ISSN 2616-5902 Vol. 8, Issue 1: 147-154, DOI: 10.25236/AJBM.2026.080120

A Novel Evaluation and Prediction Model of Innovation Efficiency Based on GCA-GDEA Index Screening and AdaBoost Integration of Machine Learning

Xuanyi Meng[1,a], Jingjie Li[1,b,*], Yuqi Wang[1,c], Shanwei Li[1,d]

[1]School of Science, Tianjin University of Commerce, Tianjin, 300134, China
[a]1684632922@qq.com, [b]lxyljj@tjcu.edu.cn, [c]1289427601@qq.com, [d]lishanwei99@126.com
[*]Corresponding author

Abstract: *Manufacturing is the foundation of a country. Since the 21st century, China's manufacturing industry has entered a new stage of rapid growth and faces higher development requirements. Therefore, to better evaluate manufacturing enterprises' innovation efficiency, this paper innovatively constructs a systematic enterprise performance evaluation and classification model. Firstly, a preliminary efficiency assessment of the enterprise based on the generalized data envelopment analysis method(GDEA) was to obtain the relative efficiency value. Subsequently, grey relational analysis was used to quantify the degree of correlation between each output indicator and the efficiency value, screen out the key influencing factors, and achieve feature selection. On this basis, a secondary GDEA analysis was implemented to optimize the efficiency assessment results and enhance the ability to identify sample heterogeneity. Next, the efficiency evaluation results are labeled and divided to construct a binary classification dataset. The K-nearest neighbor, support vector machine, and logistic regression model are used for training, respectively, and the parameters are optimized through grid search. Finally, the AdaBoost ensemble learning method is utilized to conduct a weighted fusion of multiple base classifiers, construct a strong learner, and accurately identify high-efficiency enterprises. The results show that: (1) The efficiency evaluation platform is based on the grey correlation degree, and the AdaBoost integrated model has better evaluation and prediction capabilities than the traditional GDEA model. (2) Classify and predict the future development of the research enterprises and conduct specific strategic analysis for enterprises of different classifications.*

Keywords: *Manufacturing Champion Enterprises; Indicator Screening; GDEA; GCA; Adaboost*

1. Introduction

The global economic landscape has changed significantly, and international economic and trade rules are constantly adjusted. In order to promote the development of the manufacturing industry, many developed countries in Europe and America have increased their financial support for the manufacturing industry[1]. Manufacturing strategy has an important role in business competitive strategy because it connects performance indicators to company goal[2].

China's strategic deployment of building a new development pattern of "dual circulation" and the accelerated implementation of the "dual carbon" goals have also put forward new and higher requirements to promote the development of the manufacturing industry[3].

2. Literature Review and Research Methods

2.1 Research Background and Problem Statement

Champion enterprises are often pioneers of new productive forces, typically concentrated in key areas that reflect the competitiveness of Chinese manufacturing, and make outstanding contributions to the stability and prosperity of the social economy. However, currently, China's single champion manufacturing enterprises still face certain issues, such as competitiveness of key technologies has not reached international leadership, the added value of products is low, and some enterprises still have

ISSN 2616-5902 Vol. 8, Issue 1: 147-154, DOI: 10.25236/AJBM.2026.080120

insufficient international market development [4]. Based on this situation, this paper innovatively proposes a method for assessing these enterprises by combining the generalized data envelopment analysis method with machine learning prediction methods, building a brand-new platform for evaluating.

2.2 Literature Review

2.2.1 Efficiency Evaluation Methods and Evolution

DEA and its related extended models are widely used to evaluate the performance of organizations, enterprises, or economies, particularly in assessing production efficiency, resource utilization, and effectiveness analysis. However, in many practical situations, the selection of the reference set is not always reasonable. Therefore, Ma Zhanxin developed a set of DEA models with generalized reference sets, called Generalized DEA (GDEA) models, allowing decision-makers to select appropriate sample units according to their needs[5]. Many scholars have also elaborated on and improved the GDEA method[6] .This paper also uses the GDEA model to analyze the input-output efficiency of leading enterprises in the manufacturing industry.

