基于深度森林的 CT 图像结直肠息肉检测研究

陈祎琼[1,2]，刘 澳[1]，范国华[1,2]，毕家泽[1]，陈 滔[3]

(1.安徽农业大学 信息与计算机学院，安徽 合肥 230036；2.安徽省北斗精准农业信息工程实验室，安徽 合肥 230036；
3.安徽农业大学 工学院，安徽 合肥 230036)

摘 要： 为提高在 CT 图像中结直肠息肉的筛选效率，提出一种基于深度森林的结直肠息肉 CT 图像检测方法。通过灰度化、归一化、中值滤波、随机旋转的手段对数据集进行预处理，将处理后的数据输入一个调整后的深度森林进行预测分类，得到输出结果。实验结果表明，该模型与其他分类算法采用不同指标对比后，具有较好的分类效果，分类精度达到了 99.67%，同时该模型具有较少的超参数，泛化能力强，有助于在医学影像领域辅助医生筛查疾病患者。

关键词： 深度森林；医学影像；结直肠息肉；图像分类

DOI：10.3969/j.issn.1674-5043.2022.01.012

中图分类号： TP391.4 **文献标识码：** A **文章编号：** 1674-5043(2022)01-0068-07

最近几年中，结直肠癌是最常见的恶性肿瘤之一。根据美国癌症协会的报告[1]，结直肠癌在美国的致死率与患病率在所有癌症中排名第二。多数结直肠癌起于黏膜息肉，其产生与多种因素相关[2]，起始于异常的腺窝，后逐渐演变成息肉，最终发展为结肠直肠癌，整个过程可持续多年。结直肠息肉的早期检测是预防结直肠癌、降低发病率、死亡率的有效途径，如何正确地判断结直肠息肉的存在成为了一个值得思考的问题。

针对 CT 图像中结直肠息肉是否存在分类的问题，国内外学者展开了深入地研究，并将演变成的息肉特征与智能算法结合，陈奕志等[3]研究出一种基于深度学习的结直肠息肉自动检测，取得了可观的判断效果。左艳等[4]提出医学影像处理智能化与人工智能手段结合。Rajpurkar 等[5]使用改进的 Dense Net[6]对 X 线胸片图像进行分类。Shen 等[7]使用卷积神经网络、随机森林与支持向量机对肺部 CT 图像进行肺结节良、恶性分类。秦喜文等[8]将深度森林算法与慢性肾病结合，众多学者为防治结直肠息肉提供医学依据。随着各行业对人工智能(artificial intelligence)需求的增加，也伴随着深度学习(Deep Learning)概念的提出，深度神经网络(Deep Neural Network，DNN)由于其强大的表征学习能力和拟合能力，成为医学等领域的研究热点之一。

DNN 也有着明显的缺陷：(1)训练通常需要大量的训练数据，即使是后来学者对于训练数据采取一系列的方式进行数据增强，但深度神经网络依旧很难直接面对数据量较小的人物，即使身处于大数据时代，许多实际任务由于大量客观因素，导致缺乏足够的标记数据量，所以深度神经网络在这些任务中的性能较差。(2)训练时间过长，在深度神经网络学习的过程中，深度神经网络是一个极为复杂的模型，在训练过程中所需硬件与软件设施的配备要求较高。

深度森林(Deep Forest)[9]一种新的决策树集成方法。此算法与深度神经网络的存在不同，该算法对训练样本数据要求不高，并且具有强大的表征学习能力，训练速度可观、参数鲁棒性好等许多优势。

人工智能技术的终极目标是让学习到的模型具有与人类相当的解决问题的能力。国内外学者对结直肠息肉研究主要集中于对结直肠息癌的诊断研究，以提高诊断成功率。本文提出利用深度森林方法对拍

收稿日期： 2021-07-29

作者简介： 陈祎琼(1982-)，女，安徽潜山人，硕士，讲师，主要从事生物信息学、图像处理方面的研究.

