Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on the Law and Prediction of Traffic Accidents Based on Value at Risk in the Context of Intelligent Transportation

Zihang Cheng[1, *], Zihan Shao[2] and Xinyi Wang[3]

[1]School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing 210094, China

[2]School of Big Data and Software Engineering, Zhejiang Wanli University of Software Engineering, Ningbo, 315100, China

[3]Chang'an Dublin International College of Transportation, Chang'an University, Xi'an, 710021, China

*Corresponding author: ChengZH@njust.edu.cn

Abstract. The research introduces traffic accident predictive models known as the Convolutional Long Short-Term Memory Model and a K-means cluster algorithm with Random Forest Network, which aims to provide precise forecasts of incident rates Through the amalgamation of the strengths inherent in Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) neural network, K-means cluster algorithm and Random Forest algorithms. These fusion models surpass the performance of the standalone LSTM model, particularly excelling in the prediction of incident rates during driving hours and the location. The study's outcomes reveal that the proposed model exhibits remarkable proficiency in specific transformation domains, underscoring its superior efficacy compared to single conventional models. These results underscore the advantages of integrating multiple algorithms within a unified framework to enhance predictive accuracy significantly. Consequently, the Convolutional Long Short-Term Memory and K-means cluster algorithm with a Random Forest Network emerge as promising solutions for advancing incident rate forecasting capabilities.

Keywords: Traffic accident; deep learning algorithm; prediction; ConvLSTM Model.

1. Introduction

Accurate and reliable traffic prediction is of great importance in intelligent transportation systems. It can not only help decision-makers formulate more scientific and reasonable management strategies but also help people flexibly adjust their travel plans. Yao reviewed the research progress and future development direction in the field of traffic flow prediction [1]. According to the different spatial scopes of prediction, the current traffic flow prediction methods are classified, and the research status of traffic flow prediction of single road sections and regional road networks at home and abroad is analyzed. The development of urban transport has made traffic accident safety a crucial issue. To prevent and reduce the occurrence of traffic accidents, research on traffic accident prediction technology has gained significant attention. Wang provides an overview of different techniques used for predicting traffic accidents. These include traditional statistical analysis methods, machine learning methods, neural network methods, and time series analysis methods. The advantages and disadvantages of each technique are analyzed [2]. Yan designed an urban traffic accident risk prediction model based on edge computing to predict traffic accident risk from the perspectives of real-time vehicle information and historical traffic accident data information [3]. Zhao proposed a traffic accident risk prediction algorithm based on a convolutional neural network, which is a combination of edge computing and deep learning, making deep learning directly computed on edge devices in a network-free environment, avoiding the latency problem caused by the network transmission [4]. By extracting the main features of short-term traffic flow data and tuning the model parameters, Wang constructed recurrent neural network models. The two models were also combined, and the BP neural network was applied for nonlinear mapping transformation, to improve the

accuracy of traffic flow prediction [5]. Zhang used the road section as the prediction unit and used graph convolution and long short-term memory network to construct a short-term risk prediction method for urban traffic accidents based on road network structure (TARPBRN) [6]. This method can predict the risk of traffic accidents in the short term of the specified road section so that it can be targeted and reduce the occurrence of traffic accidents.

Guo proposed a spatiotemporal traffic prediction network based on deep learning, ST-3DNet, which proposed to solve the problem of spatiotemporal grid data prediction [7]. This 3D convolution performs convolution operations in the third dimension, allowing it to retain the information in the time dimension and extract the features in both spatial and temporal dimensions simultaneously, leading to a better prediction of future traffic data. Ruan et al. proposed to use of adaptive graph convolutional networks to learn the spatial correlation between road sections and generate enough accident risk samples through data augmentation [8]. At the same time, the contrastive learning method is used to better distinguish between risk and non-risk samples, to achieve more accurate accident risk prediction. The results show that the proposed method has achieved significant performance improvement on the real traffic dataset of the Guilin expressway network. Wang proposed an urban traffic accident risk prediction model which based on spatiotemporal characteristics, where an improved spatiotemporal graph convolutional network was used, the graph convolutional network (GCN) was used to extract spatially correlated features, and a batch standardization layer was added to solve the gradient vanishing explosion problem [9]. Zhu preprocessed multi-source heterogeneous data with the help of machine learning and deep learning models to achieve the purpose of accurately and effectively, aimed to predict the risk of traffic accidents in urban areas [10]. Chen proposed a city-level traffic accident risk prediction method called "Stacked Noise Reduction Convolutional Auto Coding Algorithm (SDCAE)" [11]. The method integrates multi-source data such as traffic accident data, real-time traffic flow, and weather, and uses high-dimensional features for risk prediction. The algorithm performs better than the traditional algorithm in terms of prediction error and is of great significance for the traffic command and management department to deploy police forces in advance, alleviate traffic congestion, and provide risk prediction and path planning for drivers.

Based on the UK car accident 2005-2015 dataset, Yu found that driving time, geographical location, road shape, driving speed, and physical facilities are the main factors that induce traffic accidents [12]. Wang proposed STRiskNet, a traffic accident risk prediction model that integrates local and global spatiotemporal characteristics, to model traffic accident risk from local spatiotemporal correlation and global spatiotemporal correlation [13]. Wei introduced a deep learning theory to construct an urban traffic accident risk prediction model considering spatiotemporal characteristics, integrating real-time weather factors, POI, and traffic flow-related characteristics [14]. The improved Spatiotemporal Graph Convolution (ISTGCN) network is used in the model. Based on the AFC data of Xi'an rail transit in 2019, the annual rainfall data of Xi'an, and the POI data of Xi'an, Guo revealed the law between the amount of rainfall and the fluctuation of public rail transit passenger flow [15]. Huang combined ConvLSTM and Attention neural network to design a spatiotemporal attention convolutional long short-term memory network (ST-AttConvLSTMs) to model and predict the traffic accident risk of the whole city [16]. In view of the good performance of CNN and LSTM and their wide applications in different fields, the ConvLSTM network has been further proposed and successfully applied to spatial and temporal precipitation prediction [17].

