

基于深度学习的内镜肠道准备评分模型的建立

沈文娟, 徐昶, 林嘉希, 许春芳, 陆建英, 朱锦舟

苏州大学附属第一医院 消化内科, 江苏 苏州 215006

[摘要] 目的 基于深度学习算法构建内镜肠道准备评分的计算机视觉模型。方法 收集苏州大学附属第一医院消化内镜中心(600张)及HyperKvasir数据库(1794张)的内镜图片共2394张,根据Boston肠道准备量表完成肠道清洁度评分(0~3分,四分类),按6:2:2随机分为训练集(1439张)、验证集(478张)和测试集(477张)。选取3种深度学习网络(DenseNet169、DenseNet121、EfficientNet B3),利用迁移学习方式训练肠道准备分类模型,并采用测试集的混淆矩阵等指标评价模型分类能力,与高、低年资医师的分类能力进行对比。结果 成功构建3个基于深度学习的肠道准备分类模型。各模型的分类准确度均较高,平均分类准确度为0.897,近似于低年资内镜医师(0.914),低于高年资内镜医师(0.941)的分类表现。其中,DenseNet169模型表现最好,分类准确度(0.914)及平均精确度(0.892)均为最高。此外,采用梯度加权分类激活映射算法,用热力图形式对模型的分类推理进行可视化呈现。结论 运用深度学习算法构建的内镜肠道准备分类模型具有可行性,可通过多中心研究扩大样本来源进一步提高模型的分类及泛化能力。

[关键词] 深度学习; 计算机视觉; 卷积神经网络; 梯度加权分类激活映射

Computer Vision Models for Endoscopic Bowel Preparation Scoring Based on Deep Learning

SHEN Wenjuan, XU Chang, LIN Jiaxi, XU Chunfang, LU Jianying, ZHU Jinzhou

Department of Gastroenterology, The First Affiliated Hospital of Soochow University, Suzhou Jiangsu 215006, China

Abstract: **Objective** To develop computer vision models for endoscopic bowel preparation scoring based on deep learning. **Methods** A total of 2394 endoscopic images from the Gastrointestinal Endoscopy Centre of the First Affiliated Hospital of Soochow University ($n=600$) and the HyperKvasir database ($n=1794$) were collected, scored by endoscopists according to the Boston bowel preparation scale (BBPS, 0-3, four categories). They were randomly divided into training sets (1439 pieces), verification sets (478 pieces) and test sets (477 pieces) according to 6 : 2 : 2. Three deep learning networks (DenseNet169, DenseNet121, EfficientNet B3) were used to develop the bowel preparation classification models by transfer learning. Metrics such as confusion matrices in the test set were used for model evaluation. Meanwhile, the models were compared with senior and junior endoscopists. **Results** The three deep learning-based bowel preparation classification models were successfully developed. The classification accuracy of all models was high, and the average classification accuracy was 0.897, which was similar to the clarification performance of junior endoscopist (0.914) and lower than that of senior endoscopist (0.941). Among them, the best performing model was the DenseNet169 model, which had the highest classification accuracy (0.914) and the highest average precision (0.892). In addition, the visual interpretation of the models' classification results was presented in the form of heat maps by using gradient-weighted class activation mapping. **Conclusion** The endoscopic bowel preparation classification model developed using deep learning is feasible, and the classification and generalization ability of the model can be further improved by expanding the sample source through multicenter studies.

Key words: deep learning; computer vision; convolutional neural network; gradient-weighted class activation mapping

[中图分类号] R197.39; TP391.4

[文献标识码] A

doi: 10.3969/j.issn.1674-1633.2023.11.003

[文章编号] 1674-1633(2023)11-0011-05

引言

内镜检查由于疾病检出率高、安全性好,被广泛用于消化系统检查^[1-2]。研究显示,肠道准备情况对内镜检查结果影响极大,不理想的肠道准备是有效内镜检

查的主要障碍^[3-5]。《欧洲胃肠内镜学会(ESGE)指南》(2019版)^[6]更新了肠道清洁度的最低标准,目前主要根据Boston肠道准备量表(Boston Bowel Preparation Scale, BBPS)进行肠道准备评分^[7]。然而, BBPS评分易因内镜医生的主观性产生差异,影响结果可靠性。近年来,以深度学习算法为代表的人工智能(Artificial Intelligence, AI)技术飞速发展,计算机视觉模型广泛应用于临床诊疗各个环节中。在内镜肠道准备应用场景

