

一种白血细胞图像训练集扩充方法

臧宇¹, 苏洋²

1. 惠州市第一人民医院血液内科, 广东 惠州 516000; 2. 东莞城市学院计算机与信息学院, 广东 东莞 523419

【摘要】针对因训练集较小导致的白血细胞图像识别精度低以及传统的扩充训练集方法需要人工介入的问题,提出一种白血细胞图像训练集扩充方法,将图像旋转任意角度后,提取因旋转产生的黑色区域边缘,然后对黑色区域进行填充,并弱化边缘特征,得到扩充训练集。实验结果表明,使用本文方法扩充训练集对ResNet50、MobileNet与ShuffleNet 3种模型进行训练后,对比原始数据集,模型的识别精度分别提高220.18%、140.84%与88.99%,且不需要人工介入。

【关键词】白血细胞识别;机器学习;训练集扩充

【中图分类号】R318;TP3

【文献标志码】A

【文章编号】1005-202X(2023)03-0342-08

A method for white blood cell image training set augmentation

ZANG Yu¹, SU Yang²

1. Department of Hematology, Huizhou First Hospital, Huizhou 516000, China; 2. School of Computer and Information, Dongguan City College, Dongguan 523419, China

Abstract: Aiming at the low recognition accuracy of white blood cell images caused by the small training set and the need for manual intervention in the traditional method of training set augmentation, a novel method is proposed for the augmentation of the training set of white blood cell images. The edges of the black area caused by the image rotation by an arbitrary angle are extracted, and the training set augmentation is realized through filling the black area and weakening the edge features. The experimental results show that after using the training set augmented by the proposed method to train ResNet50, MobileNet and ShuffleNet, comparing with the original data set, the recognition accuracies of these models are improved by an average of 220.18%, 140.84%, 88.99%, and no manual intervention is required.

Keywords: white blood cell identification; machine learning; training set augmentation

前言

在医学界,人体中的血细胞主要分为3种类型:红细胞、白血细胞与血小板^[1]。其中白血细胞在人类免疫系统中起着重要作用,且与多种疾病密切相关^[2]。白血细胞一般被划分为中性粒细胞、嗜酸性粒细胞、嗜碱性粒细胞、单核细胞与淋巴细胞^[3],这些细胞计数的变化可作为诊断某些疾病的依据或者用于评估某些疾病的治疗效果:中性粒细胞计数是肺肿瘤血栓性微血管

病(Pulmonary Tumor Thrombotic Microangiopathy, PTTM)的预测因子,它的恢复情况可以作为诊断肿瘤患者康复情况的依据^[4],用于区分患者有无脑积水^[5],用于发现COVID-19感染与血栓栓塞之间的联系^[6]等;嗜酸性粒细胞计数可作为ICU患者死亡的预测因子^[7]、用于诊断嗜酸性胃肠疾病^[8]等;单核细胞计数的变化可用于判断妊娠期是否有糖尿病、巨结肠症或炎症^[9]等;淋巴细胞计数是尼伐单抗治疗转移性肾癌患者生存的独立预测指标^[10],也可以用于判断患者是否患有感染性胰腺坏死^[11]等。因此,很有必要对白血细胞进行精确的识别与分类。

传统的白血细胞分类方法是将细胞染色,然后在显微镜下进行人工分类,该方法效率低且分类效果容易受到人为因素的影响^[2]。目前已有较多的计算机辅助方法用于白血细胞的分类与计数,早期的计算机辅助方法主要是基于形态学的,人为分析出白血细胞的形

【收稿日期】2022-12-10

【基金项目】广东省教育厅重点领域专项(2021ZDZX1029);东莞市社会科技发展(重点)项目(2020507151144)

【作者简介】臧宇,硕士,主治医师,主要研究方向:免疫治疗、医学图像处理,E-mail: zang18762130668@163.com

【通信作者】苏洋,硕士,主要研究方向:机器学习应用、大数据处理,E-mail: 417311012@qq.com

状、颜色等特征,将白血细胞从背景中分割出来,达到分类的目的,如文献[12-15]提出的方法。随着机器学习技术的发展,出现了一些基于机器学习的白血细胞分类方法,这些方法主要工作是设计模型,然后使用训练集训练,得到在训练集上表现较好的模型,使用测试集来检验模型的分类效果。比如Patil等^[16]提出的基于长短期记忆人工神经网络(Convolutional Neural Networks-Long Short Term Memory, CNN-LSTM)网络结构的典型相关分析方法;Su等^[17]使用形态学的相关操作提取出白血细胞的特征,然后带入3种神经网络达到分类的目的;Jiang等^[18]结合批量归一化算法、残差卷积结构提出一种新的CNN模型;Liang等^[19]提出一种循环卷积神经网络(Convolutional Neural Networks-Recurrent Neural Network, CNN-RNN)框架,通过递归方式充分提取图像特征,达到更好的分类精度。当有足够的计算量与数据集时,比起复杂的编程实现人工提取图像特征,机器学习可以取代手工的提取图像特征,且效率更高^[20]。

