Advances in Applied Mathematics 应用数学进展, 2026, 15(2), 293-301
Published Online February 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.152070

Hans汉斯

多种机器学习算法在甲状腺癌复发风险评估中的对比分析

丁光琰，李　敏*

青岛大学数学与统计学院，山东 青岛

收稿日期：2026年1月10日；录用日期：2026年2月4日；发布日期：2026年2月12日

摘　要

甲状腺癌术后复发风险的准确评估对改善患者预后及优化医疗资源配置具有重要意义。针对传统统计学方法在处理复杂非线性临床数据时的局限性，本文旨在探讨并比较多种机器学习算法在甲状腺癌复发预测中的应用价值。本文采用Borzooei和Tarokhian提供的临床数据集，包含17项特征变量，在对数据进行清洗及Label Encoding数值化处理后，按3:1比例划分为训练集与测试集，构建了逻辑回归、K近邻、决策树、支持向量机(SVM)以及随机森林、XGBoost、CatBoost共七种机器学习模型。通过受试者工作特征曲线(ROC)、曲线下面积(AUC)以及准确度、灵敏度、特异度等这一多维指标体系，全面评估各模型的分类性能。实验结果显示，七种模型均表现出优异的预测性能，AUC值均超过0.93，准确度均高于0.89。其中，集成学习算法表现最为突出：随机森林(Random Forest)以0.9904的最高AUC值展现了最优的泛化能力；XGBoost与CatBoost在整体准确度上并列第一(0.9375)，且XGBoost在特异度(0.9000)上表现最佳。特征分析进一步揭示，风险等级(Risk)、治疗反应(Response)及TNM分期是影响复发预测的核心指标。机器学习技术，特别是以随机森林和XGBoost为代表的集成学习算法，能有效提升甲状腺癌复发风险预测的准确性，该模型可作为一种客观、高效的辅助诊断工具，为临床医生制定个性化随访策略提供科学依据。

关键词

甲状腺癌，复发预测，机器学习，集成学习

Comparative Analysis of Multiple Machine Learning Algorithms for Risk Assessment of Thyroid Cancer Recurrence

Guangyan Ding, Min Li*

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

*通讯作者。

文章引用: 丁光琰, 李敏. 多种机器学习算法在甲状腺癌复发风险评估中的对比分析[J]. 应用数学进展, 2026, 15(2): 293-301. DOI: 10.12677/aam.2026.152070

Received: January 10, 2026; accepted: February 4, 2026; published: February 12, 2026

Abstract

Accurate assessment of postoperative recurrence risk in thyroid cancer is of great significance for improving patient prognosis and optimizing medical resource allocation. Addressing the limitations of traditional statistical methods in handling complex nonlinear clinical data, this paper aims to explore and compare the application value of various machine learning algorithms in predicting thyroid cancer recurrence. This study utilizes the clinical dataset provided by Borzooei and Tarokhian, comprising 17 feature variables. After data cleaning and numerical processing via Label Encoding, the data was split into training and testing sets at a 3:1 ratio. Seven machine learning models were constructed, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Random Forest, XGBoost, and CatBoost. The classification performance of each model was comprehensively evaluated using a multidimensional metric system involving Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC), accuracy, sensitivity, and specificity. Experimental results demonstrated that all seven models exhibited excellent predictive performance, with AUC values exceeding 0.93 and accuracy surpassing 0.89. Notably, ensemble learning algorithms performed the best: Random Forest demonstrated optimal generalization ability with the highest AUC of 0.9904; XGBoost and CatBoost tied for the highest overall accuracy (0.9375), with XGBoost achieving the best specificity (0.9000). Feature analysis further revealed that Risk level, Response to therapy, and TNM stage were the core predictors affecting recurrence. In conclusion, machine learning techniques, particularly ensemble learning algorithms represented by Random Forest and XGBoost, can effectively improve the accuracy of thyroid cancer recurrence risk prediction. These models can serve as objective and efficient auxiliary diagnostic tools, providing a scientific basis for clinicians to formulate personalized follow-up strategies.