2.2.2 Indicator Screening Method

The accuracy of DEA analysis results largely depends on the rationality of input-output indicators. Traditional methods are subjective. In order to improve objectivity, researchers attempted to introduce methods such as AHP, PCA, and GCA into the indicator selection process.For example, some scholars combine Analytic Hierarchy Process (AHP) with DEA for evaluation analysis [7], while others combine Principal Component Analysis (PCA) with DEA for efficiency evaluation[8] .

2.2.3 Efficiency Prediction Model

To better analyze the future impact of efficiency values, this paper proposes a new approach for efficiency classification prediction based on the PCA-GCA-GDEA model. Commonly used predictive classification algorithms mainly include K-Nearest Neighbors [9], Logistic Regression[10] and Support Vector Machine[11], among others.

For most base learner models, it is often difficult to achieve high accuracy while avoiding overfitting. However, in the field of machine learning, AdaBoost can elevate weak learning algorithms with accuracy slightly better than random guessing into strong learning algorithms with arbitrary accuracy, bringing a new method and new design ideas to the design of learning algorithms [12]. In cross-applications with DEA, the supplier selection method based on the DEA-AdaBoost hybrid model is suitable for multi-objective optimization problems [13].

2.3 Research Innovations

This paper combines principal component analysis, grey relational analysis, and DEA objectively to automatically select key indicators, and then embeds them into machine learning classification and prediction, breaking through the limitations of the traditional 'self-comparison and stop-calculation' approach.

Built on Matlab, the integrated 'Statistics-Machine Learning-Integration' efficiency platform offers high reliability for comparative validation and is plug-and-play for manufacturing and other industries.

Taking Manufacturing champion enterprises in Tianjin as an example, we use PCA-GDEA-GCA to set indicators, calculate efficiency, and build a two-category ensemble model.

2.4 Process Framework

The evaluation system proposed in this article achieves a transition from 'selecting indicators based on experience' to 'screening indicators based on data,' and from 'assessing efficiency statically' to 'measuring potential dynamically,' providing a systematic and intelligent solution for evaluating and predicting innovation efficiency in manufacturing enterprises, as Figure 1.

ISSN 2616-5902 Vol. 8, Issue 1: 147-154, DOI: 10.25236/AJBM.2026.080120

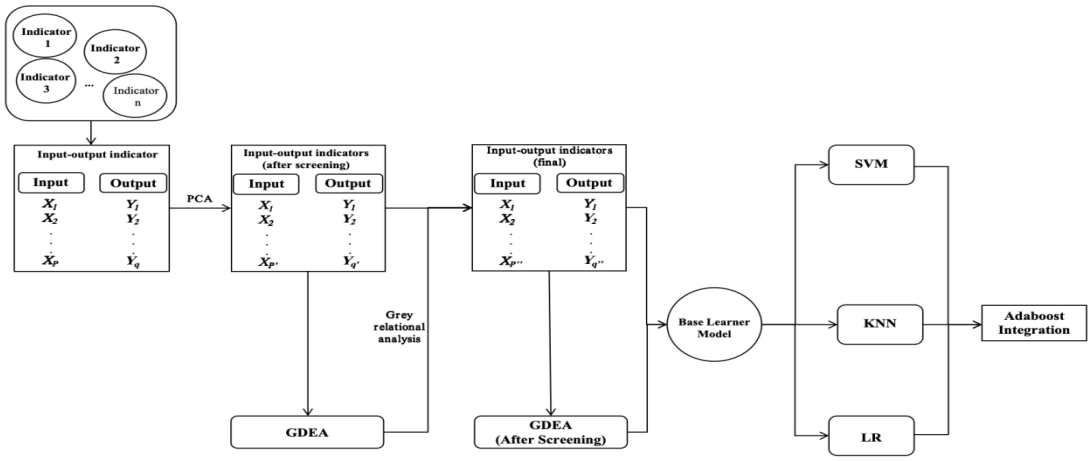

Figure 1: Research Process

3. Empirical analysis

3.1 Data Sources and Sample Description

The sample enterprises selected in this paper are drawn from the first five batches of single champion manufacturing enterprises in Tianjin announced by the state, choosing those in each batch that have publicly available data on finance, innovation, and other relevant information for the past three years. The data are sourced from CNINFO or the annual reports published by the enterprises. These sample data cover 31 champion enterprises, including both national and municipal levels, with 13 national champion enterprises and 18 municipal champion enterprises.

