Hans Journal of Computational Biology 计算生物学, 2024, 14(3), 29-38
Published Online September 2024 in Hans. https://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2024.143003

Hans汉斯

基于双向长短期记忆网络和卷积神经网络的DNA 6mA甲基化位点预测

高　伟，郭晓甜，李慧敏*

云南民族大学数学与计算机科学院，云南　昆明

收稿日期：2024年8月25日；录用日期：2024年9月18日；发布日期：2024年9月25日

摘　要

DNA N6-甲基腺嘌呤(6mA)是一种重要的表观遗传修饰，参与基因调控、DNA复制和修复等生物过程，对疾病研究也具有重要意义，准确识别DNA 6mA位点对理解其功能和机制至关重要。尽管现有的NA 6mA位点预测方法已取得较大成功，但在预测精度和跨物种泛化能力上仍有改进空间。本文提出了一种结合双向长短期记忆网络(BiLSTM)和卷积神经网络(CNN)的混合深度学习模型(BiLSTM → CNN)来提高对DNA 6mA位点预测的能力。模型首先采用one-hot、EIIP和DNA二聚体三种编码方式对DNA序列进行编码，然后在不同网络结构、层数和优化器下优化模型。通过在蔷薇科植物、水稻和拟南芥的数据集上的广泛实验表明，BiLSTM → CNN 模型在蔷薇科植物中的准确率(ACC)为94.5%，在水稻中为93.8%，在拟南芥中为86.6%。与其他方法相比， BiLSTM → CNN 模型在三个植物物种的6mA位点预测中均展现出良好的性能，并具有出色的跨物种泛化能力。

关键词

DNA 6mA位点，双向长短期记忆网络，卷积神经网络，特征编码

Prediction of DNA 6mA Methylation Sites Based on Bidirectional Long Short-Term Memory Network and Convolutional Neural Network

Wei Gao, Xiaotian Guo, Huimin Li*

School of Mathematics and Computer Science, Yunnan Minzu University, Kunming Yunnan

Received: Aug. 25th, 2024; accepted: Sep. 18th, 2024; published: Sep. 25th, 2024

*通讯作者。

文章引用: 高伟, 郭晓甜, 李慧敏. 基于双向长短期记忆网络和卷积神经网络的 DNA 6mA 甲基化位点预测[J]. 计算生物学, 2024, 14(3): 29-38. DOI: 10.12677/hjcb.2024.143003

Abstract

DNA N6-methyladenine (6mA) is an important epigenetic modification involved in biological processes such as gene regulation, DNA replication, and repair, making it significant for disease research. Therefore, accurately identifying DNA 6mA sites is crucial for understanding their functions and mechanisms. Despite notable successes with existing methods, there is still room for improvement in prediction accuracy and cross-species generalization. In this study, we propose a hybrid deep learning model (BiLSTM → CNN) that integrates bidirectional long short-term memory networks (BiLSTM) and convolutional neural networks (CNN). Firstly, the model-encoded DNA sequences employ one-hot encoding, EIIP encoding, and DNA dimer encoding. And then optimized under various network architectures, layer configurations and optimizers. We conducted experiments on datasets from Rosaceae, rice and Arabidopsis thaliana, the results indicate that the BiLSTM → CNN model achieves an accuracy (ACC) of 94.5% for Rosaceae, 93.8% for rice, and 86.6% for Arabidopsis. Compared to other methods, BiLSTM → CNN demonstrates excellent performance in predicting 6mA sites across the three plant species, and exhibits cross-species generalization capabilities.

Keywords

DNA 6mA Sites, Bidirectional Long Short-Term Memory Network, Convolutional Neural Network, Feature Encoding

1. 引言

N6-甲基腺嘌呤(6mA)是 DNA 分子中最重要的表观遗传修饰之一，它在基因表达、核小体定位、细胞周期调节、DNA 修复和复制以及限制性修饰(R-M)中起着重要作用[1]。近年来，通过高效液相色谱(HPLC)分离与串联质谱(MS/MS)联用[2]、单分子实时(SMRT)测序[3]等生物实验方法，已经鉴定出一些6mA 位点，但该类方法由于劳动密集型、耗时和昂贵等特点，其在大规模数据的应用中受到了限制。因此，发展高效的生物信息学方法来识别 6mA 位点尤为重要。

