Modeling and Simulation 建模与仿真, 2025, 14(10), 1-11
Published Online October 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.1410600

Hans汉斯

基于TPE-SVM的乳腺癌诊断可解释人工智能方法

宋柯蕾[1]，傅文翰[1,2*]，王秋阳[1]，谢彤嫣[1]

[1]上海理工大学管理学院，上海
[2]上海理工大学智慧应急管理学院，上海

收稿日期：2025年8月28日；录用日期：2025年9月21日；发布日期：2025年9月28日

摘　要

乳腺癌是女性中最常见的恶性肿瘤之一，早期精准诊断对提高患者生存率至关重要。针对传统乳腺癌诊断方法存在的主观性强和误诊率高，以及现有人工智能模型可解释性较差等问题。本文设计一个基于贝叶斯优化支持向量机(TPE-SVM)的乳腺癌智能诊断模型，并结合LIME解释方法提高诊断过程中的可解释性。方法上，构建基于径向基核函数的支持向量机模型，并利用TPE算法对关键超参数进行优化，最后引入LIME方法实现诊断结果的特征可视化与解释。采用UCI数据库中的威斯康星乳腺癌数据集进行仿真验证，结果显示，该方法在各项指标上表现优异，进一步的LIME可解释性分析也表明，模型判别依据与临床医学知识高度一致。本文所构建的诊断框架为AI在医疗场景中的可用性与可信性提供了新思路。

关键词

可解释人工智能，乳腺癌诊断，支持向量机，贝叶斯优化，可解释性分析

Explainable AI Approach for Breast Cancer Diagnosis via TPE-Optimized Support Vector Machine

Kelei Song[1], Wenhan Fu[1,2*], Qiuyang Wang[1], Tongyan Xie[1]

[1]Business School, University of Shanghai for Science and Technology, Shanghai
[2]School of Intelligent Emergency Management, University of Shanghai for Science and Technology, Shanghai

Received: August 28, 2025; accepted: September 21, 2025; published: September 28, 2025

*通讯作者。

文章引用: 宋柯蕾, 傅文翰, 王秋阳, 谢彤嫣. 基于TPE-SVM的乳腺癌诊断可解释人工智能方法[J]. 建模与仿真, 2025, 14(10): 1-11. DOI: 10.12677/mos.2025.1410600

Abstract

Breast cancer is one of the most common malignant tumors among women, and accurate early diagnosis is crucial for improving patient survival rates. To address the high subjectivity and misdiagnosis rates of traditional diagnostic methods, as well as the limited interpretability of current artificial intelligence models, this study proposes an intelligent diagnostic model for breast cancer based on a Bayesian-Optimized Support Vector Machine (TPE-SVM), integrated with the LIME explanation method to enhance interpretability during the diagnostic process. Methodologically, a Support Vector Machine with a radial basis function kernel is constructed, and its key hyperparameters are optimized using the TPE algorithm. The LIME method is subsequently employed to visualize and interpret the model's diagnostic decisions. The model is validated on the Wisconsin Breast Cancer dataset from the UCI repository. Experimental results demonstrate excellent performance across multiple evaluation metrics, and further LIME-based interpretability analysis confirms that the model's decision-making criteria are highly consistent with established clinical knowledge. The diagnostic framework proposed in this study offers new insights into the usability and trustworthiness of AI applications in medical scenarios.

Keywords

Explainable Artificial Intelligence, Breast Cancer Diagnosis, Support Vector Machine, Bayesian Optimization, Interpretability Analysis

1. 引言

乳腺癌，被称为全球第一大癌症，也是女性群体中发病率和致死率最高的癌症。在中国，I期乳腺癌患者的确诊比例接近 31.8%，美国约 54.6%，相关学者认为中国早期乳腺癌的确诊比率低于发达国家的主要原因是早期筛查不充分[1]。因此，早期乳腺癌病理诊断的准确性至关重要。

随着人工智能的发展，尤其在图像识别和数据分析领域的发展，为疾病的智能诊断提供了许多新方法，如支持向量机(SVM)、多层感知机(MLP)、随机森林(RF)或深度神经网络(DNN)等，然而，机器学习和深度学习用于智能诊断，虽然在诊断性能上取得了良好的效果，但普遍存在两个问题亟待解决。首先，模型超参数的选择高度依赖效率高，计算成本小，缺少取值准确的调参方法。其次，模型的可解释性较差，使医生和患者难以信任诊断结果[2]。在实际情况中，医生们希望得到的不仅仅是简单的良恶性分类，而是需要更多的信息指导他们编写诊断报告和进行进一步检查[3]。因此，提升诊断模型及其诊断预测结果的可解释性显得尤为重要。

