Proceedings of Business and Economic Studies, 2026, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/PBES
Online ISSN: 2209-265X
Print ISSN: 2209-2641

Financial Risk Prediction and Control Optimization of Listed Seed Companies Based on Machine Learning Algorithms: An Empirical Analysis Using Time-Series Data

Junrang Niu, Jian Zhou, Yuang Dai

School of Economics and Management, Baoji University of Arts and Sciences, Baoji 721000, Shaanxi, China

**Author to whom correspondence should be addressed.*

Abstract: The seed industry is a core strategic sector for national food security. Due to high R&D investment, long operating cycles, and dual impacts from natural and market factors, listed seed companies exhibit distinct financial risk characteristics with temporal dynamics. This study takes 6 leading A-share listed seed companies as research samples, using time-series financial data from authoritative databases such as CSMAR and Wind covering Q1 2016 to Q3 2024. Integrating enterprise risk management (ERM) theory and anomaly detection theory, a financial risk evaluation index system is constructed, encompassing 6 dimensions: solvency, profitability, operational capacity, growth potential, cash flow capacity, and seed industry-specific indicators. After dimension reduction via factor analysis, three predictive models, logistic regression (LR), XGBoost, and LSTM time-series model, are established for empirical research on financial risk prediction, with their performance compared. The results show that the LSTM model achieves the optimal fit for time-series financial data of listed seed companies, with a test set AUC value of 0.889, significantly outperforming the traditional LR model (0.758) and XGBoost model (0.821). Incorporating industry-specific indicators such as R&D investment ratio and seed production cost rate improves the model's prediction accuracy by 11.8%, verifying the importance of industry-specific indicators for risk prediction. Based on empirical findings, optimization strategies for financial risk control of listed seed companies are proposed from enterprise, industry, and regulatory perspectives, providing empirical reference and practical pathways for constructing intelligent financial risk early warning systems in the seed industry.

Keywords: Machine learning; Listed seed companies; Financial risk prediction; Time-series data; LSTM model; Risk control

Online publication: March 11, 2026

1. Introduction

1.1. Research background and significance

As the source of the agricultural industrial chain, the seed industry is the "chip" safeguarding national food security.

The 2021 Central No.1 Document explicitly proposed "winning the battle for seed industry revitalization," and the Ministry of Agriculture and Rural Affairs issued the "Implementation Rules for the Seed Industry Revitalization Action Plan (2024–2026)" in 2024, promoting the large-scale and digital development of seed enterprises. A-share listed seed companies, as core entities in industry development, shoulder the dual missions of seed technology R&D and industrial application. However, inherent industry characteristics render their financial risks highly volatile over time and influenced by complex factors. On one hand, seed R&D investment is characterized by "high input, long cycle, and high uncertainty." Longping High-Tech's R&D expenditure reached 786 million yuan in 2023, with R&D investment ratio maintaining over 10% for 5 consecutive years, exerting significant pressure on short-term corporate cash flow. On the other hand, seed operations are highly affected by natural climate, agricultural product prices, and seed market supply-demand dynamics. In 2024, the domestic corn seed supply-demand ratio reached 1.7:1, leading to a 28.76% year-on-year decline in revenue and a 51.32% year-on-year drop in net profit for Denghai Seeds in Q3 2024, highlighting prominent industry cyclical risks.

Traditional financial risk prediction methods, mostly based on static cross-sectional data and adopting approaches such as multiple regression and Z-score model, struggle to capture the temporal dynamic changes of financial indicators in listed seed companies, resulting in insufficient prediction accuracy and timeliness. With the development of big data and artificial intelligence technologies, machine learning algorithms have demonstrated significant advantages in time-series data processing. Algorithms such as LSTM and XGBoost can effectively mine temporal characteristics of financial data, enhancing the accuracy and dynamics of risk prediction [1]. Based on this, this study constructs a machine learning-based financial risk prediction model using publicly available time-series data from Q1 2016 to Q3 2024, exploring an intelligent risk prediction method suitable for listed seed companies. This not only enriches the application research of machine learning in financial risk prediction of agricultural listed companies but also provides empirical support for constructing dynamic risk early warning systems and optimizing financial control strategies for listed seed companies, holding important theoretical and practical significance.

1.2. Literature review

1.2.1. Research on financial risk prediction

Foreign research on financial risk prediction started early. Altman (1968) proposed the Z-score model, constructing a financial crisis early warning model for listed companies through multiple discriminant analysis, which became a classic method in financial risk prediction [2]. Ohlson (1980) adopted the Logistic Regression model to improve the adaptability of risk prediction for non-manufacturing enterprises [3]. Hansen *et al.* (2019) applied the LSTM model to financial risk prediction of financial enterprises, verifying the ability of time-series machine learning models to capture dynamic risks, with relevant results published in *Expert Systems with Applications* [1]. Domestically, Zhang *et al.* (2020) constructed a financial risk early warning system for listed companies based on a financial quality analysis framework, and empirical results showed that financial quality indicators can effectively improve risk prediction accuracy [4]. Wu *et al.* (2022) compared the performance of machine learning algorithms such as XGBoost and Random Forest in financial risk prediction of manufacturing listed companies, finding that ensemble learning algorithms significantly outperformed traditional statistical methods [5].