随着人工智能技术的发展，基于机器学习和深度学习的计算方法为 6mA 位点的识别提供了新的解决方案。这些方法不仅高效且成本低，还能在基因组尺度上进行大规模预测。目前，已开发了多种基于 ML 和 DL 的 6mA 预测模型，涵盖从单一特征到多特征融合的不同方法。Chen 等[4]开发了一种基于支持向量机和 DNA 序列特征编码筛选策略的 i6mA-Pred 预测器，用于识别水稻基因组中的 6mA 位点。Pian 等[5]利用马尔可夫模型计算 DNA 序列中相邻核苷酸的转移概率，提出了 MM-6 mAPred 来预测水稻物种 6mA 位点。Kong 和 Zhang [6]开发了 i6mA-DNCP，该通过二核苷酸频率和二核苷酸理化性质来表示水稻 DNA 序列，并采用启发式方法来选择最具代表性的特征。Hasan 等[7]提出了一个多特征融合的计算模型 i6mA-Fuse，使用 5 种编码方案建立单编码随机森林模型，并通过线性回归模型整合这些模型的预测概率评分，用于蔷薇科植物月季和野草莓的 6mA 位点预测。Xu 等[8]通过整合 7 种类型的序列衍生信息和 3 种类型的基于物理化学的特征开发了 6mA-Finder，用于一般和物种特异性的 6mA 位点预测。Hasan 等[9]提出的 i6mA-stack 使用递归特征消除交叉验证策略从 5 种不同的 DNA 序列编码方案中提取最优特征子

集，并使用双层堆叠模型进行预测。Ha 等[10]提出的 Meta-i6mA 方法选择了 5 种基于物理化学和位置特定信息的编码方式，并结合 6 种常用的机器学习方法生成 30 个基线模型，通过元学习预测方法整合这些模型。该模型在蔷薇科植物、水稻和拟南芥上表现出高马修斯相关系数值。He 和 Chen [11]开发了 iDNA6mA-Rice-DL，通过深度学习方法利用嵌入层和密集层自动编码和提取关键 DNA 特征，成功预测了水稻基因组中的 6mA 位点。Huang 等[12]提出了 6mA-StackingCV，一种基于交叉验证的叠加系综模型，通过元学习算法整合多个基线模型的输出，在蔷薇科植物上表现出高级性能和稳定性。Teng 等[13]提出了 i6mA-Vote 采用多数投票策略，结合 6 种不同的特征编码方案和机器学习分类器，有效预测了蔷薇科植物、水稻和拟南芥的 6mA 位点。尽管上述模型在特定物种中取得了较高的预测准确性，但是跨物种预测能力上存在不足。

本文基于长短期记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)和卷积神经网络(Convolutional Neural Network, CNN)，提出一种高效且具有泛化能力的 6mA 位点预测方法(BiLSTM→CNN)。首先将 DNA 序列使用 one-hot [14]，EIIP [15]和 DNA 二聚体(DNA 2-mer) [16] 3 种编码方式进行编码，然后将编码后的序列输入到 BiLSTM 层提取序列的长距离依赖关系，再将这些长距离依赖特征传入 CNN 层，通过卷积操作进一步提取局部特征，最后使用 Sigmoid 激活函数得到 DNA 6mA 甲基化位点的最终预测。结果表明，与现有方法相比，BiLSTM→CNN 模型具有较高的预测性能和泛化能力。

2. 模型及相关理论介绍

2.1. BiLSTM

BiLSTM [17]是一种特殊的递归神经网络，能够同时捕捉序列的前向和后向信息。

$$c_t = f_t * x_t + i_t * \tilde{c}_t \tag{1}$$

$$o_t = \delta\left(W_{xo} x_t + W_{xo} \cdot h_{t-1} + b_o\right) \tag{2}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{3}$$

其中公式(1)~(3)，i_t 代表输入门，δ 为 sigmoid 函数，f_t 代表遗忘门，h_t 为当前隐藏状态，$\tilde{c}_t$ 表示当前输入的单元状态，c_t 代表保留程度，b_o 表示偏置项。前向 LSTM 的输出为 $\overrightarrow{h_t}$，后向 LSTM 的输出为 $\overleftarrow{h_t}$。每个时间步 t 的 BiLSTM 输出 h_t 计算如公式(4)所示。

$$h_t = \overrightarrow{h_t} + \overleftarrow{h_t} \tag{4}$$

2.2. CNN

CNN [18]能够学习序列局部依赖关系，有效地处理和提取基因序列的特征。

$$Y[j] = \sum_j w[j] \cdot X[i+j] \tag{5}$$

$$Y_{pool}[i] = \max\left(X[i:i+k]\right) \tag{6}$$

公式(5)~(6)中，X 是输入序列，w 表示卷积核，$Y[i]$ 表示卷积输出，$w[j]$ 表示卷积权重，$X[i+j]$ 是输入序列的局部区域，k 是池化窗口的大小，$Y_{pool}[i]$ 是最大池化。

