网络首发时间：2025-12-03 11:27:13　网络首发地址：https://link.cnki.net/urlid/34.1065.R.20251202.1334.020

基于 CNN-LSTM-Attention 模型的缺血性脑卒中发病预测

刘佳铭[1],周　骁[1],王孚银[1],孙　晓[1],夏晓爽[1,2],李　新[1,2]
([1] 天津医科大学第二医院神经内科,天津　300211;[2] 天津市健康气象交叉创新中心,天津　300211)

摘要　目的　构建基于卷积神经网络(CNN)－长短时记忆网络(LSTM)－注意力机制(Attention)的深度学习模型,探讨气象、临床因素与缺血性脑卒中发病的关联性。**方法**　纳入缺血性脑卒中住院患者的临床资料及同期的气象数据,构建基于 CNN、LSTM 和 Attention 的融合模型 CNN-LSTM-Attention,通过最大预测偏差和均方根误差(RMSE)评估模型的预测性能。通过选择 1～7 d 的滞后天数,探讨不同滞后天数对预测性能的影响。**结果**　在短期和长期预测中,CNN-LSTM-Attention 融合模型(短期:1.5 和 0.6;长期:8.3 和 2.5)的最大预测偏差和 RMSE 均优于 LSTM 模型(短期:2.8 和 1.2;长期:19.5 和 5.5)和 CNN-LSTM 模型(短期:2.0 和 0.8;长期:11.2 和 3.3)。纳入滞后天数后,在短期和长期预测中,滞后 3 d(短期:0.7 和 0.4;长期:5.5 和 1.9)和 5 d(短期:0.8 和 0.3;长期:6.5 和 2.0)的最大预测偏差和 RMSE 均小于滞后 0 d(短期:1.5 和 0.6;长期:8.3 和 2.5)。滞后 1 d(短期:1.5 和 0.8;长期:6.8 和 2.4)和 7 d(短期:1.9 和 0.9;长期:7.5 和 2.7)的最大预测偏差和 RMSE 均大于滞后 0 d。**结论**　建立的 CNN-LSTM-Attention 融合模型对缺血性脑卒中发病具有较好的预测性,可为医疗资源合理配置提供参考。

关键词　缺血性脑卒中;气象因素;预测模型;卷积神经网络;长短时记忆网络;注意力机制
中图分类号　R 743.32
文献标志码　A　**文章编号**　1000－1492(2025)12－2353－10
doi:10.19405/j.cnki.issn1000－1492.2025.12.020

缺血性脑卒中致死和致残率高,对公共健康构成了重大挑战,精准预测其发病风险是优化脑卒中防控的关键[1-2]。基于临床特征的预测是通过统计学或机器学习算法评估缺血性脑卒中发生的可能性[3],但其未能充分考虑动态气象因素等外部环境变量的影响。基于环境因素的预测虽发现 $PM_{2.5}$、气温等与缺血性脑卒中发病显著相关[4],但缺乏对气象变化与临床特征等多因素的综合分析。因此,研究融合个体临床特征与环境动态因素的综合预测方法对于提升预测准确性与可靠性具有重要价值。

以深度学习为代表的人工智能方法在时序数据分析中表现突出,有效捕捉时间依赖关系,广泛应用于医疗序列数据的再分析[5]。脑卒中数据有显著时序特征,分析其变化规律可为发病预测提供参考。在此基础上,该研究构建了卷积神经网络(convolutional neural network,CNN)与长短期记忆网络(long short-term memory, LSTM)与注意力机制(Attention)融合的模型(CNN-LSTM-Attention),通过分析患者临床资料与空气污染、气象等数据,实现脑卒中发病预测,为优化脑卒中防控提供参考。

1　材料与方法

1.1　病例资料　本研究回顾性收集天津医科大学第二医院神经内科 2023 年 4 月 30 日—2024 年 4 月 29 日诊断为急性缺血性脑卒中的 1345 例住院患者。纳入标准:① 年龄 ≥ 18 岁;② 经头颅计算机断层扫描或磁共振成像诊断为急性缺血性脑卒中,符合《中国急性缺血性脑卒中诊治指南 2023》中的诊断标准[6];③ 临床病历资料完整。排除标准:① 入院时发病时间超过 3 d;② 合并急性心肌梗死;③ 合并严重肝肾功能障碍;④ 合并恶性肿瘤患者。最终共纳入 1 038 例符合标准的患者,年龄为(70.80 ±11.34)岁,其中男性 599 例(57.71%)。

1.2　资料采集

1.2.1　基线资料　纳入患者入院后 24 h 内采集的首次实验室检查结果,主要记录其性别、年龄、既往史、收缩压和舒张压。同时收集患者入院时的血常规、肝功能、肾功能、血脂、空腹血糖(fasting blood

2025－08－24 接收
基金项目:国家自然科学基金(编号:42275197);天津市卫生健康科技项目(编号:TJWJ2023XK007);天津市医学重点学科建设项目(编号:TJYXZDXK-065B);天津市科技计划项目(编号:21JCZDJC01230)
作者简介:刘佳铭,女,硕士研究生;
李　新,女,教授,博士生导师,通信作者,E-mail:lixinsci@126.com

glucose, FBG)及同型半胱氨酸(homocysteine, HCY)等实验室指标。其中血常规包括白细胞计数(white blood cell count, WBC)、红细胞计数(red blood cell count, RBC)、血红蛋白(hemoglobin, HGB)和血小板计数(platelet count, PLT)。肝功能包括丙氨酸氨基转移酶(alanine aminotransferase, ALT)和天门冬氨酸氨基转移酶(aspartate aminotransferase, AST)。肾功能包括血尿素氮(blood urea nitrogen, BUN)和肌酐(creatinine, Cr)。血脂包括三酰甘油(triglycerides, TG)、总胆固醇(total cholesterol, TC)、高密度脂蛋白胆固醇(high-density lipoprotein, HDL)、低密度脂蛋白胆固醇(low-density lipoprotein, LDL)和极低密度脂蛋白胆固醇(very low-density lipoprotein, VLDL)。

