Construction of an Intelligent Evaluation Model of Yield Risk Based on Empirical Probability Distribution

Zhou Yanru[1], Yang Jing[2]
Chaohu University, School of Mathematics and Big Data, Chaohu 238000, China[1]
Hefei Technology College, School of Architectural Engineering, Chaohu 238000, China[2]

***Abstract*—In order to improve the accuracy of yield risk evaluation, an intelligent evaluation model of yield risk based on empirical probability distribution is constructed. The dimensionality reduction method of risk factor based on principal component analysis is adopted. After adjusting the multiple data dimensions of risk factors that affect the rate of return to a unified dimension, the cluster-based evaluation index screening method is used to build the evaluation index set that best reflects the risk of the rate of return; The index weight vector equation method based on entropy weight and information entropy is used to set the evaluation index weight. Through the comprehensive evaluation model based on the empirical probability distribution of risk indicators, the empirical probability distribution information of risk indicators at all levels is analyzed, and the risk level of yield is intelligently evaluated. The research structure shows that the model can effectively evaluate the level of return risk and provide an effective reference for preventing and controlling investment return risk.**

Keywords*—Empirical probability distribution; yield; risk intelligence evaluation; principal component analysis; clustering; weight

I. INTRODUCTION

The characteristics of the investment itself, the complexity of the market environment and the risk management ability of investors will greatly affect the project's return on investment [1]–[3]. In order to obtain the highest return, investors should evaluate the investment risk of the product in the early stage. Because the risk always accompanies the return, and the high return must be accompanied by the high risk, so once the risk occurs, it may give investors a devastating blow [4], [5] The greater the risk of the activity is, the greater the loss of the final result if the decision is wrong, and vice versa. Risk cannot be completely avoided, but rational choice can minimize risk [6]. When conducting risk evaluation, investors should consider the source and use of funds. When examining the use of funds, that is, the investment of projects, they should also comprehensively consider the risks of financing and the overall market environment [7].

According to the analysis of the existing risk evaluation models, Authors used the NPV analysis model and @ risk software to assess the economic benefits and risks of China's carbon capture, utilization and storage projects in the context of carbon neutrality. Although the evaluation effect is effective, it is limited by the completeness of software functions. If the software is abnormal, whether the evaluation accuracy meets the standards remains to be tested [8]. Researchers built a risk evaluation model for overseas mining investment based on the structural power theory. Firstly, they built an evaluation index system of mining investment risk with the safety structure, production structure, financial structure and knowledge structure as the criterion level; then, the Topsis method and grey correlation analysis method were used to build a grey correlation risk evaluation model to complete the effective evaluation of overseas mining investment risk. However, this model does not analyze the problem of data dimension and indicator overlap, and the evaluation ability needs to be optimized [9]. Authors used the case analysis method to identify the investment risk of overseas railway construction projects and built a risk index system. Relevant methods establish the risk assessment model of overseas railway construction project investment. But the evaluation accuracy of this model is limited by the training effect of the neural network [10].

Although the above methods have made some progress, the accuracy of rate of return risk evaluation is low, and there are problems in evaluating the risk level of rate of return. Therefore this study focuses on the intelligent evaluation of yield risk. After reducing the dimension of risk factors, the clustering algorithm is introduced to cluster risk factors and build the optimal yield risk index system. On the basis of determining the weight of each index, empirical probability distribution theory is introduced to build an intelligent evaluation model of yield risk based on empirical probability distribution. In this model, the dimension reduction method of risk factors based on principal component analysis is adopted. After adjusting the data dimensions of various risk factors that affect the rate of return to a unified dimension, the evaluation index set that best reflects the rate of return risk is constructed by the evaluation index screening method based on clustering. The index weight vector formulation method based on entropy weight and information entropy is used to set the evaluation index weight. Through the comprehensive evaluation model based on the empirical probability distribution of risk indicators, the empirical probability distribution information of risk indicators at all levels is analyzed, and the risk level of return rate is intelligently evaluated. After the experimental test, the conclusions are as follows: (1) After reducing the dimension of the data of the influencing factors of yield risk,

the data dimension is obviously controlled in a unified range, which ensures the regularity of the data; The test results of cross-correlation coefficient of yield risk assessment results show that the cross-correlation coefficient is 1, and there is a significant correlation between the yield risk assessment results and the actual yield risk level; Compared with other models, the evaluation effect is the best, which can accurately evaluate the risk level of return, and the evaluation results are in line with the reality.

II. Intelligent Evaluation Model of Yield Risk

According to the classification of systematic risk and non-systematic risk, the classification of risk types of return on investment is shown in Fig. 1.

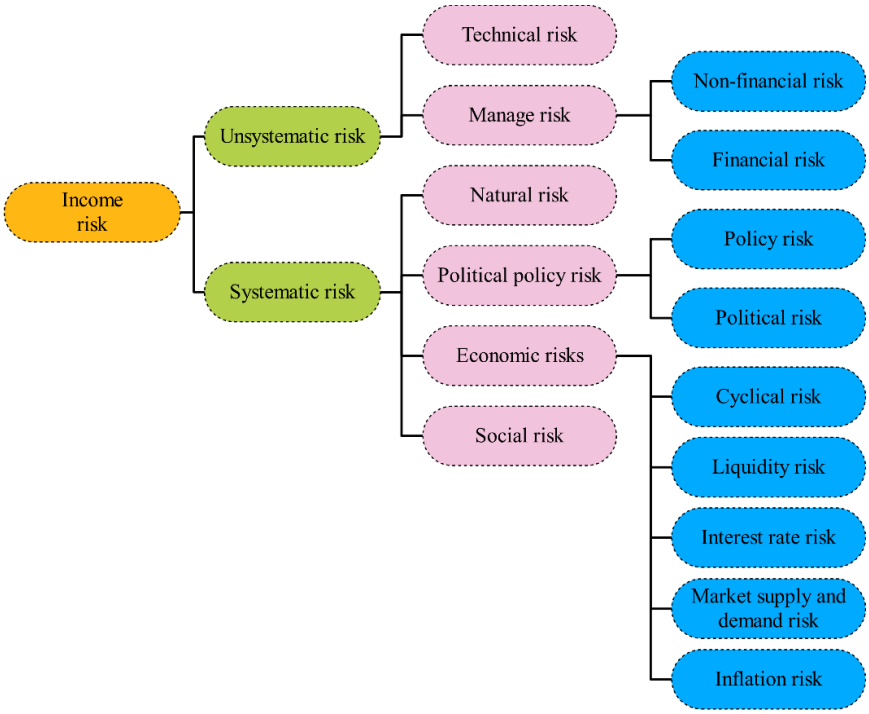


Fig. 1. Classification of risk categories of return on investment.