2.3. 激活函数

神经网络中常用的激活函数有 tanh，relu，elu，sigmoid 和 softmax [18] [19]等，这些激活函数在模型中具有不同的作用，本文激活函数的选择主要涉及 tanh 和 sigmoid 函数。

tanh 函数是一种双曲正切函数，将输入值压缩到[−1, 1]范围内，该函数是中心对称的，即 tanh (0) = 0，在二分类问题中 tanh 函数多用于隐藏层部分，具体表示为公式(7)。

$$f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{7}$$

sigmoid 函数将输入值压缩到(0, 1)范围内，该函数的输出可以解释为某个事件发生的概率，故 sigmoid 函数常用于二分类问题的输出层，将输出转换为概率值，具体表示为公式(8)。

$$f(x)=\frac{1}{1+e^{-x}} \tag{8}$$

2.4. 优化器

随机梯度下降(Stochastic Gradient Descent, SGD)是一种简单且常用的优化算法，用于在梯度下降过程中最小化损失函数。

$$\theta_{t+1}=\theta_{t}-\eta\nabla_{\theta}L\left(\theta_{t};x^{(i)};y^{(i)}\right) \tag{9}$$

在公式(9)中，$\nabla_{\theta}L\left(\theta_{t};x^{(i)};y^{(i)}\right)$ 表示在第 t 次迭代时，使用第 i 个样本 $x^{(i)}$ 和其对应的标签 $y^{(i)}$ 计算的损失函数梯度，η 为学习率。

Adam [20]通过计算梯度的一阶矩(即梯度的均值)和二阶矩(即梯度的平方均值)的指数移动平均值来动态调整学习率，从而在保证收敛速度的同时提高模型的性能，具体计算为公式(10)~(13)。

$$g_{t}=\nabla_{\theta}L\left(\theta_{t};x^{(i)};y^{(i)}\right) \tag{10}$$

$$m_{t}=\beta_{1}m_{t-1}+\left(1-\beta_{1}\right)g_{t} \tag{11}$$

$$v_{t}=\beta_{2}v_{t-1}+\left(1-\beta_{2}\right)g_{t}^{2} \tag{12}$$

$$\theta_{t+1}=\theta_{t}-\frac{\eta}{\sqrt{\hat{v}_{t}+\varepsilon}}\hat{m}_{t} \tag{13}$$

初始时刻的一阶矩和二阶矩均为 0，β_{1} 和 β_{2} 是动量参数，分别设置为 0.9 和 0.999。参数 θ 沿着修正后的一阶矩方向更新，步长由学习率 η 和修正后的二阶矩的平方根控制，ε 是一个非常小的常数，防止除零错误。

2.5. BiLSTM→CNN 模型的建立

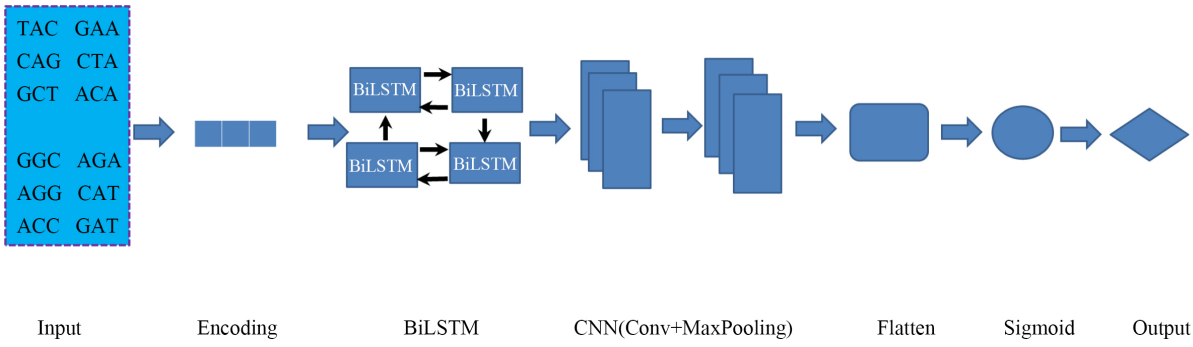

Figure 1. Overall workflow of BiLSTM + CNN model
图 1. BiLSTM → CNN 模型结构

为充分发挥 BiLSTM 在长距离依赖关系建模方面的能力，本文提出 BiLSTM → CNN 模型，将 one-hot，EIIP 和 DNA 2-mer 编码后的序列首先传入 BiLSTM 层，通过双向长短期记忆网络提取序列的长距离依赖关系，然后再将这些长距离依赖特征传入 CNN 层，通过卷积操作进一步提取局部特征，模型结构如图 1 所示。

