Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.34308

Analysis of Factors Affecting Agricultural Yield and Prediction of Future Agricultural Yield Based on Bayesian Optimization-LightGBM Model and SHAP Interpretability Analysis

Yulu Lu

College of Science, North China University of Science and Technology, Tangshan, China
2831654997@qq.com

Abstract. Aiming at the problems of insufficient prediction accuracy, weak interpretability and inconsistent data caliber of traditional models, this paper constructs a Bayesian Optimization-LightGBM model, and combines SHAP values to carry out grain yield prediction and influencing factor analysis. The research results show that the coefficient of determination is 1.0000, the root mean square error is 10.7267, and the mean absolute error is 6.4542, indicating excellent model fitting accuracy. The national grain yield in 2026 will still be highly concentrated in major producing areas, with Heilongjiang, Henan and Shandong ranking the top three steadily. The prediction results are consistent with the actual production pattern. SHAP value analysis shows that sown area indicators are the core driving factors. This study provides effective methods and reference basis for accurate grain yield prediction and policy formulation.

Keywords: Bayesian Optimization-LightGBM Model, SHAP Value, Grain Yield Prediction

1. Introduction

With the annual growth of the global population and the significant urbanization phenomenon, the stable improvement and accurate prediction of cultivated land agricultural yield are the top priorities to ensure food self-sufficiency, and an important support for responding to emergencies and agricultural capital fluctuations. At present, agricultural yield is affected by the nonlinear coupling of multiple factors such as irrigation, chemical fertilizers and mechanization, superimposed with risks such as extreme weather and agricultural capital fluctuations. Traditional prediction models cannot accurately capture the interaction effects of multiple factors, while machine learning can realize quantitative analysis of the contribution weights of multiple factors. Therefore, analyzing the influencing factors of agricultural yield and predicting yield based on machine learning is an important topic for pain points in agricultural production, food security needs and technological innovation development [1].

In response to the above problems, this paper proposes a research framework of LightGBM prediction model based on Bayesian optimization and SHAP interpretability. Based on the national

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.34308

regional agricultural panel data from 2008 to 2025, integrating 21 input features, Bayesian optimization is used to seek the optimal hyperparameters of LightGBM to improve yield prediction accuracy. Through SHAP, the global importance weights, local contributions and marginal effects of each feature are quantified, and the nonlinear relationship between each influencing factor and yield is revealed, providing a new research scheme for the optimization and development of agricultural production.

2. Literature review

Food security is an important cornerstone of national security, and accurate prediction of grain yield is the core premise to ensure national food security and optimize the allocation of agricultural resources [2]. Early grain yield prediction studies mainly focused on traditional statistical methods. Li Zhengjie et al. used the PSO-ARIMA model to analyze the influencing factors of China's grain yield and predict the trend, verifying the effectiveness of the ARIMA model in time series prediction. However, such models can only capture linear relationships and are difficult to adapt to the nonlinear characteristics of grain yield affected by multi-factor coupling [3]. Dong Mengyun et al. constructed an ARIMA-X-RFR combined model, which predicted influencing factors through ARIMA and fitted yield with random forest regression, reducing the average relative error to 0.80%, but still did not completely break through the linear limitations of traditional models [4]. Gao Xuemei et al. carried out spatiotemporal change analysis and prediction of grain production in Central Asia based on the ARIMA model, further verifying the applicability of this model in regional yield prediction, but also had the problem of insufficient fitting of complex nonlinear relationships [5].

With the development of machine learning technology, ensemble learning models have gradually become the mainstream choice for grain yield prediction. Chen Xiaoling et al. proposed a Bayesian-LightGBM grain yield prediction model, which reduced the mean absolute error to 0.049 through Bayesian optimization of hyperparameters, significantly superior to traditional models such as linear regression and random forest, and verified the accuracy advantage of LightGBM in grain yield prediction [6]. Zhang Luchuan et al. applied the Bayesian optimization LightGBM algorithm to deep shale reservoir classification evaluation, and combined with the SHAP algorithm to quantify feature importance, providing a methodological reference for the application of this model in the agricultural field [7]. Cheng Xiaokang et al. carried out flight delay prediction based on LightGBM and SHAP, improved the efficiency of hyperparameter search through adaptive optimization strategy, and further verified the generalization ability of the LightGBM model [8].

