

·临床信息学·

张亚洲<sup>1</sup>,李智伟<sup>2</sup>,农卫霞<sup>3</sup>,雷 伟<sup>1</sup>,摆文丽<sup>1</sup>,李寅臻<sup>1</sup>,李 瑞<sup>1</sup>,王 奎<sup>1</sup>

(1.石河子大学医学院预防医学系,新疆 石河子 832000;

2.新疆维吾尔自治区人民医院临床检测中心,新疆 乌鲁木齐 830001;

3.石河子大学医学院风湿血液科,新疆 石河子 832000)

**摘要:**目的 探索深度学习模型在流式细胞检测报告结果部分文本资料上的分类效果。方法 使用 CNN、LSTM 等六种深度学习模型对流式检测报告的结果部分的文字资料进行分析,并对急性白血病患者进行分类预测,最后通过综合指标 F1 值对模型进行评价。结果 CNN-BiLSTM 混合模型的精确率、召回率、F1 值最优,分别为 0.7422、0.7365、0.7361;模型在正常人、急性髓系白血病、急性 B 淋巴细胞白血病、有核红细胞异常、中性粒细胞异常、浆细胞异常、单核细胞异常这 7 类的 F1 值均达到了 70%。结论 混合模型对流式细胞术检测报告结果部分的文本资料的分类效果较好,可与前期研究联合共同构建了一个更为完整的流式细胞术自动化分析体系,进一步提高流式细胞术分析的效率 and 准确性。

**关键词:**流式细胞术;文本分类;CNN;自动化分析;深度学习

中图分类号:TP311

文献标识码:A

DOI:10.3969/j.issn.1006-1959.2025.08.003

文章编号:1006-1959(2025)08-0015-06

## Automatic Classification of Text Data for Flow Cytometry Detection of Acute Leukemia Based on Deep Learning

ZHANG Yazhou<sup>1</sup>, LI Zhiwei<sup>2</sup>, NONG Weixia<sup>3</sup>, LEI Wei<sup>1</sup>, BAI Wenli<sup>1</sup>, LI Yinzen<sup>1</sup>, LI Rui<sup>1</sup>, WANG Kui<sup>1</sup>

(1.Department of Preventive Medicine, Shihezi University School of Medicine, Shihezi 832000, Xinjiang, China;

2.Clinical Testing Center, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi 830001, Xinjiang, China;

3.Department of Rheumatology and Hematology, Shihezi University School of Medicine, Shihezi 832000, Xinjiang, China)

**Abstract:** **Objective** To explore the classification effect of deep learning model on text data of flow cytometry report results. **Methods** Six deep learning models such as CNN and LSTM were used to analyze the text data of the results of the flow cytometry report, classify and predict the patients with acute leukemia, and finally evaluate the model by the comprehensive index F1 score. **Results** The precision, recall and F1 score of the CNN-BiLSTM mixed model were the best, which were 0.7422, 0.7365 and 0.7361, respectively, and the F1 score of the model reached 70% in seven categories: normal humans, acute myeloid leukemia, acute B lymphoblastic leukemia, nucleated red blood cell abnormalities, neutrophil abnormalities, plasma cell abnormalities and monocytic abnormalities. **Conclusion** The mixed model has a good effect on the classification of text data in the results of flow cytometry test report, and can be combined with previous studies to build a more complete automated flow cytometry analysis system to further improve the efficiency and accuracy of flow cytometry analysis.