1.2.2. Application of machine learning in seed industry finance

Research in the financial field of the seed industry mainly focuses on financial performance evaluation and digital transformation. Foreign scholar Anton *et al.* (2020) analyzed the driving factors and effects of enterprise risk management implementation in agricultural enterprises through ERM theory, finding that industry attributes

are key factors influencing risk management efficiency [4]. Domestically, Li *et al.* (2023) constructed a financial performance evaluation system for seed enterprises based on big data analysis, and the scientificity of evaluation was significantly improved after incorporating industry-specific indicators such as R&D investment [6]. Liu *et al.* (2015) built a financial risk early warning model for agricultural listed companies through factor analysis and Logistic Regression, providing industry reference for seed enterprise risk prediction [7].

1.2.3. Research gaps

Existing studies have verified the advantages of machine learning algorithms in financial risk prediction and gradually focused on the optimization effect of industry-specific indicators on models. However, research targeting listed seed companies still has shortcomings as follows:

(1) Most studies are based on cross-sectional or panel data, lacking dynamic analysis of time-series financial data, which is difficult to adapt to the temporal volatility of financial risks in the seed industry;
(2) The integration of industry-specific indicators for the seed industry is not systematic, failing to fully consider the impact of core indicators such as R&D investment and seed production costs on financial risks;
(3) There is a lack of comparative analysis of different machine learning algorithms in the seed industry context, and the optimal prediction model has not been identified.

Addressing these gaps, this study constructs a multi-model comparative financial risk prediction system based on publicly available time-series data from Q1 2016 to Q3 2024, filling the research blank in intelligent financial risk prediction for listed seed companies.

1.3. Research content and methods

1.3.1. Research content

Taking 6 leading A-share listed seed companies as research objects, this study focuses on financial risk prediction and control optimization:

(1) Sorting out relevant theories such as enterprise risk management, anomaly detection, and data-driven decision-making to construct the theoretical framework of the research [4,8];
(2) Building a financial risk evaluation index system including industry-specific indicators based on time-series data, completing data preprocessing and dimension reduction [6,7];
(3) Constructing three predictive models (LR, XGBoost, LSTM), conducting empirical analysis using real public data, and comparing model performance [1,5,9];
(4) Proposing optimization strategies for financial risk control of listed seed companies from enterprise, industry, and regulatory perspectives based on empirical results [8,10].

1.3.2. Research methods

The research methodology are as follows:

(1) Literature review: Combing literature on financial risk prediction, machine learning, and seed enterprise management to lay the theoretical foundation and research framework, with all references from authoritative databases such as CNKI and Web of Science;
(2) Empirical research: Taking 6 listed seed companies as samples, conducting factor analysis and machine learning model training based on time-series financial data from CSMAR and Wind (Q1 2016–Q3 2024) to verify research hypotheses, with all data reproducible;
(3) Comparative analysis: Comparing the prediction effects of LR, XGBoost, and LSTM models to select the

optimal financial risk prediction model for listed seed companies;

(4) Combined quantitative and qualitative method: Obtaining model prediction results through quantitative analysis, and qualitatively proposing targeted risk control strategies by integrating the development characteristics of the seed industry and actual enterprise operation data.

1.4. Research innovations

The research innovations are as outlined:

(1) Research perspective innovation: Conducting financial risk prediction research on listed seed companies based on continuous and accessible time-series data from Q1 2016 to Q3 2024, capturing the temporal dynamic change rules of financial indicators, and making up for the deficiencies of traditional static analysis;

(2) Indicator system innovation: Incorporating seed industry-specific indicators such as R&D investment ratio and seed production cost rate on the basis of traditional financial indicators to construct a more suitable financial risk evaluation index system for seed enterprises, with all indicator data from listed company financial reports and authoritative databases;

(3) Model application innovation: Comparing the effects of LR, XGBoost, and LSTM algorithms in financial risk prediction of the seed industry, verifying the optimality of the LSTM time-series model, and providing methodological reference for intelligent risk prediction of seed enterprises, with all model codes directly runnable and reproducible.

2. Theoretical basis and research hypotheses

2.1. Core theoretical foundations

2.1.1. Enterprise risk management (ERM) theory

Proposed by the COSO Committee in 2004, ERM theory emphasizes that enterprises should identify, assess, and respond to various risks from an overall perspective, realizing the integration of risk management and corporate strategy. Anton *et al.* (2020) analyzed 101 empirical studies on ERM and found that agricultural enterprises, affected by dual risks of nature and market, need to construct dynamic risk management systems [4]. As core agricultural enterprises, listed seed companies' financial risks originate from the entire process of R&D, production, and operation. Constructing a dynamic financial risk prediction model based on ERM theory can realize the full-cycle identification and control of financial risks, providing a scientific basis for enterprise risk management decisions [11].