Keywords

Thyroid Cancer, Recurrence Prediction, Machine Learning, Ensemble Learning

1. 引言

近年来，随着筛查手段普及，甲状腺疾病发病率逐年上升，对患者生活质量产生显著影响[1]。在疾病诊疗中，利用机器学习进行辅助预测已成为研究热点[2]，机器学习能够通过分析影像、临床及基因数据，辅助医生进行精准诊断及风险评估。

机器学习在医学诊断与决策支持领域应用广泛。司锐[3]、于帆[4]等学者指出，人工智能已深入心血管、神经及消化系统等多个医学领域。黄浩然[5]、文宏伟[6]及王新光[7]等研究证实，通过特征提取与模型优化，机器学习在心血管疾病分型、神经影像分析及肾脏疾病术前预测等方面表现优异。在甲状腺疾病研究中，机器学习同样成果显著，孙悦[8]与卢江昆[9]分别将其用于甲状腺眼病基因筛选及结节超声预测；易捷伊[10]与马明瑞[11]验证了随机森林与神经网络在结节良恶性鉴别中的高准确性；周天晗[12]与王子柯[13]则通过梯度提升与随机森林算法，有效提升了淋巴结转移预测及临床综合诊断的效能。本文基于 Borzooei 和 Tarokhian [14]发布的甲状腺癌数据集，应用多种机器学习算法构建复发预测模型，对比各

个模型性能。

2. 数据描述与预处理

本文选取 Borzooei 和 Tarokhian [14]发布的甲状腺癌数据集，对数据集中 383 名甲状腺癌患者进行研究，这些患者在同一医疗中心接受了甲状腺癌的组织病理学诊断。本文共纳入 17 个特征变量，其名称、含义及数据类型详见表 1。在所有变量中，“Age”(年龄)为唯一的定量连续变量，可直接用于模型计算，其余 16 个变量均为定性分类变量，涵盖了人口学特征(如 Gender)、既往病史(如 Hx Radiotherapy)、临床病理指标(如 Pathology、TNM 分期)以及预后情况(Response、Recurred)。

考虑到分类变量(如“Gender”、“Risk”等)在原始数据中以文本字符形式存在，无法直接输入机器学习模型，本文对其进行了数值化编码处理(Label Encoding)。例如，在性别特征中，将“女性”映射为 0，“男性”映射为 1，对癌症分期及风险等级等变量也进行了相应的离散数值转换，以确保数据格式满足算法输入要求。

Table 1. Description of dataset variables
表 1. 数据集变量特征说明

变量名称	含义	类型
Age	患者年龄	连续变量
Gender	性别	分类变量
Smoking	吸烟状况	分类变量
Hx Smoking	既往吸烟史	分类变量
Hx Radiothreapy	既往放疗史	分类变量
Thyroid Function	甲状腺功能状态	分类变量
Physical Examination	颈部体格检查	分类变量
Adenopathy	淋巴结肿大情况	分类变量
Pathology	病理学类型	分类变量
Focality	病灶数量	分类变量
Risk	复发风险分层	分类变量
T	T 分期(原发肿瘤范围)	分类变量
N	N 分期(淋巴结转移)	分类变量
M	M 分期(远处转移)	分类变量
Stage	癌症综合分期	分类变量
Response	治疗反应	分类变量
Recurred	复发状态	分类变量

3. 描述性分析

ATA 指南作为分化型甲状腺癌复发风险评估的权威标准，将患者分为低、中、高三个风险层级[15]，本数据集亦采用此分类。为了解年龄对复发状态及风险等级的具体影响，通过核密度图和箱线图分析了数据的分布特征，如图 1 所示。

