DOI: 10.13671/j.hjkxxb.2025.0240
程洋,尹海龙.2025.基于特征提取优化的CNN-LSTM市政泵站水质预测模型研究[J].环境科学学报,45(11):120-128
CHENG Yang, YIN Hailong. 2025. Research on water quality prediction model for municipal pumping stations based on CNN-LSTM with enhanced feature extraction[J]. Acta Scientiae Circumstantiae, 45(11): 120-128

基于特征提取优化的CNN-LSTM市政泵站水质预测模型研究

程洋[1,2],尹海龙[1,*]

1. 同济大学,环境科学与工程学院,上海 200092
2. 上海市排水管理事务中心,上海 200001

摘要:市政泵站水质预测对于排水系统调度管理和雨天河道污染控制具有重要的指导意义.本文提出了一种基于特征提取优化(Enhanced Feature Extraction, EFE)、卷积神经网络(Convolutional Neural Network, CNN)和长短时记忆网络(Long Short-Term Memory, LSTM)的水质等级预测模型.首先利用CNN提取输入特征中的丰富信息,然后利用LSTM捕捉时序上的依赖关系,最后利用交叉熵作为水质等级预测的损失函数.相较于传统的卷积循环神经网络(CNN-LSTM)模型,本模型在CNN上融合了EFE结构,提升了特征提取的能力和模型的稳定性.选取上海中心城区12座市政泵站2022年7月—2024年10月的监测数据,对模型预测性能进行验证.结果表明:在12座市政泵站水质等级预测中,本文提出的EFE-CNN-LSTM模型与传统的CNN、LSTM和CNN-LSTM模型相比整体F_1-score平均值提升幅度分别超过了24%、29%和9%,验证了EFE结构的有效性.该模型在12座泵站的测试集上均取得了较高精度,具有较好的适宜性和工程价值.研究方法可为市政泵站水质预测及泵站运行控制策略的制定提供借鉴.

关键词:特征提取优化;机器学习模型;市政泵站;水质预测

文章编号:0253-2468(2025)11-0120-09　　**中图分类号**:X32,X832　　**文献标识码**:A

Research on water quality prediction model for municipal pumping stations based on CNN-LSTM with enhanced feature extraction

CHENG Yang[1,2], YIN Hailong[1,*]

1. School of Environment Science and Engineering, Tongji University, Shanghai 200092
2. Shanghai Municipal Drainage Affairs Center, Shanghai 200001

Abstract: The prediction of water quality at municipal pump stations is of significant guidance for the operational management of urban drainage systems and the control of river pollution during wet-weather events. This paper proposes a water quality level prediction model based on enhanced feature extraction (EFE), a convolutional neural network (CNN), and a long short-term memory (LSTM) network. The model first utilizes the CNN to extract rich informational features from the input data, followed by the LSTM to capture temporal dependencies. Finally, cross-entropy is employed as the loss function for the water quality level prediction. Compared to the conventional CNN-LSTM model, our proposed model integrates an EFE structure into the CNN, thereby enhancing its feature extraction capabilities and improving overall model stability. To validate the model's performance, monitoring data from 12 municipal pump stations in the central urban area of Shanghai, collected from July 2022 to October 2024, were used. The results demonstrate that in the prediction of water quality levels across the 12 pump stations, the proposed EFE-CNN-LSTM model shows an average increase in the overall *F*1-score of over 24%, 29%, and 9% compared to traditional CNN, LSTM, and CNN-LSTM models, respectively, which validates the effectiveness of the EFE structure. The model achieved high accuracy on the test sets for all 12 pump stations, indicating its excellent suitability and practical engineering value. The research methodology presented can provide a valuable reference for water quality prediction at municipal pump stations and for the formulation of pump operation and control strategies.

Keywords: enhanced feature extraction; machine learning model; municipal pumping station; water quality prediction

收稿日期:2025-05-30　　**修回日期**:2025-07-07　　**录用日期**:2025-07-09
基金项目:国家自然科学基金(No.52170103);上海市排水管理事务中心资助项目(No.ZX-ZB2025-0020)
作者简介:程洋(1993—),女,E-mail:18321243648@163.com;*责任作者,E-mail:yinhailong@tongji.edu.cn

1　引言(Introduction)

我国快速城市化进程中,城区面积不断扩大,不透水地面增长导致自然渗透减少和雨水径流量增加.在末端建设有市政泵站的排水系统,地表径流由分散排放模式变为集中排放模式,雨天时雨水管网对降雨径流进行集中收集和输送,经市政泵站排放至受纳水体.由于地表径流中污染物浓度较高,排水系统雨污分流不彻底,以及管网沉积物成分复杂等原因,导致排放至受纳水体中的污染物浓度较高,河道水质不能稳定达到功能区目标,甚至出现雨天河道黑臭现象.市政泵站雨天排放已经成为城区河道的主要污染排放来源,也是造成河道水质不稳定的主要原因.因此,非常有必要对市政泵站水质进行预测,进而探究市政泵站雨天排放的优化控制策略,减轻对受纳水体水质的污染程度.

