doi: 10.3969/j.issn.1672-4933.2026.06.014

基于GRBAS-G维度标注与支持向量机的病理嗓音自动检测初步研究

Automatic Detection of Pathological Voice Using a Support Vector Machine Trained on GRBAS Grade (G) Ratings: A Preliminary Study

杨紫良[1] 祝国隽[2] 张三国[2] 刘阳[1] 黄冬雁[1]

YANG Zi-liang, ZHU Guo-juan, ZHANG San-guo, LIU Yang, HUANG Dong-yan

【摘要】目的 构建以听感知评估GRBAS量表总体音质(grade,G)维度评分为监督标签的病理嗓音自动检测模型,为病理性嗓音筛查及嗓音质量客观评估提供辅助工具。**方法** 回顾性选取2020年1月至2022年3月解放军总医院第六医学中心病理嗓音数据库中662例语音样本,由5名专业听评人员完成标准化GRBAS听感知评估,采用Fleiss Kappa系数评价各GRBAS维度评估者间一致性,选取至少3名评估者G评分一致的443例样本构建建模数据集。以G0为正常嗓音组,G1～G3为病理嗓音组,提取梅尔频率倒谱系数(Mel-frequency cepstral coefficients,MFCC)及基于计盒维数(Db10)小波包分解的多频带非线性动力学特征,包括Hurst指数、2-Rényi熵和计盒维数Db,采用多项式核支持向量机(support vector machine,SVM)构建正常/病理嗓音二分类模型,通过分层十折交叉验证评估模型性能。**结果** G维度的Fleiss Kappa值为0.51,高于R、B、A、S维度的0.25～0.28。最优特征组合为MFCC+Hurst+2-Rényi+Db,模型平均准确率为84.4%,最高准确率为91.0%,敏感度为83.6%,特异度为85.1%,受试者工作特征曲线下面积(area under the receiver operating characteristic curve,AUC)为0.89,95%置信区间为0.86～0.92。**结论** 融合线性频谱特征与多频带非线性动力学特征的机器学习模型可较好区分正常嗓音与病理嗓音。以GRBAS-G维度进行监督标注的病理嗓音自动检测具有可行性,可为临床嗓音异常的无创筛查与客观评估提供辅助工具。

【关键词】病理嗓音;听感知评估;机器学习;梅尔频率倒谱系数;非线性动力学特征;支持向量机

【中图分类号】R767.52 **【文献标识码】**A **【文章编号】**1672-4933(2026)06-0639-05

【Abstract】 ***Objective*** To develop an automatic machine-learning model for pathological voice detection based on the G dimension of the GRBAS scale, providing an objective auxiliary tool for clinical voice assessment. ***Methods*** A total of 662 voice samples were retrospectively selected from a pathological voice database at a single center between January 2020 and March 2022. Five trained raters performed GRBAS auditory-perceptual assessments, and the interrater agreement of each GRBAS dimension was evaluated using the Fleiss kappa coefficient. A modeling dataset was constructed using 443 samples in which at least three raters agreed on the G score. G0 was defined as normal voice, whereas G1-G3 were defined as pathological voice. Mel-frequency cepstral coefficients (MFCC) and multi-band nonlinear dynamic features based on Db10 wavelet packet decomposition, including the Hurst exponent, 2-Rényi entropy, and box-counting dimension Db, were extracted. A polynomial-kernel support vector machine (SVM) was developed for binary classification, and model performance was evaluated using stratified 10-fold cross-validation. ***Results*** The Fleiss' kappa of the G dimension was 0.51, higher than those of the R, B, A, and S dimensions (0.25-0.28). The optimal feature combination was MFCC + Hurst + 2-Rényi + Db. The model achieved a mean accuracy of 84.4%, a maximum accuracy of 91.0%, a sensitivity of 83.6%, a specificity of 85.1%, and an AUC of 0.89 with a 95% CI of 0.86-0.92. ***Conclusion*** A machine learning model integrating linear spectral features and multi-band nonlinear dynamic features can effectively distinguish normal voice from pathological voice. This approach demonstrates the feasibility of automatic pathological voice detection using the GRBAS G dimension as the supervisory label and may be useful for noninvasive clinical screening and objective voice evaluation.