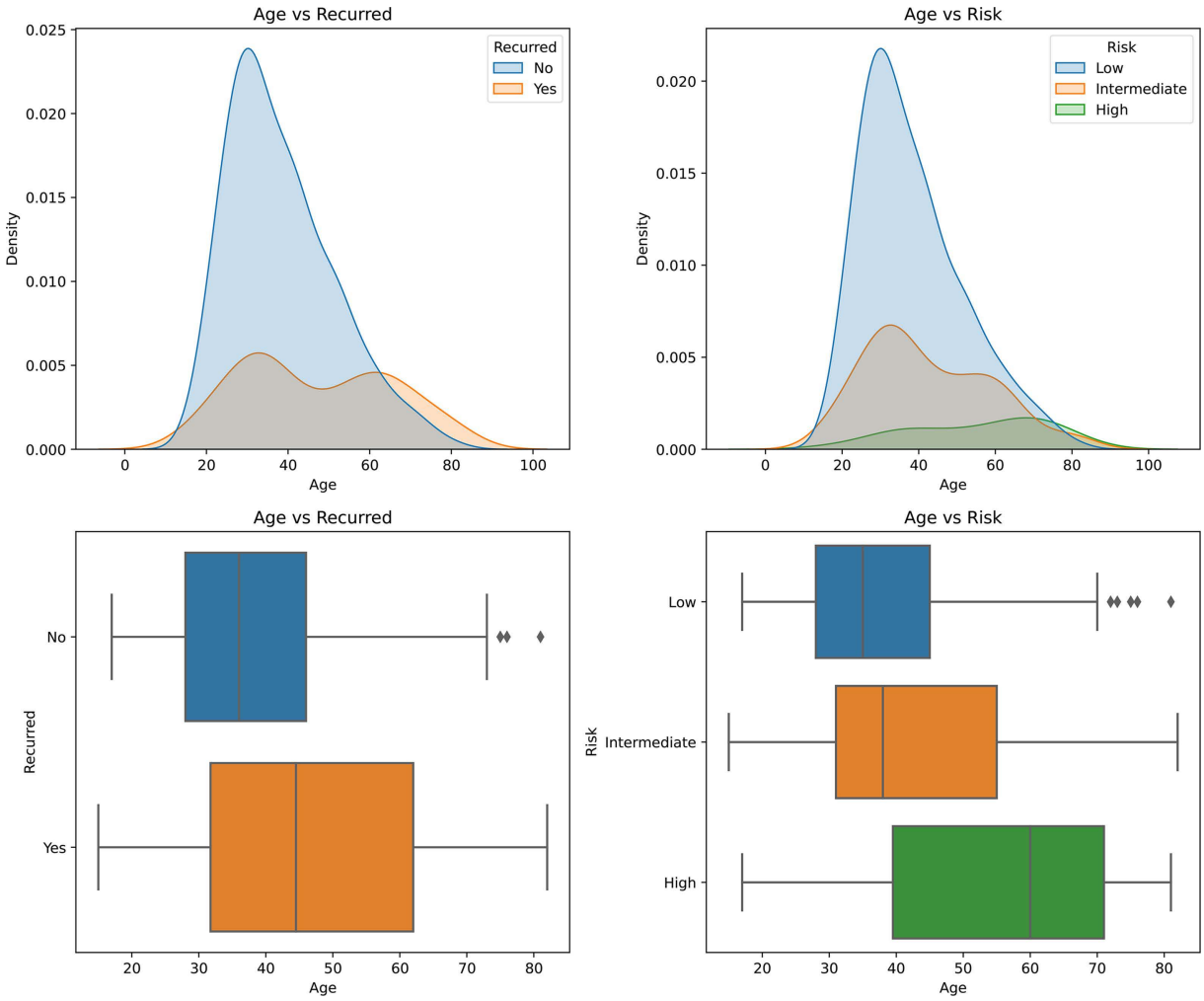

Figure 1. Age distribution characteristics across recurrence status and risk levels
图 1. 不同复发状态及风险等级下的年龄分布特征

左上图的核密度曲线显示，未复发患者集中在较年轻的年龄段，波峰窄而高。相反，复发患者的年龄分布曲线更平坦、跨度更广，虽在 30~40 岁有峰值，但在老年区间的密度显著高于未复发组，左下角的箱线图进一步印证了这一点：复发患者的年龄中位数(约 45 岁)明显高于未复发者(约 35 岁)，说明复发人群整体年龄偏大。

右上图显示，低风险患者主要集中在 30 岁左右，分布形态紧凑；中等风险患者年龄跨度大，分散在各年龄段；高风险患者则在 60~80 岁区间有较高的密度分布。右下角的箱线图表明，低风险组年龄最轻且最集中(中位数约 35 岁)，随着风险等级升高，年龄中位数上移至 60 岁左右，且高风险组包含了更多的高龄患者。上述分析表明复发患者及高风险患者的年龄分布更为广泛且整体偏高，这意味着老年群体面临更高的复发风险，建议在临床实践中给予该人群更密切的关注。

在输入机器学习模型之前，深入挖掘各临床病理特征之间的内在联系、分析特征间的多重共线性对于理解数据结构至关重要。图 2 展示了 17 个特征变量两两之间的相关性矩阵，图中颜色的深浅代表相关系数的大小，通过观察该图，可以直观地判断各临床指标之间是否存在高度耦合的情况，以及它们与复发状态的关联强度。

