Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

A Study on the Prospects of Regional Artificial Intelligence Development Based on Carbon Emission and Development Indicators

Qiaochu Li [1, *, †], Xinyu Zhuang [2, †]

[1] Sydney Smart Technology College, Northeastern University at Qinhuangdao, Qinhuangdao, China, 066003

[2] School of computer and communication engineering, Northeastern University at Qinhuangdao, Qinhuangdao, China, 066003

* Corresponding Author Email: liweiwei_wcx@163.com

† These authors contributed equally.

Abstract. Artificial Intelligence today demonstrates an astonishing productivity, and its development has become one of the directions for many countries. However, while developing, it is also necessary to consider the adverse impact of AI development on carbon emissions and to seek environments and regions suitable for AI development. To address this issue, this paper innovatively proposes the concept of "fertility" to describe the AI development potential of a region, and fully considers the adverse impact of AI development on carbon emissions. Based on the carbon emission and development data of various provinces in China in recent years, the "fertility" is modeled through PCA (Principal Component Analysis) and GB-DT model, and combined with LSTM for predicting the future AI development potential, thus deriving the future AI development potential of various provinces in China.

Keywords: New Quality Productive Forces; PCA (Principal Component Analysis); GB-DT; LSTM; Artificial Intelligence Development Potential.

1. Introduction

Artificial Intelligence (AI) is bringing more and more productive forces to social development [1], and society increasingly needs the development of AI. The integration of AI with various industries will bring immeasurable progress [2]. However, due to the huge demand for AI development [3], Xuefei once revealed the dual impact of AI development [4], especially in terms of electricity consumption [5]. Moreover, it has also been pointed out that on the key path of China's economic structure transformation in the new era, the surge of AI will not only produce a scale effect on economic growth, drive the development of information technology industry innovation clusters, but also produce industry spillover effects, bringing dividends to related industries, and thus increasing the total carbon emissions of the region [6]. Choosing the right development address can make AI development dynamic while not adversely affecting the environment.

It is worth mentioning that the development of artificial intelligence does not always have a negative impact on carbon emissions. When the development of artificial intelligence is advanced to a certain extent, it can also be applied to related technologies such as power generation [7], power saving [8], and waste gas treatment that can reduce carbon emissions. However, at present, artificial intelligence is still in the development stage, and we should always consider the carbon emissions caused by the development of artificial intelligence in its development plan.

This paper proposes the concept of "fertility" to describe the prospects of developing artificial intelligence technology in a certain province and to quantify it. At the same time, based on the research of Ma Guangwei [9], a more accurate artificial intelligence level evaluation system is established by integrating considerations in policy environment and industrial structure. Then, on the premise of considering carbon emission volume, the correlation analysis between "fertility" and "local

artificial intelligence development level" is made to prove the reference value of "fertility". Finally, according to the "fertility" index data of each province, the provinces most suitable for the development of artificial intelligence today are obtained.

2. Evaluation Model Combined with Carbon Emission Indicators

2.1. Determination of Fertility Index Related Variables

This paper mainly considers variables related to the fertility index from the following perspectives:

(1) Establish the fertility index, which needs to consider the feasibility of product structure transformation. This indicator is called the product structure upgrade coefficient, $\Delta q_{m,n}$ represents the proportion of the output value of industry or sector j in the total output value from period m to period n.

$$\Delta q_{m.n} = \frac{1}{2}\sum_{\mathrm{j}} |q_j^n - q_j^m| \times 100\% \tag{1}$$

(2) Establishing the fertility index requires considering the professional foundation of local technology companies, combined with data on the proportion of high-tech industries. In terms of cutting-edge technology collaboration, some provinces that are technologically advanced often have faster adaptability in other technical fields. In addition, the growth rate of scientific and technological paper data is calculated as a supplementary explanation.

(3) This paper innovatively introduces the total carbon emission of the province as the first environmental factor variable to consider. This paper combines the carbon emission volume from different sources in various provinces with their AI development level for principal component analysis, and the results are shown in Table 1 and Figure 1.