【Key words】 Pathological voice; GRBAS; Machine learning; Mel-frequency cepstral coefficients; Nonlinear dynamic features; Support vector machine

嗓音质量评估是嗓音疾病临床诊断、疗效评价与随访监测的重要环节,主要包括主观听感知评估和客观声

基金项目:北京市科技新星交叉课题"面向喉镜视频的喉部病变精准检测与识别研究"(20240484543)

作者单位:1 解放军总医院耳鼻咽喉头颈外科医学部/国家耳鼻喉疾病临床研究中心 北京 100048

2 中国科学院大学 北京 100190

作者简介:杨紫良 硕士;研究方向:咽喉嗓音疾病

通讯作者:黄冬雁,E-mail:huangdy301@163.com

学分析两类[1,2]。其中，听感知评估因操作简便、贴近临床直观判断，是临床嗓音障碍评估的重要方法之一[1,3]。GRBAS评分体系是目前应用广泛的听感知评估工具，可从总体音质（grade，G）、粗糙度（roughness，R）、气息音（breathiness，B）、无力感（asthenia，A）和紧张度（strain，S）5个维度对嗓音质量进行描述[1,3]。然而，听感知评价具有一定主观性，其可靠性受评估者经验、内部判断标准、评分尺度及语音样本等因素影响[3,4]。不同听感知评价维度的评估者间一致性存在差异[5~7]。在中文病理嗓音样本中，刘阳等[6]发现GRBAS总体音质G维度的评估者间一致性相对较高，而R、B、A、S各分项维度的一致性较低。近年来，机器学习与声学特征分析技术逐步应用于病理嗓音自动识别及GRBAS智能评分研究，并在正常与病理嗓音二分类及嗓音严重程度分级方面取得进展[8~11]。鉴于GRBAS-G维度评估者间一致性相对较高，本研究选取G评分作为监督标签，融合梅尔频率倒谱系数（Mel-frequency cepstral coefficients，MFCC）线性频谱特征与Db10小波包分解多频带非线性动力学特征，构建多项式核支持向量机二分类模型，以G0代表正常嗓音、G1~G3代表病理嗓音，探索基于GRBAS-G维度标注的病理嗓音自动检测方法，为临床嗓音异常的无创筛查与客观评估提供技术支撑。

1 材料与方法

本研究为单中心回顾性分析，研究方案遵循赫尔辛基医学宣言伦理准则，经解放军总医院第六医学中心医学伦理委员会批准（HZKY-PJ-2025-4）。

1.1 研究对象与语音样本采集

数据库样本均来自解放军总医院第六医学中心病理嗓音数据库2020年1月至2022年3月嗓音障碍患者资料662例，其中男性423例，女性239例，年龄12～90岁，平均年龄48.22±14.64岁。诊断覆盖声带息肉、声带麻痹、声带白斑、声带囊肿、功能性发育障碍等常见嗓音疾病，同时纳入喉镜检查正常、无嗓音及咽喉疾病史的健康正常嗓音受试者，样本病种与年龄分布具有临床代表性。

所有语音样本均在语音分析室录制，环境噪声小于40 dB（A），使用Xion公司的DiVAS v2.3嗓音分析系统录制，采样率22100 Hz，振幅分辨率16 bit，以wav文件格式无损保存。采集流程标准化，受试者保持舒适坐姿，头戴杆式麦克风置于口唇前约30 cm处。使用舒适自然的声音持续发元音/a/、/i/，要求尽量持续发音3～5秒。

1.2 评估流程

组建由5名具备5年以上嗓音听感知评估资质的专业人员构成评估小组，包括2名耳鼻咽喉头颈外科专科医师、2名言语嗓音治疗师、1名艺术嗓音专业教师，覆盖多专业、多从业背景，以降低单一专业评估偏倚。为统一对嗓音特征的理解和评估尺度，正式评估前，组建跨领域多国专家小组，成员来自中国、美国、等，专业领域包括耳鼻喉科医生、言语病理学家、嗓音治疗师、艺术嗓音教师。专家小组讨论细化听感知评估的操作细则，界定嗓音异常判读标准与分级尺度，评委以此为依据进行后续听感知评估。