机器学习的分类效果受训练集样本数量大小的影响,当训练集的样本数量较少时,会产生欠拟合的问题^[20],导致学习精度低,分类效果差。目前,白血细胞数据的来源主要有3种:(1)自有数据集^[18];(2)Shenggan/BCCD原始数据集(以下简称BCCD_O)及Shenggan/BCCD扩充数据集(以下简称BCCD_A)^[16, 19, 21-24];(3)由DC-GAN(Deep Convolutional Generative Adversarial Network)算法生成白血细胞图像与BCCD_O或BCCD_A混合而成的数据集^[24]。自有数据集相对来说比较小,且非医疗人员很难获取,BCCD_O数据集同样很小,仅有346张图像,由DC-GAN算法生成的数据集因为并非真实的白血细胞成像,仍需人工检验生成的图像是否接近真实的白血细胞,效率低。为了解决白血细胞数据集小的问题,现在普遍采用的方法是在BCCD_O数据集基础上对图像进行旋转操作,达到扩充数据集的目的;第一种方式是对图像进行水平与垂直旋转^[24],由于旋转角度有限,对数据集的扩充也比较有限;第二种方式是对图像在0°~360°中随机选择角度旋转,即BCCD_A数

据集的获取方式,这种方式会大大扩充数据集,但是旋转后的图像会出现黑色区域,这些黑色区域存在着很明显的边界,有可能会被作为图像特征用于分类,进而影响分类效果。为了消除黑色区域的影响,有的研究人员采用了剪裁的方式,将黑色区域完全去除,仅保留白血细胞图像,这种方法需要人工介入,效率比较低;第三种方式是将图像乘以旋转矩阵^[19],这种方法会引起明显的图像形变,导致白血细胞的形态特征发生较大变化,进而影响分类精度。

在实验阶段,数据集小的影响同样存在,现在的学者普遍是将扩充后的数据集按比例划分为训练集与测试集,即验证模型效果的测试集数据是按某种方法处理过的白血细胞图像,而不是原始图像,比如使用对图像进行随机旋转的方法扩充数据集,其测试集中的图像也是经过旋转的,同样包含了旋转后留下的黑色区域。而真实情况是医疗机构首先获取到的图像都是原始图像,即未经旋转的,如果想要达到实验中的效果,则需要对原始图像进行随机旋转,这样就降低了效率。

为了解决上述问题,本文提出一种新的白血细胞图像数据集扩充方法。首先将白血细胞图像进行随机旋转,使用Canny边缘检测算法^[25],提取出图像中黑色区域的边缘;然后统计出未旋转图像中出现次数最多的像素,将旋转后图像中黑色区域用该像素填充;最后对之前提取得到的边缘上每个点进行高斯模糊,最终得到消除了黑色区域特征的图像,达到扩充白血细胞图像数据集的目的,使用该数据集对机器学习模型进行训练即可。

1 方法

本节描述了扩充白血细胞图像数据集的方法,为了最大限度保证白血细胞初始特征,本文将图像围绕中心点进行整体旋转,统计得到未旋转图像中出现次数最多的像素,使用该像素对黑色区域进行填充,然后提取出因旋转产生的黑色区域与原始图像间的边缘,对提取得到的边缘周围进行高斯模糊处理,该方法的工作流程图见图1。

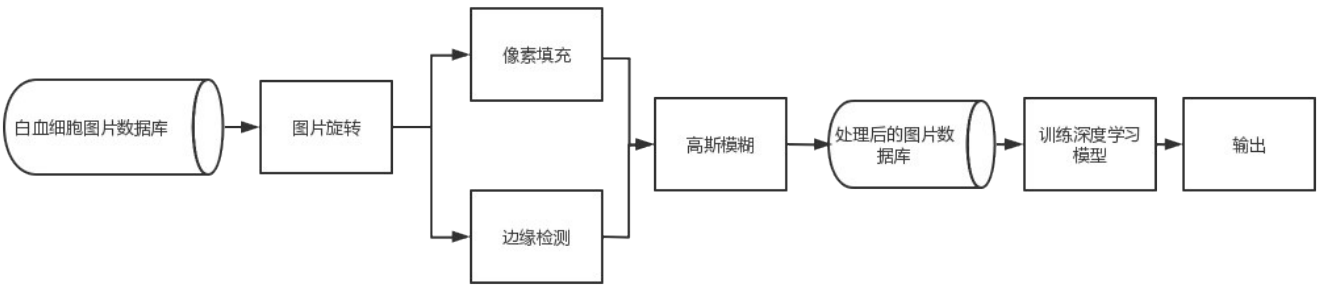


图1 数据集扩充方法概述

Figure 1 Overview of the method for data set augmentation

1.1 图像旋转

本文的旋转方法是以图像中心点为轴,旋转特定角度,公式见式(1):

$$\begin{bmatrix} x_0 & y_0 & 1 \end{bmatrix} = \begin{bmatrix} x & y & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ -0.5W\cos\theta - 0.5H\sin\theta + 0.5W & 0.5W\sin\theta - 0.5H\cos\theta + 0.5H & 1 \end{bmatrix} \quad (1)$$