In this paper, a ConvLSTM-based model with the number of traffic accidents occurring in the Manhattan downtown area for the years 2020-2023 is continued to explore the fitting of the number accurately and work on completing the predictive quantity model thereafter. In the meantime, the K-means cluster algorithm and Random Forest network were borrowed to explore the relationship between the Manhattan downtown area and the number of traffic accidents and draw conclusions.

2. Methods

2.1. Data Source

In the meantime, the traffic accident data of the Manhattan District (see Fig. 1 below) is selected as the object of study. As an economic and trade center and a densely populated area, the Manhattan district faces great traffic pressure. Therefore, it is of great practical significance to take the traffic accidents in Manhattan as the research object for urban traffic accident risk prediction.

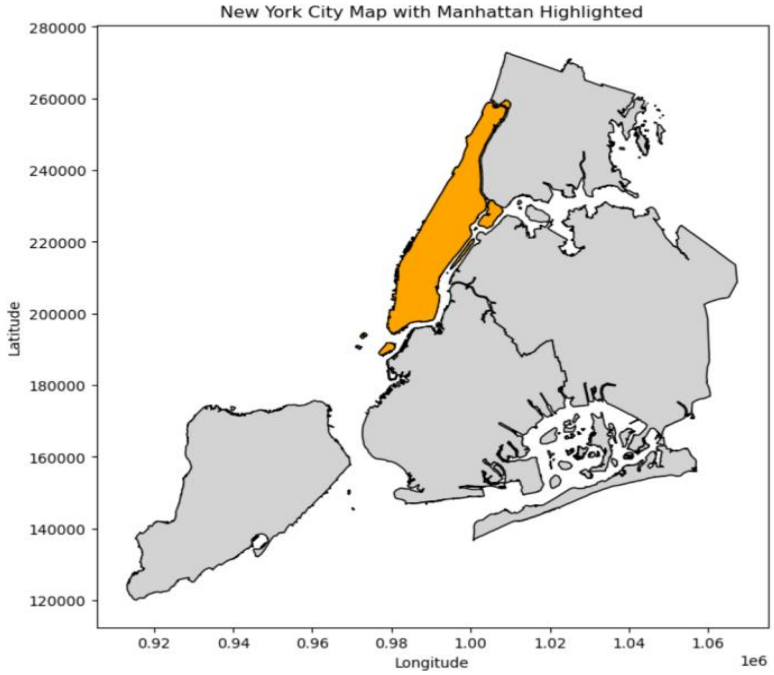

Fig. 1 New York City Map with Manhattan Highlighted

2.2. Indicator Selection and Description

This paper studies and tests all traffic accidents that occurred in Manhattan in the entire year of 2023, with 15,458 of them [18]. Each piece of data contains some basic information as shown in Table 1:

Table 1. Traffic accident data introduction

field	typical example	Data Type
Crash date	2023-11-17	Date
Crash time	21:28	Time
Latitude	40.734500	Float
Longitude	-74.001180	Float
Number of persons killed	1.0	Float
Number of persons killed	0.0	Float

According to statistics, there were a total of 15,458 traffic accidents in Manhattan in 2023, with three specific types of accidents: fatal accidents, accidents with injuries, and accidents with no injuries, as shown in Table 2:

Table 2. Statistics on the types and numbers of traffic accidents in Manhattan in 2023

Types of accidents	number	Data Type
injuries	6167	integer
fatalities	38	integer
no casualties	9253	Integer

2.3. Method Introduction

This paper focuses on the three deep learning algorithms: the convolutional neural network algorithm, the random forest neural network algorithm, and long and short-term memory neural network algorithm.

2.3.1. K-means clustering algorithm

The K-Means algorithm first needs to select the initial clustering center, then classify all the data points, and finally calculate the average value of each cluster to adjust the clustering center in a continuous iterative loop. Eventually, the object similarity within the class is maximized and the object similarity between classes is minimized. The specific process is as follows (Fig 2):

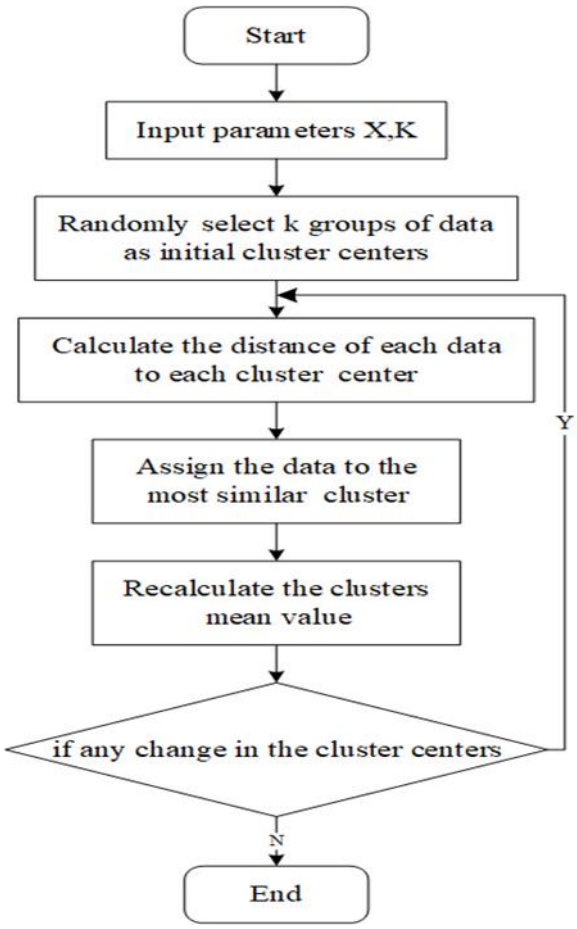

Fig. 2 Structure of K-means Clustering Algorithm

2.3.2. Convolutional neural network (CNN)

The basic unit of a Convolutional Neural Network (CNN) is a neuron, which has three basic elements: a set of links, a summation unit, and an excitation function. In addition, there is a threshold value (Fig 3), which is expressed as follows:

$$y^k = \varphi\{\sum_{i=1}^{p} w_{ij}x_j + b_k\} \quad (1)$$

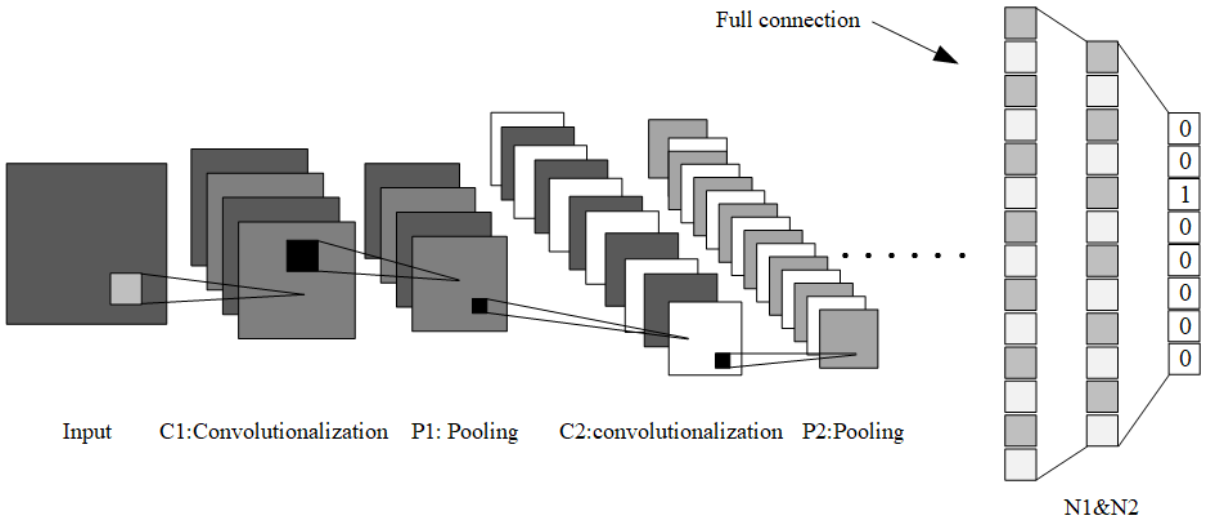

Fig. 3 Structure of Convolutional neural network [19]

Where h_i are the input signals, the connection weights between the input signals and the neuron k are denoted by the symbols b_k is the bias, and the activation function. The connection weights between the input signal and neuron k are denoted by the symbols $w_1, w_2, \cdots, w_k$ is the bias, the activation function is denoted by φ, and the output of the neuron is denoted by y_k, The activation function is denoted by φ , and the output of the neuron is denoted by y_k.

2.3.3. LSTM model

Its computational procedure uses the hidden state output h_{i-1} at the previous moment and the input x_t at the current moment and calculate the hidden state output h_t at the current moment t (Fig 4). The formula is as follows:

$$i_t = \sigma(w_i x_t + u_i h_{t-1} + b_i) \quad (2)$$

$$g_t = \tan h\,(w_g x_t + u_g h_{t-1} + b_g) \quad (3)$$

$$f_t = \sigma(w_f x_t + u_f h_{t-1} + b_f) \quad (4)$$

$$c_t = f_t^* c_{t-1} + i_t^* g_t \quad (5)$$

$$o_t = \sigma(w_o x_t + u_o h_{t-1} + b_o) \quad (6)$$

$$h_t = o_t^* \tan h\,(c_t) \quad (7)$$

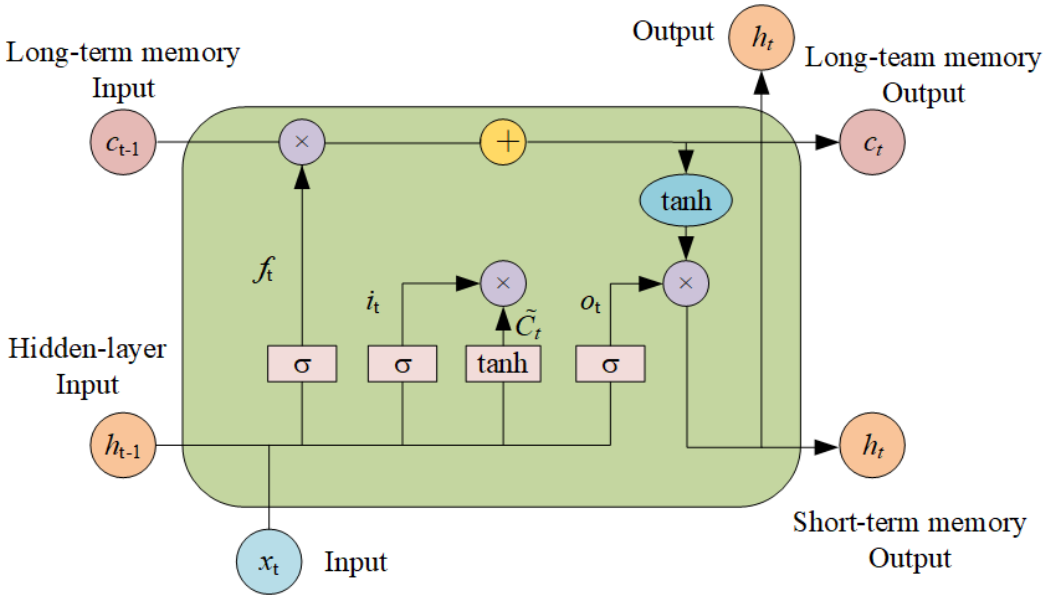

Fig. 4 Structure of LSTM Model [20]

