Frontiers in Traffic and Transportation Engineering (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ftte.2024.040101
ISSN 2616-2334 Vol. 4 Num. 1

# *Research on Traffic Congestion Prediction Based on XGBoost*

**Wei Yu[1], Feifei Xie[2]**

[1]*School of Automotive and Transportation Engineering, Jiangsu University, Zhenjiang, 212013, China*
[2]*School of Business, Anhui University, Hefei, 230601, China*

***Keywords:*** Congestion, XGBoost Algorithm, Spatio-Temporal Distribution, Feature Importance Ranking

***Abstract:*** Accurate prediction of road congestion is imperative for improving road utilization, thereby reducing economic losses and enhancing traffic management efficiency. Employing the XGBoost algorithm, this study integrates both temporal and spatial dimensions into the prediction of road congestion. Analysis of road congestion box plots across various coordinates and directions reveals significant disparities in traffic congestion coefficients, indicating a close relationship between the spatial dimension and traffic congestion conditions. Additionally, discernible variations in congestion coefficients between weekdays and non-workdays highlight a crucial association between traffic congestion conditions and time. The model incorporates spatial and temporal data to predict and simulate real Chicago road traffic conditions. Comparative analysis between actual and predicted values demonstrates the model's alignment with real data, attesting to its excellent predictive efficiency. Finally, elucidation of the influence of each variable on the traffic congestion prediction model is achieved through the feature importance ranking.

## 1. Introduction

In recent years, with the acceleration of urbanization, the problem of urban traffic congestion has become more and more serious, seriously affecting people's travel efficiency and quality of life. Therefore, the use of big data mining technology to predict traffic congestion and propose improvement measures is a current topic of concern in the field of traffic management [1]. With the complex factors of spatial and temporal variations of traffic flow, there is a need to find an effective big data mining method to predict traffic congestion, so as to improve the utilization rate of traffic roads.

At present, in the research of traffic congestion prediction, our scholars Tang Zhikang et al. design a traffic congestion prediction model based on the Bagging integrated learning method, which considers the influence of various environmental factors on the traffic condition and improves the comprehensiveness of the prediction [2]; GU Li-qiong et al. construct a MM-SVR model based on the congestion indicator to improve the support vector machine model for road congestion prediction, which greatly improves the predictive accuracy [3].

This paper employs the XGBoost algorithm to systematically analyze road congestion, considering

both spatial and temporal dimensions to forecast future traffic conditions. Spatially, discernible disparities in road congestion across various coordinates and directions are evident, while temporally, significant differences in traffic congestion coefficients between weekdays and non-weekdays are observed. Subsequently, the XGBoost algorithm is applied to assess feature importance, identifying key variables influencing traffic congestion prediction accuracy. Model simulation and prediction indicate a strong alignment between actual and predicted values, with prediction accuracy assessed by calculating the Mean Absolute Error. Results demonstrate the effectiveness of the prediction method in accurately forecasting road conditions, providing valuable insights for urban transportation management and construction in China.

## 2. Principles of the methodology

XGBoost is a model based on GBDT, and its model structure is similar to GBDT in that it is based on a decision tree, and integrates weak classifiers into strong classifiers through continuous iteration (e.g., the Figure 1 shown). [4] On the one hand, compared to GBDT which only uses first-order derivatives, the XGBoost algorithm carries out second-order derivatives to make the loss function more accurate and adds a regular term in the objective function, $\Omega(f_k)$ which is used to control the complexity of the model to avoid overfitting; on the other hand, the XGBoost algorithm is an advanced version of the gradient boosting decision tree, which grows the decision tree by constant feature splitting, and in the process of learning the decision tree, it fits the error between the actual value and the predicted value of the model to improve the accuracy of prediction as shown in Figure 1 and Figure 2.[5]