A. The Dimensionality Reduction Method of Risk Factor Based on Principal Component Analysis

According to the risk types described in Section II, many risk factors affect the rate of return. These risk factors are highly related to economic data, and the data dimensions are inevitably different [11], [12]. If such factors are directly used in the rate of return risk evaluation, it will increase the number of evaluation tasks. As a statistical analysis method for data dimensionality reduction, the principal component analysis method can transform multiple dimensions of the rate of return risk factors into a unified dimension on the premise of retaining most of the original information so as to improve the efficiency of data analysis [13], [14].

Principal component analysis (PCA) is a statistical analysis method with the main method of reducing data dimensions. On the premise of losing little original information, it transforms multiple influencing factors into several comprehensive factors (principal components) by calculating covariance, explains the internal structure of multiple influencing factors through a few principal components, catches the main contradictions and reduces the number of variables, and achieves the purpose of data compression and improving the efficiency of analysis [15], [16]. The main steps of the dimensionality reduction method for risk factors based on principal component analysis are as follows:

1) Standardize the data of yield risk factors. The risk factor data of the original yield is standardized to eliminate the impact of the yield risk factor data dimension and order of magnitude [17]. The standardization equation is:

$$a'_{ij} = \frac{a_{ij}-\beta_j}{R_j} \quad (1)$$

In the formula, a'_{ij} is the risk factor data of the standardized rate of return; a_{ij} is the original factor data; β_j and R_j are the sample mean and standard deviation of the jth yield risk factor; i=1,2,3,...,n, j=1,2,3,...,m, and are the number of samples and the number of factors.

2) Calculate the correlation coefficient matrix of standardized factor data [18]. The correlation coefficient matrix $S = (s_{ij})_{m\times m}$ is the m -order symmetric matrix, and the correlation coefficient s_{ij} represents the degree of correlation between the i -th factor and the j -th factor. The calculation equation of s_{ij} is:

$$s_{ij} = \frac{\sum_{h=1}^{n}(a_{hi}-\beta_i)(a_{hj}-\beta_j)}{\sqrt{\sum_{h=1}^{n}(a_{hi}-\beta_i)^2}\sqrt{\sum_{h=1}^{n}(a_{hj}-\beta_j)^2}} \quad (2)$$

In the formula, a_{hi} and a_{hj} are the h standardized data of the i -th and j -th risk factors respectively. β_i is the mean value of the second risk factor sample.

3) For the characteristic vector v_i of the yield risk factor, $\sum_{j=1}^{m} v_{ij}^2 = 1$ is required, where v_{ij} represents the j -th component of the characteristic vector v_i[19], [20].

4) Calculate the variance contribution rate and determine the principal component. Variance contribution rate F_i represents the proportion of variance of principal component G_i in the total variance, and its calculation equation is:

$$G_i = \frac{\varepsilon_i}{\sum_{i=1}^{m} \varepsilon_i} \times S_{ij} \tag{3}$$

Generally, the first h principal components of the cumulative contribution rate $\sum_{i=1}^{h} G_i \geq 85\%$ can contain most of the original factor information [21].

5) Calculate the principal component load (principal component coefficient matrix). The relationship between the principal component load value z_{ij} and the eigenvector v_i is:

$$\Omega_{ij} = \frac{v_{ij}}{\sqrt{\varepsilon_i}} \times G_i \tag{4}$$

6) Determine the dimension of risk evaluation factors of return rate [22]. Principal component load value Ω_{ij} and variance contribution rate F_i of the principal component jointly determine the final dimension of yield rate risk assessment factors. Then:

$$\Psi_j = \frac{\sum_{i=1}^{h} |\Omega_{ij}| G_i}{\sum_{j=1}^{h} \sum_{i=1}^{h} |\Omega_{ij}| G_i} \tag{5}$$

In the formula, Ψ_j is the dimension of the j -th yield risk factor.

B. Cluster-based Evaluation Index Screening Method

There is usually "coverage" and "overlap" of information between evaluation indicators. In order to obtain more accurate and objective evaluation results, this paper first determines and eliminates the indicators whose information is covered and then adjusts the weight of evaluation indicators according to the amount of information overlap.

Determining the evaluation index set usually includes two stages: rough selection and simplification of the evaluation index. The rough selection indicator set mainly considers the comprehensiveness of the evaluation indicators and is determined by the experts in the field and the evaluators through consultation; the selected indicator set mainly considers the representativeness of the evaluation indicators, which can be determined by statistical analysis methods such as correlation analysis based on the evaluation indicator value. There is no relevant quantitative method for the rough selection of evaluation indicators. Here, only a few main principles to be followed are given:

1) Purpose principle. To select evaluation indicators, it must first clarify the purpose of the evaluation. The evaluation indicators concerned vary with different purposes.

2) The principle of comprehensiveness. This principle needs to be followed in the rough selection stage of evaluation to ensure that the information contained in the evaluation index set reflects the effectiveness of weapons and equipment as fully and comprehensively as possible. As a result, there is also a phenomenon of coverage and overlap between evaluation indicators.

3) Principle of independence. This is a slightly conflicting principle with the principle of comprehensiveness. In the process of rough selection of evaluation indicators, it is necessary to compromise with the principle of comprehensiveness. It is to avoid the coverage and overlap of evaluation indicators as much as possible and ensure that certain characteristics of weapon equipment effectiveness will not be repeatedly reflected in multiple evaluation indicators.

4) Feasibility principle. The selected evaluation index must have a clear meaning, which can not only be understood and recognized by most people but also determines the evaluation index value based on sufficient and reliable data.

In this section, in the risk factor set after dimension reduction in Section II(A), the "very close" risk indicator factors can be determined through cluster analysis. At this time, these indicators can be considered to cover each other. Here, the distance between the yields risk evaluation indicators A_i and A_j is given as follows:

$$D_{ij} = 1 - \frac{\sum_{h=1}^{m} \left(A_i^h - \bar{A}_i\right)\left(A_j^h - \bar{A}_j\right)}{\sqrt{\left[\sum_{h=1}^{m} \left(A_i^h - \bar{A}_i\right)\left(A_j^h - \bar{A}_j\right)\right]^2}} \tag{6}$$

In the formula, D_{ij} represents the distance between indicators A_i and A_j; $\bar{A}_i = \sum_{h=1}^{m} A_i^h / m$, $\bar{A}_j = \sum_{h=1}^{m} A_j^h / m$; m is the number of indicators.

