A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction

Honglei Ren[*], You Song[*], Jingwen Wang[*], Yucheng Hu[+], and Jinzhi Lei[+]

[*]School of Software, Beihang University, Beijing, China,
songyou@buaa.edu.cn, renhongleiz@buaa.edu.cn,wangjingwen@buaa.edu.cn
[+]Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, China,
huyc@tsinghua.edu.cn, jzlei@mail.tsinghua.edu.cn

Abstract— With the rapid development of urbanization, the boom of vehicle numbers has resulted in serious traffic accidents, which led to casualties and huge economic losses. The ability to predict the risk of traffic accident is important in the prevention of the occurrence of accidents and to reduce the damages caused by accidents in a proactive way. However, traffic accident risk prediction with high spatiotemporal resolution is difficult, mainly due to the complex traffic environment, human behavior, and lack of real-time traffic-related data.Based on the quantitative analysis of big traffic accident data, this paper first introduced an important characteristic of traffic accidents — the spatiotemporal correlation, and then constructed a high accurate deep learning model for traffic accident risk prediction based on spatiotemporal correlation pattern. The predictive accident risk can be potential applied to the traffic accident warning system. The proposed method can be integrated into an intelligent traffic control system toward a more reasonable traffic prediction and command organization.

I. INTRODUCTION

In modern society, the rapid development of urbanization has resulted in the boom of vehicles, causing a number of problems, such as traffic congestion, air pollution, and traffic accidents. These problems have caused huge economic loss as well as human casualties. According to *Global Status Report on Road Safety*, published by World Health Organization in 2015, about 1.25 million people were killed in traffic accidents every year. With the help of big traffic data and deep learning, real-time traffic flow prediction has enabled people to avoid traffic jam by choosing less congested routes. Big traffic data and deep learning may also provide a promising solution to predict or reduce the risk of traffic accidents.

One important task in traffic accident prevention is to build an effective traffic accident risk prediction system. If the traffic accident risk in a certain region can be predicted, we can disseminate this information to the nearby drivers to alert them or make them choose a less hazardous road. However, accurate prediction of traffic accident risk is very difficult because many related factors could affect traffic accident. For example, different regions have tremendous difference on traffic accident rate. In addition, poor weather condition such as snow or fog can reduce road visibility and traffic capacity, thus increase the change of traffic accidents. Traffic accident

rate varies at different time of a day, possibly related to the physical condition of the drivers. Although many researchers have focused on the identification of key factors associated with traffic accident [1], effective prediction of the traffic accident risk dynamically remains to be a challenge problem.

Based on the historical statistic of traffic accident collected over the years, one can get a rough overview of the risk regions. However, the risk distribution varied a lot according to the time of day, day of week, month of year, and complex factors can affect the accident risk such as the crowd density, traffic flow, weather, events, etc. Besides, the historical statistic can not achieve the fine-grained and dynamic accident risk prediction. Therefore, to forecast the dynamic accident risk change in a fine-grained and quantitative way, it is essential to introduce the deep learning method into traffic accident risk prediction. Chen use human mobility features extracted from Stack denoise Autoencoder to infer traffic accident risk in Japan [2]. However, they did not consider the periodical patterns and the spatial distribution patterns of traffic accidents.

In this paper, we collected big traffic accident data and built a deep model for traffic accident risk prediction based on recurrent neural network. The workflow of our prediction method can be illustrated by Figure 1. First, we discretized the big traffic accident data in space and time, so that it can be processed by machine learning algorithm. By analyzing the spatial and temporal patterns of traffic accident frequency, we presented the spatiotemporal correlation of traffic accidents. Based on the patterns, we constructed a deep model on the basis of recurrent neural network to infer traffic accident risk, and input the processed data into it. After the data training, we input the recent traffic accident frequency data into the trained model, and then obtained the predicted accident risk map from the output. As a potential application, the traffic accident prediction system based on our method can be used to help traffic enforcement department to allocate police forces in advance of traffic accidents.