随着机器学习的迅猛发展,越来越多的研究者开始将机器学习技术应用到水环境领域(程兵芬等,2023;刘杰等,2024;郭利进等,2025).众所周知,机器学习是数据驱动的,而数字水务的发展积累了大量的数据,这促使机器学习技术成为分析预测水质的有力工具.前馈神经网络由于其简单易用的结构特征,前期被广泛应用于排水系统水质预测研(陈威等,2020;Noori *et al.*,2020;李雪清等,2021).基于卷积神经网络(Convolutional Neural Network,CNN)(Khan *et al.*,2020)在特征提取方面的优异性能,肖明君等(2024)构建了三层卷积网络用于河湖水体水质预测.考虑到水质预测往往要依赖历史数据,例如,历史降雨量、历史水位等,长短时记忆网络(Long Short-Term Memory, LSTM)(Graves,2012)在水质预测领域逐渐流行起来(陈湛峰等,2024;尚旭东等,2024;Yin *et al.*,2025).为了充分提取输入水质指标中的信息和捕捉水质指标时序上的周期性,一些研究将CNN与LSTM网络级联起来构建了基于CNN-LSTM的水质预测模型.Zhang等(2024)提出了一种注意力机制增强的CNN-LSTM河湖水体水质预测模型.Barzegar等(2020)提出了三层卷积神经网络的CNN-LSTM河湖水体水质预测模型.Yang等(2021)构建了两层卷积神经网络级联两层LSTM的CNN-LSTM河湖水体水质预测模型.传统的多层CNN网络结构(Barzegar *et al.*,2020;Pan *et al.*,2020;肖明君等,2024)中,输入数据经过卷积层和激活函数后直连下一个卷积层.这种模型结构虽然简单易于搭建,但在特征提取能力和模型训练稳定性方面存在一些缺陷.在特征提取能力上,经过多层卷积和激活函数运算后,最终抽取的特征可能已经无法准确地包含初始输入特征的上下文信息.在模型训练稳定性方面,多层卷积层经过激活函数运算后直接级联可能会引发梯度异常,造成模型训练无法收敛.

本文提出了一种基于特征提取优化的CNN-LSTM水质预测模型.与传统的多层CNN网络不同的是,本文在CNN上融合EFE结构,用以提升特征提取的能力和模型稳定性.EFE结构包括残差连接和层归一化两部分.残差连接能够有效地缓解多层CNN网络中潜在的梯度消失问题,同时还能保证经过多层特征抽取后的输出变量依然能够准确地包含初始输入特征的上下文信息.层归一化操作稳定了中间层输入的分布,进一步缓解了梯度异常.本文以上海中心城区12座市政泵站2022年7月—2024年10月的监测数据为研究对象,选取降雨量、前池水位、瞬时流量、截流泵开机个数、防汛泵开机个数和历史水质数据作为输入特征,分析本研究模型预测当前时刻水质等级的性能.在计算过程中,与传统的CNN、LSTM和CNN-LSTM模型进行对比,以验证EFE结构的有效性.在测试集上,本研究模型在12座泵站均取得了较高精度,研究方法可为市政泵站水质预测及泵站运行控制策略的制定提供借鉴

2　材料与方法(Materials and methods)

2.1　数据来源

研究对象为上海中心城区排水系统.在综合分析泵站类型、截流设施、排水片区、运行模式等情况的基础上,选取了12座市政泵站作为代表性泵站进行研究,包括6座雨水泵站和6座合流泵站,主要集中在黄浦江、苏州河、淀浦河、桃浦河、西泗塘等河道.

本文的数据来自12座泵站的在线监测数据,时间跨度为2022年7月—2024年10月.数据主要包括降雨量、前池水位、瞬时流量、截流泵开机个数和防汛泵开机个数等指标,以及化学需氧量(Chemical Oxygen Demand, COD)和氨氮(Ammonia Nitrogen, NH_3-N)等指标浓度.

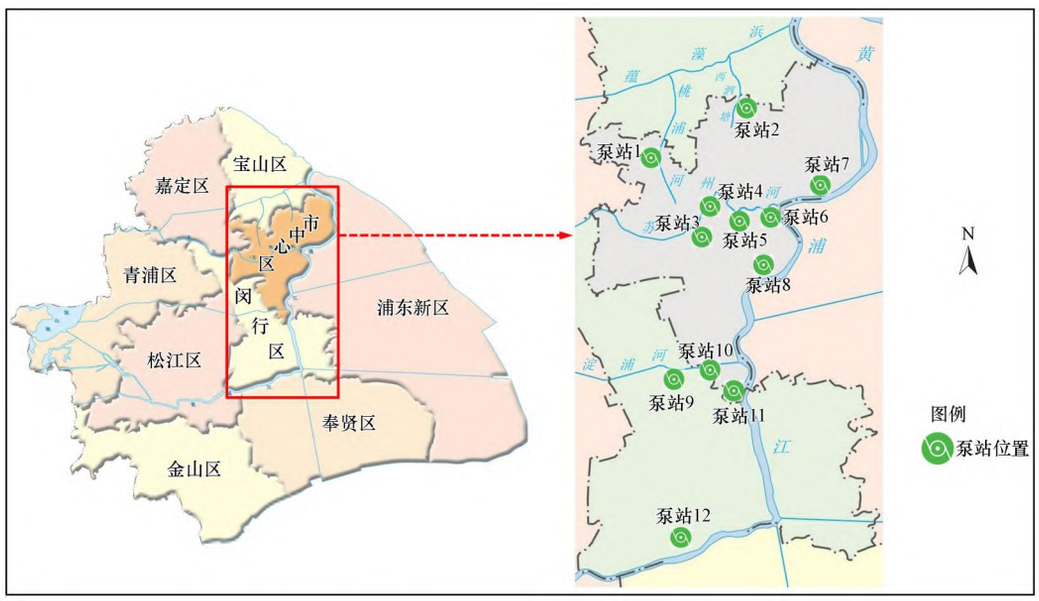

图1 上海中心城区12座代表性市政泵站分布

Fig.1 Distribution of 12 typical municipal pumping stations in Shanghai downtown area

