020,23(4):432-436.

[10] ANISHA P R, KISHOR K R C, APOORVA K, et al. Early diagnosis of breast cancer prediction using random forest classifier [J]. IOP Conference Series: Materials Science and Engineering, 2021,1116(1):012187.

[11] ZORGANI A, MOHAMED M, MEHMOOD I, et al. Learning transferable features for diagnosis of breast cancer from histopathological images [C]//International Conference on Medical Imaging and Computer-Aided Diagnosis. Singapore: Springer,2021:124-133.

[12] 吴泽琪,马梦伟,刘仁懿,等. 基于影像特征建立乳腺癌腋窝淋巴结转移机器学习预测模型[J]. 国际医学放射学杂志,2023,46(3):255-260.

[13] 商强,林赐云,杨兆升,等. 基于变量选择和核极限学习机的交通事件检测[J]. 浙江大学学报(工学版),2017,51(7):1339-1346,1445.

[14] KE Guolin, MENG Qi, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]// Advances in Neural Information Processing Systems. Long Beach, USA: NIPS Foundation, 2017, 30:3149-3157.

[15] 王静莹. 服务互联网价值建模与优化分析方法研究[D]. 哈尔滨:哈尔滨工业大学,2021.

[16] MOLNAR C. Interpretable Machine Learning[EB/OL]. [2022-01-16]. https://christophm. github. io/int-erpretable-ml-book/shaply. html.

[17] 束鹃. 基于可解释机器学习的城市道路交通事故严重程度预测[D]. 西安:长安大学,2021.