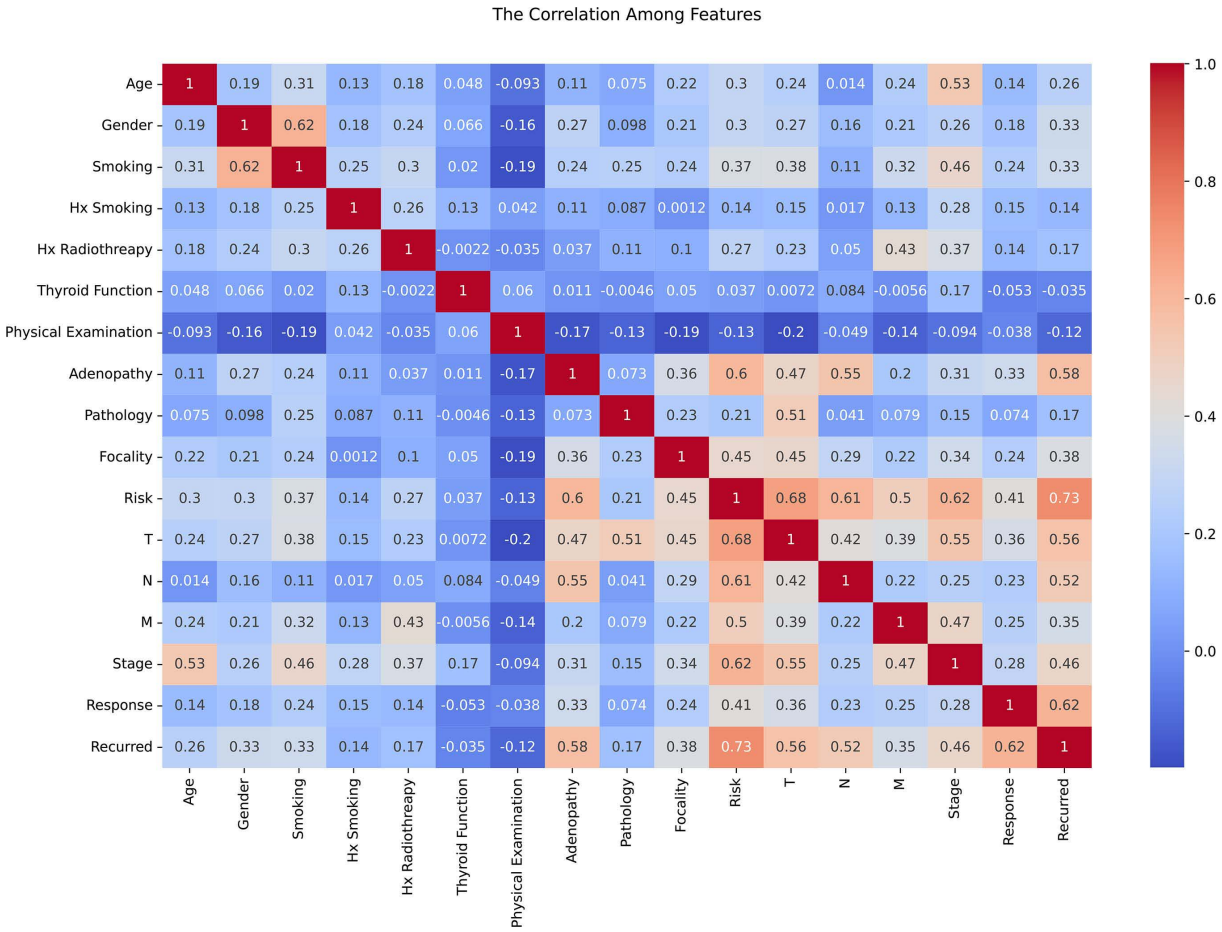

Figure 2. Pearson correlation coefficient heatmap of feature variables
图 2. 特征变量皮尔逊相关系数热力图

