·甲状腺及甲状旁腺专题论著·

CT影像特征机器学习预测模型对甲状腺乳头状癌的预测价值

朱翰林[1]　冯波[1]　张海峰[1]　张梅花[1]　田敏[2]　章彤[2]　魏培英[2]　韩志江[3]
[1]浙江省杭州市第九人民医院放射科,杭州　311225;[2]西湖大学医学院附属杭州市第一人民医院放射科,杭州　310006;[3]浙江大学医学院附属邵逸夫医院放射科,杭州　310016
通信作者:韩志江,Email:hzj1022@zju.edu.cn

【摘要】 目的　建立甲状腺乳头状癌(PTC)CT影像特征3种机器学习预测模型,运用SHAP值分析最佳模型中各CT特征在模型中的贡献度。**方法**　回顾性分析2016年1月至2021年1月西湖大学医学院附属杭州市第一人民医院门诊和住院收治,且经病理证实426例440枚PTC的CT影像特征,与467例528枚结节性甲状腺肿(NG)对比,评估咬饼征、增强后范围缩小/模糊、微钙化和形态不规则4个CT特征在2者中的分布。将PTC和NG的CT影像资料以8∶2的比例随机分为训练集和测试集,采用极端梯度提升(XGBoost)、随机森林(RF)和支持向量机(SVM)构建3个机器学习模型。通过受试者工作特征曲线下面积(AUC)、准确率、F1评分等,筛选出最佳模型。使用SHAP值解释最佳模型中各CT特征对模型贡献度。**结果**　440枚PTC和528枚NG的CT特征中,咬饼征、增强后范围缩小/模糊、微钙化和形态不规则分别为326枚和30枚(χ^2=483.05, P<0.001)、363枚和106枚(χ^2=374.45, P<0.001)、158枚和53枚(χ^2=94.24, P<0.001)、354枚和52枚(χ^2=491.34, P<0.001)。XGBoost、RF和SVM构建的机器学习模型在训练集上的AUC、准确度、F1评分范围分别为0.884~0.925、0.867~0.873、0.844~0.854,在测试集上为0.869~0.923、0.845~0.871、0.803~0.845,其中XGBoost模型在测试集上诊断效能最高。4个CT特征中,形态不规则的绝对SHAP值最高,对诊断PTC为正向贡献。**结论**　XGBoost机器学习模型诊断PTC的效能最高;CT特征中,形态不规则对诊断PTC贡献度最高且为正向贡献。

【关键词】 甲状腺结节;　甲状腺乳头状癌;　X线计算机;　体层摄影术;　SHAP;　机器学习模型
基金项目:浙江省医药卫生科技项目(2020RC091,2021RC024);杭州市医药卫生科技计划项目(A20220121);杭州市医药卫生科技项目(20211231Y058,A20220841)
DOI:10.3760/cma.j.cn115807-20231018-00112

Predictive value of machine learning models based on CT imaging features for papillary thyroid carcinoma

Zhu Hanlin[1], Feng Bo[1], Zhang Haifeng[1], Zhang Meihua[1], Tian Min[2], Zhang Tong[2], Wei Peiying[2], Han Zhijiang[3]
[1]Department of Radiology, the Ninth People's Hospital of Hangzhou, Hangzhou 311225, China; [2]Department of Radiology, Affiliated Hangzhou First People's Hospital, Westlake University School of Medicine, Hangzhou 310006, China; [3]Department of Radiology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 31006, China
Corresponding author: Han Zhijiang, Email: hzj1022@zju.edu.cn