Taking the results of principal component analysis as weights, the carbon emissions from various sources are weighted and added to represent the emission bearing capacity of the region.

$$CA = \omega_1 \times ca_1 + \omega_2 \times ca_2 + \cdots + \omega_n \times ca_n \tag{2}$$

Where CA represents the emission bearing capacity of the region, ω_n represents the weight derived from the principal component analysis of different source carbon emissions, and ca_n represents the carbon emissions from different sources. In the following research, CA will serve as one of the important indicators describing the "fertility" of the region.

Table 1. Principal Component Analysis Structure of Different Sources of Carbon Emission

Sources of carbon emissions	Factor load factor	Commonality (common factor variance)
Raw coal	0.257	0.066
Coke	0.605	0.366
Petrol	0.831	0.69
Paraffin	-0.309	0.095
Diesel oil	0.847	0.718
Fuel oil	0.237	0.056

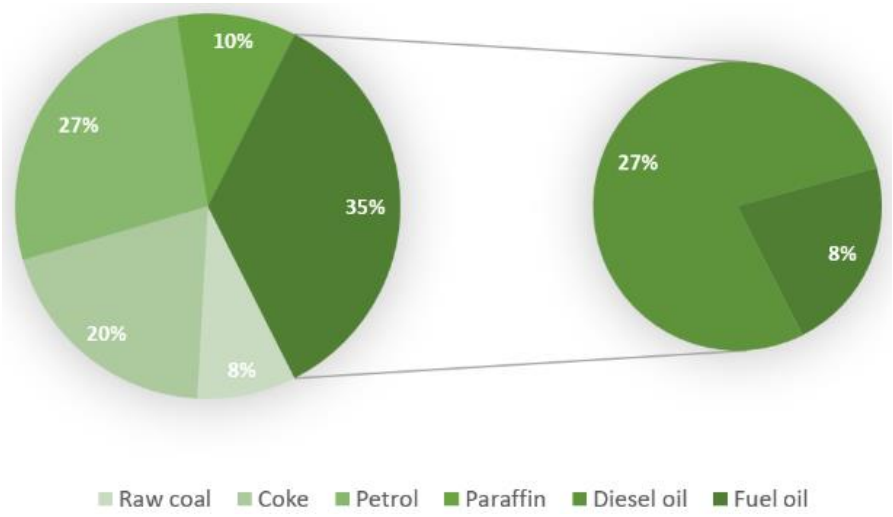

Figure 1. Principal Component Analysis of Carbon Emission

(4) Whether the technology can be implemented in specific provinces in China depends on the local policy environment and development direction. Therefore, the following variables are established to represent the degree of government support for AI development:

$$a_{9(x)} = \begin{cases} \frac{a_{1(x)} - a_{1(x-1)}}{\sum_{i=1}^{279}(a_{1(i)} - a_{1(i-1)}}, x \ mod \ 9 \neq 0, \\ 0, x \ mod \ 9 = 0 \end{cases} \tag{3}$$

Where $a_{9(x)}$ represents the value of the government's support for AI development data in the data set in the order of x, and l_x represents the value of the government's new product development fund data in the data set in the order of x. The growth rate of urban scientific and technological papers over the years:

$$a_{2(x)} = \begin{cases} \frac{l_x - l_{x-1}}{\sum_{i-1}^{279}(l_i - l_{i-1})}, x \ mod \ 9 \neq 0, \\ 0, \ x \ mod \ 9 = 0. \end{cases} \tag{4}$$

Where $a_{2(x)}$ represents the value of the city's scientific and technological paper growth rate in the data set in the order of x, and l_x represents the value of the city's scientific and technological paper release quantity in the data set in the order of x.

2.2. Eliminating the Impact of Original Data

Before officially modeling, it is necessary to conduct validity testing on the processed data. The selection of scale validity refers to the artificial intelligence industry evaluation index system established by Yang Mengcheng [10] in China, and combines the current key carbon emission index to make a reasonable weight distribution. On this premise, KMO test and Bartlett test are used to ensure that the measurement dimensions of the variables are strongly related to the problems we need to study.

Table 2. KMO Test and Bartlett Test Results

KMO value		0.616
Bartlett test	Approximate chi-square	2140.42
	df	45
	P	0.002***

The KMO test results are shown in Table 2, where the KMO value is 0.616. At the same time, the Bartlett's sphericity test shows a significant P-value of 0.002***, which is significant at the level, rejecting the null hypothesis, indicating that there is a correlation between the variables, and the principal component analysis is valid. Subsequently, the eigenvalues are solved, representing the original data as an m*n matrix, where m is the number of variables, and n is the number of original data points. The mean of the original data is used as the eigenvalue decomposition to obtain multiple eigenvectors and their corresponding eigenvalues.