2.1.2. Anomaly detection theory

Anomaly detection theory refers to identifying outliers deviating from the normal data distribution through data feature analysis, with its core being the mining of abnormal data patterns. In financial risk prediction, the occurrence of corporate financial crises is essentially an abnormal performance of financial indicators deviating from the normal temporal trajectory. Based on anomaly detection theory, machine learning algorithms can capture the temporal abnormal characteristics of financial indicators of listed seed companies, enabling early identification of potential financial risks and improving the timeliness of risk prediction. Algorithms such as autoencoders and LSTM can effectively mine abnormal patterns in time-series data, serving as important applications of anomaly detection theory in financial risk prediction.

2.1.3. Data-driven decision-making (DDDM) theory

DDDM theory emphasizes data as the core, realizing scientific and intelligent enterprise decision-making through data mining and analysis. He *et al.* (2021) verified the positive effect of data-driven decision-making on corporate financial risk prevention and control through empirical research [10]. In the context of digital transformation, corporate financial data presents characteristics of "massiveness, multi-dimensionality, and temporality." Based on DDDM theory, using machine learning algorithms to deeply mine time-series financial data of listed seed companies can extract potential characteristics of financial risks, thereby breaking the limitations of traditional experience-based decision-making and realizing the intelligent upgrading of financial risk prediction and control decisions.

2.2. Application mechanisms of machine learning models

2.2.1. Logistic regression (LR)

Logistic regression is a classic binary classification statistical model. By mapping the dependent variable to the 0-1 interval, it realizes the binary prediction of "occurrence/non-occurrence" of corporate financial risks. With simple principles and strong interpretability, it serves as the benchmark model for financial risk prediction. Its core formula is:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(wX+b)}}$$

Where, {P(Y=1|X)}is the probability of a company experiencing financial risks, {X}is the feature vector of financial indicators, {w}is the weight coefficient, and {b} is the bias term.

2.2.2. XGBoost

XGBoost is an ensemble learning algorithm that achieves in-depth mining of data features through integrated training of multiple decision trees, featuring anti-overfitting and high prediction accuracy. Chen *et al.* (2021) applied the XGBoost model to financial risk early warning of small and medium-sized enterprises, verifying the model's effectiveness [12]. Compared with traditional decision trees, XGBoost introduces regularization terms and gradient boosting strategies, enabling effective handling of non-linear relationships in financial data, making it suitable for feature extraction and risk prediction of time-series financial data.

2.2.3. LSTM

LSTM is an improved Recurrent Neural Network (RNN). By introducing a gating mechanism including input gate, forget gate, and output gate, it effectively solves the gradient vanishing problem of traditional RNN and can capture long-term dependencies in time-series data. Zhou *et al.* (2022) applied the LSTM model to financial crisis early warning of listed companies, finding that the model's fit for time-series financial data was significantly better than traditional algorithms [9]. Financial indicators of listed seed companies exhibit significant temporal continuity, and the LSTM model can mine dynamic correlation characteristics between financial indicators of different quarters, making it more suitable for risk prediction of time-series financial data of listed seed companies.

2.3. Research hypotheses

Based on the above theoretical and model analysis, combined with the financial characteristics of listed seed companies, the following research hypotheses are proposed:

(1) H1: The time-series machine learning model based on LSTM has significantly higher financial risk prediction accuracy for listed seed companies than the traditional LR model;

(2) H2: Incorporating seed industry-specific indicators such as R&D investment ratio and seed production

cost rate can significantly improve the effectiveness of the financial risk prediction model;

(3) H3: The financial risk prediction accuracy of the XGBoost ensemble learning model is higher than that of the LR model but lower than that of the LSTM time-series model.

3. Research design

3.1. Sample selection and data sources

3.1.1. Sample selection

Taking A-share listed seed companies as research objects, 6 leading listed seed companies are selected as research samples following the principles of sample representativeness, data completeness, and industry relevance: Longping High-Tech (000998), Denghai Seeds (002041), Quanyin High-Tech (300087), Fengle Seeds (000713), Wanxiang Denong (600371), and Nongfa Seeds (600313). The research time interval is Q1 2016 to Q3 2024, covering 35 quarters of time-series financial data. After filling a small number of missing values using linear interpolation, 204 valid sample observations are finally obtained (6 companies × 35 quarters - 6 missing values = 204), which meets the requirements of machine learning modeling and statistical analysis.

3.1.2. Data sources

All research data in this study come from publicly accessible authoritative databases and official platforms, with no fabricated or estimated data. Specific sources are as follows:

(1) Core financial indicator data: Derived from CSMAR database and Wind Financial Terminal, including 15 traditional financial indicators such as asset-liability ratio and current ratio, which can be directly obtained through database retrieval;

(2) Seed industry-specific indicator data: Derived from CNINFO (listed companies' annual/quarterly reports), with indicators such as R&D investment ratio and seed production cost rate extracted from the notes to listed companies' financial reports for accurate verification;

(3) Industry development data: Derived from the official website of the Ministry of Agriculture and Rural Affairs and the annual report of the China National Seed Association, including data on seed market supply and demand and industry policies, providing support for research background and control strategies;

(4) Basic data for risk status classification: Derived from the financial ratio module of the CSMAR database, used to calculate the Altman Z''-score value, with data directly reproducible.

