Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3022726, 11 pages
https://doi.org/10.1155/2022/3022726

Research Article

A Hybrid Approach for Identification of Deficiencies in Enterprise Internal Control

Zhihua Huang

School of Foreign Languages, Southwestern University of Finance and Economics, China

Correspondence should be addressed to Zhihua Huang; hzh@swufe.edu.cn

Received 8 June 2022; Accepted 30 July 2022; Published 9 September 2022

Academic Editor: Kapil Sharma

Enterprise management and internal control are used to prevent and control risks and promote enterprises to achieve development strategy. This paper adopts comprehensive multidisciplinary research method to study the basic theory of prediction of major defects in enterprise internal control. Firstly, this paper proposes the prediction index system and sample selection standard of internal control major defects. Totally, 630 listed company and 12 indicators are collected and then use the random forest classification method based on principal component analysis. The parameters of random forest are optimized by genetic algorithm. Finally, the prediction model of major internal control defects of listed companies is established. The experimental results show that the average score of PCA-RF model in TPR value reaches 85%, which is nearly 20% higher than the 65% of RF model, proving that the PCA and GA can significantly improve the classification accuracy of ST Company and has important practical significance. Therefore, the proposed method system can reasonably solve the prediction problem of major defects in internal control.

1. Introduction

An important internal mechanism in business management is the internal control that corrects mistakes, prevents fraud, and ensures the organization's ability to develop functions. Internal control has existed since ancient times. In the era of increasingly complex industrial organizational structure and more serious financial fraud, the importance of internal control is becoming more and more prominent.

The goal of internal control is to prevent and control risks and promote enterprises to achieve development strategy. Internal controls are always flawed because companies cannot predict all future risks and are constrained by management perceptions and cost-effectiveness principles. According to the signal transmission theory, internal control defects, especially major defects, are important signals to judge the existence of business risks of enterprises. Disclosing major defects of internal control will produce negative effects [1], so enterprises have insufficient endogenous motivation to disclose defect information. Under the requirement of national compulsory disclosure, the public internal control information gradually becomes the game product between technical operation and regulatory rules. Many concealment and insufficient disclosure phenomena occur frequently, such as the events of the long-life biological vaccine, the failure of Zhangzidao scallop, and the financial fraud of Yabet. As a result, the principal-agent mechanism, orderly flow of property rights, and effective allocation of resources in the social and economic system cannot run smoothly. The prediction of major defects in internal control is a process of identifying and warning possible major defects by using appropriate methods according to the financial or nonfinancial data and other relevant information disclosed by enterprises. It aims to reduce the information asymmetry between enterprises and shareholders, creditors, investors, and government regulatory departments and reduce transaction costs paid by stakeholders to collect true information or falsify disclosed information. Predictive research provides a feedforward control method based on cybernetics. By integrating the characteristic information of data mining and referring to the path of "taking expectation as the standard-measuring reality-comparing reality and expectationdetermining deviation-analyzing the cause of deviation," major defects of internal control are actively identified to guide future actions.

The research on internal control defect prediction is mainly carried out from three aspects: defect identification, defect influencing factors, and prediction methods. The identification results of internal control defects answer the question of whether there are defects in enterprises and which are major defects and can reflect the implementation effect of internal control construction of enterprises. However, China's internal control standards have not yet made a clear description of how to divide the defects of different levels, nor for financial reporting and nonfinancial reporting internal control to make a reasonable distinction, policy, and practice level generally exist lack of norms, vague concept, and insufficient attention. Scholars generally believe that enterprises with financial fraud and significant financial restatement have major defects in internal control. However, with the change of external operating environment and the reconstruction of internal organizational structure, factors affecting the existence of internal control defects range from accounting and auditing to organizational structure and business operations. It is further expanding in the macrodirection such as marketization level, legal environment, and government regulation, and nonfinancial factors are playing an increasingly significant role [2, 3]. Therefore, both financial and nonfinancial factors should be considered comprehensively in the construction of prediction models. In terms of prediction methods, the prediction accuracy achieved by logistic regression and discriminant analysis in existing studies is about 70%~80% [4], and the prediction performance still has great room for improvement. With the rapid introduction of artificial intelligence technology into public view, the major defect prediction method has been transformed from statistical measurement to machine learning. Prediction models based on machine learning, such as support vector machine, BP neural network, and integration algorithm, have stronger ability to learn empirical knowledge from data than traditional econometric models and can extract useful information from a large number of nonlinear, high-dimensional, and high-noise data to obtain better prediction effect. Some existing studies have shown that artificial intelligence algorithm can effectively analyze and predict the financial situation and other field [5, 6], management ability of enterprises [7], and risk management [8].

Scholars establish indicators to monitor internal control in terms of effectiveness, audit efficiency, and timeliness [9]. Some scholars have analyzed the decisive factors of major internal control defects [10] or tried to establish a model to distinguish major internal control defects by testing and discriminant methods [11]. Compared with traditional statistical methods such as univariate and multivariable discrimination and regression, neural network does not need to preset standardized function formulas and give hypotheses of statistical distribution characteristics of variables in the model, and it can be used for identification and prediction of variables and models changing over time [12]. At present, in addition to the aforementioned application of neural network in enterprise internal control, many scholars have used neural network methods to explore financial crisis, enterprise bankruptcy, financial market trend prediction [13–15], and bank performance evaluation [16]. However, the NN-based model is easy to overfit. Based on the above analysis, this paper tries to answer the following questions: how to establish a suitable index system for predicting major internal control defects of listed enterprises in China; by comparing different machine learning algorithms, the prediction model based on ensemble algorithm is better than the existing research. Which factors have a greater impact on the forecast results? The possible contributions of this paper are as follows: first, starting from the key elements of internal control and combining empirical evidence, the predictive index system is constructed from the four dimensions of internal governance mechanism, external environmental risk, financial status, supervision, and information communication. Second, machine learning algorithm is used to build prediction models, and the best prediction effect is found by comparing six models including logistic regression, support vector machine, decision tree, BP neural network, random forest, and XGBoost. Most existing literatures are based on linear regression method, with strong explanatory ability but weak prediction ability. This paper demonstrates the advantages of machine learning, especially integrated learning, in prediction performance, and tests the practicality of prediction index system. Third, further explore the improvement space of prediction model through feature contribution analysis.