【Abstract】 Objective　To establish three machine learning prediction models based on CT imaging characteristics of papillary thyroid carcinoma (PTC), and use SHAP (shapley additive explanations) analysis to investigate the contribution of each CT image features in the best model. **Methods**　CT imaging features in 426 cases of 440 PTCs confirmed pathologically from Jan. 2016 to Jan. 2021 at the affiliated Hangzhou First People's Hospital of Westlake University Medical School were retrospectively analyzed. compared with 467 cases of 528 nodular goiter (NG), evaluating the distribution of four CT characteristics: cookie bite sign, enhanced range of narrowing/blur (ERNB), microcalcifications, and irregular shape. We split the data into 8∶2 ratio for training and testing sets, then constructed three machine learning models using XGBoost, RF, and SVM. Based on AUC, accuracy, F1 score, and other metrics, we selected the best model. Lastly, we used SHAP values to assess each CT feature's contribution

and positive/negative effects on the model. **Results** Among 440 PTC and 528 NG nodules, CT features like cookie bite sign, ERNB, microcalcifications, and irregular shape occurred in 326 and 30 (χ^2=483.05, P<0.001), 363 and 106 (χ^2=374.45, P<0.001), 158 and 53 (χ^2=94.24, P<0.001), and 354 and 52 (χ^2=491.34, P<0.001) nodules, respectively. The machine learning models built using XGBoost, RF, and SVM had AUC, accuracy, and F1 scores ranging from 0.884~0.925, 0.867~0.873, and 0.844~0.854 respectively on the training set. On the test set, the scores ranged from 0.869~0.923, 0.845~0.871, and 0.803~0.845. Among them, the XGBoost model demonstrated the highest diagnostic performance on the test set. Among the four CT features, irregular shape had the highest absolute SHAP value, positively contributing to PTC diagnosis. **Conclusion** XGBoost model showed the highest PTC diagnostic performance. Irregular shape had the greatest positive impact on PTC diagnosis.

【Key words】 Thyroid nodules; Papillary thyroid carcinoma; Tomography, X-ray; SHAP; Machine learning model

Fund program: Medical Science Research Program of Zhejiang Province (2020RC091, 2021RC024); Medical Science Research Program of Hangzhou City (A20220121); Hangzhou Medical and Health Science and Technology Project (20211231Y058 and A20220841)

DOI: 10.3760/cma.j.cn115807-20231018-00112

甲状腺结节为最常见的内分泌疾病[1],发病率呈逐年上升趋势[1-2]。据统计,甲状腺结节患病率为20%~76%,约8%~16%为恶性[3-4]。CT扫描可显示甲状腺结节的形态、密度等特征,对甲状腺乳头状癌(papillary thyroid carcinoma, PTC)和结节性甲状腺肿(nodular goiter, NG)鉴别具有重要价值[5]。既往研究采用逻辑回归中的*OR*值(odds ratio)来评估咬饼征、微钙化、增强后范围模糊/缩小和形态不规则等CT特征与PTC的关联强度[6-7],但其无法分析这些特征在个体或全局层面上的权重。本文使用SHAP(shapley additive explanations)值[8-9]对机器学习模型中各CT特征的权重行量化和可视化,旨在明确各特征在模型中的正/负向贡献,为临床诊断PTC提供依据。

1 资料与方法

1.1 临床资料

回顾性分析2016年1月至2021年1月西湖大学医学院附属杭州市第一人民医院门诊和住院收治的经手术病理证实的甲状腺肿瘤患者临床和CT资料。本文纳入PTC组共426例,男104例,女322例,年龄(46.6±13.7)岁;NG组共467例,男87例,女380例,年龄(49.7±11.5)岁。

1.2 纳入标准和排除标准

纳入标准:①术后病理结果为PTC或NG;②所有患者均经颈部CT平扫和增强检查;③年龄>18岁。排除标准:①结节≤10 mm;②CT强化程度无法评估粗大钙化为主或孤立性粗钙化结节;③弥漫性病变或锁骨伪影遮盖,及存在运动伪影而结节显示不清者;④同时合并PTC和NG者。本研究遵守《赫尔辛基宣言》,获西湖大学医学院附属杭州市第一人民医院伦理委员会批准(批号:ZN-20230131-0014-01),并免除受试者知情同意。