It is assumed that the risk evaluation index of return rate, class O_N, is obtained by combining class O_H and class O_Z, and the distance between it and class O_I is defined by the middle distance method as follows:

$$d_{nI} = \sqrt{\frac{1}{2} d_{HI}^2 + \frac{1}{2} d_{ZI}^2 - \frac{1}{4} d_{HZ}^2} \times D_{ij} \tag{7}$$

where, d_{nI} represents the distance between the class O_N and O_I; d_{HZ} represents the distance between classes O_H and O_I; d_{ZI} represents the distance between classes O_Z and O_I; d_{HI} represents the distance between classes O_H and O_Z.

If O_N and O_I only contain one yield risk evaluation index A_i and A_j, there are:

$$d_{nI} = D_{ij} \tag{8}$$

Further, the evaluation index screening method based on clustering is as follows:

1) Set the distance threshold σ between the evaluation indicators, and treat the yield risk evaluation indicator $A_1, A_2, \ldots, A_m$ as m different categories;

2) Calculate the distance between classes according to Eq. (6) to Eq. (8);

3) Determine whether the minimum distance $d_{nI} = d_{min}$ is less than σ;

4) If yes, merge classes O_H and O_Z into a new class O_N, and skip to step (2), otherwise execute step (5);

5) Output evaluation index class O_N.

Assume that O_N contains the set of yield risk evaluation indicators $\{l_1, l_2, \ldots, l_z\} \subseteq \{A_1, A_2, \ldots, A_m\}$, and z is the number of yield risk evaluation indicators included in O_N. At this point, it can be considered that $\{l_1, l_2, \ldots, l_z\}$ has basically covered each other, and it is further necessary to select the most representative yield risk evaluation index from them. Since the complex correlation number S_i reflects the degree of correlation between l_i and other indicators, the larger the value is, the higher the degree of coverage between l_i and other indicators is; that is, the more representative $l_i$is. The complex correlation coefficient between l_i and other indicators are as follows:

$$S_i = \sqrt{\frac{1}{z-1}\left(\sum_{j=1}^{1} s_{ij}^2 - 1\right)}, i \in z \tag{9}$$

where, s_{ij} is the correlation coefficient between indicators.

Further, the complex correlation coefficient corresponding to the representative index l_s of O_N is $S_s = \underset{1 \le i \le z}{max} S_i$.

After the reduction of yield risk indicators, the set of yield risk evaluation indicators obtained is $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_u\} \in \{A_1, A_2, \ldots, A_m\}$, u is the number of yield risk evaluation indicators after the reduction, and the set of indicators $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_u\}$ is the evaluation indicator system that best reflects the yield risk after the reduction.

C. Index Weight Vector Equation Method based on Entropy Weight and Information Entropy

1) Equation ion of weight vector in index layer considering data entropy weight: Entropy weight is an indicator to measure the degree of information provided by indicator data. By evaluating the degree of variation of data, we can measure the impact of the indicator data on the final rate of return risk evaluation results [23].

Firstly, the following steps are adopted to equation the weight vector between indicators within each indicator layer:

a) Through the expert scoring method, it can get the importance matrix $B(t) = \left(b_{ij}\right)_{m \times m}$ between the indicators in the k -th index layer of $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_u\}$, where $k = 1,2, \ldots, u$, b_{ij} is the importance between the i -th index and the j -th index. The maximum eigenvalue and eigenvector of $B(t)$ are calculated as shown in Eq. (10).

$$\gamma(k) \frac{1}{\hat{A}_u} \sum_{1}^{\hat{A}_u} \frac{(B(t) \cdot \varpi(k))}{b_i}_{max} \tag{10}$$

In the formula, $\varpi(k) = (\varpi_1, \varpi_2, \ldots, \varpi_u)^T$ is the approximate eigenvector of the k -th index layer, which is the initial weight. A consistency check is performed on $\gamma(k)_{max}$, and if it is satisfied, it can proceed to the next step; otherwise, return to step (1) to re-evaluate $B(t)$ [24].

b) The information entropy evaluation is carried out for the jth index of the k -th index layer in $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_u\}$. The evaluation method is as follows (11):

$$f(k,j) = \frac{-\sum_{z=1}^{Q(k,j)} q(k,j,z) \ln\left(2, q(k,j,z)\right)}{ln\left(Q(k,j)\right)} \tag{11}$$

In the formula, $f(k,j)$ is the information entropy of the j -th index in the k -th index layer; $Q(k,j)$ is the number of factor data of the j -th index of the k -th index layer; $q(k,j,z)$ is the satisfaction index of the z -th index factor data in the k -th index layer.

c) The improvement of $\varpi(k)$ taking into account information entropy is shown in Eq. (12):

$$\bar{\varpi}(k) = \varpi(k)^{T} f(k) \tag{12}$$

In the formula, $\bar{\varpi}(k)$ is the weight vector of the k -th index layer taking into account the information entropy; $f(k)$ is the information entropy vector of the k -th index layer; It can normalize the above equation:

$$\bar{\varpi}(k,j) = \frac{\bar{\varpi}(k,j)}{\sum_{j=1}^{m} \bar{\varpi}(k,j)} \tag{13}$$

In the formula, $\bar{\varpi}(k,j)$ is the weight of the j -th index in the k -th index layer after normalization.

2) Formulation of weight vector of criterion layer considering weight information entropy

However, the impact of different criteria levels on the final rate of return risk evaluation results is different, which cannot be reflected by the traditional expert scoring method and needs to be improved by using weight information entropy [25].