基金项目： 国家重点研发计划项目(2017YFD0301303)；安徽省高校自然科学研究项目(KJ2019A0211)；安徽省北斗精准农业信息工程实验室开放基金项目(AHBD201904).

摄的 CT 图像中是否含有结直肠息肉进行分类，并采用了 7 种方法进行建模比较研究，经实验研究发现使用深度森林方法在结直肠息肉 CT 图像数据集中具有较好的分类效果。

1　深度森林模型原理

深度森林是以决策树为基础构建的深层模型，是一个基于树的新的集成学习方法，与深度神经网络不同，深度森林整体包括多粒度扫描（multi-grained scanning，MGS）和级联森林（cascade forest，CF）两个阶段。在多粒度扫描阶段进行医学影像的特征学习，提取其样本特征之后，再送入级联森林进行分类预测。

1.1　随机森林

随机森林（Random Forests，RF）[10] 是一种基于集成学习的算法，是一个包含多棵相互独立决策树的分类器，将多棵决策树集成起来，其输出的类别是由个别树输出的类别的众数而决定。

随机森林预测模型最终输出结果通过投票方式获得：

$$F(X) = \arg\max_{R} \sum_{i=1}^{n} I(f_i(x) = R) \tag{1}$$

其中：$F(X)$ 为组合预测模型；f_i 为单棵决策树预测模型；R 为预测结果；$I(\cdot)$ 为示性函数；n 为 RF 中决策树的数量。

RF 算法步骤：

1）假设原始训练集为 D，利用自助采样方法有放回地抽取 k 个样本集，循环 k 次，生成 k 个样本集 $D^* = \{D_1, D_2, \ldots, D_k\}$。

2）设每个样本有 M 个原始变量，则在 M 中随机抽取 m_{try} 个变量（$m_{try} \leqslant M$），当每个决策树的节点进行分裂时，需要在 m_{try} 中选择分类能力最优的变量作为该节点的分裂属性，重复该步骤，生成 k 棵决策树；

3）由上述步骤生成的 k 棵决策树组成随机森林，随机森林的分类结果由公式（1）得出。

随机森林分类模型的实现过程如图 1 所示。

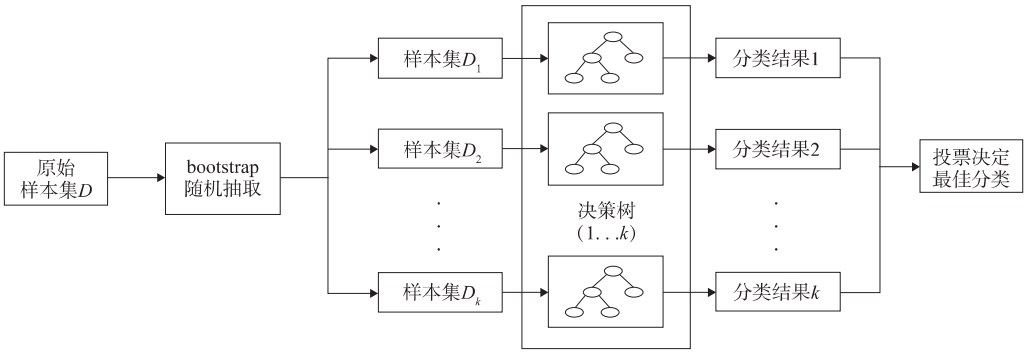


图 1　随机森林算法流程

1.2　多粒度扫描模块

多粒度扫描模块是在深度森林模型中引入了类似卷积神经网络的多个滑动窗口。MGS 通过使用不同大小的滑动窗口扫描原始输入图像，进行特征提取，生成多个维度的特征实例。将每个特征实例送入一个随机森林和一个完全随机森林进行训练提取类概率向量，最后通过拼接各森林的类概率向量得到转换特征向量。多粒度扫描模块的实现过程如图 2 所示。

