

Article

Research on Company Financial Risk Early Warnings Based on FA-LSTFormer Model

Miao Cheng ^{1,*}  and Ning Wu ²¹ Faculty of New Commercial Science, Anhui Sanlian University, Hefei 230601, China² Faculty of Engineering, Anhui Sanlian University, Hefei 230601, China

* Correspondence: cmiao5568@126.com

Abstract

Timely identification of company financial risks is crucial for investors and regulators. However, existing studies overlook the class imbalance caused by the scarcity of high-risk samples, and the interpretability of deep models is insufficient, making it difficult to meet the practical needs. To address these problems, we propose a classification model named FA-LSTFormer, which models a company's financial risk as low-, medium- and high-warning tasks. FA-LSTFormer employs LSTM and Transformer to decouple the short-term continuity and long-term dependency inherent in financial data. To further attend to indicator-level nuances, we incorporate a Risk-Sensitive Hierarchical Indicator Attention (RSHA) module. Moreover, given the pronounced class imbalance where high-risk events are substantially underrepresented, we further propose a Class-Imbalance-Aware Focal Loss (CIFL) function to prioritize these minor yet critical samples and suppress false negatives. On the dataset of Chinese A-share manufacturing listed companies, experimental results show that our FA-LSTFormer achieves superior performance in accuracy, precision, recall, F1-score and AUC, achieving 92.76%, 93.13%, 91.84%, 92.48%, and 95.27%, respectively. Compared to the suboptimal LTR-Net, it improves these metrics by 1.64–3.60%. Compared to the LSTM–Transformer baseline, FA-LSTFormer improves on it by 4.30–9.13%. In the risk-oriented decision evaluation, FA-LSTFormer achieves a warning ROC of 0.954 for the high-risk class and lowers the error rate to 9.82%. It maintains an accuracy rate of 84.46% even after three years of early warning and exhibits strong robustness across different warning thresholds and company sizes. These results verify the advantages of FA-LSTFormer in both algorithmic performance and practical early-warning applications.

Keywords: financial risk; early warning; LSTM–Transformer; RSHA; CIFL

Academic Editor: Chenliang Liu

Received: 5 June 2026

Revised: 15 July 2026

Accepted: 16 July 2026

Published: 23 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The financial health of a company serves as the cornerstone of its survival and growth. When a company encounters financial risks, operational difficulties may trigger a chain reaction, including debt defaults, layoffs and bankruptcy, damaging the interests of investors and creditors and undermining the stability of the wider economic system. Therefore, accurately predicting financial risks is one of the core issues in corporate finance and financial management. An effective early-warning method provides managers with a window in which to take corrective action, enables creditors and investors to reprice risk, and supports regulators in maintaining market order. In recent years, with the rapid development of deep learning and time-series modeling, research on financial risk early warning has been evolving from static discrimination toward dynamic, interpretable and gradable methods.

Research on financial risk prediction has broadly evolved through three phases. The first phase is dominated by statistical discriminant methods. Altman [1] pioneers the Z-score model, using multiple discriminant analyses to combine accounting ratios to distinguish between healthy enterprises and potential bankruptcy candidates. Beaver [2] examined the predictive ability of individual financial ratios for company failure with a univariate analysis. Ohlson [3] employs logistic regression to relax distributional assumptions such as normality and provides easily rankable probability estimates. These linear models [4] are transparent and easy to apply, yet they rest on strict statistical assumptions and treat each reporting period in isolation rather than capturing the joint evolution of indicators across years [5]. The second phase introduces machine learning classifiers such as support vector machines, decision trees, random forests [6] and gradient-boosted trees [7], which can capture nonlinear interactions among financial metrics. These often outperform classical statistical models in cross-sectional data and are further combined with imbalance mitigation techniques to enhance minority class recognition, such as SMOTE [8]. Many works [9,10] point out that, although this stage markedly improves predictive accuracy, most studies still compress multi-period information into hand-crafted summary features, discarding the temporal structure that risk evolution should inherently reveal. The third phase applies deep learning to the inherently temporal nature of financial data [11]. Subsequently, representative works such as Informer [12], Temporal Fusion Transformer [13] and Autoformer [14] emerged in the field of long sequence prediction.

For financial risk warning tasks, Cheng [15] proposes a DUAL-LSTM with embedded attention that confirms the value of attention mechanisms in improving warning accuracy. Mai et al. [16] integrate textual disclosures into deep bankruptcy prediction. Hosaka [17] visualizes financial ratios and applies convolutional neural networks. LTR-Net [18] fuses LSTM, Transformer and residual modules and reports that the joint design outperforms any single architecture in capturing temporal dependencies, global information and nonlinear relationships. Wang et al. [19] propose a fusion deep neural network for multi-period annual report data. Mohamed [20] constructs a deep warning model for multidimensional analysis. Yang [21] explores how CNN-LSTM supports company financial forecasting. Beyond Transformers, selective state space models represented by Mamba [22] have emerged as efficient sequence learners. Recent frameworks such as TSCMamba [23], Mamba-hypergraph [24] and MambaSL [25] report significant improvements on time-series classification benchmarks, offering competitive baselines for sequence modeling.