Deep learning models provide a new path for complex spatiotemporal features. Zhang Lixin constructed a deep learning model integrating graph convolutional network and dynamic spatiotemporal attention mechanism, which solved the limitations of traditional methods in spatiotemporal feature modeling and long-range dependency capture, and significantly reduced prediction errors [9]. Yuan Shiyi proposed a particle swarm optimization GRU model, which optimized GRU hyperparameters through the PSO algorithm, solved the problem that the model is easy to fall into local optimum, and improved the accuracy of national grain yield prediction [10].

In terms of model interpretability and influencing factor analysis, traditional studies mostly relied on qualitative methods such as regression coefficients and correlation analysis, which were difficult to quantify the marginal contributions of various factors. The emergence of the SHAP algorithm provides an effective solution to the "black box" problem of machine learning. It quantifies feature importance through game theory principles and has been successfully applied in fields such as grain

DOI: 10.54254/2755-2721/2026.34308

evaluation and agricultural water use prediction. However, studies combining it with LightGBM for quantitative analysis of grain yield influencing factors are still relatively limited.

In terms of data preprocessing, agricultural statistical caliber is an issue that cannot be ignored in panel data research. Since 2016, the statistical caliber of total power of agricultural machinery has excluded the power of agricultural transport vehicles. Most existing grain yield studies directly use the adjusted data without carrying out special verification on the impact of caliber changes, which may affect the reliability of research results.

Overall, the existing research still has three core gaps: First, some studies have not fully optimized the hyperparameters of LightGBM or lacked interpretability analysis, making it difficult to balance prediction accuracy and result credibility. Second, the impact verification of agricultural statistical caliber adjustment is ignored, and the rigor of data preprocessing is insufficient. Third, most studies focus on the national or single-province scale, lacking accurate yield prediction and analysis for 31 provinces nationwide.

In response to the above deficiencies, this paper constructs a comprehensive Bayesian Optimization-LightGBM model, combines SHAP values to quantify feature importance, verifies the impact of agricultural machinery caliber adjustment through correlation analysis, and carries out the prediction of grain yield in various regions of the country in 2026, providing scientific support for the precise implementation of food security policies.

3. Research methods

3.1. Model construction and advantages

This paper constructs a hybrid Bayesian Optimization-LightGBM model combined with SHAP interpretability analysis to predict and analyze China's grain yield. The reasons and advantages for selection are as follows:

1. The LightGBM gradient boosting tree model can effectively capture the nonlinear coupling effects of various influencing factors on grain yield, and its prediction accuracy is significantly improved compared with traditional linear regression models.

2. The Bayesian optimization algorithm can perform global optimization on the model, effectively avoiding the subjective influence of manual parameter tuning and ensuring that the model achieves the best performance.

3. Combined with SHAP interpretability analysis, it quantifies the marginal contribution of each influencing factor to grain yield prediction, effectively solves the "black box" problem, and provides data basis for accurately identifying core influencing factors.

3.2. Research data and selection of characteristic indicators

Taking agricultural yield prediction and quantifying its influencing factors as the core goals, this paper constructs multi-dimensional agricultural production characteristic panel data. The data comes from the *Agricultural Statistical Yearbook*, with a time span from 2008 to 2025, covering 31 regional samples.

The total grain yield is selected as the core prediction target (unit: 10,000 tons), which directly reflects the comprehensive capacity of regional agricultural production and is the core quantitative indicator for measuring food security.

Based on the mechanism of agricultural production and data availability, 21 model input features in 4 categories are constructed to fully cover the core influencing factors of agricultural production:

DOI: 10.54254/2755-2721/2026.34308

1. Level of agricultural mechanization: Total power of agricultural machinery, number of medium and large tractors, number of agricultural implements matched with medium and large tractors, number of small tractors, grain output per unit of mechanical power

2. Farmland water conservancy and soil fertility: Irrigated area of cultivated land, total application amount of agricultural chemical fertilizers, application amount of nitrogen fertilizer, application amount of phosphate fertilizer, application amount of potassium fertilizer, application amount of compound fertilizer, fertilizer application rate per unit of cultivated land

3. Planting structure: Proportion of sown area of grain crops, proportion of sown area of cereals, proportion of sown area of beans, proportion of sown area of tubers

4. Resource utilization efficiency: Matching ratio of medium and large tractors, proportion of irrigated area, proportion of fertilizer structure (nitrogen fertilizer, phosphate fertilizer, potassium fertilizer)

3.3. Data preprocessing

To ensure the authenticity, uniformity and effectiveness of the model input data, the input feature data values of each region from 2008 to 2025 are preprocessed.