其中, x_0 、 y_0 、1分别表示旋转后像素点的横坐标、纵坐标及维度, x 、 y 、1代表旋转前的横坐标、纵坐标及维度, W 、 H 分别代表图像的宽与高, θ 代表旋转角度。旋转前和旋转后(IMG_Rotated)的图像见图2、图3。

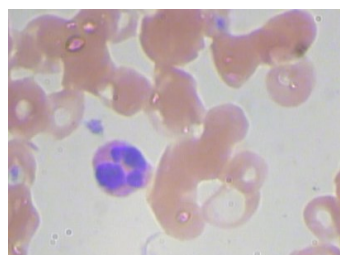


图2 原始图像
Figure 2 Original image

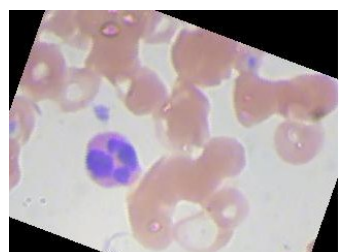


图3 旋转后图像
Figure 3 Image after rotation

1.2 图像填充

白血细胞图像中存在着大量颜色单调的背景,见图4,其中方框内即是背景。本文首先统计得到原图像中出现次数最多的像素值,然后将旋转后图片中出现的黑色区域用该像素值替换,以达到消除黑色区域的目的,像素填充后的图像(IMG_Filled)效果见图5。



图4 白血细胞图像背景示意图
Figure 4 White blood cell image background

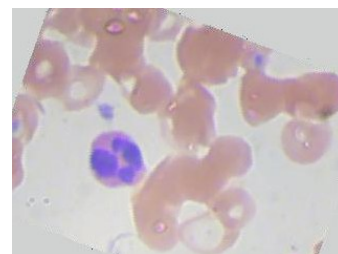


图5 像素填充后的白血细胞图像
Figure 5 Pixel filled white blood cell image

1.3 边缘提取

本文使用Canny算法对IMG_Rotated进行边缘提取,分为以下几个步骤。

(1)高斯滤波。对于位置在 (x,y) 的像素点,其灰度值为 $f(x,y)$,高斯滤波后的灰度值变为:

$$g_{\sigma}(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \cdot f(x,y) \quad (2)$$

(2)计算梯度值和梯度方向。分别计算水平与垂直方向的梯度,综合得到最终梯度值与梯度方向,见式(3)、式(4):

$$G(x,y) = \sqrt{g_x(x,y)^2 + g_y(x,y)^2} \quad (3)$$

$$\theta = \arctan\left(\frac{g_y(x,y)}{g_x(x,y)}\right) \quad (4)$$

其中, $g_x(x,y)$ 与 $g_y(x,y)$ 分别为水平与垂直方向梯度。

(3)过滤非最大值。

(4)使用上下阈值来检测边缘。设置两个阈值,分别为maxVal和minVal。图中像素值大于maxVal的点都被检测为边缘,而低于minVal的点都被检测为非边缘。对于中间的像素点,如果与确定为边缘的像素点邻接,则判定为边缘;否则为非边缘。使用Canny算法提取出的边缘见图6。

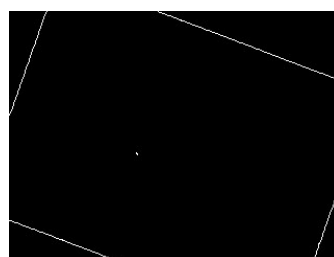


图6 黑色区域边缘提取
Figure 6 Black area edge extraction

1.4 高斯滤波

在IMG_Filled中以上述提取到的边缘点为中心,进行高斯滤波,二维高斯滤波函数见式(5):

$$h(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (5)$$

其中, (x,y) 为像素点的坐标, σ 是标准差, 确定高斯模板窗口大小 N , 然后将边缘像素作为中心点, 进行高斯滤波即可, 高斯滤波后的图像见图 7。

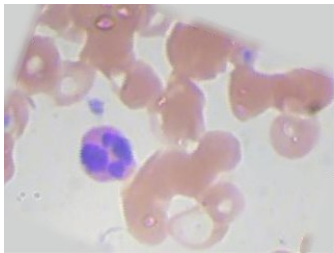


图 7 对边缘进行高斯模糊后的图像
Figure 7 Image after Gaussian blurring of edges

将 BCCD_O 数据集进行上述 4 步操作后, 即可得到新的白血细胞图像的扩充数据集。

2 结果与分析

2.1 实验环境

本文使用的实验环境如下所示, CPU: Inter Core i5-11600KF; RAM: 16 G; GPU: NVIDIA GEFORCE 1050Ti; Operating System: Windows10 64 bit; Python: 3.9; Tensorflow: 2.6.0。