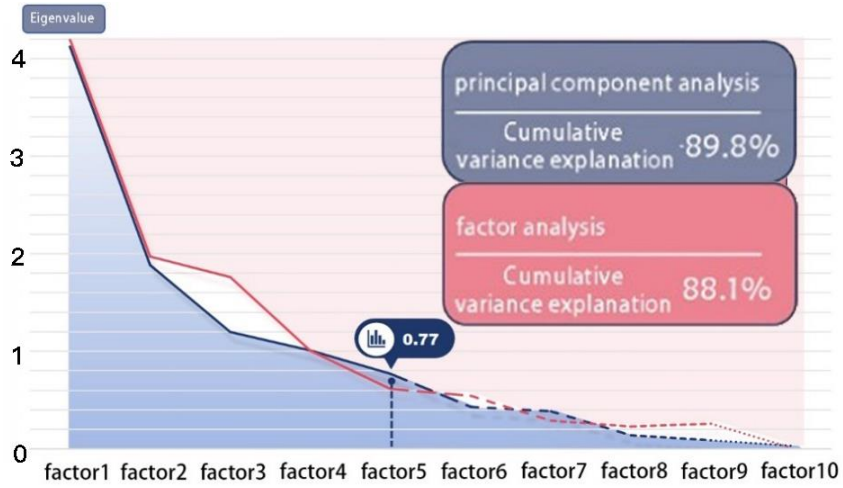

Figure 2. Principal Component Analysis Comparison Scree Plot

According to the degree of data variation explained by each principal component, the two models are different in the slope of the characteristic value decline. By looking for the "slope tends to be flat" mutation point in Figure 2, the number of principal components can be determined to be 5.

2.3. Model Solution

Adjust the number of principal components to 5, and re-perform the principal component analysis to analyze the importance of hidden variables in each principal component, and combine specific business for the hidden variable analysis of each factor. The heat map of the load matrix derived from the fertility index of each component is shown in Figure 3.

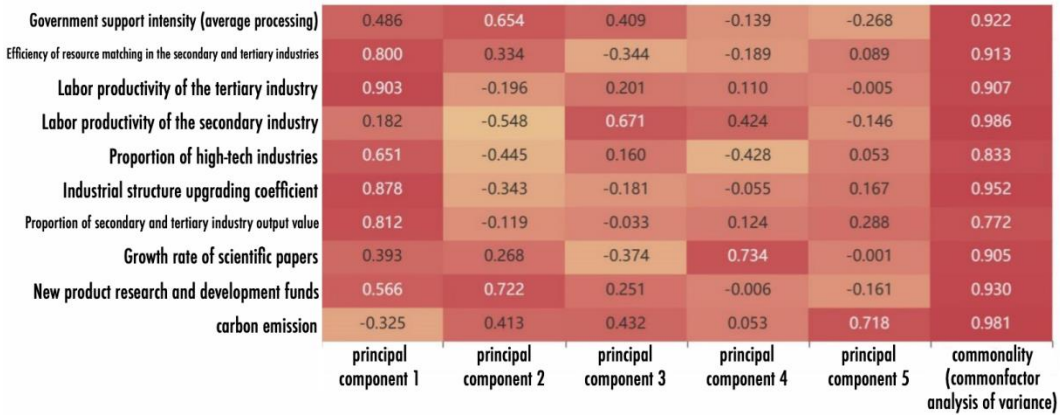

Figure 3. Principal Component Analysis Correlation Heat Map Analysis

Based on the dimensionality reduction of variables, the artificial intelligence development potential evaluation system considering carbon emission factors established in this paper is simplified to 5 principal components and has a high degree of commonality. After the evaluation scoring system of SPSS26.0, the development potential scores of various provinces from 2014 to 2022 were organized and collected, obtaining the changes in "fertility" of various provinces in China over nine years.

