Research on Sustainability of Property Insurance Based on LSTM Model

Wenchao Zhang [1, *], Jiangwen An [2], Mingzhao Li [3], Zhiyu Cheng [1]

[1] School of Mathematics and Statistics, Hainan University, Haikou, China, 570228

[2] Institute of Tropical Agriculture and Forestry, Hainan University, Haikou, China, 570228

[3] School of Computer Science and Technology, Hainan University, Haikou, China, 570228

* Corresponding Author Email: zwc18370943196@163.com

Abstract. Losses from extreme weather events are increasing in recent years, and climate change will drive premiums higher. At present, the adjustment of insurance policies by insurance companies, as well as the protection of real estate siting and ancient buildings, are the problems that we need to solve now. To explore the circumstances under which insurance companies undergo insurance, this paper identified two severely affected areas and collected monthly insurance claims data from 2010 to 2023. Based on the collected data, an LSTM is established to derive the amount of compensation, and then determine whether to underwrite. The research centers on "how and where to choose a site and whether to build in a certain place", we determined 5 aspects of research in natural conditions, transportation conditions, infrastructure, surrounding environment and economic factors, and selected 17 secondary indicators; firstly, the paper quantified the secondary indicators into 3 parts of proximity, remoteness and Neutral factor 3 parts. Firstly, using EDA and feature engineering & insurance factor quantitative measurement to find out the required insurance factor F. Then using Kmeans cluster analysis to classify the 19 regions, and using Random Forest to find out the weights of the 17 indicators. Secondly, using TOPSIS comprehensive evaluation model to find out the scores of the 19 regions, and finally determining the highest-scoring the "leader" with the highest score is finally determined as the optimal region, For the research of developing building protection model, this paper identifies four aspects of research from basic information, economic contribution, historical contribution and architectural information. Finally, the Word to vector word embedding vector model is used to number the words after the text is transformed; then the LSTM is used to do the text binary classification to get the emotional percentage of each area, and the emotional percentage is used as a new column of indicators; the PCA is used to extract the variance contribution rate of each indicator; then the MinMax variance contribution rate is used to carry out the normalization and the sigmoid function for weighting; finally, use AHP to find out the score, and from the results, it can be seen that the regional conservation model of Yellowstone National Park is the most reasonable.

Keywords: LSTM model, EDA, TOPSIS, Kmeans cluster analysis.

1. Introduction

Extreme weather events brought on by climate change pose a serious threat to homeowners and insurance companies. In recent years, the world has suffered more than $1 trillion in losses due to "more than 10,000 extreme weather events". In 2022, the number of natural disaster youth claims was 115% higher than the average of the past 30 years. Damage from severe weather-related events is expected to continue to expand as a result of factors such as floods, hurricanes, cyclones, drought and wildfires [1-3].

Not only is property insurance becoming more expensive, it's also becoming harder to find as insurers change the way they write and what they write. Climate change is expected to drive premiums up by 30-60% by 2040. Weather-related events drive up the cost of property insurance premiums by different amounts in different parts of the world [4]. Moreover, the global gap in insurance coverage averages 57% and is widening, underscoring the industry's dilemma.

There is a crisis in the profitability of insurance companies and the affordability of homeowners. Insurers are changing how and where they write insurance to accommodate this change.

COMAP's Catastrophe Insurance Modeler (ICM) is interested in the sustainability of the property insurance industry. Climate change has led to a rise in the likelihood of more severe weather and natural disasters, so ICM wants to determine the best way to deploy property insurance to make the system resilient enough to cover the cost of future claims while ensuring the long-term health of the insurer [5]. If insurance companies refuse to write policies too often, they will go out of business for lack of customers. Conversely, if they underwrite policies that are too risky, they may pay out too much on claims. We explored the circumstances under which insurance companies should underwrite policies and when they should choose to take risks [6].

2. Long Short Term Memory (LSTM)

2.1. The Introduction to the LSTM model

LSTM neural network model is a special RNN (Recurrent Neural Network) model. In traditional RNN, the training algorithm uses LSTMTT, when the time is longer, the residuals that need to be returned will fall exponentially, resulting in slow updating of the network weights, which cannot reflect the effect of long-term memory of RNN, and a memory unit is needed to store the memory [6]. The difference between RNN and LSTM model is shown in Fig 1-2.