2.2 预测模型构建

2.2.1 模型总体设计框架 本文提出的EFE-CNN-LSTM模型结构如图2所示.输入数据是前120个时刻的降雨量、前池水位、瞬时流量、截流泵开机个数、防汛泵开机个数、COD或NH_3-N指标浓度,输出数据是当前时刻的COD或NH_3-N指标浓度.考虑到COD和NH_3-N没有相互影响的关系,所以本文中COD和NH_3-N指标浓度是分开训练,分开预测.

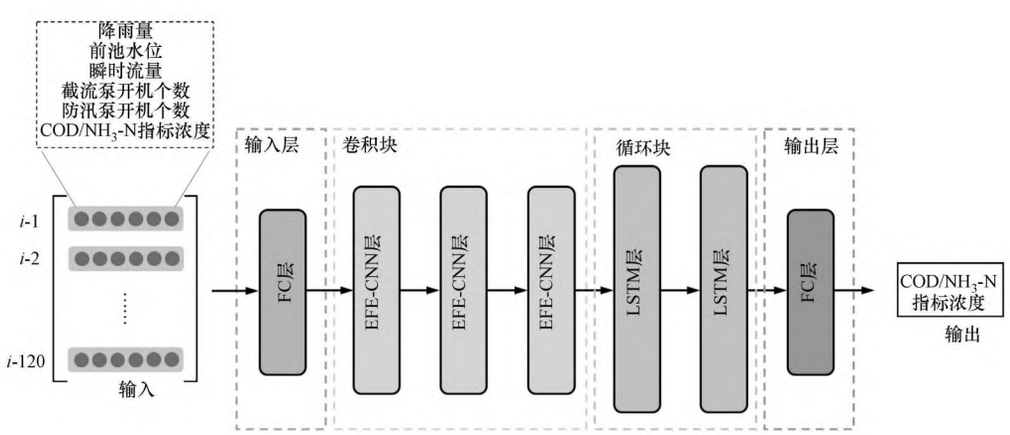

图2 EFE-CNN-LSTM模型结构

Fig.2 EFE-CNN-LSTM model structure

模型一共由四大块组成,分别是输入层、卷积块、循环块和输出层.输入层是一个全连接(Fully Connected,FC)层,它将低维数据转换成高维数据,方便后续特征提取.卷积块由三层EFE-CNN结构级联组成,可以有效地提取输入序列数据在时序维度上的特征.经过卷积块处理后,数据进入由两层LSTM网络级联组成的循环块.循环块可以捕捉数据在时序上的依赖关系.最后,输出层的FC层用来计算预测的目标值.EFE-CNN-LSTM模型通过融合卷积神经网络与长短期记忆网络的架构优势,在保留各自特征提取与时序建模能力的基础上,显著提升了模型的鲁棒性.

2.2.2 EFE-CNN结构 本文提出的EFE-CNN结构如图3所示.每个EFE-CNN模块包括一个一维卷积层Conv1D层、一个RELU激活函数、残差连接和一个LayerNorm层.选择一维卷积网络作为卷积层,作用是将输入

数据与卷积核做卷积运算.RELU激活函数能够学习数据间的非线性关系,并加速模型收敛.假设输入数据为X,那么经过激活函数后的中间状态$H(X)$为式(1).

$$H(X)=\mathrm{RELU}(\boldsymbol{W}\odot X+\boldsymbol{b}) \tag{1}$$

式中,RELU为激活函数,$\boldsymbol{W}$为卷积层权重矩阵,$\odot$为卷积运算,$\boldsymbol{b}$为偏置向量.图3中灰色部分为EFE结构,包括残差连接和层归一化两部分.残差连接的核心公式可以表示为式(2).

$$F(X)=H(X)+X \tag{2}$$

其构建了梯度直流通路,通过跳跃连接使得梯度可通过短路路径直接回传,缓解了链式求导导致的梯度消失问题;同时,通过输入的线性叠加确保了深层特征对初始输入上下文信息的完整性保留.批归一化和层归一化是两种常用的归一化技术,批归一化是沿批次维度(跨样本)归一化,而层归一化是沿特征维度(单样本内)归一化.因本文中样本之间并无关联,所以选用层归一化.层归一化动态校准了中间层输入的统计分布,通过样本内均值和方差归一化抑制了梯度异常.残差连接和层归一化二者协同形成了稳定的梯度流拓扑,既保障了深层特征无损传递,又增强了网络对深度结构的适应性.

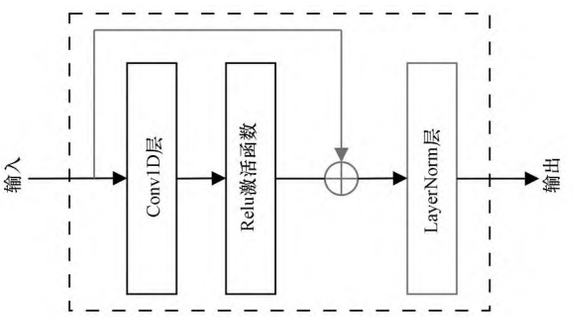

图3　EFE-CNN结构图

Fig.3　The structure of EFE-CNN

2.2.3　LSTM网络　LSTM网络是循环神经网络(Recurrent Neural Network, RNN)的一种变体,它能够有效解决RNN网络梯度消失和梯度爆炸的问题.LSTM网络的记忆单元结构如图4所示,其中,C表示LSTM的单元状态,h表示单元的隐藏层状态.每个记忆单元包含3类门控结构:遗忘门、输入门和输出门.记忆单元负责存储单元状态信息,门控结构则负责单元状态的更新和维护.