评估者采用独立盲法进行评分，评估时位于安静无干扰房间内，使用高质量音响或耳机播放语音样本。播放完一段语音样本后，立即记录评估结果。为评价GRBAS听感知评分的评估者间一致性，基于全部662例语音样本的5名评估者原始评分采用Fleiss Kappa系数分别评价总体音质（G）、粗糙度（R）、气息音（B）、无力感（A）和紧张度（S）5个维度的评估者间一致性。在此基础上，为降低监督标签中的评估者差异，选取5名评估者中至少3名G评分一致的样本用于后续模型构建。最终纳入443例，包括正常嗓音G0样本161个，病理嗓音G1样本115个，G2样本106个，G3样本61个，建立原始数据集。

1.3 声学特征提取

1.3.1 MFCC特征提取 基于Python语言环境，依托Librosa及自编程序完成所有语音样本预处理与特征提取。原始语音信号均经过统一预处理流程，预加重系数设置为0.97，分帧参数为帧长25 ms，帧移10 ms，采用汉明窗函数进行加窗平滑，经快速傅里叶变换完成时域—频域转换，通过26阶梅尔滤波器组将线性频谱映射至人耳非线性梅尔频率，进一步倒谱运算生成26维梅尔倒谱系数，剔除低阶冗余分量，保留第2～13维共12维倒谱参数，同步叠加单帧对数能量构成13维基础特征向量；最终以整段语音各帧特征全局均值作为单样本MFCC表征特征。

1.3.2 多频带非线性参数特征 为刻画病理嗓音的复杂动力学特征，本研究引入3类经典非线性分析指标，包括Hurst指数（采用R/S重标极差法求解）、计盒维数（Db）及Rényi熵。对输入的原始语音信号使用Db10小波进行4层分解，分解后生成16个相互正交的独立频率子带；逐个子带分别提取Hurst参数、Db、2-Rényi熵3类非线性指标，单类特征经多子带分解后形成16维多尺度特征向量，最终整合得到高维非线性动力学特征集。

1.4 模型开发与评估

以G0为正常嗓音标签、G1～G3合并为病理嗓音标签，构建二分类多项式核支持向量机（support vector machine，SVM）分类模型。首先在全部443例建模样本上采用网格搜索优化模型参数，包括多项式阶数、惩罚系

数C与核系数γ，确定最优参数组合；随后固定该参数组合，采用分层十折交叉验证评估模型性能。评价指标包括平均准确率、最高准确率、敏感度、特异度、受试者工作特征曲线下面积（area under the receiver operating characteristic curve，AUC）及95%置信区间。通过比较不同特征组合的分类效能，筛选最优特征融合方案。

2 结果

5名评估者对662例语音样本的GRBAS评分者间一致性分析显示，G、R、B、A和S的Fleiss Kappa值分别为0.51、0.28、0.25、0.28和0.27。其中，G维度的Fleiss Kappa值最高，R、B、A、S各维度均低于0.30，提示G维度的评估者间一致性相对较高。

经分层十折交叉验证，不同单一特征及多特征融合组合的二分类效能见表1。单一特征中，MFCC性能最优（平均准确率76.7%，AUC=0.82），非线性单一特征Hurst效果最弱，提示传统频谱特征可有效表征嗓音总体音质。多特征融合后模型性能持续提升，线性特征与非线性特征存在显著互补效应。最优特征组合为MFCC+Hurst+2-Rényi+Db，模型达到最佳性能：平均准确率84.4%，最高准确率91.0%，敏感度83.6%，特异度85.1%，AUC=0.89，95%置信区间0.86～0.92，表明该融合特征可稳定、有效地区分正常嗓音与病理嗓音。

3 讨论

本研究基于GRBAS听感知评估体系，选取评估者间一致性相对较高的总体音质G维度评分作为监督标签，融合MFCC线性频谱特征与小波包多频带非线性动力学特征构建多项式核SVM模型，实现正常嗓音与病理嗓音的自动区分。最优特征组合下模型平均准确率为84.4%，AUC为0.89，提示利用GRBAS-G维度进行监督标注，结合声学特征与机器学习开展病理嗓音自动检测具有可行性，可为临床异常嗓音筛查与客观评估提供辅助工具。在不同特征组合中，多特征融合优于单一特征，说明频谱特征与非线性特征在表征病理嗓音方面具有互补作用。

