Advances in Applied Mathematics 应用数学进展, 2023, 12(10), 4330-4337
Published Online October 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1210426

Hans汉斯

基于K近邻算法和混合模型的短期交通流预测

赵丽雅，周文学*

兰州交通大学数理学院，甘肃 兰州

收稿日期：2023年9月17日；录用日期：2023年10月10日；发布日期：2023年10月17日

摘 要

道路拥堵情况由于车辆数量的不断增加而一直存在，及时、准确地进行交通流预测仍是研究的重点。由于交通流数据是庞大的、复杂的，本研究使用KNN算法对数据进行挑选，选择与目标监测点相关性更高的数据，将其输入到CNN-GRU-ATT模型中，对交通流数据进行预测。模型中的CNN层提取特征，GRU层描述时间趋势，ATT层实现对关键信息的关注。实验发现：该模型与其他基线模型相比，模型精度更高，MAPE最高降低了28.33%；与未引入KNN算法相比，模型拟合优度有所提升，达到了97.79%。

关键词

KNN算法，卷积神经网络，门限循环网络，注意力机制，混合模型

Short-Term Traffic Flow Forecast Based on K near Neighbor Algorithm and Hybrid Model

Liya Zhao, Wenxue Zhou*

School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou Gansu

Received: Sep. 17th, 2023; accepted: Oct. 10th, 2023; published: Oct. 17th, 2023

Abstract

Road congestion has always existed due to the increasing number of vehicles, and traffic flow forecasting in time and accurately is still the focus of research. Because traffic flow data is huge and complex, this study uses the KNN algorithm to select the data, select data with higher correlation with the target monitoring point, and enter it into the CNN-GRU-ATT model. Perform predictions. The CNN layer extracts feature in the model, the GRU layer describes time trends, and the ATT

*通讯作者。

文章引用: 赵丽雅, 周文学. 基于K近邻算法和混合模型的短期交通流预测[J]. 应用数学进展, 2023, 12(10): 4330-4337.
DOI: 10.12677/aam.2023.1210426

layer achieves attention to key information. The experiment found that compared with other baseline models, the model has higher accuracy, and MAPE has reduced up to 28.33%; compared with the KNN algorithm, the model fitting superiority has improved, reaching 97.79%.

Keywords

KNN Algorithm, Convolutional Neural Network, Gated Recurrent Unit, Attention Mechanism, Hybrid Model

1. 引言

随着社会科技、经济的发展，汽车数量迅速增加，交通拥堵情况也随之出现。而智能交通系统(ITS)的建设可以有效缓解道路拥堵，缩短出行时间。及时、准确地进行交通流预测必不可少。

早期的交通流预测主要是基于统计分析的，例如自回归、移动平均等模型，这些模型易于理解，但对数据缺失情况难以处理。目前常用的模型分为单一模型的优化和混合模型。其中 XGBoost [1]、SVR [2]、BP [3]、LSTM [4]、BILSTM [5]、GRU [6]模型受到广泛使用。但单一模型的精度可进一步提升，因此出现了一系列优化算法，例如自适应天牛须搜索算法(BAS) [7]、灰狼算法[8] [9]、遗传算法[10]、蚁群算法[11]、鲸鱼算法[12]等，均发现模型精度有所提升。当前混合模型更是交通流预测研究的主要潮流。卷积神经网络能够用来提取特征，与上述单一方法混合有较好的模型精度[2] [4] [13]。基于交通流数据的复杂性和冗余性，K 均值聚类[14]被用于数据的预处理，选取与目标监测点相关性更高的监测点数据。Zhuang 等[15]充分考虑交通流的空间相关性，提出先使用 K 近邻算法对站点数据进行空间筛选，选取相关性高的点，再输入模型进行预测，性能较 LSTM 模型提高了 19%。

自 2014 年注意力机制的提出，其迅速成为研究热点。它能聚焦关键信息，减少干扰。Li 等[16]在模型中加入了自注意力机制，可以抑制长期时间序列信息的丢失，有效提高预测精度。除考虑时间、空间维度外，有的文章也考虑了其他因素的影响。Ma 等[17]将温度、降水量等气象数据与交通流数据进行融合，发现模型效果较好。