遗忘门控制历史信息对当前单元状态的影响,通过Sigmoid函数计算遗忘系数f_t,用于调节前一时刻单元状态C_{t-1}的保留程度,具体见式(3).

$$f_t=\sigma\left(w_f\times\left[h_{t-1},x_t\right]+b_f\right) \tag{3}$$

式中,h_{t-1}为前一时刻的隐藏层状态,x_t为当前时刻的输入值,w_f和b_f分别为输入门的权重和偏置.输入门则与遗忘门协同更新单元状态,分别见式(4)~(6).

$$i_t=\sigma\left(w_i\times\left[h_{t-1},x_t\right]+b_i\right) \tag{4}$$

$$\tilde{C}_t=\tanh\left(w_c\times\left[h_{t-1},x_t\right]+b_c\right) \tag{5}$$

$$C_t=C_{t-1}\times f_t+\tilde{C}_t\times i_t \tag{6}$$

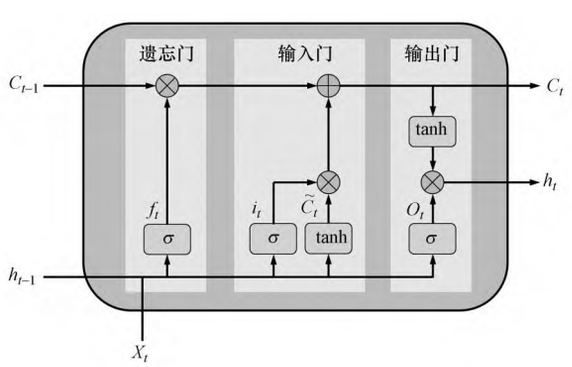

图4　LSTM记忆单元结构

Fig.4　LSTM memory cell structure

式中,$\tilde{C}_t$为候选状态,C_t为新单元状态,w_i和w_c分别为输入门和当前单元候选状态的权重,b_i和b_c分别为对应的偏置.输出门通过Sigmoid函数调控经tanh激活函数后的单元状态输出,见式(7)~(8).

$$o_t=\sigma\left(w_o\times\left[h_{t-1},x_t\right]+b_o\right) \tag{7}$$

$$h_t=o_t\times\tanh\left(C_t\right) \tag{8}$$

式中,h_t为新隐藏层状态,w_o和b_o分别为输出门的权重和偏置.LSTM记忆单元结构通过门控机制实现了对时序信息的动态选择与长期依赖建模.

3　实验设计(Experiment setup)

3.1　数据预处理

3.1.1　不同时间尺度的数据处理　泵站的不同指标数据通常采集频率不同,因此需要统一时间尺度.由于

本文采用的水质等级数据为小时尺度(一个小时一个样本),所以取1 h内降雨量总和作为小时降雨量.对于前池水位、瞬时流量、截流泵开机个数、防汛泵开机个数、COD和NH_3-N浓度,取1 h内的平均值作为对应的小时指标.本文中瞬时流量是指雨水泵和截流泵瞬时流量总和.

3.1.2 数据样本异常和缺失处理 受到泵站监测设备掉线及人为操作失误的影响,收集到的数据不可避免地存在异常值或者缺失值.对于监测数据中的异常值,进行删除.对于数据中的缺失值,如果仅缺失一次,就用前一次的监测值填补;如果出现连续缺失,则删除这些监测时刻的数据以避免人为填补造成误差影响模型学习的效果.

3.1.3 标准化处理 为解决输入特征间尺度差异问题,加速模型收敛并提升训练稳定性,本文采用Z-score标准化方法对输入数据进行预处理,具体见式(9).

$$x' = \frac{x - \mu}{\sigma} \tag{9}$$

式中,x'为标准化后的输入数据,x为输入数据,μ和σ分别为输入数据的平均值和标准差.

3.2 数据集划分

本文模型的输入数据是前N×24小时(包括当前小时)的降雨量、前池水位、瞬时流量、截流泵开机个数、防汛泵开机个数以及前N×24小时的COD或NH_3-N指标浓度,预测目标值是当前小时的COD或NH_3-N指标浓度.经数据预处理后,得到"输入数据-目标值"配对为样本格式的数据集.本文将2022年7月—2024年7月的数据用于训练集和验证集,并在全局时间顺序上对训练集和验证集进行划分,以此构建时间序列样本.同时,将2024年8—10月的数据用于测试集.最终,本文的训练集和验证集样本数为15360,测试集样本数为1216.

3.3 建模环境

本文模型使用Pytorch搭建(Paszke *et al.*,2019).一维卷积核的尺寸设置为3,并采用等长填充same模式以保持输入序列的长度不变.两层LSTM网络隐藏层的维度设置为384.采用Adam优化器,初始学习率设为0.002,权重衰减设为2.5×10^{-5}.使用ReduceLROnPlateau工具来调节学习率变化.为实现正则化,dropout系数设为0.2. Batch size设为64,在训练集上训练模型30遍.在验证集上表现最好的模型用来评估测试集.对于N的取值,本文尝试了集合{1,2,3,4,5,6,7}中的数值,发现N取5时模型在验证集上表现最好,因此N设为5.模型是在一个NVIDIA RTX 3090Ti显卡上训练.