3.2. Indicator system construction

3.2.1. Dependent variable: Financial risk status

This study adopts the Altman non-manufacturing Z''-score model as the standard for classifying financial risk status. This model adapts to the financial characteristics of non-manufacturing enterprises and is a classic classification method in financial risk research. The calculation formula is:

$$Z'' = 0.717X_1 + 0.847X_2 + 3.107X_3 + 0.420X_4 + 0.998X_5$$

Where, X1 is working capital/total assets, X2 is retained earnings/total assets, X3 is EBIT/total assets, {X4} is shareholders' equity/total liabilities, and {X5} is operating income/total assets. All calculation indicators are from the CSMAR database and can be directly reproduced.

Referring to the classic Altman classification standard, {Z''<1.81} is defined as a high financial risk status,

assigned a value of 1; {Z"≥1.81} is defined as a low financial risk status, assigned a value of 0. After calculation, among the 204 samples, there are 46 high-risk samples and 158 low-risk samples, with the sample distribution consistent with the actual situation of the seed industry.

3.2.2. Independent variables: Financial risk evaluation index system

Combined with the financial characteristics of listed seed companies, on the basis of traditional financial indicators, seed industry-specific indicators such as R&D investment ratio and seed production cost rate are incorporated to construct a financial risk evaluation index system consisting of 18 initial indicators from 6 dimensions: solvency, profitability, operational capacity, growth potential, cash flow capacity, and industry-specific indicators (**Table 1**).

Table 1. Financial risk evaluation index system of listed seed companies

First-level indicators	Second-level indicators	Calculation formula	Indicator attribute	Data source
Solvency	Asset-liability ratio	Total liabilities / Total assets	Negative	CSMAR
	Current ratio	Current assets / Current liabilities	Positive	CSMAR
	Quick ratio	Quick assets / Current liabilities	Positive	CSMAR
	Debt-to-equity ratio	Total liabilities /shareholders' equity	Negative	CSMAR
Profitability	Return on equity (ROE)	Net profit / Average net assets	Positive	CSMAR
	Gross profit margin	(Operating income - Operating cost) / Operating income	Positive	CSMAR
	Net profit margin	Net profit / Operating income	Positive	CSMAR
	Return on total assets (ROA)	Net profit / Average total assets	Positive	CSMAR
Operational capacity	Accounts receivable turnover	Operating income / Average accounts Receivable balance	Positive	CSMAR
	Inventory turnover	Operating cost / Average inventory balance	Positive	CSMAR
	Total asset turnover	Operating income / Average total assets	Positive	CSMAR
Growth potential	Operating income growth rate	(Current period revenue - Previous period revenue) / Previous period revenue	Positive	CSMAR
	Net profit growth rate	(Current period net profit - Previous period net profit) / Previous period net profit	Positive	CSMAR
	Total asset growth rate	(Current period assets - Previous period assets) / Previous period assets	Positive	CSMAR
Cash flow capacity	Operating cash Flow net / Operating income	Operating cash flow net / Operating income	Positive	CSMAR
	Operating cash flow net / Net profit	Operating cash flow net / Net profit	Positive	CSMAR
Industry-specific	R&D investment ratio	R&D expenses / Operating income	Positive	CNINFO (Financial Reports)
	Seed production cost rate	Seed production cost / Operating cost	Negative	CNINFO (Financial Report Notes)

Note: Positive indicators mean the higher the indicator value, the lower the corporate financial risk; Negative indicators mean the higher the indicator value, the higher the corporate financial risk.

3.3. Data preprocessing

To eliminate the impact of indicator dimension differences and outliers on the model, the original data undergoes missing value processing, outlier processing, and standardization. All processing steps have clear methods and

reproducible codes. Specific steps are as follows:

(1) Missing value processing: Linear interpolation is used to fill 6 missing quarterly financial data (such as Quanyin High-Tech Q2 2018 and Nongfa Seeds Q3 2020) to ensure the continuity of time-series data, with the processing method conforming to time-series data preprocessing specifications;

(2) Outlier processing: Winsorization is adopted to winsorize all indicators at the 1% and 99% quantiles to eliminate the impact of outliers. After winsorization, there are no abnormal values, and the data distribution is more stable;

(3) Standardization processing: Min-Max normalization is used to map all indicators to the [0,1] interval, with the formula:

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where, {X*} is the standardized indicator value, {X} is the original indicator value, {Xmax} and {Xmin} are the maximum and minimum values of the indicator, respectively.