**Key words:** Flow cytometry; Text classification; CNN; Automated analysis; Deep learning

流式细胞术(flow cytometry, FCM)是一种能精确且快速分析细胞或者生物微粒理化性质的检测技术,被业内称为生物实验室的“CT”<sup>[1-3]</sup>。目前多数的研究都侧重于使用机器学习方法实现医学文本分类的自动化<sup>[4]</sup>。随着 FCM 技术的广泛应用,流式细胞仪检测能力得到了显著提升,但也带来了海量的数据,大大加重了流式细胞检测实验室检验人员的工作负

荷<sup>[5]</sup>。然而,要培养一个合格的流式细胞术分析师却需要较长的时间。为了解决这一问题,前期研究提出了流式细胞数据分析全程自动化的想法<sup>[6,7]</sup>,并利用机器学习方法复现了人工分析的全过程。该过程不仅包括数据的补偿<sup>[8]</sup>、转化、去粘连细胞、去细胞碎片以及对细胞聚类的自动化<sup>[9,10]</sup>,也对急性白血病患者

的多管数据细胞亚群进行了统一标注,实现了对细胞主要亚群统计描述的自动化。然而,前期研究的自动化分析主要集中在对流式细胞仪报告数据的处理和统计描述上,并未涉及流式检测报告的结果部分的文字资料。为了实现流式检测报告的自动化,除了前期研究中直接利用流式细胞仪报告数据进行分

基金项目:国家自然科学基金项目(编号:81860374)

作者简介:张亚洲(1995.12-),男,山西运城人,硕士研究生,主要从事流行病与卫生统计学研究

通讯作者:王奎(1968.3-),男,重庆人,博士,硕士生导师,副教授,主要从事卫生统计学研究

析和自动化分类外,将流式细胞检测报告中结果部分和结论部分文字资料的分析作为数字资料分析结果的补充也是有益的。因此,本研究提出以流式检测报告结果部分的文字资料为输入,以结论部分为分类依据,利用深度学习方法训练模型,对检测报告结果部分的文字资料进行分类预测,以期更好的对流式细胞检测报告结果部分的文字资料进行分析并

对急性白血病患者进行分类,现报道如下。

## 1 资料与方法

1.1 模型设计与方法 本研究所用模型如图 1 所示,包含词嵌入层、CNN<sup>[11]</sup>层、Bi-LSTM<sup>[12]</sup>层和 softmax 层。使用 one-hot 词嵌入来自定义 embedding 的权重矩阵,方便矩阵输入模型。

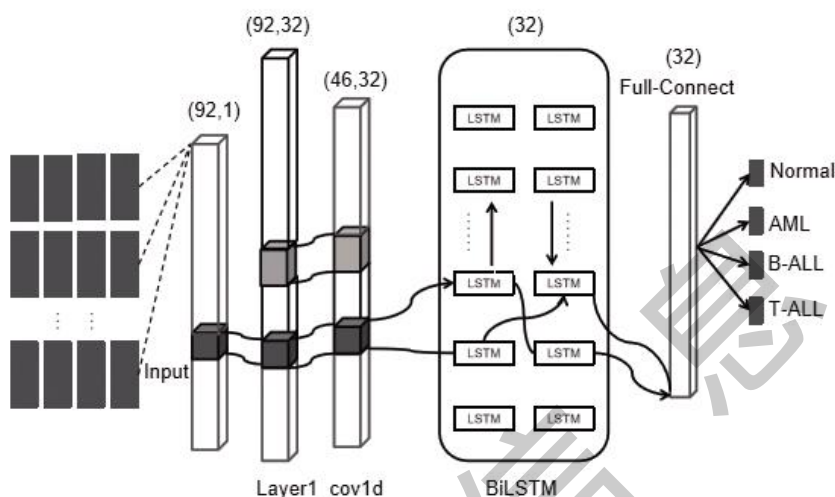


图 1 CNN-BiLSTM 网络结构

1.2 卷积神经网络 卷积神经网络(convolutional neural network, CNN)主要由输入层(input layer)、卷积层(con-volution layer)、池化层(pooling layer)、全连接层(fullconnect layer)组成。将文本数据输入 CNN 层后,卷积层通过在文本表示矩阵上上下滑动来对文本数据进行特征提取,得到的完整局部特征向量。卷积后的向量维度较高,还需要进行池化来对向量维度降低<sup>[13]</sup>,再利用全连接层将池化后的特征向量拼接成新的特征向量,输出表征更加丰富的局部特征并用于分类。