Despite the continuous advancements in predictive accuracy achieved by deep learning and hybrid architectures, prominent limitations persist when these models are applied to real-world corporate risk warning. To begin with, most works formulate the task as a binary classification between distressed and healthy firms [16,17]. However, managers actually make decisions in a hierarchical warning manner. It is necessary to distinguish between companies in a state of concern and those in acute danger in order to apply differentiated intervention measures. Moreover, high-risk companies are extremely rare, and this class imbalance biases models trained with ordinary cross-entropy toward the majority class, making the high-risk firms whose misclassification cost is the highest precisely the most likely to be missed. Finally, deep models often sacrifice transparency while improving performance, and regulatory agencies' requirements for model interpretability are increasing. More works are integrating interpretability tools into financial warning models, such as attention weights and Shapley attribution, enabling a warning to be traced back to its underlying risk drivers. To address the above three limitations, we propose FA-LSTFormer. This model employs the LSTM-Transformer as the temporal encoding backbone to capture both short-term continuity and long-term dependencies in risk evolution, which supports a three-level early-warning system. FA-LSTFormer introduces the Risk-Sensitive Hierarchi-

cal Indicator Attention (RSHA) module to impose sample-adaptive sparse attention over financial indicators, making key risk drivers traceable and enhancing interpretability. FA-LSTFormer further adopts the Class-Imbalance-Aware Focal Loss (CIFL) function to correct the class imbalance caused by scarce high-risk samples and reduce their misclassification.

Our contributions are outlined as follows:

- (1) FA-LSTFormer consists of an LSTM–Transformer, RSHA module and CIFL loss function. The LSTM–Transformer maps multi-year financial indicator sequences into contextual representations that combine short-term continuity and cross-period long-term dependencies, thereby characterizing the gradual evolution trajectory of risk. The RSHA adaptively focuses on drivers of key financial risk factors. CIFL is an imbalanced loss function that corrects the class bias caused by the scarcity of high-risk companies during training, reducing false negatives.
- (2) We design RSHA to operate on the contextual representations by splitting them along the financial indicator dimension and applying Softmax normalization, thereby enabling the model to dynamically prioritize key risk-driven indicators. RSHA both suppresses the interference of irrelevant indicators on the warning result and makes the result traceable to specific financial factors, thereby enhancing interpretability.
- (3) We propose CIFL to address the scarcity of high-risk enterprise samples and the high cost of underreporting. CIFL collaboratively corrects the bias of enterprise category distribution in three ways: introducing a focus modulation factor in the loss layer, focusing training on high-risk samples that are difficult to distinguish, using category weights based on the number of effective samples in the sample layer, and introducing a minority class synthesis mechanism in the training layer.
- (4) On the dataset of Chinese A-share manufacturing companies listed from 2015 to 2025, we compare FA-LSTFormer with nine methods, including Random Forest, LSTM, DUAL-LSTM, Transformer, CNN-LSTM, LSTM–Transformer, TSCMamba, MambaSL and LTR-Net. Results show that FA-LSTFormer achieves optimal performance in both model performance and financial risk warning applications. The former is manifested in superior accuracy, F1 and AUC value. The latter is reflected in the lowest-cost-weighted misjudgment, the longest warning lead time, and the strongest cross-industry robustness.

2. Company Financial Data and Problem Description

2.1. Making a Dataset

Raw financial data are obtained from the CSMAR database [26] and the Wind Financial Terminal [27]. The initial sample consists of A-share listed companies in the manufacturing sector during the period 2015–2025. Based on the raw financial data, we manually construct the experimental dataset by performing sample screening, handling missing values and applying Winsorization for outlier treatment. The manufacturing sector is selected because it accounts for a large proportion of the A-share market, ensuring sufficient sample size and strong comparability among firms. Standardized reporting and cross-firm comparability, together with the special treatment (ST) mechanism for firms with going-concern uncertainties, provide institutional grounds for defining objective financial risk labels.

We operationalize a company's financial risk warning into three classification tasks, namely low-risk, medium-risk and high-risk. The high-risk level corresponds to companies that are subject to ST or *ST in the prediction year or the following year, both of which belong to the category of companies issued a risk warning by the exchange. ST indicates that a company exhibits financial or operational abnormalities. *ST indicates that a company has triggered the criteria for delisting risk warning and faces potential compulsory delisting. For a target prediction year t , the model takes the financial indicators of the company from

year $t - 1$ and the preceding T consecutive years as input features, and predicts whether the company will be subject to ST or *ST designation in year t or $t + 1$. This temporal alignment ensures that the model uses only historically available financial information for prediction, eliminating the possibility of future information leakage.

The medium-risk level refers to companies that have not triggered special treatment but show clear deterioration in financial indicators. Following common threshold practices in financial risk early-warning research, the deterioration criteria are defined as any of the following: negative return on total assets for two consecutive years, indicating substantial impairment of going-concern capability; debt-to-asset ratio above the 75th percentile of the industry in the current year, indicating significantly higher leverage than peers; or negative operating profit, indicating core business losses. These thresholds are widely adopted in the financial distress literature and are consistent with exchange-based risk monitoring practices, with clear economic interpretations. The low-risk level covers all companies that are not under special treatment and meet none of the above deterioration conditions. To preserve sample quality, companies with prolonged trading suspensions and firm-year observations with critical missing values are excluded; extreme outliers are further winsorized. After filtering, 536 companies are retained, of which 300 are low-risk, 128 are medium-risk and 108 are high-risk.

Furthermore, we select eight financial indicators as training features for FA-LSTFormer, as shown in Table 1. These indicators cover the main aspects of a company's financial condition while keeping inter-indicator redundancy low and, together with the three labels, provide high-quality input features for the subsequent training and evaluation of the FA-LSTFormer.

Table 1. Eight financial indicators and their meanings.