2.6. 损失函数

该模型采用二进制交叉熵作为其损失函数，适用于二元分类任务，公式如下：

$$loss = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(y_{pred,i}\right) + \left(1-y_i\right)\log\left(1-y_{pred,i}\right)\right] \tag{14}$$

y_i 表示真实标签，$y_{pred,i}$ 表示预测标签。

2.7. 性能评价指标

为了评估模型性能，我们使用准确度(ACC)，马修斯相关系数(MCC)，灵敏度(SN)和特异度(SP)作为模评价指标，具体计算如公式(15)~(28)所示。

$$ACC = \frac{TP + TN}{TN + FP + TP + FN} \tag{15}$$

$$MCC = \frac{TP \times TN + FP \times FN}{\sqrt{\left(TP + FP\right)\left(TN + FN\right)\left(TP + FN\right)\left(TN + FP\right)}} \tag{16}$$

$$SN = \frac{TP}{TP + FN} \tag{17}$$

$$SP = \frac{TN}{TN + FP} \tag{18}$$

其中，TP (真阳性)和 TN (真阴性)分别表示正确预测的 6mA 和非 6mA 样本的数量；FP (假阳性)和 FN (假阴性)分别是错误预测的非 6mA 和 6mA 样本的数量。

3. 实验设计

3.1. 实验数据集

本文使用拟南芥(Arabidopsis)、水稻(Rice)和蔷薇科(Rosaceae)三种植物的 DNA 6mA 数据进行模型的训练和结果的评估。三个物种的数据集中共包含 1 个训练集和 3 个独立测试集，分别是 Rosaceae 训练集，Rosaceae 测试集，Rice 测试集和 Arabidopsis 测试集。本文使用 Rosaceae 训练集训练模型，将其分成两部分，80%的数据用于训练，20%的数据用于验证，通过对模型的训练，使得该模型能够泛化到 Arabidopsis 和 Rice 物种上。四个数据集的详细信息见表 1。

Table 1. Dataset information
表 1. 数据集信息

数据集	阳性样本个数	阴性样本个数
Rosaceae 训练集	29,237	29,433
Rosaceae 测试集	7298	7300
Rice 测试集	153,635	153,629
Arabidopsis 测试集	31,414	31,843

3.2. 特征编码方式

本文将 one-hot、EIIP 和 DNA 2-mer 3 种特征组合进行组合，对给定的 DNA 序列进行编码。在 one-hot 编码中，每个核苷酸被编码为一个 4 维二进制向量，即核苷酸 A、C、G 和 T 分别被表示为 $A=[1,0,0,0]$、$C=[0,1,0,0]$、$G=[0,0,1,0]$ 和 $T=[0,0,0,1]$。EIIP 根据核苷酸在给定 DNA 序列中的电子－离子能量分布来表达核苷酸，四个核苷酸分别表示为 A = 0.1260、C = 0.1340、G = 0.0806 和 T = 0.1335。DNA 二聚体编码(DNA 2-mer Encoding)的方法将 DNA 序列转换为数值数组，即

$$AA\to 0, AT\to 1, AG\to 3, TA\to 4, TT\to 5, TC\to 6, TG\to 7, CA\to 8,$$
$$CT\to 9, CC\to 10, CG\to 11, GA\to 12, GT\to 13, GC\to 14, GG\to 15.$$

当然，为了处理 DNA 序列中的缺失值(表示为“N”)，在 One-Hot 编码和 EIIP 编码过程中，对于遇到的“N”碱基，我们分别使用了零向量 $[0,0,0,1]$ 和数值 0.0 进行填充；在 2-mer 编码过程中，缺失碱基分别编码为 $AN\to 16, CN\to 17, GN\to 18, TN\to 19, NN\to 20$。

3.3. 参数选择

3.3.1. BiLSTM 模块参数选择

BiLSTM 模块的参数主要涉及 BiLSTM 的层数以及优化器，我们选取常用的 BiLSTM 层数 1，2 和 3，优化器为 SGD 和 Adam，并在 Rosaceae 训练集上对模型进行调试，调试结果见表 2，表 3。