模型解释研究目前仍处于起步阶段，有一定的理论研究和模型研究成果，从可解释性的角度，人工智能模型可分为透明盒(Transparent Box)模型和黑盒(Black Box)模型[4]。对于一些复杂的黑盒模型，其本身模型结构过于复杂，不便理解，决策过程不可见，则需要对输入信息和输出信息进行分析，用额外的研究方法和模型去拟合复杂模型的决策过程，这种方法被称为归因解释方法。根据 Gunning 等[5]对可解释人工智能解释方法的分类，解释方法可以分为深度解释、可解释模型、模型归纳。本文所运用的 LIME 代理模型方法就属于模型归纳。为了更好理解模型预测的原因，提高模型的可信任度，Ribeiro 等人[6]提

出了一种新颖解释方法——LIME，通过在预测结果周围局部学习一个可解释的模型，来解释任何分类器的预测。Kumarakulasinghe 等[7]基于医生的反馈评估 LIME 在临床领域的可解释性，考虑了 LIME 解释和医生对模型预测的一致性。Jalali 等[8]对比分析了 LIME、SHAP、Attribute-Wise 三种可解释性技术在智能诊断模型中的应用，试验结果表明三种不同的代理模型方法对大部分黑盒模型的解释是一致的。Onchis 等[9]将 LIME 方法与 SHAP 方法结合，对用于桥梁损伤检测的神经网络进行特征重要性排序。Zhu 等[10]开发一种机器学习预测模型，来评估急性冠脉综合征患者术后发生不良心血管事件的风险，并利用 SHAP 和 LIME 方法分析特征权重。

综上，本文提出了一种基于贝叶斯优化支持向量机(TPE-SVM)的乳腺癌诊断模型，并引入 LIME 可解释性方法进行结果解释。首先，考虑到在医学诊断任务中，不同类型分类错误所带来的后果不同。因此本研究在设计 SVM 时引入类别加权惩罚因子，提高模型对恶性肿瘤的识别能力。然后，采用树结构 Parzen 估计器(TPE)算法对 SVM 模型的关键超参数进行高效优化。最后，结合局部代理解释方法 LIME，实现对模型分类依据的特征级可视化，增强模型诊断的透明性与可信度。由于已有的技术相对成熟，本文针对乳腺癌早期诊断中的关键需求，构建了一个系统化的诊断框架：结合类别加权 SVM 与 TPE 算法，形成高效的集成优化机制；引入局部解释方法 LIME，实现对诊断结果的特征级可视化；最后在标准医学数据集上进行系统性实验验证。

2. 相关工作

随着人工智能及大数据技术的快速发展及广泛应用，越来越多的计算机辅助诊断方法被提出和应用。杨萌宇等[11]提出了一种基于动态粒子群优化(DPSO)算法和支持向量机的 DPSO-SVM 诊断模型，与主流诊断模型相比，该模型有着更优越的性能，提升了诊断效率的同时也降低了诊断误差。Kim [12]提出了一个对类不平衡多模态数据的成本敏感的支持向量机，通过提取不同模态数据的特征而获得融合多网络结构，并使用成本敏感的支持向量机作为分类器，该方法在解决不平衡数据问题上表现出色。Lotter 等[13]提出了一种具有注释效率的 DL 方法，该方法在乳房 X 光片分类等方面实现了最先进的性能，相较于乳腺影像专家，灵敏度平均增加 14%。Lu 等人[14]借助多实例学习框架，融合 ResNet [15]和注意力机制[16]，实现对结直肠癌病理切片不同区域诊断分数的预测，在结直肠癌二分类问题上取得 98%的准确率。Witowski 等[17]基于 MRI 图像采用增强前、后的乳腺体积构建深度学习模型，来预测 13,463 例可疑患者患乳腺癌的概率，深度学习模型的诊断准确率与 5 名乳腺放射科医生相当。Shafi 等[18]提出了一种基于深度学习支持向量机的癌症诊断模型。通过测量和比较患者，以及对照患者诊断的 CT 图像，对代表早期肺癌的肺结节检测产生了 94%的准确率。龚安等[19]基于深度学习提出了一种多模型融合机制方法，使用六个 CNN 通过迁移学习策略进行训练，提取多网络特征并保存，选出最优两个模型 ResNet50 和 Inception_v3 进行融合；实现病理图像在不同放大倍数下的多级分类。徐坤财等[20]提出一种基于卷积神经网络和 Transformer 的两阶段乳腺癌病理图像分类方法。经过验证，该方法具有较高的诊断性能，可以作为分析乳腺癌病理图像的一种通用方法。