Financial Indicator	Evaluation Dimension	Economic Meaning
Return on assets (ROAs)	Profitability	Profit generated per unit of total assets
Return on equity (ROE)	Profitability	Return delivered to shareholders
Debt-to-asset ratio	Solvency	Leverage level and long-term solvency pressure
Current ratio	Solvency	Short-term debt-paying ability
Total asset turnover	Operating efficiency	Efficiency of asset utilization
Revenue growth rate	Growth	Momentum of business expansion
Operating cash flow to total debt ratio (CFTD)	Cash flow	Cash coverage of total debt
Operating cash flow	Cash flow	Quality of earnings measured in cash

2.2. Formalization of Financial Risk Early Warning in Companies

For each company, we denote its historical sequence of length T as $X = (x_{[t-T]}, \dots, x_{[t-1]})$. Each $x(i) \in \mathbb{R}^d$ is a vector composed of d financial indicators observed by the company in year i , and t is the predicted year. We formalize the financial risk early warning of a company as a mathematical modeling problem. Given X , learn a mapping $f: X \rightarrow y$, where $y \in Y = \{\text{low risk, medium risk, high risk}\}$ represents the risk level in the predicted year t . And the three levels are arranged in ascending order of risk severity to form an ordered output space. Specifically, the mapping f is implemented by a deep neural network with parameter θ , and the model produces predictions of risk level in the form of conditional probabilities. The risk level with the maximum posterior probability is taken as the warning result, as given in Equation (1):

$$P(y = c | X; \theta) = \text{softmax}(f_{\theta}(X))c, \quad \hat{y} = \text{argmax} P(y = c | X; \theta) \{c \in Y\}, \quad (1)$$

where f_{θ} is the unnormalized scores output by the model at three risk levels, then a Softmax operator converts them into a valid probability distribution, and $\hat{y}$ is the final predicted level.

The model parameter θ is optimized by minimizing the loss function on the training set. This formulation casts the financial risk early-warning problem as a supervised sequence classification task amenable to deep learning models.

3. Method

3.1. Overall Framework

The design motivation of the FA-LSTFormer originates from two characteristics of the company's financial risk evolution process. Along the time dimension, short-term continuity exists between adjacent reporting periods, while the path to distress often depends on short-term signals across years. Along the indicator dimension, the contribution of different financial indicators to risk is not uniform, and their relative importance varies with the company and risk scenario. Along the sample dimension, high-risk companies are extremely rare, but their underreporting incurs the highest cost. If any one of these characteristics is ignored, the warning model fails in practice. On this basis, FA-LSTFormer builds an end-to-end warning framework, which assigns the above three types of characteristics to interconnected components for processing and embeds two targeted innovative mechanisms. The overall architecture is shown in Figure 1.

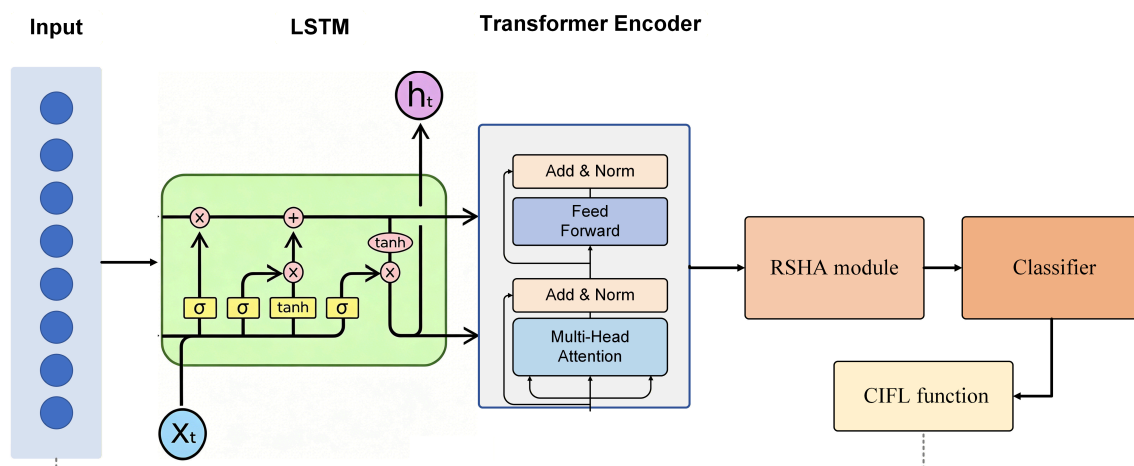


Figure 1. The overall architecture of FA-LSTFormer. Circles denote gated operations (LSTM), squares denote linear and feed-forward layers, dotted lines indicate the flow of hidden states across time steps. And different colors distinguish the LSTM branch from the Transformer branch.

The overall process of FA-LSTFormer consists of five stages. The first stage is input preprocessing. Standardize the eight financial indicators of the company to form a multivariate time-series tensor. The second stage is LSTM temporal encoding, where the indicator vectors are read year by year in chronological order, short-term continuity between adjacent reporting periods is captured through gating mechanisms, and hidden state sequences are output accordingly. Next is Transformer global modeling. After applying positional encoding, multi-head self-attention models the long-term dependencies and outputs a high-dimensional representation z that integrates the context. Based on this, our RSHA module decomposes z into eight indicator dimensions and assigns learnable normalized weights to each indicator according to the specific risk scenario of the sample. We further obtain a weighted discriminant representation of the indicators and generate the corresponding early-warning outputs. The fifth stage is training optimization. Then, the three-risk probability is output using Softmax, and the CIFL is taken as the loss function. Then, end-to-end gradient feedback drives the whole network jointly toward optimization that favors high-risk identification.