The main contributions in this paper is as followings: (1) this paper constructed a prediction index system including four aspects, including the company's internal governance, risk, financial situation, and supervision. In order to quantitatively study the enterprise's internal control and management ability, this paper successfully constructed a new data set. (2) This paper successfully applied the machine learning algorithm to the traditional enterprise management field and verified the effectiveness and prediction effect of the algorithm. (3) The model in this paper can not only predict enterprise management risk but also find several important indicators that influence enterprise management risk.

2. Model Structure

RF is an ensemble model which was proposed by Breiman (2001) first. RF mainly employ the concept of bagging model and simple decision trees, to promote the generalization performance, RF use a bootstrap sampling algorithm and a CARTs algorithm to generate multiple unrelated decision trees [17]. Figure 1 indicates a structure of the RF algorithm.

The proposed model in this article is mainly divided into four parts: data processing, feature extraction, recognition, and evaluation. The structure of the proposed algorithm in this paper for prediction of deficiencies in enterprise internal control system is introduced in Figure 2. In Figure 2, China Stock Market & Accounting Research Database (CSMAR) is based on the academic research needs of Shenzhen CSMAR Data Technology Co., Ltd. Based on professional standards of authoritative databases such as CRSP, COMPUSTAT, TAQ, and THOMSON and combined with China's actual

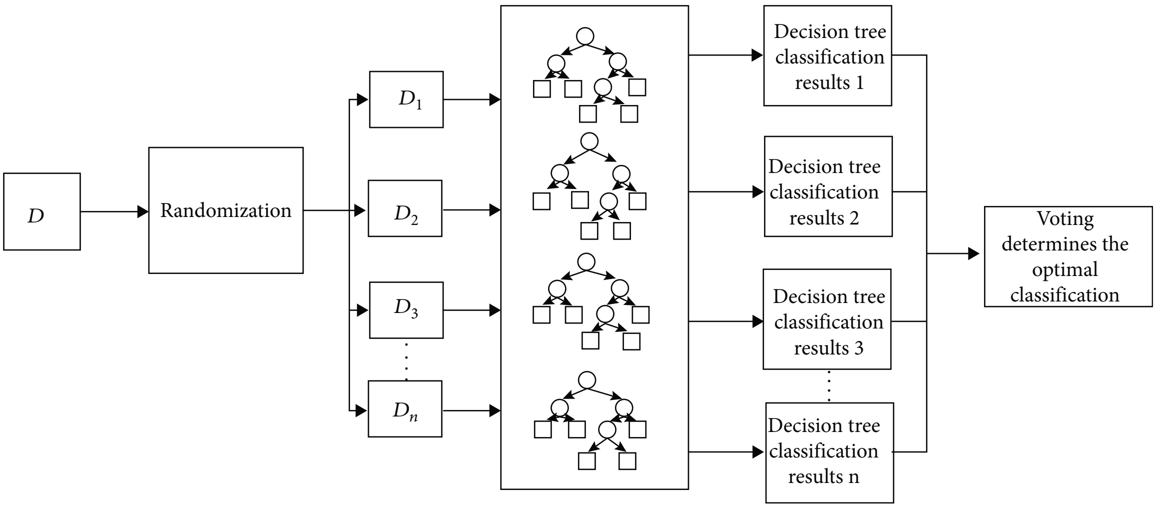
D
Randomization
D_1
D_2
D_3
D_n
Decision tree classification results 1
Decision tree classification results 2
Decision tree classification results 3
Decision tree classification results n
Voting determines the optimal classification

Figure 1: The structure of the RF algorithm.