3.2 Indicator System and Preprocessing

Due to the complexity of constructing the evaluation index system and the current lack of research on the evaluation of the competitiveness of single-champion enterprises, there is limited relevant literature. The index system constructed in this paper references several previous studies[14] and integrates prior experience in manufacturing evaluation[15]to ultimately develop a preliminary index system and evaluation approach. The index system in this paper consists of three levels: the first-level indicators include two dimensions, input indicators and output indicators; the second-level indicators consist of six categories. Under the input indicators, there are three secondary indicators: scale capability, technical resource capability, and operational capability. Under the output indicators, there are three secondary indicators, namely market capacity, innovation capacity and operational efficiency. The details of the tertiary indicators under each secondary indicator are shown in Table 1.

Table 1: Division of Indicators at All Levels

First-level Indicator	Secondary Indicators	Variable Number	Variable Name	Indicator Attribute
Input	Scale	X1	Fixed assets	Positive
		X2	Number of employees	Positive
	Technical Resource	X3	Proportion of r&d personnel	Positive
		X4	Proportion of R & D investment	Positive
		X5	R & D personnel quality	Positive
		X6	Corporation age	Positive
	Operational	X7	Turnover of total assets	Positive
		X8	Quick ratio	Positive
		X9	Assets-liability ratio	Negative
Output	Market	X10	Operating income growth rate	Positive
		X11	Operating profit growth rate	Positive
		X12	Net profit ratio	Positive
		X13	ROE	Positive
	Business Operation	X14	Province relative market share	Positive
		X15	Domestic relative market share	Positive
	Innovation	X16	Number of intellectual property rights (number of patents)	Positive

The PCA method was applied for the preliminary index screening, and the principal components

ISSN 2616-5902 Vol. 8, Issue 1: 147-154, DOI: 10.25236/AJBM.2026.080120

with a cumulative contribution rate of 95% were retained. The results are shown in the following figures 2-5.

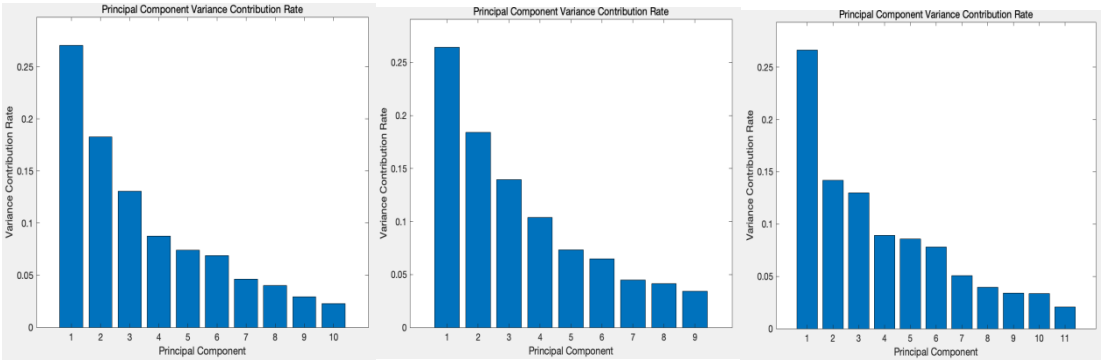

Figure 2 Quantity and Contribution Rate of Principal Components from 2020 to 2022.