3.4 模型评估方法

出于水质管理目标的考虑,本文将水质数据划分成不同的等级(Umair *et al.*,2019;石晴宜等,2021;薛亚婷等,2023;程婉清等,2023),训练模型预测水质等级.因此,本文选用分类预测模型中常用的F_1值(F_1-score)、精确率(Precision)和召回率(Recall)来评估模型的预测性能(Chen *et al.*,2020;Nasir *et al.*,2022;Guo *et al.*,2024).这3种评估方法定义为式(10)~(12).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{11}$$

$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{12}$$

式中,TP、FP和FN分别为真正例数量、假正例数量和假负例数量.本文用F_1-score来表征模型的预测准确率,F_1-score越大,表明模型的预测准确率越高,性能越好.

4 结果与讨论(Results and discussion)

4.1 泵站水质健康状况

本文参照《地表水环境质量标准》(GB 3838-2002)和《污水综合排放标准》(GB 8978-1996),将COD和NH_3-N浓度划分为不同的等级.具体分级标准如下:COD划分为5个等级(I~V类),其阈值区间分别为0~40、

40~100、100~150、150~500和>500 mg·L^{-1};NH_3-N划分为4个等级(I~IV类),其阈值区间分别为0~2、2~15、15~25和>25 mg·L^{-1}. 12座泵站在2022年7月—2024年10月期间COD和NH_3-N指标浓度各类的样本占比如表1和表2所示.从表1可以看出,6号泵站V类等级样本的占比较高,达到了14.8%;8号和11号泵站IV类等级样本的占比均超过了70%.因此,6、8和11号泵站存在COD超标风险.从表2可以看出,1、8、10和11号泵站IV类等级样本的占比均超过了30%,存在NH_3-N浓度超标风险.

表1　中心城区12座泵站COD指标各等级样本占比

Table 1　Proportions of sample categories for COD at 12 selected municipal pumping stations

泵站编号	样本占比						泵站编号	样本占比					
	I	II	III	IV	V	总体		I	II	III	IV	V	总体
1	11.7%	18.8%	24.8%	38.8%	5.9%	100.0%	7	20.1%	5.0%	3.1%	63.9%	7.9%	100.0%
2	39.1%	10.6%	5.7%	43.6%	1.0%	100.0%	8	12.0%	4.6%	9.5%	73.5%	0.4%	100.0%
3	12.8%	5.6%	30.5%	46.1%	5.0%	100.0%	9	6.4%	20.0%	12.4%	55.2%	6.0%	100.0%
4	12.6%	10.4%	9.5%	63.9%	3.6%	100.0%	10	25.2%	5.4%	7.2%	61.5%	0.7%	100.0%
5	18.9%	15.0%	6.1%	59.3%	0.7%	100.0%	11	9.4%	6.8%	7.5%	76.2%	0.1%	100.0%
6	10.9%	9.2%	13.9%	51.2%	14.8%	100.0%	12	13.4%	10.5%	32.2%	37.1%	6.8%	100.0%

表2　中心城区12座泵站NH_3-N指标各等级样本占比

Table 2　Proportions of sample categories for NH_3-N at 12 selected municipal pumping stations

泵站编号	样本占比					泵站编号	样本占比				
	I	II	III	IV	总体		I	II	III	IV	总体
1	12.5%	28.9%	21.3%	37.3%	100.0%	7	15.1%	35.3%	22.7%	26.9%	100.0%
2	10.5%	9.2%	60.3%	20.0%	100.0%	8	2.4%	27.3%	22.8%	47.5%	100.0%
3	10.0%	41.1%	26.4%	22.5%	100.0%	9	5.3%	62.8%	22.0%	9.9%	100.0%
4	3.2%	25.1%	59.8%	11.9%	100.0%	10	7.0%	11.4%	46.3%	35.3%	100.0%
5	8.1%	25.7%	57.1%	9.1%	100.0%	11	10.1%	13.3%	30.2%	46.4%	100.0%
6	10.7%	12.5%	56.7%	20.1%	100.0%	12	7.8%	30.8%	37.3%	24.1%	100.0%

4.2　模型预测分析

由于12座泵站的类型、服务范围内用地类型、运行模式等存在差异,导致不同泵站获取数据中的非线性关系也不同,因此,针对12座泵站在数据集上各自独立地训练CNN、LSTM、CNN-LSTM和EFE-CNN-LSTM模型并比较它们的预测性能.其中,作为对比的CNN模型是三层没有EFE结构的一维卷积网络,LSTM模型是两层LSTM网络,CNN-LSTM模型除了没有EFE结构,其余结构与EFE-CNN-LSTM模型相同.在12座泵站的测试集上,COD和NH_3-N指标浓度分级预测的整体F_1-score结果如表3和表4所示.从表3和表4中可以看出,EFE-CNN-LSTM模型在COD和NH_3-N指标浓度分级预测上都取得了最高的F_1-score.与CNN和LSTM模型相比,CNN-LSTM模型取得了更高的F_1-score,说明CNN-LSTM模型结合了CNN提取特征和LSTM捕捉时序依赖关系的优点,这也与Barzegar等(2020)的实验结论一致.相比CNN、LSTM和CNN-LSTM模型,在COD指标分级预测上,EFE-CNN-LSTM模型在12座泵站的平均值上将整体F_1-score分别提升了24.1%、29.6%和9.0%;在NH_3-N指标浓度分级预测上,EFE-CNN-LSTM模