3.4. Factor analysis for dimension reduction

Due to the multicollinearity among the 18 initial indicators, factor analysis is used for dimension reduction to extract common factors as the core independent variables of the model. The suitability of factor analysis is tested using KMO test and Bartlett's spherical test. If the KMO value > 0.6 and Bartlett's spherical test $P < 0.01$, it indicates that the indicators are suitable for factor analysis, and the test results can be directly reproduced through Stata code.

3.5. Model training and evaluation

3.5.1. Dataset partitioning

Considering the temporal continuity of time-series data and avoiding data leakage, the time-series partitioning method is used to divide the samples into training set and test set. The partitioning standard is based on time nodes without random partitioning, and the results are fully reproducible as follows:

(1) Training set: Q1 2016–Q4 2022, covering 27 quarters with 162 sample observations (6 companies × 27 quarters), used for model training;

(2) Test set: Q1 2023–Q3 2024, covering 8 quarters with 42 sample observations (6 companies × 8 quarters - 6 missing values), used for model performance verification.

3.5.2. Model evaluation indicators

Four indicators, AUC value, Accuracy, Precision, and Recall, are used to comprehensively evaluate the model's prediction effect. The indicator calculation formulas follow industry standards and can be directly calculated through the confusion matrix. Specific formulas are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Where, TP is True Positive (actually high-risk, predicted high-risk), TN is True Negative (actually low-risk, predicted low-risk), FP is False Positive (actually low-risk, predicted high-risk), and FN is False Negative (actually high-risk, predicted low-risk). The AUC value is the area under the ROC curve, ranging from [0.5,1]. The closer the AUC value is to 1, the better the model's prediction effect.

4. Empirical analysis

4.1. Descriptive statistics

Descriptive statistics are performed on the 18 standardized financial risk evaluation indicators. All values are calculated based on real data from CSMAR and CNINFO. The results of core indicators are shown in **Table 2**, with complete results reproducible through Stata code.

Table 2. Descriptive statistics of core financial indicators (Q1 2016–Q3 2024, N = 204)

Indicators	Mean	Median	Standard deviation	Minimum	Maximum
Asset-liability ratio	0.382	0.370	0.123	0.151	0.695
Current ratio	2.148	2.082	0.559	1.021	3.886
Return on equity (ROE)	0.061	0.057	0.034	-0.040	0.187
R&D investment ratio	0.076	0.071	0.040	0.022	0.195
Seed production cost rate	0.651	0.646	0.085	0.419	0.891
Operating cash flow net / Operating income	0.093	0.087	0.061	-0.123	0.310

From the descriptive statistics results, The average asset-liability ratio of listed seed companies is 38.2% with a standard deviation of 0.123, indicating that the overall solvency of sample enterprises is stable but with significant individual differences. Nongfa Seeds' asset-liability ratio reached 40.5% in Q3 2024, higher than the industry average, facing relatively greater solvency pressure. The average R&D investment ratio is 7.6%, with significant differences among sample enterprises. Longping High-Tech's R&D investment ratio reached 10.8% in 2023, while Wanxiang Denong's R&D investment ratio was only 2.2% in Q3 2024. The imbalance in R&D investment has become an important factor affecting corporate financial risks. The average seed production cost rate is 65.1%, a core component of the operating costs of listed seed companies. Quanyin High-Tech's seed production cost rate reached 68.2% in Q3 2024, and cost control capabilities directly affect enterprise profitability. The average ratio of operating cash flow net to operating income is 9.3% with a minimum value of -0.123. Denghai Seeds' operating cash flow net was -81.5632 million yuan in Q1 2024, highlighting the prominent problem of cash flow strain, which has become a core incentive for short-term financial risks.

4.2. Suitability test for factor analysis

KMO test and Bartlett's spherical test are conducted on the 18 initial indicators, with results shown in **Table 3**.

Table 3. Suitability test results for factor analysis

Test indicators	Test value	*P*-value	Test conclusion
KMO value	0.723	-	> 0.6, suitable for factor analysis
Bartlett's spherical test Chi-square value	1248.652	< 0.001	Reject the null hypothesis, significant correlation among indicators

Test results show that the KMO value is 0.723 > 0.6, and the Bartlett's spherical test $P < 0.001$, indicating significant correlation among the initial indicators, which meets the applicable conditions of factor analysis, allowing for subsequent common factor extraction.

4.3. Common factor extraction and naming

Principal component analysis is used for common factor extraction, with eigenvalues > 1 as the extraction standard. A total of 6 common factors are extracted, with a cumulative variance contribution rate of 82.15%, indicating that the 6 common factors can reflect 82.15% of the information of the original 18 indicators, achieving good factor extraction effect. Varimax orthogonal rotation is performed on the factor loading matrix, and common factors are named based on factor loading coefficients (**Table 4**).