图 2 展示了甲状腺癌数据集中 17 个特征变量之间的皮尔逊相关系数热力图，通过颜色梯度的变化直观量化了各变量间线性关系的强弱与方向，其中深红色代表强正相关，深蓝色代表强负相关。在针对目标变量“Recurred (复发)”的分析中，Risk (风险等级)显示出最强的正相关性，相关系数高达 0.73，表明临床风险评估是预测复发的首要关键指标；其次是 Response (治疗反应) (0.62)以及 Adenopathy (淋巴结情况)、T 分期和 N 分期，其相关系数均超过 0.5，有力证实了病理分期与治疗反馈对于复发预测的重要性。相比之下，虽然人口学特征如 Age (年龄)与复发呈现正相关(0.26)，但强度不及病理指标，而 Thyroid Function (甲状腺功能)与复发几乎无线性关联(−0.035)。此外，图表揭示了特征内部显著的多重共线性，特别是 Risk 与 T、N、Stage 之间存在高度相关(系数在 0.6 至 0.7 之间)，这符合医学逻辑，即风险分级本质上是由这些分期指标综合推导而来的；同时，Gender (性别)与 Smoking (吸烟)之间也表现出显著关联(0.62)。

4. 实验结果与分析

4.1. 评价指标

本文采用随机抽样将数据集按 3:1 的比例划分为训练集与测试集，其中 75%的样本用于模型的构建与参数训练，剩余 25%的样本作为独立的测试集，用于评估模型的泛化能力。

为全面评估各模型的分类效能，本文选取了准确率(Accuracy)、灵敏度(Sensitivity)、特异度(Specificity)、

阳性预测值(Positive Predictive Value, PPV)和阴性预测值(Negative Predictive Value, NPV)作为基础评价指标。此外，ROC (Receiver Operating Characteristic)及 AUC (Area Under the Curve)来判断模型效果，ROC 曲线通过描绘不同阈值下的真阳性率与假阳性率的关系，直观展示分类器性能，模型性能越优，其 ROC 曲线越凸向左上角(即偏离对角线越远)，AUC 值介于 0 到 1 之间，数值越接近 1，表明模型的分类鉴别能力越强。

为了给出公式，首先给出混淆矩阵(Confusion Matrix)的基础元素：

TP (True Positive)：真阳性(实际为阳性，预测也为阳性)；

TN (True Negative)：真阴性(实际为阴性，预测也为阴性)；

FP (False Positive)：假阳性(实际为阴性，误报为阳性)；

FN (False Negative)：假阴性(实际为阳性，漏报为阴性)；

准确率(Accuracy)衡量模型整体预测正确的比例。

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

灵敏度(Sensitivity)衡量模型识别出所有阳性样本的能力(即“不漏诊”的能力)。

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

特异度(Specificity)衡量模型识别出所有阴性样本的能力(即“不误诊”的能力)。

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

阳性预测值(Positive Predictive Value, PPV)模型预测为阳性的样本中，真正为阳性的比例(预测准不准)。

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

阴性预测值(Negative Predictive Value, NPV)模型预测为阴性的样本中，真正为阴性的比例。

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{5}$$

ROC 曲线坐标：

纵轴(Y 轴)：真阳性率(True Positive Rate, TPR) = 灵敏度 = $\frac{\text{TP}}{\text{TP} + \text{FN}}$；

横轴(X 轴)：假阳性率(False Positive Rate, FPR) = 1 − 特异度 = $\frac{\text{FP}}{\text{TN} + \text{FP}}$；

AUC (Area Under Curve)：AUC 是 ROC 曲线下的积分面积。

4.2. 模型性能对比分析

本节对比分析了七种不同的机器学习模型，其中包括逻辑回归(Logistic Regression, logreg)、K 近邻(KNN, knn)、支持向量机(SVM, svm)、决策树(Decision Tree, DT)传统监督学习算法，以及随机森林(Random Forest, RF)、XGBoost (xgb)、CatBoost (cat)等基于集成学习策略的先进算法。通过绘制 ROC 曲线(如图 3 所示)，直观地评估了各模型在甲状腺癌复发风险预测任务中的分类效能，并计算了曲线下面积(AUC)作为衡量模型优劣的关键指标。