表3　4种机器学习模型的COD指标分级预测整体F_1-score

Table 3　The overall F_1-score of COD class prediction across four machine learning models

泵站编号	F_1-score			
	CNN	LSTM	CNN-LSTM	EFE-CNN-LSTM
1	60.1%	59.7%	72.4%	82.1%
2	65.5%	63.2%	69.5%	81.0%
3	63.2%	66.5%	78.8%	81.7%
4	62.7%	60.5%	70.3%	83.4%
5	70.9%	65.3%	73.7%	80.3%
6	66.4%	61.9%	76.7%	82.5%
7	67.9%	62.3%	78.1%	83.4%
8	57.6%	61.9%	75.9%	79.4%
9	72.1%	65.8%	80.6%	83.2%
10	67.4%	63.4%	74.2%	80.5%
11	65.0%	60.9%	71.5%	80.9%
12	66.1%	64.7%	73.9%	82.6%
平均	65.4%	63.0%	74.6%	81.7%

型将整体F_1-score分别提升了26.8%、29.9%和11.6%.这验证了EFE结构对模型预测性能提升的有效性.

4.3 12座泵站的模型运行结果

EFE-CNN-LSTM模型在12座泵站的测试集上的水质等级预测结果如图5所示,其中,横坐标表示COD和NH_3-N指标浓度的水质等级,纵坐标表示3种模型评估方法的得分,用百分比表示.在COD指标分级预测中,模型在IV类等级水质表现最优,F_1-score达85.7%~90.6%,召回率高达87.5%~92.5%,表明其对重度污染水体具有出色的识别能力;在V类等级水质预测中性能显著下降,精确度最低降至64.3%,显示模型对极端污染情况的识别存在挑战;在II~III类等级水质预测表现中等,其中III类等级水质的召回率波动较大(75.6%~85.7%).在NH_3-N指标浓度分级预测中,模型在II类和III类等级水质的预测最为稳定,F_1-score维持在79.7%~87.3%之间,显示出对中等浓度NH_3-N的良好检测能力,说明模型对中等浓度NH_3-N的泛化能力较强;在IV类等级水质预测存在明显的精确度-召回率差异;在I类等级水质预测中表现相对均衡但F_1-score较低.横向对比显示,模型对COD的整体预测性能(F_1-score均值81.1%)略优于NH_3-N(F_1-score均值81.0%),但两者在极低或极高浓度区间上均表现较弱,可能与极端浓度样本的噪声干扰或数据不平衡有关,反映出小样本学习困难.

表4 4种机器学习模型的NH_3-N指标浓度分级预测整体F_1-score

Table 4 The overall F_1-score of NH_3-N class prediction across four machine learning models

泵站编号	F_1-score			
	CNN	LSTM	CNN-LSTM	EFE-CNN-LSTM
1	62.3%	61.0%	68.7%	81.4%
2	60.2%	61.5%	72.5%	80.3%
3	65.5%	64.7%	77.6%	80.5%
4	64.7%	62.6%	75.0%	82.7%
5	61.3%	60.9%	68.2%	81.7%
6	65.9%	65.0%	76.8%	81.5%
7	59.8%	60.3%	68.9%	80.5%
8	64.6%	62.7%	71.7%	80.1%
9	66.3%	61.6%	75.5%	80.2%
10	61.9%	59.4%	70.0%	81.0%
11	70.3%	68.8%	72.4%	83.0%
12	63.6%	62.3%	71.5%	81.4%
平均	63.9%	62.5%	72.4%	81.2%