本研究基于英国高速公路数据，建立了 KCNN-GRU-ATT 模型来更好地预测短期交通流，KNN 算法对大量数据进行挑选，选择与目标监测点距离更近的监测点数据，CNN 层用于提取特征，GRU 层用于时间趋势的描述，ATT 层用于聚焦关键信息。

2. 模型及算法介绍

2.1. KNN 算法

KNN 的核心思想是计算不同特征值之间的距离，找到最接近目标点的点，并通过加权平均得到结果。本研究采用欧式距离来选择交通流的相关性。KNN 方法是一种典型的基于数据挖掘的方法，传统的未经处理的数据库是巨大的，难以处理的。因此适当地对数据进行预处理，能较好的压缩数据库，并提高预测性能。

KNN 方法步如下：给定一个数据库，包括历史数据和当前数据；确定 K 的初始值；计算新目标监测

点数据与其他数据之间的欧式距离，并按升序排序；根据误差函数，选取最优 K 值，得到与目标监测点数据 K 个最近的邻居作为后续模型的输入。

欧式距离的计算公式如下：

$$d_i = \sqrt{\sum_k \left(x_0(k) - x_i(k)\right)^2} \tag{1}$$

其中，$x_0(k)$ 为目标监测点在 k 时刻监测到的交通流，$x_i(k)$ 为第 i 个监测点在 k 时刻监测到的交通流。

2.2. KCNN-GRU-ATT 预测模型

本文通过建立 KCNN-GRU-ATT 模型来对交通流进行预测，模型整体框架如图 1 所示。

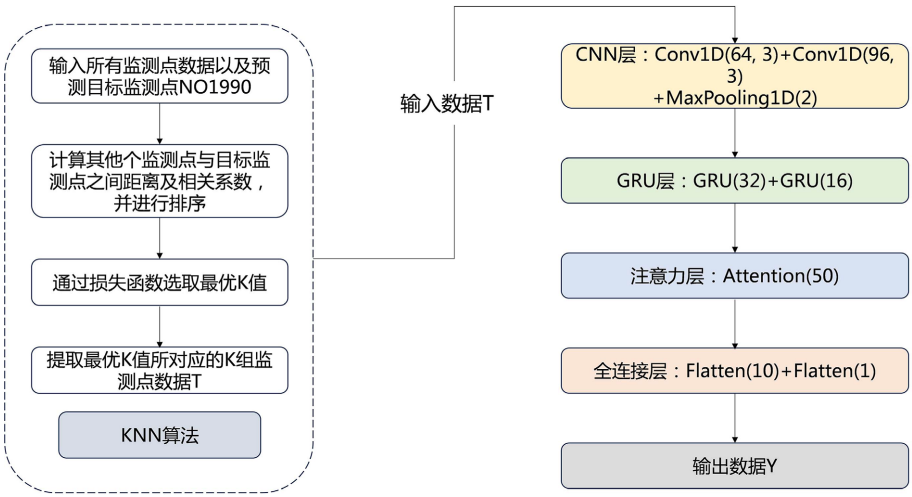

Figure 1. Model process and parameter chart
图 1. 模型流程及参数图

1) KNN 算法。通过选取最优 K 值，得到后续模型的输入数据。

2) CNN 层。主要包括两层一维卷积层、一层最大池化层。卷积层可以提取特征，而最大池化层对上述特征进行挑选，防止过拟合；

3) GRU 层。主要接收上层的输出数据，使用两层 GRU 层可以更加充分的提取潜在信息；

4) 注意力层。GRU 层会随着时间长度的增加而出现信息丢失和梯度消失问题。注意力机制可以保留重要信息，减少其他信息的干扰；

5) 输出层。本文建立两层全连接层，最终输出预测数据。

3. 数据来源及预处理

3.1. 数据来源

本文使用数据为 2023 年 1 月 1 日到 2 月 28 日共 2 个月的英国高速公路 M25 上 46 个摄像头监控的交通流数据，以 15 min 为时间间隔，每个摄像头各采集 5664 条数据。

3.2. 数据预处理

由于收集到的数据中存在缺失、异常等问题，需要对数据进行预处理。删除缺失严重的 6 个监测点数据，使用箱线图进行异常值检验时，发现并不存在异常值。但是查看数据之后，发现存在一些不切实