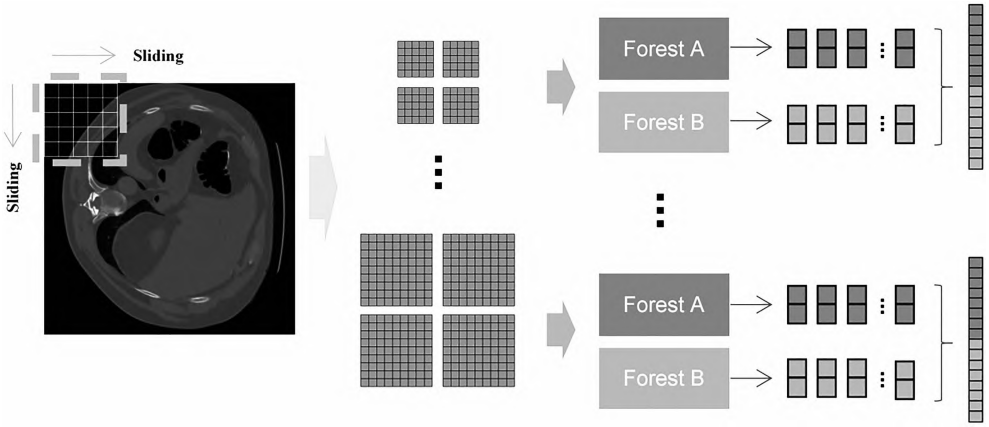


图2 多粒度扫描模块实现过程

1.3 级联森林模块

级联森林为一个拥有自主生长结构的深度树集成方法，每个级联层包含两个随机森林和两个完全随机森林，通过多个森林多层级联得出分类预测结果。相比神经网络，级联森林具有训练过程效率高、可扩展性强且对超参数不敏感等特性。

级联森林的输入特征向量是由多粒度扫描模块产生的样本转换特征向量，级联层学习训练过程中输出的类概率向量结果与原始特征向量拼接作为下一层的输入，每次扩展新的级联层后，模型会评估当前级联森林在验证集上的性能，若无显著性能增益，训练过程将终止，有效减少模型的复杂度。在最后一层将级联森林产生的所有类向量计平均值作为预测类别概率结果，其中最大值对应的类别作为深度森林模型的分类结果。

为了避免级联森林在训练过程中产生过拟合的现象，对每个森林的训练都采用 k 折交叉验证后产生的类向量。将数据分割成 k 份，其中一份作为验证数据，其他 $k-1$ 份用于训练，交叉验证重复 k 次。每个实例将产生 $k-1$ 个类向量，对其取平均值后得到下一个级联层的输入类向量。级联森林模块的实现过程如图3所示。