The rest of this paper is organized as follows: Section 2 introduces some previous works that are related with the present one. Section 3 describes the data source and the pattern analysis result of traffic accidents. Section 4 introduces our deep learning model for traffic accident risk
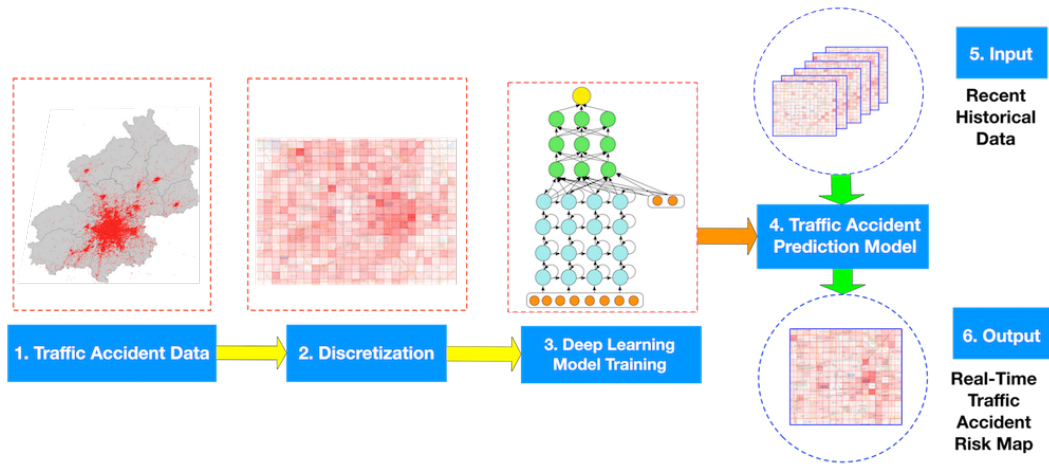
Fig. 1. Workflow of our traffic accident risk prediction method. First, the big traffic accident data is collected. Second, the data is discretized in space and time, and then feed to the deep model for training. After training, we feed recent historical data to the trained model and then obtained the real-time traffic accident risk prediction.

prediction. Section 5 shows the results of experiment. Section 6 gives the conclusions and future works.

II. RELATED WORK

1) Identification of Traffic Accident Trigger: Tremendous efforts have been devoted to the identification of key conditions or particular traffic patterns that could lead to traffic accident. For instance, Oh proposed the assumption that disruptive traffic flow is a trigger to crash [3]. Based on the loop detector data and crash data, they found that 5-min standard deviation of speeds right before a traffic accident is an effective indicator of crash. Although different crash indicators have been proposed, they could not meet the requirement of accurate accident prediction because numerous factors have complex connections with traffic accidents.

2) Spatiotemporal Traffic Prediction: The spatiotemporal dependency is a significant feature of traffic, and can be utilized to improve accuracy of traffic prediction. Yue investigates the spatiotemporal dependency of traffic flow by cross-correlation analysis, and shows its importance in traffic forecastability evaluation [4]. Asif proposed an unsupervised learning methods to infer the spatiotemporal patterns in a large-scale traffic speed prediction[5]. Pan proposed a model to predict the spatiotemporal impact of happend incidents on its surrounding traffic based on the real time traffic and incident data [6]. Yu developed a spatiotemporal recurrent convolutional networks to capture the spatial dependencies and temporal dynamics of network-wide traffic [7].

3) Real-time Traffic Accident Prediction: With the development of machine learning, many researchers start to focus on real-time traffic accident prediction. Lv chose feature variables based on Euclidean metric and utilized k-nearest neighbor method to predict traffic accident [8]. Park collected big traffic accident data of highway in Seoul and build a prediction workflow based on k-means cluster analysis and logistic regression [9]. Recently, Chen used human mobility data in Japan and build a Stack denoise Autoencoder to infer

the real-time traffic risk [2]. One limitation of these works is that, they did not consider the temporal patterns of traffic accident itself into their model. Without these information, the predictive power of the model could be weakened.