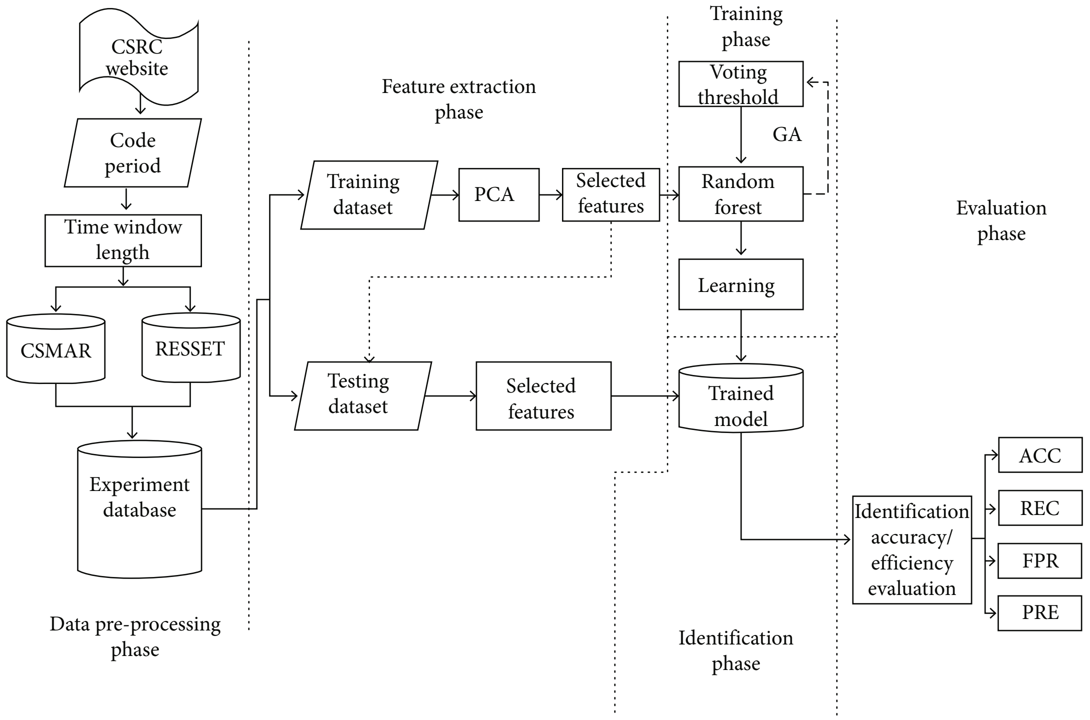

Figure 2: The structure of the PCA-RF algorithm for prediction of deficiencies in enterprise internal control system.

national conditions, the research-based accurate database in the economic and financial field is developed [18]; RESSET database is a professional data platform for model testing, investment research, and more [19].

In this paper, we also use some classical machine learning algorithms as benchmark models to verify the effectiveness of our proposed model.

3. The Theoretical Analysis

3.1. Identification and Determination of Internal Control Defects. Institutional theory holds that it is necessary to use rights and obligation arrangement or authority to limit the boundary of enterprises' pursuit of their goals and coordinate imperfect conflicts. The identification of internal control defects comes from the widely recognized "internal control-integration framework" and "enterprise risk management-integration framework" issued by COSO, which defines control defects as perceived, potential, or actual defects that will adversely affect the ability of enterprises to achieve goals. According to the degree of impact, the defects are classified into major defects, important defects, and general defects.

Based on the consideration of limited resources and costeffectiveness, internal control defects should be judged not only by the existence of defects or deficiencies in the control system but also by the extent to which such defects or deficiencies hinder the development of enterprises. Therefore, it is more urgent to identify major defects from the possibility and degree of deviation from the target. However, creditors, investors, and government regulatory departments mainly learn major defect information through the internal control evaluation report, internal control audit report, and financial report publicly disclosed by enterprises and passively receive the identification results.

In general, when judging whether an enterprise has major internal control defects, in addition to referring to the signs of major defects, reliable and objective identification results should be actively sought from other public channels, which is conducive to the subsequent model construction. The CSRC published since 2001, for example, the announcement of administrative punishment and markets in the disclosure of the listed company financial statement fraud, major guarantees, or significant related party transactions disclosure not according to stipulations or intentional omission, executives illegal behavior such as securities trading can be a conclusive that internal control is weak performance and sign complementary with other major defects.

3.2. Factors Affecting Internal Control Defects. Reference to the "internal control-integrated framework," "the listed company's internal control guidelines," "enterprise internal control basic norms," and form a complete set of guidelines, "publicly issued securities company information disclosure and reporting Rules No. 21-general provisions of the annual internal control evaluation report," and other documents issued marked the basic formation of China's internal control standard system. Enterprises on the basis of internal environment, risk assessment, control activities, information and communication, and internal supervision of the five elements build their own internal control system, covering the development strategy, corporate governance, organizational structure, corporate culture, social responsibility, financial activities, and many other factors, but the tedious internal control system is still unable to prevent fraud, corruption, and major risks, even leads to more defects in practice. Therefore, the construction of the prediction index system of major internal control defects should be based on the basic national conditions and enterprise needs, combined with existing empirical studies, and give priority to solving the urgent practical problems existing in internal control, rather than endlessly expanding the connotation and boundary of internal control and incorporating all relevant influencing factors.

Based on the above analysis and from the key elements of internal control, exploring the causes of defects along the wave source can guide the construction process of the prediction index system, in which the control environment is the basis, the control activities are the means of implementation, and the supervision is the dynamic feedback of the control activities, which is completed through information transmission and communication. Control environment is a general term for various factors that influence the establishment, strengthening or weakening of specific policies, procedures, and their efficiency. Internal control of any enterprise exists in a certain control environment, a good control environment can essentially enhance the execution effect of internal control. Control activities run through the entire organization, throughout all levels, business units, processes, and technical environment; the most basic control activity is transaction control. Financial performance to some extent reflects the closed-loop flow of important transactions, and the existence of defects can be detected from abnormal changes in financial data. The supervision process can timely evaluate the effectiveness of internal control and improve the defects found, which is the dynamic feedback of control activities, while information and communication can ensure that the feedback can be effectively communicated within the enterprise and between the enterprise and the external. This paper constructs the prediction index system of internal control major defects from three aspects: control environment, control activities, supervision, and information communication. Based on the importance of control environment, it is further divided into internal governance mechanism and external environmental risks.

3.3. Prediction Model Based on Machine Learning. Machine learning has been widely used in financial fraud identification, credit risk assessment, financial distress, and financial fraud identification, etc., providing a lot of help for making management decisions [20–23]. The prediction method of internal control defects has also been transformed from statistical learning to machine learning. Before the application of machine learning methods, relevant studies generally use econometric models to conduct causal analysis on variables, and these two methods have different trade-offs in model interpretability and prediction ability: econometric model focuses on explaining phenomena and finding the laws behind phenomena, requires to clarify the reasons for good model fitting and the interaction between variables, and pursues relatively simple function form and easily explained model estimation results. Machine learning is not limited to interpretability, it can learn more empirical knowledge from data, discover useful information in a large number of nonlinear, high-dimensional, and high-noise data, and flexibly select function forms to fit data, so it has strong predictive ability.