3.2. LSTM–Transformer Backbone Network

LSTM can update the tight encoding of short-term continuity between adjacent reporting periods. Transformer aggregates cross-temporal long-term dependencies globally through the self-attention mechanism. The combination of these two can simultaneously characterize the local and global characteristics of financial time-series data. If x_t is the input vector at time t and h_{t-1} is the hidden state at the previous time, then the state update Equation (2) for the three gates of LSTM is as follows:

$$[f_t, i_t, o_t] = \sigma(W_g[h_{t-1}, x_t] + b_g), g_t = \tanh(W_c[h_{t-1}, x_t] + b_c), c_t = f_t \odot c_{t-1} + i_t \odot g_t, h_t = o_t \odot \tanh(c_t) \quad (2)$$

where σ is the logistic Sigmoid function and f_t, i_t, o_t are the forget gate, input gate, and output gate, respectively. c_t and h_t are the updated cell state and hidden state, respectively.

After traversing the input company financial data, the hidden state sequence $\{h_1, h_2, h_3\}$ is obtained. It encodes short-term continuity between adjacent reporting periods as input to the Transformer encoder. The Transformer encoder adds positional encoding to the hidden-state sequence to preserve temporal information and performs pairwise comparisons for each year through a multi-head self-attention mechanism. The core operator is scaled dot product attention, as shown in Equation (3):

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \text{sqrt}(d_k)) V \quad (3)$$

where Q, K and V are the query, key and value matrices, respectively. d_k is the dimension of the key vector. The multi-head mechanism obtains a linear transformation by running multiple independent attention heads in parallel and concatenating their results. Each encoder block consists of multiple self-attention sub-layers and position-wise feedforward sub-layers, with residual connections and layer normalization applied. After processing by the encoder block, the model outputs a contextual representation for each reporting year that combines local continuity and global long-range dependencies. For the convenience of subsequent description, the global representation of the entire sequence is denoted as z , which is the input of the RSHA mechanism.

3.3. RSHA Module

The evolution of financial risk exhibits clear heterogeneity in the indicator. Solvency-related indicators dominate in leverage-driven risks, while cash flow-related indicators play a pivotal role in operational distress. This observation highlights the sparsity and indicator-specific nature of risk signals. RSHA incorporates learnable indicator-level attention distributions under the premise of integrating contextual global representations, enabling the model to dynamically and sparsely focus on key risk drivers based on sample-specific risk contexts. This approach enhances discriminative accuracy while maintaining a priori verifiable interpretability. RSHA decomposes the global representation z into eight indicator dimensions, obtaining sub-representations associated with each financial indicator, as shown in Equation (4):

$$r_k = W_p^{(k)} z + b_p^{(k)}, \quad k = 1, 2, \dots, 8 \quad (4)$$

where r_k is the associated sub-representation, while $W_p^{(k)}$ and $b_p^{(k)}$ are the learnable projection parameters exclusive to this indicator. This hierarchical decomposition is equivalent to constructing a dedicated perspective for each indicator dimension atop the shared global representation, enabling the subsequent attention mechanism to apply weighting at the indicator level rather than the overall level.

Next, the learnable attention distribution is calculated. Nonlinearly fuse the global representation with each sub-representation of the metrics and score them, then normalize them into a distribution via a temperature-adjustable Softmax, as shown in Equation (5):

$$e_k = v^\top \tanh(W_a z + U_a r_k + b_a), \alpha_k = \frac{\exp(e_k / \tau)}{\sum_{j=1}^8 \exp(e_j / \tau)} \quad (5)$$

where W_a , U_a , $v^\top$ and b_a are learnable parameters, and the τ coefficient is used to regulate the sparsity of attention distribution. When τ approaches zero, the distribution tends towards hard selection, and the model focuses more on a very small number of samples. During the training process, appropriate sparsity is obtained through joint optimization, ensuring that the attention distribution remains moderately sparse while being dependent on the samples.

We further integrate the weighted indicators with sparse regularization. This yields a weighted fusion representation of the indicators, along with its associated attention entropy regularization term, as formulated in Equation (6):

$$\hat{z} = \sum_{k=1}^8 \alpha_k r_k, \mathcal{R}_{\text{attn}} = -\sum_{k=1}^8 \alpha_k \log \alpha_k \quad (6)$$

where $\hat{z}$ is the weighted discriminant representation of financial indicators. Quantifying the uncertainty of the current attention distribution with $\mathcal{R}_{\text{attn}}$ encourages the model to form a sparse focus rather than a uniformly scattered attention distribution.

3.4. CIFL Function

In real-world scenarios, the number of high-risk companies is small, the sample ratio is low, and the cost of underreporting is high. Among existing class-balancing loss functions, cross-entropy exhibits a pronounced majority-class bias. Weighted cross-entropy tends to over-penalize minority samples, often at the cost of training stability. Focal loss, although effective at focusing on hard examples, does not account for prior class information.

In contrast, CIFL integrates focal modulation with effective-sample-number-based reweighting and synthetic minority generation. The CIFL function reshapes the class prior weights through the effective sample size, reallocates the gradient contributions of difficult and easy samples, and couples the minority class synthesis mechanism to expand the support domain of high-risk manifolds, enabling the model to synergistically correct imbalanced biases in the loss layer, sample layer, and training layer.