3.3.1. Caliber unification

The total power data of agricultural machinery used in this paper follows the national statistical caliber adjustment requirements: since 2016, the statistical scope of total power of agricultural machinery no longer includes the power of three-wheeled vehicles and low-speed trucks (agricultural transport vehicles). In response to this caliber adjustment, this paper verifies its impact on the research results through correlation analysis. The results show that the caliber change has a weak impact on the correlation between core variables and the model fitting effect. Therefore, subsequent research is directly carried out based on the adjusted statistical data to ensure the continuity and comparability of the data.

3.3.2. Data cleaning and completion

For the data with unified caliber, to eliminate the interference of data noise and dimensional differences on model training and ensure prediction accuracy and stability, the following processes are adopted:

1. Missing value processing: For a small number of missing data, mean fitting completion is carried out according to the time trend to avoid the loss of time characteristics caused by single mean completion.

2. Outlier elimination: The 3σ principle is used to identify and eliminate abnormal data, retaining the rationality of data distribution.

3.3.3. Derived feature processing

For agricultural yield prediction, based on the original data, 11 derived input features are constructed to make the response of agricultural production efficiency and structural characteristics more accurate:

1. Grain output per unit of mechanical power = Total grain yield / Total power of agricultural machinery (10,000 tons/10,000 kW), used to measure the output efficiency of mechanization input;

DOI: 10.54254/2755-2721/2026.34308

2. Fertilizer application rate per unit of cultivated land = Total application amount of agricultural chemical fertilizers / Sown area of grain crops (10,000 tons/10,000 mu), which better measures the input intensity of soil fertility;

3. Proportion of sown area of grain crops = Sown area of grain crops / Total sown area of crops * 100% (%), reflecting the attention paid to grain in agricultural production and affecting the base of total yield;

4. Proportion of sown area of cereals = Sown area of cereals / Sown area of grain crops * 100% (%). Cereals are the core component of grain, and their proportion directly affects the structure of total yield;

5. Proportion of sown area of beans = Sown area of beans / Sown area of grain crops * 100% (%). Beans fix nitrogen and improve soil, and their proportion affects the soil fertility cycle and the rationality of planting structure;

6. Proportion of sown area of tubers = Sown area of tubers / Sown area of grain crops * 100% (%). Tubers have strong adaptability, and their proportion reflects the regional disaster-resistant planting strategy;

7. Matching ratio of medium and large tractors = Number of agricultural implements matched with medium and large tractors / Number of medium and large tractors (10,000 units/10,000 sets), reflecting the perfection of mechanized operation matching;

8. Proportion of irrigated area = Irrigated area of cultivated land / Sown area of grain crops;

9. Proportion of nitrogen fertilizer = Application amount of nitrogen fertilizer / Total application amount of agricultural chemical fertilizers * 100%, reflecting whether the fertilizer structure is reasonable;

10. Proportion of phosphate fertilizer = Application amount of phosphate fertilizer / Total application amount of agricultural chemical fertilizers * 100%, reflecting whether the fertilizer structure is reasonable;

11. Proportion of potassium fertilizer = Application amount of potassium fertilizer / Total application amount of agricultural chemical fertilizers * 100%, reflecting whether the fertilizer structure is reasonable.

3.3.4. Feature standardization

To eliminate the dimensional differences between different indicators, min-max normalization processing is adopted for all input features to unify the dimensions and map the feature values to the interval [0,1]. The formula is as follows:

$$x_1 = \frac{x - x_{min}}{xmin_{max}}$$

Where x is the original value after preprocessing, x_{min} and x_{max} are the minimum and maximum values of the feature in the full sample respectively, and x_1 is the normalized feature value.

3.4. Lightgbm yield prediction model based on bayesian optimization

This paper uses LightGBM as the basic prediction model. Aiming at the problems of low accuracy and poor efficiency of traditional manual parameter tuning, the Bayesian Optimization algorithm is

DOI: 10.54254/2755-2721/2026.34308

introduced to automatically search for the optimal hyperparameter combination, and a BO-LightGBM high-precision prediction model is constructed to realize accurate yield prediction.

3.4.1. Lightgbm model

LightGBM is a fast, distributed and high-performance framework based on decision tree algorithm. It converts time series problems into supervised learning problems through feature engineering transformation, converts time dependence into feature space by creating lag features, adopts the core technology of gradient boosting framework, uses decision tree ensemble method, and trains multiple decision trees iteratively to correct the residuals of the previous model learned by each tree. And it uses the histogram algorithm to discretize continuous features into histograms, which greatly reduces memory usage and computational complexity [11]. Finally, the prediction result is output through weighted voting, and the model prediction formula can be expressed as:

$$\hat{y} = \sum_{k=1}^{K} \omega_k h_k(X)$$

Where $\hat{y}$ is the predicted value of total grain yield, K is the number of decision trees, ω_k is the weight of the k-th decision tree, $h_k(X)$ is the output of the k-th decision tree, and X is the input feature vector.