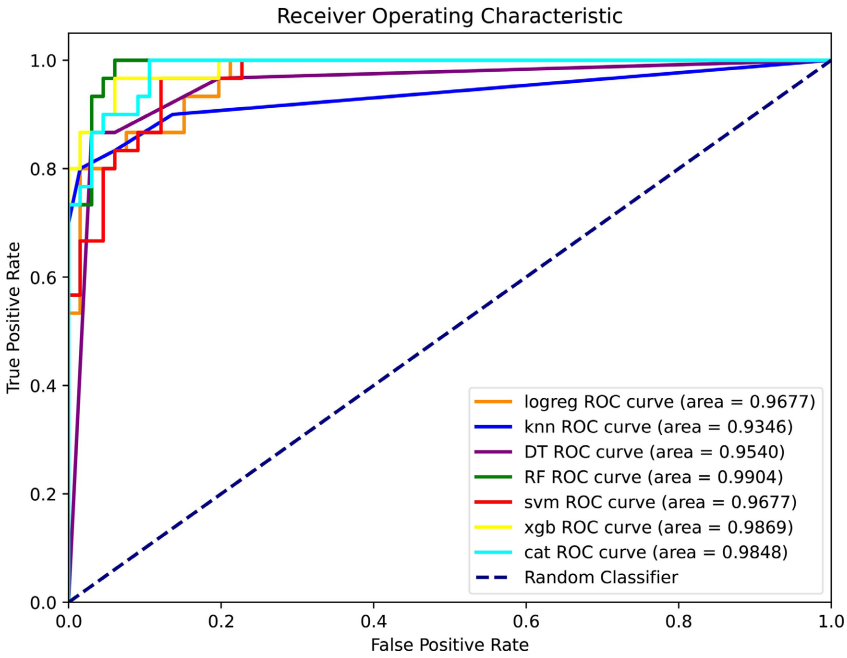

Figure 3. Comparison of ROC curves for seven machine learning models
图 3. 七种机器学习模型的 ROC 曲线对比

图中横坐标代表假阳性率(False Positive Rate)，纵坐标代表真阳性率(True Positive Rate)，深蓝色对角虚线则作为随机分类器的基准(AUC = 0.5)。从图中可以清晰地看出，所有模型的 ROC 曲线均显著凸向左上角，远离随机基准线，且 AUC 值均在 0.93 以上，表明所有参与评估的模型都具有极高的预测准确性和泛化能力。在具体模型表现上，随机森林(RF，绿色曲线)以 0.9904 的最高 AUC 值位居榜首，展现出最优的分类效能；紧随其后的是 XGBoost (黄色曲线，AUC = 0.9869)和 CatBoost (青色曲线，AUC = 0.9848)，这三者均为集成学习模型，显示出此类算法在处理该数据集时的显著优势。逻辑回归(logreg)和支持向量机(svm)也表现出色，AUC 均为 0.9677，而决策树(DT)和 K 近邻(knn)虽然略逊一筹，但 AUC 也分别达到了 0.9540 和 0.9346，有力证明了机器学习模型在复发预测中的有效性，尤其是以随机森林为代表的集成算法表现最为突出。

为了系统评估机器学习模型在甲状腺癌复发预测中的实际临床效能及增量价值，本文设计了多层次的对比实验，如图 4。首先，为了确立当前临床标准下的预测基线，我们构建了仅包含单一临床指标的 Logistic 回归模型(蓝色点线)，该模型的 AUC 值为 0.8848，其目的是量化仅依赖传统专家经验分层所能达到的预测上限，作为衡量复杂模型收益的参考标杆。其次，为了验证机器学习算法是否具备独立于专家先验知识的特征挖掘能力，我们特意剔除“Risk”分层变量，仅利用年龄、TNM 分期等原始参数训练了模型(绿色虚线)，结果显示其 AUC 仍达到 0.9818，表明算法能够成功从原始病理数据中自主捕捉非线性高危模式，具有不依赖人为预判的独立诊断价值。最终，集成了原始特征与临床分层信息的全变量模型(红色实线)实现了最优性能(AUC = 0.9869)，其与基线模型之间的性能差距展示了机器学习技术的边际收益，证明了本研究提出的模型并非简单复述现有的风险评分，而是通过整合多维数据特征，提供了优于传统临床指南的更精准的预后评估依据。

虽然 ROC 曲线和 AUC 值直观地反映了各模型的整体分类效能，但为了更全面、细致地评估模型在临床应用中的实际表现，本文进一步统计了准确度、灵敏度、特异度等具体评价指标。表 2 详细汇总了逻辑回归、随机森林、XGBoost 等七种模型在测试集上的多维性能数据，以便深入对比不同算法的优势