1.2.2 气象和污染物资料 收集2023年4月30日—2024年4月29日期间天津市的气象和污染物数据,气象因素数据来源于国家气象科学数据中心(https://data.cma.cn/),主要包括日最高、最低和平均温度、日平均风速和日平均相对湿度。同期大气污染物数据来源于天津市空气质量数据GBQ(https://citydev.gbqyun.com/data/tianjin),主要包括细颗粒物(fine particulate matter 2.5, $PM_{2.5}$)、可吸入颗粒物(inhalable particulate matter 10, PM_{10})、臭氧(ozone, O_3)、二氧化硫(sulfur dioxide, SO_2)、一氧化碳(carbon monoxide, CO)和二氧化氮(nitrogen dioxide, NO_2)。在建立模型时,将当日入院患者的临床资料匹配当日气象数据,在验证滞后效应时,分别将临床资料与前1~7 d的气象数据进行匹配。

1.2.3 算法模型原理

1.2.3.1 CNN CNN是一种深度学习算法,通过卷积层提取输入数据的局部特征,并通过池化层进行降维,从而实现数据的层级抽象[7]。CNN在图像处理和时序数据中表现出色,能够自动从复杂数据中提取重要特征。对于卒中临床时序数据,CNN能有效捕捉其时空特征,确保特征提取过程的自洽性。具体而言,卷积核是CNN结构执行特征提取的关键,每个卷积层通过权值共享来提取输入样本的特征,从而实现局部特征的提取和降维。其中,第1层的卷积运算输出值为:

$$y^l = \sum_{i=1}^{c^{l-1}} w_{i,c}^l x_i^{l-1}$$

式中,c^{l-1}:上一层的第c个通道;x_i^{l-1}:第i个通道的输入量;$w_{i,c}^l$:第l层卷积核的权重矩阵。

池化层通过减少特征向量的维度,帮助防止模型在预测时出现过拟合现象。以全局平均池化为例,假设输入特征的尺寸为宽度(W) × 高度(H) × 通道数(C)。在这种情况下,全局平均池化对于每个通道,会对该通道内的所有元素进行平均计算。具体的计算公式为:

$$y^{l(i,j)} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} x^{l(i,j)}$$

式中,$x^{l(i,j)}$:l层中第i个通道的第j个神经元的输入值。

此外,在CNN结构的具体应用中,通常会在网络末端添加一个全连接层。全连接层的作用是将提取到的局部特征整合为全局特征,并通过加权组合后,生成最终的输出。具体而言,第$l+1$层第j个神经元输出值为:

$$y^{l+1}j = \sum_{i=1}^{n} w_{i,j}^l x^{l(i)}$$

式中,$x^{l(i)}$:l层中第j个神经元的输入值;$w_{i,c}^l$:l层中第i个神经元和下一层中第j个神经元之间的权重。

1.2.3.2 LSTM LSTM是一种特殊的循环神经网络,能够有效解决传统循环神经网络在处理长时序数据时的梯度消失和爆炸问题[8]。该模型通过引入遗忘门、输入门和输出门来控制信息的传递和保留,使其能记住长期依赖关系。其具体结构如图1所示。对于卒中临床时序数据和环境危险因素数据,LSTM能够有效捕捉数据中的长期时间依赖性,并精准提取关键特征,展现出较强的分析和学习能力,特别适用于处理临床时间序列数据。

根据上图可知,在每个训练时间步中,LSTM网络首先接受当前时刻t的输入和上一时刻$t-1$的隐藏状态h_{t-1},通过Sigmoid函数经过遗忘门进行处理,计算公式如下:

$$f^t = \sigma(w_f[h_{t-1}, x_t] + b_t), \sigma(x) = \frac{1}{1+e^{-x}}$$

式中,w_f:遗忘门的权重矩阵;b_t:遗忘门的偏置矩阵。

输入门根据当前时刻t的输入x_i和上一时刻$t-1$的隐藏状态h_{t-1}的信息,选择性决定将目标信储层到细胞状态C_t中,计算公式如下:

$$q_t = \sigma(w_q[h_{t-1}, x_t] + b_q)$$

$$a_t = \tanh(w_c[h_{t-1}, x_t] + b_c), \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$C_t = f_t C_{t-1} + q_t + a_t$$

式中,q_t:输入门的输出;w_q:输入门的权重矩阵;b_q:输入门的偏置矩阵;a_t:输入节点的输出;w_c:

输入节点的权重矩阵;b_c:输入节点的偏置矩阵;C_t和C_{t-1}分别为t时刻和$t-1$时刻的单元状态。

输出门确定当前细胞状态中有多少信息将被用作当前的输出或隐藏状态h_t,计算公式如下:

$$\sigma_t = \sigma(w_o[h_{t-1}, x_t] + b_0)$$

$$h_t = O_t \tanh(C_{t-i})$$

式中,O_t:输出门的输出;w_o:输出门的权重矩阵;b_0:输出门的偏置矩阵。

1.2.3.3 Attention Attention是一种通过为输入序列中的每个元素分配动态权重的技术,使模型能够专注于最相关的信息[9]。它通过动态调整关注点来优化信息处理,尤其在处理复杂数据时具有显著优势。对于缺血性脑卒中临床数据和同期气象数据和污染物数据,注意力机制能够有效识别并聚焦于关键因素,从而提升特征学习性能。注意力机制结构见图2所示。注意力机制的结构通常包括四个关键子模块:① 查询向量:代表当前步骤的隐状态,用于向模型提供当前时刻的上下文信息。② 键向量:表示输入序列的特征,用于与查询向量进行匹配,以决定哪些输入信息更为相关。③ 值向量:表示输入序列中的具体信息,用于最终生成模型的输出。④ 注意力权重:通过加权点积得分函数计算查询向量和键向量的相似度来获得,反映了每个输入元素在当前任务中的重要性。