3. Prediction Model of Artificial Intelligence Technology Level and Fertility Correlation

3.1. Selection of Ensemble Algorithm Starting from Decision Tree

The decision tree may be simple and easy to perform in the initial stage of data analysis and is easy to interpret to explore the relationship between multiple factors and the target variable (i.e., AI development level). To further ensure the accuracy and reliability of the results, we first introduced two ensemble algorithms based on the decision tree, including random forest and Cat Boost (Categorical Boosting). The performance of these two algorithms during training is shown in Figures 4, 5, 6, and 7:

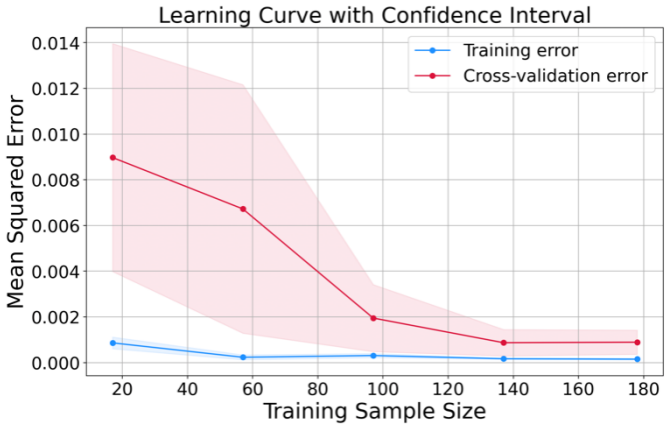

Figure 4. Random Forest Learning Curve

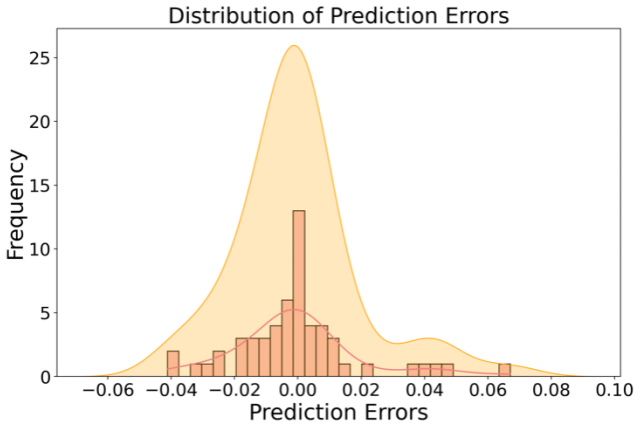

Figure 5. Random Forest Error Chart

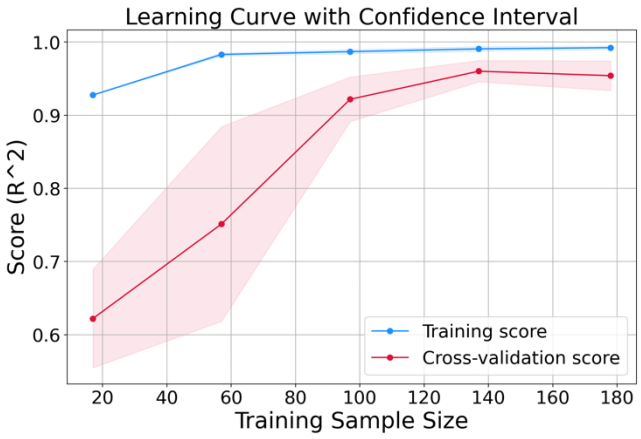

Figure 6. Cat Boost Learning Curve

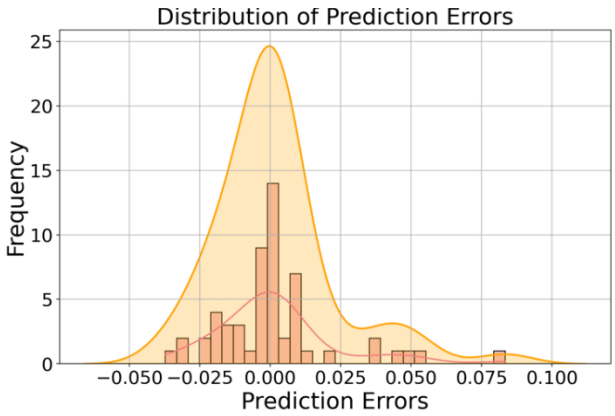

Figure 7. Cat Boost Error Chart

The random forest model and Cat Boost share the underlying logic of the decision tree and apply different algorithms, each with its own advantages and disadvantages. Using Mean Squared Error, MSE, as an indicator, the accuracy of the prediction is evaluated to roughly determine which type of algorithm is more suitable for our program prediction. The formula for solving MSE is as follows:

$$MSE = \frac{1}{2}\sum_{i=1}^{n}((y_i - \hat{y}_i)^2 \tag{5}$$

Where n is the total number of samples, y_i is the actual value of the i-th observation, and $\hat{y}_i$ is the predicted value. The smaller the MSE value of the model, the more accurate the prediction result. By calculation, it can be determined which of the random forest model and Cat Boost is more suitable for our program prediction. The results are shown in Table 3.

Table 3. Random Forest and Cat Boost

The name of the model	MSE
Random forest	0.000660602
Cat Boost	0.000480439

From the table, it can be concluded that the data involved in this paper is more suitable for prediction using the Cat Boost metho.

3.2. Precise Processing from Gradient Boosting Approximation

Similar to or improved upon the basis of Cat Boost are the XGBoost model and the GBDT model. The XGBoost model adds regularization terms to the loss function on the basis of Cat Boost:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\sum_{j-1}^{T}\omega_j^2 \tag{6}$$

Where T is the number of leaf nodes in the tree, ω_j is the weight of the leaf node, and γ and λ are regularization parameters. This can greatly prevent model overfitting. Although the GBDT model is not as advanced as XGBoost, its sensitivity to "outliers" in the data far exceeds the above models. The three models each have their own characteristics, and to obtain the best prediction results, these two models are trained, and their predictive capabilities are represented by their corresponding MSE values, as shown in Table 4.

Table 4. Prediction Validation of Random Forest, GBDT, and Cat Boost

The name of the model	MSE
XG Boost	0.000609232
GBDT	0.000336472
Cat Boost	0.000480439

At the same time, during the training process, the comparison charts between the predicted values and actual values, as well as the learning curves of XGBoost and GBDT models were also obtained. As shown in Figures 8, 9, 10, and 11.

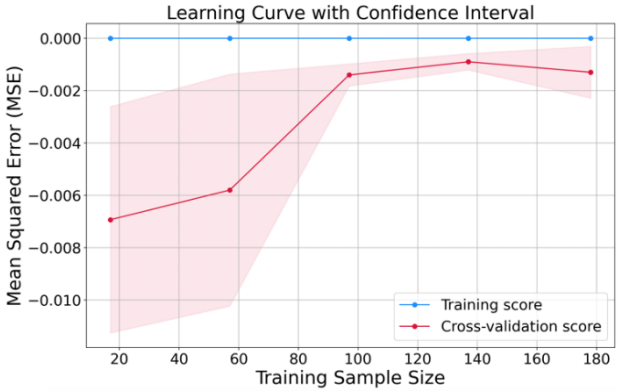

Figure 8. XG Boost Learning Curve

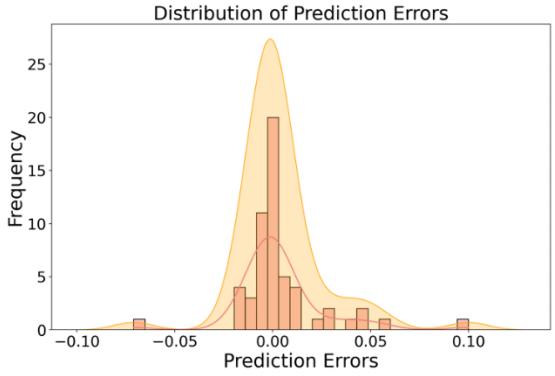

Figure 9. XG Boost Error Char

Figure 10. GBDT Learning Curve

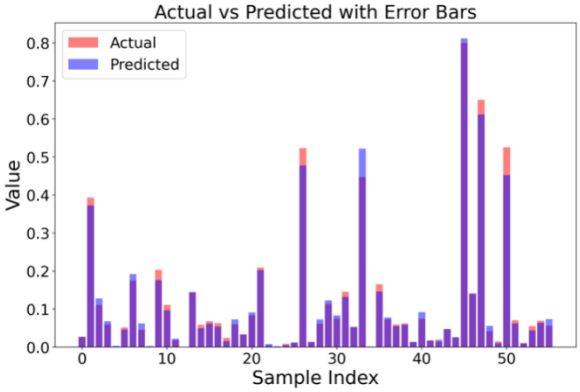

Figure 11. GBDT Difference Comparison Chart

3.3. Correlation of Prediction Results and "Fertility"

The prediction results of the GBDT model for the AI development level of various cities in each year are compared with the "fertility" index of each province in each year for Kendall's consistency test. The calculation formula for the rank variance is as follows:

$$s_j^2 = \frac{\sum_{i=1}^{n} r_{ij}^2 - \frac{n(n+1)^2}{4}}{\frac{n(n-1)}{2}} \tag{7}$$

Where γ_{ij} represents the rank given by the i-th evaluator to the j-th object. Subsequently, the Kendall coefficient is used to quantify the level of consistency, and its formula is as follows:

$$W = \frac{12}{k^2(n^3-n)} \sum_{j=1}^{k} s_j^2 - 3(n-1) \tag{8}$$

Where k is the number of evaluators, n is the number of objects to be sorted, and S_j^2 is the rank variance of the *j*-th object. The results of Kendall analysis are shown in Table 5.

Table 5. Kendall's W Analysis Results

Kendall's W analysis results					
Name	Rank mean	median	Kendall's W coefficient	X^2	P
fertility	1.99	0.356	0.96	96.04	0.000***
Amount of change in development level	1.01	0.006			

From Table 5, it is concluded that the "fertility" index of the city can well represent the AI development potential of the region.

4. Adjustment of "Fertility" Status Combined with Time Series

4.1. Neural Network Model Combined with Time Series

Next, the "fertility" index of various provinces in China in the future is predicted to combine the previous research results to determine the future fertility level of various provinces in China. This paper aims to achieve the prediction by training the LSTM model.

LSTM (Long Short-Term Memory) model is a special type of recurrent neural network (RNN) that performs well in processing and predicting time series data. The model introduces three gating structures (input gate, forget gate, output gate) to control the flow of information.

The forget gate is used to decide which information should be forgotten from the cell state, using the sigmoid activation function to output a value between 0 and 1, indicating the degree of forgetting of each cell state

$$f_t = \sigma\left(W_f \times [h_{t-1}, x_t] + b_f\right) \tag{9}$$

Where f_t is the output of the forget gate at time step *t*; h_{t-1} is the hidden state at the previous time step; x_i is the input at the current time step; W_f and b_f are the weight matrix and bias vector of the forget gate; σ is the sigmoid activation function. In addition, the input gate decides which new information should be stored in the cell state.

$$i_t = \sigma(W_i \times [h_{t-1}, x_i] + b_i) \tag{10}$$

$$\widetilde{C_t} = \tanh(W_c \times [h_{t-1}, x_i] + b_c) \tag{11}$$

Where i_t is the output of the input gate at time step *t*; $\widehat{C_t}$ is the candidate cell state. W_i, W_c and b_i, b_c are the weight matrix and bias vector of the input gate and the candidate cell state. The output gate determines the value of the hidden state, which is a function based on the current cell state.

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{12}$$

$$h_t = o_t \times \tanh(C_t) \tag{13}$$

Where o_t is the cell state at the current time step; W_o and b_o is the hidden state at the previous time step.

4.2. Obtaining the Best Results with Hyperparameter Search and Early Stopping Callback

When training the LSTM model, to ensure the accuracy and rationality of the prediction results as much as possible, the most suitable LSTM model parameters for the province (the optimal parameters for the province) will be found first, and then the future data of the province will be predicted. The optimal parameters for the LSTM model for each province are sought through Random Search parameter optimization.

First, the parameter space of the LSTM model is defined according to the structure of the model, which is the set of all possible parameter combinations. It can be represented as:

$$\Omega = \{(p_1, p_2, \cdots, p_n) | p_i \in searchSpace(p_i)\} \tag{14}$$

Where p_i is the i-th hyperparameter; $searchSpace(p_i)$ is the search range of the hyperparameter. Then, random sampling is performed in the parameter space to draw a random parameter combination in the space.

$$\omega = (p_1', p_2', \cdots, p_n') \tag{15}$$

After sufficient random sampling results are tested, the best set of parameters with the highest accuracy is determined.

$$P_{best} = \max\big(P(\omega_1), P(\omega_2), \cdots, P(\omega_k)\big) \tag{16}$$

$$P_{best} = P(\omega_{best}) \tag{17}$$

Where ω_i is the hyperparameter combination obtained by the j-th random sampling; $P\big(\omega_j\big)$ is the accuracy of the LSTM model with the j-th set of random parameters; k is the total number of trials; ω_{beat} is the set of parameters with the highest accuracy among all random parameter combinations.