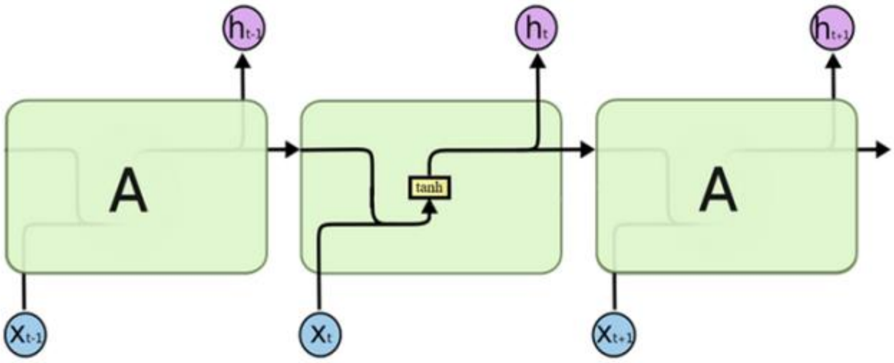

Figure 1. RNN neural network concept

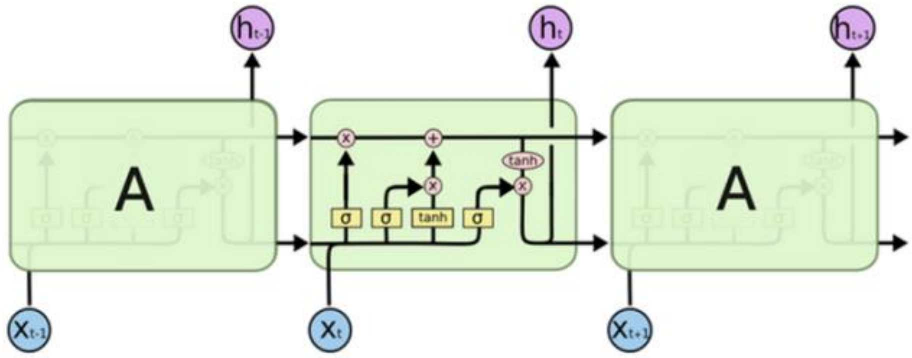

Figure 2. LSTM neural network concept

2.2. LSTM model establishment

The core of LSTM is the "cell state", which is actually the memory space that changes over time throughout the model. The conveyor belt itself does not control what information is memorized or not, but rather the control gates, which are described below Fig 3:

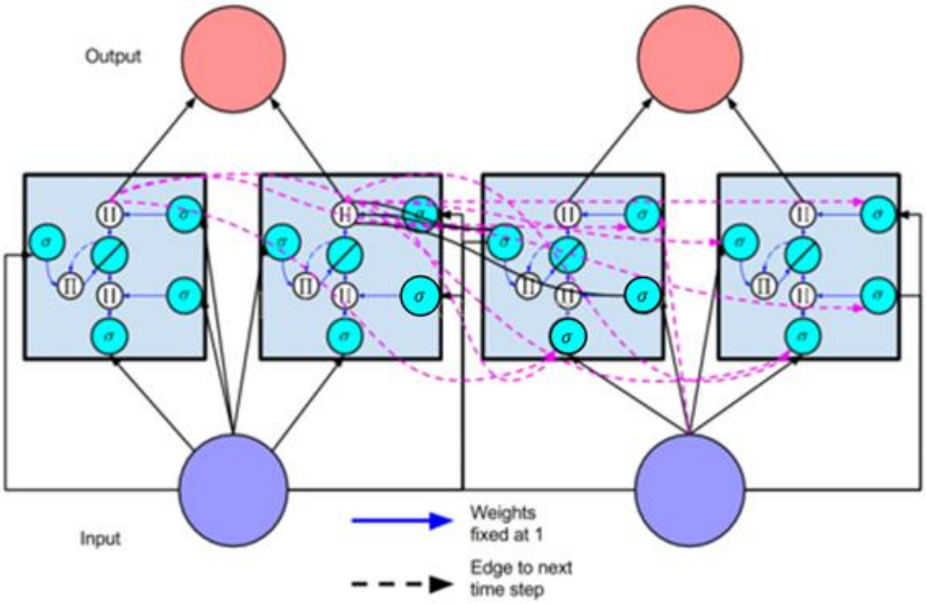

Figure 3. LSTM neural network modeling structure

The control gate consists of a sigmoid function followed by a dot product operation; the value of the sigmoid is [0, 1], when it is 0, no transmission; when it is 1, all transmissions are made. There are three control gates in LSTM: input gate, output gate, and memory gate. Now the following model is built based on the data:

(1) Forget-gate modeling: choosing to forget certain data information from the past.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

(2) Memory gate model: memorizing certain data information in the present.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

(3) Memory merger model: merging past and present memories.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{4}$$

(4) LSTM output model

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

(5) Increase the signal for cell state: for each memory gate and forget gate add.

$$f_t = \sigma\left(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f\right) \tag{7}$$

$$i_t = \sigma(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i) \tag{8}$$

$$o_t = \sigma(W_o \cdot [C_t, h_{t-1}, x_t] + b_o) \tag{9}$$

(6) Combining forget gate and input gate.

$$C_t = f_t * C_{t-1} + (1 - f_t) * \tilde{C}_t \quad (10)$$

(7) LSTM memory layer model
The memory layer iteration formula:

$$g^{(t)} = \phi\left(W_{gx}x^{(t)} + W_{ih}h^{(t-1)} + b_g\right) \quad (11)$$

$$i^{(t)} = \sigma\left(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + b_i\right) \quad (12)$$

$$f^{(t)} = \sigma\left(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + b_f\right) \quad (13)$$

$$o^{(t)} = \sigma\left(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + b_o\right) \quad (14)$$

$$s^{(t)} = g^{(t)} \odot i^{(i)} + s^{(t-1)} \odot f^{(t)} \quad (15)$$

$$h^{(t)} = s^{(t)} \odot o^{(t)} \quad (16)$$

The output layer iteration formula:

$$g^{(t)} = \phi\left(W_{gx}x^{(t)} + W_{ih}h^{(t-1)} + b_g\right) \quad (17)$$

$$i^{(t)} = \sigma\left(W_{ix}x^{(t)} + W_{ih}h^{(t-1)} + b_i\right) \quad (18)$$

$$f^{(t)} = \sigma\left(W_{fx}x^{(t)} + W_{fh}h^{(t-1)} + b_f\right) \quad (19)$$

$$o^{(t)} = \sigma\left(W_{ox}x^{(t)} + W_{oh}h^{(t-1)} + b_o\right) \quad (20)$$

$$s^{(t)} = g^{(t)} \odot i^{(i)} + s^{(t-1)} \odot f^{(t)} \quad (21)$$

$$h^{(t)} = s^{(t)} \odot o^{(t)} \quad (22)$$

2.3. LSTM model computation

The matrix of connected weights from I_m to H_m is denoted:

$$\omega_{mm} = (\omega_{11}, \cdots\cdots \omega_{1m}; \cdots\cdots; \omega_{m1}, \cdots\cdots \omega_{mm}) \quad (23)$$

The matrix of connection weights from H_m to O_m is denoted:

$$\delta_{mm} = (\delta_{11}, \cdots\cdots \delta_{1m}; \cdots\cdots; \delta_{m1}, \cdots\cdots \delta_{mm}) \quad (24)$$

Denote the input and output of each layer by *U* and *V*, respectively, with the output of the input layer equal to the input:

$$V_1^m(n) = I_m(n) \quad (25)$$

The inputs of the mth neuron of the hidden layer get weighted sums:

$$U_H^m(n) = \sum_{m-1}^{M} \omega_{mm}(n) V_1^m(n) \quad (26)$$

M in Equation (26) is the input vector length. The implicit layer transfer function must be differentiable, and this method uses a nonlinear action *Sigmoid* function:

$$Logsig(x) = \frac{1}{1+e^{-x}} \tag{27}$$

Then the output of the mth neuron of the hidden layer:

$$V_H^m(n) = Logsig[U_H^m(n)] \tag{28}$$

According to the data in this question, the input layer transfer function is a linear *pureline* function, then the mth neuron output:

$$V_O^m(n) = pureline[U_O^m(n)] \tag{29}$$

After *n* iterations, the total error of the network is denoted as:

$$E(n) = \frac{1}{2}\sum_{m-1}^{M}[T_m(n) - V_O^m(n)]^2 \tag{30}$$

The error histogram is trained by the neural network is shown in Fig 4-5.