2.3.4. Random forest network

This algorithm is an integrated learning algorithm that can be used for both classification and regression, which is performed on multiple decision trees in a mutually independent manner. When the results are obtained, the voting principle is used for the classification problem, and which category appears the most in the results of all the decision trees is the best one for classification. When the results are obtained, the voting principle is used for the classification problem, and the result of all the decision trees is the classification result of which category occurs the most.ID3, C4.5 and CART are the basic methods for constructing random forests. The G_{ini} coefficient of the sample is:

$$G_{ini}(S) = 1 - \sum_{i=1}^{m} {p_i}^2 \tag{8}$$

where p_i represents the probability of the category appearing in the sample set S. The G_{ini} value is used as a criterion for the effectiveness of random forest classification. The G_{ini} value of feature x_j at node m is:

$$G_{jm} = G_m - G_l - G_r \tag{9}$$

where G_m is the G_{ini} value of node m. G_l and G_r represent the G_{ini} values of the left and right child nodes of the current node respectively. x_j has nodes in the decision tree set M, x_j in the decision tree with G_{ini} value:

$$G_{jmi} = \sum_{m \in M} G_{jm} \tag{10}$$

There are n trees in the random forest, then the G_{ini} value of x_j:

$$G_j = \sum_{i=1}^{n} G_{jmi} \tag{11}$$

Finally, normalization can also be done. The smaller the G_{ini} value, the higher the purity of the dataset. The specific process is as follows (Fig 5):

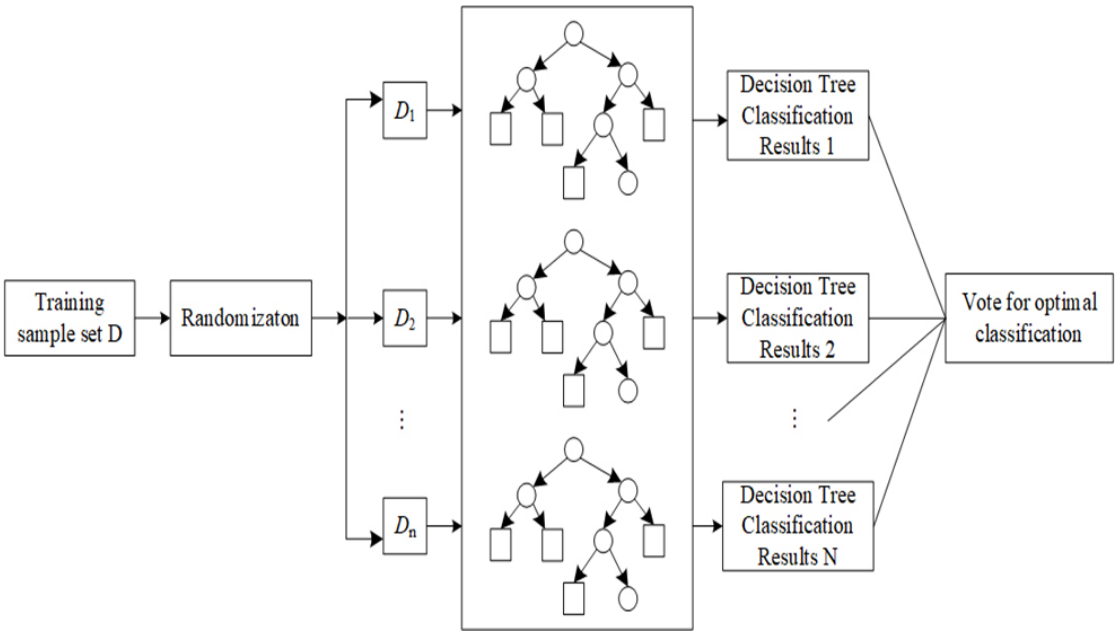

Fig. 5 Structure of Random Forest Network Algorithm

2.3.5. ConvLSTM model

The basic structure is as follows: based on the convolutional neural network, in the feature extraction Based on the convolutional neural network, in the feature extraction, according to the characteristics of the extracted objects, the convolutional neural network, the long and short-term memory neural network, and the random forest neural network are adopted respectively. Short-term memory neural network as the feature extractor according to the characteristics of the extracted object. The multi-layer full connectivity and classifier of the convolutional neural network are replaced by a random forest neural network. The structure is shown in Fig 6.