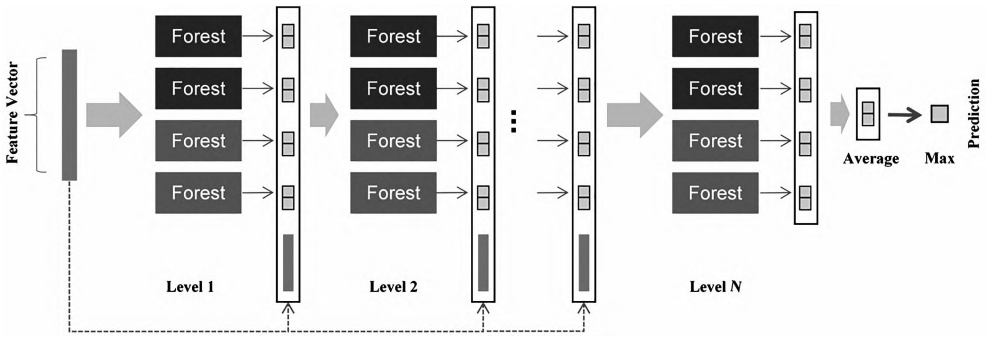


图3 级联森林模块实现过程

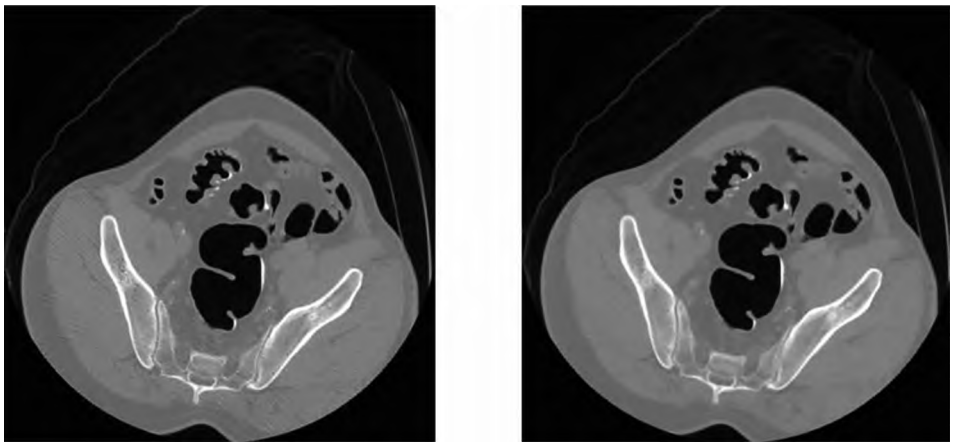
2 数据预处理

实验所使用的图像数据均来自百度 AI Studio 公开的数据集，该数据集包含了 1 500 张结直肠 CT 图像，其中 857 张存在息肉，数据按照 8∶2 的比例划分为训练集与测试集。由于数据集结直肠息肉 CT 图像的病灶区域不同、图像摄片时受到各种噪声的干扰所以本文对图像进行预处理操作[11]。预处理包含两个阶段，第一阶段对每张图片采用随机旋转 90°倍数的方法进行数据增强，旋转图片不会造成 CT 图像信息的损失，适合作为本实验数据集扩充的手段，扩充后的数据集为 2 700 张。数据集扩充结果如表1所示。

表 1　数据集扩充结果

数据集	实验样本数/张
原始数据集	1 500
扩充数据集	2 700

第二阶段对 CT 图像进行灰度化、归一化处理并使用中值滤波算法对 CT 图像的噪声进行过滤，提高图像清晰度，图像处理前后的对比如图 4 所示。



(a) 原始图像　　(b) 处理后图像

图 4　图像处理前后对比

3　实验结果及分析

3.1　评估指标

CT 图像中结直肠息肉检测为二分类问题，有 E 和 N 两个分类，分类模型中评估指标的混淆矩阵(Confusion Matrix) 计算如表 2 所示。

表 2　二分类问题混淆矩阵

真实类别	预测类别	
	E	N
E	True Positive(TP)	False Negative(FN)
N	False Positive(FP)	True Negative(TN)

由混淆矩阵计算出的模型评估指标包括分类精度(Accuracy)、召回率(Recall)、精确率(Precision)和 F_1 值，各指标定义如下:

分类精度(准确率)

$$A = \frac{TP + TN}{TP + FN + FP + TN} \tag{2}$$

召回率(查全率)

$$R = \frac{TP}{TP + FN} \tag{3}$$

精确率(查准率)

$$P = \frac{TP}{TP + FP} \tag{4}$$

F_1 值(F_1 Score)

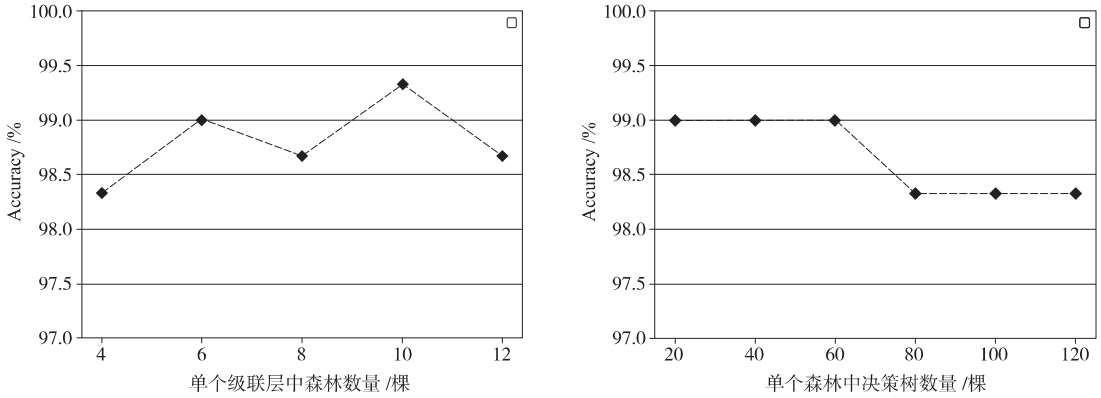
$$F_1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{n + TP - TN} \tag{5}$$