2.2 实验数据集

本文使用 3 组数据集, 分别为 BCCD_O、

BCCD_A 与本文算法处理后的图像数据集, 为了还原真实应用场景, 在训练集与测试集划分上, BCCD_O 按细胞类型以 9:1 比例随机划分为训练集与测试集, BCCD_A 与本文算法处理后的图像数据集作为训练集时, BCCD_O 作为测试集, 这样可以确保在验证效果时, 使用的测试图像均是原始未经处理的白血细胞图像。数据集见表 1, 数据集中包含 4 种不同类型白血细胞, 见图 8, 3 组数据集中包含的图像示例见图 9。

表 1 数据集
Table 1 Data sets

数据集		BCCD_O	BCCD_A	本文算法扩充后的数据集
训练集	嗜酸性粒细胞	79	2 497	2 497
	淋巴细胞	30	2 483	2 483
	单核细胞	18	2 478	2 478
	中性粒细胞	185	2 499	2 499
	合计	312	9 957	9 957
测试集	嗜酸性粒细胞	8	87	87
	淋巴细胞	3	33	33
	单核细胞	2	20	20
	中性粒细胞	21	206	206
	合计	34	346	346

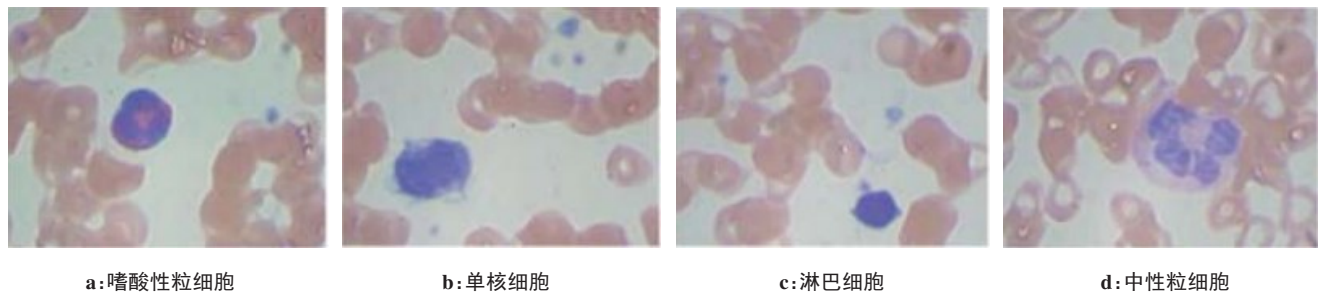


图 8 4 种不同类型的白血细胞图
Figure 8 Four different types of white blood cells

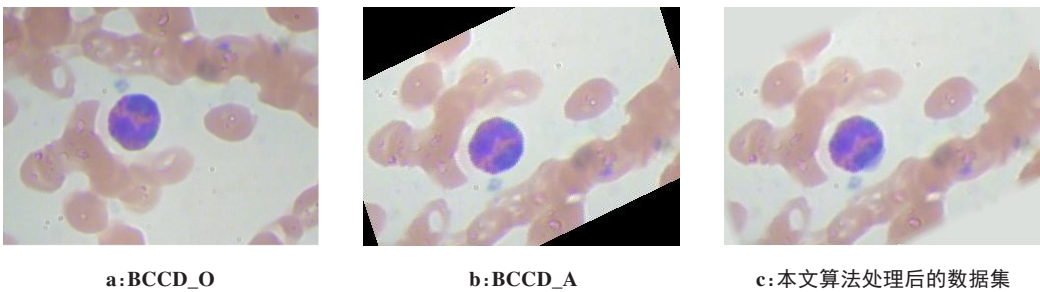


图 9 3 组数据集图像
Figure 9 Images from 3 data sets

2.3 评测指标

本文使用交叉熵损失函数、精度(PA)、准确率(P)与召回率(R)4个参数作为评价指标:

$$VA = \frac{TruePositive + TrueNegative}{TotalSamples} \quad (6)$$

$$P = \frac{TruePositive}{TruePositive + FalsePositive} \quad (7)$$

$$R = \frac{TruePositive}{TruePositive + FalseNegative} \quad (8)$$

其中, TruePositive、TureNegative、FalsePositive、FalseNegative与TotalSamples分别代表检测出来的真阳样本数、真阴样本数、假阳样本数、假阴样本数与总样本数量。