际的“0”值，因此将上述异常值当作缺失值进行处理。考虑到交通流数据的交替情况，使用均值替代有所不妥，本文使用前值进行填补。本文对数据进行归一化，使用的是 min−max 方法，公式如下：

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{2}$$

其中，x_i' 为归一化后数据，x_i 为原始数据，$x_{\min}$、$x_{\max}$ 分别为最小、最大值。

4. 实验结果与分析

4.1. 评价指标

本文采用四种回归算法中常用的评价指标来评价模型的效果，分别是 R^2 、RMSE、MAE 和 MAPE，计算公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y - \hat{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2}, \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|, \tag{5}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|. \tag{6}$$

4.2. 实证分析

本文采用 PYTHON 软件进行模拟研究。为了验证 K 近邻算法中 K 值是否对交通流预测结果产生影响，因此选择不同 K 值进行预测。结果显示，不同的 K 值对预测性能有显著影响。由图 2 可以看出，当 K = 5、6、9，即相关检测点个数为 5、6 或者 9 时，预测效果最好，损失值达到最低。

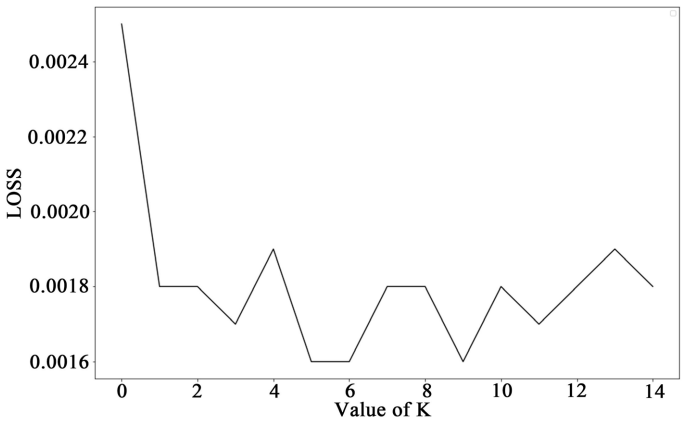

Figure 2. Different K values corresponding to loss values
图 2. 不同 K 值对应的损失值

为了更直观地观察多个监测点之间的数据相关性，将所有监测点的一天数据共 96 个样本进行可视化。从图 3 可以观察到，NO550、NO5066、NO10349 与其他监测点的数据差异较大。之后对所有监测点

的与目标监测点的相关性进行分析，从图 4 可以发现相关性系数在 98%以上的有 5 个监测点，分别是 NO547、NO5875、NO4145、NO2097、NO3437。