具体而言,加权点积得分函数是点积得分函数的变体,通过对点积结果进行缩放来避免数值过大

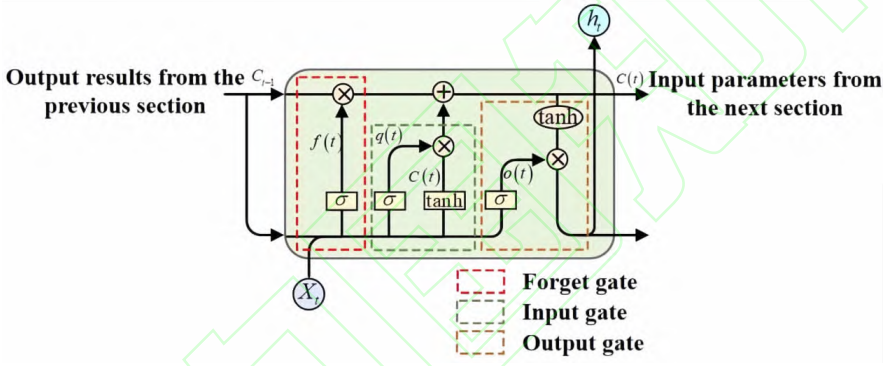

图1 LSTM框架原理

Fig. 1 Principle of LSTM framework

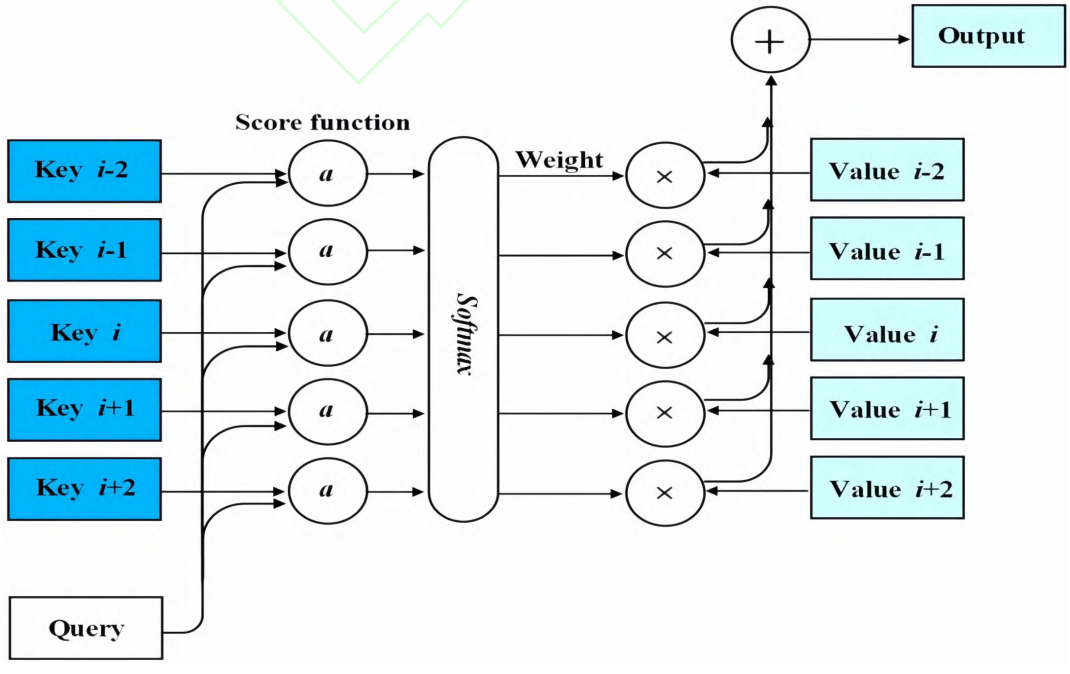

图2 Attention结构图

Fig. 2 Structure diagram of attention mechanism

或过小的问题，计算公式为：

$$score(Q,K)=\frac{Q\cdot K^{T}}{\sqrt{d_k}}$$

式中，Q：查询向量；K：键向量；d_k：键向量的维度。

通过查询向量、键向量、值向量和注意力权重四个模块的协同工作，注意力机制能够动态地选择性关注输入序列中最相关的部分，尤其在处理卒中临床数据、气象数据和污染物数据时，能够自动识别并聚焦于关键因素，从而提升模型的特征学习和预测性能。这使得模型在复杂的时序数据分析和多模态数据处理中，能够更精确地提取有价值的信息，改善预测结果。

1.2.3.4 CNN-LSTM-Attention 融合模型 结合 CNN、LSTM 和 Attention 模型的优势，建立一种高效的融合模型。具体而言，CNN 层用于提取临床数据、气象数据中的空间特征，能够自动识别原始数据中的关键局部模式；LSTM 层则主要用于捕捉气象数据与临床数据之间的时间相关性，有效建模时序数据的长期依赖关系；注意力机制进一步提升了模型的特征学习能力，通过为输入变量动态分配权重，帮助模型聚焦于最具影响力的特征，从而增强了对关键因素的识别和预测精度。该 CNN-LSTM-Attention 融合模型能够结合气象变量与临床数据的时空特征，准确预测未来周期内的缺血性脑卒中发病风险。通过这种综合性方法，模型不仅提高了预测性能，还能够提供更为细致的预测信息。该模型的具体预测流程见图 3 所示，展示了各个模块如何协同工作，以实现对缺血性脑卒中发病风险的精准预测。

1.3 模型评价指标 通过均方根误差（root mean squared error，RMSE）和最大预测偏差进行模型效果的评价，表示模型预测值与实际值之间的误差大小，误差越小表示模型效果越好。

1.4 统计学处理 Python 3.9.0 用于建立预测模型。SPSS 26.0 用于统计学分析，分别用均数 ± 标准差（$\bar{x}\pm s$）、中位数（P_{50}）、上四分位数（P_{25}）和下四分位数（P_{75}）描述连续变量，使用频数（n）和百分比（%）描述分类变量表示。

2 结果

2.1 一般特征 使用描述性统计方法对收集数据进行总结和量化描述，以获取研究样本的整体数据趋势和分布特征。根据纳入排除标准，在 2023 年 4 月 30 日—2024 年 4 月 29 日观察期间内，急性缺血性脑卒中日平均入院人次为（4.3 ± 2.0）人次。平均温度为（14.58 ± 12.81）℃，平均相对湿度为（62.84 ± 23.78）%，SO_2、NO_2、$PM_{2.5}$、PM_{10} 和 O_3 日均浓度分别为（7.48 ± 2.15）、（34.26 ± 17.00）、（40.95 ± 32.31）、（76.83 ± 50.81）和（109.68 ± 56.16）μg/m^3，CO 日均浓度为（0.69 ± 0.25）mg/m^3。基本情况详见表 1。

