

SHAP 值在 XGBoost 超声模型中诊断最大径 >1 cm 甲状腺乳头状癌的临床价值

钱佳红¹, 谢亚羽¹, 朱翰林², 王盼³, 胡春锋⁴, 韩志江⁴

摘要 **目的** 探讨 SHAP 值在 XGBoost 超声模型中诊断最大径 >1 cm 甲状腺乳头状癌(PTC)的临床价值。**方法** 选取我院经手术病理证实的 138 例 PTC 患者(145 枚结节)和 127 例结节性甲状腺肿(NG)患者(141 枚),将 286 枚结节以 8:2 的比例随机分为训练集(PTC 114 枚,NG 115 枚)和测试集(PTC 31 枚,NG 26 枚),比较训练集与测试集,以及训练集中 PTC 与 NG 超声检查结果的差异。基于训练集构建 XGBoost 超声模型;绘制受试者工作特征(ROC)曲线分析 XGBoost 超声模型对训练集和测试集中 PTC 的诊断效能,校准曲线分析模型的校准度。通过 SHAP 值解析 XGBoost 超声模型,明确各超声因素诊断 PTC 的权重。**结果** 训练集与测试集结节实性占比比较差异有统计学意义($P<0.05$),边缘模糊/不规则/腺外侵犯、纵横比(A/T)>1、超声灰阶比值(UGSR)<0.83、微钙化占比比较差异均无统计学意义。训练集中 PTC 与 NG A/T>1、UGSR<0.83、微钙化、边缘模糊/不规则/腺外侵犯、实性占比比较差异均有统计学意义(均 $P<0.001$)。ROC 曲线分析显示,XGBoost 超声模型诊断训练集和测试集中 PTC 的曲线下面积分别为 0.941(95% 可信区间:0.895~0.987)和 0.921(95% 可信区间:0.846~0.996),准确率、灵敏度、特异度分别为 88.0%、92.9%、83.2% 和 86.0%、93.5%、77.5%。校准曲线分析显示,XGBoost 超声模型诊断训练集和测试集中 PTC 的校准度均较好。通过 SHAP 值解析 XGBoost 超声模型,结果显示各超声因素对诊断 PTC 均为正向贡献,其权重从高到低依次为:UGSR<0.83、实性、边缘模糊/不规则/腺外侵犯、微钙化和 A/T>1。**结论** 利用 SHAP 值对 XGBoost 超声模型进行解析可以实现各超声因素诊断效能的量化,在诊断最大径 >1 cm PTC 中具有较好的临床应用价值。

关键词 超声检查;甲状腺乳头状癌;超声灰阶比值;SHAP 值;XGBoost
[中图分类号]R445.1;R736.1 [文献标识码]A

Clinical value of SHAP value in the diagnosis of papillary thyroid carcinoma with a maximum diameter>1 cm by XGBoost ultrasound model

QIAN Jiahong¹, XIE Yayu¹, ZHU Hanlin², WANG Pan³, HU Chunfeng⁴, HAN Zhijiang⁴

1.Department of Ultrasound, the Third People's Hospital of Xiaoshan, Hangzhou 311251, China. 2.Department of Radiology, 3.Department of Ultrasound, the Ninth People's Hospital of Hangzhou, Hangzhou 311225, China. 4.Department of Radiology, Hangzhou First People's Hospital Affiliated to West Lake University, Hangzhou 310006, China

ABSTRACT **Objective** To evaluate the clinical value of SHAP value in XGBoost-based ultrasound model for diagnosing papillary thyroid carcinoma (PTC) with a maximum diameter>1 cm. **Method** A total of 138 PTC patients (145 nodules) and 127 nodular goiter(NG) patients(141 nodules) confirmed by surgical pathology in out hospital were enrolled. The 286 nodules were randomly divided into a training set(114 PTC, 115 NG) and a testing set(31 PTC, 26 NG) at a ratio of 8:2. Ultrasound features were compared between the training and testing sets, as well as between PTC and NG nodules in the training set. A XGBoost ultrasound model was constructed using the training set. Receiver operating characteristic(ROC) curve was drawn to assess the diagnostic performance, and calibration curve evaluated the calibration of the model. SHAP value was used to interpret the model and quantify the contribution of each ultrasound feature for PTC diagnosis. **Results** There was significant difference in solid composition proportion between the training and testing sets ($P<0.05$), while there were no significant differences in the proportion of ill-defined/irregular/extrathyroidal extension margins, aspect ratio(A/T)>1, ultrasound grayscale