1.3 检查方法

扫描采用美国GE公司Light-Speed 16排CT,扫描参数:电压120 kV,电流250 mA,准直0.625 mm×16,螺距0.938 mm,机架旋转时间0.5 s。扫描范围从颅底至主动脉弓上缘,重建层厚为3.75 mm。对比剂为德国Bayer公司的碘普罗胺注射液80 mL,碘浓度300 mg/mL,高压注射器经肘静脉团注,速率2~3 mL/s,注射后50 s扫描。

1.4 CT图像分析方法

医生A和B(工作经验分别为9和5年)在不知道病理结果的情况下,通过PACS系统对CT影像进行评估,当意见不统一时,协商达成共识。评价标准:①咬饼征[10],平扫时不规则瘤体最大径位于瘤-甲交界区或甲状腺外,正常甲状腺轮廓局部缺损,状如"咬饼";②微钙化,钙化最大径≤2 mm;③形态不规则;④增强后范围缩小/模糊,缩小为增强后瘤体最大径较平扫≤2 mm,模糊为测量平扫及增强后结节边缘和其对应甲状腺CT值,增强后差值<平扫差值[6,10]。

1.5 机器学习模型构建与SHAP值分析

将纳入影像数据以8∶2随机分为训练集和测试集。使用极端梯度提升(extreme gradient boosting, XGBoost)、随机森林(random forest, RF)和支持向量机(support vector machine, SVM)构建3个机器学习模型以诊断PTC。评估指标包括曲线下面积

(area under the curve, AUC)、准确度等,对比获得最佳模型。使用SHAP值对最佳模型进行解释,并可视化。

1.6　统计学方法

采用Python(版本3.7.1)和R语言(版本4.0.1)统计分析。Shapiro-Wilk检验连续变量的正态性,正态分布变量以平均数±标准差($\bar{x}\pm s$)表示,使用t检验分析;非正态分布的变量使用中位数和四分位数(interquartile range, IQR)表示,使用Wilcoxon检验分析。分类变量以构成百分比的形式表示,使用χ^2检验分析。P<0.05差异有统计学意义。

2　结果

2.1　甲状腺结节CT特征

咬饼征、增强后范围缩小/模糊、微钙化和形态不规则在PTC组中更常见(P值均<0.001),诊断PTC的敏感和特异度分别为74.1%和94.3%、82.5%和79.9%、35.9%和90.0%、80.5%和90.2%(表1)。

2.2　模型诊断效能比较

本研究构建基于4个CT特征的XGBoost、RF和SVM三种机器学习模型,以诊断PTC。训练集中XGBoost、RF和SVM模型的AUC及95%置信区间(confidence interval, CI)分别为0.925(95% *CI*: 0.901, 0.950)、0.925(95% *CI*: 0.901, 0.949)、0.884(95% *CI*: 0.861, 0.907),准确度分别为0.867、0.871和0.873;测试集中AUC为0.923(95% *CI*: 0.874, 0.973)、0.918(95% *CI*: 0.867, 0.969)、0.869(95% *CI*: 0.815, 0.923),准确度为0.871、0.851、0.845,其中XGBoost模型表现最佳(表2,图1)。

2.3　最佳模型SHAP值解释

SHAP值对XGBoost模型分析显示,CT特征权重由大到小依次为形态不规则、增强后范围缩小/模糊、咬饼征和微钙化,绝对平均SHAP值分别为1.29、0.88、0.56、0.27,且均为正向贡献(图2)。在SHAP值个体预测可视化方面,本文从测试集中选择2个病例利用SHAP值行个体化预测(图3~4)。

3　讨论

3.1　CT特征机器学习模型预测PTC效能分析

本研究发现咬饼症、增强后范围缩小/模糊、微钙化和形态不规则4个CT特征在PTC和NG患者间存在差异,对PTC诊断具有重要价值。增强后范围缩小/模糊诊断PTC的敏感度(82.5%)最高,但特异度(79.9%)最低,而咬饼征的敏感度较低(74.1%),特异度最高(94.3%),故依赖单一CT特征难以实现对PTC最优诊断。本文通过4个CT特征