本研究选取GRBAS-G维度评分作为监督学习标签对662例语音样本进行评估者间一致性分析发现，G维度Fleiss Kappa值为0.51，高于R（0.28）、B（0.25）、A（0.28）和S（0.27），表明在GRBAS各维度中，G评分具有较好的评估者间一致性。该结果与刘阳等[6]对中文病理嗓音样本的研究一致，其报道G、R、B、A和S维度的Fleiss Kappa值分别为0.48、0.27、0.26、0.29和0.25，同样表现为G维度一致性较高，其他分项维度一致性较低。本研究G维度的一致性仍然有限，因此进一步采用至少3名评估者G评分一致作为建模样本筛选标准，以减少评估者差异所造成的监督标签噪声。最终443例样本进入模型训练与验证。基于该标注策略构建的模型能较好区分正常嗓音与病理嗓音，为采用G维度进行监督标注提供了依据。

在特征构建方面，MFCC作为核心倒谱特征，能够有效刻画语音谱包络信息，既往研究已被广泛用于病理语音分析，其倒谱分量与发声障碍严重程度具有良好的相关性[12~16]。然而，传统线性声学指标，如基频微扰（jitter）、振幅微扰（shimmer）、谐噪比（harmonics-to-noise ratio，HNR）等在重度嘶哑或非周期性增强时，稳定性下降[17~19]，本研究引入小波包多频带非线性特征已被用于刻画病理嗓音中的复杂非周期性行为，并在嗓音障碍检测中表现出较好的区分能力[20~22]。其中，Hurst指数用于表征语音信号的长程相关性与稳定性，计盒维数用于描述信号的复杂性差异，Rényi熵则用于衡量信号的不确定性与信息复杂度。上述非线性特征的引入，显著提升了模型区分效能。

在模型选择方面，SVM在本研究数据集上取得了较好的分类性能，最优特征组合的平均准确率为84.4%，AUC为0.89。既往研究采用人工智能神经网络、深度学习及其他机器学习方法开展病理嗓音自动检测与质量评估[23~25]。然而，不同研究在数据来源、研究对象、声学特征、样本构成及模型验证策略等方面存在差异，因此，不同研究间的模型性能不宜进行直接横向比较。本研究结果表明，在当前样本规模及特征工程框架下，基于SVM的传

表1 不同特征和特征组合的对G0和G1～G3二分类模型效能

特征组合	平均准确率(%)	最高准确率(%)	敏感度(%)	特异度(%)	AUC	95%CI
MFCC	76.7	87.8	75.2	78.1	0.82	0.78～0.86
Hurst	66.3	72.7	64.8	67.5	0.71	0.66～0.76
MFCC + Hurst	80.5	90.3	79.3	81.6	0.86	0.82～0.89
MFCC + 2-Rényi	80.8	90.6	79.5	82.0	0.86	0.83～0.89
MFCC + 2-Rényi + Db	81.9	87.5	80.6	83.1	0.87	0.84～0.90
MFCC + Hurst + 2-Rényi	83.5	90.6	82.3	84.5	0.88	0.85～0.91
MFCC + Hurst + 2-Rényi + Db	84.4	91.0	83.6	85.1	0.89	0.86～0.92

统机器学习方法仍可获得较好的病理嗓音判别效能，并可为后续开展不同模型比较及深度学习研究提供基础。

从临床应用价值分析，GRBAS评分受评估者经验、心理状态与环境影响较大，评估者间与评估者内一致性偏低。本研究采用基于语音声学信号的病理嗓音自动检测模型，有助于实现标准化、可重复的量化评估，是对听感知评估客观化的积极探索，为嗓音质量的自动化评估提供了可行路径。

喉镜检查是喉部器质性病变确诊的金标准，但存在有创、费用高、医疗资源消耗大等局限。本模型可作为无创初筛与风险分层工具，通过筛查区分正常与病理嗓音，高风险者建议进一步行喉镜检查明确病因，低风险者可暂缓侵入性检查、定期随访或保守干预，从而辅助临床决策，有助于优化医疗资源配置、减少不必要有创检查。