The formulation steps of the weight vector of the criterion layer considering the weight information entropy are as follows:

a) Based on the expert scoring method and the feature vector method [26], the weight vector between the criteria layers is obtained as $\varpi_{m \times m}$, and the specific method is the same as the weight vector between the indicators within the specified criteria layers.

b) The entropy weight is reflected in the criterion layer as follows: the weight information entropy of different evaluation indicators in the criterion layer, and the weight information entropy of the k -th criterion layer is calculated as shown in Eq. (14):

$$f'(k) = \frac{-\sum_{j=1}^{m(k)} \bar{\varpi}(k,j) \ln\left(2, \bar{\varpi}(k,j)\right)}{ln\left(m(k)\right)} \tag{14}$$

where, $f^{'}(k)$ is the weight information entropy of the k -th criterion layer; $m(k)$ is the number of indicators in the k -th criterion layer.

c) The weight vector $\varpi_{m \times m}$ between the criteria layers is improved by using the weight information entropy, as shown in equation (15).

$$\bar{\varpi}(k) = \varpi(k)^{T} f'(k) \tag{15}$$

where, $\bar{\varpi}(k)$ is the improved weight vector of the criterion layer. To sum up, the k -th criterion layer can be obtained, and the comprehensive weight of the j -th index is shown in Eq. (16):

$$\varpi(k,j) = \bar{\varpi}(k)f'(k) \quad (16)$$

D. Comprehensive Evaluation Model Based on the Empirical Probability Distribution of Risk Indicators

In order to comprehensively evaluate the return risk of each proposed investment project, we adopted an expert scoring method based on questionnaire. Considering that the rate of return risk index system consists of multiple groups of indicators, and each group of indicators contains different numbers of factors, a questionnaire scoring table is tailored for each project.

The design of the questionnaire carefully considered various risk indicators to ensure that quantitative and qualitative data were covered, so that experts could comprehensively and accurately assess risks. Specially invited 20 experts from the industry to participate in the grading. These experts have profound academic background and rich practical experience in related fields, and their grading has high authority and reference value. In the questionnaire, experts need to tick the corresponding level according to the data collected by the index system and their cognition of the project risk level. This process ensures that every expert can score according to a unified standard, thus increasing the objectivity and accuracy of scoring.

After this process, 20 answers from different experts will be obtained for each project. These answers provide valuable data to construct the empirical probability distribution of risk indicators. Through statistical analysis, we can understand the distribution of various risk indicators and the overall risk assessment of the project by experts. In order to ensure the validity and reliability of the questionnaire, a strict validity test is carried out. The content validity test is used to ensure that the questions and options in the questionnaire can fully reflect the rate of return risk index system. Secondly, the structural validity test is carried out, and the factor structure of the questionnaire data is analyzed to ensure that the measured results of the questionnaire are consistent with the expected risk factor structure.

1) Calculate the empirical probability distribution of risk indicators at each level:

$$q_{1ij}(Y_h) = \frac{1}{20}\varpi(k,j) \quad (17)$$

In the formula, $q_{1ij}(Y_h)$ refers to the probability distribution column of the risk indicators at the criterion level among the risk indicators at each level; Y_h is the risk level. Similarly, there are similar expressions for risk indicators at the indicator level, namely:

$$q_{2ij}(Y_h) = \frac{1}{20}\bar{\varpi}(k,j) \quad (18)$$

In the formula, $q_{2ij}(Y_h)$ represents the probability distribution column of the risk indicators at the indicator level among the risk indicators at each level.

Among the risk indicators at all levels, the empirical probability distribution of the j -th risk factor of the i -th risk indicator at the criterion level is:

$$G_{1ij}(y) = \sum_{h=1}^{5} q_{1ij}(Y_h) \quad (19)$$

At this time, the total empirical probability distribution of risk indicators at the indicator level can be expressed as:

$$G_{1i}(y) = \sum_{j=1}^{\bar{\varpi}(k,j)} G_{1ij}(y) \quad (21)$$

In the formula, $\bar{\varpi}(k,j)$ is the comprehensive weight of risk indicators at the indicator level.

2) Based on the empirical probability distribution of risk indicators at all levels, the risk level of the rate of return is evaluated:

$$Q^{(1)} = \left(q_{ij}(Y_1), q_{ij}(Y_2), q_{ij}(Y_3), q_{ij}(Y_4), q_{ij}(Y_5)\right) \quad (22)$$

In the formula, $q_{ij}(Y_1)$ is the empirical probability distribution column data of indicators in risk level 1. The total value of empirical probability distribution is 1. It is mainly used to judge the degree of risk level based on the proportion of empirical probability distribution of risk indicators in each risk level. In the issue of income risk evaluation, the risk level is mainly divided into five levels, namely, lower risk, low risk, medium risk, high risk and higher risk.

III. Experimental Analysis

In order to analyze the use effect of the model in this paper, the risk level of the return rate of two enterprises invested by a private equity fund is evaluated. A comprehensive data set is needed to train and verify the model, and the data set of yield risk is selected as the experimental data set, which is collected from publicly available financial databases, including historical yield data of financial markets such as stocks, bonds and futures. Collect GDP growth rate, inflation rate, interest rate, etc. from economic databases or official institutions, and collect income, profits, assets, etc. from company financial reports or public databases.

Data preprocessing is an important step to build the model. Firstly, data cleaning is carried out to remove the repeated, abnormal and wrong data points in the data set to ensure the accuracy and reliability of the data. Secondly, in order to eliminate the influence of dimension and range in feature data, feature normalization is carried out, so that different features can be compared and calculated fairly. Finally, the missing values are filled by interpolation or regression to ensure the integrity and continuity of the data. These pretreatment measures can improve the efficiency and accuracy of model training, and provide a solid data foundation for building an intelligent evaluation model of return risk based on empirical probability distribution. Table I is the information on yield risk indicators constructed by this model.

Table II shows the setting details of the weight of the yield risk evaluation indicators in this model:

The model in this paper analyzes the investment risk of two enterprise projects using the risk evaluation of the rate of return. Table III is the empirical probability distribution column of risk indicators of Enterprise 1 and Enterprise 2.