表1　PTC和NG CT特征分析

CT特征	结节数(枚)	长径(mm)中位数[P25,P75]	发病位置[枚(%)]		瘤体数量[枚(%)]		咬饼症[枚(%)]	
			右	左	单发	多发	有	无
PTC	440	15.0[12.0, 22.0]	242(55.0)	198(45.0)	412(93.6)	28(6.4)	326(74.1)	114(25.9)
NG	528	18.0[14.0, 25.0]	259(49.1)	269(50.9)	467(88.4)	61(11.6)	30(5.7)	498(94.3)
统计值		-27.37[a]	3.40[b]		7.13[b]		483.05[b]	
P值		<0.001	0.075		0.007		<0.001	

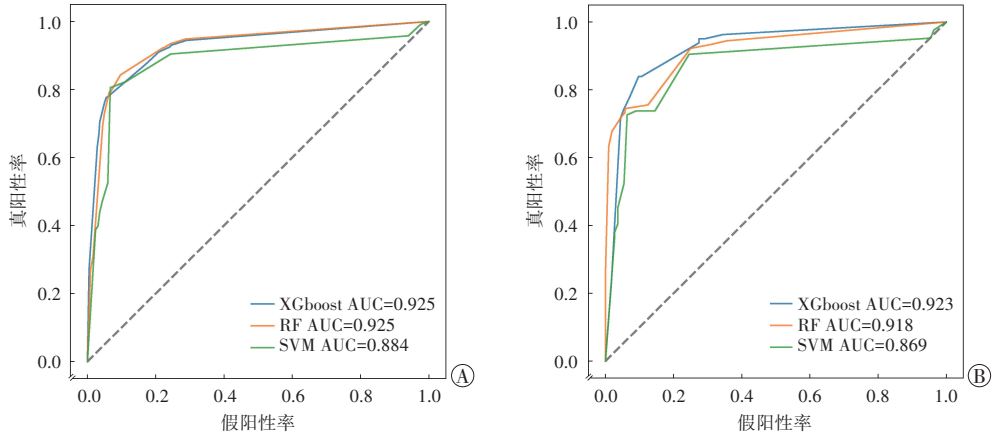
CT特征	结节数(枚)	增强后范围缩小/模糊[枚(%)]		微钙化[枚(%)]		形态不规则[枚(%)]	
		有	无	有	无	有	无
PTC	440	363(82.5)	77(17.5)	158(35.9)	282(64.1)	354(80.5)	86(19.5)
NG	528	106(20.1)	422(79.9)	53(10.0)	475(90.0)	52(9.8)	476(90.2)
统计值		374.45[b]		94.24[b]		491.34[b]	
P值		<0.001		<0.001		<0.001	