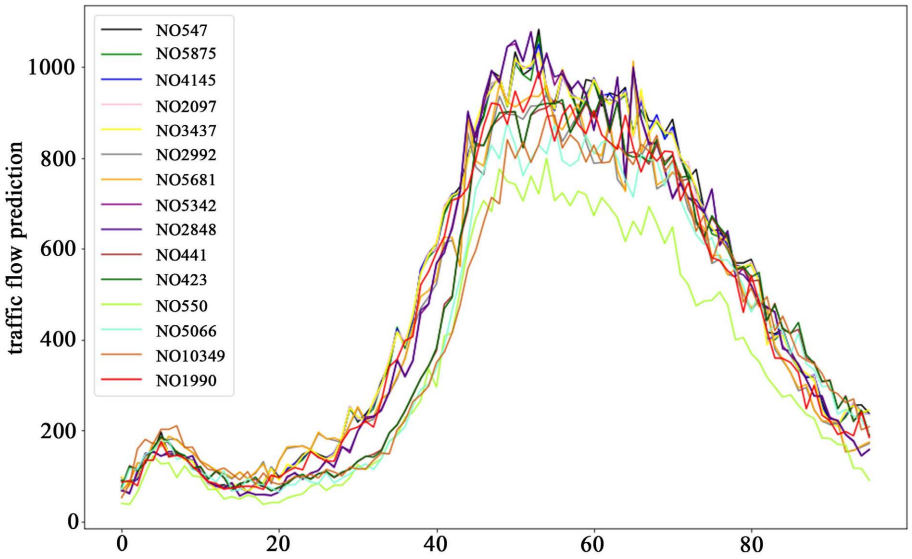

Figure 3. All monitoring points one day data distribution map
图 3. 所有监测点一天数据分布图

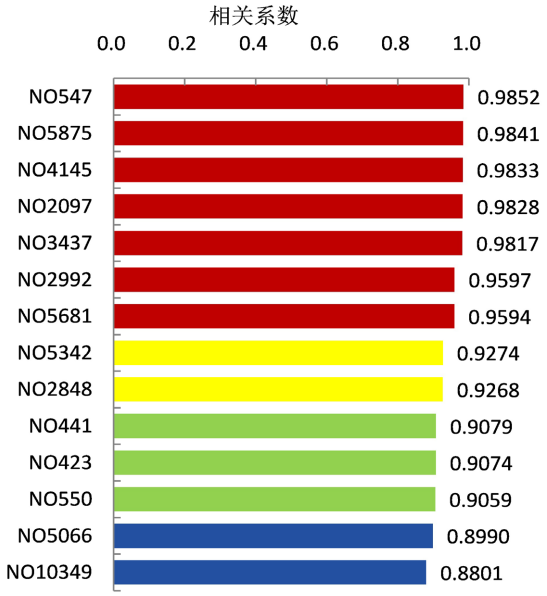

Figure 4. Analysis of the correlation point of all monitoring points and target monitoring points
图 4. 所有监测点与目标监测点的相关性分析

为了找出最优 K 值，将 K = 5、K = 6、K = 9 分别进行建模，得到的模型结果见表 1。从表 1 可以看出，K = 6 时，R^2、RMSE、MAE、MAPE 分别为 0.9779、59.7029、40.0641、0.0898，预测结果略优于 K = 5 和 K = 9 时。从表 2 可以看出，K = 6 时，对应的监测点编号分别为 NO547、NO5875、NO4145、NO2097、NO3437、NO2992。

Table 1. Evaluation indicators corresponding to different K values
表 1. 不同 K 值对应的评价指标

K 值	R^2	RMSE	MAE	MAPE
K = 5	0.9773	60.4967	40.2288	0.0898
K = 6	**0.9779**	**59.7029**	**40.0641**	**0.0886**
K = 9	0.9777	59.9726	40.2436	0.0919

Table 2. Monitoring point combination corresponding to different K values
表 2. 不同 K 值对应的监测点组合

K 值	监测点组合
目标监测点	NO1990
K = 1	NO1990、NO547
K = 2	NO1990、NO547、NO5875
K = 3	NO1990、NO547、NO5875、NO4145、NO2097、NO3437、NO2992
K = 4	NO1990、NO547、NO5875、NO4145、NO2097、NO3437、NO2992
K = 5	NO1990、NO547、NO5875、NO4145、NO2097、NO3437、NO2992
K = 6	**NO1990、NO547、NO5875、NO4145、NO2097、NO3437、NO2992**
⋮	
K = 14	NO1990、NO547、NO5875、NO4145、NO2097、NO3437、NO2992、NO5681、NO5342、NO2848、NO441、NO423、NO550、NO5066、NO10349

为了进一步评价KCNN-GRU-ATT模型的有效性，本文选择SVR、LSTM、GRU、CNN-GRU、GRU-ATT、CNN-GRU-ATT 模型作为基线模型进行对比。从图 5 可以看出，SVR、LSTM 对数据的拟合效果较差，之后绘制一天的数据拟合效果图(见图 6)，发现 GRU 模型对最高值处的拟合效果较优，而对持续上升的数据拟合效果较差，其他模型在拟合图中差别较小。