1.3 双向长短期记忆网络(BiLSTM) 双向循环网络由 1 个正向 LSTM<sup>[14]</sup>和 1 个反向 LSTM 构成<sup>[15]</sup>。LSTM 只保留过去的信息,而 BiLSTM 同时保存了过去和将来的信息。BiLSTM 层由遗忘门、输入门、输出门组成,这三个门的存在缓解了神经网络在处理中远距离依赖的序列数据中发生梯度弥散现象。为了充分发现当前时刻与前一时刻和后一时刻的联系,本研究拼接前向(forward)LSTM 和后向(backward)LSTM 形成 BiLSTM,来进一步挖掘流式细胞术数据的全局特征。利用 BiLSTM 模型提取词的上下文语义信息,提取文本中词的全局特征后,进入全连接层。该全连接层归纳全局的隐状态的输出,即向量融

合后通过全连接层。最后,使用 Softmax 激活函数进行分类,找到概率最大的标签作为预测的分类结果。

## 1.4 实验环境及数据

1.4.1 实验环境与模型参数设置 采用 python3.6.8 开发工具,第三方库选用 TensorFlow1.12.0 和 Keras2.2.4 版本进行模型训练。采用 Windows11 家庭中文版 64 位操作系统,处理器为 Intel(R)Core(TM)i5-12500H。模型参数设置如下:词向量的维度为 32,CNN 卷积核尺寸设为 3,步长为 1,见表 1。

表 1 模型的主要参数

Input(92, 1)	描述
ConvID1	32
MaxPool1D	2
BiLSTM1_hidden	32
BiLSTM2_hidden	32
LSTM	32
Loss	Categorical_crossentropy
Optimizer	Adam
Batch_size	128
Dropout	0.25
学习率	0.0001
激活函数	Softmax

1.4.2 数据来源及预处理 数据来源于新疆维吾尔自治区人民医院流式实验室 2019 年 6 月-2021 年 12 月流式细胞检测报告结果部分的文字资料,以人工诊断的结论作为金标准。数据如图 2 所示。将 2019 年和 2021 年的数据划分为训练集和验证集,2020 年的数据作为外部测试集。按金标准将数据分为正常人、急性髓系白血病(AML)、急性 T 淋巴细

胞白血病(B-ALL)、急性 B 淋巴细胞白血病(T-ALL)、有核红细胞异常、成熟 T 淋巴细胞异常、成熟 B 淋巴细胞异常、嗜碱性粒细胞异常、嗜酸性粒细胞异常、中性粒细胞异常、浆细胞异常、单核细胞异常共 12 类,本研究将 2019 年和 2021 年的数据合并后按 7:3 划分训练集和验证集,见表 2。

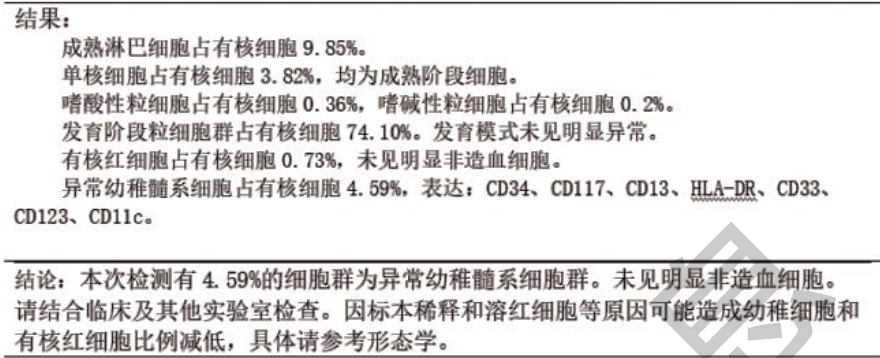


图 2 部分数据展示

表 2 实验数据划分

项目	正常人	AML	B-ALL	T-ALL	有核 红细胞 异常	成熟 T 细胞 异常	成熟 B 细胞 异常	嗜碱性 粒细胞 异常	嗜酸性 粒细胞 异常	中性 粒细胞 异常	浆 细胞 异常	单核 细胞 异常	合计
训练集	176	77	47	9	68	31	27	40	41	247	14	105	882
验证集	75	32	19	4	28	12	11	16	17	105	6	45	370
测试集	120	66	34	5	66	17	27	41	22	186	13	90	687