注:PTC为甲状腺乳头状癌,NG为结节性甲状腺肿;[a]为Wilcoxon检验,[b]为χ^2检验

表2　机器学习模型诊断效能比较

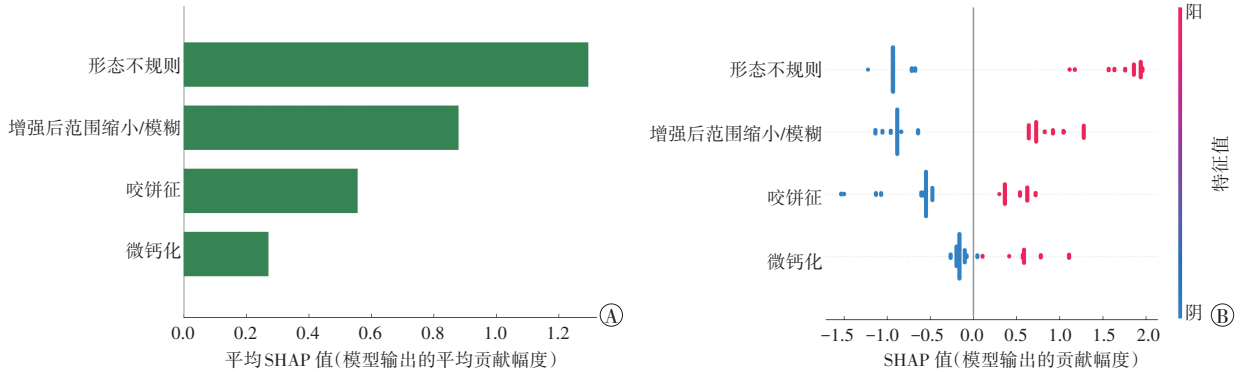
模型	训练集					测试集				
	AUC	准确度	F1评分	敏感度	特异度	AUC	准确度	F1评分	敏感度	特异度
XGBoost	0.925	0.867	0.844	0.777	0.945	0.923	0.871	0.845	0.840	0.894
RF	0.925	0.871	0.849	0.806	0.925	0.918	0.851	0.822	0.744	0.942
SVM	0.884	0.873	0.854	0.803	0.933	0.869	0.845	0.803	0.726	0.936

注:XGBoost为极端梯度提升,RF为随机森林,SVM为支持向量机,AUC为曲线下面积



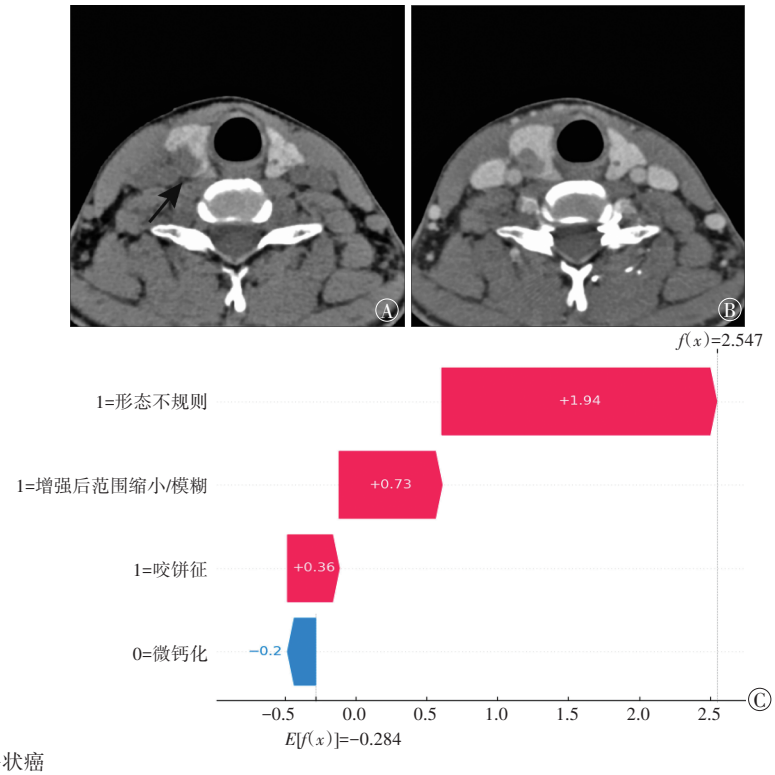
注:AUC为曲线下面积,XGBoost为极端梯度提升、RF为随机森林、SVM为支持向量机