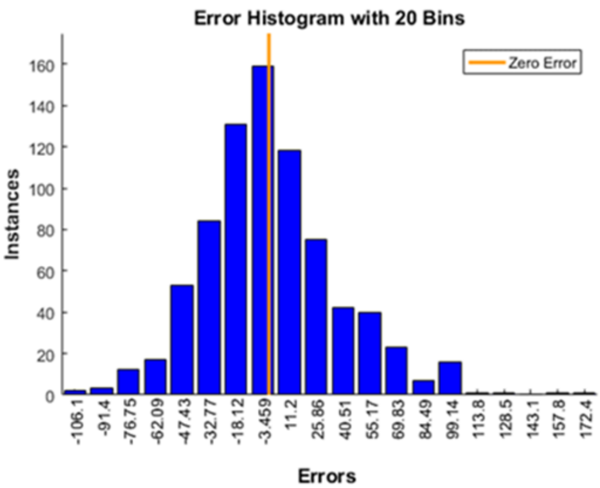

Figure 4. Normal distribution chart

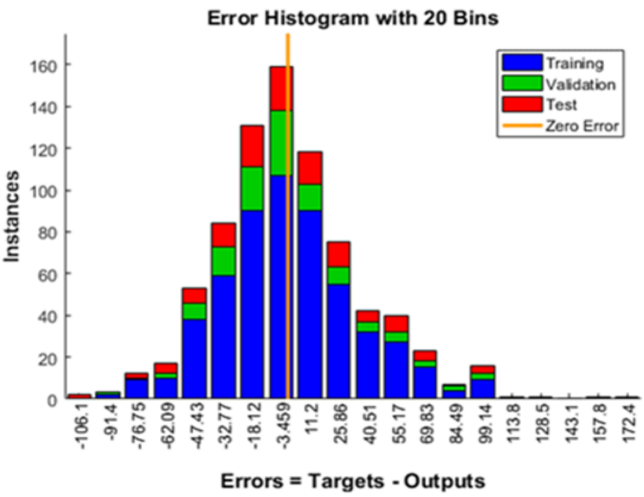

Figure 5. Error histogram data

Smoothness test was performed by constructing LSTM in fold plot and ADF unit root test. The unit root test is performed for smoothness testing. It is assumed that there is a unit root in the data, i.e., it is not smooth; and the original hypothesis is strictly rejected when the level of significance is 1%, when it is 5%, and 10% by analogy [7]. The level of significance is evaluated by looking at the p-value and the magnitude of the significance level a [8-9]. The smaller the p-value, the smaller the level of significance, the original hypothesis is rejected and the series is considered to be smooth; and the larger the level is, the hypothesis cannot be rejected and the series is considered to be unsteady. The measurement statistic p-value = 0.03, which is less than 0.05. Its training status is shown in Fig 6.