3. 基于 TPE-SVM 的乳腺癌可解释诊断模型

本文提出基于支持向量机(Support Vector Machine, SVM)的乳腺癌诊断模型，结合 TPE 算法(Tree-Structured Parzen Estimator)进行超参数的优化，提高模型性能，最后使用 LIME (Local Interpretable Model-agnostic Explanations)方法进行预测结果的可解释性分析。基于 TPE-SVM 的乳腺癌可解释性诊断模型如图 1 所示。

本研究的模型主要包括以下四个步骤：

a) 不平衡数据预处理，在医学诊断领域，经常存在样本类别分布不均的情况，通常表现为阳性病例(癌症患者)远少于阴性病例。这种数据不平衡的问题通常会导致分类器预测偏向于多数类样本，对少数类样本的识别能力较弱。为应对该问题，本研究选取 SMOTE (Synthetic Minority Over-Sampling Technique)方法对训练集进行平衡处理，合成新的少数类样本，使得输入数据变得均衡化。

b) 分类器，本研究选择支持向量机作为智能诊断模型分类器。支持向量机在医疗领域中样本数小，在特征数多的“高维小样本”场景中表现优异，能够在保障精度的同时降低误诊风险。此外，在医学诊断领域中，不同的分类错误带来的后果是不同的，以乳腺癌诊断为例，将恶性肿瘤误诊为良性会导致错过最佳治疗时间，其代价远高于将良性肿瘤误诊为恶性，因此，有必要在模型中引入代价敏感机制，对不同类别设置不同的误分类代价。

c) 参数优化，支持向量机用于诊断分类时，受到惩罚参数 C 和超参数 γ 的影响，选取不同的参数值，会影响准确率。在实际应用时，往往要对参数进行合理选取。在本研究中，为了有效地优化支持向量机的超参数，我们选取贝叶斯优化(Bayesian Optimization)作为调参方法。贝叶斯优化通过构建代理模型与采集函数，在全局范围内高效搜索最优超参数，适用于高计算成本、非线性或黑盒模型调参，具备采样效率高、样本利用率强等优势，特别适合支持向量机等模型的优化任务。

d) 可解释分析，为了增强癌症诊断模型输出的可理解性，本研究选取了 LIME 作为解释方法。LIME 能够在不依赖模型结构的前提下，对任意复杂模型提供局部预测解释。LIME 具有广泛的实用性，其局部解释能力更加贴合每一个样本的实际情况，其输出的特征贡献直观明了，便于用户理解。

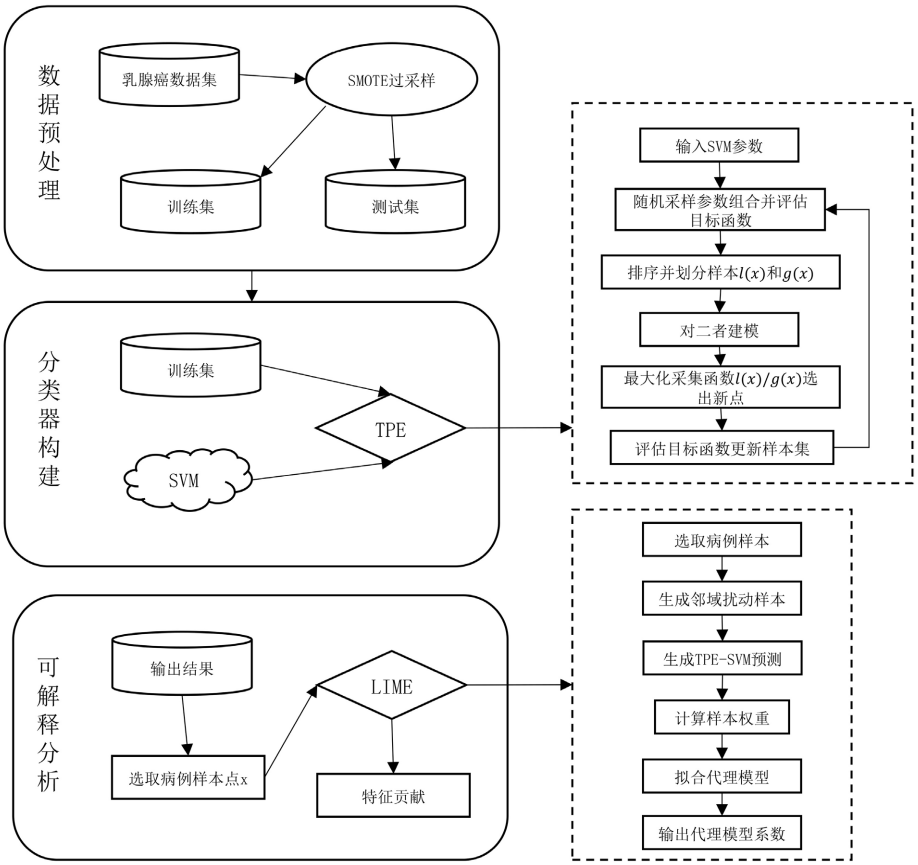

Figure 1. Structure of the interpretable breast cancer diagnostic model based on TPE-SVM
图 1. 基于 TPE-SVM 的乳腺癌可解释性诊断模型结构