作者单位:1. 杭州市萧山区第三人民医院超声科,浙江 杭州 311251;2. 杭州市第九人民医院放射科,3. 超声科,浙江 杭州 311225;4. 西湖大学附属杭州市第一人民医院放射科,浙江 杭州 310006

通讯作者:韩志江, Email: hzj1022@zju.edu.cn

ratio(UGSR)<0.83, and microcalcifications. In the training set, PTC nodules exhibited significantly higher proportion of A/T>1, UGSR<0.83, microcalcifications, ill-defined/irregular/extrathyroidal extension margins, and solid composition compared to NG nodules (all $P<0.001$). ROC curve analysis showed the area under the curve (AUC) of XGBoost model in the training and testing sets were 0.941 (95%CI: 0.895~0.987) and 0.921 (95%CI: 0.846~0.996), respectively, with accuracies of 88.0% and 86.0%, sensitivities of 92.9% and 93.5%, and specificities of 87.2% and 77.5%, respectively. Calibration curve indicated good model fit for both sets. SHAP interpretation revealed that all ultrasound features positively contributed to PTC diagnosis, ranked by weight as follows: UGSR<0.83, solid composition, ill-defined/irregular/extrathyroidal extension margins, microcalcifications, and A/T>1.

Conclusion SHAP-based interpretation of the XGBoost ultrasound model quantifies the diagnostic impact of ultrasound features, demonstrating good clinical value for diagnosing PTC with a maximum diameter>1 cm.

KEY WORDS Ultrasonography; Papillary thyroid carcinoma; Ultrasound grayscale ratio; SHAP values; XGBoost

甲状腺乳头状癌(papillary thyroid carcinoma, PTC)具有相对惰性的生物学行为,患者首诊时淋巴结转移率约为20%~90%,其中最大径>1 cm的PTC患者淋巴结转移率达78.4%^[1],准确对其进行定性诊断具有重要的临床意义。目前,甲状腺结节超声诊断指南已有多个版本^[2],如美国放射学会发布的甲状腺影像报告与数据系统(American College of Radiology Thyroid Imaging Reporting and Data Systems, ACR-TIRADS)、韩国放射学会和甲状腺放射学会联合发布的甲状腺影像报告与数据系统(Korean Thyroid Imaging Reporting and Data Systems, K-TIRADS)及中华医学会超声医学分会发布的甲状腺影像报告与数据系统(Chinese Thyroid Imaging Reporting and Data Systems, C-TIRADS),尽管这些指南通过结构化评估提高了诊断标准化程度,然而不同版本对超声征象的解读规则和权重分配存在显著差异。C-TIRADS虽更适用于我国人群^[3],但对结节回声性质的判断存在较大的主观差异,且边缘模糊/不规则/腺外侵犯、内部极低回声、实性、纵横比(A/T)>1和微钙化这5项超声征象在甲状腺良恶性结节诊断中的贡献各不相同,但C-TIRADS赋予其相等的分值,故临床应用存在局限。近年来,可解释性机器学习技术为弥补上述不足提供了新思路。基于博弈论的SHAP值能够量化各特征变量对模型预测结果的贡献,而XGBoost算法凭借其高效的集成学习机制,在医学影像特征分析中表现出卓越性能。本研究通过构建XGBoost超声模型,以SHAP值解析超声因素对最大径>1 cm PTC的判别贡献度,旨在为优化C-TIRADS的权重分配提供参考依据。