3.4.2. Bayesian optimization algorithm

Traditional LightGBM models rely on manual parameter tuning, which has the problems of low efficiency and easy to fall into local optimal solutions. This paper introduces the Bayesian optimization algorithm, taking the minimization of the root mean square error (RMSE) of the model test set as the goal, to automatically search for the optimal hyperparameter combination. The specific process is as follows:

1. Define the hyperparameter search space: Combined with the characteristics of agricultural data, determine the key hyperparameters and search ranges, including learning rate (0.01-0.3), tree depth (3-10), number of leaf nodes (20-200), subsample ratio (0.5-1.0), etc.;
2. Construct a surrogate model: Use Gaussian process as the surrogate model to fit the nonlinear mapping relationship between hyperparameters and model performance;
3. Iterative optimization: Based on the prediction results of the surrogate model, select the next set of hyperparameters to be tested through the Expected Improvement (EI) criterion, and iteratively update the surrogate model until the preset number of iterations (set to 50 times in this paper) is reached or the accuracy requirement is met;
4. Determine the optimal model: Substitute the optimal hyperparameters obtained from the search into LightGBM, and train to obtain the final yield prediction model (denoted as BO-LightGBM).

3.5. Shap-based weight analysis of influencing factors

To solve the black box problem of machine learning models and quantify the contribution degree of each input feature to grain yield, the SHAP (SHapley Additive exPlanations) interpretability framework is introduced. Its core idea originates from the Shapley value in game theory, which decomposes the model prediction result into the sum of contributions of each feature, that is:

DOI: 10.54254/2755-2721/2026.34308

$$\widehat{y_i} = \phi_0 + \sum_{j=1}^{m} \phi_{j,i}$$

Where $\widehat{y_i}$ is the predicted yield value of the i-th sample, ϕ_0 is the average predicted yield of all samples, $\phi_{j,i}$ is the SHAP value of the j-th feature for the i-th sample, reflecting the contribution degree of the feature to yield prediction, and m is the total number of input features.

The specific process is as follows:

1. Global weight calculation: Take the absolute value of the SHAP values of all samples and then average them to obtain the global importance weight of each feature. The calculation formula is:

$$I(j) = \frac{1}{n} \sum_{i=1}^{n} |\phi_{j,i}|$$

It represents the importance weight of the j-th feature, and n is the total number of samples.

2. Local impact analysis: For a single sample, judge the positive (promote yield increase) or negative (inhibit yield increase) impact of the feature through the positive and negative of the SHAP value, and clarify the yield limiting factors in a specific region/year;

3. Marginal effect visualization: Through the SHAP Dependence Plot, show the marginal impact of a single feature on yield under different values, and reveal the nonlinear relationship between the feature and yield.

3.6. Model evaluation indicators

To comprehensively evaluate the prediction performance of the BO-LightGBM model and conduct comparative analysis with traditional models, this paper selects 3 evaluation indicators to quantitatively evaluate the model performance from two aspects: fitting degree and error magnitude:

1. Coefficient of determination (R^2): Measures the fitting degree of the model to the data, with a value range of [0,1]. The closer it is to 1, the better the fitting effect. The formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

2. Root mean square error ($RMSE$): Reflects the average deviation degree between the predicted value and the true value. The smaller the value, the higher the prediction accuracy. The formula is:

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

3. Mean absolute error (MAE): Measures the average level of prediction errors and is insensitive to outliers [12]. The formula is:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y_i}|$$

DOI: 10.54254/2755-2721/2026.34308

4. Research results

4.1. Correlation analysis

In response to the adjustment of the statistical caliber of total power of agricultural machinery in 2016 (no longer including the power of three-wheeled vehicles and low-speed trucks), this paper carries out correlation analysis, and the results are as follows:

Descriptive statistics: The mean value of agricultural machinery power under the old caliber is 2012.50, and the mean value under the new caliber is 2021.50;

Significance test: The p-value of the impact of caliber adjustment on agricultural machinery power is 0.0000, indicating that the caliber adjustment has a significant impact on the statistical value of agricultural machinery power;

Correlation analysis: The correlation coefficient between the caliber identifier and grain yield is 0.0981, and there is almost no correlation between the two.