收稿日期: 2023-06-28

基金项目: 国家自然科学基金(82000540); 苏州市科教兴卫项目(KJXW2019001)。

通信作者: 朱锦舟, 副主任医师, 博士, 主要研究方向为AI在内镜诊疗中的应用。

通信作者邮箱: jzzhu@zju.edu.cn

中,既往有数项研究对此进行了报道,但均因自身局限性,包括二分类而不是四分类^[8]以及样本量过少^[9]等原因,不能很好地反映深度学习技术在该领域的优势。为了建立更加高效客观的肠道准备评估过程,本研究收集苏州大学附属第一医院消化内镜中心及 HyperKvasir 数据库^[10] 肠镜图片,通过深度卷积神经网络迁移学习,将训练好的深度卷积神经网络模型的信息包括参数、层次权重、架构等迁移到目标域,即本研究收集的肠镜图片,根据本研究四分类任务要求调整结构和参数,建立针对内镜肠道准备评分的计算机视觉模型。

1 资料与方法

1.1 一般资料

本研究为多中心的回顾性分析,共收集不同肠道清洁度的内镜图片 2394 张,其中,600 张来自苏州大学附属第一医院消化内镜中心,1794 张来自 HyperKvasir 数据库。

1.2 图片标注及预处理

由两名具有 10 年以上工作经验的内镜医生对肠镜图片根据 BBPS^[7] 打分。BBPS 具体评分为:0 分:未进行肠道准备的肠段,内有无法清除的固体粪而不能看清黏膜(562 张);1 分:肠段内黏膜部分可见,另一部分由于粪便、不透明液体残留显示不清(347 张);2 分:结肠内有少量小块粪便及不透明液体,黏膜显示清楚(387 张);3 分:肠段内所有黏膜显示清楚,结肠内没有粪便、不透明液体残留(1098 张)。

所纳入图片均进行了去身份化处理,去除所有图片中包含的患者信息。为提高深度卷积神经网络运算效率,将纳入的 2394 张图片统一调整尺寸至 224×224 像素。所有图片按 6:2:2 分为训练集(60%)、验证集(20%)、测试集(20%)。

1.3 构建肠道准备分类模型

本研究使用了 3 种不同的深度卷积神经网络架构包括 DenseNet169、DenseNet121 及 EfficientNet B3,利用深度卷积神经网络迁移学习方式,建立肠道准备分类模型。架构训练流程如下:① 架构调整:3 种架构输出层均被调整为四分类,以适应 BBPS 分类任务需求;② 载入预训练权重:架构除全连接层及输出层外,其余层权重均采用来自 ImageNet 数据库中预训练的权重。全连接层及输出层权重采用随机权重;③ 模型训练:模型训练采用迁移学习中微调策略,即训练过程中更新全连接层及输出层权重,冻结其余层权重。同时,为提高架构泛化性,训练时采用线上图片增强方式,具体为训练过程中,随机对训练集图片进行旋转,翻折等非刚性变化;④ 训练超参数及策略:架构学习率为 0.001,采用批量

化载入训练框架中,每批(batch)数量为 128,训练轮次(epoch)为 100。架构采用 Stochastic gradient descent 作为优化器,采用 Softmax 函数作为分类器。本研究采用早期停止(early-stop)策略(10 轮次)。模型训练过程如图 1 所示。