2.2 算法可行性实验与分析 本研究对所建立的

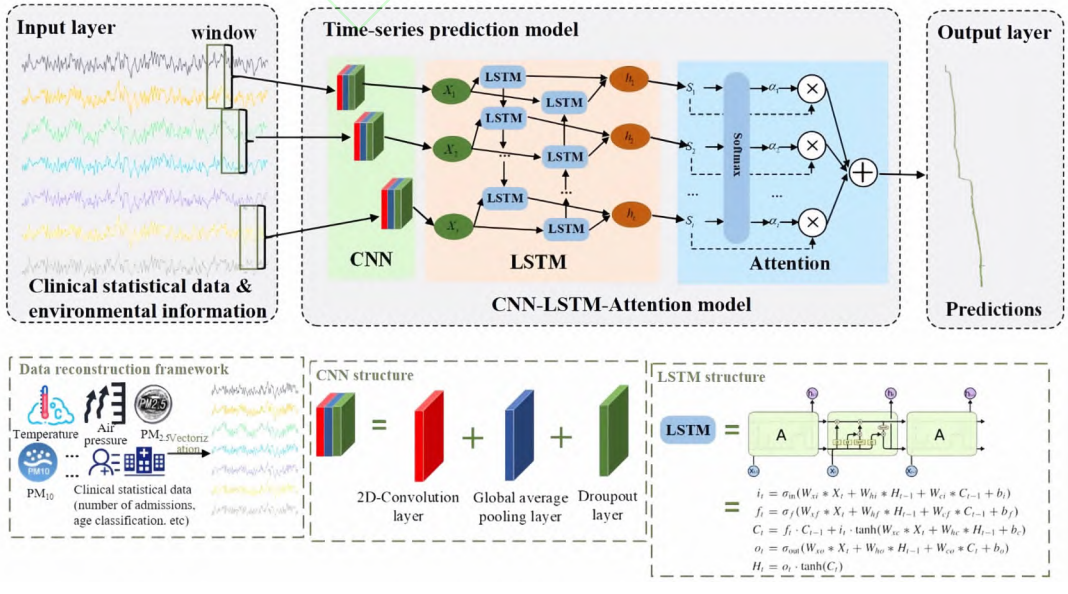

图 3 CNN-LSTM-Attention 模型框架原理

Fig. 3 Principle of CNN-LSTM-Attention model framework

表1 人口学特征、实验室指标、污染物及气象参数基本情况

Tab.1 Baseline characteristics of demographics, laboratory measures, pollutants, and meteorological factors

Factors	n (%)	$\bar{x} \pm s$	P_{25}	P_{50}	P_{75}
Individual basic feature					
Male	599 (57.71)	–	–	–	–
Hypertension	895 (86.22)	–	–	–	–
Diabetes	436 (42.00)	–	–	–	–
Coronary heart disease	336 (32.37)	–	–	–	–
cerebral infarction	313 (30.15)	–	–	–	–
Age (years)	–	149.48 ± 22.83	135	149	165
SBP (mmHg)	–	86.05 ± 25.88	77	85	95
DBP (mmHg)	–	86.15 ± 24.22	78	82	94
Blood testing indicators					
WBC ($\times 10^9$/L)	–	7.60 ± 2.99	5.64	6.96	8.58
RBC ($\times 10^{12}$/L)	–	4.22 ± 0.67	3.86	4.27	4.66
HGB (g/L)	–	127.67 ± 21.51	116	130	142
PLT ($\times 10^9$/L)	–	217.94 ± 75.15	173	207	255
TG (mmol/L)	–	1.50 ± 0.96	0.94	1.25	1.77
TC (mmol/L)	–	4.61 ± 1.32	3.72	4.55	5.38
HDL (mmol/L)	–	1.14 ± 0.35	0.90	1.10	1.32
LDL (mmol/L)	–	2.96 ± 1.07	2.19	2.88	3.63
VLDL (mmol/L)	–	0.51 ± 0.31	0.32	0.46	0.63
FBG (mmol/L)	–	7.10 ± 3.52	4.89	6.00	7.87
ALT (U/L)	–	20.89 ± 42.39	10.40	14.60	21.73
AST (U/L)	–	21.41 ± 33.19	13.00	16.10	21.20
BUN (mmol/L)	–	7.08 ± 4.83	4.80	5.90	7.63
Cr (μmol/L)	–	97.23 ± 127.46	58.80	72.90	90.60
HCY (μmol/L)	–	327.00 ± 110.26	252.20	317.10	383.85
Air pollutant indicators					
$PM_{2.5}$ (μg/m^3)	–	40.95 ± 32.31	19.00	32.00	50.75
PM_{10} (μg/m^3)	–	76.83 ± 50.81	42.00	62.50	92.75
O_3 (μg/m^3)	–	109.68 ± 56.16	65.50	98.00	149.00
SO_2 (μg/m^3)	–	7.48 ± 2.15	6.00	7.00	9.00
CO (mg/m^3)	–	0.69 ± 0.25	0.50	0.60	0.80
NO_2 (μg/m^3)	–	34.26 ± 17.00	21.00	30.50	42.75
Meteorological indicators					
Daily average temperature (℃)	–	14.58 ± 12.81	2.30	17.10	26.00
Daily maximum temperature (℃)	–	15.08 ± 12.79	2.80	17.70	26.50
Daily minimum temperature (℃)	–	14.08 ± 12.85	1.80	16.50	25.60
Daily average relative humidity (%)	–	62.84 ± 23.78	43.00	65.00	83.00
Daily average wind speed (m/s)	–	2.08 ± 1.44	1.00	1.70	2.80

模型在实际应用中的可行性进行了验证。通过结合验证集,评估了模型在处理临床数据、气象数据和污染物数据等多元信息时的适应性和稳健性。在本研究的实验中,纳入2023年4月30日—2024年4月29日的数据作为研究数据,将其划分为训练集(2023年4月30日—2024年3月8日,黑线所示)和验证集(2024年3月8日—4月29日,蓝线所示),用于预测模型的验证。具体验证结果如图4所示。

从图4可以看出,蓝线与黑线趋势较为接近,表明所提出模型对验证集中日入院数的预测效果总体较好。进一步分析预测结果可见,距离初始预测时间较近的预测值与实际值更为吻合,而随着预测时间延后,预测值逐渐偏离实际值。因此,可以认为该模型在验证集上表现出前期预测性能较好,后期预测性能有所下降的特点。根据这一结果,本研究以2024年3月8日作为预测起点,将前期预测性能相对稳定的阶段定义为短期预测,后期相对波动的阶段定义为长期预测。从图4可见,蓝线自3月23日左右开始波动较前明显,因此将3月8日—3月23日这15 d作为短期预测期,此后的时间段则划分为长期预测期。初步结果表明,CNN-LSTM-Attention