2.4 实验结果与讨论

本文使用ResNet50^[26]、MobileNet^[27]与ShuffleNet^[28]3种模型来验证本文算法,各模型的损失值、精度、准确率与召回率验证结果见图10~图12。

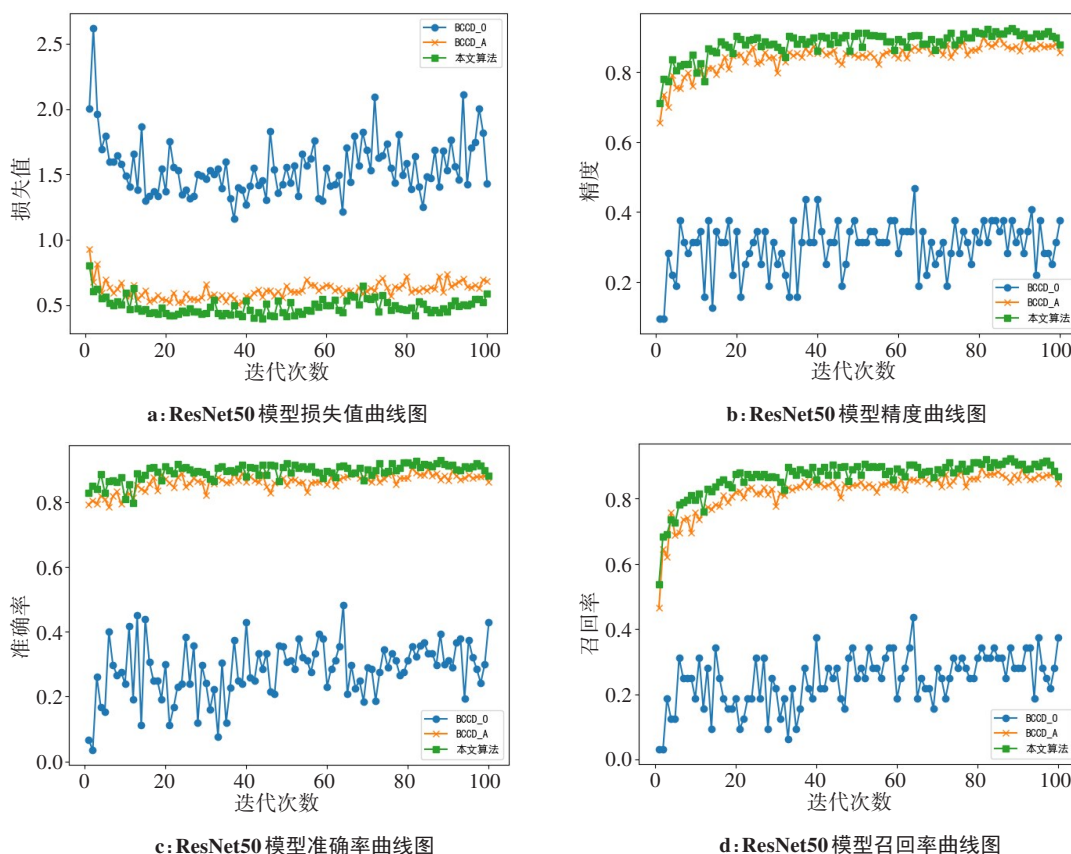


图10 ResNet50模型验证结果

Figure 10 Verification results of ResNet50

由图10a、图11a可看出,使用本文提出的算法生成的数据集训练ResNet50与MobileNet后,测试的损失值曲线平稳,可以获得较好的收敛,且损失值最小。由图12a可看出,本文提出的算法生成的数据集训练ShuffleNet模型后,虽然损失值曲线波动幅度较BCCD_O训练后的损失值曲线波动大,但损失值多数情况下是最小的。由图10a可看出,在使用BCCD_O训练ResNet50时,测试的损失值无法获得较好的收敛,精度、准确率与召回率均有较大的波动,无法获得稳定的预测效果。由图11a可以看出,BCCD_O在训练MobileNet模型时,在13代训练时损失函数发生了梯度爆炸,损失值陡升,其精度、准确

率与召回率均有大幅下降。由图12a可看出,使用BCCD_A训练ShuffleNet模型时,在73代训练时,损失函数发生了梯度爆炸,损失值陡升,其精度、准确率与召回率均出现较大幅度的下降。纵观图10~图12,使用本文算法生成的数据集训练3种模型没有发生梯度爆炸的情况,在经过多代的训练后,损失值、精度、准确率与召回率均可以达到比较稳定的状态,具有较好的鲁棒性,且在训练ResNet50、MobileNet模型时,损失值最小,精度、准确率与召回率值最大,在训练ShuffleNet模型时,损失值大多数时间也是最小的,且精度、准确率与召回率最大。