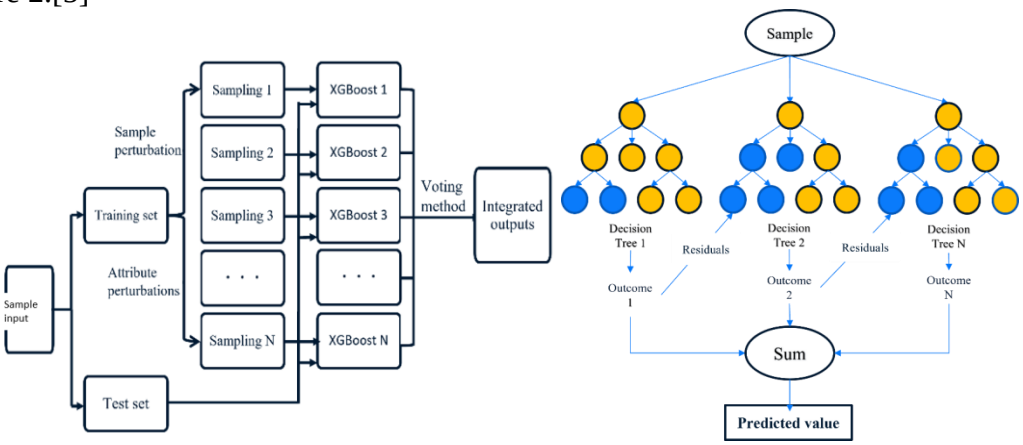



Figure 1: XGBoost algorithm integration flow (Left)

Figure 1: Flowchart of the gradient boosting decision tree model (Right)

### 2.1 Construct the objective function

The XGBoost algorithm is an additive model consisting of$k$ an additive model consisting of individual decision tree models, and the prediction accuracy improves as the number of model iterations increases. Assuming that the first$k$ iteration of the tree model to be trained is $f_k(x)$ The final prediction for the first$i$ the final prediction for the first sample is:

$$\hat{y}_i^{(k)} = \sum_{k=1}^{k} f_k(x_i) = \hat{y}_i^{(k-1)} + f_k(x_i) \tag{1}$$

Eq:$\hat{y}_i^{(k)}$ --Previous$k$ round model prediction value;$\hat{y}_i^{(k-1)}$--- previous$k-1$ round of model predictions;$f_k(x_i)$--th$k$ decision tree model.

To optimize the objective function, Taylor series, regularized expansion is used to combine the primary and secondary function coefficients and k iterations are performed to obtain the formula for the final prediction:

$$Obj_k = \Omega(f_k) + \sum_{k=1}^{n} \left[ l\left(y_i, \hat{y}_i^{(k-1)}\right) + g_i \cdot f_k(x_i) + \frac{1}{2} h_i \cdot f_k^2(x_i) \right] \tag{2}$$

Eq: $\hat{y}_i^{(k-1)}$ --Previous $k-1$ round model prediction value; $l\left(y_i, \hat{y}_i^{\langle k-1 \rangle}\right)$ --sample $x_i$ of the training error; $g_i, h_i$--pre-training $k-1$ Residuals at tree; $\Omega(f_k)$--the $k$ regular term of the tree.where the $\Omega(f_k)$ The canonical term indicates the complexity of the structure of this model, the smaller the result, the more accurate the prediction.

## 3. Result

### 3.1 Data sources

The data in this paper is from Chicago Traffic Tracker-Historical Congestion Estimates by Segment Specific parameters are as follows: Row_id--a unique identifier for each record. Time--20 minutes of time for each measurement. x--Coordinates of the east-west midpoint of the roadway. y-- Coordinates of the north-south midpoint of the roadway. Direction--the direction of travel of the road. For example: East Boulevard (EB) means "east" direction of travel, North Boulevard (NB) means "north" direction of travel, South Boulevard (SB) means "south" direction of travel, West Boulevard (WB) means "west" direction of travel, and Northeast (NE) means "west" direction of travel. South Boulevard (SB) means "southbound" direction of travel, West Boulevard (WB) means "westbound" direction of travel, Northeast (NE) means "northeast" direction of travel, and Northwest (NW) for "northeast" direction of travel. Northwest (NW) means the "northwest" direction of travel, Southeast (SE) means the "southeast" direction of travel, and Southwest (SW) means the "southwest" direction of travel. Southeast (SE) means "Southeast" direction of travel and Southwest (SW) means "Southwest" direction of travel. Congestion--the hourly level of congestion on roads, normalized to a range of 0 to 100.

### 3.2 Data processing

The Chicago traffic data for a specific road section is systematically organized. Utilizing a 20-minute time interval, the statistics detailing the road conditions for this section are presented in the table. Specifically, the data spans from April 1 to September 30 of a given year. Table 1 displays both the congestion factor distribution for this road, while Figure 3 provides a visual representation of the congestion factor distribution.