TABLE I. INFORMATION ON YIELD RISK INDICATORS

Criterion layer	Indicator layer	Factor
Financial risk	Solvency	Current ratio
		Quick ratio
		Asset-liability ratio
		Liabilities/EBIT
	Profitability	Net interest rate of equity
		Profit margin of main business
		Net asset interest rate
	Operational capacity	Total asset turnover
		Inventory cycle rate
		Accounts receivable turnover rate
	Cash payment ability	Free cash flow
		Cash flow debt ratio
		Sales cash ratio
	Growth ability	Sales growth rate
		Profit growth rate in the past two years
Non-financial risk	Industry risk	Political and economic fluctuation risk
		Policy and regulatory risks
		Industry life cycle risk
	Market risk	Sales risk
		Supply chain risk
		Market competition risk
		Market development risk
	Product risk	Property right risk
		Product substitution risk
		Technical environmental risk
		Product technical risk
		Product economic risk

TABLE II. WEIGHT SETTING DETAILS OF YIELD RISK EVALUATION INDICATORS

Criterion layer	Indicator layer	Weight	Factor	Weight
Financial risk	Solvency	0.140	Current ratio	0.273
			Quick ratio	0.296
			Asset-liability ratio	0.176
			Liabilities/EBIT	0.255
	Profitability	0.312	Net interest rate of equity	0.487
			Profit margin of main business	0.315
			Net asset interest rate	0.198
	Operational capacity	0.179	Total asset turnover	0.312
			Inventory cycle rate	0.302
			Accounts receivable turnover rate	0.386
	Cash payment ability	0.114	Free cash flow	0.4
			Cash flow debt ratio	0.302
			Sales cash ratio	0.298
	Growth ability	0.255	Sales growth rate	0.491
			Profit growth rate in the past two years	0.509
Non-financial risk	Industry risk	0.258	Political and economic fluctuation risk	0.322
			Policy and regulatory risks	0.24
			Industry life cycle risk	0.438
	Market risk	0.271	Sales risk	0.288
			Supply chain risk	0.227
			Market competition risk	0.251
			Market development risk	0.234
	Product risk	0.471	Property right risk	0.204
			Product substitution risk	0.194
			Technical environmental risk	0.154
			Product technical risk	0.253
			Product economic risk	0.195

TABLE III. THE EMPIRICAL PROBABILITY DISTRIBUTION OF RISK INDICATORS OF TWO ENTERPRISE PROJECTS

Risk level	Item 1	Item 2
1	0.450	0.173
2	0.347	0.290
3	0.151	0.263
4	0.052	0.143
5	0.000	0.131

As shown in Table III, the empirical probability distribution column value of Project 1 accounts for the largest proportion in risk level 1, followed by risk level 2, and 0.00 in risk level 1, while the empirical probability distribution column value of Project 2 accounts for the largest proportion in risk level 2, followed by risk level 3. There are empirical probability distribution columns in five risk levels, indicating that the risk is greater than Project 1.

Fig. 2 and Fig. 3 show the details of the data distribution dimensions before and after the dimensionality reduction of the data of the factors influencing the yield risk.

As shown in Fig. 2 and Fig. 3, the data dimension of the model in this paper is significantly different before the dimensionality reduction of the data of the factors influencing the yield risk. If such data is directly used in the intelligent risk evaluation, it will increase the difficulty. However, after reducing the dimension of the data of the factors influencing the yield risk in this model, the data dimension is obviously controlled in a unified range, ensuring the regularity of the data.

Whether the risk evaluation results are credible when the model is used to assess the project yield risk of two enterprises is tested, and the cross-correlation coefficient reflects the test results. The analysis method of correlation number Γ_{ij} is:

$$\Gamma_{ij} = \frac{cov\left(Q_i^{(1)}, Q_j^{(1)}\right)}{Q_i^{(1)} Q_j^{(1)}} \tag{22}$$

In the formula, $Q_i^{(1)}$ and $Q_j^{(1)}$ are the standard deviation between the risk level of the i -th and j -th project rate of return and the risk level of the actual rate of return. The value of Γ_{ij} is 0, indicating that there is an error in the yield risk evaluation result; If the value of Γ_{ij} is close to 1, it indicates that there is a significant correlation between the yield risk evaluation result and the actual yield risk level. Then the cross-correlation coefficient of the yield risk evaluation results of the model in this paper is shown in Fig. 4.

As shown in Fig. 4, the cross-correlation coefficient test results of the yield risk evaluation results of the model in this paper show that the cross-correlation number Γ_{ij} is 1, indicating that the yield risk evaluation results are significantly correlated with the actual yield risk level, and the evaluation results are reliable.

In order to highlight the mining effect of the model in this paper, it compares the model in reference [8], model in reference [9] and model in reference [10] to determine whether the model in this paper has application advantages. Fig. 5 shows the comparison test results of the four models.

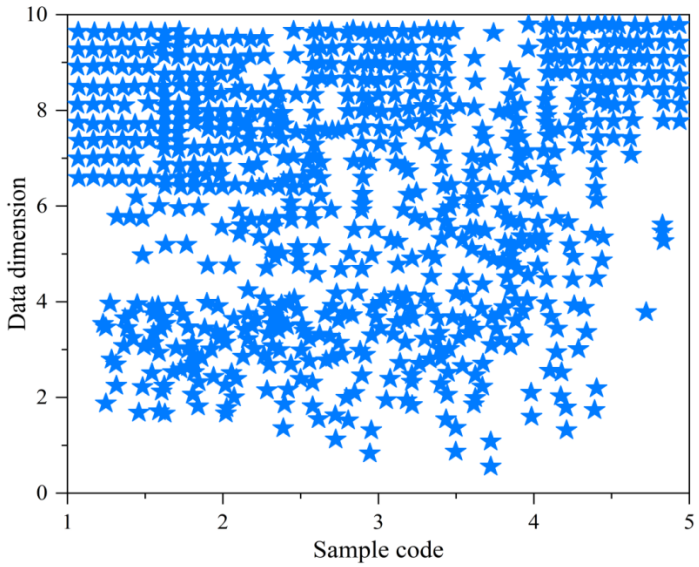


Fig. 2. Before dimensionality reduction of the data of yield risk influencing factors.