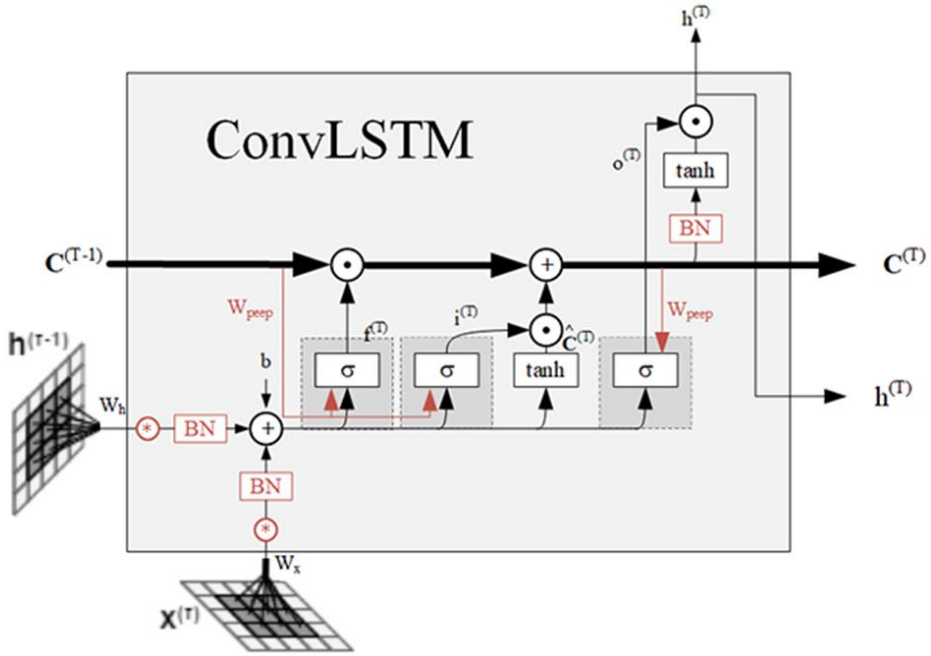

Fig. 6 Structure of ConvLSTM Model

3. Results and Discussion

3.1. Distribution of Traffic Accident Risk

In this paper, the severity of fatal accidents, accidents with injuries, and accidents without casualties are defined as a_1, a_2, and a_3, respectively, according to the severity types of traffic accidents, so the risk of traffic accidents in a region over a period of time can be expressed by equation (12):

$$S_{i,t} = a_1 D_{i,t}^1 + a_2 D_{i,t}^2 + a_3 D_{i,t}^3 \tag{12}$$

Where i denotes a region, t denotes a period of time, $D_{i,t}^1$, $D_{i,t}^2$, and $D_{i,t}^3$ denote the number of fatal accidents of persons, accidents of persons with injuries, and accidents with no injuries in region i in time t, respectively, and $S_{i,t}$ denotes the value of the risk of traffic accidents in region i in time t,which is obtained from the calculation. Based on the reality that the severity of life-threatening accidents is the highest, in this paper, we set a_1=3, a_2 =2 and a_3=1. Therefore, if three fatal accidents, one injury accident, and two no-injury accidents occur in a region during a period of time, its traffic accident risk value is $3 * 3 + 2 * 1 + 1 * 2 = 13$.

In addition, this paper combines the K-means and random forest approach to predict the value of traffic accident risk in the Manhattan area in 2023. However, in clustering algorithms, the similarity of data points is usually measured by their eigenvalues rather than considering the weights, therefore, in the actual implementation, we expand the number of traffic accidents on each coordinate point (x, y) in Manhattan region by the ratio of $3: 2: 1$ as described in the above article, and then clustering, so that the coordinate points (x, y) within a certain range together form the region of i , so as to achieve the prediction of the accident risk. For example, if there are three fatal accidents, one injury accident and two non-injury accidents at (x, y) , the number of traffic accidents at that point is considered to be $3 * 3 + 2 * 1 + 1 * 2 = 13$, which is equivalent to assigning weights to the types of traffic accidents.

The results of the experimental predictions are shown in Fig 7, which is a visualization of the total crash risk segment predictions for the Manhattan area in 2023 using the methodology described above. Its horizontal coordinates represent longitude, vertical coordinates represent latitude, the "X" pattern represents the position of the center of mass of each cluster, and the bar on the right side represents the measure of the risk value, with the darker red color on the graph representing the higher risk value of traffic accidents in the area.

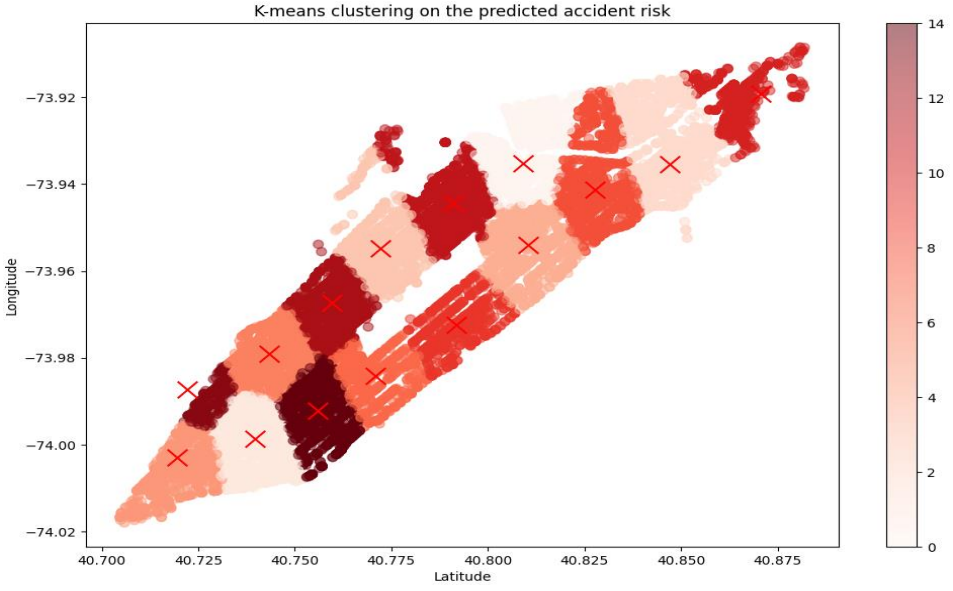

Fig. 7 K-means clustering on the predicted accident risk in Manhattan 2023

The 2023 real crash risk value data for the Manhattan area is shown in Fig 8, which has all the same labels as Fig 7.