III. PATTERN ANALYSIS OF TRAFFIC ACCIDENT

A. Big Traffic Accident Data

In this study, to predict traffic accident risk, the traffic accident records of Beijing in 2016 and 2017 was collected. Each record contains the time, GPS (Global Positioning System) coordinate of the accident event.

B. Data Preprocessing

Before we analyze the pattern of accident, and build machine learning model, a proper data structure is necessary. Therefore, we first preprocess our raw data by discretization.

The traffic accident data was first discretized in space and time. The temporal resolution was 1 hour for different time horizon of prediction, and spatial resolution dimension was 1000m×1000m in uniform grids.

After discretization, we obtained a matrix S whose element $S_{r,t}$ is the count of traffic accidents happened within region r and time slot t.

C. Spatial Distribution of Traffic Accident

To explore whether traffic accident frequency is associated with the geographical position of a region, we plot the heatmap of traffic accident frequency in Beijing in 2016 (Figure 2). As shown in Figure 2, the traffic accident frequency is not uniform distributed, and it is highly related with the geographical position of a region. Usually, the highest traffic accident region lies in the major commercial and business areas.
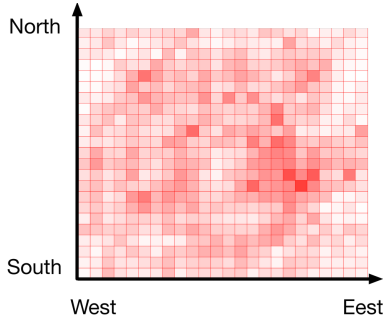
Fig. 2. The heatmap of traffic accident frequency in Beijing in 2016 with 1000m*1000m spatial resolution. Deeper red indicates higher frequencies of traffic accident.

D. Temporal Pattern of Traffic Accident

To explore the temporal patterns of the traffic accident frequency, we first checked whether everyday's traffic accident count varies in different time period. Figure 3 gives the scatter and box-plot of the everyday's traffic accident count for different time periods of Beijing. Obviously, the traffic accident patterns change drastically for different time period of a day. Specifically, traffic accident is more frequent at rush hours than that at off-peaks.

The time periods in Figure 3 is defined according to working time pattern and Chinese lifestyle [1]: 00:00–06:59 (mid-night to dawn), 07:00–08:59 (morning rush hours), 09:00–11:59 (morning working hours), 12:00–13:59(lunch break), 14:00–16:59 (afternoon working hours), 17:00–19:59 (afternoon rushing hours), and 20:00–23:59 (nighttime).
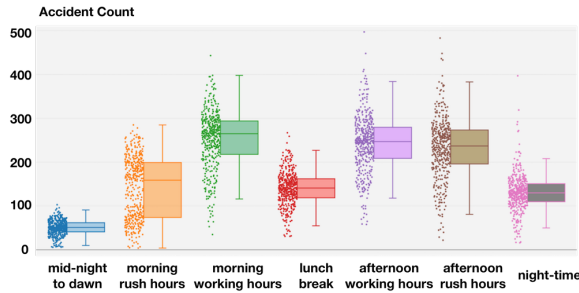

Fig. 3. Scatter and box-plot of everyday's traffic accident count for different time periods in Beijing.

To quantify the spatio-temporal correlation between traffic accident, we first defined the spatial correlation for a given time t as follows:

$$C(k,t) = \frac{\sum_{i,j}(a_{i,j,t} - \overline{a}_t)(a_{i',j',t} - \overline{a}_t)}{\sum_{i,j}(a_{i,j,t} - \overline{a}_t)^2} \quad (1)$$

where

$$\overline{a}_t = \frac{\sum_{i,j} a_{i,j,t}}{M * N}$$

The $C(k,t)$ in Eq.(1) is the spatial correlation with a k Manhattan distance for a given time t. $a_{i,j,t}$ is the traffic accident count happened in grid (i,j) and time t. $\overline{a}_t$ is the

average traffic accident count of all grids at time t. M and N are the number of grids along the longitude and latitude.