本研究存在一定局限，其一，采用MFCC等声学特征作为主要表征，特征提取依赖持续元音的相对稳态段，因此，目前建模主要基于持续元音样本，且在连续语流（如朗读或自发语音）场景下的适用性仍需进一步验证[26]；其二，本研究以专家听感知评分作为监督学习标签，尽管已通过评分一致性，提高标注质量，但主观评估的个体差异仍可能影响目标值的真实性，从而限制模型泛化能力，后续可在扩大样本量的同时引入更多评估者，开展统一培训与一致性评估，以进一步提升标注可靠性；其三，本研究仅以G评分作为监督标签，将G0定义为正常嗓音、G1～G3合并为病理嗓音，因此，模型实现的是正常与病理嗓音的二分类检测，而非G0、G1、G2、G3严重程度的自动分级，未涉及R、B、A、S各维度的自动预测。后续研究需进一步扩大样本量并优化模型，实现G维度严重程度分级及GRBAS多维度自动评估。本研究病理嗓音组包含声带息肉、声带麻痹、声带白斑、声带囊肿及功能性发育障碍等多种疾病，不同病种的病理生理机制及声学表现存在差异。本研究将其统一合并为病理嗓音进行二分类建模，尚未开展按具体病种的分层验证，因此不能排除部分模型利用某些特定病种的声学特征进行判别的可能，对训练数据中未覆盖病种的泛化能力有待验证。后续研究将在扩大不同病种样本量的基础上，进一步开展病种分层及逐病种留一验证，结合多中心独立外部队列，评价模型对不同疾病类型及未知病种的泛化性能。此外，本研究超参数寻优在全部建模样本上先行完成，随后采用分层十折交叉验证评价模型性能，未采用严格的嵌套交叉验证。因此，超参数选择阶段已利用全部样本信息，当前报告的模型性能可能存在一定程度的乐观偏倚[27,28]。后续研究将采用嵌套交叉验证，即在外层交叉验证中评价模型性能、在内层训练数据中独立完成超参数寻优，结合独立外部验证队列，严格评价模型的泛化性能。最后，本研究为单中心回顾性数据，缺乏外部队列验证，模型的普适性需进一步检验[29]。未来研究需扩大多中心临床数据规模，纳入连续自然语音数据，实现GRBAS多维度自动评分；结合模型可解释性分析、校准优化[30]，进一步提升模型透明度与临床适用性，形成可落地的临床嗓音智能评估系统。

融合MFCC线性频谱特征与小波包多频带非线性动力学特征的多项式核SVM模型，可较好区分正常嗓音与病理嗓音。基于GRBAS-G维度标注的病理嗓音自动检测具有可行性，可为临床嗓音疾病的无创筛查及客观评估提供辅助工具。未来仍需通过多中心外部验证及G0～G3严重程度分级研究进一步评价其临床应用价值。

参考文献

[1] 中华医学会耳鼻咽喉头颈外科学分会嗓音学组, 中华耳鼻咽喉头颈外科杂志编辑委员会咽喉组. 嗓音功能评估专家共识(2024)[J]. 中华耳鼻咽喉头颈外科杂志, 2024, 59(12):1267-1278.

[2] Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function[J]. American Journal of Speech-Language Pathology, 2018, 27(3):887-905.

[3] Oates J. Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions[J]. Folia Phoniatrica et Logopaedica, 2009, 61(1):49-56.

[4] Takemoto N, Sanuki T, Kozaki A, et al. Reliability of auditory-perceptual voice evaluation: an investigation of GRBAS scale reproducibility among expert raters[J/OL]. Journal of Voice, 2026.

[5] Özcebe E, Aydınlı FE, Tığrak TK, et al. Reliability and validity of the Turkish version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) [J]. Journal of Voice, 2019, 33(3): 382.e1-382.e10.

[6] 刘阳, 刘恒鑫, 屈歌, 等. GRBAS和CAPE-V量表中文听感知评估一致性分析[J]. 听力学及言语疾病杂志, 2025, 33(1):23-28.