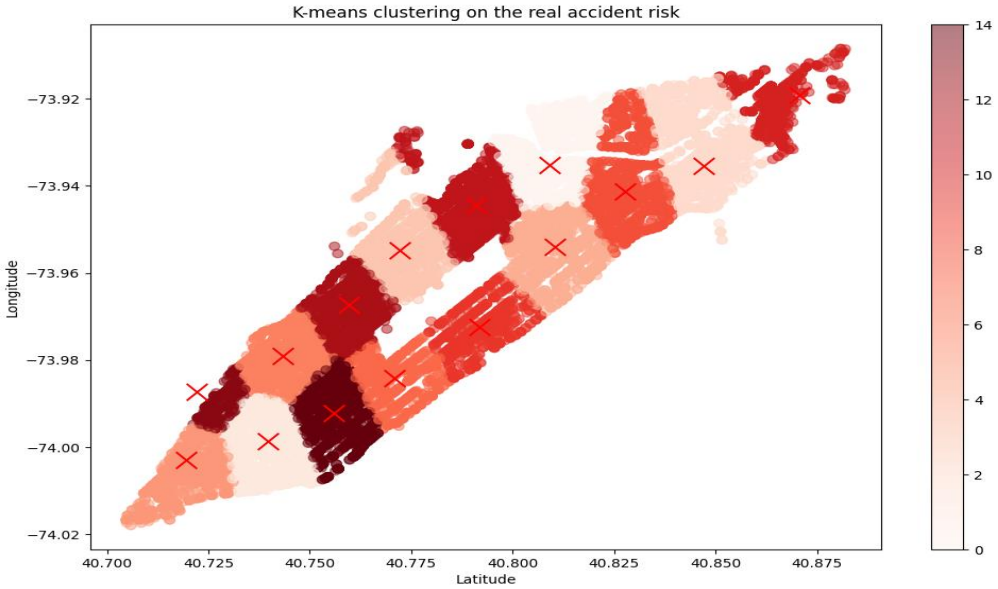

Fig. 8 K-means clustering on the real accident risk in Manhattan 2023

The mean square error (MSE) between the prediction results and the real data is about: 1.7045, which achieves a better result. In addition, it can be seen from the figure that the risk value of traffic accidents is higher in the southeastern and central regions of the Manhattan area, which coincides with the reality of the fact that Midtown Manhattan is the most important commercial and cultural center of Manhattan, and the southern end of the area is the financial center of New York City.

3.2. Forecast of Traffic Accidents

In this paper, the ConvLSTM model, a neural network structure that combines a convolutional neural network (CNN) and a long-short-term memory network (LSTM), was used to predict the number of accidents in Manhattan for the year 2023. Different weight values were assigned to accidents based on whether the accident occurred during the morning and evening peak hours. Among them, the morning and evening peak hours were 7:00 AM-9:00 AM and 5:00 PM-7:00 PM. Considering the training results and the fact that vehicles are more crowded in the morning and evening peaks, the speed of vehicles is slow, and it is not easy for traffic accidents to occur, the weight ratio between morning and evening peaks and ordinary hours is set to be 1:6.

Considering all these situations above, a hardware and software environment suitable for the operation of Pytorch framework Matlab, Keras has been built, chosen PyCharm development platform, used Python 3.8 development language and the traffic accident dataset of Manhattan, New York from 2016-2023, and processed accordingly, 80% of the data in the dataset as a training set (2016-2022), the remaining 20% of the data as a test set (2023). The significant factors inducing traffic accidents are obtained as follows: driving time, driving geographic location, etc.; the accident-prone area is southwest of Manhattan; the accident-prone period is the beginning and end of each month; and the accident-prone time is the morning and evening rush hours.

The predicted results of the experiment without the introduction of morning and evening peak weights are shown in Fig.9, where the horizontal coordinates represent the individual months in 2023, the vertical coordinates represent the number of accidents, the blue dash represents the true value of the number of accidents in Manhattan in 2023, and the red dash represents the predicted value of the number of accidents in Manhattan in 2023. Fig 10 has all the same labels as Fig 9, but it is a plot of the predicted results with the introduction of the morning and evening peak weights against the real data.