Based on the spatial correlation defined by Eq.(1), the spatio-temporal correlation can be written as Eq.(2).

$$f(k,\tau) = \frac{\sum_t (C(k,t) - \overline{C}(k))(C(k,t+\tau) - \overline{C}(k))}{\sum_t (C(k,t) - \overline{C}(k))^2} \quad (2)$$

where $f(k,\tau)$ is the correlation of two grids with a k Manhattan distance and time interval τ.
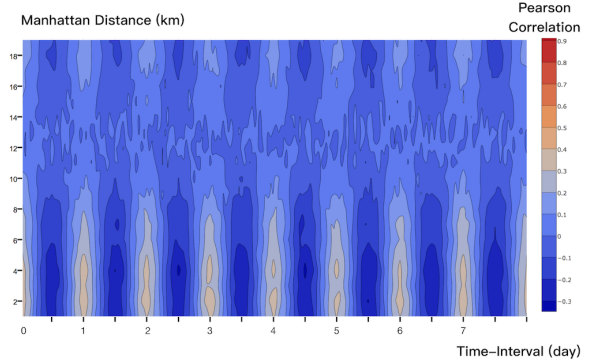

Fig. 4. The contour map of the spatio-temporal correlation of traffic accident. The horizontal and vertical axis are the time-interval and Manhattan distance of two grids, respectively.

Figure 4 shows the contour map of the spatio-temporal correlation of traffic accident. It can be observed that the correlation shows a strong temporal periodic pattern, and the period is around 24 hours. Traffic accidents have about 0.4 ∼ 0.5 correlation if their Manhattan distance is within 4 km, and time interval is the multiples of 24 hours.

IV. DEEP MODEL OF TRAFFIC ACCIDENT RISK PREDICTION

As Chen et. al. has documented, after some analysis of traffic accident data, we find it is difficult to predict whether traffic accident will happen or not directly, because complex factors can affect traffic accident, and some factors, such as the distraction of drivers, can not be observed and collected in advance [2]. Figure 5 gives the dimensionality reduction result by t-SNE when we predict whether traffic accident happen or not directly, the input data is the sequence of traffic accident count for a region. Obviously, the red points (accident) and black points (non-accident) are inseparable, and that means it is hard to predict whether traffic accident happen or not directly. Therefore, we try to predict traffic accident frequency(risk), that is average traffic accident count per hour for the same time of recent days (3 days, 7 days, 30 days, etc.). For instance, if 5 accidents happened during 8:00-9:00 a.m. in last 3 days, then today's traffic accident frequency during 8:00-9:00 a.m is $\frac{5 \ times}{3 \ hours} \approx 1.67 \ times/hour$. The results in the following sections are all based on 3-day traffic accident frequency prediction, and the results of other frequency periods are similar with that of 3-day.
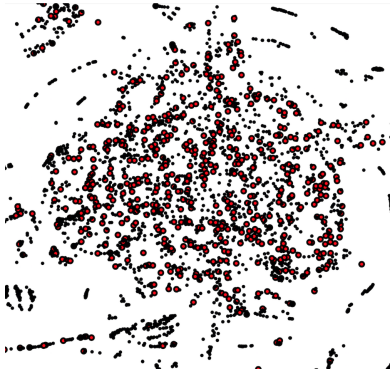
Fig. 5. Dimensionality reduction and visualization with t-SNE, the original data is the sequence of traffic accident count for a region. The red points are the data with traffic accident, and the black points are the traffic accident free data
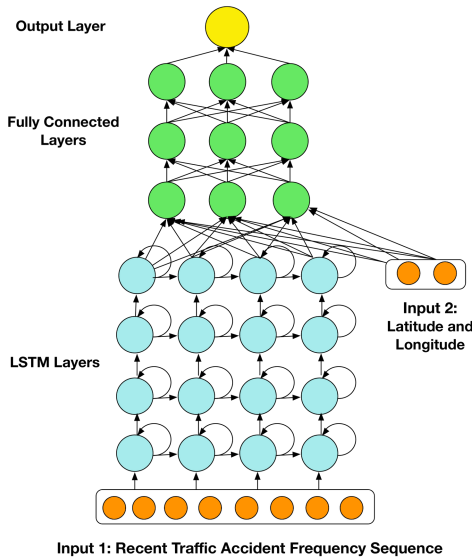