资料与方法

一、研究对象

选取2022年6月至2023年11月西湖大学附属杭州市第一人民医院经病理确诊的PTC患者138例

(145枚结节)和结节性甲状腺肿患者127例(141枚结节)。PTC中男34例,女104例,年龄16~78岁,平均(43.1±12.9)岁,结节最大径11~65 mm,中位数17.0(13.0, 21.0)mm; NG中男26例,女101例,年龄21~77岁,平均(54.7±11.3)岁,结节最大径11~75 mm,中位数23.0(16.0, 35.5)mm。PTC与NG患者一般资料比较差异均无统计学意义。排除标准:①结节最大径≤1 cm; ②结节以囊性成分或钙化为主。将286枚结节以8:2的比例随机分为训练集(PTC 114枚, NG 115枚)和测试集(PTC 31枚, NG 26枚)。本研究经我院医学伦理委员会批准(批准号:ZN-20230828-0188-01),所有患者均知情同意。

二、仪器与方法

1. 超声检查:使用GE Logiq E11彩色多普勒超声诊断仪,探头频率4~13 MHz。患者取仰卧位,充分暴露甲状腺区域,依次扫查甲状腺左右叶及峡部,记录结节数量、最大径、A/T、边缘、内部回声及钙化情况。通过RADinfo阅片系统,在同一增益水平测量甲状腺结节及周围正常甲状腺组织的灰阶值,测量结节感兴趣区(ROI)时需避开钙化和囊性区域;测量周围正常甲状腺组织的灰阶值时,其ROI应与结节的ROI一致。计算超声灰阶比值(UGSR),公式为:UGSR=同一增益水平甲状腺结节灰阶值/甲状腺周围正常组织灰阶值。

2. XGBoost超声模型的构建及应用:比较训练集中PTC患者与NG患者超声检查结果的差异,采用XGBoost算法基于差异有统计学意义的超声因素构建XGBoost超声模型;通过SHAP值解析XGBoost超声模型,明确各超声因素诊断PTC的权重。

三、统计学处理

应用SPSS 25.0统计软件和Python语言(3.71版本),计数资料以频数或率表示,采用 χ^2 检验。基于受试者工作特征(ROC)曲线获取训练集中UGSR诊断PTC的截断值,将其转为二分类变量进行统计分析;绘

制 ROC 曲线分析 XGBoost 超声模型对训练集和测试集中 PTC 的诊断效能;校准曲线评估模型的校准度。 $P<0.05$ 为差异有统计学意义。

结 果

一、UGSR 诊断训练集中 PTC 的截断值及效能分析
ROC 曲线分析显示,UGSR 诊断训练集中 PTC 的曲线下面积(AUC)为 0.814,对应的截断值为 0.83,诊断灵敏度和特异度分别为 91.7% 和 63.5%。本研究将 $UGSR<0.83$ 赋值为 1, $UGSR\geq 0.83$ 赋值为 0,进行分类变量比较。见图 1。

二、训练集与测试集超声检查结果比较
训练集与测试集结节实性占比比较差异有统计学意义($P<0.05$),边缘模糊/不规则/腺外侵犯、 $A/T>1$ 、 $UGSR<0.83$ 、微钙化占比比较差异均无统计学意义。见表 1。

三、训练集中 PTC 与 NG 超声检查结果比较
训练集中 PTC 与 NG $A/T>1$ 、 $UGSR<0.83$ 、微钙化、

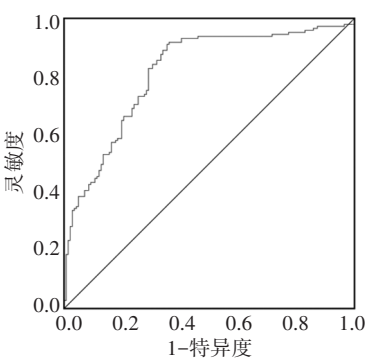


图1 UGSR 诊断训练集中 PTC 的 ROC 曲线图

边缘模糊/不规则/腺外侵犯、实性占比比较差异均有统计学意义(均 $P<0.001$)。见表 2 和图 2。

四、XGBoost 超声模型的构建及诊断效能分析
纳入 $A/T>1$ 、 $UGSR<0.83$ 、微钙化、边缘模糊/不规则/腺外侵犯、实性构建 XGBoost 超声模型。ROC 曲线分析显示,XGBoost 超声模型诊断训练集和测试集中 PTC 的曲线下面积(AUC)分别为 0.941(95% 可信区间: 0.895~0.987)和 0.921(95% 可信区间: 0.846~0.996),准