Table 1: Roadway congestion data within the period April 1-September 30

| row_id | Time | x | y | direction | congestion |
|---|---|---|---|---|---|
| 0 | 04-01 00:00:00 | 0 | 0 | EB | 70 |
| 1 | 04-01 00:00:00 | 0 | 0 | NB | 49 |
| ... | ... | ... | ... | ... | ... |
| 848833 | 09-30 11:40:00 | 2 | 3 | SW | 17 |
| 848834 | 09-30 11:40:00 | 2 | 3 | WB | 24 |

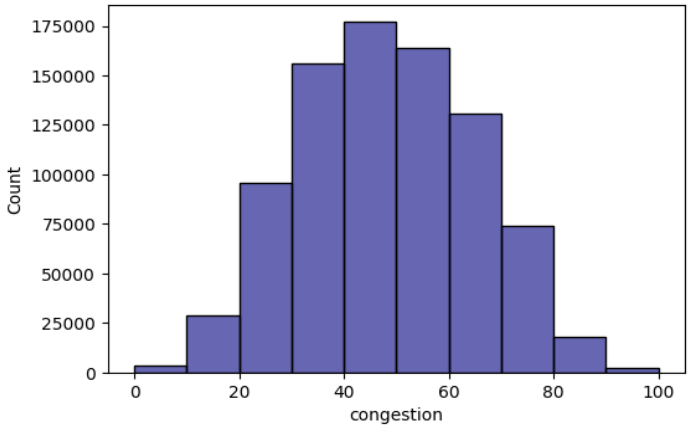



Figure 2: Histogram of road congestion factor

Figure 3 reveals a roughly normal distribution of images, suggesting that the congestion factor is predominantly concentrated within the range of 40-50. Due to space constraints, only pertinent data for the eastbound route on April 1 of a specific year is selected for display in Table 2.

Table 2: Data for eastbound travel routes on April 1, 00:00:00

| row_id | time | x | y | direction | congestion |
|---|---|---|---|---|---|
| 0 | 04-01 | 0 | 0 | EB | 70 |
| 3 | 04-01 | 0 | 1 | EB | 18 |
| ... | ... | ... | ... | ... | ... |
| 51 | 04-01 | 2 | 2 | EB | 42 |
| 59 | 04-01 | 2 | 3 | EB | 39 |

### 3.3 Spatial regularity

Table 3 presents road condition data for all roads with different coordinates and directions on April 1st.

Table 3: Orientation of different road directions

| row_id | x | y |
|---|---|---|
| 0 | 0 | 0 |
| 3 | 0 | 1 |
| ... | ... | ... |
| 51 | 2 | 2 |
| 59 | 2 | 3 |

By studying congestion across various coordinates and directions in space, it becomes plausible to predict future congestion at the nearest point in a given direction. For instance, congestion at (0-1-EB) may exhibit correlation with future congestion at (1-1-EB). Recognizing the bidirectional nature of this correlation, a reverse prediction can be attempted as well: congestion at (1-1-EB) could be correlated with past congestion at (0-1-EB). Figure 4, depicted below, illustrates the roadways at 65 different coordinates and directions.

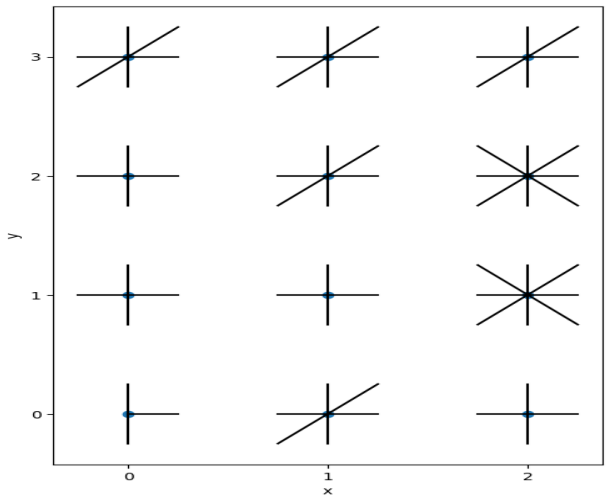
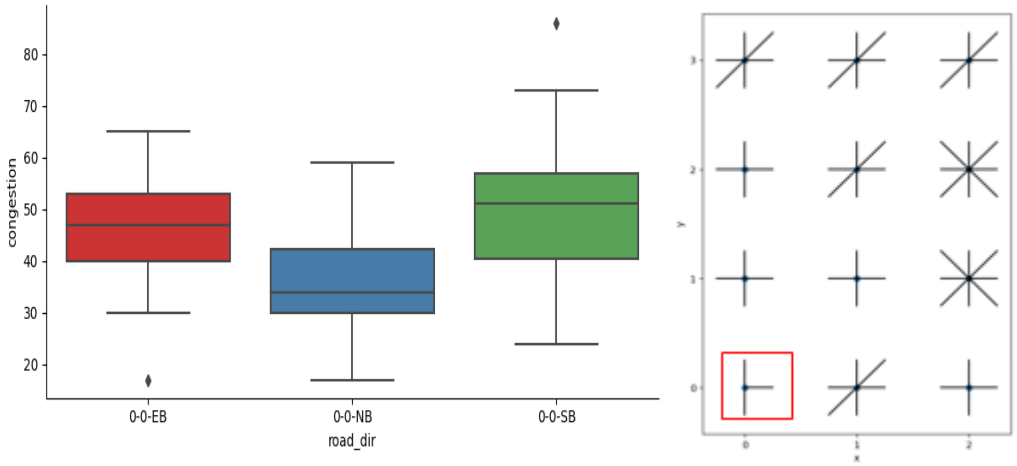