Among the mainstream machine learning algorithms, logistic regression has better fitting effect on linear relationship and is suitable for data with strong linear relationship between features and variables. Support vector machine (SVM) is a small sample learning method that adheres to mathematical principles. Based on kernel method, input data is mapped to a high-order vector space to solve classification problems, which requires high sample balance. The decision tree consists of a series of tree structures organized by "divide and conquer." The data are divided into different subsets according to different characteristics, and the information gain or Gini coefficient is used as the evaluation criteria. BP neural network is a kind of multilayer feedforward network trained according to the error back propagation algorithm, which constantly adjusts the weights and thresholds of the whole network through back propagation.

However, there are many parameters, it is difficult to train, and the output results are difficult to explain. Random forest is a forest composed of many unrelated decision trees in a random way, and the purity of data set divided by a feature is measured by information gain or Gini index, to determine the partition feature. XGBoost uses a greedy algorithm to enumerate all possible partitioning cases of features and then determines the optimal feature set based on which the final predicted value is the sum of predicted values of each base learner. It is worth noting that no algorithm can perfectly solve all problems, and all kinds of machine learning algorithms have their own good data sets, and the prediction model needs to be optimized through continuous practice.

4. Research Design

4.1. Data Sources and Sample Selection. Since 2007, listed companies in China have been required to disclose self-assessment reports of internal control, and the submission of internal control guidelines was voluntary disclosure at the initial stage. However, the proportion of disclosure increased year by year to 46% in 2011, and the proportion of disclosure is not high, and the disclosure of major defects of internal control is even less. In 2012, the internal control of listed companies in China entered the stage of comprehensive compulsory disclosure, and the proportion of disclosure in Shanghai Stock Exchange increased to 75.32%. In the initial stage of internal control information disclosure in China, there are mandatory disclosure and voluntary disclosure, and their disclosure motives are different, so they cannot be compared with each other. To sum up, this paper selects Chinese A-share listed companies (excluding the financial industry, companies subject to ST and companies with incomplete data) during the period of comprehensive compulsory disclosure from 2012 to 2020 as the sample basis. According to the selection criteria of major defect samples, 630 companies with 818 records of major defect samples are finally obtained. A total of 18,901 records were identified as significant defects in the control sample for eight years.

4.2. Definition of Variables

4.2.1. Dimension of Internal Governance Mechanism. Referring to relevant research results, internal governance mechanism can be measured by equity balance degree, institutional shareholding ratio, board size, proportion of independent directors, management power, and executive compensation. Specific indicators are listed as follows:

(1) Degree of equity balance (Tang and Xu [24])

The ratio of the sum of the shareholding of the second to tenth largest shareholders to the shareholding of the first largest shareholder.

(2) Institutional shareholding (Hermanson and Ye [25])

The ratio of the number of shares held by institutional investors to the total shares of listed companies.

(3) Board size (Hoitash and Bedard [2])

The natural log of the total number of board members.

(4) Proportion of independent directors (Krishnan [26])

Ratio of the number of independent directors to the total number of directors.

(5) Management power(Hermanson and Ye [25])

For general manager, the value is 1; for general manager and director, it is 2; for general manager and vice chairman, it is 3; for general manager and chairman, it is 4.

4.2.2. Dimension of External Environmental Risk. This paper selects the following indicators to measure:

(1) Degree of internationalization (Hammersley and Bedard [2])

Take 1 if you have overseas income; otherwise, take 0

(2) Listing years (Fethi and Pasiouras [16], Skaife et al. [27], Ge and Mcvay [28])

4.2.3. Dimension of Financial Status. This paper selects the following indicators to measure (Doyle et al. [29], Ge and Mcvay [28], Rice and Weber [30]):

(1) Z-score

Altman bankruptcy risk prediction model

(2) ROA

$$\frac{\text{Net profit}}{[(\text{ending balance of assets} + \text{beginning balance of assets})/2]} \quad (1)$$

(3) ROE

$$\frac{\text{Net profit}}{[(\text{ending value of net assets} + \text{initial value of net assets period})/2]} \quad (2)$$

4.2.4. Supervision and Information Communication. This paper selects the following indicators to measure:

(1) Changes in audit fees (Hoag and Hollingsworth [31])

$$\frac{(\text{Audit fee of current period} - \text{audit fee of previous period})}{\text{audit fee of previous period}} \quad (3)$$

(2) Whether the "top ten" (Chen et al. [32])

If it is one of the top ten audit institutions evaluated by CICPA, the value is 1; otherwise, the value is 0.