其中：TP（True Positive）为真正例；FP（False Positive）为假正例；TN（True Negative）为真反例；FN（False Negative）为假反例[12]；$TP + TN + FP + FN = n$，n 为样本容量。

分类精度可以从总体上评价模型的性能，但在医学领域并不能满足模型评估需求。对于本文结直肠息肉的识别问题，分类精度表明有多少 CT 图像被正确识别出是否存在息肉，在医学模型评估中，被模型识别出患者中确实存在息肉的患者所占比例也是评价模型性能的重要指标[13]，所以本文把 F_1 值（精确率和召回率的调和均值）作为评估指标之一。

3.2 深度森林模型调整

深度森林中级联森林模块具有良好的自适应能力，在训练过程中可以根据模型性能自行确定级联层数量[14]，需要手动调整的参数较少且对参数的敏感性低。该模块主要参数有级联层的最大数量、级联层中森林的数量（随机森林与完全随机森林）、单个森林中决策树的数量、基本估计量的选择等。级联森林模块中级联层森林数量与单个森林中决策树数量对结直肠息肉 CT 图像的分类精度如图 5 所示。



（a）森林的数量对分类精度的影响　　（b）森林中决策树数量对分类精度的影响

图 5　参数对级联森林分类精度的影响

由图 5 可以看出在本实验数据集上，模型性能整体相对稳定。在单个森林中决策树的数量增加到 120 棵时，模型的分类精度只有 98.33%，在决策树的数量设置为 20~60 时，分类精度达到了 99%，且此时模型拥有较少计算量与内存需求。增加级联层中森林的数量时，模型的性能也有所改变，当级联层中设置 10 个森林（5 个随机森林与 5 个完全随机森林）时，模型达到了 99.33%的分类精度。

综合以上实验结果以及模型实际运行效率，将级联森林模型级联层中森林的数量设置为 10 个（5 个随机森林与 5 个完全随机森林）、每个森林中决策树的数量设置为 60，并添加 XGBoost 作为模型的基本估计量，同时对级联层的最大层数不作要求。经过参数调整后的级联森林模型在测试集上的分类精度达到了 99.67%。

完整的深度森林模型包含多粒度扫描模块与级联森林模块两部分，在级联森林模块调整为上述参数后，添加多粒度扫描模块进行实验。实验所使用 CT 图片大小为 512×512，采用两种不同的窗口尺度集合与滑动步长进行实验，每个尺度对应 1 个随机森林与 1 个完全随机森林，森林中决策树数量均设置为 20，实验结果如表 3 所示。

表 3　多粒度扫描模块参数实验

多粒度扫描模块参数	分类精度/%
窗口尺度集合{143,288,430}，滑动步长 60	97.00
窗口尺度集合{143,200,266}，滑动步长 40	86.33