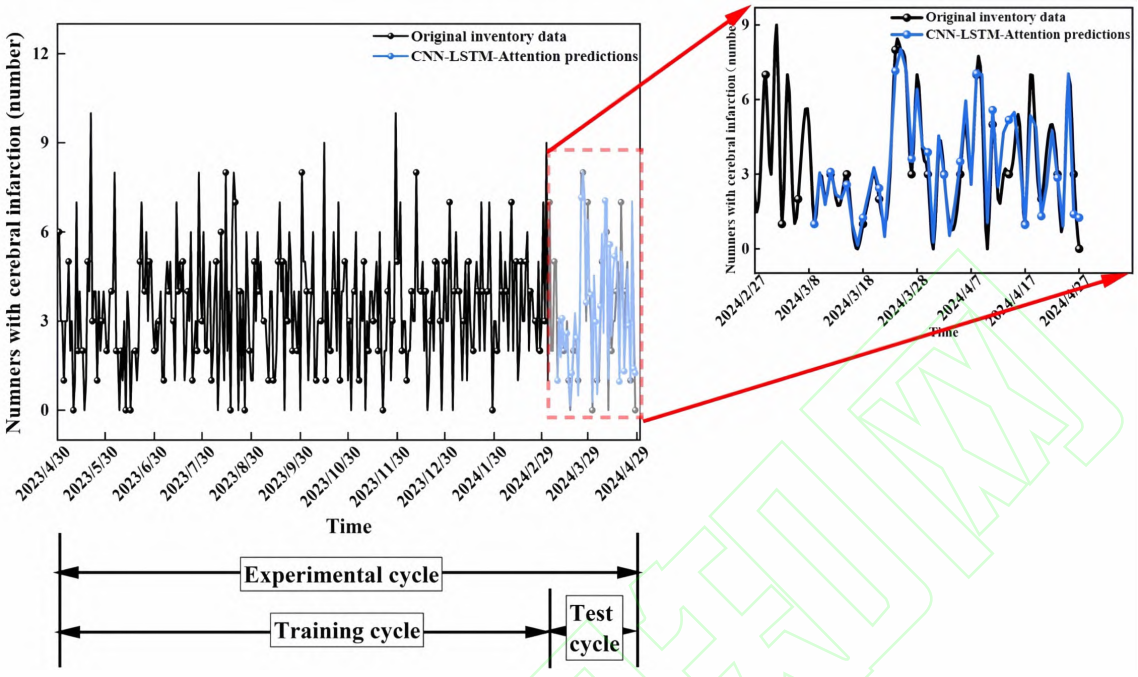

图 4 CNN-LSTM-Attention 模型的预测值与实际值分布图

Fig. 4 Distribution plot of predicted and actual values of CNN-LSTM-Attention model

融合模型既适用于短期预测，也能够应对较长周期的风险预测。这一验证结果说明，该模型在实际应用中具备较强的适应性和稳健性，能够处理多种类型的数据并提供有效的预测支持。

2.3 算法预测性能实验与分析 三种模型（CNN-LSTM-Attention、LSTM 和 CNN-LSTM）对缺血性脑卒中日入院人数的预测值与实际值之间的差异见图 5。其中，深蓝线、浅蓝线和绿线分别表示 CNN-LSTM-Attention 模型、LSTM 模型和 CNN-LSTM 模型基于 2023 年 4 月 30 日—2024 年 3 月 8 日的数据训练后，对 2024 年 3 月 8 日—4 月 29 日期间日入院数的预测结果。预测性能通过最大预测偏差和 RMSE 进行量化评估（表 2）。

结果显示，短期预测中，LSTM 模型最大预测偏差和 RMSE 分别为 2.8 和 1.2，CNN-LSTM 模型分别为 2.0 和 0.8，而 CNN-LSTM-Attention 融合模型分别为 1.5 和 0.6。在长期预测中，LSTM 模型最大预测偏差和预测 RMSE 分别为 19.5 和 5.5，CNN-LSTM 模型分别为 11.2 和 3.3，CNN-LSTM-Attention 融合模型分别为 8.3 和 2.5。结果表明，无论在短期还是长期预测中，本研究建立的融合模型均表现最佳，CNN-LSTM 模型次之，LSTM 模型的预测误差最大。

模型训练的收敛性见图 6 所示，其中训练轮次指模型训练的总次数，损失指单轮次中模型预测值与实际之间的损失值。结果显示，所建立模型的收敛速度显著优于其他两个模型，在相同数据集上，建立的模型不仅收敛边界点（即图中损失曲线拐点）出现得较早，而且收敛的下限也低于其他两种模型，表明该模型在训练过程中的稳定性和效率较高。

2.4 不同滞后时间窗口对预测性能的影响实验与分析 图 7 展示了滞后尺度为 1、3、5 和 7 d 时，所

表 2 3 种模型在长短周期内的预测性能

Tab. 2 Prediction performance of three models in long and short terms

Prediction model	Maximum prediction errors		RMSE	
	Short term	Long term	Short term	Long term
LSTM	2.8	19.5	1.2	5.5
CNN-LSTM	2.0	11.2	0.8	3.3
CNN-LSTM-Attention	1.5	8.3	0.6	2.5

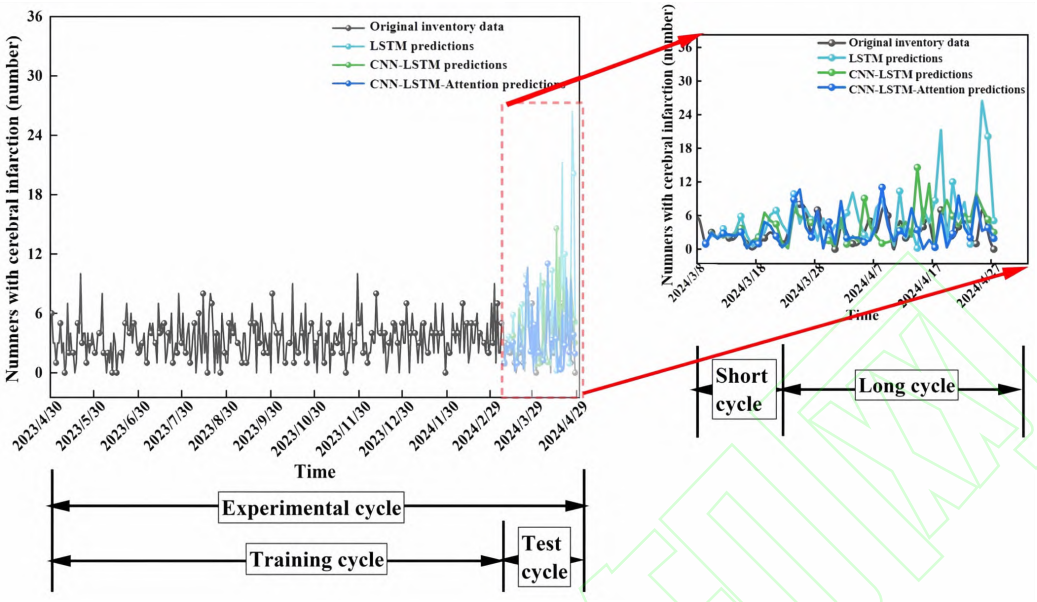

图 5　三种模型在长期和短期内预测值与实际值的差异表现

Fig. 5　Predictive performance comparison of three models in long-term and short-term forecasting against actual values