Figure 3: Schematic diagram of different coordinates and directions

As shown in Figure 4, a total of 12 coordinate points and 8 directions ('EB,' 'NB,' 'SB,' 'WB,' 'NE,' 'SW,' 'NW,' 'SE') form 65 distinct road combinations involving various geographic locations and directions. Each geographic location is associated with a minimum of 3 and a maximum of 8 directions. The dataset comprises 65 road combinations, offering a substantial volume of data for model training.

To illustrate, Figure 5 showcases a block diagram representing the congestion situation for roads with different directions at geographic coordinates x=0, y=0 on April 8.



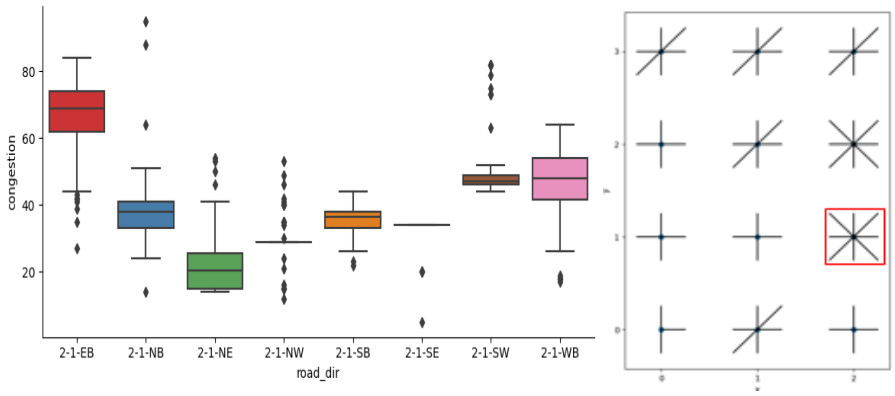

(a) (b)

Figure 4: Distribution of road congestion factor in different directions on April 8 (x=0, y=0)

In Figure 5, the analysis highlights the median of the boxplot as indicative of the concentration trend within the dataset. Specifically, considering the medians, (0-0-SB) exhibits the highest congestion factor but is situated in the upper portion of the box center, while (0-0-NB) has the lowest median but is positioned in the lower part of the box center. Regarding the median's position relative to the box, (0-0-SB) and (0-0-EB) have upper-half medians, signifying a left-skewed distribution where most congestion coefficients are smaller than the median. Conversely, (0-0-NB) displays a lower-half median, indicating a right-skewed distribution where most congestion coefficients surpass the median. Outlier data, beyond the upper and lower extents of the boxes in the boxplots, can be identified by observing outliers in different directional boxplots. Overall, notable differences in congestion among these three directions suggest that road congestion is not concentrated in the

directional dimension.

As an illustrative example, on April 8, at the geographic location x=2, y=1, statistical data from various road directions are utilized to construct a congestion box plot, presented in Figure 6.




(a) (b)

Figure 5: Distribution of roadway congestion factors in different directions on April 8 (x=2, y=1)

In Figure 6, the analysis indicates that the median of the boxplot serves as a key indicator of the concentration trend within the dataset. Specifically, among the medians, (2-1-EB) displays the highest congestion factor, while (2-1-NE) exhibits the lowest median. Considering the median's position relative to the box, (2-1-NE), (2-1-EB), and (2-1-WE) have medians in the middle of the box, suggesting a roughly normal distribution of the congestion factor for these directions. This implies that most congestion factors are approximately equal to the median. On the other hand, the medians of (2-1-SB) and (2-1-NB) are in the upper half of the box, indicating left-skewed distributions, where most congestion coefficients are smaller than the median. Conversely, the median of (2-1-SW) is in the lower half of the box, signaling a right-skewed distribution, with most congestion coefficients surpassing the median. Outlier data beyond the upper and lower extents of the boxes in the boxplots are identified, notably more prevalent in (2-1-NW), as observed through careful examination of the boxplots in different directions. Overall, marked differences in congestion across various directions underscore the complexity of the congestion prediction space.