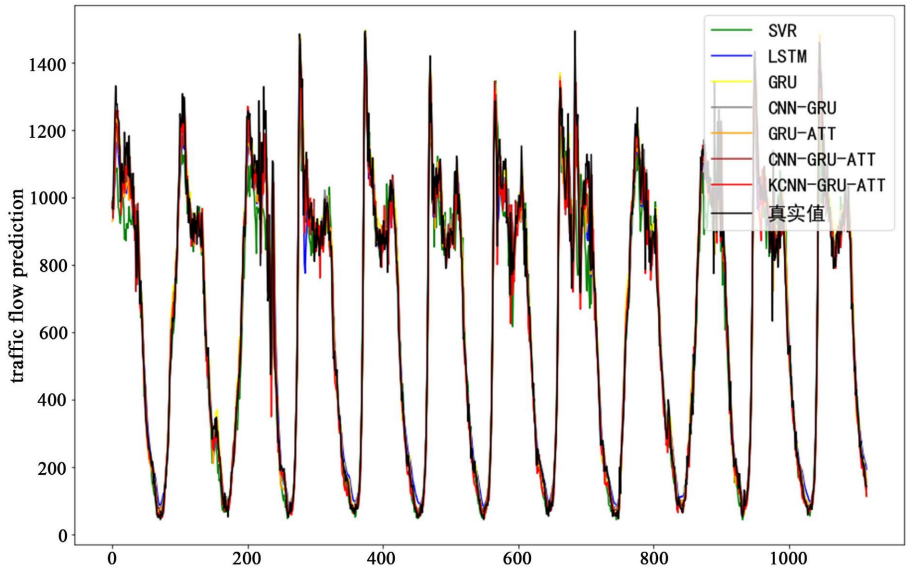

Figure 5. Fitting renderings of each model
图 5. 各模型的拟合效果图

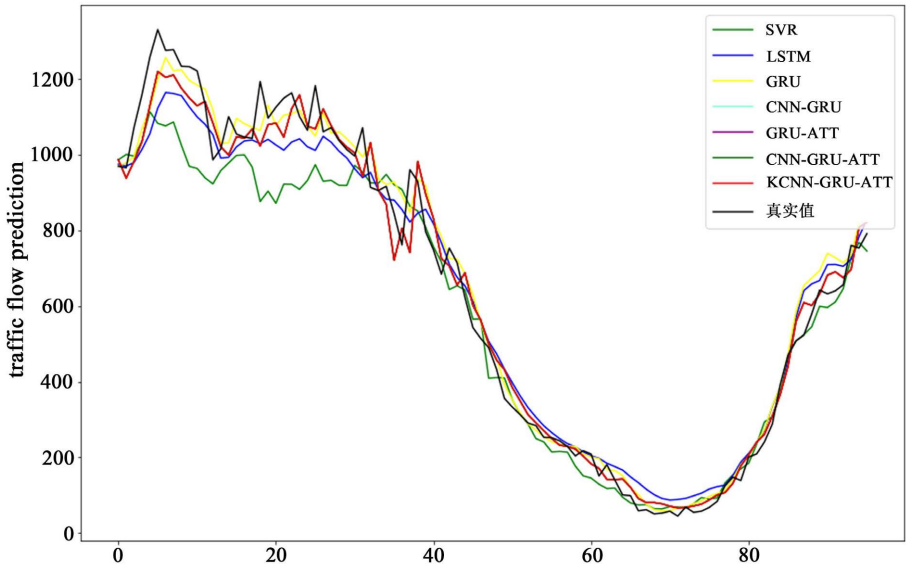

Figure 6. The fitting renderings of each model one day
图 6. 各模型一天的拟合效果图

因此从表 3 来看，文章提出模型的 MAPE 较 CNN-GRU、GRU-ATT、CNN-GRU-ATT 模型分别提高 8.35%、7.45%、9.14%、0.68%。在均方误差上有较大的降低，最少减低了 18.54%，本文模型较好。

Table 3. Evaluation indicators of different models
表 3. 不同模型的评价指标

模型	R^2	RMSE	MAE	MAPE
SVR	0.9423	96.6341	63.6032	1.9684
LSTM	0.9610	79.2601	53.5860	0.1137
GRU	0.9667	73.2808	49.2921	0.0960
CNN-GRU	0.9678	71.9996	47.4408	0.0952
GRU-ATT	0.9669	73.0234	47.7496	0.0967
CNN-GRU-ATT	0.9689	70.7743	45.4032	0.0892
KCNN-GRU-ATT	**0.9779**	**59.7029**	**40.0641**	**0.0886**

5. 结论

本研究采用 KCNN-GRU-ATT 组合模型对短期交通流进行预测。在之前的文章中，已建立 SVR、LSTM、GRU、CNN-GRU、GRU-ATT、CNN-GRU-ATT 多个模型进行预测，这次加入 K 近邻算法，来检验监测点的空间相关性。通过不同 K 值的选取，对与目标监测点之间的相关性进行排序，最终选择最适合的 K 值，将数据输入本模型。实验结果显示，该模型较之前模型在均方误差上有较大降低，其他评价指标也表现更好。本文提出模型虽然预测效果较好，但考虑因素比较单一，之后的研究中可以加入气象数据进行分析。

参考文献

[1] 焦朋朋, 安玉, 白紫秀, 等. 基于 XGBoost 的短时交通流预测研究[J]. 重庆交通大学学报(自然科学版), 2022,

41(8): 17-23+66.

