symmetry

MDPI

Article

Prediction of High-Speed Traffic Flow around City Based on BO-XGBoost Model

Xin Lu [1], Cai Chen [2,*], RuiDan Gao [3] and ZhenZhen Xing [4]

[1] Experimental Testing Institute, Xi'an Highway Research Institute Co., Ltd., School of Materials Science and Engineering, Chang'an University, Xi'an 710064, China; 2020031007@chd.edu.cn
[2] College of Highways, Chang'an University, Xi'an 710064, China
[3] School of Transportation Engineering, Changsha University of Science and Technology, Changsha 410205, China; 21201060175@stu.csust.edu.cn
[4] School of Information Engineering, Chang'an University, Xi'an 710064, China; 2022224013@chd.edu.cn
* Correspondence: 2020621003@chd.edu.cn

Abstract: The prediction of high-speed traffic flow around the city is affected by multiple factors, which have certain particularity and difficulty. This study devised an asymmetric Bayesian optimization extreme gradient boosting (BO-XGBoost) model based on Bayesian optimization for the spatiotemporal and multigranularity prediction of high-speed traffic flow around a city. First, a traffic flow dataset for a ring expressway was constructed, and the data features were processed based on the original data. The data were then visualized, and their spatiotemporal distribution exhibited characteristics such as randomness, continuity, periodicity, and rising fluctuations. Secondly, a feature matrix was constructed monthly for the dataset, and the BO-XGBoost model was used for traffic flow prediction. The proposed model BO-XGBoost was compared with the symmetric model bidirectional long short-term memory and integrated models (random forest, extreme gradient boosting, and categorical boosting) that directly input temporal data. The R-squared (R^2) of the BO XGBoost model for predicting TF and PCU reached 0.90 and 0.87, respectively, with an average absolute percentage error of 2.88% and 3.12%, respectively. Thus, the proposed model achieved an accurate prediction of high-speed traffic flow around the province, providing a theoretical basis and data support for the development of central-city planning.

Keywords: traffic flow prediction; spatiotemporal characteristics; XGBoost algorithm; bayesian optimization; Bi-LSTM

Citation: Lu, X.; Chen, C.; Gao, R.; Xing, Z. Prediction of High-Speed Traffic Flow around City Based on BO-XGBoost Model. *Symmetry* **2023**, *15*, 1453. https://doi.org/10.3390/sym15071453

Academic Editor: Alice Miller

Received: 17 May 2023
Revised: 8 July 2023
Accepted: 18 July 2023
Published: 20 July 2023

1. Introduction

Traffic flow is an important indicator of traffic congestion in transportation, particularly in densely populated central cities. Predicting traffic flow around urban highways is essential for economic development planning and cargo transportation in a city. Accurate prediction of bypass highway traffic flow is crucial for urban economic development planning, traffic resource utilization, and cargo transportation development; therefore, bypass highway traffic flow prediction in central cities has become a topic that must be studied [1–3]. Traffic flow data are influenced by several factors and exhibit strong randomness and uncertainty [4].

Compared with other traffic flow prediction studies, such as predicting the long- and short-term traffic flow at a certain intersection or predicting the traffic flow on general highways [5,6], the difficulty in analyzing traffic flow around urban highways lies in the following points:

(1) The distribution of entry and exit stations on up-and-down highways is dense, and the traffic flow varies significantly within short spatial distances.

(2) Owing to the uneven economic development of central cities and different regional functional positionings, significant differences exist in the composition of traffic components and traffic flow between different stations.
(3) Traffic flow around urban highways is easily influenced by various policies and important urban activities in the central city, and the traffic flow at the same station varies considerably at different times.

Therefore, several prediction models have been proposed based on the improvement of evaluation models. To improve the prediction accuracy of traffic flow, this study entailed the development of an asymmetric Bayesian optimization extreme gradient boosting (BO-XGBoost) model for the spatiotemporal and multigranularity prediction of high-speed traffic flow around a key city. The overall framework is shown in Figure 1, highlighting the following contributions:

(1) First, a dataset of high-speed traffic flow around a city was constructed, and the data were dimensionalized to expand the data features. By combining the inherent attributes of randomness, continuity, periodicity, and volatility of high-speed traffic flow around the city, the data were visualized and their spatiotemporal distribution characteristics were studied. Different spatial (stations and high-speed entire lines) and temporal granularities (monthly and annually) were analyzed using annual periodicity and the intra-year variability of time characteristics of the dataset.
(2) Second, different feature matrices were constructed for traffic flow data and then input into symmetric bidirectional long short-term memory (Bi-LSTM) and integrated extreme gradient boosting (XGBoost) models. The prediction accuracy of inputting time series data into the ensemble learning model after a monthly feature-matrix analysis was higher than that of directly inputting time series data into the symmetric Bi-LSTM model. To further improve the prediction accuracy of the integrated XGBoost model, a Bayesian algorithm was used to optimize the model and obtain optimal network parameters.

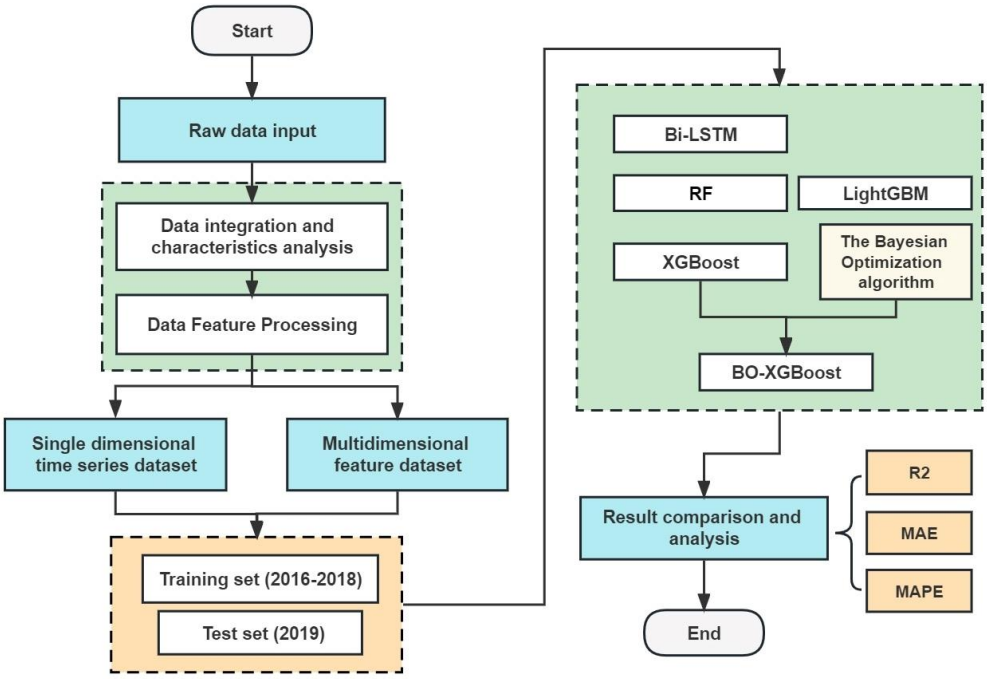


Figure 1. Research framework.

Prediction analysis was conducted using the proposed model for several traffic dispatch stations in Xi'an, a central city in Shaanxi Province in central China, and finally compared with that by the symmetric Bi-LSTM and ensemble learning models (random forest, XGBoost, and categorical boosting (Catboost)). The proposed model can provide a

theoretical basis and data support for the prediction and development planning of high-speed traffic flows around a central city.

The remainder of this paper is organized as follows. Section 2 presents the literature review. Section 3 introduces the datasets used in this study. Section 4 introduces the methods employed in this study, including the XGBoost algorithm, Bayesian optimization, and the BO-XGBoost models. Section 5 analyzes the results, and Section 6 presents the conclusions and future work.

2. Literature Review

Several studies have contributed to the field of traffic flow prediction [7–9]. Traditional machine learning methods have been used [10,11]. Polson et al. developed an architecture that uses a combination of linear models fitted with L1 regularization and tanh layer sequences to efficiently handle special events, Chicago Bear soccer games, and traffic forecasting during snowstorms. The results indicated that the proposed deep learning approach can capture these nonlinear spatiotemporal effects well [12]. Alajali et al. accurately predicted intersection traffic by introducing additional data sources to road-traffic volume data into the prediction model and investigating two types of learning schemes: batch learning and online learning. They performed their evaluations using publicly available datasets published by the Victorian government in Australia. They indicated that their proposed method can improve the accuracy of intersection-traffic prediction by combining information on accidents and roads near intersections [13]. These analyses can reflect changes through quantitative indicators, which are more convincing than subjective experience.