1.4.3 实验评价指标 选用精确率 (precision, P)、召回率(recall, R)和 F1 值作为文本分类的评价指标,计算所需的混淆矩阵见表 3。其中,TP 表示真正例,指的是实际为正例且被模型预测为正例的样本数量;FN 表示假负例,指的是实际为正类,但被模型错误预测为负类的样本数;FP 表示假正例,指的是实际为负类,但被模型错误预测为正类的样本数;TN 表示真负例,指的是实际为负例,且被模型预测为负例的样本数。精确率 P 指在所有预测为正类的样本中,实际为正类的比例,它反映了模型预测正例的能力。精确率计算公式如式(1)所示:

$$P=\frac{TP}{TP+FP}$$
 (1)

召回率 R 指在所有实际为正类的样本中,被正确预测为正类的比例,它反映了模型的完整性和灵

敏度。召回率计算公式如式(2)所示:

$$R=\frac{TP}{TP+FN}$$
 (2)

精确率和召回率通常存在一定的矛盾关系:提高精确率可能会降低召回率,反之亦然。为了平衡精确率和召回率,引入了 F1 值,它是精确率和召回率的调和平均。F1 值计算公式如式(3)所示:

$$F1=\frac{2PR}{P+R}$$
 (3)

本研究将流式细胞检测报告结果部分的文本资料分为 12 类,分别将每个类别视为“正类”,该类之外的其他所有类别则视为“负例”,根据混淆矩阵计算一个该类的精确率和召回率,从而评估模型在该特定类别上的表现。

表 3 混淆矩阵

项目	预测为正例	预测为负例
实际为正例	TP	TN
实际为负例	FP	FN

## 2 结果

**2.1 模型训练过程** 模型在训练过程中设置了调停函数,在损失函数不再下降时停止训练并保存训练结果。在训练过程中,损失函数随着迭代次数的增加逐渐下降,表明模型在持续从训练集中学习有用特征。而准确率在迭代过程的逐渐上升并趋于稳定,说明模型拟合效果较好。

**2.2 效果验证** 表 4 为不同模型在数据集的整体效果对比,结果显示本研究所选模型 CNN-BiLSTM 的精确率、召回率、F1 值以及 AUC 值都最高,分别为 0.7422、0.7365、0.7361、0.80,提示本文模型流式细胞术检测报告结果部分的文本资料具有较好的分类效

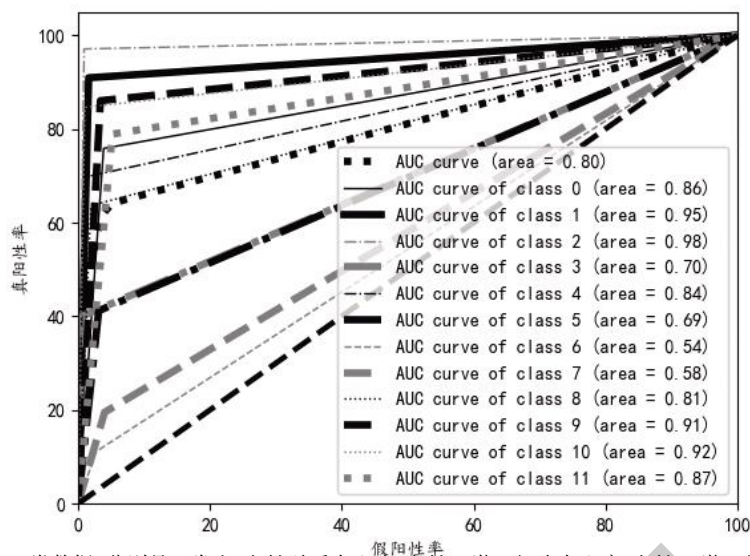
果。表 5 为 CNN-BiLSTM 模型在各类别分类效果对比,结果显示本研究模型在正常人、急性髓系白血病、急性 B 淋巴细胞白血病、有核红细胞异常、中性粒细胞异常、浆细胞异常、单核细胞异常这 7 类的 F1 值均达到了 70%,且本研究模型对急性 B 淋巴细胞白血病这一类的分类效果最好,F1 值达到了 0.9041。图 3 为 CNN-BiLSTM 模型的 ROC 曲线图,结果显示本研究模型对 12 个类的 AUC 值均大于 0.5,且急性髓系白血病、急性 T 淋巴细胞白血病、中性粒细胞异常、浆细胞异常这 5 类的 AUC 值均大于 0.9,对急性 T 淋巴细胞白血病的 AUC 值最大,为 0.98。