Table 2. The effect of BiLSTM layers on the BiLSTM module
表 2. BiLSTM 层数对模型的影响

层数	SN	SP	ACC	MCC	AUC
1	0.943	0.926	0.934	0.869	0.978
2	0.926	0.930	0.928	0.856	0.970
3	0.920	0.929	0.925	0.850	0.970

Table 3. The influence of the optimizer on the BiLSTM module
表 3. 优化器对 BiLSTM module 预测结果的影响

优化器	SN	SP	ACC	MCC	AUC
SGD	0.425	0.659	0.542	0.088	0.562
Adam	0.943	0.926	0.934	0.869	0.978

从表 2 和表 3 可以看出，当 BiLSTM 层数为 1 层，优化器为 Adam 时，BiLSTM 模块性能达到最优。因此 BiLSTM 的层数和优化器分别选取 1 和 Adam。

3.3.2. CNN 模块参数选择

对 CNN 模块的选择，我们同样选取常用的参数，并在 Rosaceae 训练集上进行调试，结果见表 4~6。

Table 4. The effect of convolution layers on the CNN module
表 4. 卷积层数对模型的影响

层数	SN	SP	ACC	MCC	AUC
1	0.933	0.952	0.942	0.885	0.982
2	0.927	0.925	0.926	0.852	0.974
3	0.929	0.924	0.927	0.853	0.973

Table 5. The effect of activation function on the CNN module
表 5. 激活函数对模型的影响

卷积层	压平层	SN	SP	ACC	MCC	AUC
relu	relu	0.935	0.946	0.940	0.881	0.981
elu	elu	0.938	0.948	0.943	0.886	0.982
relu	tanh	0.942	0.944	0.943	0.887	0.982
elu	relu	0.937	0.944	0.941	0.882	0.981
tanh	tanh	0.940	0.947	0.944	0.888	0.982
relu	elu	0.932	0.952	0.942	0.885	0.982
tanh	relu	0.941	0.942	0.941	0.883	0.981
elu	tanh	0.940	0.946	0.943	0.887	0.982
tanh	elu	0.937	0.947	0.942	0.885	0.982

Table 6. The influence of the optimizer on the CNN module
表 6. 优化器对 CNN 模块预测结果的影响

优化器	SN	SP	ACC	MCC	AUC
SGD	0.901	0.892	0.896	0.793	0.958
Adam	0.939	0.949	0.944	0.889	0.983

从表 4~6 可以看出，当卷积层数为 1 层，卷积层和压平层的激活函数均为 Tanh，优化器为 Adam 时，CNN 模块达到了最优性能。因此取 CNN 模块的卷积层数为 1，各部分激活函数均为 Tanh，优化器选择 Adam。

4. 结果与分析

4.1 与现有方法的性能比较

基于上述 3 个物种的独立测试集，我们将 BiLSTM → CNN 方法与现有方法进行比较。表 7 列出了与目前在相应物种上预测能力较好模型的比较结果，主要有 Meta-i6mA [10]，iDNA6mA-Rice [11]，MM-6mAPred [5]和 6mA-StackingCV [12]。

在 Rice 物种中，尽管 BiLSTM → CNN 模型的 SN 值(0.946)略低于几种方法，但在其他三 3 个评估指标上表现出色。与目前对 Rice 6mA 位点预测的最优方法 Meta-i6mA 相比，ACC 提高 0.058，MCC 提高 0.108，SP 提高 0.128。与最新的 6mA-StackingCV 相比，上述 3 个指标分别高 0.093，0.169 和 0.204。表明 BiLSTM → CNN 模型在 Rice 物种中具有卓越的预测性能，尤其是在减少假阴性方面表现优异。

在 Arabidopsis 物种中，BiLSTM → CNN 模型在 4 个评估指标中依然表现出色。与目前对 Arabidopsis 中 6mA 位点预测最好和最新的方法 6mA-StackingCV 相比，其 ACC 提高 0.084，MCC 提高 0.154，SN 提高 0.180，并且在所有方法中均为最优。其 SP 值为 0.873，仅低于 Meta-i6mA 的 0.936，排名第二。表明，BiLSTM → CNN 模型在对 Arabidopsis 物种的 6mA 位点预测方面具有显著的优势。

尽管 BiLSTM → CNN 模型在 Rosaceae 独立测试数据集上的表现性能略低于 6mA-StackingCV 和 Meta-i6mA，但是 4 个指标的值在所有方法中均位于前列。并且与最好的方法 6mA-StackingCV 相比，各项指标的差异最大为 0.029，展示了其在该物种上同样具有竞争力。