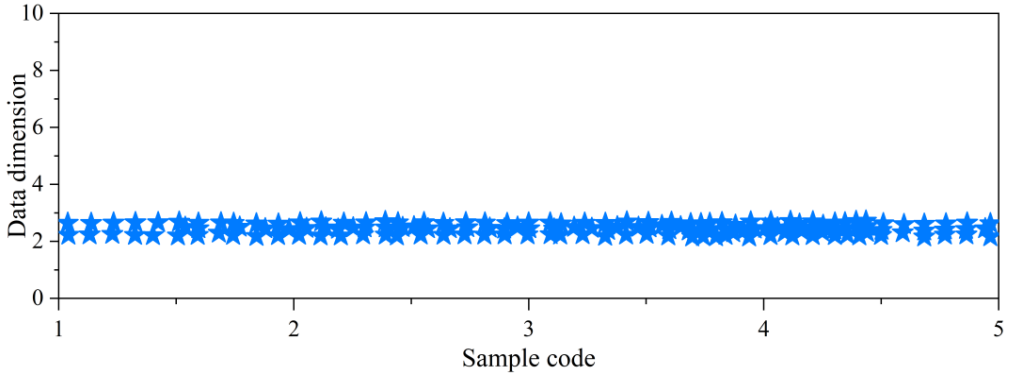
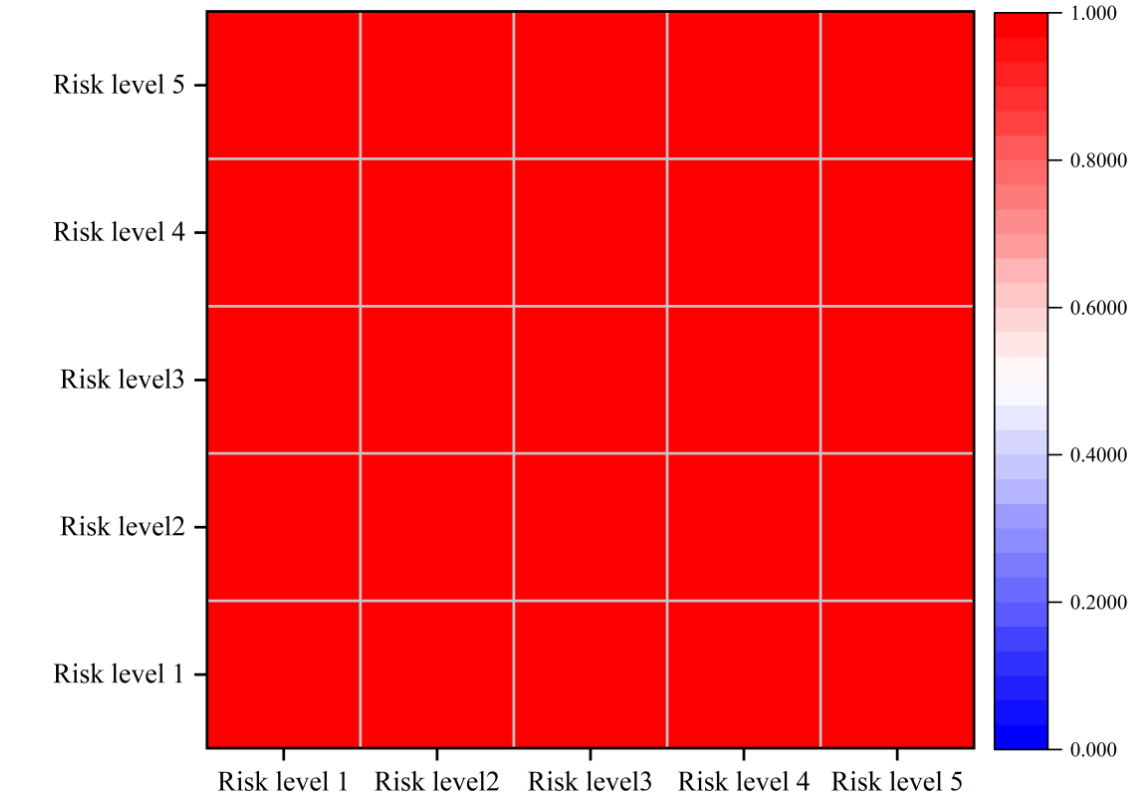
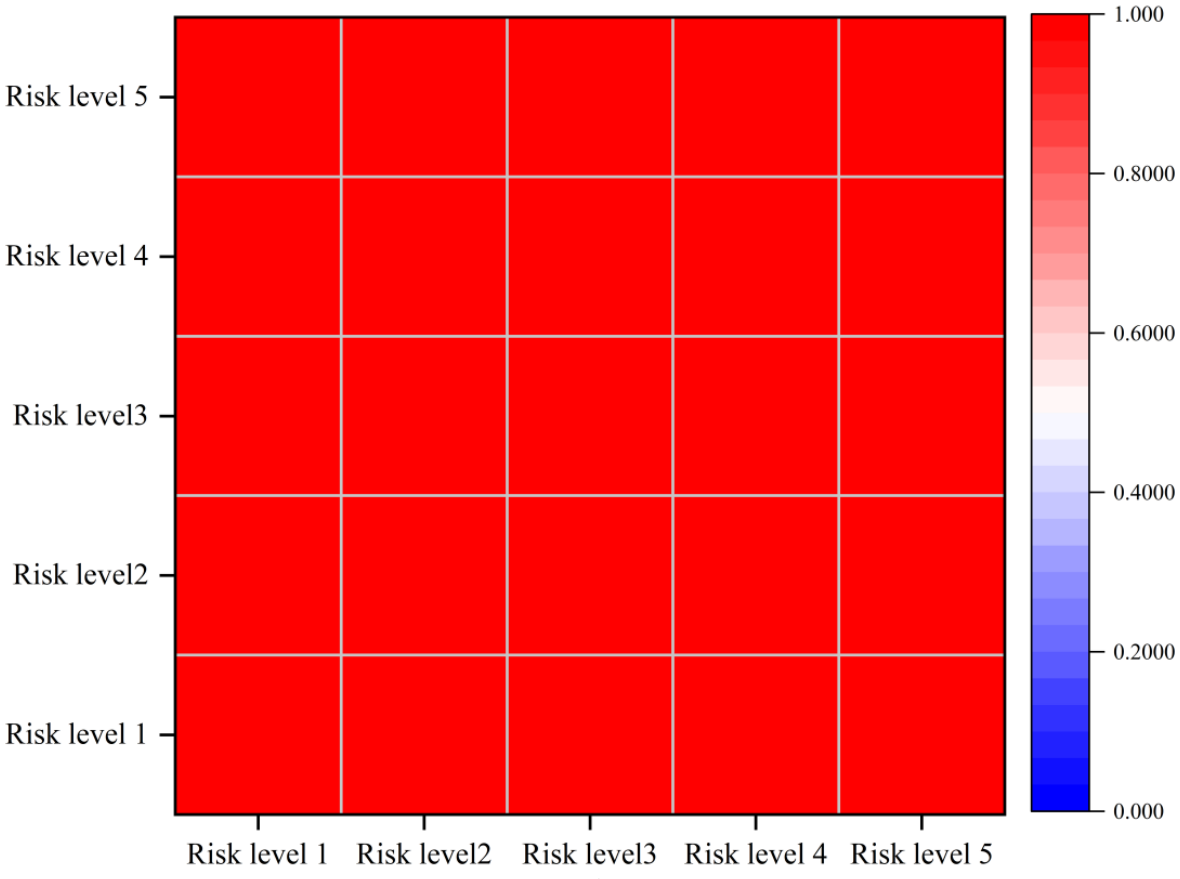


Fig. 3. After the dimensionality reduction of the data on the influencing factors of yield risk.