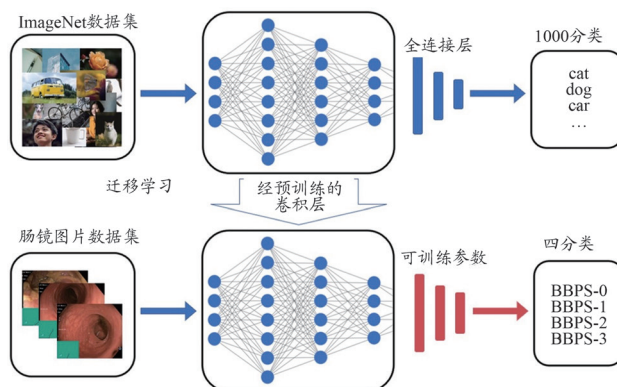


图1 肠道准备分类模型的迁移学习示意图

1.4 可视化解释肠道准备分类模型分类结果

深度卷积神经网络有较高的准确度,但其内部机制较难以理解,致使分类结果的可信度降低。本研究使用梯度加权分类激活映射(Gradient-Weighted Class Activation Mapping, Grad-CAM)算法,对模型分类推理过程进行可视化呈现。该方法通过提取网络模型末端的特征层并对所有特征图进行加权求和,建立得到最终的激活热力图,可视化呈现图片各部位对模型决策的重要程度。之后,将类激活热力图与初始内镜图片叠加,便能实现对模型分类结果的可视化解释。

1.5 肠道准备分类模型评价并与内镜医师分类结果对比

使用混淆矩阵(confusion matrix)呈现上述构建的 3 个肠道准备分类模型和高、低年资医师对测试集数据的分类结果,进行对比评估模型分类水平;同时,以准确度(Accuracy)、精确度(Precision)、召回率(Recall)、特异性(Specificity)、F1 值(F1-score)为评价指标,对各个模型的分类能力进行评价。

2 结果

2.1 肠道准备分类模型评价

基于深度卷积神经网络框架迁移学习后的 3 个模型均拥有较高的分类准确度,平均分类准确度为 0.897。其中,表现最好的模型是 Densenet169 模型,相较另外两种模型,其分类准确度及平均精确度均为最高,分别为 0.914 和 0.892,其分类能力等同于低年资医师(0.914),低于高年资医师(0.941)。三分类模型和高、低年资医师在测试集数据表现的混淆矩阵和效能的总体评价如图 2~3 所示。基于深度卷积神经网络框架迁移学习后的 3 个模型对于测试集 477 个数据均拥有