表 4 不同模型在测试集上的分类效果对比

项目	精确率	召回率	F1 值	AUC 值
CNN	0.6218	0.6376	0.6236	0.73
LSTM	0.5245	0.5488	0.5154	0.71
BiLSTM	0.52	0.5371	0.5213	0.73
LSTM-CNN	0.6924	0.6929	0.6889	0.77
CNN-LSTM	0.6957	0.6885	0.6837	0.77
CNN-BiLSTM	0.7422	0.7365	0.7361	0.80

表 5 CNN-BiLSTM 模型在各类别分类效果对比

分类	召回率	特异度	精确率	阴性预 测率(NPV)	误识别 率(FPR)	拒识率 (FNR)	误发现 率(FDR)	准确率 (ACC)	F1 值
正常人	0.7583	0.9612	0.8053	0.9495	0.0388	0.2417	0.1947	0.7583	0.7811
急性髓系白血病	0.9091	0.9839	0.8571	0.9903	0.0161	0.0909	0.1429	0.9091	0.8824
急性 B 淋巴细胞白血病	0.9706	0.9908	0.8462	0.9985	0.0092	0.0294	0.1538	0.9706	0.9041
急性 T 淋巴细胞白血病	0.4000	0.9941	0.3333	0.9956	0.0059	0.6000	0.6667	0.4000	0.3636
有核红细胞异常	0.6970	0.9903	0.8846	0.9685	0.0097	0.3030	0.1154	0.6970	0.7797
B 淋巴细胞异常	0.4118	0.9687	0.2500	0.9848	0.0313	0.5882	0.7500	0.4118	0.3111
T 淋巴细胞异常	0.1111	0.9727	0.1429	0.9640	0.0273	0.8889	0.8571	0.1111	0.1250
嗜碱性粒细胞异常	0.1951	0.9598	0.2353	0.9495	0.0402	0.8049	0.7647	0.1951	0.2133
嗜酸性粒细胞异常	0.6364	0.9805	0.5185	0.9879	0.0195	0.3636	0.4815	0.6364	0.5714
中性粒细胞异常	0.8602	0.9661	0.9040	0.9490	0.0339	0.1398	0.0960	0.8602	0.8815
浆细胞异常	0.8462	0.9896	0.6111	0.9970	0.0104	0.1538	0.3889	0.8462	0.7097
单核细胞异常	0.7889	0.9481	0.6961	0.9675	0.0519	0.2111	0.3039	0.7889	0.7396



注: class 0~11 依次对应 12 类数据, 分别是正常人、急性髓系白血病、急性 T 淋巴细胞白血病、急性 B 淋巴细胞白血病、有核红细胞异常、成熟 T 淋巴细胞异常、成熟 B 淋巴细胞异常、嗜碱性粒细胞异常、嗜酸性粒细胞异常、中性粒细胞异常、浆细胞异常、单核细胞异常。