CIFL builds upon focal loss to redistribute gradient contributions between easy and hard samples. During training, it jointly corrects class distribution bias from three complementary levels: at the loss level, a focal modulation factor is introduced to focus training on hard-to-classify samples; at the sample level, class weights based on the effective number of samples are adopted; and at the training level, an online synthesis mechanism generates minority-class samples to enrich the support of high-risk samples, as shown in Equation (7):

$$\mathcal{L}_{\text{CIFL}} = -\alpha_y (1 - p)^\gamma \log p_y, \alpha_y = \frac{1 - \beta}{1 - \beta^{n_y}} \quad (7)$$

where p_y denotes the predicted probability assigned by the model to the true class y . γ is the focus parameter, and the modulation factor $(1 - p)^\gamma$ rapidly decays when p approaches 1. The class weight α_y draws on the idea of effective sample size, where n_y is the number of training samples for category y , and β is a decay constant close to 1. This weight can more robustly reflect the effective information of minority classes, avoiding training instability caused by excessive weighting under extreme imbalance.

4. Experiments and Results

4.1. Experimental Setting

Table 2 lists the key hyperparameters and their search ranges for FA-LSTFormer. The LSTM sets have 64 hidden units. The Transformer encoder comprises two blocks, each with four attention heads. The RSHA module operates across eight metric dimensions. Optimization using Adam, an initial learning rate of 0.001, batch size of 64, and regularization using 0.2 Dropout. γ is set to 2.0, and the class balance parameter β is set to 0.999. These hyperparameters are selected via grid search, combined with early stopping to prevent overfitting. The search ranges are determined based on common practices in time-series deep learning and preliminary experiments. The dataset is chronologically split into training, validation, and test sets. The dataset is split chronologically into three subsets: 2015–2021 for training, 2022 for validation and 2023–2025 for testing. This chronological splitting strategy prevents future information leakage and ensures the authenticity of model evaluation.

Table 2. Key hyperparameters and search ranges of FA-LSTFormer.

Hyperparameter	Search Range	Selected Value	Description
LSTM hidden units	{16, 32, 64, 128, 256}	64	Number of hidden units in LSTM layer
Transformer blocks	{1, 2, 3, 4}	2	Number of stacked Transformer encoder blocks
Attention heads	{2, 4, 8, 16}	4	Number of parallel attention heads
Dropout rate	{0.1, 0.2, 0.3, 0.5}	0.2	Dropout ratio for regularization
Batch size	{16, 32, 64, 128}	64	Mini-batch size for training
Learning rate	$\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}\}$	0.001	Initial learning rate for Adam optimizer

We compare FA-LSTFormer with other methods, including random forest, LSTM, DUAL-LSTM, Transformer, CNN-LSTM, LSTM-Transformer, LTR-Net and TSCMamba. All models share the same chronological data split and the same set of eight financial indicators. Given the class imbalance inherent in corporate financial risk early-warning tasks, we adopt evaluation metrics beyond accuracy, including precision, recall, F1-score and AUC value. Special attention is given to the geometric mean of per-class recall rates, as it is particularly sensitive to minority-class performance. The training process of FA-LSTFormer is shown in Algorithm 1.

Algorithm 1. Training procedure of FA-LSTFormer

Input: Training set $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}_{i=1}^N$, where $\mathbf{X}_i \in \mathbb{R}^{T \times d}$ and $y_i \in \{\text{low, medium, high}\}$; number of epochs E ; batch size B .

Output: Trained model parameters Θ .

- 1: Initialize model parameters Θ randomly.
 - 2: **for** $epoch = 1$ to E **do**
 - 3: Shuffle and partition training set into mini-batches of size B .
 - 4: **for each** mini-batch $\{(\mathbf{X}_j, y_j)\}_{j=1}^B$ **do**
 - 5: **LSTM-Transformer Encoding:** Encode each $\mathbf{X}_j$ to obtain contextual representation $\mathbf{z}_j$.
 - 6: **RSHA Weighting:** Compute indicator-level attention weights α_j and obtain weighted representation $\hat{\mathbf{z}}_j$.
 - 7: **Classifier Output:** Compute risk probability distribution $\mathbf{p}_j = \text{softmax}(f(\hat{\mathbf{z}}_j))$.
 - 8: **Loss Computation:** Compute CIFL $\mathcal{L}_{\text{CIFL}}(\mathbf{p}_j, y_j)$.
 - 9: Backpropagate the loss and update parameters Θ via Adam optimizer.
 - 10: **end for**
 - 11: Evaluate on validation set; save the best-performing model.
 - 12: **end for**
 - 13: **return** Trained model with performance.
-

4.2. The Overall Performance of Financial Risk Early Warning

Table 3 compares the overall performance of various methods for financial risk early warning. FA-LSTFormer achieves superior performance across all metrics, with an F1-score of 92.48% and an AUC of 95.27%. Compared to the suboptimal LTR-Net, the F1-score and AUC improve by 1.67% and 3.31%, respectively. Relative to the LSTM–Transformer baseline, the F1-score and AUC increase by 7.29% and 5.09%, respectively. Other methods, such as LSTM, Transformer, DUAL-LSTM, and CNN-LSTM lag, significantly behind the top three methods across all evaluation metrics.

Table 3. Performance of different methods in financial risk warning for companies.