With an increasing focus on machine learning technology in the field of traffic flow prediction, numerous new methods and techniques have been applied to prediction [14–16]. Based on the established results, further explorations have been conducted based on practical situations. Using neural network models trained on a significant amount of historical data, data fluctuations can be accurately learned and data features can be better extracted, vastly improving prediction accuracy [17]. Wang et al. proposed a short-term traffic flow prediction method based on a genetic algorithm (GA)-optimized back-propagation neural network (BPNN); they processed the weights and thresholds of the BP neural network using a GA and applied the BP neural network to optimally adjust the real-time data prediction values [18]. Zhang et al. proposed a hybrid prediction framework based on support vector regression (SVR), which uses random forest (RF) to select the feature subset with the largest amount of information and enhanced GA with chaotic characteristics to identify the optimal prediction-model parameters [19]. Wang et al. used SVR as the main regression model for traffic flow prediction and Bayesian optimization for parameter selection. They proposed a short-term traffic flow prediction regression framework with automatic parameter adjustment. Their experimental results indicated that the accuracy of this method is superior to classic seasonal autoregressive integrated moving average (SARIMA), multi-layer perceptron neural network (MLP NN), and Adaboost methods [20]. Lu et al. formulated the road network as a dynamically weighted graph through the attention mechanism, searching for spatial and semantic neighbors to establish more connections between road nodes, thus proposing a new spatiotemporal adaptive gated graph convolutional network (STAG-GCN) to predict future traffic conditions at several time-steps [21]. However, the graph convolutional network is highly suitable for predicting urban traffic flow and traffic flow with origin–destination (OD) data.

Time series forecasting models have been used for traffic flow predictions using historical data with time series characteristics [22]. Fang et al. relaxed the assumption of prediction error to an arbitrary distribution using a negative bootstrap mixed correlation entropy criterion and constructed a long short-term memory (LSTM), equipped with a robust loss function called a free-LSTM network, for short-term traffic flow prediction [23]. Lan Tianhe et al. proposed a prediction model (GWO Attention LSTM) based on the combination of the optimized attention mechanism and LSTM. The results indicate that the GWO attention LSTM model has good model performance and can provide effective

assistance for traffic management control and traffic flow theory research [24]. Symmetry-based Bi-LSTM networks can overcome the drawback of one-way LSTM networks of being only able to learn unidirectional information [25,26]. Zhuang et al. proposed a multistep prediction model based on a convolutional neural network and the Bi-LSTM model. Experimental results indicated that the Bi-LSTM model improved the prediction accuracy and reduced the mean absolute error, mean absolute percentage error, and root mean square error by 30.4%, 32.2%, and 39.6%, respectively, compared with the SVR and gated recurrent unit models [27].

In the field of integrated learning, recently, XGBoost has performed well in many research areas because of its good performance and generalization capabilities [28]. Du et al. proposed a feature-enhanced XGBoost based on LSTM (XGBoost-LSTM) for base-station traffic prediction. By observing the predicted values, they observed that a simple combination of XGBoost and LSTM could achieve significant improvements [29]. Sun et al. proposed and improved their prediction of high-resolution traffic states using the OD relationship of flow data from upstream and downstream sections of highways. Two improved models were proposed, namely independent XGBoost (XGBoost-I) with different interval independent adjustment parameters and static XGBoost-S with overall parameter adjustment [30].

Several research methods have been adopted to solve traffic flow prediction problems. Different models have their own merits and limitations, while the spatial and temporal distribution characteristics of traffic flow and the change patterns of different regions and different levels of roads also vary. The aforementioned studies on traffic flow prediction are summarized in Table 1. These methods provide good solutions for traffic flow prediction [31,32].

Table 1. Comparison of traffic flow prediction methods.

Literature	Method	Overall Evaluation of the Method
[9]	ST-3DGMR	Compared with the state-of-the-art, ST-3DGMR has significantly lower RMSE on the BikeNYC, TaxiBJ, and TaxiCQ datasets.
[18]	GA-BPNN	Experimental results indicated that the GA-BPNN algorithm had a better prediction effect and provided a certain reference value for short-time traffic-flow prediction.
[22]	RFCGASVR	The results indicated that the proposed RFCGASVR model performed better than other methods.
[23]	SVR	The experimental results indicated that the accuracy of this method is superior to classic SARIMA, MLP NN, ERT, and Adaboost methods.
[24]	GWO-attention-LSTM	It has better performance and can provide effective help for traffic management control and traffic flow theory research.
[26]	Bi-LSTM- Attention	Bidirectional long short-term memory (Bi-LSTM)—Attention models can effectively improve the prediction accuracy of traffic flow.
[30]	XGBoost-LSTM	The XGBoost-LSTM model was used to predict base-station traffic and performed better than competing algorithms.
[31]	XGBoost-I XGBoost-S	Two improved models have been proposed: independent XGBoost (XGBoost-I) with different interval independent adjustment parameters and static XGBoost-S with overall parameter adjustment.

3. Materials

3.1. Data Characteristic Analysis

This study focused on the prediction of Xi'an bypass traffic flow data obtained from the Xi'an Traffic Management Department. Twelve major sites were selected for the study: FangJiaCun, FangJiaCunLiJiao, HeChiZhaiLiJiaoDong, HeChiZhaiLiJiaoXi, LiuCunBaoLiJiao, LvXiaoZhaiLiJiaoXi, MaoErLiuLiJiaoNan, QuJiangLiJiaoDong, QuJiangLiJiaoXi, XiGaoXin, XiangWangLiJiao, and XieWangLiJiao. In the follow-up study, we used A, B, C, D, E, F, G, H, I, J, K, and L to denote the individual sites. Owing to the impact of the global pandemic in 2020, most urban residents in China were quarantined at home, and traffic flow in major central cities declined rapidly. Therefore, the traffic data from January 2020 to December 2022 are different from the spatiotemporal-characteristic distribution of the

data under normal conditions and do not have the cyclical and representative nature of the past. Therefore, we used data from January 2016 to December 2019, and each site had approximately 4 × 12 × 15 = 720 (year × month × indicators) data counts; therefore, the data contained 720 × 12 = 8640 samples.

The dataset contains multiple feature data for 12 months per year. The primary classification node for the original intermodulation data was time (year, month), and the secondary classification node was the location (different sites).

The feature abbreviations and their meanings in the original data are listed in Table 2. The raw sample data are presented in Table 3.

Table 2. Explanations of abbreviations.

Abbreviation	Full Name	Type of Information	Abbreviation	Full Name	Type of Information
ObName	Observatory Name	position information	PCTF	Passenger Car Traffic Flow	real statistical indicators
NO.	Start Stake Number	position information	TTF	Truck Traffic Flow	real statistical indicators
SMPTF	Small and Medium Passenger Traffic Flow	real statistical indicators	TF	Traffic Flow	real statistical indicators
LBTF	Large Bus Traffic Flow	real statistical indicators	PU	Passenger Unit	calculated indicators
STTF	Small Truck Traffic Flow	real statistical indicators	TU	Truck Unit	indicators calculated
MTTF	Medium Truck Traffic Flow	real statistical indicators	PCU	Passenger Car Unit	indicators calculated
LTTF	Large Truck Traffic Flow	real statistical indicators	Speed	Motor Vehicle Speed	real statistical indicators
ELCTF	Extra Large Cargo Traffic Flow	real statistical indicators	TVR	Traffic Volume Ratio	indicators calculated
CTF	Container Traffic Flow	real statistical indicators	/	/	/

Table 3. January 2016 original dataset (sample).