综上所述，BiLSTM → CNN 在 3 个独立测试集上均达到了较好的性能，并且相对于其他方法，具有更好的泛化能力。

Table 7. Comparison results with existing methods
表 7. 与现有方法结果比较

独立测试集	模型	ACC	MCC	SN	SP
Rosaceae	Meta-i6mA*	0.953	0.905	0.954	0.951
	iDNA6mA-Rice*	0.878	0.764	0.951	0.805
	MM-6mAPred*	0.873	0.758	0.961	0.785
	6mA-StackingCV*	0.960	0.920	0.959	0.961
	BiLSTM + CNN	0.945	0.891	0.945	0.945
Rice	Meta-i6mA*	0.880	0.768	0.957	0.802
	iDNA6mA-Rice*	0.755	0.561	0.960	0.547
	MM-6mAPred*	0.834	0.689	0.958	0.710
	6mA-StackingCV*	0.845	0.710	0.963	0.726
	BiLSTM + CNN	0.938	0.876	0.946	0.930
Arabidopsis	Meta-i6mA*	0.787	0.600	0.636	0.936
	iDNA6mA-Rice*	0.734	0.473	0.655	0.812
	MM-6mAPred*	0.765	0.531	0.784	0.747
	6mA-StackingCV*	0.782	0.576	0.677	0.866
	BiLSTM + CNN	0.866	0.730	0.857	0.873

星号(*)表示结果来自文献[12]。

4.2. 模型分析

为了验证各模块的有效性以及 BiLSTM 和 CNN 顺序的合理性，我们还比较了仅使用 CNN 或者 BiLSTM，以及 BiLSTM 和 CNN 不同的组合方式在不同数据集上的预测结果。这几种模型分别记为 CNN，BiLSTM，CNN → BiLSTM (表示 CNN 在前，BiLSTM 在后)以及 CNN ↔ BiLSTM (表示 CNN 和 BiLSTM 并联)，本文模型记为 BiLSTM → CNN 。表 8 列出了比较结果。

从表 8 可以看出，BiLSTM → CNN 模型在 3 个独立测试集上的 ACC 均达到最优，在 Rosaceae 中的 SN，Rice 中的 MCC 和 SP，以及 Arabidopsis 中的 MCC 略低于 CNN，Arabidopsis 中 SP 略低于 CNN → BiLSTM，其他指标均达到最优。这表明相对于只使用单一的 CNN 或者 BiLSTM 模块，或者其他的 BiLSTM 和 CNN 的结合顺序，BiLSTM → CNN 模型无论是在 6mA 位点的预测准确性方面，还是模型的泛化能力，都表现出优越的性能。图 2 所示的 ROC 曲线进一步支持了这些发现。

Table 8. Performance comparison of five models on independent test data sets
表 8. 五种模型在独立测试数据集上的性能比较

独立测试集	模型	ACC	MCC	SN	SP
Rosaceae	BiLSTM	0.912	0.825	0.914	0.910
	CNN	0.942	0.884	0.948	0.937
	CNN ↔ BiLSTM	0.940	0.879	0.939	0.940
	CNN → BiLSTM	0.941	0.882	0.940	0.942
	BiLSTM → CNN	0.945	0.891	0.945	0.945

续表

Rice	BiLSTM	0.938	0.876	0.945	0.931
	CNN	0.938	0.877	0.943	0.934
	CNN ↔ BiLSTM	0.937	0.875	0.941	0.933
	CNN → BiLSTM	0.937	0.875	0.942	0.933
	BiLSTM → CNN	0.938	0.876	0.946	0.930
Arabidopsis	BiLSTM	0.847	0.694	0.826	0.867
	CNN	0.865	0.733	0.849	0.884
	CNN ↔ BiLSTM	0.864	0.729	0.844	0.883
	CNN → BiLSTM	0.863	0.727	0.840	0.886
	BiLSTM → CNN	0.866	0.730	0.857	0.873