图1　机器学习模型受试者工作特征曲线分析。A:训练集;B:测试集



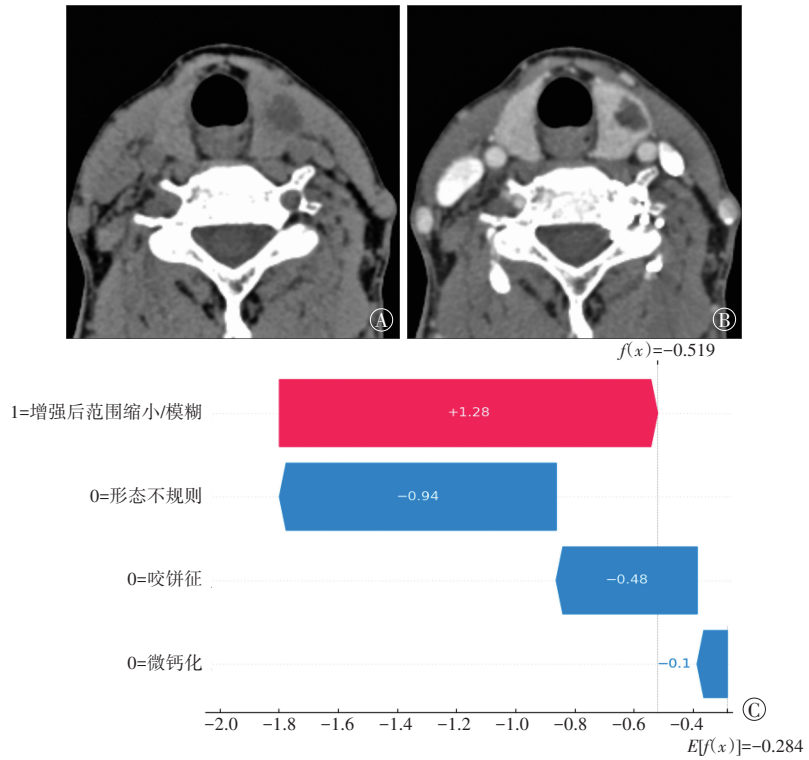
注:XGBoost为极端梯度提升

图2　XGboost模型SHAP值分析。A:SHAP值柱状图,x轴为绝对平均SHAP值,y轴为CT特征,按SHAP值降序排列;B:SHAP值散点图,每个点为单个患者的特征值,点位于右侧对模型为正贡献,位于左侧为负贡献,蓝色为CT特征阴性,红色为阳性



注:PTC为甲状腺乳头状癌

图3　真阳性个案SHAP值展示。A:CT平扫示右侧叶甲状腺不规则病灶(黑箭头);B:增强后病灶范围缩小/模糊,有咬饼征,无微钙化;C:XGboost模型的$f(x)$为2.547,即模型预测结果为PTC,最终病理回报为PTC



注：NG为结节性甲状腺肿，XGBoost为极端梯度提升

图4　真阴性个案SHAP值展示。A：CT平扫示左侧叶甲状腺低密度病灶；B：增强后病灶范围缩小，病灶形态规则，无咬饼征，无微钙化；C：XGboost模型的f(x)为-0.519，即模型预测结果为NG，最终病理回报为NG

分别构建了XGBoost、RF和SVM三种机器学习模型，XGBoost模型增加诊断PTC效能，测试集中XGBoost诊断效能最佳，三者AUC分别为0.923、0.918和0.869，准确度为0.871、0.851、0.845，F1评分为0.845、0.822、0.803。SHAP值对XGBoost模型各CT特征贡献度分析显示，形态不规则的权重最高，且为正向贡献，此外，个体水平预测过程的可视化，初步解决PTC机器学习模型决策过程“黑盒子”问题[8]。

3.2　CT特征机器学习模型在PTC诊断中的优化

目前已有学者利用CT特征构建PTC预测模型[7,11]，如吴寒冰等[7]利用形态不规则、咬饼征、微钙化等构建逻辑回归模型诊断PTC，其AUC为0.94，但该研究无测试集验证，存在一定过拟合风险，很难评估模型在新数据上的泛化能力。本文采用大样本量，并通过测试集进行验证，此方法可减小过拟合风险，同时可评估模型泛化能力。此外，本文对比3种机器学习模型的诊断能力，最佳XGBoost模型在测试集中诊断效能最高，有效区分PTC和NG，使用SHAP值对XGBoost模型行可视化，增加模型解释性。SHAP值作为一种模型事后解释方法，其是通过计算每个变量对模型输出结果的贡献度来进行解释[12]，与*OR*值相比，SHAP值提供局部和全局性解释，弥补*OR*值易高估关联强度的缺陷[8-9]。