Through the above method, combined with early stopping, RandomSearch parameter optimization method, and LSTM model training method, the optimal parameters of the LSTM model unique to each province have been successfully determined, and the prediction results of future data have been obtained.

5. Conclusion

This paper establishes a model based on the development status and environmental carrying capacity of various provinces in China, combined with the development of AI in recent years, to describe the size of the AI development prospects of a region. The most suitable provinces for the development of AI in China over the next five years have been identified, which are: Jiangsu Province, Guangdong Province, Shandong Province, Zhejiang Province, and Sichuan Province. By observing the level of AI development and "fertility" in Jiangsu Province, it is found that a high level of AI development does not necessarily indicate a high level of fertility. Too rapid and hasty development may lead to a decrease in fertility. Therefore, to maintain the momentum of AI development in the future, it is essential to establish better carbon emission treatment mechanisms or adopt cleaner energy structures. The research also reveals that the development of AI has regional driving effects. The "fertility" of the provinces adjacent to areas with higher levels of AI development will also improve to some extent, which is particularly evident in Anhui Province. Moreover, the level of economic development does not directly determine the level of AI development fertility of a province. To a large extent, the development of AI depends on the sufficient supply of electricity. If an area is energy-constrained and has a high demand for electricity, this may become a disadvantage for the development of AI, as reflected by comparing Sichuan Province and Jiangsu Province. Therefore, in the future where the development of AI is urgent, ensuring the power supply for AI development with minimal carbon emissions is an important prerequisite.

Finally, it is worth mentioning that the development of AI does not always have a negative impact on carbon emissions. When AI technology is developed to a sufficient extent, it can also be applied to technologies that can reduce carbon emissions, such as power generation, power saving, and waste gas treatment. However, as AI is still in the development stage, we should always consider the carbon emissions caused by the development of AI in its development planning.

References

[1] Zhu Mingjie. AI applications may become a new productive force leading industrial transformation [J]. Shanghai People's Monthly, 2024, (06): 52.

[2] Feng D, Shengnan Z, Jiao Z, et al. The Impact of the Integrated Development of AI and Energy Industry on Regional Energy Industry: A Case of China [J]. International Journal of Environmental Research and Public Health, 2021, 18 (17): 8946-8946.

[3] Jiang Hongde. Developing computing power as an important engine for AI large models [J]. China Informatization, 2024, (06): 18-20.

[4] Xue Fei, Liu Jiaqi, Fu Yamei. The impact of artificial intelligence technology on carbon emissions [J]. Science and Technology Progress and Countermeasures, 2022, 39 (24): 1-9.

[5] Chen Yongwei. Beyond ChatGPT: Opportunities, Risks, and Challenges of Generative AI [J]. Journal of Shandong University (Philosophy and Social Sciences), 2023, (03): 127-143.

[6] Shi Bo. The path selection of artificial intelligence promoting the transformation and upgrading of China's economic structure in the new era [J]. Journal of Northwest University (Philosophy and Social Sciences), 2019, 49 (05): 14-20.

[7] Zhang Xiaoshun, Li Jincheng, Guo Zhengxun. Large model-assisted topology optimization of collection system for large offshore wind farm [J]. High Voltage Technology, 2024, 50 (07): 2894-2905.

[8] Ding Lifu, Chen Ying, Xiao Tannan, et al. Preliminary exploration of a new power system generative intelligent application model based on large language models [J/OL]. Automation of Electric Power Systems, 1-16 [2024-08-01].

[9] Ma Guangwei, Zhong Yuting, Zhong Jian. Construction and empirical measurement of China's artificial intelligence development evaluation index system [J]. Science and Technology Management Research, 2023, 43 (18): 55-61.

[10] Yang Mengcheng. Research on the evaluation index system of China's artificial intelligence industry [J]. Value Engineering, 2021.