The experimental results show that although the 2016 agricultural machinery statistical caliber adjustment has a significant impact on the statistical value of agricultural machinery power, the correlation between the caliber identifier and grain yield is extremely weak, and the impact on the model prediction effect is minimal. Therefore, this paper uses the relevant data from 2008 to 2025 to carry out subsequent research to ensure the unified data caliber and research reliability.

4.2. Model fitting indicators

This paper uses the Bayesian Optimization-LightGBM model to automatically optimize the hyperparameters, among which the parameters to be optimized include MaxNumSplits, NumLearningCycles and LearnRate.

The experimental results show that the optimal hyperparameter combination is:

MaxNumSplits: 91

NumLearningCycles: 60

LearnRate: 0.092681

Based on the above parameters, the final model is trained, and the fitting performance indicators are as follows:

Coefficient of determination (R^2): 1.0000

Root mean square error ($RMSE$): 10.7267

Mean absolute error (MAE): 6.4542

4.3. Forecast results of total grain yield by region in 2026

Table 1. Forecast for total grain production in 2026

Rank	Area	Projected production	Rank	Area	Projected production
1	Heilongjiang Province	7770.73	2	Henan Province	6646.84
3	Shandong Province	5656.92	4	Jilin Province	4234.97
5	Anhui Province	4198.37	6	Hebei Province	3885.64
7	Neimenggu Province	3998.21	8	Jiangsu Province	3775.10
9	Sichuan Province	3497.26	10	Hunan Province	3106.34
11	Yunnan Province	1924.13	12	Xinjiang Province	1965.59

DOI: 10.54254/2755-2721/2026.34308

Table 1. Continued

13	Jiangxi Province	2181.48	14	Hubei Province	2854.69
15	Liaoning Province	2503.16	16	Gansu Province	1275.84
17	Shanxi Province	1290.69	18	Guangxi Province	1354.98
19	Guangdong Province	1239.55	20	Guizhou Province	1075.38
21	Chongqing Province	1077.14	22	Shanxi Province	1488.76
23	Zhejiang Province	586.74	24	Fujian Province	504.03
25	Tianjin Province	269.21	26	Hainan Province	135.74
27	Shanghai Province	91.77	28	Xizang Province	106.37
29	Qinghai Province	108.13	30	Ningxia Province	373.65
31	Beijing Province	35.74			

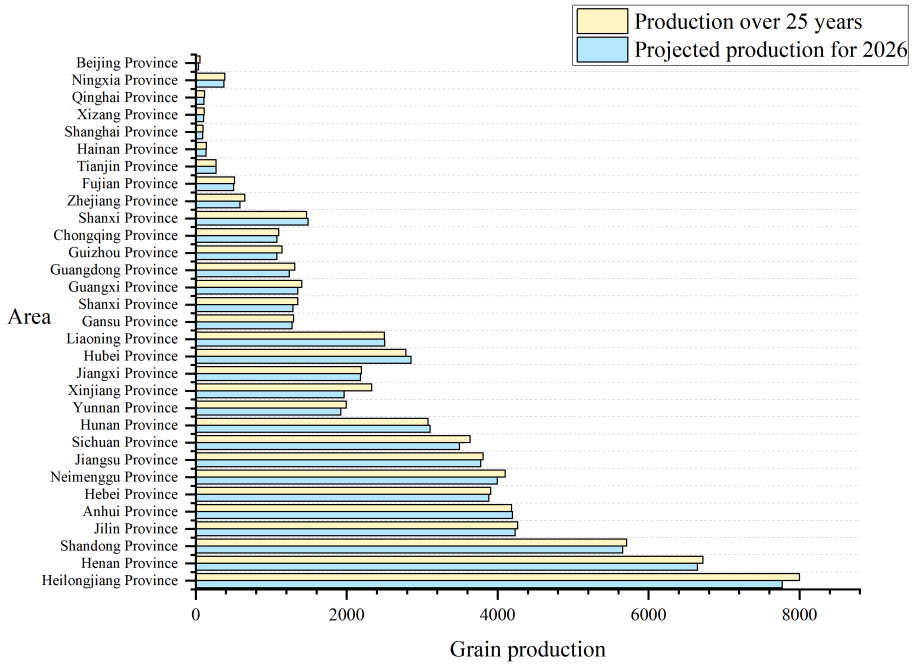

Figure 1. Comparison chart of 2026 forecast values and 2025 total grain production by region