(a) Project of enterprise 1



(b) Project of enterprise 2

Fig. 4. The test results of the cross-correlation coefficient of the yield risk evaluation results of this model.

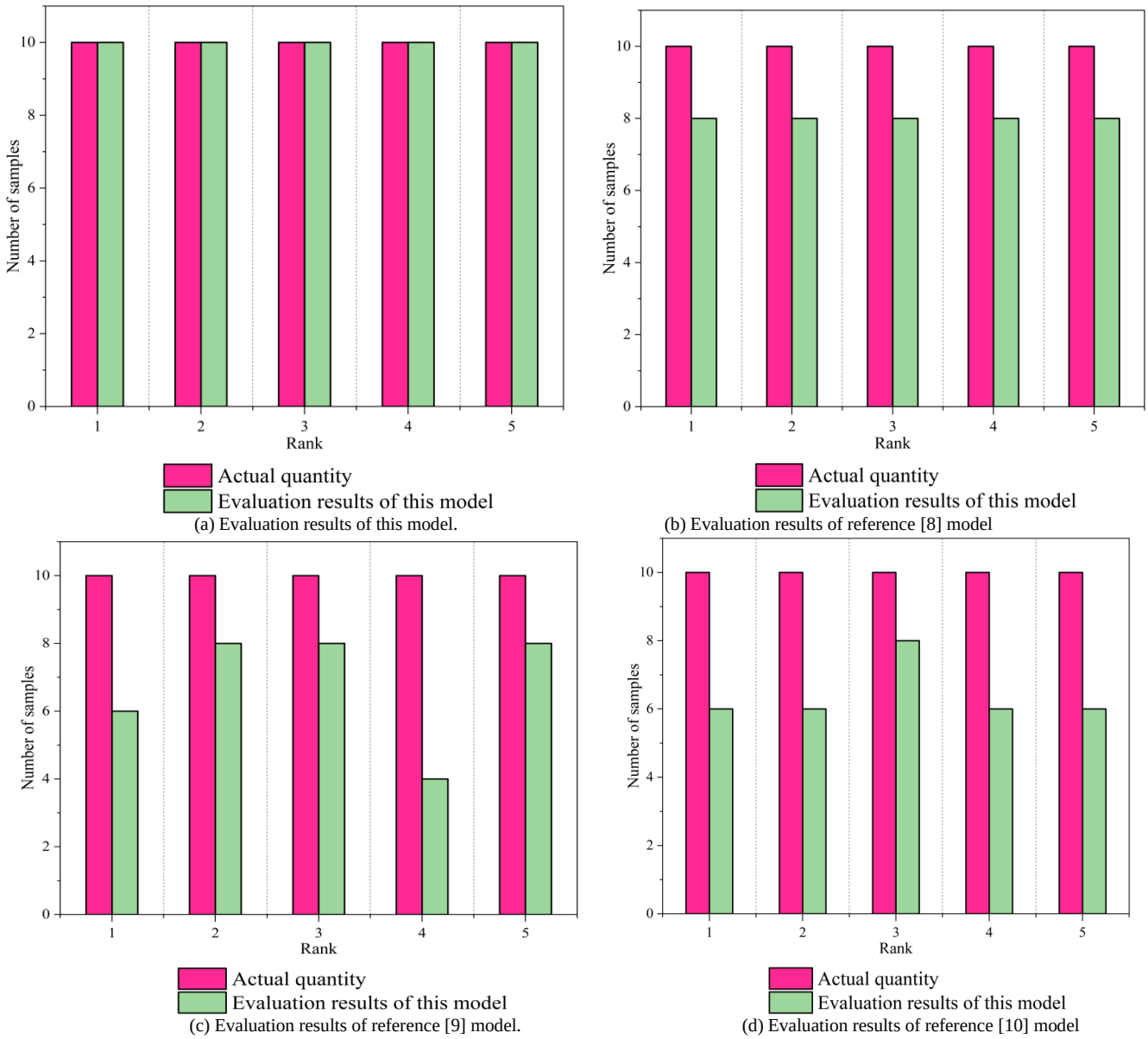


Fig. 5. Comparative test results of four models.

As shown in Fig. 5, after the model in this paper, the model in reference [8], the model in reference [9] and the model in reference [10] evaluate the same yield risk target, the evaluation effect of the model in this paper is the best, which can accurately evaluate the yield risk level, and the evaluation results are in line with the reality. However, the evaluation results of the model in reference [8], the model in reference [9], and the model in reference [10]are biased. The reason is that the model in this paper can reduce the dimension of risk indicators and screen indicators before evaluation so as to ensure the rationality of indicators.

IV. Conclusion

This paper constructs an intelligent evaluation model of yield risk based on empirical probability distribution, comprehensively analyzes the yield risk from multiple perspectives, analyzes it from financial and non-financial perspectives, constructs an intelligent evaluation index of yield risk, and designs a comprehensive evaluation model based on the empirical probability distribution of risk indicators. After in-depth performance testing in the experiment, the test conclusions are as follows:

1) After the dimensionality reduction of the data of the factors influencing the yield risk in the model of this paper, the data dimension is obviously controlled in a unified range, ensuring data regularity.

2) The cross-correlation coefficient test results of the yield risk evaluation results of the model in this paper show that the correlation number is 1, the yield risk evaluation results have a significant correlation with the actual yield risk level, and the evaluation results are reliable.

3) Compared with other models, the model in this paper has the best evaluation effect and can accurately evaluate the

risk level of the rate of return. The evaluation results are in line with reality.

For the future research work of the intelligent evaluation model of return risk based on empirical probability distribution, the following are some prospects:

1) Establish a more accurate model: Future research can explore a more accurate model and introduce market sentiment and macroeconomic indicators to improve the prediction accuracy of the model.

2) Consider the nonlinear relationship: In fact, there may be a nonlinear relationship between stock returns, and future research can explore how to consider the nonlinear relationship to better describe and predict the return risk.

3) Consider the change of market conditions: The change of market conditions is one of the important factors affecting the rate of return. Future research can explore how to adjust model parameters according to different market conditions and provide corresponding risk assessment.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing of interests.