表1 训练集与测试集超声检查结果比较 枚

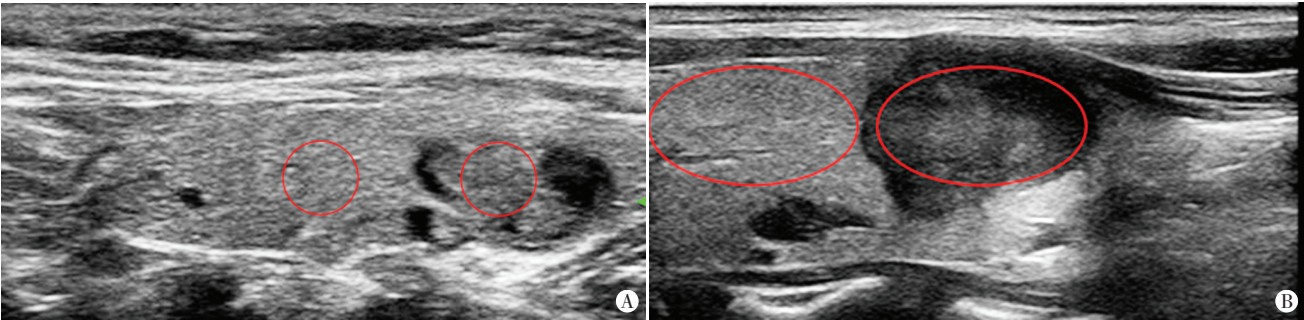
分类	A/T>1		UGSR<0.83		微钙化		边缘模糊/不规则/腺外侵犯		实性	
	是	否	是	否	是	否	是	否	是	否
训练集(229)	42	187	144	85	49	180	47	182	161	68
测试集(57)	11	46	39	18	16	41	18	39	48	9
χ^2 值	0.03		0.61		1.37		3.18		4.49	
P值	0.87		0.44		0.50		0.08		0.03	

A/T:纵横比;UGSR:超声灰阶比值

表2 PTC 与 NG 结节超声检查结果比较 枚

病理结果	A/T>1		UGSR<0.83		微钙化		边缘模糊/不规则/腺外侵犯		实性	
	是	否	是	否	是	否	是	否	是	否
PTC(145)	45	100	132	13	58	87	58	87	141	4
NG(141)	8	133	51	90	7	134	7	134	68	73
χ^2 值	30.454		93.378		50.588		49.965		87.290	
P值	<0.001		<0.001		<0.001		<0.001		<0.001	

A/T:纵横比;UGSR:超声灰阶比值



A:NG患者(女,52岁),甲状腺左侧叶囊实性等回声结节,边缘光整,UGSR为1.0;B:PTC患者(女,51岁),甲状腺右侧叶实性低回声结节,边缘模糊,UGSR为0.6

图2 两组 UGSR 检测图(红圈示测量同一增益水平 ROI 的灰阶值)

确率、灵敏度、特异度分别为 88.0%、92.9%、83.2% 和 86.0%、93.5%、77.5%。见图 3。校准曲线分析显示, XGBoost 超声模型诊断训练集和测试集中 PTC 的预测曲线与理想曲线贴合, 校准度均较好。见图 4。

五、XGBoost 超声模型的 SHAP 值解析

各超声因素对诊断 PTC 均为正向贡献, 其权重从高到低依次为: UGSR<0.83、实性、边缘模糊/不规则/腺外侵犯、微钙化和 A/T>1, 绝对 SHAP 值分别约为 1.31、1.12、0.72、0.57、0.25。见图 5。