3.3　CT特征在机器模型中的贡献度分析

PTC肿瘤组织可侵犯甲状腺及周围组织，致瘤体形态不规则[13]；NG与甲状腺组织存在纤维分隔，呈膨胀性生长，形态近似圆形[14]，使XGBoost模型易学习到与该特征相关规律，所获得绝对SHAP值可达1.29，高于增强后范围缩小/模糊＋微钙化(1.15)、咬饼征＋微钙化(0.83)。增强后范围缩小/模糊特征在XGBoost模型中排名第二，仅次于形态不规则特征重要性，其病理机制与瘤体边缘血供丰富有关[15]，CT增强扫描时瘤体边缘会出现明显强化，从而致其与周围甲状腺间的密度差异降低，边界模糊，相较于病灶形态不规则特征，增强后范围缩小/模糊在模型中的价值相对局限，故贡献度较低。咬饼征和微钙化的SHAP值较低，这与它们在数据集中分布有关，尽管微钙化在诊断PTC时具有较高特异性，但其敏感度相对较低[6]。同样，咬饼征判断易受个人认知水平和经验影响，特别是在处理较大瘤体时，致CT特征判读困难，并影响其权重稳定性。

3.4　不足与结论

本研究存在一定局限性：① 单中心回顾性研

究,难以避免选择偏差影响;②本文仅探讨>1 mm的PTC,未对≤10 mm的PTC行探索研究。总之,基于CT特征的XGBoost模型在诊断PTC方面效能最佳,形态不规则对模型存在最高正向贡献度,为诊断PTC提供临床应用价值。

利益冲突　所有作者声明不存在利益冲突

作者贡献声明　朱翰林、冯波:实验操作、论文书写;张海峰、张梅花:数据整理、统计学分析;田敏、章彤:数据整理、文献回顾;魏培英、韩志江:研究指导、论文修改、经费支持

参　考　文　献

[1] La Vecchia C, Malvezzi M, Bosetti C, et al. Thyroid cancer mortality and incidence: a global overview[J]. Int J Cancer,2015,136(9):2187-2195. DOI:10.1002/ijc.29251.

[2] 中华医学会内分泌学分会, 中华医学会外科学分会内分泌学组, 中国抗癌协会头颈肿瘤专业委员会, 等. 甲状腺结节和分化型甲状腺癌诊治指南(第二版)[J]. 中华内分泌代谢杂志, 2023,39(3):181-226. DOI:10.3760/cma.j.cn311282-20221023-00589.
Chinese Society of Endocrinology, Thyroid and Metabolism Surgery Group of the Chinese Society of Surgery, China Anti-Cancer Association, et al. Guidelines for the diagnosis and management of thyroid nodules and differentiated thyroid cancer (second edition)[J]. Chin J Endocrinol Metab, 2023, 39(3): 181-226. DOI: 10.3760/cma.j.cn311282-20221023-00589.

[3] Megwalu UC, Moon PK. Thyroid cancer incidence and mortality trends in the United States: 2000-2018[J]. Thyroid, 2022, 32(5): 560-570. DOI:10.1089/thy.2021.0662.

[4] Li Y, Teng D, Ba J, et al. Efficacy and safety of long-term universal salt iodization on thyroid disorders: epidemiological evidence from 31 provinces of mainland China[J]. Thyroid,2020,30(4):568-579. DOI:10.1089/thy.2019.0067.

[5] Traylor KS. Computed tomography and MR imaging of thyroid disease[J]. Radiol Clin North Am, 2020, 58(6): 1059-1070. DOI: 10.1016/j.rcl.2020.07.004.