3.1. 支持向量机

支持向量机是一种有监督的机器学习方法，常常用来进行分类和回归问题。对于一个给定的样本集，它的目标是找到一个超平面对样本进行分割。为了使样本到超平面的间隔最大，找到最优超平面。本文在 SVM 模型中引入类别加权惩罚因子，用于调整不同类别的代价差异。

改进后的求解问题如下：

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C^+ \sum_{i/y_i=+1}^{N} \xi_i + C^- \sum_{i/y_i=-1}^{N} \xi_i$$
$$s.t.\ y_i\left(w^T x_i + b\right) \geq 1-\xi, i = 1,2,\cdots,N$$
$$\xi_i \geq 0$$

其中，C^+ 表示对恶性样本的惩罚权重，C^- 表示对良性样本的惩罚权重。在实际训练中，本文设置类别权重比为 $\alpha = \frac{C^+}{C^-}$，并作为超参数纳入 TPE 优化流程中。

此外，对于线性不可分的样本数据，将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。因此引入核函数 $K\left(X_i, X_j\right)$，通过核函数将低维数据映射到高维空间。本研究使用径向基核函数作为 SVM 的核函数，如下：

$$K\left(X_i, X_j\right) = \exp\left(-\gamma\left\|X_i - X_j\right\|^2\right)$$

其中，γ 是超参数，控制高维空间的投影方式，γ 的改变会影响 SVM 的准确性。

3.2. 贝叶斯优化

贝叶斯优化(Bayesian Optimization, BO)是一种基于贝叶斯定理的全局优化算法，其核心思想是通过建立目标函数的概率模型，在每次迭代中选择最有可能带来优化的点进行评估，以此来寻找使目标函数达到最优值的参数组合。

贝叶斯优化框架主要包含两个核心部分——概率代理模型(Probabilistic Surrogate Model)和采集函数(Acquisition Function)。贝叶斯优化通常假设目标函数服从一个先验分布，常用模型为高斯过程，通过此模型预测函数值，并在每次迭代中依据当前代理模型和选择函数确定新的评估点，新的数据用于更新概率模型，该过程持续进行直至满足预设终止条件。

设 $f(x)$ 表示目标函数，χ 是定义域。通过高斯过程来给函数的概率建模，以估计目标函数 $f\left(x\right)$ 的后验分布：

$$f\left(x\right) \sim GP\left(\mu\left(x\right), K\left(x, x'\right)\right)$$

其中，$\mu\left(x\right)$ 表示均值函数，$K\left(x, x'\right)$ 为核函数，用来衡量两个输入值之间的相似度。由此得到 $f\left(x\right)$ 在不同输入值上的概率分布，采集函数基于该分布评估候选点的性能，选取合适的 x 点，使 $f\left(x\right)$ 更接近目标。

TPE 算法是贝叶斯优化的一种具体实现方式，与高斯过程建模目标函数不同，TPE 采概率密度估计的方法在参数空间采样。TPE 算法使用两个不同的概率密度函数来建模参数的条件概率分布，分别对应参数空间中表现良好和表现较差的区域。通过调整这两个概率密度函数，TPE 方法可以更好地进行搜索，从而找到最优解。

TPE 方法对目标函数的条件概率分布建模：

$$p\left(x|y\right)$$

其中，χ 是超参数组合，y 是评估模型结果。

TPE 将历史结果分为两个子集：

$L=\{X|y<y*\}$：性能好的参数集；

$G=\{X|y\geq y*\}$：性能差的参数集；

$y*$通常选择为历史观测值的某个分位数。

TPE 构建两个密度函数：

$$l(x)=p(x|y<y*)$$

$$g(x)=p(x|y<y*)$$

然后在选择下一个采样点时，优化以下目标：

$$x_{\text{next}}=\arg\max g\frac{l(x)}{g(x)}$$

如果这个比值高说明当前的参数组合出现在好模型中的概率更高，因此 TPE 会倾向于探索好的参数组合所在的区域。

3.3. 可解释性分析

LIME 是一种模型无关的局部解释方法，其核心思想是通过用可解释的模型对样本点附近对复杂模型进行局部近似，可以忠实地解释任何分类器或回归器的预测。假设对于需要解释的黑箱模型，LIME 取其中的实例样本，在其附近扰动生成新的样本点，并得到黑箱模型的预测值，使用新的数据集训练诸如线性回归和决策树等可解释的模型，最终得到对黑箱模型良好的局部近似。