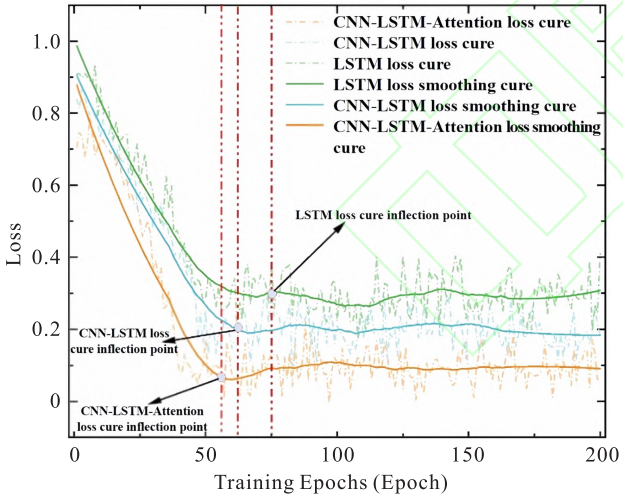

图 6　三种模型的训练损失分布

Fig. 6　Training loss distribution of the three models

建立的模型在长期预测中的预测偏差分布情况。可以看出,在 1 d 和 7 d 的滞后尺度下,预测值与实际值之间的偏差呈现随机不均匀分布,且最大预测偏差较大,相同偏差值对应的出现频率也高于其他滞后窗口;而在滞后 3 d 和 5 d 的尺度下,模型预测值和实际值之间的偏差分布较为均匀,最大偏差较小。此外,表 3 中列出了 5 个滞后尺度下最大预测偏差和 RMSE 的量化评估结果。结果显示,纳入滞后天数后,短期预测中滞后 3 d(最大预测偏差:0.7,RMSE:0.4)和 5 d(最大预测偏差:0.8,RMSE:0.3)的最大预测偏差和 RMSE 均小于滞后 0 d(1.5 和 0.6);而滞后 1 d(1.5 和 0.8)和 7 d(1.9 和 0.9)的两项指标均大于滞后 0 d。在长期预测中,滞后 3 d(5.5 和 1.9)和 5 d(6.5 和 2.0)的最大预测偏差和 RMSE 同样小于滞后 0 d(8.3 和 2.5);滞后 1 d(6.8 和 2.4)和 7 d(7.5 和 2.7)的则均大于滞后 0 d。Attention-CNN-LSTM 融合模型在滞后 3 d 的时间尺度上表现最佳,滞后 5 d 次之,滞后 1 d 和 7 d 表现最差。这一结果进一步地说明了气象因素的滞后效应与缺血性脑卒中发病之间存在较强的相关性。

3　讨论

通过构建基于 CNN-LSTM-Attention 融合模型的缺血性脑卒中发病风险预测模型,实现对多元信息的深度融合和时空特征的充分挖掘。结果表明,该模型能够精准预测缺血性脑卒中风险的变化趋势,与传统时间序列预测方法相比,本模型在相同预测周期内展现出显著的精度优势,RMSE 值显著降低,充分验证了其在脑卒中预测中的高效性。通过纳入气象因素的滞后效应,研究进一步验证了污染物与气象数据对脑卒中发病的潜在影响,增强了模型的实用价值和预测能力。

三种模型在性能上的差异主要源于其对数据特征的处理方式不同。LSTM模型是一种特殊类型的

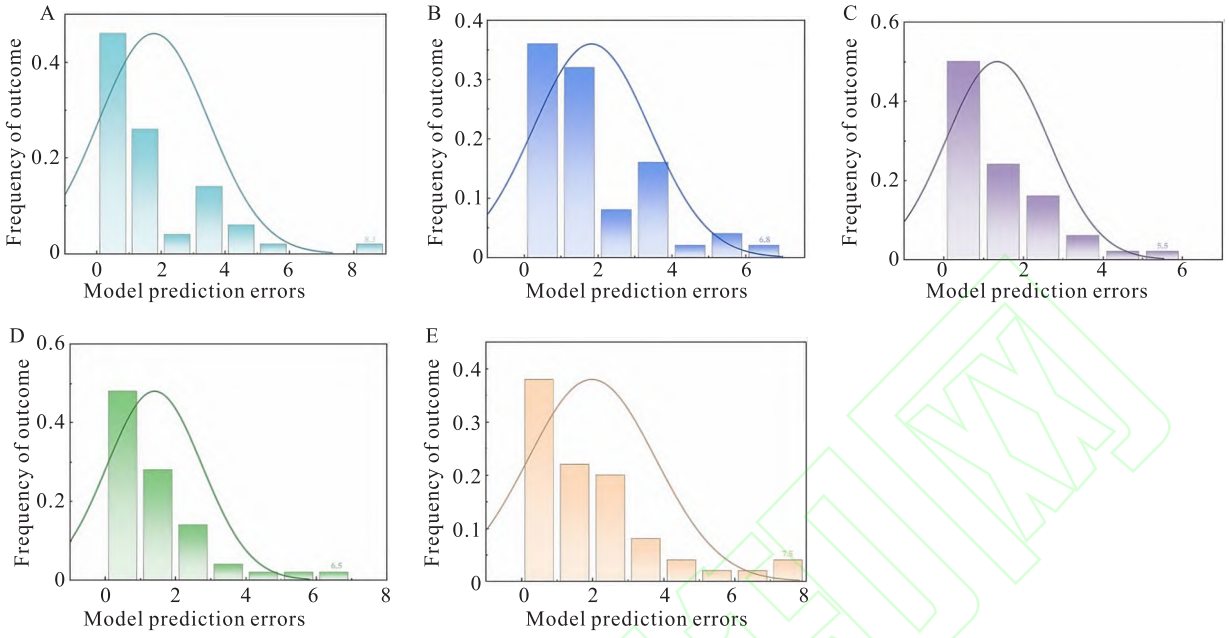

图 7 CNN-LSTM-Attention 模型在预测周期内的预测偏差

Fig. 7 Prediction errors of the CNN-LSTM-Attention model during the prediction period

Figures A – E show the histograms of the model prediction errors for lag step sizes of 0, 1, 3, 5, and 7 days. The bar chart shows the distribution of prediction errors, with error values on the horizontal axis and occurrence frequency on the vertical axis, while the curves represent the normal distribution of the fitted errors.