图 3 CNN-BiLSTM 模型 ROC 曲线

### 3 讨论

FCM 检测能力提升的同时,也给流式细胞检测实验室检验人员的分析效率带来了挑战,而培养一名优秀的流式细胞术分析师需要较高的时间成本,因此寻求一种流式细胞术自动化分析的方法变得尤为必要。虽然目前已经有一些自动分析方法在 FCM 数据取得不错的效果,但由于其操作复杂及自动化不彻底等原因并未被广泛使用<sup>[16]</sup>,因此在实际临床工作中仍以人工分析为主。

本研究通过深度学习模型对流式细胞检测报告结果部分的文字资料进行分析,并对急性白血病患者进行自动分类,探索文本分类方法联合前期对流式细胞仪报告数据通过自动化分析方法实现白血病预测的效果,通过观察 CNN 和 LSTM 基线模型发现,CNN 比 LSTM 的模型性能高,主要原因是原始数据集由人工进行预筛选,且为描述细胞占比的文本句子,每个文本为冗杂且相关性不强的短文本描述句子,上下文相关性不强,且含有众多临床术语。短文本的特征一般独立存在于句子的某个局部,CNN 擅长捕捉短文本的局部特征信息,而 LSTM 捕捉的多为冗杂且相关性不高的上下文特征信息。因此,相较于本研究的流式细胞检测报告结果部分的文本资料,CNN 的分类效果优于 LSTM。

以 CNN 作为基线模型,对 CNN、CNN-LSTM 和 CNN-BiLSTM 进行对比发现,CNN 混合模型比 CNN

的模型性能高。由于传统 CNN 模型的卷积神经网络直接与全连接层相连,而混合模型是在 BiLSTM 或 LSTM 后连接全连接层。由于全连接层会造成部分空间文本信息的丢失,从而忽略了部分上下文的关系。因此,CNN 的准确率低于其混合模型,且混合模型的 F1 值较基线模型高了 11.25%。以 LSTM 作为基线模型,对 LSTM、BiLSTM 进行对比发现,BiLSTM 较 LSTM 好。对于 BiLSTM 模型而言,LSTM 只能处理后向的文本序列,而 BiLSTM 可以同时拼接前向和后向两个方向的输出,能对前后文语义进行更高的表征,提升了模型计算的复杂度和精确度。因此,BiLSTM 较 LSTM 模型的 F1 值提升 0.59%。

对于 CNN 与 LSTM 的混合模型,可得出 LSTM-CNN 比 CNN-LSTM<sup>[17]</sup>的串联模型性能更加优越。对比 CNN-LSTM 和 LSTM-CNN 文本分类模型,两者在网络结构上有所不同,主要体现在卷积层和长期记忆(LSTM)层的顺序。CNN-LSTM 模型首先使用 CNN 对输入的文本进行特征提取,然后将提取到的特征序列输入到 LSTM 层进行序列建模和分类预测。而 LSTM-CNN 模型则是先使用 LSTM 层对文本进行序列建模,然后将 LSTM 输出的特征序列输入到卷积层进行局部特征提取和分类预测。在文本分类任务中,卷积层能够提取出局部特征,但同时也造成了信息的丢失,在只对 LSTM 进行后向序列信息计算时,会造成一定的信息差异;而 LSTM 向后传递



的语句信息是完整的,再将提取的信息传入 CNN 中提取局部关键信息,所以 LSTM-CNN 的分类效果略高于 CNN-LSTM。而 CNN-BiLSTM 由于可以拼接前向和后向两个方向的输出,可以明显提前模型分类效果,对 LSTM-CNN 和 CNN-LSTM 模型的 F1 值分别提高 4.72% 和 5.24%。

本研究也存在一定的局限性:在文本分类模型中,本研究选用的样本量较少,且类别不均匀,因此还未能充分发挥各深度学习模型在大数据分析上的优势;本研究只选择了正常人、急性髓系白血病患者和急性淋系白血病患者等 12 类文本资料进行分析,并未包含急性白血病数据的全部资料;本研究只是对 FCM 的结果部分的文字资料做了文本分类,缺乏更详细的疾病信息,因此仅探讨了各深度学习模型在白血病初步诊断中的应用;为了保证结果的客观性和可信度,本研究未对数据进行增强,也未深入对模型参数的设置进行研究,而是尽可能选择工具包默认参数,因此训练好的模型并不是最好的,可在将来使用过程中进一步完善;由于本研究所选的文字资料存在多标签问题,而本研究只做了单标签文本分类,因此后期将从多标签文本分类角度对数据进行进一步研究。