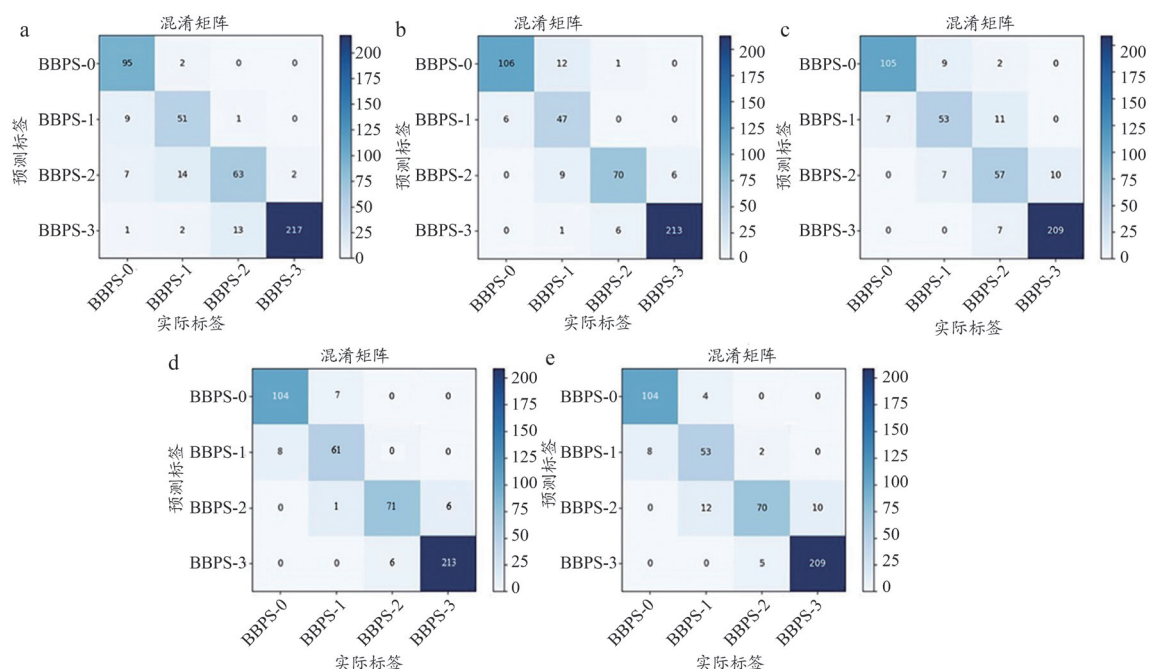


图2 肠道准备分类模型和高、低年资医师在测试集的混淆矩阵

注: a. DenseNet121 模型; b. DenseNet169 模型; c. EfficientNet B3 模型; d. 高年资医师; e. 低年资医师。

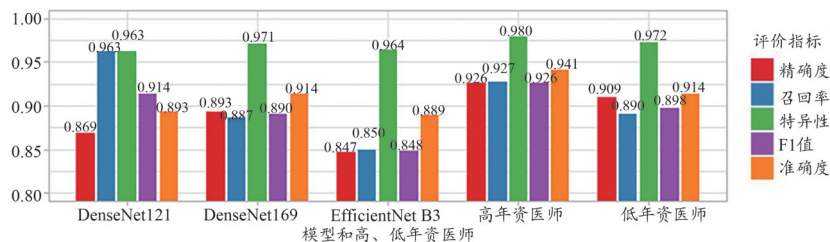


图3 肠道准备分类模型和高、低年资医师在测试集的总体评价指标

较高的分类准确度, 平均分类准确度为 0.897。其中, Densenet169 模型表现最优, 准确度为 0.914, 精确度为 0.893, 特异性为 0.971, 召回率为 0.887, F1 值 0.890, 相较另 2 种模型, 其分类准确度及平均精确度均为最高, 为最佳模型, 其分类能力等同于低年资医师 (0.914), 低于高年资医师 (0.941)。模型及高低年资内镜医师分类 BBPS 评分具体表现如图 4 所示。统计效能在分类 BBPS-2、BBPS-3 分时均有不同程度的下降, 模型及低年资医师的精确度和召回率下降较为显著。同时, 为进一步比较模型与高低年资内镜医师 BBPS 评分分类能力, 3 个模型的平均性能与医师分类表现进行的比较结果如图 5 所示, 可见深度卷积神经网络模型的平均分类性能与内镜医师十分接近。

2.2 肠道准备分类模型分类结果的可视化解

本研究使用 Grad-CAM 算法对肠道准备分类模型进行可视化解, 通过加权求和特征图得到类激活热力图, 结合内镜图片生成高亮热力图 (图 6)。热力图标注的红色区域是在模型判读图片中影响决策权重较高的区域, 浅蓝色区域则是影响权重较低的区域, 可见模型较为准确地识别了图片的特征位置, 即粪便、不透明液体 (红

色区域), 与其余显示清楚的黏膜 (蓝色区域) 进行区分, 并基于识别的特征作出分类判断。因此, 热力图中的不同颜色可体现模型在进行图片分类时的分析过程, 即大致判断肠道内粪便、不透明液体的量从而分类肠道清洁度, 实现了对模型分类结果的可视化解。

3 讨论

随着 AI 技术的逐步发展和深入, 其在医学领域展现了巨大的潜力^[11-12], 深度学习作为 AI 领域的一个热门主题, 在图像分类任务中的卓越表现使其在消化内镜领域的运用范围越来越广泛, 如识别直结肠腺瘤、诊断 Barrett 食管、肠镜检查等^[13-15]。迄今为止, 已有多个研究使用 AI 技术建立关于肠道准备的分类模型。Zhou 等^[9]收集 120 张内镜图片, 使用深度卷积神经网络建立肠道准备分类模型, 模型准确度为 0.933。Zhou 等^[16]收集 67411 张内镜图片开发了一个基于深度学习的自动化肠道准备评分系统用于细分 BBPS 评分为 0、1 分的图片, 模型准确度为 0.886。Lee 等^[8]收集 73304 张内镜图片, 运用深度卷积神经网络建立 BBPS 评分二分类模型, 结果显示该模型 AUC 值为 0.918, 准确度为 0.853, 敏感