表 3 CNN-LSTM-Attention 模型在长短周期内的预测性能

Tab. 3 Prediction performance of the CNN-LSTM-Attention models in long and short terms

Prediction model	Maximum prediction errors		RMSE	
	Short term	Long term	Short term	Long term
Lag window = 0	1.5	8.3	0.6	2.5
Lag window = 1	1.5	6.8	0.8	2.4
Lag window = 3	0.7	5.5	0.4	1.9
Lag window = 5	0.8	6.5	0.3	2.0
Lag window = 7	1.9	7.5	0.9	2.7

循环神经网络,适用于捕捉时间序列数据中的时序依赖关系。在气象因素对脑卒中发病预测的研究中,Yang et al[10]研究对比了 LSTM 模型和随机森林模型的预测效果,发现 LSTM 模型在 RMSE 指标上表现更优,表明其在结合气象数据进行脑卒中发病预测时具有显著优势。然而,LSTM 模型固有的固定记忆单元结构在处理涉及多元因素的复杂数据时存在局限性,可能限制模型性能。为克服这一局限,本研究引入 CNN 模块来增强对多元数据的处理能力。Yu et al[11]研究证实了这一改进的有效性,利用 CNN 模型成功分析缺血性脑卒中患者发病前后 12 导联心电图数据,实现了对潜在心房颤动患者的有效识别。这种结合 CNN 和 LSTM 优势的混合模型能够同时考虑多元数据特征和时间序列特征。另一项研究在使用脑电图数据预测脑卒中发病的研究中,通过对比单一 LSTM 模型和 CNN-LSTM 混合模型的性能,进一步验证了混合模型的优越性[12]。然而,尽管 CNN-LSTM 能够捕捉这些信息,它在学习和提取关键特征方面存在一定的不足,尤其是在重点特征的关注与提炼上表现较弱。因此,进一步通过引入 Attention 模型,不仅充分考虑了数据的多元性,还通过 Attention 对关键特征进行了强化学习。这种设计使模型能够更有效地捕捉数据中的潜在规律与关键模式,从而在预测精度、模型稳定性和收敛速度等方面展现出明显优势。

缺血性脑卒中作为一种可防可控的疾病,其发病受到空气污染物和气象因素的影响[13]。Liu et al[14]对中国宁夏 22 个县区的缺血性脑卒中患者进

行分析,发现高温、低相对湿度和 CO 对缺血性脑卒中发病有协同作用。Wang et al[15]研究证实了同时暴露于热浪与 O3 会协同增加缺血性卒中死亡风险。Liu et al[16]研究表明寒潮对缺血性卒中入院的影响具有显著滞后效应,单日效应在滞后第 7 天出现,并于滞后第 14 天达到最大值。同样,Zhao et al[17]发现极端低温在滞后 14 d 时与缺血性脑卒中的发生相关性最大。因此,本研究在分析滞后效应时,从滞后 1 d 开始探索气象因素对缺血性脑卒中发病的影响规律。研究结果显示,随着滞后时间的延长,模型预测效能呈现先上升后下降的趋势,并在滞后 7 d 时出现明显下降。为确保研究结果的科学性和可靠性,本研究最终确定 0 ~ 7 d 的时间尺度作为分析窗口,这与上述研究相符合。该时间窗口既体现了缺血性脑卒中发病与气象因素的累积效应,也有效避免因滞后时间过长导致的预测性能下降,从而提升模型的预测性能和实用性。

本研究构建的 CNN-LSTM-Attention 融合模型,分析了气象数据、临床特征与缺血性脑卒中发病的关联性,提升了缺血性脑卒中风险的预测准确性。通过进一步引入不同滞后时间尺度进行模型预测实验,验证了气象因素对缺血性脑卒中发病的时间滞后效应,为医疗资源合理配置提供更合理的参考依据。该模型综合考量了临床特征与环境因素,在预测性能上优于传统方法,表现出良好的应用潜力。未来可通过融入影像数据及多中心数据,进一步探索其在临床辅助决策与公共卫生干预中的应用价值。

参考文献

[1] Hilkens N A, Casolla B, Leung T W, et al. Stroke[J]. Lancet, 2024, 403(10446): 2820 - 36. doi:10.1016/S0140 - 6736(24)00642 - 1.

[2] 张 艳, 周 霞, 王幼萌, 等. 急性缺血性脑卒中机械取栓术后出血转化及其对预后的影响[J]. 安徽医科大学学报, 2022, 57(6): 987 - 90. doi:10.19405/j.cnki.issn1000 - 1492.2022.06.027.

[2] Zhang Y, Zhou X, Wang Y M, et al. Hemorrhagic transformation after mechanical thrombectomy for acute ischemic stroke and its effect on prognosis[J]. Acta Univ Med Anhui, 2022, 57(6): 987 - 90. doi:10.19405/j.cnki.issn1000 - 1492.2022.06.027.

[3] Zhou L, Wang J, Wu H, et al. Serum levels of vitamin B12 combined with folate and plasma total homocysteine predict ischemic stroke disease: a retrospective case-control study[J]. Nutr J, 2024, 23(1): 76. doi:10.1186/s12937 - 024 - 00977 - 7.

[4] Pan K, Lin F, Huang K, et al. The effect of short-term exposure to air pollution on the admission of ischemic stroke and its interaction with meteorological factors[J]. Public Health, 2025, 239: 103 - 11. doi:10.1016/j.puhe.2024.12.042.

[5] 徐 佩, 樊重俊, 朱人杰, 等. 基于 Prophet-LSTM-PSO 组合模型的医院住院量预测研究[J]. 上海理工大学学报, 2021, 43(1): 68 - 72. doi:10.13255/j.cnki.jusst.20200308003.