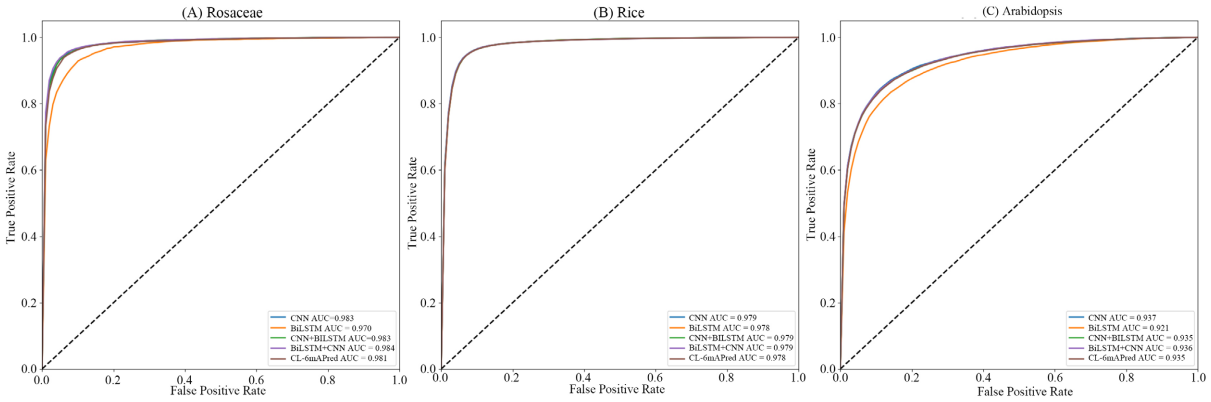

Figure 2. The roc curves of different models on three independent test datasets
图 2. 不同模型在三个独立测试数据集上的 ROC 曲线

5. 结论

本文提出了一种基于 BiLSTM 和 CNN 的 DNA 6mA 甲基化位点混合预测模型 BiLSTM → CNN 。首先利用 one-hot、EIIP 和 DNA 二聚体 3 种编码方式对 DNA 序列进行编码，然后在不同网络结构、层数、优化器和学习率下优化模型，并在拟南芥、水稻和蔷薇科三种植物上进行了实验验证。本文通过对比单一CNN模型、单一BiLSTM模型、CNN↔BiLSTM模型、CNN→BiLSTM模型以及本文提出的BiLSTM→CNN模型在 Rosaceae、Rice 和 Arabidopsis 三个独立测试数据集上的预测结果，表明 BiLSTM→CNN 模型在总体准确率(ACC)方面均达到最优，且在多数性能指标上也表现出色。尽管在 Rosaceae 中的敏感性(SN)、Rice 中的马修斯相关系数(MCC)和特异性(SP)，以及 Arabidopsis 中的 MCC 和 SP 方面略低于某些模型，但其综合性能和泛化能力仍然显著优于其他模型。因此，BiLSTM→CNN 模型在 6mA 位点预测任务中展现出优越的性能。与现有方法相比，本文模型在 DNA 6mA 甲基化位点的预测上表现出更优的性能。

值得注意的是，DNA 6mA 甲基化位点的预测工作相当复杂，算法的预测能力不仅取决于模型本身，还与位点附近的序列和位点分布相关，有效结合这些特征可以显著提升预测效果。随着研究手段和计算方法的不断进步，大量新的 DNA 6mA 甲基化位点将被确定，为识别这些位点提供新的数据支持。在未来的研究中，我们还考虑进一步优化特征编码融合方式，集成多种深度学习方法，以及在更广泛的数据集进行试验评估，从而得到更优的 DNA 6mA 预测模型。

基金项目

云南省研究生优质课程建设项目“高等数理统计”(云学位[2022] 8 号)。

参考文献

[1] 杜轲. DNA 表观遗传修饰 6mA 抑制 DNA 聚合酶 eta 催化 DNA 复制的动力学研究[D]: [硕士学位论文]. 延安: 延安大学, 2019.

[2] Ye, Q., Belabed, H., Wang, Y., Yu, Z., Palaniappan, M., Li, J., *et al.* (2022) Advancing ASMS with LC-MS/MS for the Discovery of Novel PDCL2 Ligands from DNA-Encoded Chemical Library Selections. *Andrology*, **11**, 808-815. https://doi.org/10.1111/andr.13309

[3] Adhikari, S., Erill, I. and Curtis, P.D. (2021) Transcriptional Rewiring of the GcrA/CcrM Bacterial Epigenetic Regulatory System in Closely Related Bacteria. *PLOS Genetics*, **17**, e1009433. https://doi.org/10.1371/journal.pgen.1009433

[4] Chen, W., Lv, H., Nie, F. and Lin, H. (2019) i6mA-Pred: Identifying DNA N6-Methyladenine Sites in the Rice Genome. *Bioinformatics*, **35**, 2796-2800. https://doi.org/10.1093/bioinformatics/btz015

[5] Pian, C., Zhang, G., Li, F. and Fan, X. (2019) MM-6mAPred: Identifying DNA N6-Methyladenine Sites Based on Markov Model. *Bioinformatics*, **36**, 388-392. https://doi.org/10.1093/bioinformatics/btz556