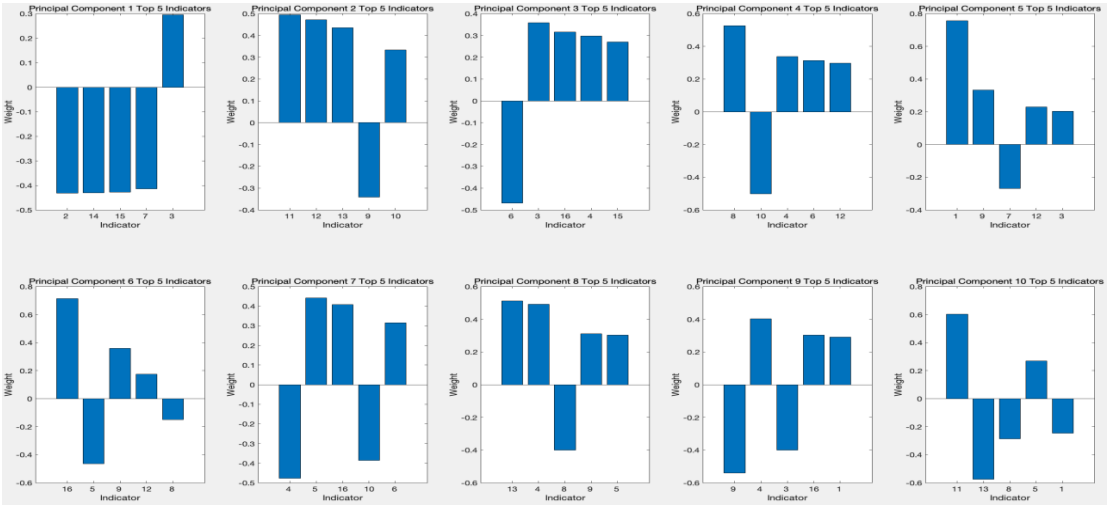

Figure 3 Top five principal components explaining 95 % of the cumulative variance in 2020.

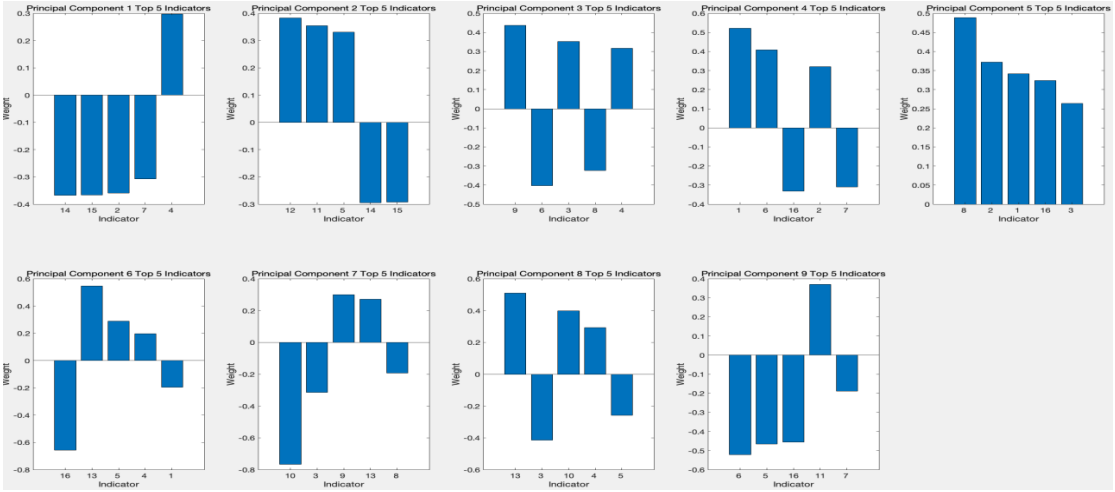

Figure 4 Top five principal components explaining 95 % of the cumulative variance in 2021.