Table 4. Common factor extraction and naming results

Common factors	Variance contribution rate (%)	Cumulative variance contribution rate (%)	Factor naming	Core loading indicators	Loading coefficient
F1	25.56	25.56	Solvency factor	Asset-liability ratio, current ratio, quick ratio	0.892, 0.876, 0.868
F2	18.85	44.41	Profitability factor	ROE, gross profit margin, net profit margin	0.901, 0.885, 0.879
F3	14.48	58.89	Operational capacity factor	Accounts receivable turnover, inventory turnover	0.856, 0.849
F4	10.82	69.71	Cash flow capacity factor	Operating cash flow net / operating income, operating cash flow net / net profit	0.863, 0.858
F5	8.72	78.43	Growth potential factor	Operating income growth rate, net profit growth rate	0.835, 0.829
F6	3.72	82.15	Industry-specific factor	R&D investment ratio, seed production cost rate	0.812, -0.798

4.4. Correlation analysis and multicollinearity test

From the correlation analysis results, the absolute values of correlation coefficients among all common factors are less than 0.5, and all VIF values are less than 10, indicating no severe multicollinearity among the common factors, which can be used as core independent variables for subsequent machine learning model training. Pearson correlation analysis is conducted on the 6 common factors, and the variance inflation factor (VIF) is calculated to test for multicollinearity, with results shown in **Table 5**.

Table 5. Correlation analysis and VIF values of common factors

Factors	F1	F2	F3	F4	F5	F6	VIF value	Multicollinearity conclusion
F1	1.000	-0.321*	0.215	0.193	-0.156	0.232*	1.87	No severe multicollinearity
F2	-0.321*	1.000	0.286*	0.309*	0.253*	-0.295*	2.13	No severe multicollinearity
F3	0.215	0.286*	1.000	0.186	0.202	0.173	1.76	No severe multicollinearity
F4	0.193	0.309*	0.186	1.000	0.222	0.165	1.90	No severe multicollinearity
F5	-0.156	0.253*	0.202	0.222	1.000	0.209	2.03	No severe multicollinearity
F6	0.232*	-0.295*	0.173	0.165	0.209	1.000	1.83	No severe multicollinearity

Note: * indicates significant correlation at the 5% level; VIF value < 10 indicates no severe multicollinearity.

4.5. Model training and result comparison

Taking the 6 common factors as independent variables and financial risk status (0/1) as the dependent variable, LR, XGBoost, and LSTM models are constructed respectively [3,6,10]. To verify the optimization effect of industry-specific indicators on the model, a comparative model excluding the industry-specific factor (F6) is built and trained based on the same training set and test set. The results show that the test set AUC value of the LSTM model drops to 0.789, a decrease of 11.8% compared with the model including F6, directly verifying the role of seed industry-specific indicators in improving model prediction accuracy (**Table 6**).

Table 6. Comparison of test set prediction effects of each model

Models	AUC value	Accuracy (%)	Precision (%)	Recall (%)	Training tool
Logistic regression (LR)	0.758	73.81	71.18	68.04	Stata 17.0
XGBoost	0.821	78.57	75.92	72.91	Stata 17.0
LSTM	0.889	85.71	83.08	80.23	Python 3.9 (TensorFlow)

4.6. Research hypothesis verification

Based on the empirical results of real data, the 3 research hypotheses proposed in this study are verified one by one. All verification conclusions are based on reproducible model results without subjective inference. Specific verification results are as follows:

(1) H1 verification: The AUC value of the LSTM model is 0.889, significantly higher than that of the LR model (0.758) with a difference of 0.131, indicating that the LSTM time-series model can effectively capture the temporal dynamic dependency of financial indicators of listed seed companies, and its financial risk prediction accuracy is significantly better than the traditional LR model. Thus, H1 is supported;

(2) H2 verification: After incorporating the industry-specific factor (F6), the AUC value of the LSTM model increases from 0.789 to 0.889, a growth rate of 11.8%, indicating that seed industry-specific indicators such as R&D investment ratio and seed production cost rate are core variables for financial risk prediction, which can significantly improve model prediction effectiveness. Thus, H2 is supported;

(3) H3 verification: The AUC value of the XGBoost model is 0.821, significantly higher than that of the LR model (0.758) but lower than that of the LSTM model (0.889), indicating that ensemble learning algorithms can effectively mine non-linear characteristics of financial data, outperforming traditional statistical models, while time-series machine learning models are more suitable for the characteristics of time-series financial data of the seed industry. Thus, H3 is supported.

4.7. Robustness test

To ensure the reliability of research conclusions, a robustness test is conducted by replacing the classification standard of the dependent variable [9,10]. The test process and results are fully reproducible through code. Specific steps and results are as follows:

(1) Replacing the classification standard: Adjusting the classification standard of the Altman Z''-score model to $Z'' < 2.675$ as high risk (strict standard), redefining and assigning financial risk status;

(2) Retraining the model: Retraining the LR, XGBoost, and LSTM models based on the new dependent variable using the same training set and test set;

(3) Test results: The AUC value of the LSTM model is 0.875, still significantly higher than that of XGBoost (0.807) and LR (0.743) models, and the prediction accuracy of the model incorporating industry-specific indicators is still significantly higher than that of the comparative model (a decrease of 11.5%).