度为 100%。邓少琦等^[17]纳入 455 例研究对象,运用神经网络建立肠道准备预测模型,结果显示,模型 AUC 值 0.918。此外, AI 算法也被用于病例对照试验中比较小肠清洁度^[18]。可见大部分研究为了提高模型分类能力,主要聚焦于建立肠道准备结果的二分类模型,少有文献报道收集千张图片建立肠道清洁度分类模型的研究。

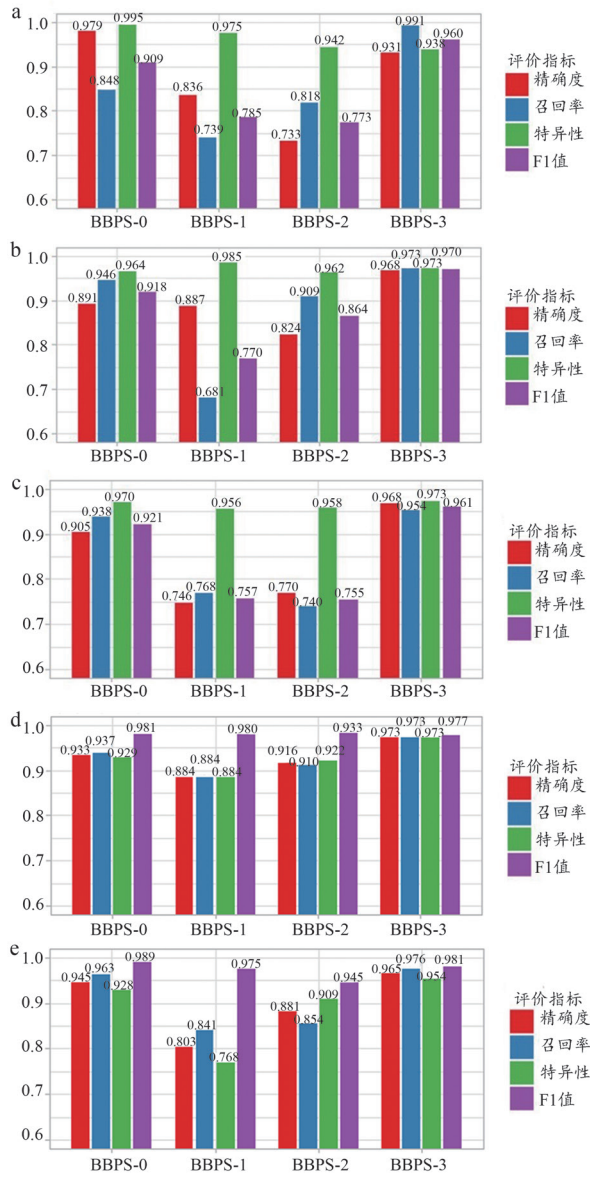


图4 3个模型和高、低年资医师分别在测试集的四分类表现
注: a. DenseNet121模型; b. DenseNet169模型; c. EfficientNet B3模型; d. 高年资医师; e. 低年资医师。

肠道清洁度主要根据内镜医生判断打分,评分准确性易受到观察者的主观性限制。同时,时间限制或疲劳等外部因素也会影响内镜医生的状态从而影响评分结果^[3]。另外,有研究发现,在实际的临床应用中,内镜医师基于术后回忆进行评分可能会导致报告的不准确甚至遗漏^[5]。因此可自动化辅助肠道准备评分的 AI 有重