ISSN 2616-5902 Vol. 8, Issue 1: 147-154, DOI: 10.25236/AJBM.2026.080120

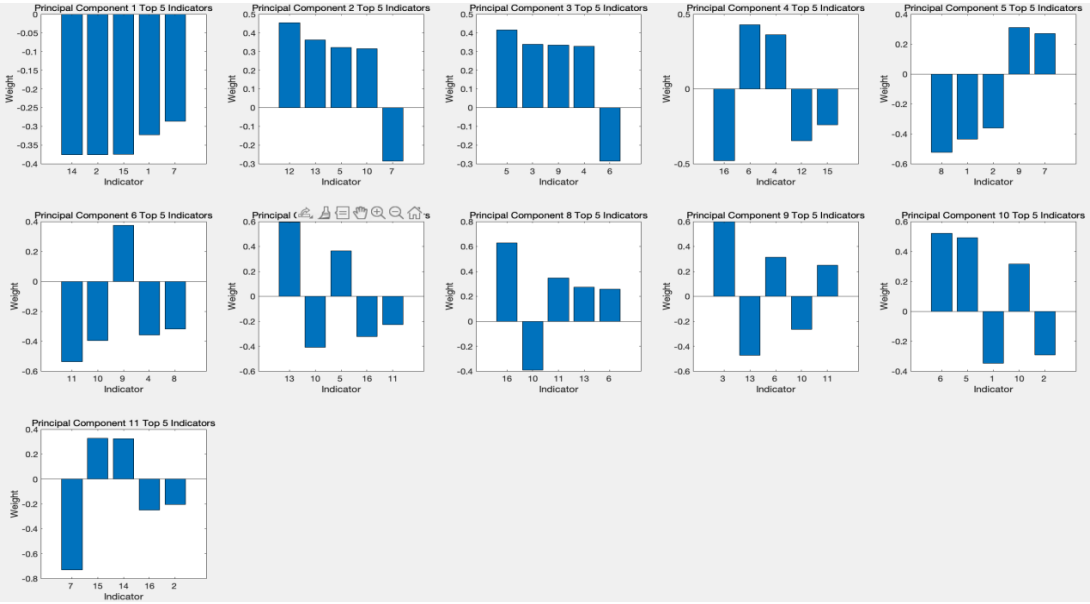

Figure 5 Top five principal components explaining 95 % of the cumulative variance in 2022..

3.3 Analysis of Efficiency Evaluation Results

After the first analysis using the generalized DEA, the efficiency of each of the 16 indicators can be determined. This study takes the efficiency values as the target sequence and calculates the degree of association between each input-output indicator and the overall efficiency value using the GCA method, in order to construct a set of key characteristic indicators that determine the final evaluation efficiency. The GCA analysis was conducted on the data of 31 leading companies from 2020 to 2022, and the average values were calculated. The results are summarized in Table 2.

Table 2: Average grey Correlation degree of each indicator over three years

Variable	GCA-Value	Sorting	Variable	GCA-Value	Sorting
X1	0.76350	13	X9	0.83003	4
X2	0.76320	15	X10	0.80000	6
X3	0.79157	8	X11	0.87747	1
X4	0.76893	11	X12	0.85267	2
X5	0.85037	3	X13	0.79600	7
X6	0.81370	5	X14	0.76303	16
X7	0.76710	12	X15	0.76320	14
X8	0.76967	10	X16	0.77460	9

It can be seen that X1, X2, X7, X14 and X15 are the last five indicators. The gray correlation values of these five indicators in the GCA each year are relatively low, so they are excluded. The steps of eliminating some indicators are in line with the actual situation and therefore have important practical significance. The remaining indicators after screening can reflect the operating conditions of the enterprise more effectively. Based on these remaining indicators, the generalized DEA model was reconstructed, and its indicators are shown in Table3.

Table 3: Division of indicators at all levels after excluding the indicators

First-level Indicator	Secondary Indicators	Variable Number	Variable Name
Input	Technical Resource	*X3*	Proportion of r&d personnel
		X4	Proportion of R & D investment
		X5	R & D personnel quality
		X6	Corporation age
	Operational	*X8*	Quick ratio
		X9	Assets-liability ratio
Output	Market	*X10*	Operating income growth rate
		X11	Operating profit growth rate
		X12	Net profit ratio
		X13	ROE
	Innovation	*X16*	Number of intellectual property rights (number of patents)

The data after eliminating the indicators was subjected to positive processing. The second DEA analysis can be conducted on the retained indicators

ISSN 2616-5902 Vol. 8, Issue 1: 147-154, DOI: 10.25236/AJBM.2026.080120

3.4 Base learner classification prediction and Adaboost classification prediction

The innovation efficiency value calculated based on the generalized DEA model, with 1 as the dividing point of the efficiency value, the classification and representative significance are shown in Table 4

Table 4 Definition of Broad DEA Efficiency

Efficiency value	Meaning
Less than 1	Non-valid Unit
Greater than or equal to 1	Valid Unit

Binary categorical variables can be divided according to the efficiency value, namely "0 - non-effective unit" and "1 - effective unit", to further construct the base learner. The evaluation metrics of the three base learner models are shown in Table5.