Fig. 6. The Deep Model of Traffic Accident Risk Prediction Method based on LSTM. The model consisted of 2 separated input layers, 4 LSTM layers, 3 fully connected layers and 1 output layer. The first input layer is made up of the sequence of recent traffic accident frequency. The second input layer contains the longitude and latitude of the region center that we expected to predict.

A. Model

In this subsection, we will introduce our Traffic Accident Risk Prediction Method based on LSTM (TARPML), and Figure 6 illustrates the deep model of TARPML. The input layers are consisted of two parts. The first input is the sequence of recent traffic accident frequency, and it is input to the first LSTM layer. The second input contains the longitude and latitude of the region center that we expected to predict, and it directly input into Fully-Connected layers. The hidden layers of deep model is consisted of 4 LSTM layers and 3 Fully-Connected layers sequentially. Here, the number of LSTM layers and Fully-Connected layers are decided by performance comparison with different parameters. Parameters with the best performance and the least network size is chosen according to Table I. The last layer of model

is output layer, which outputs the predicted traffic accident risk(frequency) for the given input.

Since the network is always queried with one particular grid cell location, when we perform the prediction for the whole city, we do a full sweep over all grids, and output the prediction for each grid cell.

To avoid overfitting, we add a dropout layer with 0.5 dropout rate between each two fully connected layers. The activation function of fully connected layers and output layer is Rectified Linear Units (RELU), which can be denoted as $\max(0, x)$ mathematically.

The reason why we chose LSTM is that LSTM can capture the periodic feature of traffic accident, and traditional RNNs shows poor performance and intrinsic difficulties in training when it has long time period. On another hand, the explicit memory cell in LSTM can avoid the problems of gradient vanish or gradient explosion existed in traditional RNNs.

V. EXPERIMENTS AND RESULTS

In this section, we compare our TARPML method with several baseline models, including Lasso, Support Vector Regression, Random Forest Regression, etc. All of the experiments are performed by a PC (CPU: Intel Xeon(R) CPU E5-2609, 32GB memory, GPU: Tesla K20C).

A. Experimental Setup

Because our model is temporal related, we arrange the data chronologically. We chose the data from 2016-01-01 to 2017-04-01 as the training data, and the data from 2017-04-01 to 2017-08-20 is for testing. The last 20% of training data is used as validation data. The sample size of training, validation, testing is 1590958, 397740 and 233850 respectively.

The architecture of TARPML are built upon Keras, which is a Python Deep Learning library. We chose mean squared error as objective function of optimization, and selected RMSProp as the optimizer.

By comparing the Root Mean Square Error (RMSE) of TARPML method with different input sequence lengths and number of neurons per layer (Table II), we finally chose 100 as the best sequence length to input, and 200 as the number of neurons per layer except the first LSTM layer.

B. Performance Evaluation

1) Evaluation Metrics: To evaluate the accuracy and precision of the prediction, we selected Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as our metrics. They are defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |r_i - \widehat{r_i}| \tag{3}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (r_i - \widehat{r_i})^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (r_i - \widehat{r_i})^2} \tag{5}$$

TABLE I

PERFORMANCE OF TARPML METHOD WITH DIFFERENT LSTM AND FULLY-CONNECTED LAYER NUMBER

| Number of LSTM layers (right) | 1 | 2 | 3 | 4 | 5 | 6 |
Number of Fully-Connected layers (below)						
1	0.046	0.044	0.039	0.038	0.039	0.038
2	0.044	0.040	0.038	0.036	0.036	0.036
3	0.042	0.039	0.037	0.034	0.034	0.035
4	0.041	0.028	0.037	0.035	0.035	0.034

TABLE II

PERFORMANCE OF TARPML METHOD WITH DIFFERENT INPUT SEQUENCE LENGTHS AND NEURON NUMBERS

| Input sequence length (right) | 10 | 20 | 50 | 100 | 150 | 200 |
Number of neurons per layer (below)						
100	0.049	0.047	0.042	0.040	0.041	0.040
200	0.042	0.041	0.038	0.034	0.035	0.035
300	0.042	0.042	0.039	0.037	0.037	0.037
500	0.044	0.044	0.041	0.039	0.038	0.039

where n is the sample size, r_i and $\widehat{r_i}$ are real and predicted risk (accident frequency) respectively.