笔者使用式(8)对损失值、精度、准确率与召回

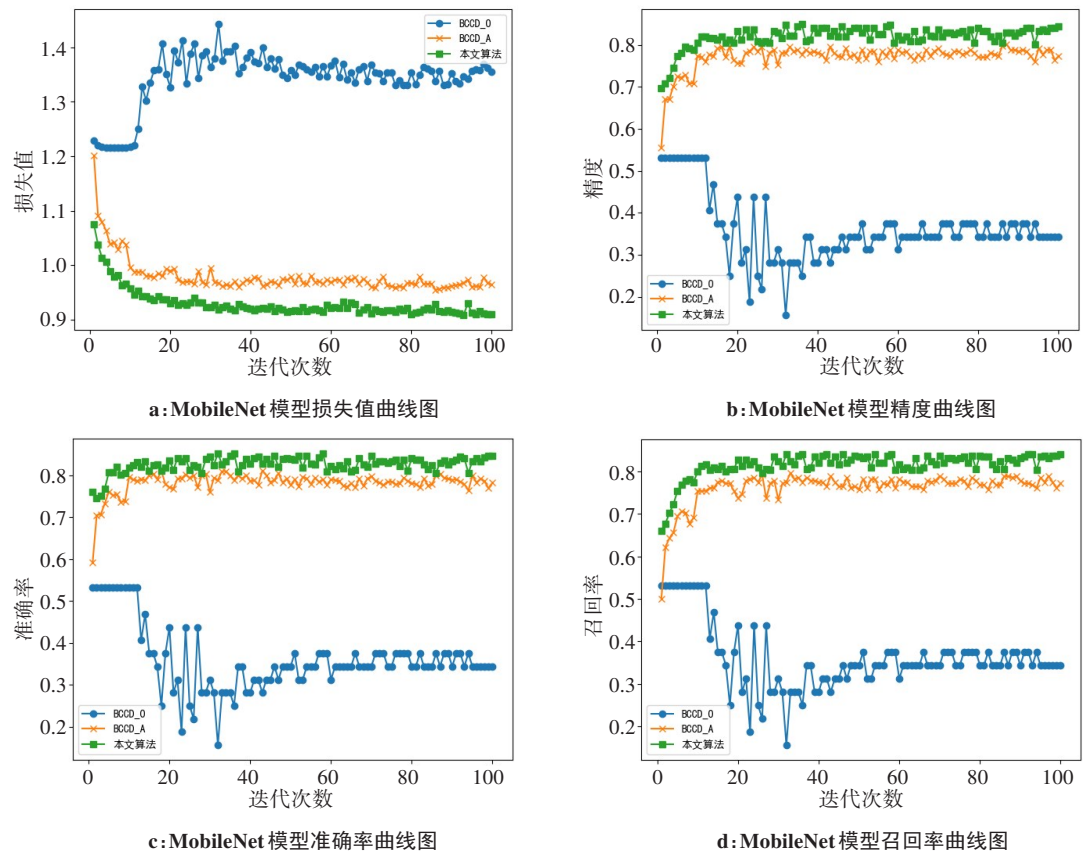


图 11 MobileNet 模型验证结果
Figure 11 Verification results of MobileNet

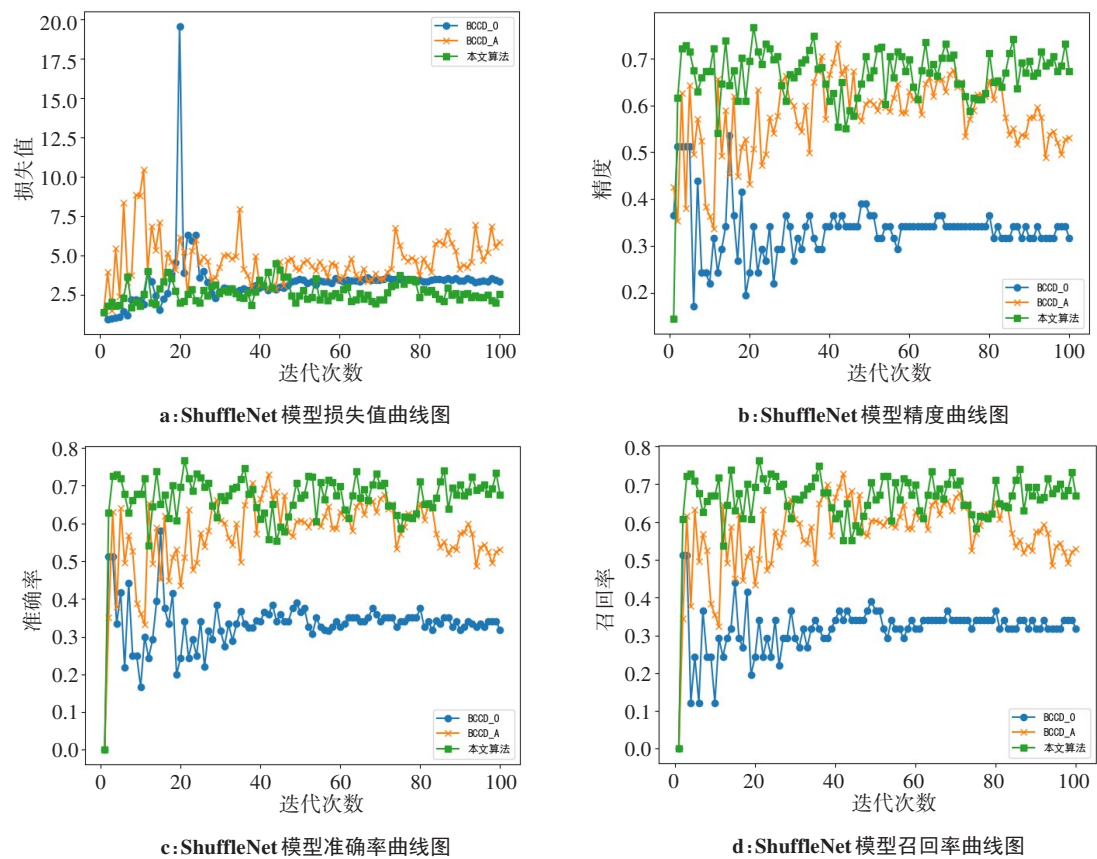


图 12 ShuffleNet 模型验证结果
Figure 12 Verification results of ShuffleNet

率的提高做一个定量的评估,以BCCD_O训练3种模型得到的损失值、精度、准确率与召回率为基准,计算通过100代训练后,各项参数提高的平均值,其中 Pa_n 为使用BCCD_A或本文算法生成的数据集在第 n 代时损失值、精度、准确率与召回率, B_n 为使用BCCD_O在第 n 代时损失值、精度、准确率与召回率,epochs为总的训练代数,结果见表2~表4。

$$I = \frac{\sum_{n=1}^{\text{epochs}} (Pa_n - B_n) / B_n}{\text{epochs}}$$

(8)

由表2~表4可以看出,BCCD_A与BCCD_O相比,除了在ShuffleNet模型上损失值没有减少,精度、准确率与召回率均提升明显。使用本文算法生成的数据集来训练3种模型,损失值、精度、准确率与召回率较BCCD_A又有了进一步提升。

表2 ResNet50提升效果(%)
Table 2 Improvements by ResNet50 (%)

参数	使用BCCD_A数据集	使用本文算法扩充后的数据集
损失值	-60.00	-70.46
精度	204.02	220.18
准确率	262.51	278.34
召回率	305.57	331.65

表3 MobileNet提升效果(%)
Table 3 Improvements by MobileNet (%)

参数	使用BCCD_A数据集	使用本文算法扩充后的数据集
损失值	-26.90	-30.69
精度	124.90	140.84
准确率	127.62	143.66
召回率	122.22	138.87

表4 ShuffleNet提升效果(%)
Table 4 Improvements by ShuffleNet (%)

参数	使用BCCD_A数据集	使用本文算法扩充后的数据集
损失值	61.64	34.59
精度	77.11	88.99
准确率	75.89	87.98
召回率	86.27	99.91

3 结束语