由表 3 可以看出，加入多粒度扫描模块后，深度森林模型分类精度最高达到了 97.00%，并且随着滑

动窗口与滑动步长的减小，模型分类精度降低。取该实验分类精度最高的模型与不使用多粒度扫描模块的深度森林模型进行性能对比，结果如图 6 所示。

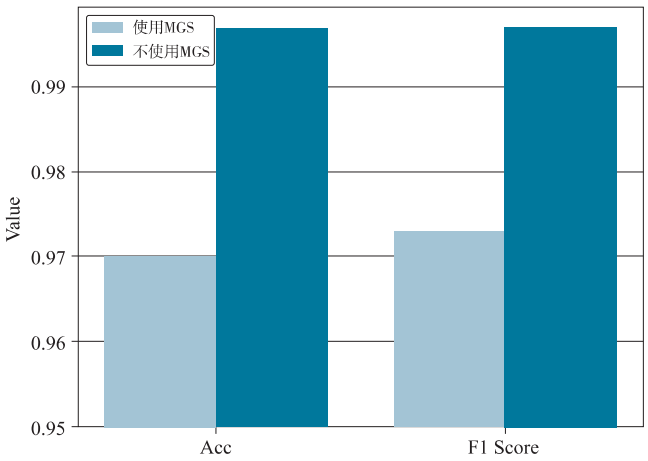


图 6　深度森林模块选择实验结果

由图 6 可以看出，在本文数据集上，使用多粒度扫描模块的模型相比只使用级联森林的模型在测试集上 A 值与 F_1 值分别下降了 2.7%与 2.4%。产生这种情况可能是因为多粒度扫描模块会产生多个只包含噪声的训练样本，对于结直肠息肉 CT 图像来说，有意义的图像仅在于息肉存在的部分，其他无意义的图像部分作为训练样本反而会造成模型泛化性能的下降，所以本文只使用深度森林模型中的级联森林模块。

3.3　不同分类模型比较

本文使用深度森林算法对结直肠息肉 CT 图像进行分类，为了比较模型的性能，采用 K-最近邻（KNN）、随机森林（RF）、卷积神经网络（CNN）、支持向量机（SVM）、BP 神经网络（BP）、朴素贝叶斯模型（NBM）与决策树（CART）7 种算法进行实验比较，实验均使用处理后的数据集。

其中 KNN、SVM 与 CART 均采用基于十折交叉验证的网格搜索方法选择最优建模参数；对 NBM 中多项式模型、高斯模型与伯努利模型 3 种常见模型采用十折交叉验证方式分别进行性能评估后，选择多项式模型进行实验；对于 CNN 中主流模型，本文使用 VGG16 进行实验，其具有很好的 CNN 特征提取能力[15]。由于数据集规模大小对 VGG16 网络训练结果有较大影响，所以在图像预处理第 1 阶段的基础上，采用翻转、旋转的手段进一步扩大数据集，且第 2 阶段不进行灰度化处理。处理后数据集容量达到7 500张，将训练集中五分之一的数据作为验证集，总共进行了 100 次训练。

表 4　不同实验方法比较结果

实验方法	A/%	R/%	P/%	F_1
KNN	99.33	100	98.81	0.994 0
RF	96.67	98.19	95.88	0.970 2
VGG16	94.67	95.18	95.18	0.951 8
SVM	99.33	100	98.81	0.994 0
BP	98.67	100	97.65	0.988 1
NBM	73.67	80.72	74.03	0.772 3
CART	91.33	93.98	90.70	0.923 1
Deep Forest	99.67	100	99.40	0.997 0

从表 4 可以看出，深度森林模型在分类精度、召回率、精确率、F_1 值均高于或等于其他 7 种分类算法。从分类精度上看，深度森林达到 99.67%，其次为支持向量机与 K-最近邻模型达到了 99.33%的精

度；从召回率上看，深度森林、BP神经网络、支持向量机、与K-最近邻模型均达到了100%，说明所有存在息肉的CT图像均被识别出来。从精确率上看，深度森林识别出存在息肉的图片中识别正确的图片占比达到99.40%，最低的为朴素贝叶斯模型，仅仅只有74.03%。

4 结 语

基于深度森林算法对结肠息肉CT图像进行识别分类，并与7种分类算法使用多项评估指标对比，结果表明该方法可提升结直肠息肉检测分类的效果。此外，深度森林在模型鲁棒性、可解释性、训练难易程度上具有优势，这使得本文提出的结直肠息肉CT图像检测方法在医学影像领域具有良好的通用性。本文只对结直肠息肉CT图像进行了分类检测，如何对息肉存在位置进行标注并对检测体系做进一步的优化，提供更加全面精准的检测报告，辅助医生根据CT影像筛查存在疾病的患者并为临床医生提供理论指导，将是下一步的研究重点。

参考文献:

[1] SOCIETY A C.Cancer Facts & Figures 2018 [M] .Atlanta: American Cancer Society,2018: 30-35.