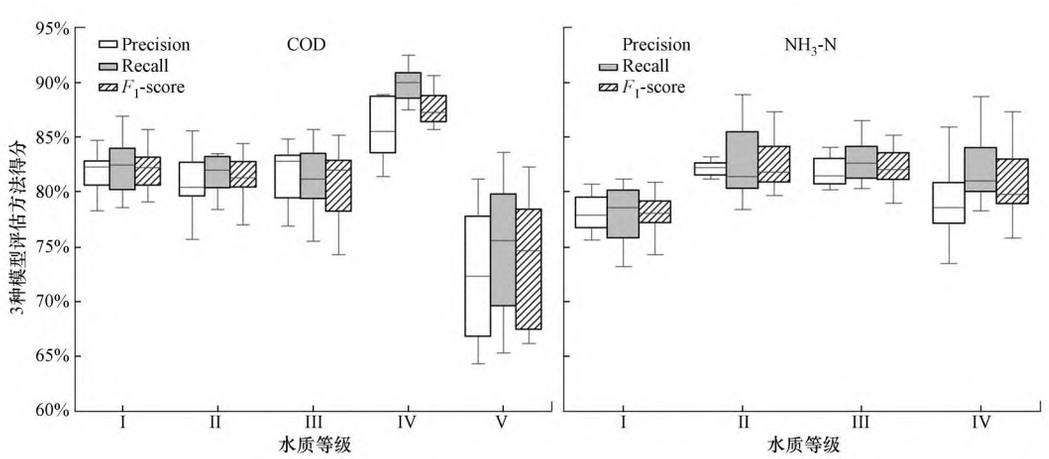

图5 中心城区12座泵站水质等级预测结果

Fig.5 Water quality class prediction results of 12 municipal pumping stations

12座泵站的COD和NH_3-N指标浓度分级预测的混淆矩阵(百分比)如图6所示,12座泵站的混淆矩阵依序按首字母顺序排列.在COD预测中,IV类等级的识别准确率最高,平均值达到了89.8%,但普遍存在V类等级被误判成IV类等级误判率较高的现象,这与IV类等级样本占比普遍较高(均值55.8%)以及V类等级样本稀缺性(均值4.4%)直接相关.类似地,在NH_3-N预测中,9号泵站II类等级样本占比高达62.8%,I类等级样本仅占5.3%,导致很大一部分I类等级样本被误判成II类等级.这说明模型会产生对占比较高类别等级的强烈偏倚,导致少数类别等级的识别能力下降、误判率上升.而在类别等级分布较为均衡的泵站中,比如6号泵站的COD,混淆矩阵的预测结果在各类别等级之间分布更加均匀,没有单一类别等级占据主导地位,从而整体

分级预测准确率更高,系统性误判较少.上述现象在COD和NH_3-N两项指标中均有体现,表明类别等级严重失衡会导致模型对主导类别等级的过度预测,而均衡的数据分布则有助于提升模型对各类别等级的辨别能力.在未来的研究中,可以尝试针对稀疏样本设计数据增强策略,从而平衡训练样本中各分级的占比来减少误判.

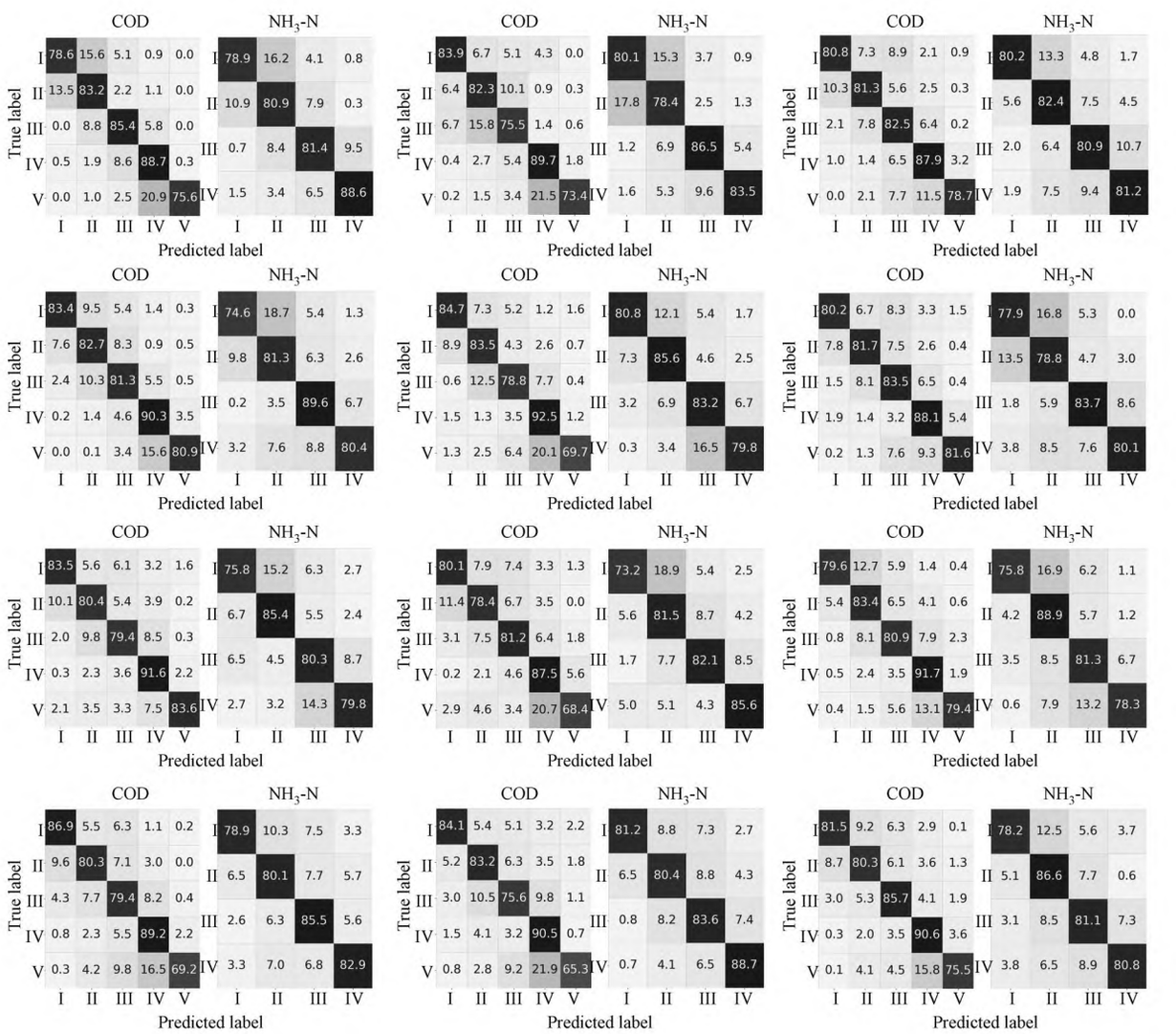

图6　中心城区12座泵站COD和NH_3-N指标浓度分级预测混淆矩阵

Fig.6　Confusion matrices of COD and NH_3-N class prediction at 12 municipal pumping stations

5　结论(Conclusions)

1)相比CNN和LSTM模型,CNN-LSTM模型能够结合CNN提取特征和LSTM捕捉时序依赖关系的优点,取得更高的预测准确率.