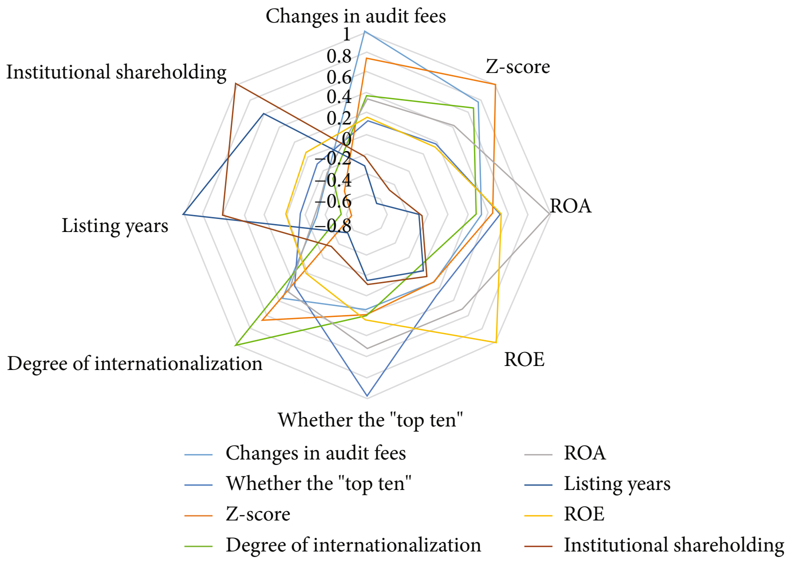

Figure 3: Correlation analysis of the indicators.

5. Model Process

5.1. Feature Engineering. Figure 3 is the correlation analysis diagram of an indicator, which shows the correlation between indicators of continuous value selected by us. We use this radar diagram to show it.

Figure 4 shows the classification results for internal control defects. In Figure 4, companies with internal control deficiencies are shown in red, and companies with no internal control deficiencies are shown in blue. The double-significance test can only test whether the differentiation of different indicators is significant between two groups of samples, but cannot solve the problem of multicollinearity between data, which often makes the model less accurate or even completely distorted. To address this problem, Principal component analysis (PCA) is adopted in this paper for feature reduction based on the double significance test. PCA is a linear combination transformation of the initial indicators to obtain uncorrelated principal components and achieves the purpose of variable screening and dimensionality reduction while retaining most of the original information.

For each dynamic data set obtained, principal component analysis was conducted based on double significance test. The 80% cumulative variance contribution rate was used as the screening criterion for the target principal components. Table 1 shows that we have obtained three main principal components after processing by principal component analysis algorithm (these three principal components contribute 86% of information content in total). After dimensionality reduction of principal component analysis, the final input index dimension of our model has changed from the original 12 indicators to the present 3 indicators.

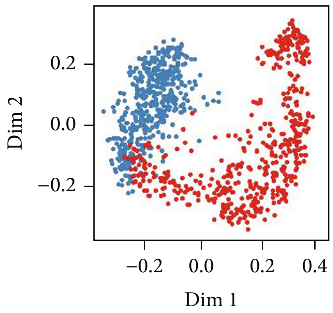

Figure 4: The classification results for internal control defects.

Table 1: Principal component rotation matrix.

Indicator	PC1	PC3	PC12
Degree of equity balance	0.048	0.009	0.023
Institutional shareholding	0.002	0.611	0.04
Board size	0.03	0.048	0.137
Proportion of independent directors	0.125	0.061	0.083
Management power	0.015	0.059	0.286
Degree of internationalization	0.12	0.004	0.056
Listing years	0.037	0.07	0.061
(1) *Z*-score	0.031	0.088	0.044
(2) ROA	0.426	0.054	0.124
(3) ROE	0.065	0.255	0.212
Changes in audit fees	0.004	0.261	0.053
Whether the “top ten”	0.061	0.097	0.056

5.2. Model Evaluation. The prediction results were divided into true positive, false positive, true negative, and false negative according to the combination of the true category and the prediction category, which were represented by the confusion matrix, as shown in Table 2.

In the verification process, after the input test data, the model will generate the prediction probability of a certain

Table 2: Confusion matrix.

The actual situation	Predicted results No significant defects are predicted	 No significant defects are predicted
Positive example (significant defect)	Real-positive example	False counterexample
Counterexamples (no significant defects)	False-positive example	True counterexample

Table 3: Result of the proposed method and benchmark method.

	Accuracy	FR rate	Precision	DR	$f1$
RF	88.92%	11.12%	88.96%	88.96%	88.96%
gbdt	85.77%	11.12%	88.84%	82.39%	85.51%
PCA-RF	87.86%	10.00%	85.45%	90.00%	87.67%
nn	84.18%	16.67%	85.24%	81.95%	83.57%
knn	85.50%	16.59%	87.25%	84.64%	85.93%
Bayes	85.81%	10.00%	81.30%	79.00%	85.45%
Logistic	85.40%	17.29%	88.72%	81.67%	85.06%
Tree	85.72%	17.07%	88.86%	82.55%	85.60%
Ctree	83.55%	19.90%	87.65%	79.25%	83.26%

label (such as the probability of an enterprise is predicted to have major defects), the test samples in descending order of the probability value, and the classification process is to find the threshold and "truncate" for two categories, greater than a threshold for one class, otherwise another class. If the threshold is large, truncate at the backward position, and if the threshold is small, the forward position can be used to check the positive examples. The ROC curve (receive-operating characteristic curve) and the AUC value (area under ROC curve) are the best measures of the model generalization performance from the above perspective. The abscissa FPR (false-positive rate) of the ROC curve represents the case of majority class error, and the ordinate TPR (true-positive rate) represents the model's ability of the model to capture minority classes, which can measure how the model misjudge majority classes when it tries to capture minority classes. The area enclosed by the ROC curve is the AUC value. The larger the AUC value is, the closer the ROC curve is to the upper left corner, and the better the prediction effect is.