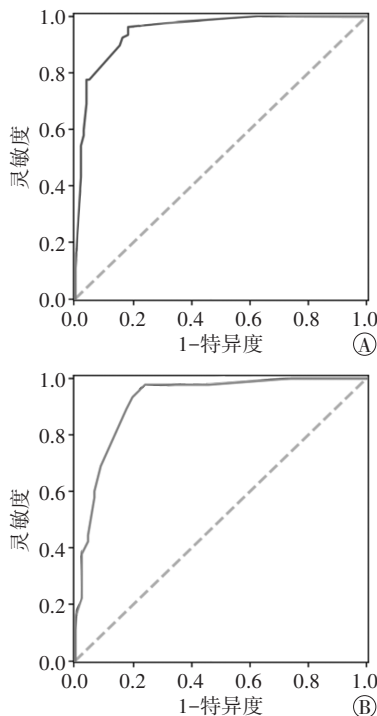


图 3 XGBoost 超声模型诊断训练集(A)和测试集(B)中 PTC 的 ROC 曲线图

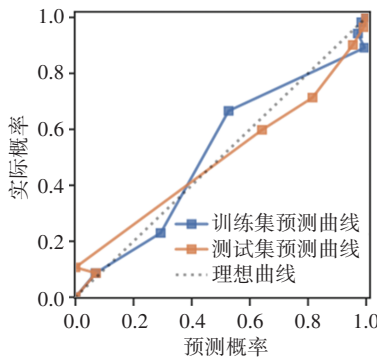


图 4 XGBoost 超声模型诊断训练集和测试集中 PTC 的校准曲线图

讨 论

目前, 不同版本甲状腺结节诊断指南中的超声征象评估标准存在差异, 且对其权重的定义不够明确^[2-4]。基于最大径>1 cm PTC 相较于微小结节具有更高的淋巴结转移率^[1]这一特征, 本研究区别于既往聚

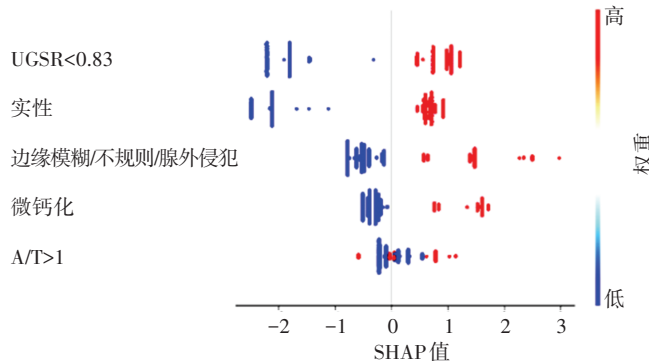


图 5 XGBoost 超声模型分析图及 SHAP 散点图(散点为单个样本, 红蓝代表是否具有该特征, 横轴的右边代表正向贡献)

焦甲状腺微小癌的研究^[5-7], 旨在系统探讨超声因素(回声、微钙化、结构、形态及边缘)在鉴别最大径>1 cm PTC 与 NG 中的权重分配机制。本团队前期^[5-7]研究已表明 UGSR 对甲状腺微小癌具有重要的鉴别诊断价值, 但在最大径>1 cm PTC 中, 其诊断价值尚未明确, 因此需要引入新的定量指标分析各超声因素权重。本研究通过构建 XGBoost 超声模型, 并结合 SHAP 值^[8]解析多特征交互效应, 实现超声诊断最大径>1 cm PTC 的可解释量化。