Table 5 Evaluation Metrics of Base Learner Model

		Training Accuracy	Verification accuracy	Overfitting Gap	Overall Accuracy
2020	SVM	0.7737	0.7429	0.0308	0.8065
	LR	0.8310	0.8095	0.0215	0.7742
	KNN	0.5880	0.6143	0.0263	0.7742
2021	SVM	0.6130	0.6143	0.0013	0.6129
	LR	0.7577	0.7429	0.0148	0.7419
	KNN	0.6453	0.6476	0.0023	0.6774
2022	SVM	0.7343	0.7095	0.0248	0.6452
	LR	0.7260	0.7429	0.0169	0.7097
	KNN	0.6930	0.7381	0.0451	0.7419

From the experimental results from 2020 to 2022, the three base learners exhibited relatively small gaps between their training and validation accuracy, with no significant overfitting observed overall, indicating that the feature selection and parameter tuning strategies were effective. Among them, logistic regression demonstrated the strongest stability and generalization capability, maintaining consistently high accuracy with the smallest training-validation gap. Although the support vector machine performed best in 2020, it proved more sensitive to data structure, with greater performance fluctuations in subsequent years. The K-nearest neighbors algorithm showed relatively low training accuracy but unique advantages in handling local features on the validation set. The three models displayed complementary characteristics across different years.

The experimental results are shown in Table 6, integrating logistic regression, support vector machines, and K-nearest neighbors as base learners into the AdaBoost framework can effectively combine their respective strengths. Logistic regression provides a stable baseline, support vector machines capture complex boundaries, and K-neighbors handle local structures, with their complementary roles enhancing overall classification performance. In particular, AdaBoost, by integrating multiple weak learners, effectively mitigates the bias and variance of individual models. While maintaining high accuracy (reaching 0.8710 in 2020), the overfitting gap remained extremely low (only 0.0113 in 2022), demonstrating excellent generalization capability and resistance to overfitting. Overall, AdaBoost outperformed any single base learner.

Table 6 Evaluation Indicators of the Adaboost model

	Training Accuracy	Verification accuracy	Overfitting Gap	Overall Accuracy
2020	0.7610	0.7206	0.0404	0.8710
2021	0.7011	0.6651	0.0360	0.8065
2022	0.6157	0.6270	0.0113	0.8387

Table 7 shows the weights of each base learner in the Adaboost model from 2020 to 2022,In the 2020–2022 experiments, the weight shifts among SVM, LR, and KNN models reflected varying relative importance each year, with SVM dominating in 2020, LR in 2021, and KNN in 2022, while LR dropped to zero weight in the final year, highlighting yearly model performance differences. Throughout this period, AdaBoost consistently demonstrated high accuracy, robustness, and low overfitting—maintaining strong generalization even as individual model contributions fluctuated—and overall outperformed each base learner in stability, accuracy, and overfitting resistance.

Table 7 Base Learner Weights

	KNN	SVM	LR
2020	0.2936	0.6938	0.0127
2021	0.2454	0.0328	0.7218
2022	0.6557	0.3443	0.0000

ISSN 2616-5902 Vol. 8, Issue 1: 147-154, DOI: 10.25236/AJBM.2026.080120

4. Conclusion and Outlook

4.1 Conclusion

This paper constructs a systematic classification and prediction model for the operation status of enterprises based on the input-output index data of 31 companies from 2020 to 2022. Firstly, Principal component analysis (PCA) was used to screen the original indicators. Although all the results were retained, by further analyzing the variance contribution rate of the principal components, a basis was provided for the subsequent screening of indicators. Then, the generalized efficiency was calculated based on the retained indicators, and combined with the grey correlation degree analysis, five indicators with relatively low correlations to the operational efficiency of the enterprise were successfully screened out. This process effectively improved the pertinence of the indicators and the interpretability of the model. On this basis, the base learner models such as SVM, KNN, and logistic regression were constructed respectively using the remaining indicators, and the parameters of each model were adjusted and optimized through methods such as cross-validation to ensure the performance of the base learner.