In addition, the TPR and TNR are introduced to evaluate the predictive classification ability of the model for minority and majority samples, respectively. The AUC value is a general performance indicator of the classifier, which refers to the area under the ROC curve and the axis, and its value is not affected by the distribution of positive and negative samples in the data set. The F-value is a combination of recall and precision which are positively correlated with F-value. The G value is a combination of the TPR and TNR values, and the G value will get larger only when both TPR and TNR values are high. Both the F and G values take into account the classification of minority samples, so they can be used as evaluation criteria for the unbalanced financial warning model. Each indicator can be derived from the confusion matrix, and the specific formulas are shown in Equations (4)–(9), respectively.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}, \tag{4}$$

$$\mathrm{TPR}(\mathrm{Recall}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{5}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{6}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{7}$$

$$F = \frac{2 \times (\mathrm{Recall} \times \mathrm{Precision})}{\mathrm{Recall} + \mathrm{Precision}}, \tag{8}$$

$$G = \sqrt{\mathrm{TPR} \times \mathrm{TNR}} = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}}. \tag{9}$$

6. Analysis

The performance of prediction models with different algorithms in each data set is shown in Table 3, which preliminarily indicates the availability of the prediction index system. In comparison, the prediction effect of RF and PCA-RF ensemble models on data sets with different time spans is stronger than that of the other 7 individual learners, and the PCA-RF model is slightly better than the random forest model, and the prediction result is more robust, indicating that compared with the individual learners, the integration model represented by PCA-RF has better predictive performance and application value in predicting major defects of internal control in enterprises. One of the fundamental tenets of machine learning is that no algorithm can

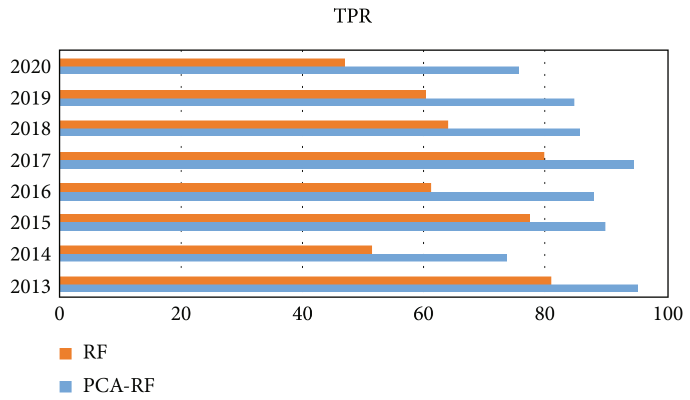

Figure 5: Result of TPR.

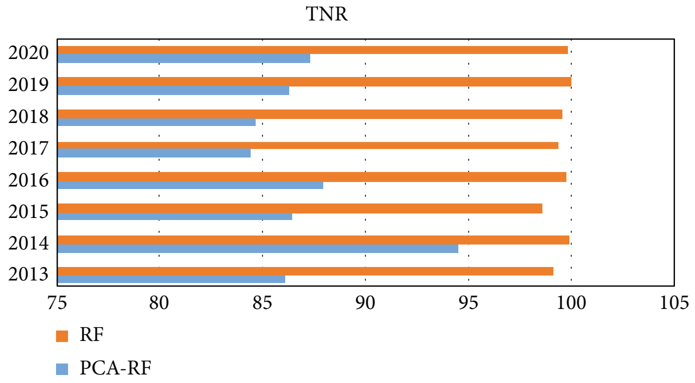

Figure 6: Result of TNR.

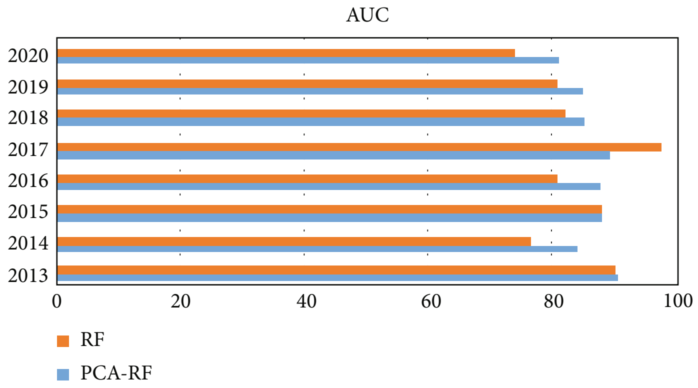

Figure 7: Result of AUC.

solve every problem perfectly, and every algorithm has a problem area in which it excels. The above results preliminarily indicate that the prediction index system of major internal control defects suitable for listed enterprises in China is established in this paper, and through comparison, the prediction effect of integrated model based on PCA-RF algorithm is the best.

6.1. Model Comparison Analysis. In order to verify whether the PCA-RF model can deal with the data imbalance in dynamic prediction, the RF model is used as a comparative model. In addition, since the RF is a strong classifier, a relatively stable test result can be obtained with the number of cycles $U = 20$. In each cycle, the optimal parameters of the RF models are obtained by the GA. After that, 50 iterations are performed to avoid the randomness interference on the model accuracy. The classification performance of the two models in each year is shown in Figures 5–10.

It can be seen that the average score of PCA-RF model in TPR value reaches 85%, which is nearly 20% higher than the 65% of RF model, proving that the former can significantly improve the classification accuracy of ST Company (a few

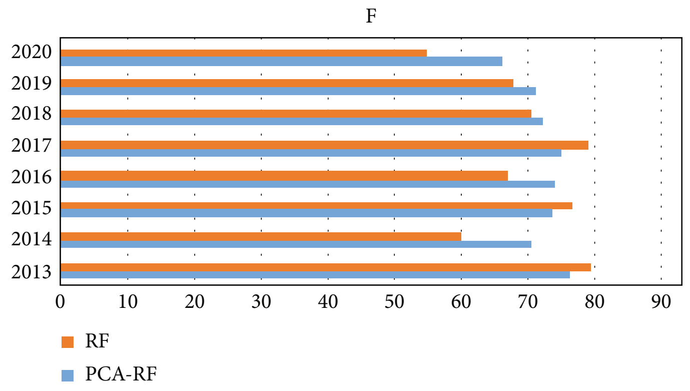

FIGURE 8: Result of *F*.