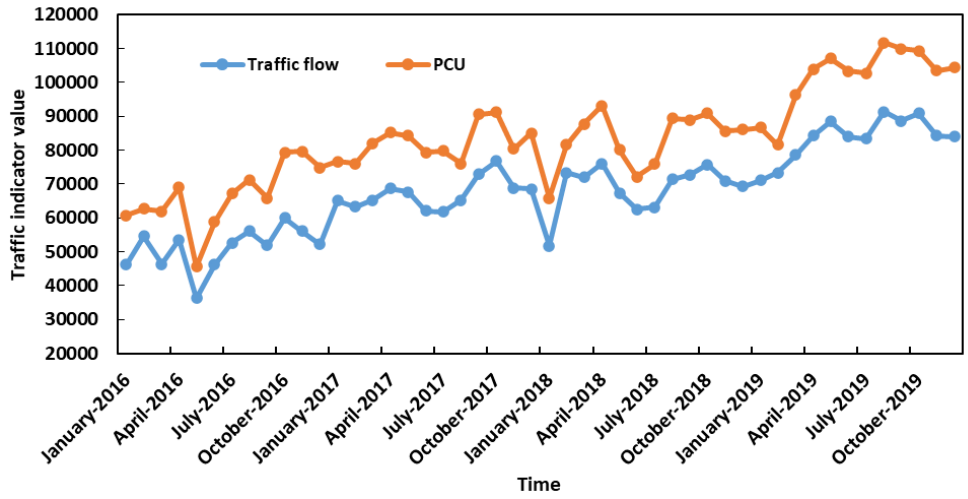
Year	Month	Ob Name	NO.	SMPTF	LBTF	STTF	MTTF	LTTF	ELCTF	CTF	PCTF	TTF	TF	PU	TU	PCU *	Speed	TVR *
2016	01	FangJiaCun	0	10,655	150	401	976	571	2125	130	10,805	4203	15,008	10,880	12,598	23,478	79.9	0.293
2016	01	FangJiaCunLiJiao	77.5	16,448	352	1061	786	911	2901	567	16,800	6226	23,026	16,976	18,845	35,821	83.7	0.448
2016	01	HeChi ZhaiLiJiao	46.8	38,904	4871	4320	3242	876	335	3043	43,775	11,816	55,591	46,211	25,323	71,534	70.8	0.894
2016	01	HeChi ZhaiLiJiaoXi	39.6	30,513	2375	3389	1581	459	342	1376	32,888	7147	40,035	34,076	14,010	48,085	64.9	0.601
2016	01	LiuCunBaoLiJiao	18.6	22,160	641	2185	946	1254	2252	427	22,801	7064	29,865	23,122	18,082	41,204	72.8	0.515
2016	01	LvXiao ZhaiLiJiaoXi	15.6	24,602	3351	4101	2888	6058	903	2083	27,953	16,033	43,986	29,629	38,551	68,180	71.7	0.852
2016	01	MaoEr LiuLiJiaoNan Qu	34	26,393	310	2030	975	778	1637	203	26,703	5623	32,326	26,858	13,187	40,045	77.8	0.501
2016	01	Jiang LiJiaoDong Qu	57.5	42,650	554	3261	1539	970	2156	515	43,204	8441	51,645	43,481	19,164	62,645	66.8	0.783
2016	01	Jiang LIJiaoXi	63.1	37,747	423	1831	1274	1211	3003	751	38,170	8070	46,240	38,382	22,391	60,773	81	0.76
2016	01	XiGao Xin	52.7	24,659	430	1748	680	457	1155	191	25,089	4231	29,320	25,304	9523	34,827	61.7	0.435
2016	01	Xiang WangLiJiao	73	33,517	1325	3721	5288	812	412	3728	34,842	13,961	48,803	35,505	30,649	66,154	74.4	0.827
2016	01	XieWangLiJiao	4.96	19,106	3326	4775	5410	1029	7871	3064	22,432	22,149	44,581	24,095	59,717	83,812	73.1	1.048

* Passenger Car Unit (PCU): standard vehicle equivalent, also known as equivalent traffic, is the equivalent traffic volume of a standard model that converts the actual motor and non-motor vehicle traffic volume into a certain standard model based on a certain conversion. TVR: ratio of the total traffic volume of the road network to the total allowable capacity of the road network.

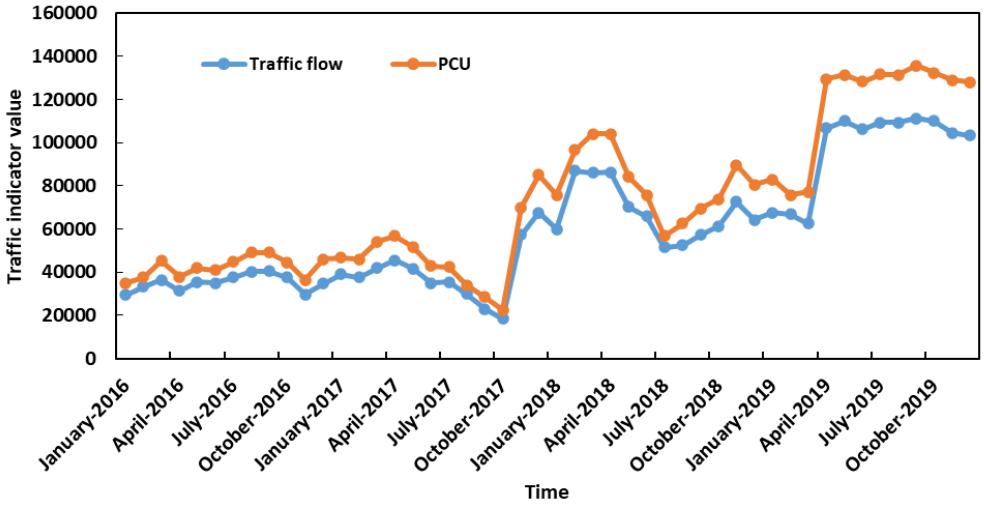
Table 3 lists the passenger unit, truck unit, passenger car unit, and congestion as indicators calculated using a formula based on real indicators.

To conduct time series-related research, datasets from different years and months should be integrated to form a training dataset that can be input into the model. Considering space (different stations) as the primary key identification, the spatial position was first

aligned, and then the spatial-location data from January 2016 to December 2019 were arranged in time series order to obtain the traffic flow time series data of the 12 stations of the central city ring expressway. Using stations I and J as examples, we visualized the changes in their traffic flow and passenger car unit (PCU), as shown in Figure 2. Stations I and J had significant differences in traffic volume around the city expressway. From Figure 2, it can be seen that the data waveforms of different stations are different; however, recently, they have exhibited a fluctuating upward trend, which is noticeable. For station I, although the monthly granularity data have a certain randomness and volatility, they have a certain periodicity and continuity. For example, troughs occur annually between May and July. However, owing to the influence of policies, economic development, and road maintenance, a significant gap occurred between the traffic data change curves of stations J and I, which primarily indicate a stepwise upward distribution.