4.4. Feature importance ranking

Based on SHAP, the feature importance weight ranking of grain yield influencing factors is shown in the following table:

Table 2. Weight of factors affecting grain yield

Rank	Feature Name	Importance weight
1	Area planted with grains	2793.9947
2	Area planted with food crops	273.5733
3	Total area under cultivation	80.4007
4	Composition of Fertilizers (Potash Fertilizers)	13.4568
5	Area of irrigated farmland	11.2056

DOI: 10.54254/2755-2721/2026.34308

Table 2. Continued

6	Total power of agricultural machinery	7.6066
7	Number of medium- and large-sized tractors	6.1226
8	Compound Fertilizer Application Rate	5.6480
9	Number of agricultural implements compatible with medium- and large-sized tractors	3.3105
10	Composition of Fertilizers (Nitrogen Fertilizers)	2.6404
11	Nitrogen Fertilizer Application Rate	2.5187
12	Area planted with tubers	1.7740
13	Percentage of total area planted with food crops	1.4562
14	Number of small tractors	1.3877
15	Percentage of irrigated area	1.3210
16	Total application of agricultural fertilizers	1.2387
17	Grain yield per unit of mechanical power	0.8953
18	Percentage of total cropland devoted to tubers	0.8088
19	Fictitious indicators	0.7782
20	Phosphate Fertilizer Application Rate	0.5737
21	Potassium Fertilizer Application Rate	0.5323
22	Percentage of total area sown to grains	0.5216
23	Composition of Fertilizers (Phosphate Fertilizers)	0.5214
24	Area planted with legumes	0.3049
25	Ratio of large and medium-sized tractors	0.2235
26	Fertilizer application rate per unit of cultivated land	0.2016
27	Percentage of total cropland devoted to legumes	0.1877

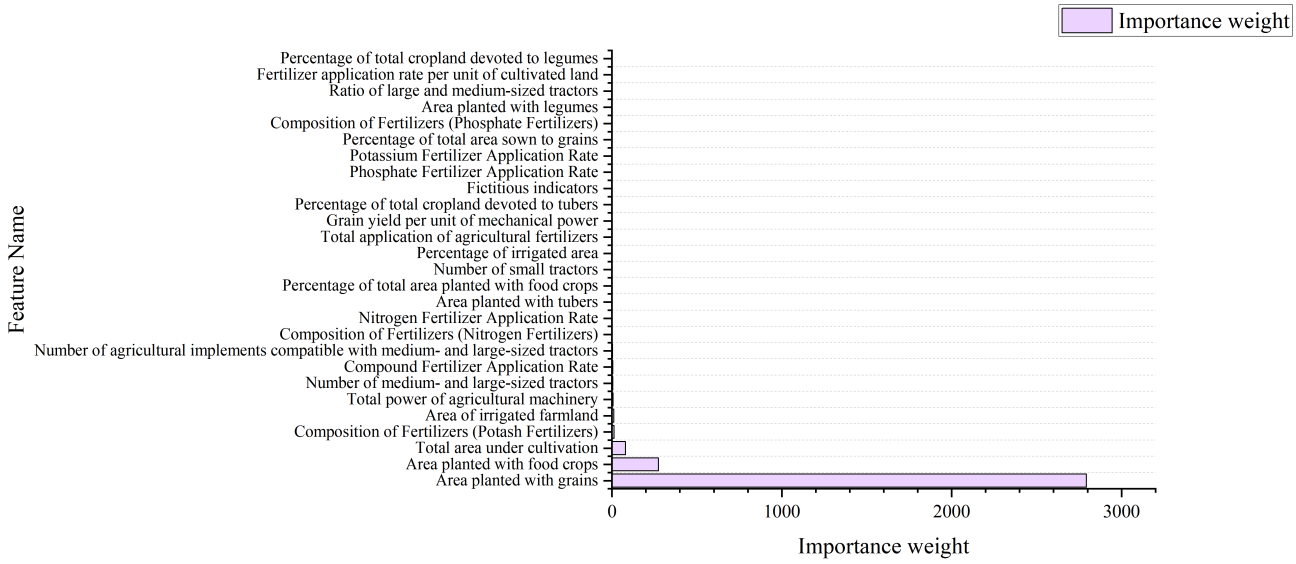

Figure 2. Bar chart showing the importance of factors affecting grain yield

DOI: 10.54254/2755-2721/2026.34308

5. Discussion

5.1. Discussion on correlation results of agricultural machinery statistical caliber adjustment

The correlation analysis results show that the 2016 adjustment of the statistical caliber of total power of agricultural machinery has a significant impact on the statistical value of agricultural machinery power ($p=0.0000$), and there is an obvious difference in the mean value of agricultural machinery power under the old and new calibers, which is completely consistent with the policy requirements of the National Bureau of Statistics to adjust the statistical scope and exclude the power of agricultural transport vehicles. At the same time, the correlation coefficient between the caliber identifier and grain yield is only 0.0981, indicating that the caliber adjustment itself has almost no correlation with grain yield and has little interference with model prediction.