假设有一个训练好的模型 f，对一个输入样本 $x\in R^d$ 做出预测 $f(x)$。但是由于 f 难以解析，LIME 引入一个简单的可解释模型 $g\in G$ (线性模型或决策树模型)，在 x 的邻域内近似 f，目标函数如下：

$$\xi(x)=\arg\min_{g\in G}\mathcal{L}(f,g,\pi_x)+\Omega(g)$$

其中，$\mathcal{L}(f,g,\pi_x)$ 表示局部损失函数，用来衡量黑盒模型 f 和简单模型 g 在样本点邻域内的拟合程度。π_x 表示邻域权重函数，衡量新样本点 x' 与 x 的相似程度。$\Omega(g)$ 表示 g 模型的复杂程度。

LIME 方法的流程如下：

1) 选择待解释乳腺癌病例样本 x；

2) 生成邻域扰动样本 z，构建新样本集 $\{z_i\}$；

3) 计算黑盒预测，将这些扰动样本输入 TPE-SVM 模型，获取对应的分类预测结果；

4) 计算样本权重，根据 z_i 到 x 的距离，对于乳腺癌的数值型临床指标，使用欧氏距离作为度量标准。权重函数定义如下：

$$\pi_x(z)=\exp\left(-\frac{D(x,z)^2}{\sigma^2}\right)$$

其中，$D(x,z)^2$ 表示距离度量。

5) 拟合代理模型，以新样本集 $\{z_i\}$ 的特征为输入，以 SVM 模型的预测结果为输出，结合样本权重 $\pi_x(Z)$，训练一个加权可解释的模型 g；

6) 输出代理模型 g 的系数 β_j，表示各乳腺癌指标在该诊断结果中所起的正负影响与重要程度。

4. 实验设计

4.1. 数据集介绍

本文采用 UCI 数据库中具有代表性的威斯康星医学院乳腺癌数据集(Wisconsin Diagnostic Breast Cancer, WDBC)为实验对象，该数据集总共包含 569 条病例记录，其中良性肿瘤 357 例，恶性肿瘤 212 例。每一个病例包含 30 个特征，涵盖 10 个实值特征的平均值、标准差和最大值，如表 1 所示：

Table 1. Dataset description
表 1. 数据集描述

NO	特征信息
1	ID
2	Diagnosis
3…12	特征(平均值)
13…22	特征(标准差)
23…32	特征(最大值)

4.2. 数据预处理

该数据集存在明显的类别不平衡特征，在处理静态非平衡医疗数据的分类任务中，决策平面通常更倾向于多数类别样本，导致少数类别样本的识别率较低，为了增强模型的鲁棒性，本文采用 SMOTE (Synthetic Minority Over-Sampling Technique)过采样方法平衡数据集。处理后数据集类别分布如表 2 所示：

Table 2. Sample class distribution after data balancing
表 2. 数据平衡后样本类别分布

类别	数量
样本总数	714
良性样本	357
恶性样本	357
特征数量	30

本研究以 8:2 比例划分测试集与数据集。类别分布一致。

4.3. 模型选择与超参数优化

本文采用支持向量机(SVM)作为分类器，选取径向基核函数。为了提高模型性能，采用贝叶斯优化方法中的 TPE 算法对 SVM 的关键超参数进行优化，主要包括：惩罚参数 C：控制间隔大小和误差衡量；核函数参数 γ：控制 RBF 核函数的宽度。

超参数搜索空间设定如表 3 所示：

Table 3. Hyperparameter search space settings
表 3. 超参数搜索空间设定

超参数	搜索范围	类型
惩罚参数 C	[10, 200]	连续型
核函数参数 γ	[0.0001, 20]	连续型
类别权重比 α	[1, 10]	连续型

采用 TPE 方法进行目标下的参数搜索，设置迭代次数为 100 次，确保找到最优解。

4.4. 实验结果与分析

为了验证本研究提出模型的性能，本文分别从模型效果评估、参数优化评估、可解释性分析三个层面进行结果分析。评价指标包括准确率(ACC)、F1-score、AUC 值、Recall 值，用于衡量模型在乳腺癌良恶性预测中的综合表现。乳腺癌良恶性对应混淆矩阵如表 4 所示，由表 4 得各指标计算方式如下：