与侧重点。

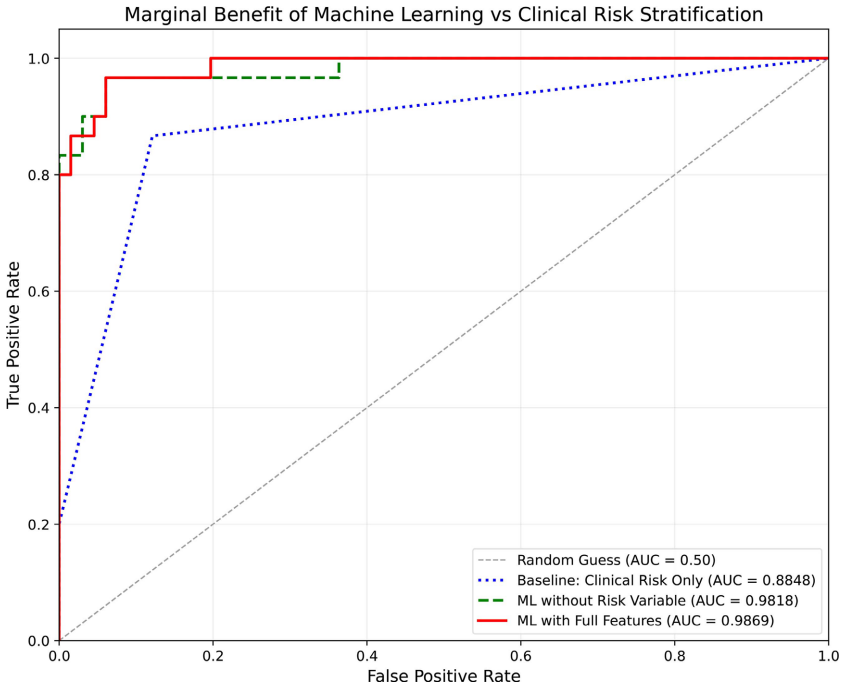

Figure 4. Comparison of ROC among the clinical baseline, the model excluding expert scores, and the full-feature machine learning model
图 4. 临床基线、剔除专家评分模型与全特征机器学习模型的 ROC 曲线比较

Table 2. Comparison of performance metrics for seven machine learning models
表 2. 七种机器学习模型性能评价指标对比

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Logistic 回归	0.9063	0.9393	0.8333	0.9254	0.8621	0.9677
K 近邻	0.9271	**0.9848**	0.8000	0.9155	**0.9600**	0.9346
决策树	**0.9375**	0.9697	0.8667	0.9412	0.9286	0.9540
随机森林	0.8958	0.9697	0.7333	0.8889	0.9167	**0.9904**
SVM	0.8958	0.9545	0.7667	0.9000	0.8846	0.9677
XGBoost	**0.9375**	0.9545	**0.9000**	**0.9545**	0.9000	0.9869
CatBoost	**0.9375**	0.9697	0.8666	0.9412	0.9286	0.9848

该表汇总了逻辑回归、K 近邻、决策树、随机森林、SVM、XGBoost 及 CatBoost 七种模型在甲状腺癌复发预测任务中的多维性能评估结果，涵盖准确度、灵敏度、特异度、PPV、NPV 及 AUC 六大关键指标。数据表明，决策树、XGBoost 与 CatBoost 在整体预测准确度上并列第一，均达到 0.9375，展现了极高的分类精度。在细分指标上，各模型呈现出不同的优势侧重：K 近邻模型虽然整体准确度居中，但录得了最高的灵敏度(Sensitivity, 0.9848)和阴性预测值(NPV, 0.9600)，提示其在减少漏诊方面具有显著优势；而 XGBoost 则在特异度(Specificity, 0.9000)和阳性预测值(PPV, 0.9545)上表现最佳，说明其误报率控制得当。值得注意的是，尽管随机森林的准确度(Accuracy, 0.8958)并非最高，但其 AUC 值高达 0.9904，位列

所有模型之首，表明其在综合区分正负样本的能力及模型的鲁棒性上具有不可替代的优势，因此集成学习算法(特别是 XGBoost 和随机森林)在各项指标的综合平衡上表现最为优异，是构建复发预测模型的理想选择。