(**a**) Station I

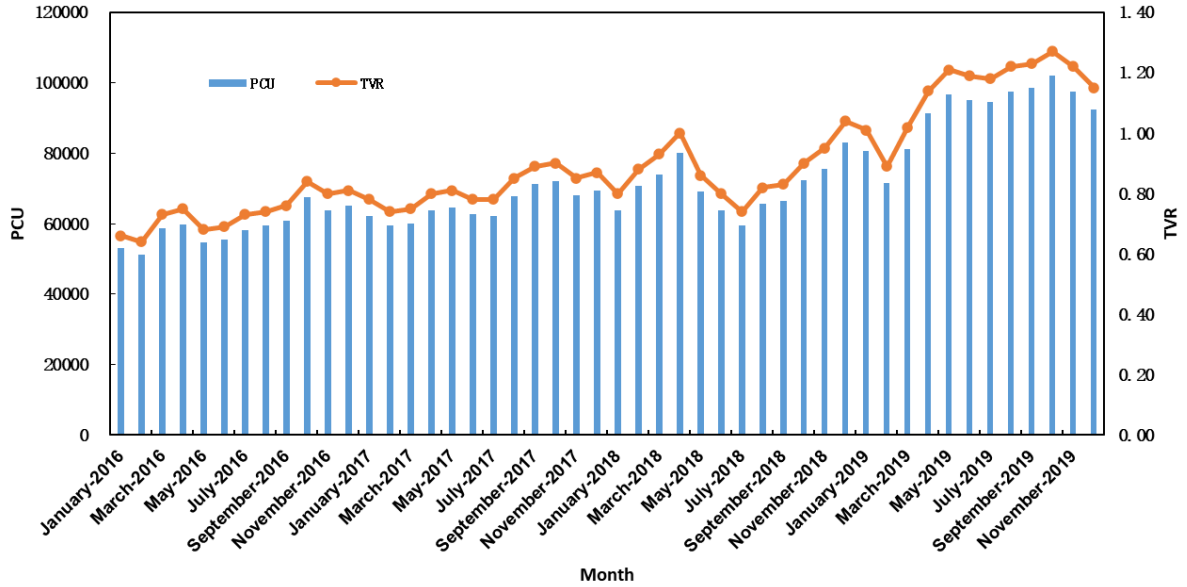


(**b**) Station J

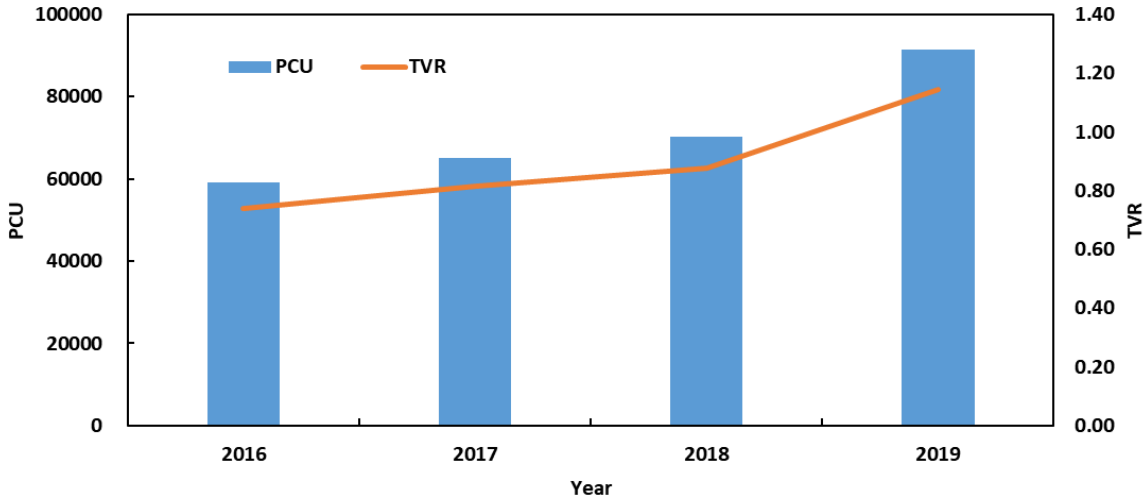
Figure 2. Trend of traffic flow and PCU monthly granularity data from 2016 to 2019.

To analyze the annual changes in the expressway traffic around the city, the data from each station were averaged, and the traffic equivalent PCU and ratio of the total traffic volume of the road network to the total allowable capacity of the road network (TVR) time series data were visualized by monthly and annual granularities, respectively, as shown in Figure 3. Figure 3a demonstrates that the traffic flow changes around the urban expressway have the characteristics of continuity, randomness, periodicity, and a fluctuating rise. From

the annual granularity data in Figure 3b, the PCU can be observed to steadily increase from 2016 to 2018, which can be characterized by a linear model. The PCU rate significantly increased in 2019.



(**a**) Monthly granularity



(**b**) Annual granularity

Figure 3. Visualization of PCU and TVR with different time granularities.

3.2. Data Feature Processing

3.2.1. Removing Invalid Features

Each determined time and location had 15 attributes. The attributes that were irrelevant to this data prediction, such as "No.", "station", and "speed", were deleted during data pre-processing.

3.2.2. Removing Strongly Correlated Features

Some comprehensive attributes were calculated from the basic attributes, and a strong correlation occurred between the features. The symmetrical correlation matrix between the various features is shown in Figure 4.

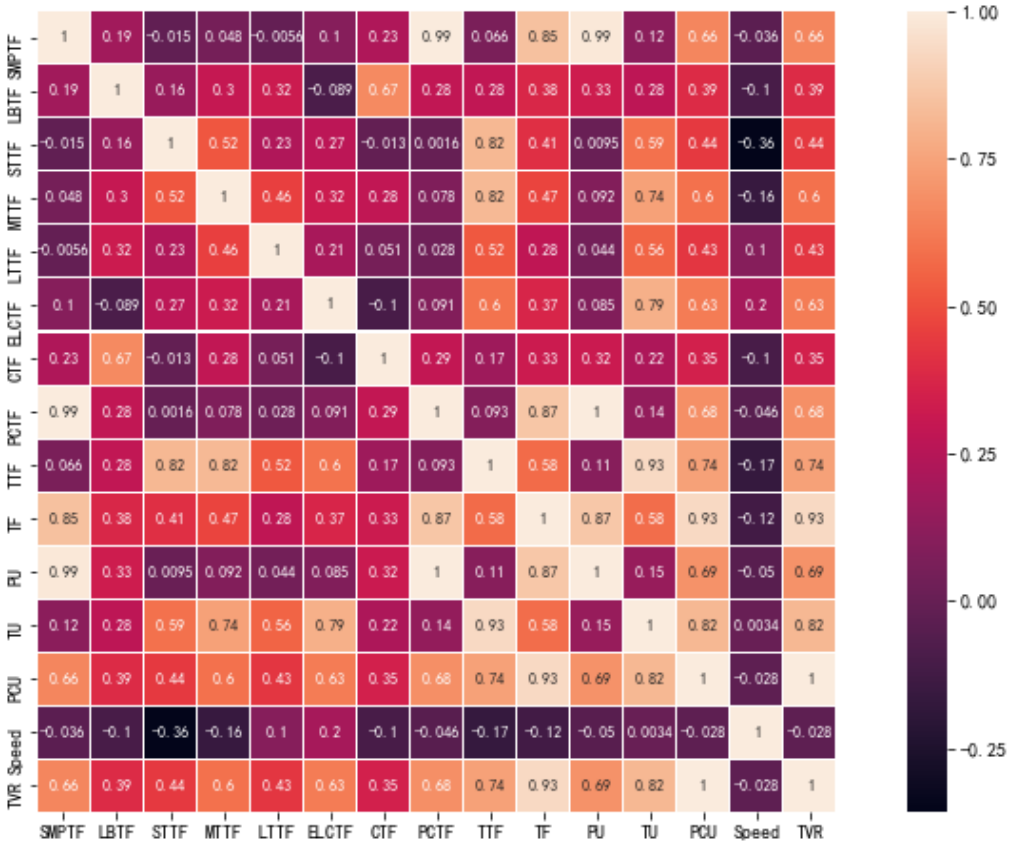


Figure 4. Correlation matrix between features.

As shown in Figure 4, the symmetrical correlation between many features was greater than 0.6. For example, the correlation between the PCU and TVP was 1, while the correlation between the MTTF and TU was 0.74, the correlation between the TTF and TU was 0.79, the correlation between the TF and TU was 0.93, and the correlation between the TVR and TU was 0.82. If attributes with high correlation coefficients are input into the model, its accuracy is significantly affected. Therefore, these basic attributes were removed. The removed features were MTTF, TTF, TF, TVR, and TU.

3.2.3. Fusing Features

Considering that the study objectives are in time series, the features "year" and "month" in the original table were combined and modified to "time". Sequential arrangement was used to characterize the traffic flow detection time; however, it was not used as the feature input for the subsequent model and did not participate in the network training.

4. Methodology

4.1. XGBoost

XGBoost trees are similar to random forests in that they are based on multiple decision trees; however, unlike random forests, multiple decision trees are assembled using a boosting method [33]. The basic idea of modeling pavement service performance evolution prediction based on XGBoost is to improve the overall prediction accuracy by constructing a new function to learn the residuals of the previous round, that is, by generating a new decision tree to compensate for the error of the previous tree until the error of the objective function is less than the predetermined error. Combined with a gradient boosting algorithm to improve the objective function, parallel processing and regularization are supported.

The objective function of each round is defined as the loss function + regularization term, as in Equation (1).

$$L(y, f(x)) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \tag{1}$$

where y_i represents the true output of the sample x_i, $\hat{y}_i$ the predicted output of the sample x_i, f_k the k-th boosted decision tree, and $l(y_i, \hat{y}_i)$ the loss function representing the difference between the predicted and true outputs. $\Omega(f_k)$ is the canonical term for the function. The objective function is analyzed to obtain the base learner with the best performance. The most critical aspect is the selection of the split node function, which is used to determine the split node of the decision tree by calculating the loss reduction after splitting. Controlling the complexity of the tree when splitting and setting a certain threshold value to select splitting only when the score is greater than this threshold value, which plays a role in pre-pruning, is necessary.