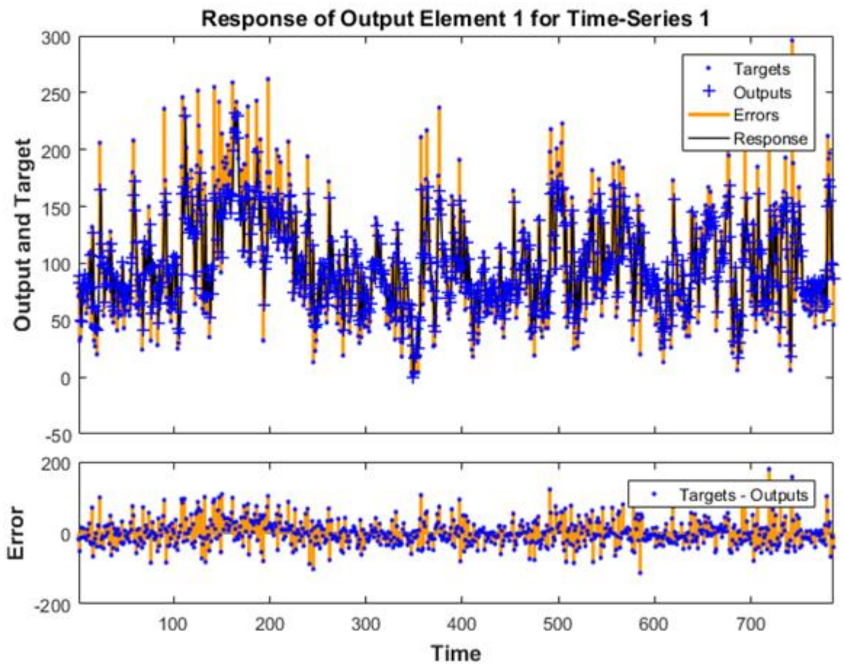

Figure 6. LSTM neural network training data error effect plot

A moving smoothing window test was performed on the above smoothed series to explore the correlation between the variables, which is shown in Fig 7. The total correlation value for the ozone concentration values reached 84%. This dataset shows a weak correlation in the pre-training period, and in the post-training period, the correlation tends to increase, indicating a trend that the model is tending to become more accurate [10].

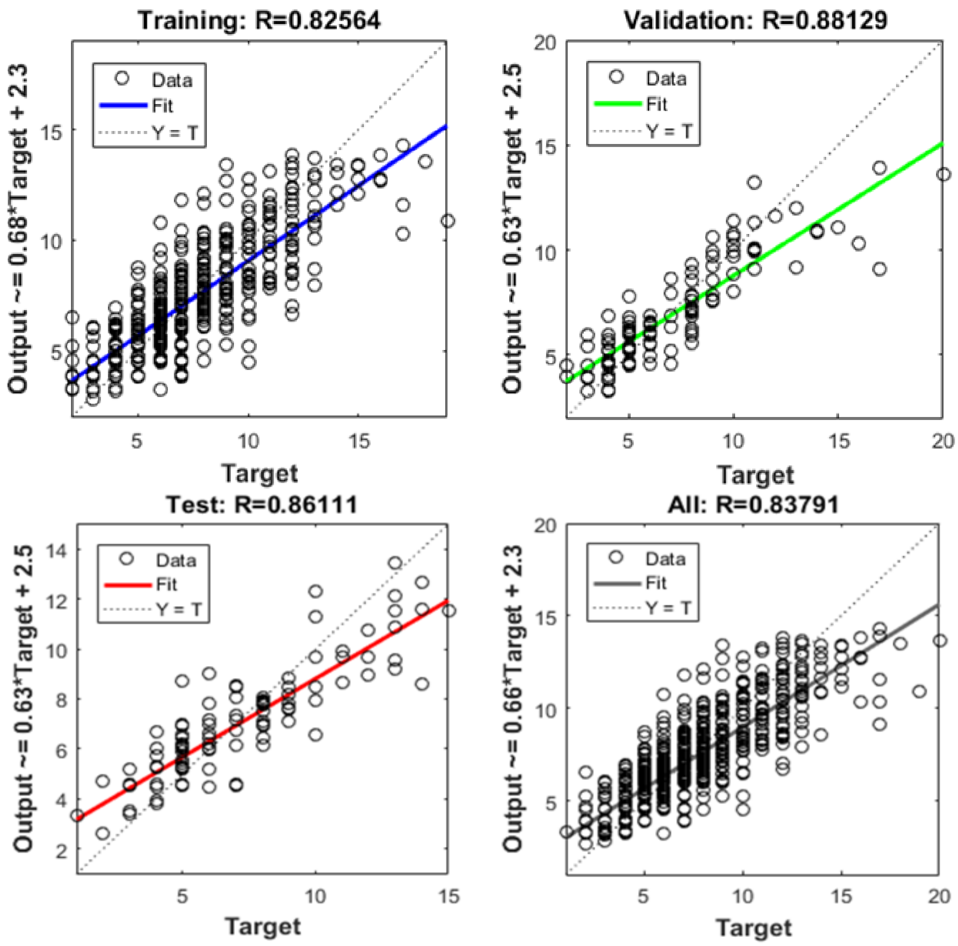

Figure 7. Regression superiority analysis of LSTM neural network training data

2.4. LSTM model solving

In this paper, the collected insurance payout amounts are brought into the LSTM model and the results are shown in Table 1 and Fig 8 below.