[7] De Bodt MS, Wuyts FL, Van de Heyning PH, et al. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality[J]. Journal of Voice, 1997, 11(1):74-80.

[8] Sáenz-Lechón N, Godino-Llorente JI, Osma-Ruiz V, et al. Automatic assessment of voice quality according to the GRBAS scale[C]. 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2006,2478-2481.

[9] Arias-Londoño JD, Godino-Llorente JI, Sáenz-Lechón N, et al. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients[J]. IEEE Transactions

on Biomedical Engineering, 2011, 58(2):370-379.

[10] Berteloot L, Sismono F, Maertens L, et al. AI-driven detection and classification of voice disorders using acoustic recordings[J/OL]. Journal of Voice, 2026.

[11] Au YC, Yan N, Ng ML. Automatic assessment of voice quality using machine learning[J]. Logopedics Phoniatrics Vocology, 2026, 51(3): 167-184.

[12] Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4):357-366.

[13] Godino-Llorente JI, Gómez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors[J]. IEEE Transactions on Biomedical Engineering, 2004, 51(2):380-384.

[14] Awan SN, Roy N, Jetté ME, et al. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: comparisons with auditory-perceptual judgements from the CAPE-V[J]. Clinical Linguistics & Phonetics, 2010, 24(9):742-758.

[15] Yagnavajjula MK, Alku P, Rao KS, et al. Detection of neurogenic voice disorders using the Fisher vector representation of cepstral features[J]. Journal of Voice, 2025, 39(3):757-763.

[16] Maryn Y, Roy N, De Bodt M, et al. Acoustic measurement of overall voice quality: a meta-analysis[J]. The Journal of the Acoustical Society of America, 2009, 126(5):2619-2634.

[17] Baken RJ, Orlikoff RF. Clinical measurement of speech and voice[M]. 2nd ed. San Diego: Singular Thomson Learning, 2000. 130-137, 190-213.

[18] Brockmann-Bauser M, Bohlender JE, Mehta DD. Acoustic perturbation measures improve with increasing vocal intensity in individuals with and without voice disorders[J]. Journal of Voice, 2018, 32(2):162-168.

[19] Ma EPM, Yiu EML. Suitability of acoustic perturbation measures in analysing periodic and nearly periodic voice signals[J]. Folia Phoniatrica et Logopaedica, 2005, 57(1):38-47.

[20] Little MA, McSharry PE, Roberts SJ, et al. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection[J]. BioMedical Engineering OnLine, 2007, 6:23-23.

[21] Zhang XH, Hu WP. Pathological voice classification based on wavelet packet multiscale analysis[C]. Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. New York: ACM, 2018,1-6.

[22] Dalalana GJP, Guido RC, Honorato ES, et al. Application of wavelet analysis and paraconsistent feature extraction in the classification of voice pathologies[J/OL]. Journal of Voice, 2025.

[23] 庞宇峰, 黄娟, 徐蓓峥, 等. 病态嗓音的定量分析及人工神经网络识别[J]. 临床耳鼻咽喉头颈外科杂志, 2017, 31(2):100-102.

[24] Fang SH, Tsao Y, Hsiao MJ, et al. Detection of pathological voice using cepstrum vectors: a deep learning approach[J]. Journal of Voice, 2019, 33(5):634-641.

[25] Godino-Llorente JI, Gómez-Vilda P, Blanco-Velasco M. Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters[J]. IEEE Transactions on Biomedical Engineering, 2006, 53 (10):1943-1953.

[26] Maryn Y, Corthals P, Van Cauwenberge P, et al. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels[J]. Journal of Voice, 2010, 24(5):540-555.

[27] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection[J]. BMC Bioinformatics, 2006, 7: 91-91.

[28] Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation[J]. Journal of Machine Learning Research, 2010, 11:2079-2107.

[29] Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation[J]. Journal of Clinical Epidemiology, 2016, 69:245-247.

[30] Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics[J]. BMC Medicine, 2019, 17: 230-230.

收稿日期 2026-05-27
修回日期 2026-08-25
录用日期 2026-09-28
责任编辑 薛　静