4.2. Bayesian Optimization Algorithm

The Bayesian optimization (BO) algorithm consists of two main steps: (i) employing a probabilistic model to replace the evaluation scores of the hyperparameters in the original objective function and (ii) constructing the payoff function using the posterior probability information transformed from the prior probabilities of the probabilistic agent model. The choice of the probabilistic model must be generally determined on a case-by-case basis [34].

Because the probability distribution of the hyperparameters in the objective function of the algorithm used in this study was unknown, the commonly used Gaussian agent model was chosen for parameter optimization. Assuming that the objective function $f(x)$ optimized by hyperparameter combination x obeys the Gaussian stochastic process, the posterior distribution $p(f(x)|x)$ of $f(x)$ is calculated to estimate $f(x)$ as follows:

$$p(f(x)|x) = p(f(x))p(x|f(x))/p(x), \tag{2}$$

where $p(f(x))$ is the prior probability distribution of $f(x)$; $p(x|f(x))$ the likelihood distribution of hyperparameter combination x on the objective function $f(x)$; and $p(x)$ the normalization factor. Therefore, when the optimized objective function $f(x)$ obeys the Gaussian stochastic process, the posterior probability distribution is obtained using Bayes' theorem, that is, the confidence level of the objective function after adjusting the correction for the prior probability.

The cost of calculating the objective function corresponding to the hyperparameter combination in hyperparameter optimization is high. To make the corrected $p(f(x)|x)$ rapidly approach the true distribution of the objective function, the payoff function is constructed through the posterior probability distribution, and the payoff function is calculated and maximized to obtain the next hyperparameter combination that can better improve the objective function [35].

The gain function searches for the global optimal solution that maximizes the gain of the gain function $\alpha(x)$ based on the current search point in the new parameter space and near the local optimal solution. The set of optimal solutions x^* in the hyperparameter full set A is given by the following equation:

$$x^* = \text{argmax}(\alpha(x)), \quad x \in A. \tag{3}$$

Commonly used gain functions include the probability of improvement (PI), expected improvement (EI), and Gaussian process upper confidence bounds (GP-UCB). Among them, the GP-UCB has a gain function in Set A as

$$\alpha(x, \theta, A) = \mu(x, \theta, A) + \kappa\sigma(x, \theta, A), \tag{4}$$

where $\mu(x)$ and $\sigma(x)$ are the mean and variance, respectively, of the objective function obeying the Gaussian process, κ the upper confidence bound, and θ the hyperparameter.

4.3. BO-XGBoost Model

The combination of the Bayesian and XGBoost networks was used to search for the optimal parameter combination of the XGBoost model using BO. BO can calculate the gain function based on the mean and variance of the objective function corresponding to the current search point, add a new hyperparameter search point with the maximum gain function to the evaluation point set, and update the probabilistic surrogate model through a new set until the number of iterations to obtain the optimal combination of the hyperparameter of XGBoost is reached. The proposed hyperparameter optimization method was optimized throughout the entire model training process, rather than during the model testing process. Training data were input into the network and the training data were trained using the set network parameters. The loss function value of the validation data was used for feedback, and the model parameters were constantly adjusted. Finally, the model parameters with the highest accuracy in the training set were obtained, and then the model parameters were output. The test set data were input into the trained model for testing, and the accuracy of the test set was obtained. The specific process is illustrated in Figure 5.

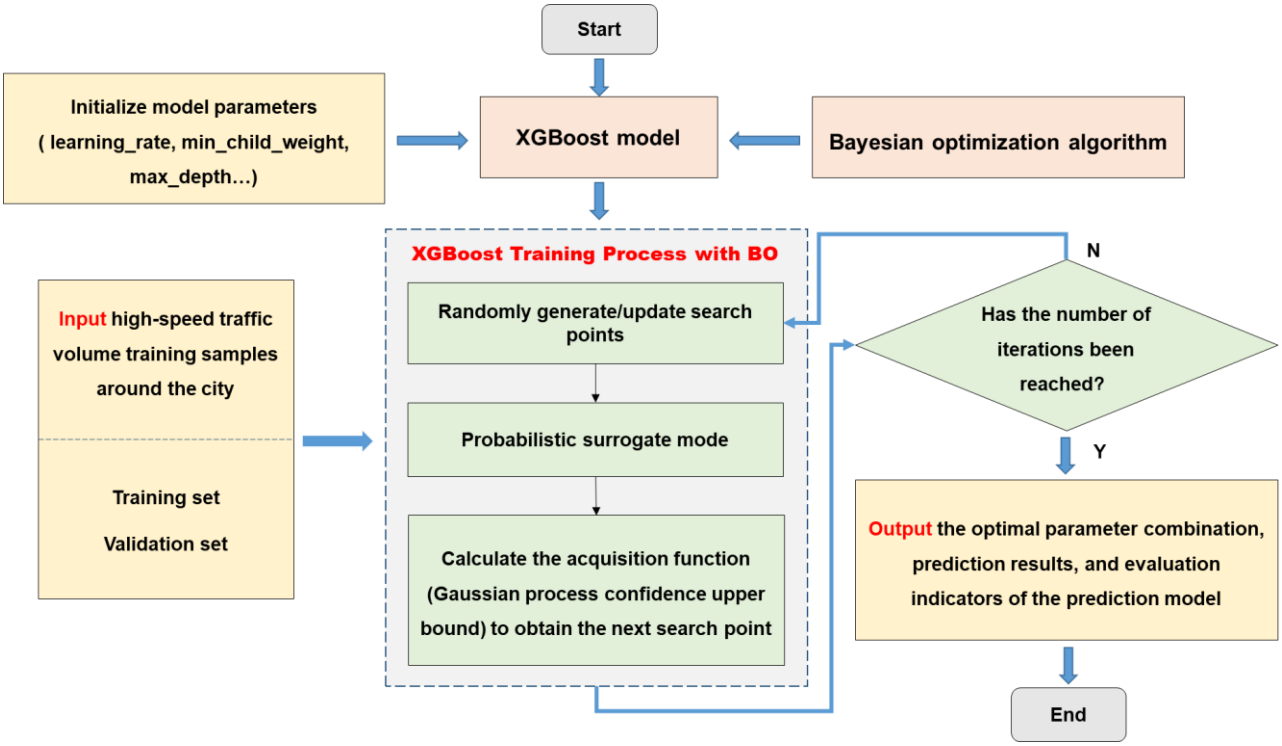


Figure 5. BO-XGBoost model process.

The XGBoost model must set more parameters, and each parameter has a certain connection with the others, which is not suitable for manual repeated tuning of parameters to train the comparison. With more continuous parameters, such as gamma, subsample, colsample_bytree, reg_alpha, and reg_lambda, its optimal solution may be decimal with an accuracy of 0.01, which is difficult for the grid-search algorithm by enumeration because continuous values can be infinitely subdivided; therefore, a BO algorithm can be used to iteratively find the optimal solution for the hyperparameters of the XGBoost model.

After optimizing the main parameters of the XGBoost-based pavement service performance prediction model using Bayesian methods, the optimal parameter values of the model were obtained, as listed in Table 4.

Table 4. Bo-XGBoost model parameters table.

Model Parameters	Parameter Explanation	Optimal Values
max_depth	Limiting the maximum depth of a tree	4
gamma	Penalty term coefficient, the minimum loss function reduction required to split the node	0.3
min_child_weight	Minimum leaf node sample weights sum	1
subsample	The rate at which samples are sampled when the decision tree is built	0.8
colsample_bytree	The rate at which features are sampled when building a decision tree	0.8
reg_alpha	L1 regularization parameters	0.2
reg_lambda	L2 regularization parameters	0.7
learning_rate	The learning rate, which is the model learning step	0.035
n_estimators	The number of iterations raised	300

5. Traffic Data Prediction Based on the BO-XGBoost

5.1. Experimental Setup

5.1.1. Experimental Datasets

In the experiment, we used the dataset after it had undergone the processing stage described in Section 3. Before inputting the BO-XGBoost model, performing feature reconstruction on the monthly data was necessary. Subsequently, using historical traffic data from 2016 to 2018 as features, the TF and PCU were predicted for 2019. Eighty per cent of the 2019 data were used as the training set input model, and the remaining 20% were used as test data. After the model training was completed, the model was input for testing.