2) Baseline Models: We selected several traditional machine learning models as our baseline models to compare the prediction performance with our TARPML method. The baseline models we selected are Lasso, Ridge, Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Multilayer Perceptron(MLP) and Autoregressive Moving Average Model (ARMA) .

3) Performance Evaluation: We compared the predictability of our model with that of baselines on test set, and Table III demonstrates their MAE, MRE and RMSE values for 3-day traffic accident frequency(Its definition can be found at Section IV). The table shows that our model outperforms than other models, and have less prediction errors.

All the baseline models were implemented by scikit-learn, and the most important parameters we used for models are listed in Table IV, the default values were applied for the not listed parameters of scikit-learn models.

4) Simulation Results: To evaluate the effectiveness of our model, here we selected 2017-07-10 (Monday) as a example for comparing the prediction results of different models. Figure 7 (a) are the real traffic accident risk map (the leftmost one) and the predicted results of different models, and it can be clearly seen that TARPML model is far better than other models. Figure 7(b) are the predicted risk curve from different models and its corresponding real traffic risk curve. It can also observed that the predicted curve from TARPML model are more accurate than others.

VI. CONCLUSION

In this paper, we collected big traffic accident data, and built a deep learning model based on LSTM for predicting

TABLE III

PERFORMANCE FOR 3-DAY TRAFFIC ACCIDENT FREQUENCY PREDICTION

Method	MAE	MSE	RMSE
Lasso	0.046	0.006	0.076
Ridge	0.024	0.002	0.045
SVR	0.066	0.006	0.075
DTR	0.021	0.004	0.058
RFR	0.020	0.002	0.047
ARMA	0.058	0.049	0.169
MLP	0.023	0.034	0.060
TARPML	0.014	0.001	0.034

traffic accident risk. Based on the pattern analysis result, it can be observed that the traffic accident risk are not uniformly distributed in space and time. It shows strong periodical temporal patterns and regional spatial correlation. According to the dimensionality reduction result (t-SNE, Figure 5), it is hard to predict traffic accident directly. Therefore, we defined the traffic accident risk by its frequency, and built a deep learning model based on LSTM to capture its spatial and temporal patterns. The performance comparison based on RMSE (Table III) and the predicted risk map (Figure 7) shows the accuracy and effectiveness of our model. Our method can be easily applied to the traffic accident warning system and help people avoiding traffic accident by choosing safer regions.