要的临床价值。利用自动评分系统,一方面提高了肠道评分的客观性,另一方面可作为内镜报告质量控制系统的潜在组成部分,降低内镜报告的错误率^[19]。

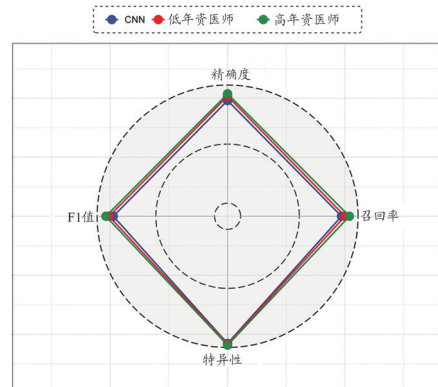


图5 3个模型平均性能与内镜医师比较
注: CNN: 卷积神经网络。

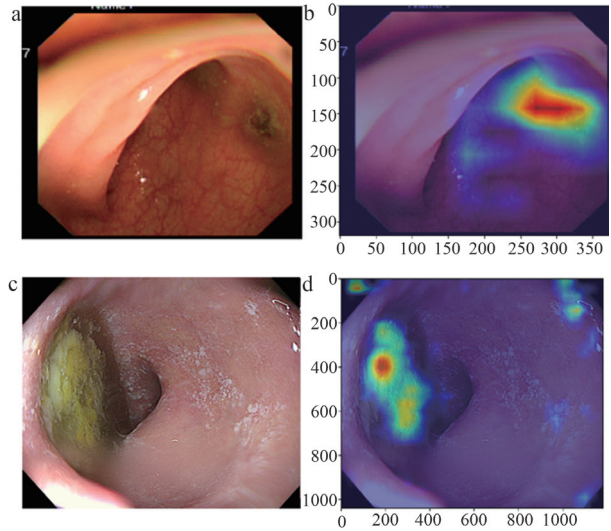


图6 基于Grad-CAM的模型可视化解释

注: a、c. 肠镜图片; b、d. 肠镜图片叠加类激活热力图后的高亮热力图。

本研究成功构建了一系列自动评价肠道清洁度 BBPS 评分的模型,且有着较好的分类准确度,与 Zhou 等^[9]构建的肠道准备分类模型思路相似,但本研究收集了来自多中心的更大的样本量,增加模型的泛化性,同时尝试多种网络架构,进一步提升了模型性能。AI 的高可重复性将允许对肠道准备进行客观和一致的评估。经过训练部署后的模型可以快速、准确地进行图像判读,性能近似于低年资医师,有利于改进工作流程和减少医疗差错^[20],具有可重复性高、节省人工和时间的优点。

但本研究尚存在一些局限性:首先,本研究收集的样本量较少,来源较单一。肠道清洁度评分,尤其是 BBPS 评分为 1 分和 2 分,包含观察者较强的主观性,需要大量多样化的样本来抵消,样本量不足会降低模型的准确度,致使本研究构建的 3 个模型在分类 1 分和 2 分

的图片时,相较于0分和3分,模型精确度和召回率降低,需要后续研究增加来源不同的样本,进一步提高模型的分分类能力和泛化性。其次,本研究合并了来源不同的两个图片集,缺乏外部数据集来验证临床价值,后续需要对该模型进行多中心验证,探讨模型的泛化性,不断优化模型。

4 结论

本研究使用于ImageNet数据库中预训练的多个计算机视觉模型,包括DenseNet169、DenseNet121及EfficientNet B3 3个深度卷积神经网络框架,构建了具有较好分类效能的肠道准备四分类模型,分类准确度均较高。其中,DenseNet169模型表现最好,分类准确度(0.914)及平均精确度(0.892)均为最高。此外,采用Grad-CAM算法可视化解释了模型分类结果,证明了深度学习建立图片多分类模型的可能性。该模型有助于提高内镜医师肠道准备评分的准确度,为临床内镜检查提供便利,促进建立高效、客观的内镜检查流程,为未来构建更加可靠的肠道准备评分系统提供参考。

[参考文献]