基于机器学习的白血细胞图像识别在临床上有着重要的意义,但对于非医疗从业人员,用于训练与

学习的白血细胞图像数据集获取困难,而数据集的大小影响着模型的训练与验证效果。本文提出一种白血细胞数据集扩充方法,利用图像旋转后出现的黑色区域边缘与原图像像素统计以及填充黑色区域,实现降低因旋转产生的黑色区域作为特征影响分类效果的可能。实验证明,使用本文方法得到的数据集用于ResNet、MobileNet与ShuffleNet训练,训练得到的模型具有较好的鲁棒性,且预测精度有着明显的提高。

本文方法的主要思路是利用白血细胞图像有着大量单调背景,图像旋转会产生黑色区域,而该区域与原图像之间存在明显边缘,以边缘为出发点对黑色区域进行像素填充,后续可对其他领域有着相同特征的图像进行研究,研究该方法在其它领域应用的可能性,而且本文没有涉及机器学习模型改进,后续模型改进也是研究方向之一。

【参考文献】

[1] Young IT. The classification of white blood cells[J]. IEEE Trans Biomed Eng, 1972, 19(4): 291-298.

[2] Duan YF, Wang JS, Hu MH, et al. Leukocyte classification based on spatial and spectral features of microscopic hyperspectral images[J]. Opt Laser Technol, 2019, 112: 530-538.

[3] Bikheth SF, Darwish AM, Tolba HA, et al. Segmentation and classification of white blood cells [C]//2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2000, 4: 2259-2261.

[4] Picca A, Wahlquist AE, Hudspeth M. Incorporating absolute phagocyte count with absolute neutrophil count as a measure for safe discharge for pediatric oncology febrile neutropenia: a pilot study[J]. J Pediatr Hematol Oncol, 2021, 43(7): e1000-e1002.