Method	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	79.06	74.23	71.45	71.67	78.87
LSTM	82.14	78.65	76.92	77.77	85.43
DUAL-LSTM	82.37	78.96	77.03	78.12	85.80
Transformer	83.06	79.88	78.10	79.98	86.72
CNN-LSTM	83.27	80.13	81.33	79.12	87.23
TSCMamba	84.79	82.05	83.47	82.75	89.94
MambaSL	85.33	82.71	81.26	81.97	90.51
LSTM–Transformer	88.91	85.34	85.05	85.19	90.18
LTR-Net	90.42	91.27	90.36	90.81	91.96
FA-LSTFormer (Ours)	92.76	93.13	91.84	92.48	95.27

Figure 2 further shows the ROC of the top three models across three risk levels. It can be observed that low corporate financial risk is the easiest to identify, with all three models achieving high and closely clustered AUC values. The performance gap among the three methods widens for medium financial risk early warning. For financial high-risk warning, the AUC of FA-LSTFormer is 0.954, which is significantly better than LTR-Net’s 0.913 and LSTM Transformer’s 0.895. Moreover, the curve shows a particularly significant lead in the low false positive rate area on the left side. This is precisely the work range that financial risk warning values the most. On the premise of strictly controlling false positives and avoiding disturbing healthy companies, FA-LSTFormer can still capture a higher proportion of real high-risk companies. The advantage of three consistent levels of curves indicates that our FA-LSTFormer effectively enhances the model’s ability to identify and warn of three types of financial risks, especially high-risk companies, which face the most severe consequences.

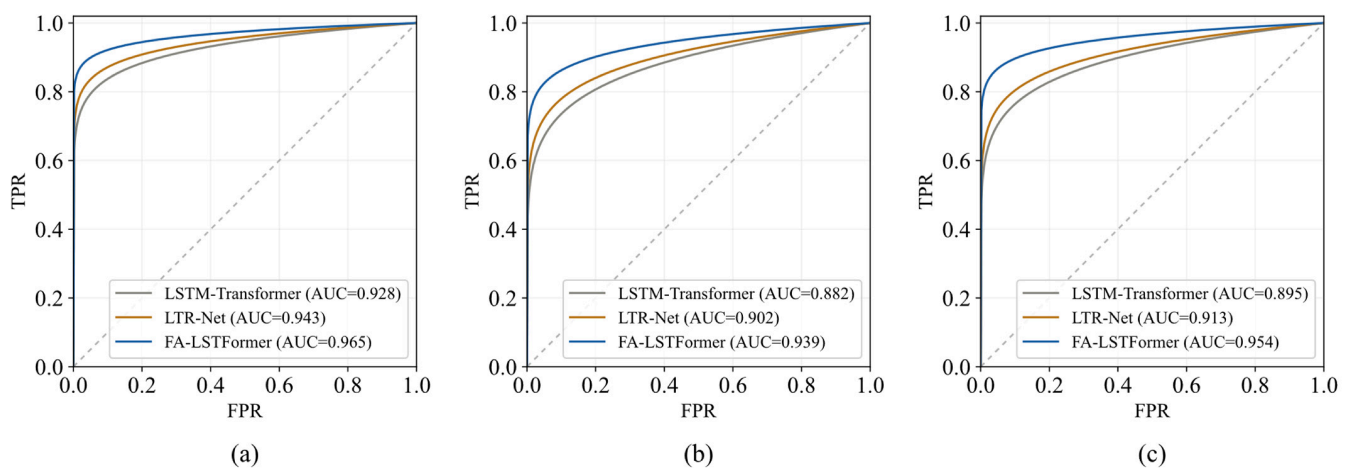


Figure 2. ROC curves for three financial risk warning levels. (a) Low-risk level. (b) Medium-risk level. (c) High-risk level.

4.3. Early-Warning Risk Results for Company Decision-Making

Misclassifications in financial risk early warning typically entail two types of errors with asymmetric consequences. The first is a false alarm, which means that the health company is classified as at risk. The cost of misreporting mainly includes additional due diligence costs, unnecessary attention, and a minor impact on the company's reputation. The second type of error is underreporting, which means that companies that are actually at risk are judged as healthy. The cost of underreporting is very serious, as it may cause investors and creditors to miss the opportunity to cut losses due to failure to receive timely warnings, and regulators may also overlook it as a result. Table 4 shows the evaluation results of five methods for two types of errors. According to the consensus of financial risk management, we calculate the total cost based on a weight of 5:1 between the second-type error and the first-type error. FA-LSTFormer reduces the second-type error rate to 9.82% and achieves the lowest total cost of 60.18, which is far lower than LTR-Net, LSTM Transformer, MambaSL, and TSCMamba. This indicates that FA-LSTFormer not only leads in statistical indicators but also significantly reduces actual decision risks in the most critical dimension of underreporting.

Table 4. Cost evaluation of misclassification errors for different methods.

Method	Error I	Error II	Cost
TSCMamba	18.62	23.08	134.02
MambaSL	17.95	21.9	127.45
LSTM-Transformer	15.41	17.95	105.16
LTR-Net	13.27	15.64	91.47
FA-LSTFormer (Ours)	11.08	9.82	60.18

Figure 3 shows the accuracy of early warnings under different prediction lead times, measuring the accuracy of making predictions one year, two years, and three years before the occurrence of risk events. The accuracy of all models decreases with the extension of the early stage, as earlier signals are weaker. The attenuation of FA-LSTFormer is the smoothest. It maintained an accuracy of 85.46% three years ahead of schedule, while the accuracy of the LSTM Transformer is only 78.05%. Early and reliable warnings are crucial for providing financial risk managers with sufficient decision-making and action time and have significant practical value.