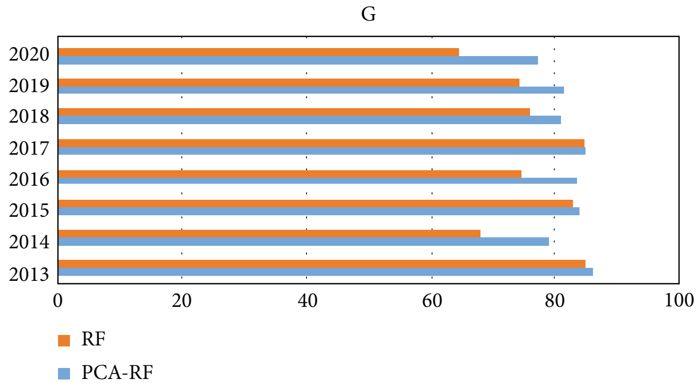

FIGURE 9: Result of *G*.

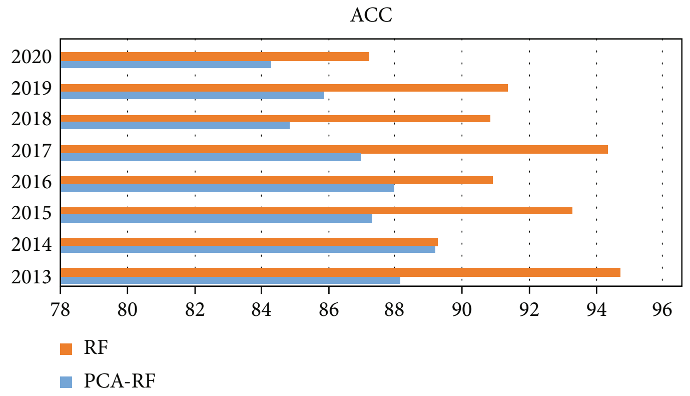

FIGURE 10: Result of ACC.

types of samples) and has important practical significance. In addition, the performance of PCA-RF model is superior to RF model in AUC and *G* value. Although the improvement in *F* value is limited, it can still maintain relatively stable, which proves the excellent comprehensive performance of the former.

7. Conclusion

This paper adopts comprehensive multidisciplinary research method to study the basic theory of prediction of major defects in enterprise internal control. Firstly, this paper proposes the prediction index system and sample selection standard of internal control major defects. Totally, 630 listed company and 12 indicators are collected and then use the random forest classification method based on principal component analysis. The parameters of random forest are optimized by genetic algorithm. Finally, the prediction model of major internal control defects of listed companies is established. The prediction model of major defects in internal control is designed to identify enterprises with major defects in internal control, so that they can get early warning hints in decision-making of management, investment, evaluation, and supervision and correct and improve the defects pertinently. Furthermore, from the perspective of the prediction index system, it is beneficial to identify the root causes of internal control defects, formulate solutions in time, and prevent the expansion of losses by looking for the unreasonable internal governance mechanism, serious external environmental risks, and abnormal financial conditions that may exist in enterprises. The main conclusions of this paper are as follows: (1) it is feasible to establish the prediction index system of major internal control defects. In this paper, the prediction index system of major defects of internal control is established from the four dimensions of internal governance mechanism, external environmental risk, financial status, supervision, and information communication, and 12 prediction features are screened out after feature engineering processing. The AUC values of prediction models based on different time spans and different algorithms are basically above 0.8. (2) PCA-RF model is slightly better than random forest model. Due to different algorithm principles, RF model can learn more feature information and have a wider range of feature selection.

Abbreviations

PCA:	Principal component analysis
SVM:	Support vector machine
ANN:	Artificial neural network
KNN:	*K* nearest neighbors
DT:	Decision tree
NB:	Naive Bayesian
CARTs:	Classification and regression trees
RF:	Random forest
REC:	Recall
FPR:	False-positive rate
PRE:	Precision
ACC:	Accuracy.

Data Availability

Data that can be provided by the corresponding author has no reserves.

Conflicts of Interest

The author declares that there is no conflict of interest in this work.

References

[1] J. S. Hammersley, L. A. Myers, and J. Zhou, "The failure to remediate previously disclosed material weaknesses in internal controls," *Auditing: A Journal of Practice and Theory*, vol. 31, no. 2, pp. 73–111, 2012.

[2] U. Hoitash and H. Bedard, "Corporate governance and internal control over financial reporting: a comparison of regulatory regimes," *Accounting Review*, vol. 84, no. 3, pp. 839–867, 2009.

[3] K. Johnstone, L. Chan, and K. H. Rupley, "Changes in corporate governance associated with the revelation of internal control material weaknesses and their subsequent remediation," *Contemporary Accounting Research*, vol. 28, no. 1, pp. 331–383, 2011.

[4] M. Franklin, *Sarbanes Oxley Section 404: Can Material Weakness be Predicted and Modeled? An Examination of the Variables of the ZETA Model in Prediction of Material Weakness*, Dissertations & Theses-Gradworks, 2007.

[5] G. C. Florea, "Financial statements forecast," *International conference Knowledge-Based Organization*, vol. 26, no. 2, pp. 19–22, 2020.