[6] Kong, L. and Zhang, L. (2019) i6mA-DNCP: Computational Identification of DNA N6-Methyladenine Sites in the Rice Genome Using Optimized Dinucleotide-Based Features. *Genes*, **10**, Article 828. https://doi.org/10.3390/genes10100828

[7] Hasan, M.M., Manavalan, B., Shoombuatong, W., Khatun, M.S. and Kurata, H. (2020) i6mA-Fuse: Improved and Robust Prediction of DNA 6 Ma Sites in the Rosaceae Genome by Fusing Multiple Feature Representation. *Plant Molecular Biology*, **103**, 225-234. https://doi.org/10.1007/s11103-020-00988-y

[8] Xu, H., Hu, R., Jia, P. and Zhao, Z. (2020) 6mA-Finder: A Novel Online Tool for Predicting DNA N6-Methyladenine Sites in Genomes. *Bioinformatics*, **36**, 3257-3259. https://doi.org/10.1093/bioinformatics/btaa113

[9] Khanal, J., Lim, D.Y., Tayara, H. and Chong, K.T. (2021) i6mA-Stack: A Stacking Ensemble-Based Computational Prediction of DNA N6-Methyladenine (6mA) Sites in the Rosaceae Genome. *Genomics*, **113**, 582-592. https://doi.org/10.1016/j.ygeno.2020.09.054

[10] Hasan, M.M., Basith, S., Khatun, M.S., Lee, G., Manavalan, B. and Kurata, H. (2020) Meta-i6mA: An Interspecies Predictor for Identifying DNA N^6-Methyladenine Sites of Plant Genomes by Exploiting Informative Features in an Integrative Machine-Learning Framework. *Briefings in Bioinformatics*, **22**, bbaa202. https://doi.org/10.1093/bib/bbaa202

[11] He, S., Kong, L. and Chen, J. (2021) iDNA6mA-Rice-DL: A Local Web Server for Identifying DNA N6-Methyladenine Sites in Rice Genome by Deep Learning Method. *Journal of Bioinformatics and Computational Biology*, **19**, Article ID: 2150019. https://doi.org/10.1142/s0219720021500190

[12] Huang, G., Huang, X. and Luo, W. (2023) 6mA-StackingCV: An Improved Stacking Ensemble Model for Predicting DNA N6-Methyladenine Site. *BioData Mining*, **16**, Article No. 34. https://doi.org/10.1186/s13040-023-00348-8

[13] Teng, Z., Zhao, Z., Li, Y., Tian, Z., Guo, M., Lu, Q., *et al.* (2022) i6mA-Vote: Cross-Species Identification of DNA N6-Methyladenine Sites in Plant Genomes Based on Ensemble Learning with Voting. *Frontiers in Plant Science*, **13**, Article 845835. https://doi.org/10.3389/fpls.2022.845835

[14] Kelley, D.R., Snoek, J. and Rinn, J.L. (2016) Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks. *Genome Research*, **26**, 990-999. https://doi.org/10.1101/gr.200535.115

[15] Alakuş, T.B. (2023) A Novel Repetition Frequency-Based DNA Encoding Scheme to Predict Human and Mouse DNA Enhancers with Deep Learning. *Biomimetics*, **8**, Article 218. https://doi.org/10.3390/biomimetics8020218

[16] Matsuki, M., Lago, P. and Inoue, S. (2019) Characterizing Word Embeddings for Zero-Shot Sensor-Based Human Activity Recognition. *Sensors*, **19**, Article 5043. https://doi.org/10.3390/s19225043

[17] Farid, A.B., Fathy, E.M., Sharaf Eldin, A. and Abd-Elmegid, L.A. (2021) Software Defect Prediction Using Hybrid Model (CBIL) of Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM). *PeerJ Computer Science*, **7**, e739. https://doi.org/10.7717/peerj-cs.739

[18] Yasin, M., Sarıgül, M. and Avci, M. (2024) Logarithmic Learning Differential Convolutional Neural Network. *Neural Networks*, **172**, Article ID: 106114. https://doi.org/10.1016/j.neunet.2024.106114

[19] 王双印, 滕国文. 卷积神经网络中 ReLU 激活函数优化设计[J]. 信息通信, 2018(1): 42-43.

[20] 邢波涛. 基于全卷积神经网络的 MR 脑肿瘤图像分割算法研究[D]: [硕士学位论文]. 天津: 天津大学, 2018.