Because the comparative experiments were conducted using different models that have different requirements, the input and output variables must be set to data that satisfy the requirements of the model. The models used to solve traffic flow prediction problems generally include those that handle time series and those that handle multiple regression problems.

Therefore, two datasets were established for this experiment.

Dataset 1: The input was a time series variable composed of time-steps and independent variables (TF and PCU), and the output was a target prediction value composed of a one-dimensional array.

Dataset 2: The input was a multivariate feature set composed of multiple features (TF, PCU, and other selected feathers for the same month over the years), and the output was a target prediction value composed of a one-dimensional array.

Dataset 1 was used as an input for the symmetric Bi-LSTM model. Dataset 2 was used as an input for the proposed asymmetric model and other comparative models.

5.1.2. Experimental Environment

The environment configurations used for the experiments are presented in Table 5.

Table 5. Experimental environment.

Items	Configuration
Software	Anaconda, Jupyter Notebook
Hardware	Win11, 12th Gen Intel(R) Core (TM) i5-12500 3.00 GHz, 16 GB of memory
Language and Frames	Python and TensorFlow

5.1.3. Evaluating Indicator

Based on the measured and predicted traffic values, the prediction results of the model were evaluated based on commonly used prediction metrics such as the mean absolute error (MAE), mean absolute percentage error (MAPE), and linear correlation coefficient

(R^2) [36]. The formulas, meanings, and evaluation criteria for each metric are shown in Equations (5)–(7), where n is the total number of measured values, y_i the traffic flow predicted measurement, $\widetilde{y_i}$ the predicted value of the temperature prediction, and $\overline{y_i}$ the average value of y_i.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left| \left(y_i - \widetilde{y_i} \right) \right| \tag{5}$$

$$\mathrm{APE} = \frac{100\%}{n}\sum_{i=1}^{n} \left| \frac{y_i - \widetilde{y_i}}{y_i} \right| \tag{6}$$

$$R^2 = 1 - \frac{\sum_i^n \left(y_i - \widetilde{y_i} \right)^2}{\sum_i^n \left(y_i - \overline{y_i} \right)^2} \tag{7}$$

5.2. Result Comparison and Analysis

To evaluate the effectiveness of the BO-XGBoost model, the prediction–evaluation indices of the other models were compared. The optimal parameters of each model were adjusted using a step-based grid-search method, and the optimal parameters of each comparative model were obtained as shown in Table 6. Hyperparametric optimization was performed to determine the optimal parameters suitable for the objectives of this research, rather than adjusting parameters based on other methods. The comparison of evaluation indicators for TF and PCU for different models on the training and testing sets is shown in Tables 7 and 8, respectively. Notably, the evaluation indicators here refer to the average indicator values of the 12 station experiments.

Table 6. Optimal parameters of each model.

Model	Parameter Settings
DNN	shuffle = True, epochs = 1000, batch_size = 16, verbose = 1
RF	n_estimators = 50, max_depth = 2, min_samples_split = 4, min_samples_leaf' = 1, max_features = 7, oob_score = False
XGBoost	max_depth = 4, gamma = 0.5, min_child_weight = 1, subsample = 0.6, colsample_bytree = 0.8, reg_alpha = 0.2, reg_lambda = 0.6, learning_rate = 0.05, n_estimators = 300
GA-XGBoost	Pc = 0.6, Pm = 0.01, T = 100, M = 50
Bi-LSTM	epochs = 100, batch_size = 64, validation_split = 0.15, INPUT_DIMS = 1, TIME_STEPS = 12, lstm_units = 64
CatBoost	iterations = 400, learning_rate = 0.05, depth = 5, l2_leaf_reg = 1

Table 7. Analysis of evaluation indicators for various models of monthly granularity TF.

Model	Training Set			Test Set		
	R^2	MAE	MAPE%	R^2	MAE	MAPE%
DNN	0.99	13.38	0.0169	0. 13	5289	5.98
RF	0.87	2758.24	3.56	0.84	2967.42	3.83
CatBoost	0.99	1.55	0.0001	0.22	5228.98	4.25
Bi-LSTM	0.90	2254.21	2.91	0.88	2596.44	3.35
XGBoost	0.91	2196.56	2.83	0.88	2238.15	2.89
GA-XGBoost	0.91	2048.63	2.71	0.89	2213.57	2.87
BO-XGBoost	0.92	1963.26	2.53	0.90	2231.65	2.88

Table 8. Analysis of evaluation indicators for various models of monthly granularity PCU.

Contrast Model	Training Set			Test Set		
	R^2	MAE	MAPE%	R^2	MAE	MAPE%
DNN	0.99	5.56	0.0089	0.12	6352	6.81
RF	0.79	4448	4.02	0.78	4691.42	4.24
Catboost	0.99	0.248	0.0002	0.42	5353.74	4.84
Bi-LSTM	0.85	3652.73	3.22	0.81	4356.12	3.93
XGBoost	0.87	3885.32	3.51	0.86	3712.91	3.35
GA-XGBoost	0.88	3562.62	3.12	0.86	3521.66	3.41
BO-XGBoost	0.89	3125.51	2.82	0.87	3447.14	3.12

Tables 7 and 8 demonstrate that the prediction accuracy of the symmetric Bi-LSTM is higher than that of RF, indicating that the performance of this symmetric time series model is superior to the classical machine learning RF model in predicting TF and PCU.

Catboost and DNN models have an overfitting phenomenon when training TF and PCU data; that is, the fitting accuracy is significantly high on the training set and considerably low on the test set (below 0.5). Especially when the DNN model predicts TF and PCU, the accuracy of the test set is less than 0.15 on both types of targets. Probably owing to the slightly complex construction of neural network models, complex models have a strong fitting ability to data and a strong learning ability for noise, making it easier to cause overfitting. For the Catboost model, it may be because it is more suitable for the training set, whose input features are categorical. No categorical features were present in the dataset in this study, and the amount of data was not large, which caused the overfitting of the model.

In the prediction of TF, the difference in accuracy between XGBoost and symmetric LSTM models was insignificant, and XGBoost prediction results indicated a decrease in error. In terms of accuracy and error on the PCU test set, XGBoost outperformed the Bi LSTM model in all aspects. Compared with XGBoost, the proposed BO-XGBoost improved the prediction accuracy of TF and PCU while the error decreased. Especially on the training set of PCU, the accuracy was significantly improved and the error was significantly reduced. In terms of test sets, the accuracy was slightly improved compared with XGBoost. This may be because of the limited searchable range of parameter combinations in the fixed-step grid-search method. Compared with the prediction accuracy of GA-XGBoost, BO-XGBoost had a higher R2 and slightly lower overall error.

To compare the accuracy of the models, we plotted the average R2 and MAPE of each model (excluding overfitting models) for TF and PCU testing sets, as shown in Figures 6 and 7.

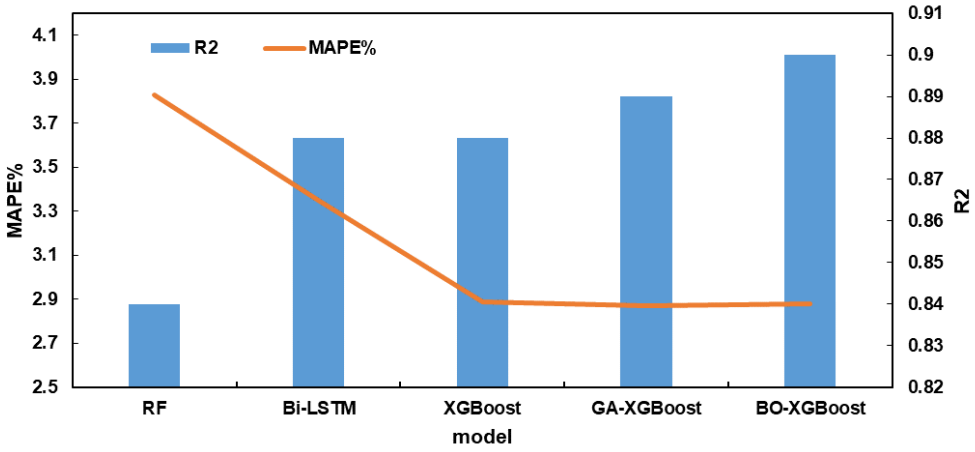


Figure 6. Comparison of R^2 and MAPE% metrics of TF by different models.