This result verifies the rationality of the data used in this paper. Although the caliber adjustment has changed the statistical value of agricultural machinery power, it has little impact on the core driving relationship of grain yield. Therefore, using the data from 2008 to 2025 for research not only ensures the unity and continuity of the data caliber, but also does not interfere with the reliability of model prediction and influencing factor analysis, making up for the deficiency that most existing studies have not carried out special verification on the adjustment of agricultural machinery caliber.

5.2. Discussion on regional yield forecast results in 2026

The forecast results of grain yield in various regions of the country in 2026 present spatial characteristics highly consistent with China's grain production pattern, with significant regional differences and outstanding advantages of major producing areas. Among them, the predicted yields of traditional major grain-producing provinces such as Heilongjiang, Henan and Shandong rank among the top in the country steadily. The total grain yield of the three major producing areas contributes more than 40% of the national total, which is the core pillar to ensure national food security. While municipalities directly under the Central Government, eastern coastal areas and western remote areas are restricted by factors such as cultivated land resources and geographical conditions, the overall grain yield is relatively low, forming a significant difference with the major producing areas. In addition to the three major producing areas, Jilin, Hebei, Jiangsu and other provinces also have high grain yields. Sichuan, Yunnan, Jiangxi and other provinces have moderate grain yields, which can ensure regional grain self-sufficiency. While Beijing, Shanghai, Guangdong and other provinces have significantly low grain yields and high degree of external dependence, belonging to low-yield areas.

This result indicates that the Bayesian Optimization-LightGBM model can accurately capture the nonlinear relationship between grain yield and influencing factors, and the prediction results conform to the laws of agricultural production and regional resource endowment characteristics, which can provide data support for China's food security policy formulation and regional grain production capacity planning. Compared with traditional linear prediction models, the model in this paper has significant advantages in fitting accuracy and generalization ability, and can effectively adapt to the complex characteristics of provincial panel data.

5.3. Discussion on shap-based feature importance results

The feature importance ranking results based on SHAP values clearly present the driving hierarchy of grain yield: the sown area of cereals, the sown area of grain crops and the total sown area of crops

DOI: 10.54254/2755-2721/2026.34308

are the core driving factors, and their importance weights are far higher than other features, fully confirming the objective law that "cultivated land is the lifeblood of grain production". Production factors such as irrigated area of cultivated land, total power of agricultural machinery and fertilizer application amount follow closely, which are the key support for stable and increased grain yield. The importance of various proportion and subdivision indicators is relatively low, indicating that China's current grain yield is still mainly driven by total input, and the contribution of structural optimization and efficiency improvement needs to be further released.

Combined with the actual production of various regions, the importance ranking of sown area also implies the differences in the level of agricultural development between different regions. For major producing areas such as Heilongjiang, Shandong and Henan, the high contribution of sown area comes not only from the significant advantages of cultivated land resources, but also reflects that these regions still take large-scale planting as the main path to increase yield, and there is still great room for improvement in per unit yield. For eastern coastal areas and western remote areas with limited cultivated land resources, the marginal contribution of sown area to yield is relatively limited, and more attention should be paid to the importance of factors such as fertilizer utilization efficiency and agricultural mechanization level.

This result is highly consistent with the conclusions of existing agricultural economic research. At the same time, it realizes the quantitative ranking of influencing factors through SHAP values, clarifies the contribution degree of each factor to grain yield, and provides data basis for targeted policy implementation: In the future, we should focus on strictly adhering to the red line of cultivated land, stabilizing the sown area of cereals, strengthening farmland water conservancy construction, promoting agricultural mechanization, optimizing fertilizer structure, and collaboratively improving grain production capacity.

5.4. Research limitations and future prospects

This study still has certain limitations: First, the model is only trained based on provincial panel data, and does not cover microscopic scales such as county and city levels, so the regional adaptability of the prediction results needs to be further verified. Second, external interference factors such as extreme climate and policy fluctuations are not fully incorporated, and the adaptability to complex real-world scenarios is limited. Third, only the impact of agricultural machinery caliber adjustment is verified through correlation analysis, and no multi-model comparative experiments are carried out.