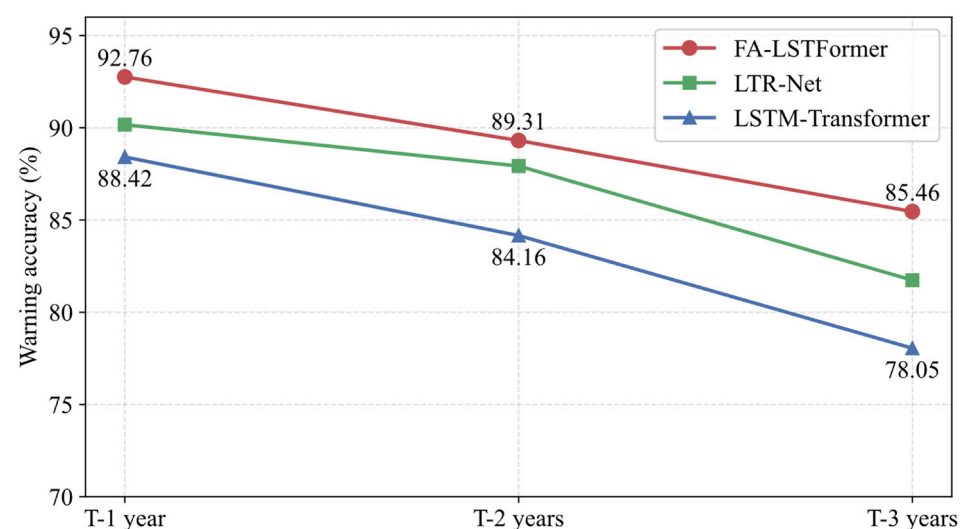


Figure 3. Accuracy of risk warning under different forecast years.

Figure 4 further presents the classification results of FA-LSTFormer for each risk level. For low-risk samples, 57 are correctly identified, two are misclassified as medium-risk, and one is classified as high-risk. For medium-risk samples, 23 are correctly identified, one is misclassified as low-risk, and two are classified as high-risk. For high-risk samples, 19 are correctly identified and two are misclassified as medium-risk, with no samples misclassified as low-risk. Notably, the fact that no high-risk samples are misclassified as low-risk validates the effectiveness of the CIFL function in reducing false negatives for the high-risk class.

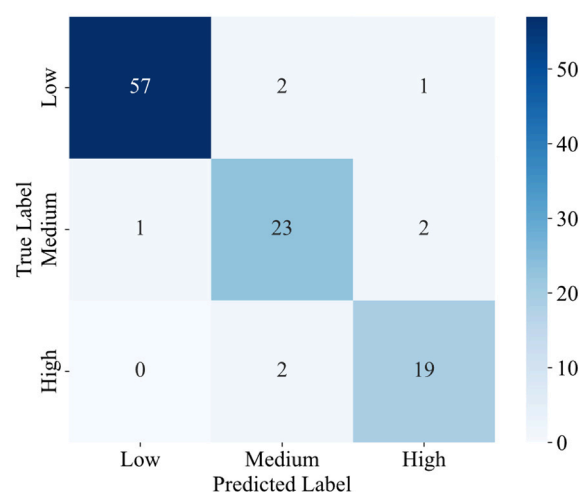


Figure 4. Confusion matrix of FA-LSTFormer on the test set.

Figure 5 visualizes the indicator-level attention weights generated by the RSHA module after taking the average within each risk level. The rows represent three levels of risk, and each column represents a financial indicator. The cell value represents the attention weight of the model on financial indicators at the risk level. The darker the color, the higher the weight. The horizontal axis represents the financial indicators that the model is most concerned about under the same risk level. The vertical axis reflects the changes in attention to the same financial indicators among different risk levels. For low-risk companies, attention is evenly distributed across various indicators. For high-risk-level companies, the model focuses on the asset–liability ratio and the ratio of operating cash flow to total liabilities. This model is economically reasonable because debt repayment pressure and insufficient cash coverage are precursors of the company’s financial risk, which indicates that the model learned to focus on the most important indicators when the risk was acute.

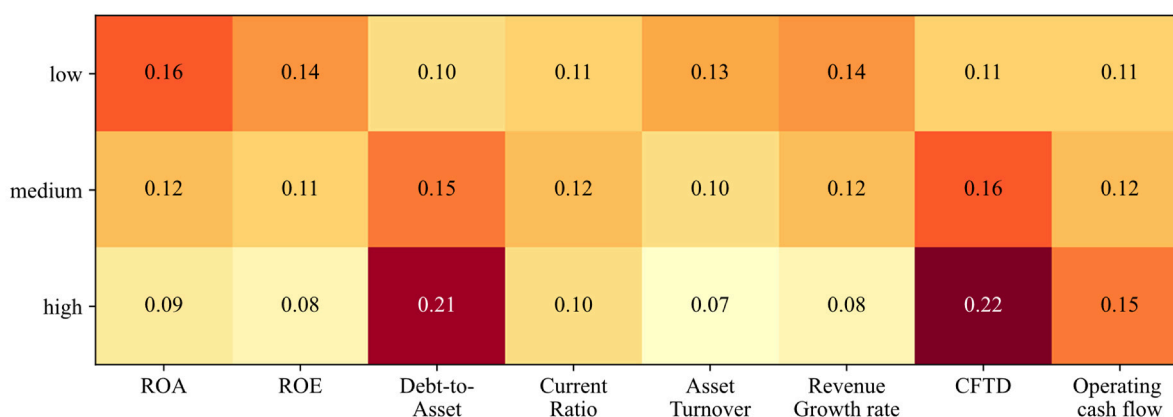


Figure 5. Model attention weights for financial indicators under three types of risks.

For high-risk companies, the model pays greatest attention to the debt-to-asset ratio and the operating cash flow to total debt ratio. This is unsurprising from a financial perspective, as rising leverage and weakening cash coverage are well-established distress signals. This observation indicates that RSHA not only improves predictive performance but also makes the warning results traceable to specific risk drivers, thereby meeting the interpretability requirements in regulatory and auditing contexts.