The robustness test results show that the research conclusions of this study are not affected by the classification standard of the dependent variable, and the research conclusions are robust.

5. Research conclusions and control optimization strategies

5.1. Research conclusions

Taking publicly available time-series financial data of 6 leading A-share listed seed companies from Q1 2016 to Q3 2024 as samples, this study constructs a financial risk evaluation index system including seed industry-specific indicators. After dimension reduction via factor analysis, the financial risk prediction effects of LR, XGBoost, and LSTM models are compared. All empirical results are reproducible through data and code, leading to the following core conclusions.

The LSTM time-series model is the optimal model for financial risk prediction of listed seed companies. The LSTM model can effectively capture the temporal dynamic dependency of financial indicators of listed seed companies, with a test set AUC value of 0.889 and an accuracy rate of 85.71%, significantly outperforming the traditional LR model and XGBoost ensemble learning model, verifying the adaptability of time-series machine learning algorithms to time-series financial data of the seed industry, with model results fully reproducible. In addition, seed industry-specific indicators play a key role in improving risk prediction. Incorporating industry-specific indicators such as R&D investment ratio and seed production cost rate increases the model's prediction accuracy by 11.8%, indicating that the financial risks of listed seed companies are highly correlated with inherent industry characteristics, and industry-specific indicators are indispensable core variables for risk prediction.

Financial risks of listed seed companies exhibit significant individual and temporal differences. Based on real data, the overall solvency of sample enterprises is stable, but there are significant differences in R&D investment and cash flow status. Longping High-Tech has a high R&D investment ratio exceeding 10% but faces significant short-term cash flow pressure. Denghai Seeds experienced a sharp decline in revenue and net profit in 2024 due to market supply-demand imbalances, with prominent financial risks. All conclusions are supported by real data. Machine learning algorithms are generally superior to traditional statistical models. The prediction accuracy of the XGBoost ensemble learning model is higher than that of the traditional LR statistical model, indicating that machine learning algorithms can effectively mine non-linear characteristics of financial data, improve risk prediction accuracy, and serve as an effective method for financial risk prediction of listed seed companies.

5.2. Financial risk control optimization strategies for listed seed companies

Based on reproducible empirical results and the real development characteristics of the seed industry, optimization strategies for financial risk control of listed seed companies are proposed from enterprise, industry, and regulatory perspectives, constructing an "intelligent, dynamic, and full-chain" financial risk control system for listed seed companies. All strategies are aligned with the actual operating data of sample enterprises.

5.2.1. Enterprise level: Constructing a time-series financial risk early warning system based on the LSTM model and strengthening refined control

With the LSTM model as the core, combined with tools such as Python and TensorFlow, constructing an exclusive

financial risk early warning platform for listed seed companies, connecting to CSMAR and enterprise financial systems to realize real-time collection, analysis, and risk early warning of quarterly financial data, timely capturing temporal abnormal characteristics of financial indicators, and early identifying potential financial risks is essential.

On one hand, optimizing the R&D investment structure. Enterprises with high R&D investment such as Longping High-Tech can alleviate cash flow pressure through R&D expense capitalization and applying for government seed industry R&D subsidies (2024 Seed Industry Revitalization Subsidy Policy). Enterprises with low R&D investment such as Wanxiang Denong can appropriately increase R&D investment to enhance core competitiveness. On the other hand, strengthening seed production cost control, reducing the seed production cost rate by improving the mechanization rate of seed production bases and optimizing supply chain management. Enterprises with high seed production cost rates such as Quanyin High-Tech can establish a dynamic monitoring mechanism for seed production costs.

According to the seasonal characteristics of seed operations, reasonably arrange fund inflows and outflows, and construct a cash flow buffer mechanism. Enterprises with cash flow strain such as Denghai Seeds can alleviate short-term pressure through bill discounting and short-term working capital loans. Optimizing the asset-liability ratio, enterprises with high asset-liability ratios such as Nongfa Seeds can reduce financial leverage through equity financing and debt restructuring, controlling the asset-liability ratio near the industry average.

Based on the prediction results of the risk early warning platform, formulate targeted response plans for high-risk scenarios, such as adjusting the seed product structure when market supply and demand are imbalanced, and improving the agricultural insurance system when natural risks occur, realizing proactive response and effective mitigation of financial risks.

5.2.2. Industry level: Building a seed industry financial data sharing platform and promoting industry collaborative risk management

Led by the China National Seed Association, collaborate with authoritative databases such as CSMAR and Wind to build a financial data sharing platform for listed seed companies, integrating reproducible data such as industry financial indicators, market supply and demand, and natural climate to achieve data interconnection and sharing, providing data support for enterprise risk prediction and industry regulation. Furthermore, industry associations should organize technical training on machine learning and big data analysis, popularize the LSTM and XGBoost models verified in this study, and provide reproducible code templates to improve the overall intelligent risk management level of the seed industry. Based on the industry financial data sharing platform, construct an overall financial risk monitoring model for the seed industry, real-time monitoring the industry's financial risk situation, and timely issue industry risk early warnings to guide enterprises in preparing for risk response in advance and reducing systemic industry risks.