- [1] Yang Y, Xiao Y, Zhang L, *et al.* Effects of different intervention methods on intestinal cleanliness in children undergoing colonoscopy[J]. *J Healthc Eng*, 2022, 2022: 1898610.
- [2] 李兆申, 张荊. 消化内镜进入新时代[J]. *中华消化杂志*, 2021, 41(6): 361-365.
Li ZS, Zhang D. New era of digestive endoscopy[J]. *Chin J Dig*, 2021, 41(6): 361-365.
- [3] Ahmad OF. Deep learning for automated bowel preparation assessment during colonoscopy: time to embrace a new approach?[J]. *Lancet Digit Health*, 2021, 3(11): e685-e686.
- [4] Bisschops R. Top tips for evaluating and cleaning up bowel preparation[J]. *Gastrointest Endosc*, 2022, 95(5): 990-995.
- [5] Jacobson BC, Calderwood AH. Measuring bowel preparation adequacy in colonoscopy-based research: review of key considerations[J]. *Gastrointest Endosc*, 2020, 91(2): 248-256.
- [6] Hassan C, East J, Radaelli F, *et al.* Bowel preparation for colonoscopy: European Society of Gastrointestinal Endoscopy (ESGE) Guideline-update 2019[J]. *Endoscopy*, 2019, 51(8): 775-794.
- [7] Lu W, Zhou K, Cai C, *et al.* Effects on BBPS score with bowel preparation time and dosage[J]. *Medicine(Baltimore)*, 2022, 101(27): e29897.
- [8] Lee JY, Calderwood AH, Karnes W, *et al.* Artificial intelligence for the assessment of bowel preparation[J]. *Gastrointest Endosc*, 2022, 95(3): 512-518.e1.
- [9] Zhou J, Wu L, Wan X, *et al.* A novel artificial intelligence system for the assessment of bowel preparation (with video)[J]. *Gastrointest Endosc*, 2020, 91(2): 428-435.e2.
- [10] Borgli H, Thambawita V, Smedsrud PH, *et al.* HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy[J]. *Sci Data*, 2020, 7(1): 283.
- [11] Bini SA. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?[J]. *J Arthroplasty*, 2018, 33(8): 2358-2361.
- [12] 蒋西然, 蒋韬, 孙嘉瑶, 等. 深度学习人工智能技术在医学影像辅助分析中的应用[J]. *中国医疗设备*, 2021, 36(6): 164-171.
Jiang XR, Jiang T, Sun JY, *et al.* Deep learning in computer aided analyses of medical images[J]. *China Med Devices*, 2021, 36(6): 164-171.
- [13] 张晨霞, 李迅, 姚理文, 等. 基于深度学习的超声内镜下消化道黏膜下肿物诊断系统[J]. *中华消化杂志*, 2022, 42(7): 464-469.
Zhang CX, Li X, Yao LW, *et al.* Deep learning-based diagnostic system for gastrointestinal submucosal tumor under endoscopic ultrasonography[J]. *Chin J Dig*, 2022, 42(7): 464-469.
- [14] Moore M, Sharma P. Updates in artificial intelligence in gastroenterology endoscopy in 2020[J]. *Curr Opin Gastroenterol*, 2021, 37(5): 428-433.
- [15] 王跃, 王卫东, 赵蕾, 等. 基于迁移学习的胃镜图像自动识别多分类系统的研究[J]. *中国医疗设备*, 2021, 36(3): 81-84.
Wang Y, Wang WD, Zhao L, *et al.* Research on multi-classification system for automatic recognition of gastroscope images based on transfer learning[J]. *China Med Devices*, 2021, 36(3): 81-84.
- [16] Zhou W, Yao L, Wu H, *et al.* Multi-step validation of a deep learning-based system for the quantification of bowel preparation: a prospective, observational study[J]. *Lancet Digit Health*, 2021, 3(11): e697-e706.
- [17] 邓少琦, 林丹丹, 刘鑫杰, 等. 人工智能算法与Logistic回归在肠道准备预测中的应用[J]. *现代消化及介入诊疗*, 2021, 26(8): 1008-1013.
- [18] Jun OD, Youngbae H, Hyung NJ, *et al.* Small bowel cleanliness in capsule endoscopy: a case-control study using validated artificial intelligence algorithm[J]. *Sci Rep*, 2022, 12(1): 18265.
- [19] 赵子夜, 王成龙, 袁捷, 等. 结肠镜技能训练、能力与质量评估的研究进展[J]. *中华消化内镜杂志*, 2022, 39(9): 686-690.
Zhao ZY, Wang CL, Yuan J, *et al.* Research advances in colonoscopic technical training, competence and quality evaluation[J]. *Chin J Dig*, 2022, 39(9): 686-690.
- [20] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence[J]. *Nature Med*, 2019, 25(1): 44-56.

本文编辑 崔丽君