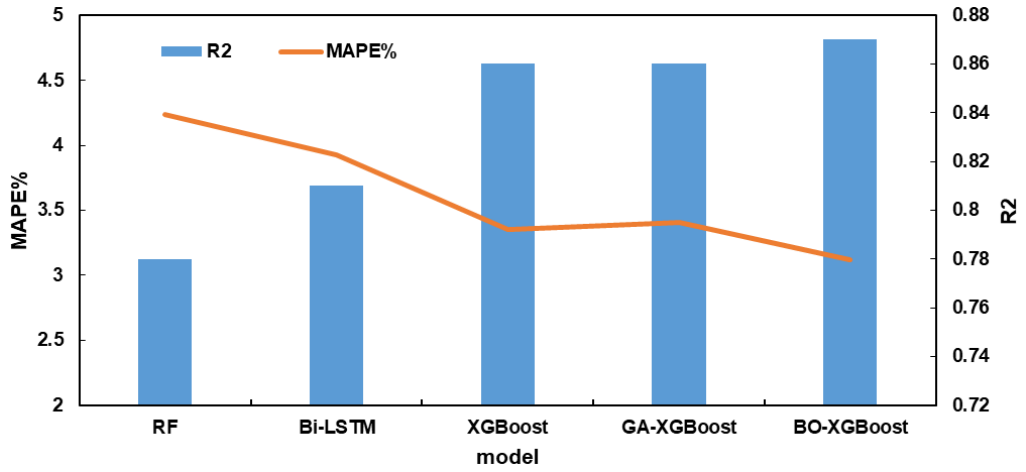


Figure 7. Comparison of R^2 and MAPE% metrics of PCU by different models.

5.3. Feature Importance Analysis Based on XGBoost

In a single decision tree, the importance of each feature is quantitatively evaluated by calculating and comparing the performance improvement of each feature splitting point. Then, the results of a feature appearing in all boost trees are calculated as the weighted average, which is the importance score corresponding to each feature. Finally, the importance of the features are ranked based on this score. Using XGBoost's weight calculation method, the number of times feature variables are used as partition variables in all trees is accumulated to obtain a feature importance score. The importance ranking of X influencing factors of input BO XGBoost using the XGBoost model is shown in Figure 8.

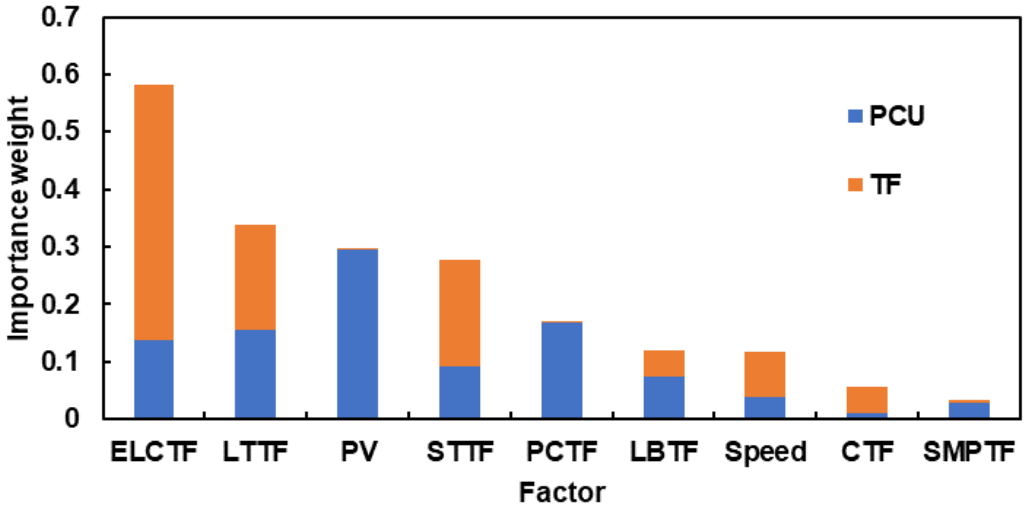


Figure 8. Importance ranking of TF and PCU influencing factors based on XGBoost.

From Figure 8, it can be observed that the two research objectives have some similar influencing factors. However, some influencing factors with very different weights exist in the two research objectives, such as PV, which has the greatest impact weight on PCU and the smallest impact weight on TF. Overall, the ELCTF, LTTF, and STTF have the greatest impact on the two objectives. The PV and PCTF have the greatest impact on PCU. ELCTF, STTF, and LTTF have a significant impact on TF. Considering the arithmetic mean of the importance weights of the two research objectives as the weight of the importance influencing factor, the impact factors with importance in the analysis results are listed in Table 9.

Table 9. Importance weights of impact factors based on XGBoost.

Impact Factors	Importance Weight of PCU	Importance Weight of TF	Average
PV	0.2949	0.0041	0.1495
PCTF	0.1670	0.0050	0.086
LTTF	0.1551	0.1845	0.1698
ELCTF	0.1375	0.4450	0.29125
STTF	0.0922	0.1865	0.13935
LBTF	0.0746	0.0460	0.0603
Speed	0.0387	0.0788	0.05875
SMPTF	0.0284	0.0046	0.0165
CTF	0.0116	0.0459	0.02875

6. Conclusions

A BO-XGBoost network with a Bayesian optimization parameter adjustment process was proposed to analyze the high-speed traffic flow and traffic equivalents around a central city. Considering Xi'an's data as an example for analysis, the main conclusions are as follows:

(1) A dataset of high-speed traffic flow around a city was constructed, and the data were dimensionalized to expand the data features. By combining the inherent attributes of randomness, continuity, periodicity, and volatility of high-speed traffic flow data, the data were visualized, and their spatiotemporal distribution characteristics were studied. We observed that the dataset had apparent annual periodicity and intra-year variability.

(2) We constructed different feature matrices for the traffic flow data and input those into the symmetric Bi-LSTM and integrated XGBoost models. The prediction accuracy of inputting time series data into the ensemble learning model after the monthly feature matrix analysis was observed to be higher than that of directly inputting time series data into the symmetric Bi-LSTM model. To further improve the prediction accuracy of the integrated XGBoost model, a Bayesian algorithm was used to optimize the model and obtain optimal network parameters. Experiments have shown that the improved XGBoost model has better accuracy, with a prediction accuracy of 0.90 and 0.87 for TF and PCU, respectively.

The following future research prospects warrant further study:

1. Collecting more data on urban highways around central cities and analyzing the impact of urban traffic flow on the economic development of central cities;
2. A hyperparameter optimization method with better accuracy and efficiency to determine the optimal parameter combination within the searchable range.

Author Contributions: Conceptualization, C.C. and X.L.; methodology, C.C. and X.L.; validation, Z.X. and C.C.; writing—original draft preparation, C.C. and R.G.; writing—review and editing, X.L., R.G. and C.C.; project administration, X.L., R.G. and C.C.; funding acquisition, X.L. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Key R&D Projects in Shaanxi Province (2022JBGS3-08).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rajeh, T.M.; Li, T.; Li, C.; Javed, M.H.; Luo, Z.; Alhaek, F. Modeling multi-regional temporal correlation with gated recurrent unit and multiple linear regression for urban traffic flow prediction. *Knowl. Based Syst.* **2023**, *262*, 110237. [CrossRef]
2. Guo, S.; Lin, Y.; Li, S.; Chen, Z.; Wan, H. Deep Spatial–Temporal 3D Convolutional Neural Networks for Traffic Data Forecasting. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3913–3926. [CrossRef]