综上所述,CNN-BiLSTM 深度学习模型对流式细胞检测报告结果部分文本资料的分类效果较好,能够辅助临床工作者在急性白血病诊断上做出更准确的诊断,提高诊断效率和准确性。

#### 参考文献:

- [1]Paul RJ,Mario R.Flow cytometry strikes gold[J].Science, 2015,350(6262):739-740.
- [2]Jaye DL,Bray RA,Gebel HM,et al.Translational applications of flow cytometry in clinical practice[J].J Immunol,2012,188(10): 4715-4719.
- [3]Suo YZ,Gu ZQ,Wei XB.Advances of In Vivo Flow Cytometry on Cancer Studies[J].Cytometry A,2020,97(1):15-23.
- [4]Cheung M,Campbell JJ,Whitby L,et al.Current trends in flow cytometry automated data analysis software [J].Cytometry A, 2021,99(10):1007-1021.
- [5]Greg F,Marc L,Maria J,et al.Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium[J].Scientific Reports,2016,6(1):20686.
- [6]郭玉娟,李智伟,芮东升,等.急性髓系白血病流式细胞术全程自动化诊断技术研究[J].石河子大学学报(自然科学版),2022,40(4):431-437.
- [7]雷伟,李智伟,芮东升,等.卷积神经网络在急性髓系白血病流式细胞术自动诊断中的应用[J].安徽医科大学学报,2023,58(7):1189-1193.
- [8]Fuda F,Chen M,Chen W,et al.Artificial intelligence in clinical multiparameter flow cytometry and mass cytometry-key tools and progress[J].Semin Diagn Pathol,2023,40(2):120-128.
- [9]Van Gassen S,Callebaut B,Van Helden MJ,et al.FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data[J].Cytometry A,2015,87(7):636-645.
- [10]Lacombe F,Lechevalier N,Vial JP,et al.An R-Derived FlowSOM Process to Analyze Unsupervised Clustering of Normal and Malignant Human Bone Marrow Classical Flow Cytometry Data[J].Cytometry A,2019,95(11):1191-1197.
- [11]Collobert R,Weston J,Bottou L,et al.Natural Language Processing (almost) from Scratch[J].CoRR,2011:2493-2537.
- [12]Pan CP,Cao HT,Zhang WW,et al.Driver activity recognition using spatial-temporal graph convolutional LSTM networks with attention mechanism [J].IET Intelligent Transport Systems,2020,15(2):297-307.
- [13]宋纯贺,李泽熙,于洪霞,等.一种基于改进 GoogLeNet 的油井故障识别方法[J].江苏科技大学学报(自然科学版),2021,35(2):52-58.
- [14]王若佳,魏思仪,王继民.BiLSTM-CRF 模型在中文电子病历命名实体识别中的应用研究[J].文献与数据学报,2019,1(2): 53-66.
- [15]Kamruzzaman M,Almazroui M,Salam MA,et al.Spatiotemporal drought analysis in Bangladesh using the standardized precipitation index (SPI) and standardized precipitation evapotranspiration index (SPEI)[J].Sci Rep,2022,12(1):20694.
- [16]马闪闪,董明利,张帆,等.基于核主成分分析的流式细胞数据分群方法研究[J].生物医学工程学杂志,2017,34(1):115-122.
- [17]Obeidat Y,Alqudah A.M.A Hybrid Lightweight 1D CNN-LSTM Architecture for Automated ECG Beat-Wise Classification[J].Traitement du Signal: Signal Image Parole,2021,38(5):1281-1291.

收稿日期:2024-02-19;修回日期:2024-03-28

编辑/杜帆