Table 4. Confusion matrix of benign and malignant breast cancer
表 4. 乳腺癌良恶性的混淆矩阵

混淆矩阵		预测值	
		恶性	良性
真实类别	恶性	TP	FN
	良性	FP	TN

$$acc = \frac{TP + FN}{TP + FN + FP + TN}, precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}, F1 = \frac{2 \times precision \times recall}{precision + recall}$$

模型效果评估：用本研究模型与现有常用进行比较，具体结果如表 5 所示：

Table 5. Experimental results on the WDBC dataset
表 5. WDBC 数据集实验结果

研究方法	ACC	F1	AUC
SVM	0.9136	0.9168	0.9139
DT	0.9164	0.9175	0.9162
DES-CIER	0.9288	0.9296	0.9247
RF	0.9121	0.9121	0.9156
PSO-SVM [21]	0.9250	-	-
DES-KNN	0.9216	0.9231	0.9216
DES-DCR-CIER [3]	0.9522	0.9533	0.9514
SAPSO-SVM	0.9283	-	0.9578
本研究方法	0.9535	0.9510	0.9898

本章所展示的最终结果采用 5 折交叉验证获得，从表 5 可知，本研究在平衡处理后的 WDBC 数据集上，采用引入代价敏感机制的 TPE-SVM 模型进行分类实验，整体分类准确率为 95.35%，F1 分数为 0.9510。如图 2 所示，模型的 ROC 曲线下面积(AUC)高达 0.9898，取得了良好的表现。如图 3 所示，该模型的 PR-AUC 达到 0.9929，模型对恶性肿瘤的预测精度为 0.9848，F1 分数为 0.9489，表明模型在保持高检测准确性的同时，具备对恶性样本的高度敏感性。

综上进行分析，通过 TPE 算法对 SVM 的核函数参数、惩罚因子以及代价敏感权重进行联合优化，同时使用 SMOTE 算法进行过采样，减少了类别不平衡对模型训练的影响，显著提升了模型性能；此外，通过设置类别权重机制使模型对恶性肿瘤赋予更高分类权重，从而有效降低“假阴性”风险。

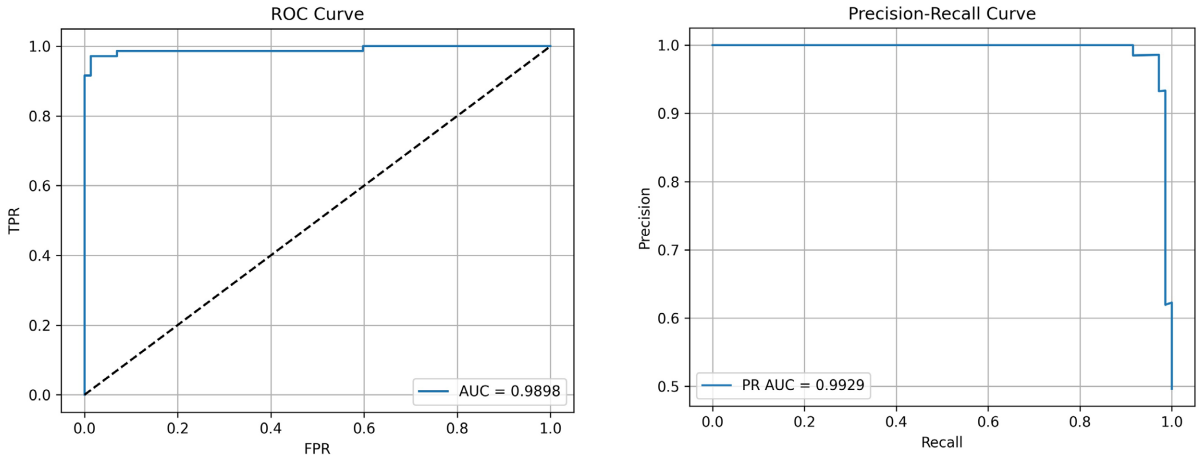

Figure 2. ROC curve and PR curve based on the TPE-SVM classification method
图 2. 基于 TPE-SVM 分类方法的 ROC 曲线和 PR 曲线