2)本文提出的基于特征提取优化(EFE)的CNN-LSTM模型提升了特征提取的能力和模型的稳定性.在12座市政泵站水质等级预测中,本文提出的EFE-CNN-LSTM模型与传统的CNN、LSTM和CNN-LSTM模型相比整体F_1-score平均值提升幅度分别超过了24%、29%和9%,验证了EFE结构的有效性.

3)分级预测结果表明,训练样本中占比较低的类别在预测时误判率会比占比较高的类别更高.未来可以尝试针对稀疏样本设计数据增强策略,从而平衡训练样本中各分级的占比来减少误判.

4)使用EFE-CNN-LSTM模型在12座泵站各自的测试集上均取得了较高精度,具有较好的适宜性和工程价值,研究方法可为市政泵站水质预测及泵站运行控制策略的制定提供借鉴.

参考文献(References):

Ahmed U, Mumtaz R, Anwar H, *et al*. 2019. Efficient water quality prediction using supervised machine learning[J]. Water, 11(11):2210

Barzegar R, Aalami M T, Adamowski J. 2020. Short-term water quality variable prediction using a hybrid CNN-LSTM deep learning model[J]. Stochastic Environmental Research & Risk Assessment, 34(2):415-433

Chen K Y, Chen H X, Zhou C L, *et al*. 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data[J]. Water Research, 171: 115454

陈威, 陈会娟, 戴凡翔, 等. 2020. 基于人工神经网络的污水处理出水水质预测模型[J]. 给水排水, 56(S1):990-994

陈湛峰, 李晓芳. 2024. 基于注意力机制优化的BiLSTM珠江口水质预测模型[J]. 环境科学,45(6):3205-3213

程兵芬, 夏瑞, 郑贵强, 等. 2023. 潮白河下游水质演变特征及预测分析[J]. 环境科学学报, 43(10):133-142

程婉清, 袁定波, 熊鹏, 等. 2023. 基于多种机器学习算法的水质指数预测模型构建与评估[J]. 环境科学学报, 43(11):144-152

Graves A. 2012. Long short-term memory[J]. Supervised Sequence Labelling with Recurrent Neural Networks, 385:37-45

Guo H, Chen Z, Teo F Y. 2024. Intelligent water quality prediction system with a hybrid CNN-LSTM model[J]. Water Practice and Technology, 19(11): 4538-4555

郭利进, 刘彦宾, 刘文哲, 等. 2025. 融合数据分解和优化门控循环单元的水质预测模型及应用[J]. 环境科学学报, 45(2):201-213

Khan A, Sohail A, Zahoora U, *et al*. 2020. A survey of the recent architectures of deep convolutional neural networks[J]. Artificial Intelligence Review, 53(8):5455-5516

李雪清, 郑航, 刘悦忆, 等. 2021. 基于多源数据机器学习的区域水质预测方法研究[J]. 水利水电技术(中英文), 52(11):152-163

刘杰, 陈前, 许妍, 等. 2024. 长江流域洞庭湖区出入湖磷通量模拟及水质预测:机器学习与传统水文模型耦合方法[J]. 地球科学, 49(11):3995-4007

Nasir N, Kansal A, Alshaltone O, *et al*. 2022. Water quality classification using machine learning algorithms[J]. Journal of Water Process Engineering, 48: 102920

Noori N, Kalin L, Isik S. 2020. Water quality prediction using SWAT-ANN coupled approach[J].Journal of Hydrology, 590:125220

Pan M, Zhou H, Cao J, *et al*. 2020. Water level prediction model based on GRU and CNN[J].IEEE Access, 8:60090-60100

Paszke A, Lerer A, Killeen T, *et al*. 2019. PyTorch: An imperative style, high-performance deep learning library[J]. Advances in Neural Information Processing Systems, 32:8024-8035

尚旭东, 段中兴, 陈炳生, 等. 2024. 基于双向长短期记忆网络组合模型的水质预测[J]. 环境科学学报, 44(7):261-270

石晴宜, 董增川, 罗赟,等.2021.基于机器学习方法的洪泽湖入湖水质评价及预测研究[J]. 中国农村水利水电, (12):53-59

肖明君, 朱逸纯, 高雯媛, 等. 2024. 基于不同人工神经网络的水质预测方法对比[J].环境科学, 45(10):5761-5767

薛亚婷, 吴升伟, 王江涛. 2023. 基于机器学习算法的水质预测及相关算法比较研究[J]. 水资源开发与管理, 9(7):67-74

Yang Y R, Xiong Q Y, Wu C, *et al*. 2021. A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism[J]. Environmental Science and Pollution Research International, 28(39):55129-55139

Yin H L, Chen Y Q, Zhou J S, *et al*. 2025. A probabilistic deep learning approach to enhance the prediction of wastewater treatment plant effluent quality under shocking load events[J]. Water Research X, 26:100291

Zhang M H, Zhang Z Y, Wang X, *et al*. 2024. The use of attention-enhanced CNN-LSTM models for multi-indicator and time-series predictions of surface water quality[J]. Water Resources Management, 38(15):6103-6119