传统依靠颈前带状肌与甲状腺回声对比来划分甲状腺病变回声(如无回声、极低回声、低回声、等回声、高回声五类)的方式受观察者主观判断影响大, 从而在一定程度上降低了回声在 PTC 诊断中的作用。本研究将回声强度以 UGSR 表示, 可有效降低不同超声仪器、操作人员、增益、动态设置等因素导致回声判断的不稳定性。本研究 ROC 曲线分析显示, UGSR 诊断 PTC 的 AUC 为 0.814, 对应的截断值为 0.83, 诊断灵敏度和特异度分别为 91.7% 和 63.5%, 提示其具有一定的诊断潜力, 为临床提供了新的参考依据。此外, 本研究中 UGSR 诊断最大径>1 cm PTC 的截断值(0.83)明显高于前期研究^[6]报道(0.68~0.70), 分析原因为随着结节的增大, 可能导致更多的结构异质性, 如血管增生在超声图像上表现为高回声或等回声, 使 UGSR 的截断值偏高。本研究结果显示, 训练集中 PTC 与 NG 在结节 A/T>1、UGSR<0.83、微钙化、边缘模糊/不规则/腺外侵犯、实性占比方面比较差异均有统计学意义(均 $P<0.001$), 反映了 PTC 的恶性生物学行为。与既往研究^[4]结果一致。分析原因可能为 PTC 组织成分排列紧凑, 纤维化明显, 存在砂粒体, 且呈浸润性生长, 与 NG 的良性生物学行为有明显差异。

AUC 是评估模型性能的重要指标, 既往 Zheng 等^[9]和 Zhang 等^[10]基于甲状腺结节的结构、回声、形态、边缘、强回声灶建立 Logistic 回归模型, 其诊断甲状

腺结节良恶性的 AUC 为 0.685~0.924。本研究基于 C-TIRADS 选取了 A/T>1、UGSR<0.83、微钙化、边缘模糊/不规则/腺外侵犯、实性共 5 个超声因素构建模型, ROC 曲线分析显示 XGBoost 超声模型诊断训练集和测试集中 PTC 的 AUC 分别为 0.941 和 0.921, 高于上述研究, 分析与以下因素有关: ①既往研究基于 TIRADS 选取超声因素, 而本研究除 UGSR 外均依据 C-TIRADS 选取超声因素, 更适合我国患者; ②本研究用 UGSR 代替传统回声, 减少了观察者主观判断的不稳定性, 提高了模型的诊断效能; ③既往研究均未设置独立测试集对模型进行验证, 导致模型的稳健性和泛化能力无法得到充分检验, 存在过拟合风险。本研究设置独立测试集, ROC 曲线分析和校准曲线分析均证实 XGBoost 超声模型具有较好的稳健性和泛化能力。

现有研究^[11-12]大多依赖比值比评估单一特征或线性模型的全局关联, 而 SHAP 值能够评估机器学习模型中多特征间的交互或非线性关系, 提供更全面的诊断信息。本研究结果显示, 各超声因素对诊断 PTC 均为正向贡献, 其权重从高到低依次为: UGSR<0.83、实性、边缘模糊/不规则/腺外侵犯、微钙化和 A/T>1。其中“UGSR<0.83”的 SHAP 值最高, 为 1.31, 其机制可能与 PTC 分化程度低、间质成分少有关, 随着结节增大, 血管和纤维组织增生, 导致回声出现变化^[13]; “实性”的 SHAP 值为 1.12, 其机制一方面为 PTC 囊变较 NG 少, 另一方面为本研究纳入样本排除了以囊性成分为主的结节, 其中大多为 NG, 导致其权重出现一定程度降低; “边缘模糊/不规则/腺外侵犯”的 SHAP 值为 0.72, 其机制与瘤体浸润性生长方式有关; “微钙化”的 SHAP 值为 0.57, 其机制为钙化是 PTC 的重要征象, 主要部分为砂粒体, 组织来源可能是发生贫血性梗死的乳头结构, 也可能是间质成分中的纤维胶原化^[14], 但操作者易将微小囊壁、浓缩胶体等误判为微钙化, 一定程度上会降低其在模型中的权重; “A/T>1”的 SHAP 值仅为 0.25, 其机制一方面与最大径>1 cm PTC 在前后方向生长时易受到甲状腺被膜的限制而向左右/上下方向生长有关, 另一方面可能与结节内部纤维胶原化的各乳头结构相互牵拉有关^[14], 从而降低了该因素在模型中的权重。在实际应用中, 临床医师可以利用 SHAP 值在制定诊断标准时区分各因素的权重。