可解释性分析：为了提升本研究模型在实际临床应用中的可接受性与可信度，本研究采用 LIME 方法对模型的输出结果进行局部可解释分析。

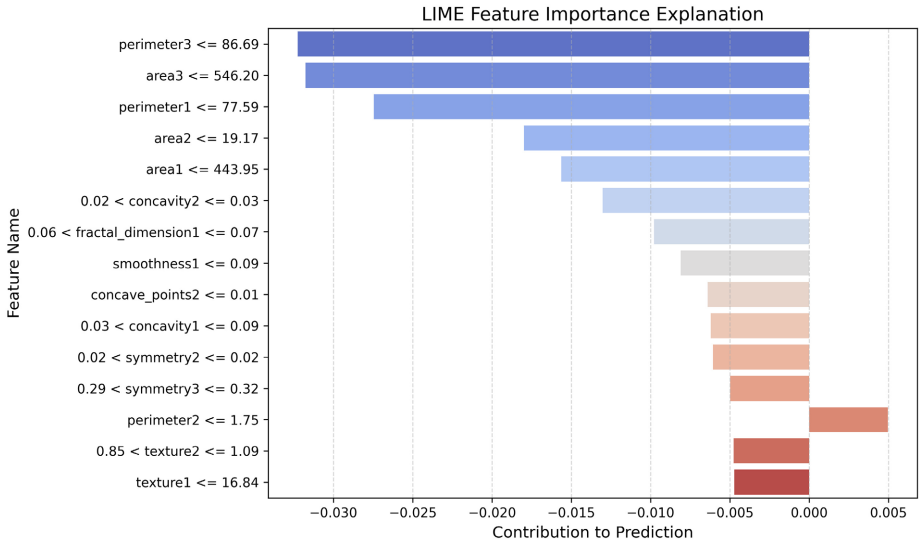

Figure 3. Feature contributions for malignant sample predictions
图 3. 恶性样本预测特征贡献

图 3 展示了模型对一个恶性肿瘤预测的可解释性分析，该样本被判定为“恶性肿瘤”的概率高达 99%，而被预测为“良性肿瘤”的概率仅为 1%，表明模型对该预测结果具有较高的置信水平。

如图 3 所示，进一步从特征层面进行分析，在 LIME 输出的局部特征重要程度中，模型主要依据以下特征诊断该样本为“恶性”：

area 3 ≤ 546.20，area 2 ≤ 19.17；

perimeter 3 ≤ 86.69，perimeter 1 ≤ 77.59。

其中“area”和“perimeter”代表肿瘤的面积和周长，该结果显示肿瘤的面积和周长是判断“恶性”分类的重要依据。尽管部分特征对“良性”分类有一定的正向作用，但其影响权重过小，难以对模型的判断产生实质性改变。此外，该样本的实际特征值，明显高于 LIME 所提供的判定阈值，进一步支持模型做出的“恶性”判定具有医学合理性。

从可解释性角度来看，LIME 所揭示的特征重要性与传统医学诊断的知识高度一致，说明模型不仅具备较好的分类性能，同时在决策逻辑上展现出与临床推理相符合的合理性。由此，本研究对决策依据的可视化和量化分析，不仅增强了模型的可理解性，也提升了模型在医学场景下的可信赖性和可采纳性。

5. 结论

针对乳腺癌智能诊断中存在的调参效率低、可解释性弱和临床实用性不足等问题，本文提出了一种集成贝叶斯优化与可解释性机制的 TPE-SVM 诊断模型。该方法在构建支持向量机分类器的基础上，结合类别加权方法实现代价敏感的分类，提高模型对恶性肿瘤的识别能力，然后通过 TPE 算法对模型的超参数进行全局优化。最后，本文采用 LIME 方法对预测结果进行特征层面的局部解释分析，揭示模型的决策依据，增强模型的可信度。

本研究在 WDBC 乳腺癌数据集上进行的数值实验表明，本文模型在准确率、F1-score、AUC 值等指标上均表现良好，特别是对恶性肿瘤具有超强的识别能力。LIME 可解释性分析可以准确识别出对分类判断影响最大的特征，这些结果与临床的医学知识高度符合，进一步佐证了模型的决策依据与医学知识的一致性，提升了模型诊断的用户信任度。

尽管本研究在性能和解释性方面均取得了良好效果，但仍存在一些局限性：首先，实验仅基于 UCI 的 WDBC 单一数据集进行验证，样本量有限，且缺乏多中心、多模态数据的支撑，这在一定程度上限制了模型的泛化能力；其次，本研究的实验场景仍属离线验证，尚未在真实临床环境中进行前瞻性评估，模型的操作性与医生的接受度仍需进一步检验。未来研究可从以下几方面展开：一是扩展至多中心大规模临床数据集，以提高模型的稳健性和普适性；二是探索 LIME 与其他解释性方法(如 SHAP、集成归因方法)的结合，提升解释结果的稳定性与可信度。

综上，本研究构建的 TPE-SVM 乳腺癌诊断框架在保障分类性能的同时，兼顾了模型诊断的可解释性和医学实用性，具有良好的实用价值，可以为乳腺癌诊断以及其他疾病的辅助诊断提供理论支持和技术参考，为实现人工智能在医疗领域的“可用且可理解”提供了有益探索。

基金项目

上海市软科学研究项目(25692116600)；上海高校青年教师培养资助计划(ZZ202203036)。

参考文献

[1] 徐兵河. 早筛可提高乳腺癌患者生存率[J]. 科学新生活, 2024, 27(8): 15.