## 3.4 Time regularity

Given the intricacies associated with traffic forecasting, an exploration of congestion patterns is undertaken at the temporal level. Time series data for congestion coefficients on weekdays (April 8) and non-weekdays (April 13) are selected and presented in Figures 7 and 8, respectively.

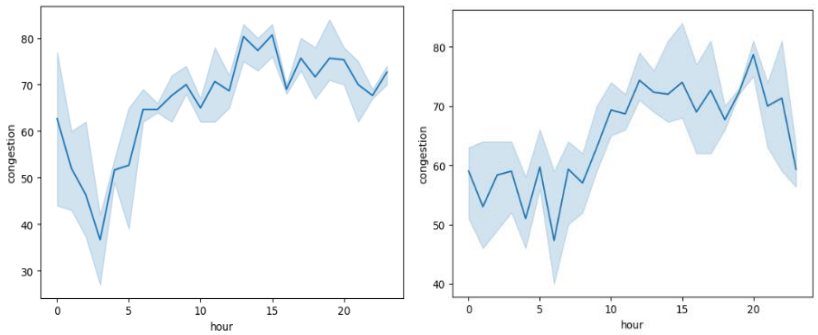



Figure 6: Time series plot for weekday (April 8) (Left)

Figure 7: Time Series Plot of Non-Working Days (April 17) (Right)

Figures 7 and 8 reveal a shared trend where congestion factors on both April 8 and April 13 begin to rise around 7:00, reaching a peak around 20:00 before gradually declining. Notably, on weekdays, the congestion coefficients peak during 8:00-9:00, 12:00-13:00, and 18:00-19:00, aligning with typical morning and evening rush hours. Conversely, on non-weekdays, the congestion coefficients peak during 19:00-20:00, corresponding to the evening peak for weekend trips. This observation underscores the efficacy of the congestion coefficient index proposed in this study, affirming its ability to accurately reflect the actual road congestion conditions.

### 3.5 Model predictions

The acquired road state variables constitute the sample set for training a model using the XGBoost algorithm. Subsequently, the trained model is applied to predict road states and trends for the following day. For efficient model training and evaluation, 80% of the total data is randomly designated as the training set, leaving the remaining 20% as the test set for data prediction. The time range for the test set is selected randomly. In this paper, to illustrate the process, only 100 data points are chosen for demonstration, as depicted in Figure 9. The predictive accuracy of the model is assessed by computing the Mean Absolute Error (MAE) between the predicted and actual values.

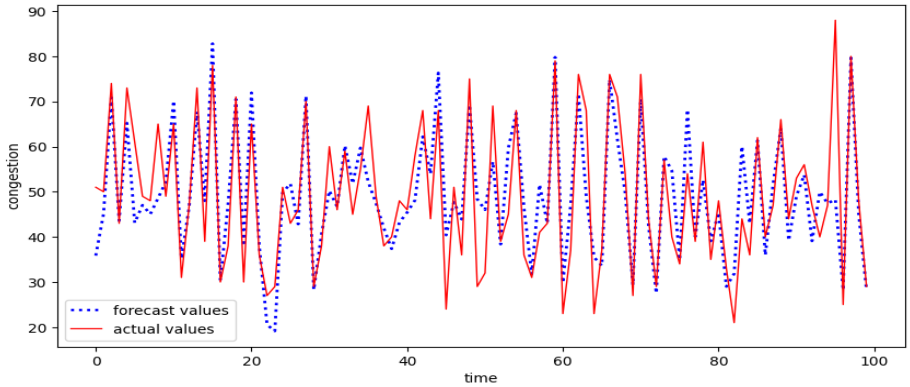



Figure 8: Comparison of predicted and real data

Figure 9 presents a comparative illustration between the test data and the actual data, focusing on the initial 100 data points to address data redundancy. Through careful comparison and analysis of the image, it is observed that the predictions generated by the model utilized in this study align well with the actual data. Simulation calculations yield a Mean Absolute Error (MAE) value of 6.1665, indicating a high level of prediction accuracy.