[5] Cuoco JA, Guillems EL, Klein BJ, et al. Neutrophil count on admission predicts acute symptomatic hydrocephalus after aneurysmal subarachnoid hemorrhage[J]. World Neurosurg, 2021, 156: e338-e344.

[6] Vallecillo G, Marti-Bonany J, Robles MJ, et al. Transient drop in the neutrophil count during COVID-19 regardless of clozapine treatment in patients with mental illness[J]. Rev Psiquiatr Salud Ment (Engl Ed), 2022, 15(2): 134-137.

[7] Xuan W, Jiang XL, Huang LL, et al. Predictive value of eosinophil count on COVID-19 disease progression and outcomes, a retrospective study of Leishenshan Hospital in Wuhan, China[J]. J Intensive Care Med, 2022, 37(3): 359-365.

[8] Reed CC, Genta RM, Youngblood BA, et al. Mast cell and eosinophil counts in gastric and duodenal biopsy specimens from patients with and without eosinophilic gastroenteritis [J]. Clin Gastroenterol Hepatol, 2021, 19(10): 2102-2111.

[9] Huang XM, Zha BB, Zhang MN, et al. Decreased monocyte count is associated with gestational diabetes mellitus development, macrosomia, and inflammation[J]. J Clin Endocrinol Metab, 2022, 107(1): 192-204.

[10] Ueda K, Suekane S, Kurose H, et al. Absolute lymphocyte count is an independent predictor of survival in patients with metastatic renal cell carcinoma treated with nivolumab[J]. Jpn J Clin Oncol, 2022, 52(2): 179-186.

[11] Zhou J, Chen WS, Liu Y, et al. Trajectories of lymphocyte counts in the early phase of acute pancreatitis are associated with infected pancreatic necrosis[J]. Clin Transl Gastroenterol, 2021, 12(9): e00405.

[12] Azam B, Qureshi RJ, Jan Z, et al. Color based segmentation of white blood cells in blood photomicrographs using image quantization[J]. Res J Recent Sciences, 2014, 3(4): 34-39.

[13] Alreza ZK, Karimian A. Design a new algorithm to count white blood

- cells for classification leukemic blood image using machine vision system [C]//2016 6th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2016: 251-256.
- [14] Safuan SN, Tomari R, Zakaria WN, et al. White blood cell counting analysis of blood smear images using various segmentation strategies [C]//AIP Conference Proceedings. AIP Publishing LLC, 2017, 1883 (1): 020018.
- [15] Safuan SN, Tomari MR, Zakaria WN. White blood cell (WBC) counting analysis in blood smear images using various color segmentation methods[J]. Measurement, 2018, 116: 543-555.
- [16] Patil AM, Patil MD, Birajdar GK. White blood cells image classification using deep learning with canonical correlation analysis [J]. IRBM, 2021, 42(5): 378-389.
- [17] Su MC, Cheng CY, Wang PC. A neural-network-based approach to white blood cell classification[J]. Sci World J, 2014. Doi: 10.1155/2014/796371.
- [18] Jiang M, Cheng L, Qin F, et al. White blood cells classification with deep convolutional neural networks[J]. Int J Pattern Recognit Artif Intell, 2018, 32(9): 1857006.
- [19] Liang G, Hong H, Xie W, et al. Combining convolutional neural network with recursive neural network for blood cell image classification[J]. IEEE Access, 2018, 6: 36188-36197.
- [20] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems. 2012: 1097-1105.
- [21] Zhang H, Zhang L, Jiang Y. Overfitting and underfitting analysis for deep learning based end-to-end communication systems [C]//2019 11th International Conference on Wireless Communications and Signal Processing (WCSP). IEEE, 2019: 1-6.
- [22] Zhao JW, Zhang MS, Zhou ZH, et al. Automatic detection and classification of leukocytes using convolutional neural networks[J]. Med Biol Eng Comput, 2017, 55(8): 1287-1301.
- [23] Islam MT, Alam MM. A machine learning approach of automatic identification and counting of blood cells[J]. Healthc Technol Lett, 2019, 6(4): 103-108.
- [24] Li M, Shuai RJ, Ran XM, et al. Combining DC-GAN with ResNet for blood cell image classification[J]. Med Biol Eng Comput, 2020, 58 (6): 1251-1264.
- [25] Canny J. A computational approach to edge detection[J]. IEEE Trans Pattern Anal Mach Intell, 1986, 8(6): 679-698.
- [26] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770-778.
- [27] Howard AG, Zhu M, Chen B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv: 1704.04861, 2017.
- [28] Ma N, Zhang X, Zheng HT, et al. Shufflenet v2: practical guidelines for efficient cnn architecture design [C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 116-131.
- (编辑: 薛泽玲)