AUTHORSHIP CONTRIBUTION STATEMENT

Yang Jing: Writing-Original draft preparation

Conceptualization, Supervision, Project administration.

Zhou Yanru: Conceptualization, Methodology, Supervision

AVAILABILITY OF DATA AND MATERIALS

On Request

REFERENCES

[1] K. M. Frias, D. L. Popovich, D. F. Duhan, and R. F. Lusch, "Perceived market risk in new ventures: A study of early-phase business angel investment screening," Journal of Macromarketing, vol. 40, no. 3, pp. 339–354, 2020.

[2] J. K. Hammitt and L. A. Robinson, "Introduction to special issue on risk assessment, economic evaluation, and decisions," Risk Analysis, vol. 41, no. 4, pp. 559–564, 2021.

[3] F. He, Y. Li, T. Xu, L. Yin, W. Zhang, and X. Zhang, "A data-analytics approach for risk evaluation in peer-to-peer lending platforms," IEEE Intell Syst, vol. 35, no. 3, pp. 85–95, 2020.

[4] E. F. Drabo et al., "A Social-Return-On-Investment Analysis Of Bon Secours Hospital's 'Housing For Health'Affordable Housing Program: Study evaluates the broader social, environmental, and economic benefits of Bon Secours Hospital's Housing for Health program.," Health Aff, vol. 40, no. 3, pp. 513–520, 2021.

[5] H. San Martín, B. Hernández, and Á. Herrero, "Social consciousness and perceived risk as drivers of crowdfunding as a socially responsible investment in tourism," J Travel Res, vol. 60, no. 1, pp. 16–30, 2021.

[6] D. Li, J. Bi, and M. Hu, "Alpha-robust mean-variance investment strategy for DC pension plan with uncertainty about jump-diffusion risk," RAIRO-Operations Research, vol. 55, pp. S2983–S2997, 2021.

[7] Q. Wu, Y. Gao, and Y. Sun, "Research on Probability Mean-Lower Semivariance-Entropy Portfolio Model with Background Risk," Math Probl Eng, vol. 2020, pp. 1–13, 2020.

[8] N. Wei, S. Liu, Z. Jiao, and X. Li, "A possible contribution of carbon capture, geological utilization, and storage in the Chinese crude steel industry for carbon neutrality," J Clean Prod, vol. 374, p. 133793, 2022.

[9] H. Shuifeng, D. Yating, and L. I. Hongdan, "Risk evaluation of China's overseas mining investment based on structural power theory," China Mining Magazine, vol. 30, no. 10, pp. 24–31, 2021.

[10] Y. Changwei, L. Zonghao, G. Xueyan, Y. Wenying, J. Jing, and Z. Liang, "Application of BP neural network model in risk evaluation of railway construction," Complexity, vol. 2019, 2019.

[11] R. Castro and J. Tapia, "Adding a social risk adjustment into the estimation of efficiency: the case of Chilean hospitals," Quality Management in Healthcare, vol. 30, no. 2, pp. 104–111, 2021.

[12] A. Garratt and I. Petrella, "Commodity prices and inflation risk," Journal of Applied Econometrics, vol. 37, no. 2, pp. 392–414, 2022.

[13] L. Lu, A. Gavin, F. J. Drummond, and L. Sharp, "Cumulative financial stress as a potential risk factor for cancer-related fatigue among prostate cancer survivors," Journal of Cancer Survivorship, vol. 15, pp. 1–13, 2021.

[14] J. Safitri, S. Suyanto, M. L. Taolin, and S. L. Prasilowati, "Inclusion of interest rate risk in credit risk on bank performance: Evidence in Indonesia," JRAP (Jurnal Riset Akuntansi dan Perpajakan), vol. 7, no. 01, pp. 13–26, 2020.

[15] W. Li et al., "Characteristic of five subpopulation leukocytes in single-cell levels based on partial principal component analysis coupled with Raman spectroscopy," Appl Spectrosc, vol. 74, no. 12, pp. 1463–1472, 2020.

[16] Y. Zhang et al., "Infrared image impulse noise suppression using tensor robust principal component analysis and truncated total variation," Appl Opt, vol. 60, no. 16, pp. 4916–4929, 2021.

[17] P. J. Atkins and M. Cummins, "Improved scalability and risk factor proxying with a two-step principal component analysis for multi-curve modelling," Eur J Oper Res, vol. 304, no. 3, pp. 1331–1348, 2023.

[18] F. B. Salling, N. Jeppesen, M. R. Sonne, J. H. Hattel, and L. P. Mikkelsen, "Individual fibre inclination segmentation from X-ray computed tomography using principal component analysis," J Compos Mater, vol. 56, no. 1, pp. 83–98, 2022.

[19] K. A. Asha, L. E. Hsu, A. Patyal, and H. M. Chen, "Improving the quality of FPGA RO-PUF by principal component analysis (PCA)," ACM J Emerg Technol Comput Syst, vol. 17, no. 3, p. 3442444, 2021.

[20] Z. Nie et al., "Using a single sensor for bridge condition monitoring via moving embedded principal component analysis," Struct Health Monit, vol. 20, no. 6, pp. 3123–3149, 2021.

[21] X. Chen, L. Wang, and Z. Huang, "Principal component analysis based dynamic fuzzy neural network for internal corrosion rate prediction of gas pipelines," Math Probl Eng, vol. 2020, pp. 1–9, 2020.

[22] T. Alharbi, "Pulse-shape discrimination of internal α-contamination in LaBr3: Ce detectors by using the principal component analysis," Journal of Instrumentation, vol. 15, no. 06, p. P06010, 2020.

[23] J. Zhang, "A study on mental health assessments of college students based on triangular fuzzy function and entropy weight method," Math Probl Eng, vol. 2021, pp. 1–8, 2021.

[24] J. Huan, D. Ma, W. Wang, X. Guo, Z. Wang, and L. Wu, "Safety-state evaluation model based on structural entropy weight–matter element extension method for ancient timber architecture," Advances in Structural Engineering, vol. 23, no. 6, pp. 1087–1097, 2020.

[25] P. Liu and Y. Li, "An improved failure mode and effect analysis method for multi-criteria group decision-making in green logistics risk assessment," Reliab Eng Syst Saf, vol. 215, p. 107826, 2021.

[26] M. Jia et al., "Network Optimization of CNT Yarn Sensor Based on NNIA Algorithm in Damage Monitoring of 3D Braided Composites," Materials, vol. 15, no. 23, p. 8534, 2022.