[2] 张姝艳, 皮婷婷. 医疗领域中人工智能应用的可解释性困境与治理[J]. 医学与哲学, 2023, 44(3): 25-29, 35.

[3] 刘子华, 郑汉东, 刘卫勇. 基于改进动态集成选择算法的乳腺肿块辅助诊断模型[J]. 计算机应用研究, 2023, 40(1): 147-154.

[4] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018) A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, **51**, 1-42. https://doi.org/10.1145/3236009

[5] Gunning, D. and Aha, D.W. (2019) Darpa's Explainable Artificial Intelligence Program. *AI Magazine*, **40**, 44-58. https://doi.org/10.1609/aimag.v40i2.2850

[6] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[7] Barr Kumarakulasinghe, N., Blomberg, T., Liu, J., Saraiva Leao, A. and Papapetrou, P. (2020) Evaluating Local Interpretable Model-Agnostic Explanations on Clinical Machine Learning Classification Models. 2020 *IEEE 33rd International Symposium on Computer-Based Medical Systems* (*CBMS*), Rochester, 28-30 July 2020, 7-12. https://doi.org/10.1109/cbms49503.2020.00009

[8] Jalali, A., Schindler, A., Haslhofer, B. and Rauber, A. (2020) Machine Learning Interpretability Techniques for Outage Prediction: A Comparative Study. *PHM Society European Conference*, **5**, 10. https://doi.org/10.36001/phme.2020.v5i1.1244

[9] Onchis, D.M. and Gillich, G. (2021) Stable and Explainable Deep Learning Damage Prediction for Prismatic Cantilever Steel Beam. *Computers in Industry*, **125**, Article ID: 103359. https://doi.org/10.1016/j.compind.2020.103359

[10] Zhu, X., Zhang, K., Li, X., Su, F. and Tian, J. (2024) An Interpretable Machine Learning Method for Risk Stratification of Patients with Acute Coronary Syndrome. *Heliyon*, **10**, e36815. https://doi.org/10.1016/j.heliyon.2024.e36815

[11] 杨萌宇, 张雷, 曾悦. 改进粒子群算法优化的 SVM 在恶性肿瘤诊断中的应用[J]. 现代电子技术, 2020, 43(15): 110-114, 118.

[12] Kim, K.H. and Sohn, S.Y. (2020) Hybrid Neural Network with Cost-Sensitive Support Vector Machine for Class-Imbalanced Multimodal Data. *Neural Networks*, **130**, 176-184. https://doi.org/10.1016/j.neunet.2020.06.026

[13] Lotter, W., Diab, A.R., Haslam, B., Kim, J.G., Grisot, G., Wu, E., *et al*. (2021) Robust Breast Cancer Detection in Mammography and Digital Breast Tomosynthesis Using an Annotation-Efficient Deep Learning Approach. *Nature Medicine*, **27**, 244-249. https://doi.org/10.1038/s41591-020-01174-9

[14] Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M. and Mahmood, F. (2021) Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nature Biomedical Engineering*, **5**, 555-570. https://doi.org/10.1038/s41551-020-00682-w

[15] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[16] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. arXiv: 1706.03762.

[17] Witowski, J., Heacock, L., Reig, B., Kang, S.K., Lewin, A., Pysarenko, K., *et al*. (2022) Improving Breast Cancer Diagnostics with Deep Learning for MRI. *Science Translational Medicine*, **14**, eabo4802. https://doi.org/10.1126/scitranslmed.abo4802

[18] Shafi, I., Din, S., Khan, A., Díez, I.D.L.T., Casanova, R.d.J.P., Pifarre, K.T., *et al*. (2022) An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network. *Cancers*, **14**, Article 5457. https://doi.org/10.3390/cancers14215457

[19] 龚安, 吕秀明. 基于卷积神经网络的乳腺癌病理图像分类方法[J]. 计算机应用与软件, 2023, 40(6): 133-139, 198.

[20] 徐坤财, 张宁, 廖益龙, 等. 基于 CNN 和 Transformer 的两阶段乳腺癌病理图像分类方法研究[J]. 医疗卫生装备, 2024, 45(12): 1-8.

[21] 孟霖宜, 刘屿鸿. 基于 PSO 优化 SVM 算法的癌症诊断方法研究[J]. 计算机仿真, 2023, 40(8): 279-283.