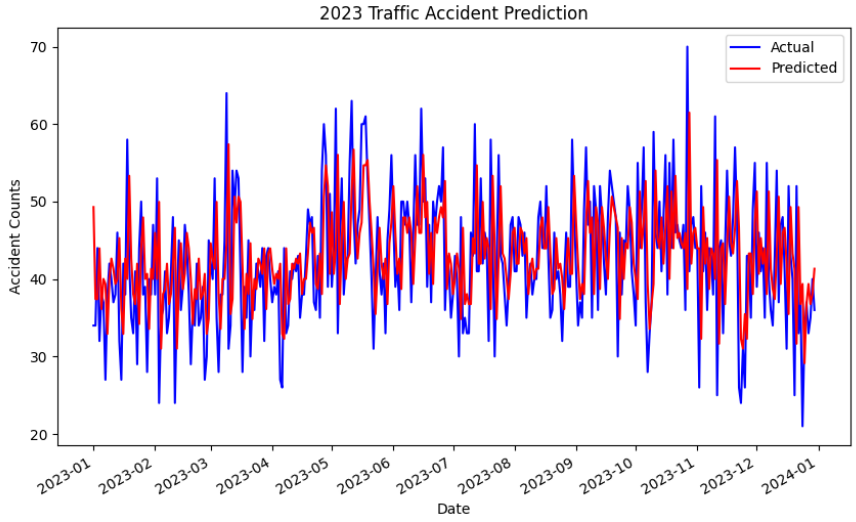

Fig. 9 Traffic accident prediction

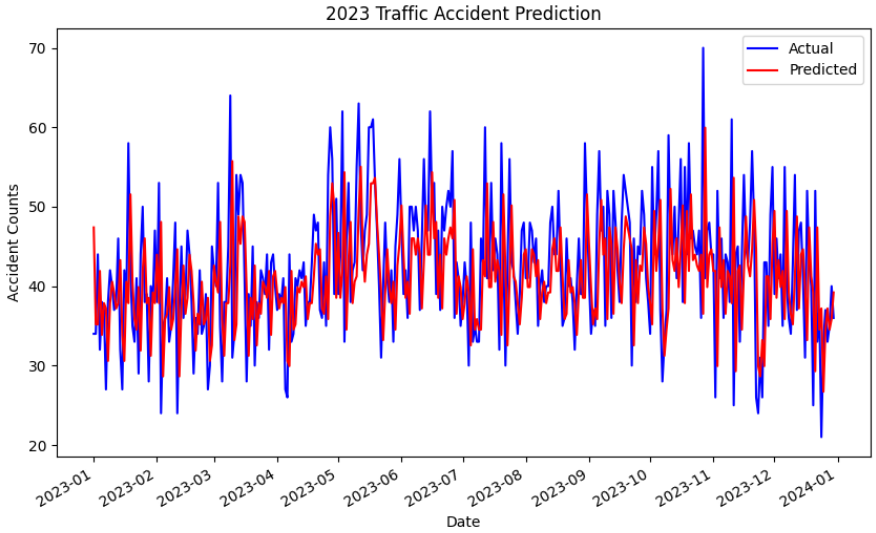

Fig. 10 Weighted Traffic accident prediction

Table 3. Comparison of the values of MSE and MAE

Method	MAE	MSE
LSTM	7.38	87.14
ConvLSTM	7.21	82.12
Weighted ConvLSTM	6.59	70.16

Despite the good fit, due to the large MSE values shown in the previous section, this paper has again visualised and compared the mean values of the predicted data and the real data (show in Fig 11 and 12), and it is clear that the data after the introduction of the weights are well fitted on the mean, so it is speculated that the large MSE values may be due to the large volume of the data (Table 3).