3. Zhan, A.; Du, F.; Chen, Z.; Yin, G.; Wang, M.; Zhang, Y. A traffic flow forecasting method based on the GA-SVR. *J. High Speed Netw.* **2022**, *28*, 97–106. [CrossRef]
4. Zhou, T.; Huang, B.; Li, R.; Liu, X.; Huang, Z. An attention-based deep learning model for citywide traffic flow forecasting. *Int. J. Digit. Earth* **2022**, *15*, 323–344. [CrossRef]
5. Kashyap, A.A.; Raviraj, S.; Devarakonda, A.; Nayak, K.S.R.; KV, S.; Bhat, S.J. Traffic flow prediction models—A review of deep learning techniques. *Cogent Eng.* **2022**, *9*, 2010510. [CrossRef]
6. Wang, S.; Shao, C.; Zhang, J.; Zheng, Y.; Meng, M. Traffic flow prediction using bi-directional gated recurrent unit method. *Urban Inform.* **2022**, *1*, 16. [CrossRef]
7. Zhang, X.; Yu, G.; Shang, J.; Zhang, B. Short-term Traffic Flow Prediction With Residual Graph Attention Network. *Eng. Lett.* **2022**, *30*, 4.
8. Wen, Y.; Xu, P.; Li, Z.; Xu, W.; Wang, X. RPConvformer: A novel Transformer-based deep neural networks for traffic flow prediction. *Expert Syst. Appl.* **2023**, *218*, 119587. [CrossRef]
9. He, R.; Xiao, Y.; Lu, X.; Zhang, S.; Liu, Y. ST-3DGMR: Spatio-temporal 3D grouped multiscale ResNet network for region-based urban traffic flow prediction. *Inf. Sci.* **2023**, *624*, 68–93. [CrossRef]
10. Razali, N.A.M.; Shamsaimon, N.; Ishak, K.K.; Ramli, S.; Amran, M.F.M.; Sukardi, S. Gap, techniques and evaluation: Traffic flow prediction using machine learning and deep learning. *J. Big Data* **2021**, *8*, 152. [CrossRef]
11. Cengil, E.; Çınar, A.; Yıldırım, M. A hybrid approach for efficient multi-classification of white blood cells based on transfer learning techniques and traditional machine learning methods. *Concurr. Comput. Pract. Exp.* **2021**, *34*, e6756. [CrossRef]
12. Polson, N.G.; Sokolov, V.O. Deep Learning for Short-Term Traffic Flow Prediction. *Transp. Res. Part C* **2017**, *79*, 1–17. [CrossRef]
13. Alajali, W.; Zhou, W.; Wen, S.; Wang, Y. Intersection Traffic Prediction Using Decision Tree Models. *Symmetry* **2018**, *10*, 386. [CrossRef]
14. Chen, Y.; Wang, W.; Chen, X.M. Bibliometric methods in traffic flow prediction based on artificial intelligence. *Expert Syst. Appl.* **2023**, *228*, 120421. [CrossRef]
15. Sun, Z.; Li, Y.; Pei, L.; Li, W.; Hao, X. Classification of Coarse Aggregate Particle Size Based on Deep Residual Network. *Symmetry* **2022**, *14*, 349. [CrossRef]
16. Xianglong, L.; Liyao, N.; Shengrui, Z. An Algorithm for Traffic Flow Prediction Based on Improved SARIMA and GA. *KSCE J. Civ. Eng.* **2018**, *22*, 4107–4115. [CrossRef]
17. Marcelino, P.; de Lurdes Antunes, M.; Fortunato, E.; Gomes, M.C. Transfer learning for pavement performance prediction. *Int. J. Pavement Res. Technol.* **2020**, *13*, 154–167. [CrossRef]
18. Wang, X.; Xiao, J.; Fan, M.; Gang, C.; Tang, Y. Short-Term Traffic Flow Prediction Based on Ga-Bp Neural Network. *Adv. Comput. Signals Syst.* **2022**, *6*, 75–82.
19. Zhang, L.; Alharbe, N.R.; Luo, G.; Yao, Z.; Li, Y. A hybrid forecasting framework based on support vector regression with a modified genetic algorithm and a random forest for traffic flow prediction. *Tsinghua Sci. Technol.* **2018**, *23*, 479–492. [CrossRef]
20. Wang, D.; Wang, C.; Xiao, J.; Xiao, Z.; Chen, W.; Havyarimana, V. Bayesian optimization of support vector machine for regression prediction of short-term traffic flow. *Intell. Data Anal.* **2019**, *23*, 481–497. [CrossRef]
21. Lu, B.; Gan, X.; Jin, H.; Fu, L.; Wang, X.; Zhang, H. Make More Connections: Urban Traffic Flow Forecasting with Spatiotemporal Adaptive Gated Graph Convolution Network. *ACM Trans. Intell. Syst. Technol.* **2022**, *13*, 1–25. [CrossRef]
22. Chai, C.; Ren, C.; Yin, C.; Xu, H.; Meng, Q.; Teng, J.; Gao, G. A Multifeature Fusion Short-Term Traffic Flow Prediction Model Based on Deep Learnings. *J. Adv. Transp.* **2022**, *2022*, 1702766. [CrossRef]
23. Fang, W.; Zhuo, W.; Song, Y.; Yan, J.; Zhou, T.; Qin, J. [formula omitted]-LSTM: An error distribution free deep learning for short-term traffic flow forecasting. *Neurocomputing* **2023**, *526*, 180–190. [CrossRef]
24. Lan, T.; Zhang, X.; Qu, D.; Yang, Y.; Chen, Y. Short-Term Traffic Flow Prediction Based on the Optimi-zation Study of Initial Weights of the Attention Mechanism. *Sustainability* **2023**, *15*, 1374. [CrossRef]
25. Zhang, Z. Prediction of traffic flow based on deep learning. *Int. J. Adv. Comput. Technol.* **2020**, *9*, 5–11.
26. Hao, X.; Liu, Y.; Pei, L.; Li, W.; Du, Y. Atmospheric Temperature Prediction Based on a BiLSTMAttention Model. *Symmetry* **2022**, *14*, 2470. [CrossRef]
27. Zhuang, W.; Cao, Y. Short-Term Traffic Flow Prediction Based on CNN-BILSTM with Multicomponent Information. *Appl. Sci.* **2022**, *12*, 8714. [CrossRef]
28. Sun, Z.; Pei, L.; Xu, L. Intelligent Detection and Restoration of Road Domain Environmental Perception Data Based on DS-LOF and GA-XGBoost. *J. China Highw. Eng.* **2023**, *36*, 15–26.
29. Du, Q.; Yin, F.; Li, Z. Base station traffic prediction using XGBoost-LSTM with feature enhancement. *IET Netw.* **2020**, *9*, 0103. [CrossRef]
30. Sun, B.; Sun, T.; Jiao, P. Spatio-Temporal Segmented Traffic Flow Prediction with ANPRS Data Based on Improved XGBoost. *J. Adv. Transp.* **2021**, *2021*, 1–24. [CrossRef]
31. Tumash, L.; Canudas-de-Wit, C.; Delle Monache, M.L. Multi-directional continuous traffic model for large-scale urban networks. *Transp. Res. Part B* **2022**, *158*, 374–402. [CrossRef]
32. Hu, Y.; Ma, T.; Chen, J. Multi-anticipative bi-directional visual field traffic flow models in the connected vehicle environment. *Phys. A Stat. Mech. Its Appl.* **2021**, *584*, 126372. [CrossRef]

33. Pei, L.; Sun, Z.; Yu, T.; Li, W.; Hao, X.; Hu, Y.; Yang, C. Pavement aggregate shape classification based on extreme gradient boosting. *Constr. Build. Mater.* **2020**, *256*, 119356–119369. [CrossRef]
34. Yuan, C.; He, C.; Xu, J.; Liao, L.; Kong, Q. Bayesian Optimization for Selecting Efficient Machine Learning Regressors to Determine Bond-slip Model of FRP-to-concrete Interface. *Structures* **2022**, *39*, 351–364. [CrossRef]
35. Zhou, J.; Qiu, Y.; Zhu, S.; Armaghani, D.J.; Khandelwal, M.; Mohamad, E.T. Estimation of the TBM Advance Rate Under Hard Rock Conditions Using XGBoost and Bayesian Optimization. *Undergr. Space* **2020**, *6*, 506–515. [CrossRef]
36. Pei, L.; Sun, Z.; Hu, Y.; Li, W.; Gao, Y.; Hao, X. Neural Network Model for Road Aggregate Size Calculation Based on Multiple Features. *J. South China Univ. Technol.* **2020**, *48*, 77–86.