[2] 李倩倩,王军,赵越,等.结直肠息肉发生相关危险因素的研究现状 [J] .医学综述,2020,26(16):3196-3200.

[3] 陈奕志. 基于深度学习的结肠息肉自动检测 [D] .上海: 上海交通大学,2019: 18-21.

[4] 左艳,黄钢,聂生东.深度学习在医学影像智能处理中的应用与挑战 [J] .中国图象图形学报,2021,26(2):305-315.

[5] RAJPURKAR P,IRVIN J,ZHU K,et al.Net: radiologist-level pneumonia detection on chest X-rays with deep learning [EB/OL] . [2020-09-13] .https: //arxiv.org/pdf/1608.06993.pdf.

[6] HUANG G,LIU Z,Maaten L,et al.Densely connected convolutional networks [EB/OL] . [2020-09-13] . https: //arxiv.org/pdf/1608.06993.pdf.

[7] SHEN W,ZHOU M,YANG F,et al.Multi-scale convolutional neural networks for lung nodule classification [J] . Inf Process Med Imaging,2015(24):588-599.

[8] 秦喜文,周红梅,吴睿,等.深度森林算法的慢性肾病识别 [J] .长春工业大学学报,2020,41(6):533-539.

[9] ZHOU ZHIHUA,FENG J.Deep Forest: towardsan alternative to deep neural networks [EB/OL] . [2019-5-28] .

[10] BREIMAN L.Random forests [J] .Machine Learning,2001,45(1):5-32.

[11] 刘志华,李丰军,严传波.卷积神经网络在肝包虫病CT图像诊断中的应用 [J] .电子技术应用,2019,45(11):17-20.

[12] 周志华.机器学习: Machine learning [M] .北京: 清华大学出版社,2016: 125-127.

[13] 刘文博,梁盛楠,秦喜文,等.基于迭代随机森林算法的糖尿病预测 [J] .长春工业大学学报,2019,40(6):604-611.

[14] 李淼,雷鸣,周挺,等.基于深度森林的电力系统暂态稳定评估方法 [J] .电测与仪表,2021,58(2):53-58.

[15] MANUEL L A,RUBEN G O,Nicolai P.Aprearance invariant place recognition by discriminatively training a convolutional neural-network [J] .Pattern Recognition Letters,2017(92):89-95.

Detection of Colorectal Polyps on CT Images Based on Deep Forest

CHEN Yiqiong[1,2], LIU Ao[1], FAN Guohua[1,2], BI Jiaze[1], CHEN Tao[3]

(1.School of Information and Computer, Anhui Agricultural University, Hefei 230036, China;
2.Anhui Provincial Engineering Laboratory for Beidou Precision Agricultural Information, Anhui Agricultural University, Hefei 230036, China; 3.School of Engineering, Anhui Agricultural University, Hefei 230036, China)

Abstract: To improve the screening efficiency of colorectal polyps in CT images, the CT image detection method for colorectal polyps is proposed based on deep forest. The data set is first preprocessed by means of grayscale, normalization, median filtering, and random rotation, and the processed data are then entered into an adjusted deep forest for predictive classification to obtain the output. The experiment results show that the model has a better classification effect compared with other classification algorithms under different indicators, and the classification accuracy reaches 99.67%, meanwhile, the model has fewer hyper parameters and strong generalization ability, which is helpful for assisting physicians to screen patients in the field of medical imaging.

Keywords: deep forest; medical image; colorectal polyps; image classification

(责任编辑: 陈白生)