[2] 罗文慧, 董宝田, 王泽胜. 基于 CNN-SVR 混合深度学习模型的短时交通流预测[J]. 交通运输系统工程与信息, 2017, 17(5): 68-74.

[3] 姚洁, 邱劲. 基于 SSA-BP 算法的道路交通流量预测研究[J]. 西南大学学报(自然科学版), 2022, 44(10): 193-201. https://doi.org/10.13718/j.cnki.xdzk.2022.10.020

[4] 赵明伟, 张文胜, 王克文, 等. 基于 EMD-PSO-LSTM 组合模型的城市轨道交通短时客流预测[J]. 铁道运输与经济, 2022, 44(7): 110-118. https://doi.org/10.16668/j.cnki.issn.1003-1421.2022.07.17

[5] 杜秀丽, 范志宇, 吕亚娜, 等. 基于双向长短期记忆循环神经网络的网络流量预测[J]. 计算机应用与软件, 2022, 39(2): 144-149+156.

[6] Jin, F. and Zhao, B. (2019) Short-Term Traffic Flow Prediction Based on Road Network Topology. *Journal of Beijing Institute of Technology*, **28**, 383-388. https://doi.org/10.15918/j.jbit1004-0579.18001

[7] 李巧茹, 刘桂欣, 陈亮, 等. 自适应 BAS 优化 RBF 神经网络的短时交通流预测[J]. 哈尔滨工业大学学报, 2023, 55(3): 93-99.

[8] 张兴辉, 樊秀梅, 阿喜达, 等. 反向学习的灰狼算法优化及其在交通流预测中的应用[J]. 电子学报, 2021, 49(5): 879-886.

[9] 张文胜, 郝孜奇, 朱冀军, 等. 基于改进灰狼算法优化 BP 神经网络的短时交通流预测模型[J]. 交通运输系统工程与信息, 2020, 20(2): 196-203.

[10] Li, Y.F., Chen, M.N., Lu, X.D., *et al*. (2018) Research on Optimized GA-SVM Vehicle Speed Prediction Model Based on Driver-Vehicle-Road-Traffic System. *Science China Technological Sciences*, **61**, 782-790. https://doi.org/10.1007/s11431-017-9213-0

[11] 蒋杰, 张江鑫. 改进 ACO 优化的 BP 神经网络短时交通流量预测[J]. 计算机仿真, 2021, 38(7): 97-101+180.

[12] 胡松, 成卫, 李艾. 一种改进鲸鱼算法及其在短时交通流预测中的应用研究[J]. 小型微型计算机系统, 2021, 42(8): 1627-1632.

[13] 蒲悦逸, 王文涵, 朱强, 等. 基于 CNN-ResNet-LSTM 模型的城市短时交通流量预测算法[J]. 北京邮电大学学报, 2020, 43(5): 9-14.

[14] Li, R., Huang, Y. and Wang, J. (2019) Long-Term Traffic Volume Prediction Based on K-means Gaussian Interval Type-2 Fuzzy Sets. *IEEE/CAA Journal of Automatica Sinica*, **6**, 1344-1351. https://doi.org/10.1109/JAS.2019.1911723

[15] Zhuang, W. and Cao, Y. (2023) Short-Term Traffic Flow Prediction Based on a K-Nearest Neighbor and Bidirectional Long Short-Term Memory Model. *Applied Sciences*, **13**, Article 2681. https://doi.org/10.3390/app13042681

[16] Li, Z., Wang, X. and Yang, K. (2023) An Effective Self-Attention-Based Hybrid Model for Short-Term Traffic Flow Prediction. *Advances in Civil Engineering*, **2023**, Article ID: 9308576. https://doi.org/10.1155/2023/9308576

[17] Ma, F., Deng, S. and Mei, S. (2023) A Short-Term Highway Traffic Flow Forecasting Model Based on CNN-LSTM with an Attention Mechanism. *Journal of Physics*: *Conference Series*, **2491**, Article ID: 012008. https://doi.org/10.1088/1742-6596/2491/1/012008