[6] L. Li, H. Li, G. Kou et al., "Dynamic camouflage characteristics of a thermal infrared film inspired by honeycomb structure," *Journal of Bionic Engineering*, vol. 19, no. 2, pp. 458–470, 2022.

[7] J. Yu, "Qualitative simulation algorithm for resource scheduling in enterprise management cloud mode," *Complexity*, vol. 2021, Article ID 6676908, 12 pages, 2021.

[8] S. Deng, C. Wang, M. Wang, and Z. Sun, "A gradient boosting decision tree approach for insider trading identification: an empirical model evaluation of China stock market," *Applied Soft Computing*, vol. 83, article 105652, 2019.

[9] A. Masli, G. F. Peters, V. J. Richardson, and J. M. Sanchez, "Examining the potential benefits of internal control monitoring technology," *The Accounting Review*, vol. 85, no. 3, pp. 1001–1034, 2010.

[10] J. Doyle, W. Ge, and S. McVay, "Determinants of weaknesses in internal control over financial reporting," *Journal of Accounting and Economics*, vol. 44, no. 1-2, pp. 193–223, 2007.

[11] J. B. Kim, B. Y. Song, and L. Zhang, "Internal control weakness and bank loan contracting: evidence from SOX section 404 disclosures," *The Accounting Review*, vol. 86, no. 4, pp. 1157–1188, 2011.

[12] K. C. Chung, S. S. Tan, and D. K. Holdsworth, "Insolvency prediction model using multivariate discriminant analysis and artificial neural network for the finance industry in New Zealand," *International Journal of Business and Management*, vol. 39, no. 1, pp. 19–28, 2008.

[13] F. Abid and A. Zouari, "Financial distress prediction using neural networks," in *Proceedings of the MS'2000 International Conference on Modeling and Simulation*, pp. 399–406, Spain, 2000.

[14] M. Perez, "Artificial neural networks and bankruptcy forecasting: a state of the art," *Neural Computing and Applications*, vol. 15, no. 2, pp. 154–163, 2006.

[15] F. J. López-Iturriaga, Ó. López-de-Foronda, and I. Pastor-Sanz, *Predicting Bankruptcy Using Neural Networks in the Current Financial Crisis: A Study of U.S. Commercial Banks*, Working Paper, University of Burgos, 2010.

[16] M. D. Fethi and F. Pasiouras, "Assessing bank efficiency and performance with operational research and artificial intelligence techniques: a survey," *European Journal of Operational Research*, vol. 204, no. 2, pp. 189–198, 2010.

[17] S. Deng, C. Wang, Z. Fu, and M. Wang, "An intelligent system for insider trading identification in Chinese security market," *Computational Economics*, vol. 57, no. 2, pp. 593–616, 2021.

[18] Resset databaseMay 2022, http://www.resset.cn/databases.

[19] China Stock Market & Accounting Research DatabaseMay 2022, https://www.gtarsc.com/.

[20] S. Y. Kim and A. Upneja, "Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models," *Economic Modelling*, vol. 36, pp. 354–362, 2014.

[21] C. C. Lin, A. A. Chiu, S. Y. Huang, and D. C. Yen, "Detecting the financial statement fraud: the analysis of the differences between data mining techniques and experts' judgments," *Knowledge-Based Systems*, vol. 89, pp. 459–470, 2015.

[22] N. Sanaz and S. Mehdi, "Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors," *Expert Systems with Applications*, vol. 110, pp. 381–392, 2018.

[23] S. M. Askari and M. A. Hussain, "IFDTC4.5: intuitionistic fuzzy logic based decision tree for E-transactional fraud detection," *Journal of Information Security and Applications*, vol. 52, article 102469, 2020.

[24] A. P. Tang and L. Xu, *Institutional Ownership, Internal Control Material Weakness and Firm Performance*, vol. 49, no. 2, 2007Social Science Electronic Publishing, 2007.

[25] D. R. Hermanson and Z. Ye, "Why do some accelerated filers with SOX section 404 material weaknesses provide early warning under section 302?," *Auditing: A Journal of Practice & Theory*, vol. 28, no. 2, pp. 247–271, 2009.

[26] J. Krishnan, "Audit committee quality and internal control: an empirical analysis," *Accounting Review*, vol. 80, no. 2, pp. 649–675, 2005.

[27] H. A. Skaife, D. W. Collins, and W. R. Kinney, "The discovery and reporting of internal control deficiencies prior to sox-mandated audits," *Journal of accounting and economics*, vol. 44, no. 1-2, pp. 166–192, 2007.

[28] W. Ge and S. E. Mcvay, "The disclosure of material weaknesses in internal control after the Sarbanes-Oxley act," *Accounting Horizons*, vol. 19, no. 3, pp. 137–158, 2005.

[29] J. T. Doyle, W. Ge, and S. Mcvay, "Accruals quality and internal control over financial reporting," *Accounting Review*, vol. 82, no. 5, pp. 1141–1170, 2007.

[30] S. C. Rice and D. P. Weber, "How effective is internal control reporting under SOX 404? Determinants of the (non-)disclosure of existing material weaknesses," *Journal of Accounting Research*, vol. 50, no. 3, pp. 811–843, 2012.

[31] M. L. Hoag and C. W. Hollingsworth, "An intertemporal analysis of audit fees and section 404 material weaknesses," *Auditing A Journal of Practice & Theory*, vol. 30, no. 2, pp. 173–200, 2011.

[32] Y. Chen, F. A. Gul, C. Truong, and M. Veeraraghavan, *Audit Quality and Internal Control Weakness: Evidence from SOX 404 Disclosures*, Electronic Publishing, 2012.