[5] Xu P, Fan C J, Zhu R J, et al. Prediction of hospital inpatients based on combined Prophet-LSTM-PSO model[J]. J Univ Shanghai Sci Technol, 2021, 43(1): 68 - 72. doi:10.13255/j.cnki.jusst.20200308003.

[6] 中华医学会神经病学分会, 中华医学会神经病学分会脑血管病学组. 中国急性缺血性卒中诊治指南 2023[J]. 中华神经科杂志, 2024, 57(6): 523 - 59. doi: 10.3760/cma.j.cn113694 - 20240410 - 00221.

[6] Chinese Society of Neurology, Chinese Society of Neurology, Vascular Diseases Group. Chinese guidelines for diagnosis and treatment of acute ischemic stroke 2023[J]. Chin J Neurol, 2024, 57(6): 523 - 59. doi: 10.3760/cma.j.cn113694 - 20240410 - 00221.

[7] Zhang S, Wang J, Pei L, et al. Interpretable CNN for ischemic stroke subtype classification with active model adaptation[J]. BMC Med Inform Decis Mak, 2022, 22(1): 3. doi:10.1186/s12911 - 021 - 01721 - 5.

[8] Wang T, Tian Y, Qiu R G. Long short-term memory recurrent neural networks for multiple diseases risk prediction by leveraging longitudinal medical records[J]. IEEE J Biomed Health Inform, 2020, 24(8): 2337 - 46. doi:10.1109/JBHI.2019.2962366.

[9] Karthik R, Menaka R, Hariharan M, et al. Contour-enhanced attention CNN for CT-based COVID-19 segmentation[J]. Pattern Recognit, 2022, 125: 108538. doi:10.1016/j.patcog.2022.108538.

[10] Yang Y, Zhang M, Zhang J, et al. Medical meteorological forecast for ischemic stroke: random forest regression *vs* long short-term memory model[J]. Int J Biometeorol, 2025, 69(2): 397 - 402. doi:10.1007/s00484 - 024 - 02818 - y.

[11] Yu C C, Peng Y Q, Lin C, et al. ECG-based machine learning model for AF identification in patients with first ischemic stroke[J]. Int J Stroke, 2025, 20(4): 411 - 8. doi:10.1177/17474930241302272.

[12] Choi Y A, Park S J, Jun J A, et al. Deep learning-based stroke disease prediction system using real-time bio signals[J]. Sensors (Basel), 2021, 21(13): 4269. doi:10.3390/s21134269.

[13] He C, Breitner S, Zhang S, et al. Nocturnal heat exposure and stroke risk[J]. Eur Heart J, 2024, 45(24): 2158 - 66. doi:10.1093/eurheartj/ehae277.

[14] Liu Z, Meng H, Wang X, et al. Interaction between ambient CO and temperature or relative humidity on the risk of stroke hospitalization[J]. Sci Rep, 2024, 14(1):16740. doi:10.1038/s41598 - 024 - 67568 - 8.

[15] Wang Z, Zhu L, Peng M, et al. Summer heatwave, ozone pollution and ischemic stroke mortality: an individual-level case-crossover study[J]. Environ Res, 2025, 268: 120818. doi:10.1016/

j. envres. 2025. 120818.

[16] Liu P, Chen Z, Han S, et al. The added effects of cold spells on stroke admissions: differential effects on ischemic and hemorrhagic stroke[J]. Int J Stroke, 2024, 19(2): 217 - 25. doi:10.1177/17474930231203129.

[17] Zhao J, Zhang Y, Ni Y, et al. Effect of ambient temperature and other environmental factors on stroke emergency department visits in Beijing: a distributed lag non-linear model[J]. Front Public Health, 2022, 10: 1034534. doi: 10.3389/fpubh.2022.1034534.

Prediction of ischemic stroke incidence based on Attention-CNN-LSTM model

Liu Jiaming[1], Zhou Xiao[1], Wang Fuyin[1], Sun Xiao[1], Xia Xiaoshuang[1,2], Li Xin[1,2]

([1]*Dept of Neurology, The Second Hospital of Tianjin Medical University, Tianjin* 300211;
[2]*Tianjin Center for Health and Meteorology Multidisciplinary Innovation, Tianjin* 300211)

Abstract ***Objective*** To construct a deep learning model based on convolutional neural network (CNN)-long short term memory network (LSTM)-Attention to explore the correlation between meteorological and clinical factors and the incidence of ischemic stroke. ***Methods*** A fusion model CNN-LSTM-Attention based on CNN, LSTM, and Attention was constructed by incorporating clinical data and meteorological data of ischemic stroke inpatients. The predictive performance of the model was evaluated by maximum prediction error and root mean square error (RMSE). The impact of different lag days on prediction performance was investigated by selecting lag periods ranging from 1 to 7 days. ***Results*** In both short-term and long-term predictions, the CNN-LSTM-Attention fusion model (short-term: 1.5 and 0.6; long-term: 8.3 and 2.5) showed superior maximum prediction bias and RMSE compared to the LSTM model (short-term: 2.8 and 1.2; long-term: 19.5 and 5.5) and the CNN-LSTM model (short-term: 2.0 and 0.8; long-term: 11.2 and 3.3). After incorporating lag days, the maximum prediction deviation and RMSE for lags of 3 days (short-term: 0.7 and 0.4; long-term: 5.5 and 1.9) and 5 days (short-term: 0.8 and 0.3; long-term: 6.5 and 2.0) in both short-term and long-term forecasts were smaller than lags of 0 days (short-term: 1.5 and 0.6; long-term: 8.3 and 2.5). The maximum prediction deviation and RMSE with a lag of 1 day (short-term: 1.5 and 0.8; long-term: 6.8 and 2.4) and 7 days (short-term: 1.9 and 0.9; long-term: 7.5 and 2.7) were both greater than those with a lag of 0 days. ***Conclusion*** The established CNN-LSTM-Attention model demonstrates significant predictive value for the onset of ischemic stroke and can provide reference for the rational allocation of medical resources.

Key words ischemic stroke; meteorological factors; prediction model; convolutional neural networks; long short-term memory networks; attention

Fund programs National Natural Science Foundation of China (No. 42275197); Health and Technology Project of Tianjin (No. TJWJ2023XK007); Key Medical Discipline (Specialty) Construction Project of Tianjin (No. TJYXZDXK-065B); Scientific and Technological Project of Tianjin (No. 21JCZDJC01230)

Corresponding author Li Xin, E-mail: lixinsci@126.com