综上所述, 利用 SHAP 值对 XGBoost 超声模型进行解析可以实现各超声因素诊断效能的量化, 在诊断最大径>1 cm PTC 中具有较好的临床应用价值。但本研究为回顾性研究, 样本量有限, 所得结果有待今后行大样本、前瞻性、多中心研究深入探讨。

参考文献

- [1] Sun H, Zhao X, Wang X, et al. Correlation analysis of risk factors for cervical lymphatic metastasis in papillary thyroid carcinoma[J]. *Diagn Pathol*, 2024, 19(1): 13.
- [2] Zhu H, Yang Y, Wu S, et al. Diagnostic performance of US-based FNAB criteria of the 2020 Chinese guideline for malignant thyroid nodules: comparison with the 2017 American College of Radiology guideline, the 2015 American Thyroid Association guideline, and the 2016 Korean Thyroid Association Guideline[J]. *Quant Imaging Med Surg*, 2021, 11(8): 3604-3618.
- [3] Qi Q, Zhou A, Guo S, et al. Explore the diagnostic efficiency of Chinese thyroid imaging reporting and data systems by comparing with the other four systems (ACR TI-RADS, Kwak-TIRADS, KSThR-TIRADS, and EU-TIRADS): a single-center study[J]. *Front Endocrinol*, 2021, 12(10): 706978.
- [4] 中华医学会超声医学分会浅表器官和血管学组, 中国甲状腺与乳腺超声人工智能联盟. 2020 甲状腺结节超声恶性危险分层中国指南: C-TIRADS[J]. *中华超声影像学杂志*, 2021, 30(3): 185-200.
- [5] Han Z, Feng N, Lu Y, et al. A control study on the value of the ultrasound grayscale ratio for the differential diagnosis of thyroid micropapillary carcinoma and micronodular goiter in two medical centers[J]. *Front Oncol*, 2021, 10(1): 625238.
- [6] Feng N, Wei P, Kong X, et al. The value of ultrasound grayscale ratio in the diagnosis of papillary thyroid microcarcinomas and benign micronodules in patients with Hashimoto's thyroiditis: a two-center controlled study[J]. *Front Endocrinol*, 2022, 13(8): 949847.
- [7] Gong Y, Yao X, Yu L, et al. Ultrasound grayscale ratio: a reliable parameter for differentiating between papillary thyroid microcarcinoma and micronodular goiter[J]. *BMC Endocr Disord*, 2022, 22(1): 75.
- [8] Li X, Zhou Y, Dvornek NC, et al. Efficient Shapley explanation for features importance estimation under uncertainty[J]. *Med Image Comput Comput Assist Interv*, 2020, 12261(9): 792-801.
- [9] Zheng Y, Xu S, Zheng Z, et al. Ultrasonic classification of multicategory thyroid nodules based on Logistic regression[J]. *Ultrasound Q*, 2020, 36(2): 146-157.
- [10] Zhang X, Ze Y, Sang J, et al. Risk factors and diagnostic prediction models for papillary thyroid carcinoma[J]. *Front Endocrinol*, 2022, 13(9): 938008.
- [11] 李子凤, 王金萍, 朱琳, 等. 常规超声联合剪切波弹性成像多参数权重法鉴别甲状腺结节的临床价值[J]. *临床超声医学杂志*, 2023, 25(9): 696-700.
- [12] 留碧丽, 夏文霞, 李敏, 等. 超声检查及引导穿刺活检在甲状腺微小结节良恶性中的鉴别诊断价值[J]. *医学影像学杂志*, 2021, 31(11): 1864-1867.
- [13] Seyrek NC, Baser H, Topaloglu O, et al. Ultrasonographical clinical and histopathological features of 1264 nodules with papillary thyroid carcinoma and microcarcinoma based on tumor size[J]. *Arch Endocrinol Metab*, 2021, 64(5): 533-541.
- [14] 虞红珍, 徐天宇, 程岚卿, 等. 甲状腺结节超声 C-TIRADS 5 征象的组织病理结构影响因素分析[J]. *临床与实验病理学杂志*, 2023, 39(4): 433-438.

(收稿日期: 2024-07-02)