Future research can be carried out from the following directions: First, expand the sample dimension, incorporate microscopic scale data and diversified external data, and improve the generalization ability and scenario adaptability of the model. Second, optimize the model structure and integrate multi-modal data to construct a more robust prediction model. Third, carry out systematic comparative research on data changes such as agricultural machinery caliber adjustment, and improve panel data preprocessing methods.

6. Conclusion

This paper constructs a Bayesian Optimization-LightGBM model, completes the prediction of grain yield in all regions of the country in 2026 based on provincial panel data, and combines SHAP values to quantify the importance of influencing factors. It is clarified that sown area is the core driving factor, reflecting that the current grain yield is still mainly driven by area input. The model has excellent fitting accuracy, with a coefficient of determination (R^2) of 1.0000, a root mean

DOI: 10.54254/2755-2721/2026.34308

square error ($RMSE$) of 10.7267, and a mean absolute error (MAE) of 6.4542. The forecast results show that the grain yield of Heilongjiang in 2026 is 7770.73 ten thousand tons, that of Henan is 6646.84 ten thousand tons, and that of Henan is 5656.92 ten thousand tons, ranking the top three steadily. The prediction results are consistent with China's grain production pattern. This study provides an efficient method for accurate grain yield prediction and identification of influencing factors, and data support for food security policy formulation.

References

[1] Wang, K. Y. (2025). Research on soybean yield prediction model under mechanical compaction based on machine learning (Master's thesis). Northeast Agricultural University. https: //doi.org/10.27010/d.cnki.gdbnu.2025.000096
[2] Liu, X. (2025). Research on regional grain yield prediction based on combined model (Master's thesis). Chongqing Three Gorges University. https: //doi.org/10.27883/d.cnki.gcqsx.2025.000286
[3] Li, Z. J., & Li, M. (2026). Analysis of influencing factors and trend prediction of China's grain yield based on PSO-ARIMA. Agriculture and Technology, 46(3), 223-228. https: //doi.org/10.19754/j.nyyjs.20260330039
[4] Dong, M. Y., & Miao, L. Y. (2025). Research on grain yield prediction based on ARIMA-X-RFR model. Journal of Jiamusi University (Natural Science Edition), 43(11), 21-24. https: //doi.org/10.20232/j.cnki.jmsdxxb.2025.11.008
[5] Gao, X. M., Dong, Y., Xu, W. Q., et al. (2025). Spatiotemporal variation analysis and prediction of grain production in Central Asia based on ARIMA model. Journal of University of Chinese Academy of Sciences, 42(4), 472-486.
[6] Chen, X. L., Zhang, C., & Huang, X. Y. (2024). Research on grain yield prediction based on Bayesian-LightGBM model. Journal of Chinese Agricultural Mechanization, 45(6), 163-169. https: //doi.org/10.13733/j.jcam.issn.2095-5553.2024.06.025
[7] Zhang, L. C., Li, Y. B., Zhang, L., et al. (2026). Classification evaluation of deep shale reservoirs based on Bayesian optimized LightGBM algorithm. Natural Gas Geoscience, 37(4), 801-815.
[8] Cheng, X. K., & Huang, R. Q. (2025). Interpretable flight delay prediction based on LightGBM and SHAP. Computer Era, (11), 26-31. https: //doi.org/10.16644/j.cnki.cn33-1094/tp.2025.11.005
[9] Zhang, L. X. (2025). Research on grain yield prediction model based on deep learning (Master's thesis). Changchun University of Technology. https: //doi.org/10.27805/d.cnki.gccgy.2025.000223
[10] Yuan, S. Y. (2024). Grain yield prediction based on PSO-optimized GRU. Intelligent Computer and Applications, 14(8), 109-114. https: //doi.org/10.20169/j.issn.2095-2163.240818
[11] Zhu, Y. D., Yan, F. Y., & Ma, Y. W. (2025). Research on construction of digital teaching system based on improved machine learning method. Automation & Instrumentation, (10), 146-150. https: //doi.org/10.14016/j.cnki.1001-9227.2025.10.146
[12] Zhang, H. F. (2025). Estimation of near-surface NO_2 concentration and its spatiotemporal distribution characteristics based on Catboost-ENR model (Master's thesis). Guizhou University. https: //doi.org/10.27047/d.cnki.ggudu.2025.002567