In order to further improve the classification accuracy and robustness of the model, this paper adopts the Adaboost algorithm to conduct ensemble learning on the above-mentioned base learners. The Adaboost algorithm combines multiple weak learners into one strong learner , giving full play to the advantages of each base learner and effectively balancing the complexity and generalization ability of the model. The final constructed enterprise operation status classification prediction platformcan provide strong support for the assessment and decision-making of enterprise operation status.

4.2 Outlook

The ideas and methods of this paper are innovative and practical to a certain extent. By screening indicators layer by layer and integrating multiple learning algorithms, the accuracy of classification and prediction of the operation status of enterprises has been effectively improved. However, the research also has some limitations, such as the relatively small sample size of the data and the need for further verification of the generalization ability of the model. Future research can consider expanding the sample size, introducing more enterprise data and indicators, and further optimizing the model structure and parameters to improve the universality and predictive ability of the model.

References

[1] Jin Y H. Enlightenment and suggestions for European and American governments to support the development of manufacturing industry[J]. Northern Economy. 2022;(2):27.
[2] Liu M,Zhu W W. Research on the spatiotemporal differentiation and countermeasures of the development quality of China's manufacturing industry[J].Journal of Chang'an University(Social Science Edition),2025,27(01):108.
[3] Ma X C. China's manufacturing industry orderly transfer of the implementation the path[J]. Modern Industrial Economy and Informationization. 2023;13(12):305.
[4] Liu M L,Tao H F. Research on the Development Issues of Single Champion Enterprises in China's Manufacturing Industry[J].China Journal of Commerce2022,(18):156
[5] Ma Z X. DEA model with generalized reference set and its properties[J]. Systems Engineering and Electronics. 2012;34(4):709.
[6] Zou Y, Tao S, Chu W H. Generalized DEA considering preference of inputs and outputs[J]. Mathematics in Practice and Theory. 2022;52(11):67.
[7] Zhang F, Wei Y J. Research on ecological sensitivity evaluation based on AHP-DEA weighting method: a case study of Shunde District[J]. Natural Resources Informatization. 2024(2).
[8] Jiang Y, Qiao Y Y. Financial Performance Evaluation of Logistics Enterprises Based on PCA-DEA Model[J].Logistics Sci-Tech,2024,47(05):44.
[9] Li G W, Qin JF, Huang J. Research on Multi-dimensional Quantitative of Product Selection Decision Making Method with Intelligent Based on KMeans[J].Computer Technology and Development, 2025, 35(03):194.
[10] Liu B Q, Lu Z F, Zhu W M, et al. Effective customer identification of Internet financial loan products based on Logistic DEA[J].Modernization of Management,2018,38(04):1.
[11] Ran M S, Zhou S, Huang L Y. Application of SVM Model Based on DEA Index in Financial Early Warning[J].Statistics & Decision,2009,(20):143.

ISSN 2616-5902 Vol. 8, Issue 1: 147-154, DOI: 10.25236/AJBM.2026.080120

[12] Cao Y, Miao Q G, Liu J-C, et al. Advance and Prospects of AdaBoost Algorithm[J]. Acta Automatica Sinica, 2013, 39(6): 745.
[13] Cheng Y, Peng J, Zhou Z, et al. A Hybrid DEA-Adaboost Model in Supplier Selection for Fuzzy Variable and Multiple Objectives[J]. IFAC-PapersOnLine, 2017, 50(1): 12255.
[14] Dang, M. J. Evaluation of the competitiveness of single champion enterprises in Hebei province. Hebei. Hebei University of Science & Technology,2020.
[15] Xu, D. D., Zeng, Z. B., & Dong, Y. Research on the realization path for classified reform of SOE's from the perspective of "efficiency evaluation": The case of high-end equipment manufacturing industry[J]. China Soft Science, 2017(07): 182.