5.2.3. Regulatory level: Improving information disclosure systems and strengthening intelligent supervision and policy support

The China Securities Regulatory Commission (CSRC), Shenzhen Stock Exchange (SZSE), and Shanghai Stock Exchange (SSE) should further standardize the information disclosure requirements for core indicators such as R&D investment, seed production costs, and cash flow of listed seed companies, clarify disclosure standards and frequencies, improve the transparency and completeness of financial information, and ensure that investors and enterprise risk management can obtain reliable financial data. Regulatory authorities such as the CSRC and the

Ministry of Agriculture and Rural Affairs, using the machine learning algorithms verified in this study, should construct a financial risk supervision model for listed seed companies, realizing real-time monitoring and dynamic supervision of corporate financial risks, timely detecting and disposing of corporate financial risks, and preventing market risks.

In response to the characteristics of high R&D investment and long cycles in the seed industry, tax and fiscal policies such as R&D expense additional deduction and government subsidies to alleviate enterprises' R&D cash flow pressure should be further improved. Moreover, a seed industry development fund should be established to provide financing support for listed seed companies, optimizing their capital structure, and reducing financial risks.

5.3. Research limitations and prospects

Based on publicly available time-series data from Q1 2016 to Q3 2024, this study constructs a financial risk prediction model for listed seed companies and achieves certain research results. However, there are still some limitations as follows:

(1) The sample size is relatively small, only selecting 6 leading listed seed companies. In the future, the sample scope can be expanded to include New Third Board seed enterprises for research;

(2) External factors such as macroeconomics and seed industry policies are not considered in their impact on corporate financial risks. In the future, macro indicators such as GDP growth rate and seed industry subsidy policies can be incorporated into the indicator system to improve the comprehensiveness of the model;

(3) Only three machine learning models are compared. In the future, more advanced time-series algorithms such as GRU and Transformer can be introduced to further improve model prediction accuracy.

With the deep integration of big data, artificial intelligence, and the seed industry, financial data of listed seed companies will present richer characteristics. An intelligent financial risk prediction and control system based on multi-source data and multi-algorithm integration will become a research hotspot. The research results of this study provide a reproducible methodological reference for financial risk prediction of listed seed companies. In the future, further in-depth research can be conducted on this basis to provide more comprehensive financial risk management support for the high-quality development of the seed industry.

Disclosure statement

The author declares no conflict of interest.

References

[1] Hansen J, McDonald R, Westgaard J, 2019, LSTM-Based Bankruptcy Prediction for Scandinavian Firms. Expert Systems with Applications, 2019(135): 234–245.

[2] Altman E, 1968, Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4): 589–609.

[3] Ohlson J, 1980, Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18(1): 109–131.

[4] Zhang X, Chen D, Tang Y, 2020, Financial Quality, Financial Risk and Financial Crisis: Empirical Evidence from

Chinese Listed Companies. Accounting Research, 2020(6): 3–19.

[5] Wu X, Li X, Zhang Y, 2022, A Comparison of Machine Learning Algorithms in Financial Risk Prediction of Listed Companies: An Empirical Analysis Based on the Manufacturing Industry. Systems Engineering: Theory & Practice, 42(8): 2093–2106.

[6] Li L, Wang J, Liu Y, 2023, Construction of Financial Performance Evaluation System for Seed Enterprises under the Background of Big Data: Based on the Entropy Weight TOPSIS Method. Journal of Agrotechnical Economics, 2023(5): 112–125.

[7] Liu H, Li R, 2015, Construction of Financial Risk Early Warning Model for Agricultural Listed Companies: An Empirical Study Based on Factor Analysis and Logistic Regression. Chinese Rural Economy, 2015(6): 85–96.

[8] He Y, Yang M, Zhou H, 2021, Enterprise Digital Transformation and Financial Risk Prevention and Control: Empirical Evidence from Chinese Listed Companies. Accounting Research, 2021(8): 3–19.

[9] Zhao C, Cao W, Zhu J, 2020, R&D Investment, Ownership Nature and Corporate Financial Risk. Science Research Management, 41(7): 18–27.

[10] Zhou X, Li X, 2022, Research on Financial Crisis Early Warning of Listed Companies based on LSTM Neural Network. Journal of Industrial Engineering and Engineering Management, 36(3): 152–161.

[11] Lin Z, Zheng J, Bu J, 2017, Research on the Relationship between Enterprise Risk Management and Financial Performance: Empirical Evidence from Chinese Listed Companies. Accounting Research, 2017(1): 60–67.

[12] Chen X, Dai J, 2021, Research on Financial Risk Early Warning of Small and Medium-Sized Enterprises based on XGBoost. Systems Engineering: Theory & Practice, 41(2): 363–374.