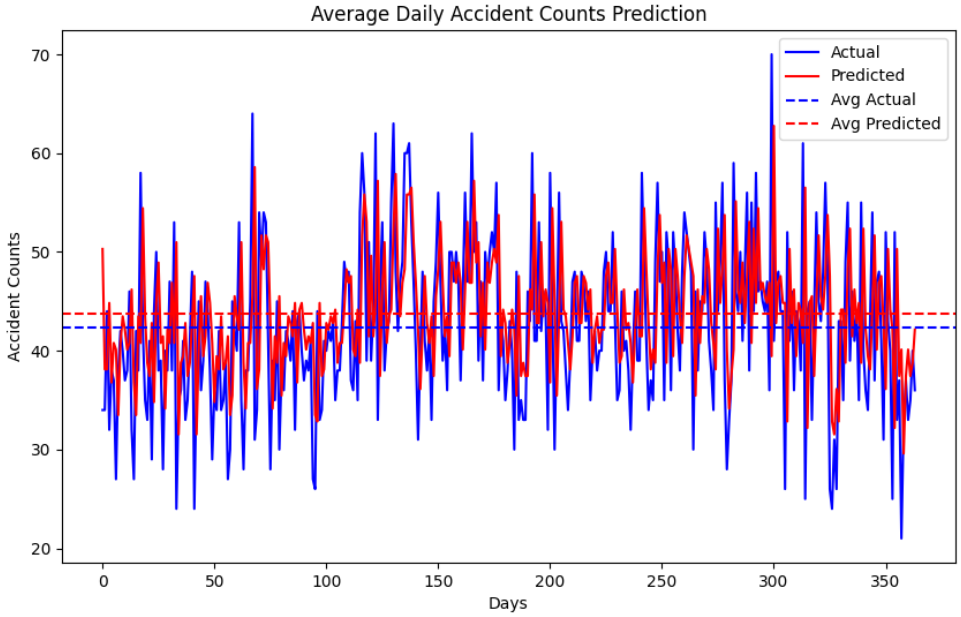

Fig. 11 Average of real data on the number of accidents in Manhattan in 2023

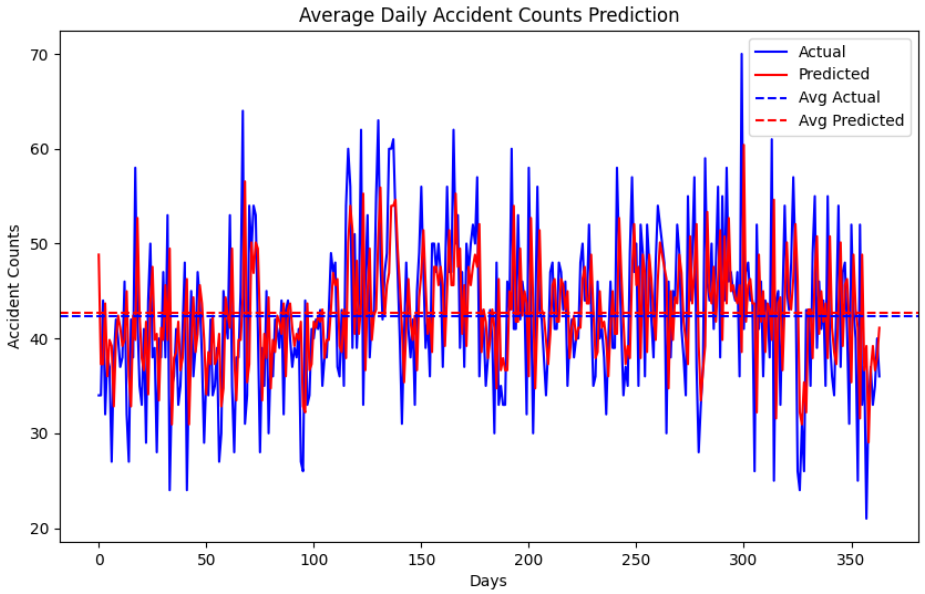

Fig. 12 Average of projected Manhattan accident volume data for 2023

Through the comparison of the line graphs and the comparison of the values of MSE and MAE (see Table 3), it can be clearly seen that the value of the error of the prediction results with the introduction of the early morning and late evening peaks weights is much smaller than that without the introduction of the early morning and late evening peaks weights.

4. Conclusion

Road traffic accident risk prediction data is an important reference basis for the scientific management of road traffic factors that induce road traffic accidents are various, while some of these factors have temporal characteristics, and others have spatial characteristics. Some of these factors have temporal characteristics, some have spatial characteristics, and some have temporal-spatial characteristics, so how to use deep learning algorithms to comprehensively extract the characteristics of these triggers and make predictions is a research topic in the field of road traffic safety.

In this paper, a convolutional long and short-term memory randomized forest neural network algorithm model was proposed. The factors inducing road traffic accidents. The factors, according to

their temporal and spatial characteristics, were extracted by using convolutional neural network and long and short-term memory neural network algorithms according to their temporal and spatial characteristics, which fully consider the characteristics of the factor variables and make the feature extraction more comprehensive and closer to reality. Prediction using the random forest neural network algorithm takes full advantage of its advantages of high accuracy, good performance on the test set, high noise immunity, and nonlinear classification models and other advantages.

Utilizing the 2016-2023 Traffic Crash Dataset for Manhattan, New York for further research study, the convolutional long and short-term memory random forest neural network algorithm model evaluation, and an in-depth comparison with other deep learning algorithm models in predicting road crash risk in an in-depth comparison with other deep learning algorithm models. In addition, in terms of inducing traffic accident variables, the selection of variables that are both temporal and spatial are explored in depth.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Yao Junfeng, et al. Review on machine learning-based traffic flow prediction methods. Journal of Traffic and Transportation Engineering, 2023, 23(3): 44-67.

[2] Wang Shunshun, et al. A Review of Road Traffic Accident Prediction Methods. American Journal of Management Science and Engineering, 2023, 8(3): 73-77.

[3] Yan Liping, et al. Deep prediction of urban traffic accident risk based on edge computing. Computer Simulation, 2022, 39(12): 226-229.

[4] Zhao Haitao, et al. Research on Traffic Accident Risk Prediction Algorithm of Edge Internet of Vehicles Based on Deep Learning. Journal of Electronics & Information Technology, 2020, 42(1): 50-57.

[5] Wang Yuan. Smart city traffic flow prediction based on SARIMA and LSTM model. Xiangtan University, 2020.

[6] Zhang Yankong, et al. A short-term risk prediction method for urban traffic accidents based on road network structure. Journal of Intelligent Systems, 2020, 15(4): 663-671.

[7] Guo Shengnan, et al. Deep Spatial-Temporal 3D Convolutional Neural Networks for Traffic Data Forecasting. IEEE Transactions on Intelligent Transportation Systems, 2019, 1-14.

[8] Ran Hongzhu, et al. A deep learning approach for highway accident risk prediction. Computer Technology and Development, 2023, 33(11): 189-195.

[9] Wang Qingrong, et al. Research on traffic accident risk prediction based on spatio-temporal graph convolutional network. Computer Engineering, 2022, 48(11): 22-29.

[10] Zhu Lei. Research and implementation of traffic accident risk prediction based on deep learning. Southwest Jiaotong University, 2019.

[11] Chen Chao. Research on urban traffic accident risk prediction algorithm based on deep learning. Xiamen University, 2019.

[12] Yu Zhiqing. A deep learning algorithm for traffic accident risk prediction. Computer Age, 2023, 8: 60-64.

[13] Wang Beibei. Research on Traffic Accident Risk Prediction Methods by Fusing Multi-source Spatio-temporal Data. Beijing Jiaotong University, 2021.

[14] Wei Yimeng. Research on traffic accident risk prediction method based on neural network. Lanzhou Jiaotong University, 2022.

[15] Guo Xiaoyi. Research on the fluctuation law and prediction of urban rail transit passenger flow under rainfall environment. Chang'an University, 2023.

[16] Huang Youzhi. Research on urban traffic accident risk prediction method based on deep learning. Zhejiang University of Technology, 2022.

[17] Shi, Wang. Convolutional LSTM Network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, 2015, 802-810.

[18] Chen Quanjun, et al. Learning deep representation from big and heterogeneous data for traffic accident inference. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, 338-344.

[19] Sherstinsky. Fundamentals of recurrent neural network and long short-term memory network. Physica D Nonlinear Phenomena, 2020, 404(8): 132306.

[20] Li Hao, et al. Survey of convolutional neural network. Journal of Computer Applications, 2016, 36(9): 2508-2515.