However, due to the complexity of traffic accident, our study has some limitations in following aspects. First, here we only utilized the traffic accident data itself for prediction. However, other related data, such as traffic flow, human mobility, road characteristic and special events, maybe significant to traffic accident risk prediction as well. Second, our
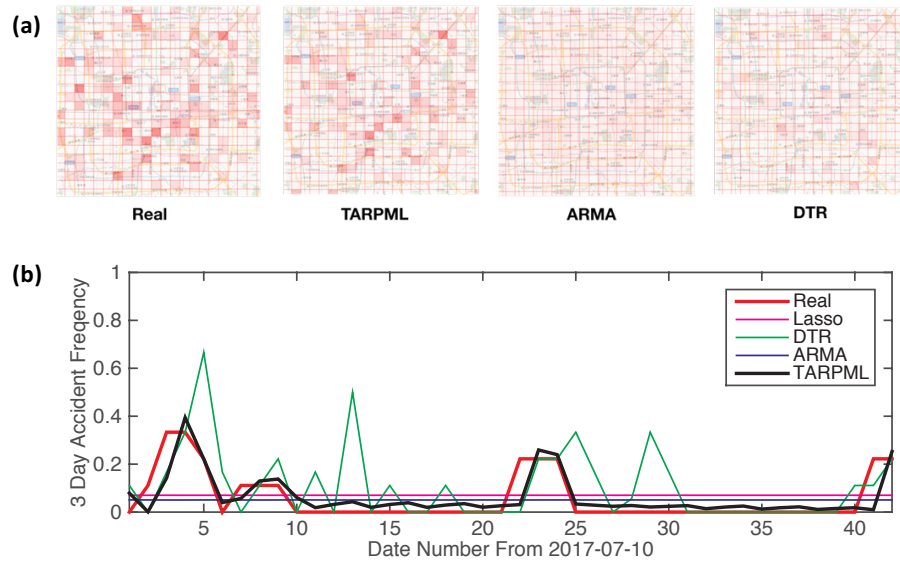
Fig. 7. (a) Comparison of the real traffic accident risk map (the leftmost one) and the predicted results of different models. (b) are the predicted risk curve from different models and its corresponding real traffic risk curve.

TABLE IV

THE PARAMETERS OF THE MODELS IN TABLE III

Method	Parameters
Lasso	alpha = 0.1
Ridge	alpha = 0.5, solver = 'auto'
SVR	kernel='linear', C=1, epsilon=0.1, shrinking=True, tol=1e-3
DTR	max_leaf_nodes=None, criterion='mse', max_features=None, max_depth =None
RFR	n_estimators = 20, criterion='mse', max_features='auto', max_depth=None
ARMA	p and q varies for different regions and time slots
MLP	solver='adam', activation = 'relu', alpha=1e-5, hidden_layer_sizes=(100, 100), random_state=1
TARPML	dropout_rate=0.2, activation='relu', loss='mse', optimizer='rmsprop'

prediction results are coarse-grained, and can not provide road level accident risk prediction. But it can be easily applied to the road network based prediction. Therefore, future work combined with structure of urban road network and comprehensive factors related with traffic accident will be promising to make better prediction result.

REFERENCES

[1] G. Zhang, K. K. W. Yau, and G. Chen, "Risk factors associated with traffic violations and accident severity in China." *Accident; analysis and prevention*, vol. 59, pp. 18–25, Oct. 2013.

[2] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, "Learning deep representation from big and heterogeneous data for traffic accident inference," 2016.

[3] C. Oh, J.-S. Oh, S. Ritchie, and M. Chang, "Real-time estimation of freeway accident likelihood," in *80th Annual Meeting of the Transportation Research Board, Washington, DC*, 2001.

[4] Y. Yue and A. G.-O. Yeh, "Spatiotemporal traffic-flow dependency and short-term traffic forecasting," *Environment and Planning B: Planning and Design*, vol. 35, no. 5, pp. 762–771, 2008.

[5] M. T. Asif, J. Dauwels, C. Y. Goh, A. Oran, E. Fathi, M. Xu, M. M. Dhanya, N. Mitrovic, and P. Jaillet, "Spatiotemporal patterns in large-scale traffic speed prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 794–804, 2014.

[6] B. Pan, U. Demiryurek, C. Shahabi, and C. Gupta, "Forecasting spatiotemporal impact of traffic incidents on road networks," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 587–596.

[7] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma, "Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks," *Sensors*, vol. 17, no. 7, p. 1501, 2017.

[8] Y. Lv, S. Tang, and H. Zhao, "Real-time highway traffic accident prediction based on the k-nearest neighbor method," in *Measuring Technology and Mechatronics Automation, 2009. ICMTMA'09. International Conference on*, vol. 3. IEEE, 2009, pp. 547–550.

[9] S.-h. Park, S.-m. Kim, and Y.-g. Ha, "Highway traffic accident prediction using vds big data analysis," *The Journal of Supercomputing*, vol. 72, no. 7, pp. 2815–2831, 2016.