[6] 王海滨, 舒艳艳, 韩志江, 等. CT在甲状腺结节良、恶性风险评估中的价值[J]. 中华医学杂志,2017,97(35):2766-2769. DOI: 10.3760/cma.j.issn.0376-2491.2017.35.012.
Wang HB, Shu YY, Han ZJ, et al. Value of CT in evaluating the risk of benign and malignant thyroid nodules[J]. Natl Med J China, 2017, 97 (35): 2766-2769. DOI: 10.3760/cma. j. issn. 0376-2491.2017.35.012.

[7] 吴寒冰, 刘翰卿, 张晓茹. 基于CT征象的模型对甲状腺乳头状癌的鉴别诊断价值[J]. 中华内分泌外科杂志,2023,17(1):52-57. DOI:10.3760/cma.j.cn.115807-20220619-00164.
Wu HB, Liu HQ, Zhang XR. Value of CT-based model in differential diagnosis of papillary thyroid carcinoma[J]. Chin J Endocr Surg, 2023, 17(1): 52-57. DOI: 10.3760/cma. j. cn. 115807-20220619-00164.

[8] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees[J]. Nat Mach Intell,2020,2(1):56-67. DOI:10.1038/s42256-019-0138-9.

[9] Zou Y, Shi Y, Sun F, et al. Extreme gradient boosting model to assess risk of central cervical lymph node metastasis in patients with papillary thyroid carcinoma: individual prediction using shapley additive explanations[J]. Comput Methods Programs Biomed,2022,225:107038. DOI:10.1016/j.cmpb.2022.107038.

[10] 韩志江, 陈文辉, 周健, 等. 微小甲状腺癌的CT特点[J]. 中华放射学杂志,2012,46(2):135-138. DOI:10.3760/cma.j.issn.1005-1201.2012.02.010.
Han ZJ, Chen WH, Zhou J, et al. CT feature of microcarcinoma of thyroid[J]. Chin J Radiol,2012,46(2):135-138. DOI:10.3760/cma.j.issn.1005-1201.2012.02.010.

[11] 张海明, 郑海格, 李振宇, 等. 结节性甲状腺肿与甲状腺乳头状癌CT征象的logistic回归分析模型的建立及其预测价值[J]. 临床放射学杂志,2021,40(7):1282-1286. DOI:10.13437/j.cnki.jcr.2021.07.008.
Zhang HM, Zheng HG, Li ZY, et al. The establishment of logistic regression analysis model of nodular goiter and papillary carcinoma and its predictive value[J]. J Clin Radiol, 2021, 40(7): 1282-1286. DOI:10.13437/j.cnki.jcr.2021.07.008.

[12] Ponce-Bobadilla AV, Schmitt V, Maier CS, et al. Practical guide to shap analysis: explaining supervised machine learning model predictions in drug development[J]. Clin Transl Sci,2024,17(11):e70056. DOI:10.1111/cts.70056.

[13] Nikiforov YE, Seethala RR, Tallini G, et al. Nomenclature revision for encapsulated follicular variant of papillary thyroid carcinoma: a paradigm shift to reduce overtreatment of indolent tumors[J]. JAMA Oncol, 2016, 2(8): 1023-1029. DOI: 10.1001/jamaoncol.2016.0386.

[14] Zhao J, Zheng X, Gao M, et al. Ultrasound features of medullary thyroid cancer as predictors of biological behavior[J]. Cancer Imaging,2021,21(1):33. DOI:10.1186/s40644-021-00402-w.

[15] Moon WJ, Baek JH, Jung SL, et al. Ultrasonography and the ultrasound-based management of thyroid nodules: consensus statement and recommendations[J]. Korean J Radiol, 2011, 12(1): 1-14. DOI:10.3348/kjr.2011.12.1.1.

(收稿日期:2023-10-18)
(本文编辑:魏琳)