### 3.6 Feature Importance Ranking

The XGBoost algorithm, a widely employed additive model for big data mining, consists of decision tree models. In contrast to "black box" algorithms like neural networks and GBDT, the XGBoost algorithm, utilized in this study, is distinctive for its foundation on the Congestion index. This allows for the prioritization of variables, rendering an interpretable analysis that specifically delineates the contribution of feature variables to road congestion prediction. The established XGBoost algorithm prediction model in this paper considers the congestion coefficient as the dependent variable, incorporating 14 feature variables such as weekdays, non-workdays, and different coordinates and directions. By computing the congestion coefficient for each feature variable, a ranking of variables is obtained, elucidating their respective contributions. This facilitates the determination of variable importance in relation to their impact on road congestion. High-importance feature variables are retained, while those with lower contributions are excluded, thereby simplifying

the model, enhancing prediction accuracy, and improving efficiency.

The paper conducts a comprehensive analysis of road congestion at spatial and temporal levels, comparing the characteristics of different roads. The resultant order of importance for these characteristics is depicted in Figure 10.

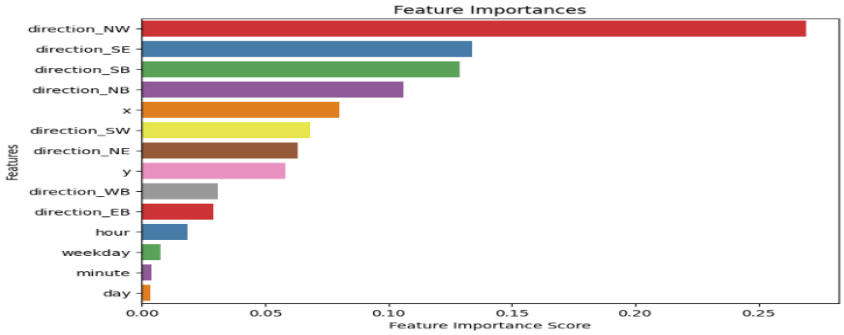



Figure 9: Importance ranking of features

From Figure 10, it can be seen that the importance of road direction and location is higher than that of time, which reflects that the prediction of road congestion mainly depends on the spatial distribution of roads, but also has a certain connection with time.

## 4. Conclusion

This paper employs the XGBoost algorithm prediction model to construct an evaluation model for assessing road congestion conditions. Specifically, it establishes a road congestion coefficient index based on historical data, enabling the prediction of road congestion conditions at specific times. The distinctive aspect of this study lies in its exploration and analysis of both temporal and spatial dimensions of road conditions. It identifies variables with high correlation to the prediction model, aiming to enhance prediction accuracy. Results from prediction experiments demonstrate the model's ability to accurately anticipate future road congestion, providing a basis for reasonable analysis. This, in turn, offers valuable insights for transportation departments to monitor real-time traffic conditions and implement timely traffic control measures.

In the current landscape of road prediction, the trend towards comprehensive consideration of various factors influencing road congestion is evident. This paper exemplifies this trend by incorporating spatial and temporal dimensions, among other factors, to improve the rationality of road congestion prediction. However, it is acknowledged that real-world complexities, such as different car models, speeds, environments, and road grades, may impact road congestion. Future research endeavors can gradually integrate these factors to further enhance the accuracy of prediction results.

## References

*[1] Lin Lichun, Liu Hua, Hong Dong. Traffic congestion prediction technology based on big data analysis [J]. Western Transportation Science and Technology, 2020, 000(009):138-141.*

*[2] Tang Zhikang, Wang Weizhi, Tan Weixin. Research on traffic congestion prediction based on Bagging [J]. Journal of Jimei University: Natural Science Edition, 2006, 11(2):5.*

*[3] Gu Liqiong, Song Zukang, Yang Yang. Traffic congestion prediction based on improved support vector machine model [J]. Software Guide, 2019, 18(12):5.*

*[4] Zhang Wei, Hu Fangrui, Qi Wei, et al. Evaluation of geologic hazard susceptibility based on XGBoost and cloud model [J]. Chinese Journal of Geological Hazards and Prevention, 2023, 34(6):1-10.*

*[5] Pahari P K, Vyas S, Aman S, et al.Therapeutic Options for the Treatment of 2019-Novel Coronavirus in India: A Review[J].Coronaviruses. 2022(2):3.*