Table 1. Table of LSTM model results

Model Metrics	Recall (%)	Accuracy (%)	F1 value	AUC
LSTM	84.42	84.67	0.8459	0.92

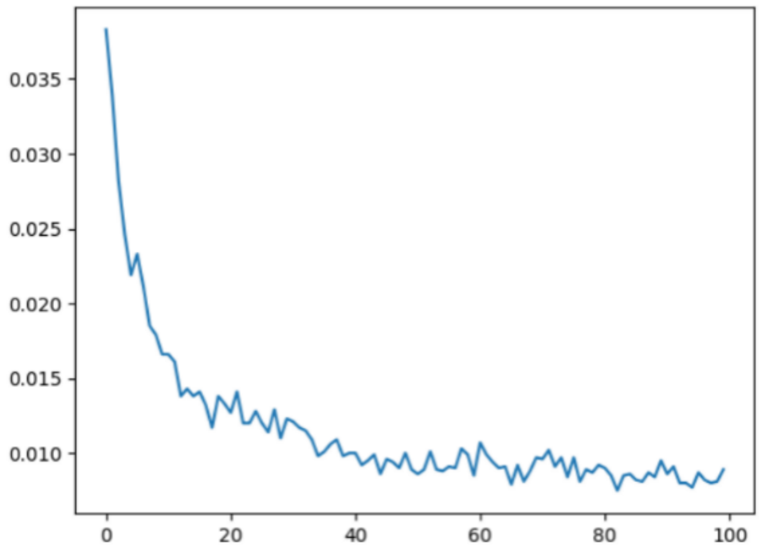

Figure 8. LSTM neural network loss plot

3. Conclusion

The entire article collected monthly insurance claim data from two severely affected areas from 2010 to 2023, and established an LSTM model based on the collected data. The research revolves around "how to choose a location, where to choose a location, and whether to build a house in a certain place". Through research on five aspects, including natural conditions, transportation conditions, infrastructure, surrounding environment, and economic factors, 17 secondary indicators were selected and quantitatively measured using EDA and characteristic engineering insurance factors to identify the required insurance factor F; Then use Kmeans clustering analysis to classify 19 regions, and use random forest to identify the weights of 17 indicators; Finally, use the TOPSIS comprehensive evaluation model to identify the scores of 19 regions. The conclusion has been drawn that Yellowstone National Park is the best area. In addition, during the training of LSTM neural networks, we found that the model tends to become more accurate.

References

[1] Huang Siying. Investment portfolio construction strategy based on LSTM and entropy TOPSIS method [J]. Modern Business, 2023 (03): 117 - 120.

[2] Xu G, Xiaoming D, Tingting H, et al. Analysis of influencing factors on excellent teachers' professional growth based on DB-Kmeans method [J]. EURASIP Journal on Advances in Signal Processing, 2022, 2022 (1).

[3] Susmaga R, Szczęch I, Brzezinski D. Towards explainable TOPSIS: Visual insights into the effects of weights and aggregations on rankings [J]. Applied Soft Computing, 2024, 153111279-.

[4] Jan P, Maciej M, Sabina L. Graph Tar: applying word2vec and graph neural networks to miRNA target prediction [J]. BMC Bioinformatics, 2023, 24 (1): 436 - 436.

[5] David W, Balaji R, Edith Z. Long-Lead Forecasting of Runoff Season Flows in the Colorado River Basin Using a Random Forest Approach [J]. Journal of Water Resources Planning and Management, 2024, 150 (4).

[6] Ozegin O K, Ilugbo O S, Adebo B. Spatial evaluation of groundwater vulnerability using the DRASTIC-L model with the analytic hierarchy process (AHP) and GIS approaches in Edo State, Nigeria [J]. Physics and Chemistry of the Earth, 2024, 134103562-.

[7] Ge Yaming, Qiu Chenguang, Xie Lirong, et al. Multi energy coupled power load prediction based on K-means clustering and LSTM model [J/OL]. Modern Power: 1-8 [2024-02-19].

[8] Fang Tao Research on Near Ground Ozone Estimation and Prediction Based on Neural Networks [D]. Shanghai Normal University, 2020.

[9] Zhang Meng Research on Short term Traffic Volume Prediction Based on LSTM Neural Network [D]. Shijiazhuang Railway University, 2023.

[10] Tang Zhenpeng, Wu Junchuan, Zhang Tingting, et al. Research on Systematic Risk Warning in China's Insurance Industry Based on EEMD-LSTM [J]. Management Review, 2022, 34 (09): 27 - 34.