5. 总结

本文基于 Borzooei 和 Tarokhian [14]提供的临床数据集，系统开展了数据预处理、特征关联性分析及七种机器学习模型的构建与评估工作。通过对逻辑回归、决策树、随机森林、XGBoost 等模型的对比分析，证实了利用机器学习挖掘临床病理特征预测复发的可行性，所有模型均展现出极高的分类精度。其中，集成学习算法表现最为显著，随机森林以 0.9904 的 AUC 值位居榜首，XGBoost 在特异度和准确度上表现优异，验证了该类算法在处理高维非线性医疗数据时的鲁棒性；同时，特征分析确认了风险等级、治疗反应及 TNM 分期为影响复发的最核心指标。

尽管本研究构建的集成学习模型在甲状腺癌复发预测中表现优异，但仍受限于部分客观条件。由于数据来源于单中心且样本规模有限，缺乏大规模外部队列的独立验证，模型在不同群体中的泛化能力尚待进一步考证。同时，现有特征仅涵盖结构化临床文本，尚未融合超声影像、病理切片或基因测序等多模态数据，导致部分深层生物学信息利用不足。鉴于此，后续工作拟引入深度学习技术处理非结构化数据，探索多维度分析模式，从而开发出更加精准且具备普适性的临床辅助决策系统。

参考文献

[1] 贾林梓. 认识甲状腺疾病为健康护航[J]. 健康向导, 2024, 30(2): 1.

[2] Sidey-Gibbons, J.A.M. and Sidey-Gibbons, C.J. (2019) Machine Learning in Medicine: A Practical Introduction. *BMC Medical Research Methodology*, **19**, Article No. 64. https://doi.org/10.1186/s12874-019-0681-4

[3] 司锐, 李文秀, 苏俊武. 人工智能在医学领域的应用进展[J]. 中国医药, 2021, 16(6): 957-960.

[4] 于帆, 何海洪, 周义文. 人工智能在检验医学领域的应用进展[J]. 国际检验医学杂志, 2023, 44(18): 2267-2273.

[5] 黄浩然. 基于机器学习的心血管疾病预测研究[D]: [硕士学位论文]. 武汉: 湖北大学, 2024.

[6] 文宏伟, 陆菁菁, 何晖光. 机器学习在神经精神疾病诊断及预测中的应用[J]. 协和医学杂志, 2018, 9(1): 19-24.

[7] 王新光. 机器学习在结石性肾积脓术前诊断及 PCNL 术后 SIRS 预测方面的应用研究[D]: [博士学位论文]. 武汉: 华中科技大学, 2021.

[8] 孙悦, 夏宁, 戴玮然. 基于机器学习算法探讨甲状腺相关性眼病的免疫相关基因[J]. 广西医学, 2023, 45(10): 1200-1207.

[9] 卢江昆, 胡纪杨, 龚建鸣等. 机器学习模型预测甲状腺结节良恶性分析[J]. 山西医药杂志, 2021, 50(20): 2899-2901.

[10] 易捷伊. 机器学习在甲状腺结节良恶性诊断中的辅助分析[D]: [硕士学位论文]. 昆明: 云南大学, 2018.

[11] 马明瑞, 马晓剑, 冷晓玲. 基于机器学习的微灶甲状腺乳头状癌超声智能诊断方法探析[J]. 中国医疗设备, 2019, 34(S2): 171-173.

[12] 周天晗, 吴凡, 陆凯宁, 等. 基于机器学习算法预测甲状腺乳头状癌右喉返神经后方淋巴结转移907例临床研究[J]. 中国实用外科杂志, 2021, 41(12): 1394-1399.

[13] 王子柯. 基于机器学习的甲状腺乳头状癌临床数据分析与诊断模型研究[D]: [硕士学位论文]. 柳州: 广西科技大学, 2022.

[14] Borzooei, S., Briganti, G., Golparian, M., Lechien, J.R. and Tarokhian, A. (2023) Machine Learning for Risk Stratification of Thyroid Cancer Patients: A 15-Year Cohort Study. *European Archives of Oto-Rhino-Laryngology*, **281**, 2095-2104. https://doi.org/10.1007/s00405-023-08299-w

[15] Lee, J., Lee, S.G., Kim, K., Yim, S.H., Ryu, H., Lee, C.R., *et al.* (2019) Clinical Value of Lymph Node Ratio Integration with the 8th Edition of the UICC TNM Classification and 2015 ATA Risk Stratification Systems for Recurrence Prediction in Papillary Thyroid Cancer. *Scientific Reports*, **9**, Article No. 13361. https://doi.org/10.1038/s41598-019-50069-4