For high-risk-level companies, the model allocates the highest attention to the debt-to-asset ratio and the operating cash flow to total debt ratio, which aligns with classical financial theory, as rising leverage and insufficient cash coverage are core precursors of financial distress. This observation indicates that RSHA not only improves predictive performance but also makes the warning results traceable to specific risk drivers, thereby meeting the interpretability requirements in regulatory and auditing contexts. Unlike post hoc explanation tools such as SHAP and LIME, RSHA is a built-in attention mechanism that simultaneously achieves feature focusing and interpretability during training, without incurring additional computational costs. Moreover, its explanations are intrinsically consistent with the model's predictions, providing a more reliable and efficient means for understanding the model's behavior.

Table 5 shows the three indicators that received the highest attention when FA-LSTFormer issued a high-risk warning for the three companies that were subjected to special treatment. Among them, the company logo underwent desensitization treatment. For the three companies, the dominant indicators are consistent with the difficulties publicly disclosed by the companies. For example, a rise in leverage or a deterioration in cash flow indicates that the warning is not only accurate but also auditable.

Table 5. Case analysis of three struggling companies.

Case	Top Financial Indicators	Risk of Disclosure
Company A	Debt-to-assets ratio, total liabilities, current ratio	Liquidity stress driven by excessive leverage
Company B	total liabilities, ROA, OCF per share	Persistent operating cash flow deficit
Company C	Debt-to-assets ratio, revenue growth rate, ROA	Revenue collapse under high leverage pressure

Figure 6 shows the robustness test results of FA-LSTFormer. Figure 6a By changing the financial risk warning threshold, it is shown that the FA-LSTFormer remains higher than the macro F1 of the base model throughout the entire range of values. Figure 6b Grouping companies by market value shows that this advantage also holds true in large, medium, and small companies, with only a slight decrease in the smallest group. This confirms that the model has generalization ability across different company sizes.

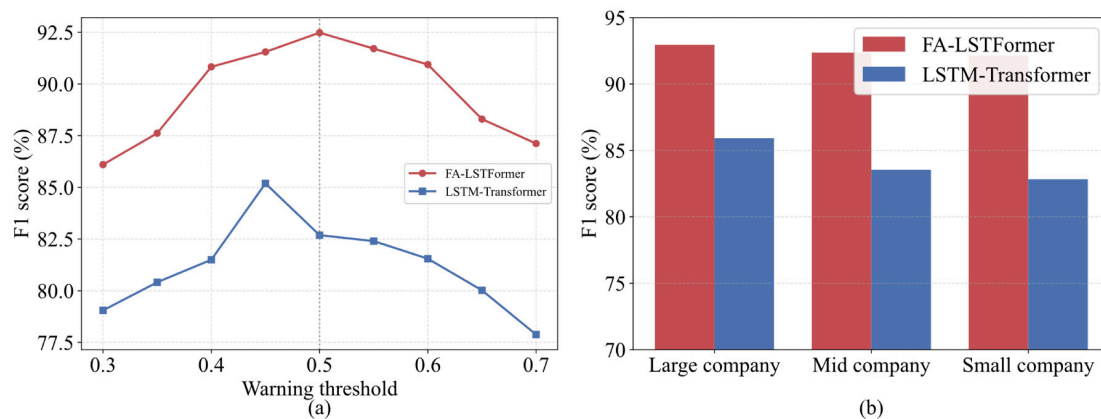


Figure 6. Threshold sensitivity and robustness across company sizes. (a) Financial risk warning threshold. (b) Company size grouping.

4.4. Ablation Experiment

We conducted comprehensive ablation experiments on the model components, as shown in Table 6. Using the LSTM–Transformer as the baseline, we sequentially added the RSHA module and the CIFL function to evaluate their individual contributions. After adding RSHA separately, the F1-score of the model increased from 85.19% to 87.12%, indicating that explicit modeling of indicator importance can effectively sharpen the model’s risk discrimination ability. After adding CIFL separately, the high-risk level recall rate significantly increased to 86.95%, confirming the targeted suppression effect of this loss function on false negatives. After synergistically integrating the two improvements, the complete FA-LSTFormer achieved optimal performance on all evaluation metrics. The results indicate that RSHA and CIFL have high complementarity at the functional level rather than redundant superposition.

Table 6. The ablation results of the two components of FA-LSTFormer.

RSHA	CIFL	Accuracy	F1-Score	AUC	High-Risk Recall
LSTM–Transformer (Baseline)		88.91	85.19	90.18	79.34
✓		90.34	87.12	92.41	83.07
	✓	89.87	86.58	92.05	86.95
✓	✓	92.76	92.48	95.27	95.40

✓ indicates adding the component.

To validate the reliability of the RSHA weights, we randomly selected 100 samples from the test set and compared the indicator-level attention assigned by RSHA with SHAP values. For each sample, we recorded the top three indicators ranked by RSHA and SHAP, and computed their overlap rate. The results show that the top three overlap rate reaches 82.7%, and 86.3% of the samples share at least two overlapping indicators. This finding demonstrates that the indicators focused on by RSHA are highly consistent with those identified as important by SHAP, providing quantitative evidence for the interpretability of RSHA. The detailed results are presented in Table 7.

Table 7. Agreement between RSHA and SHAP on top 3 indicators.

Metric	Value
Average top 3 overlap rate	82.5%
Proportion of samples with at least 2 overlapping indicators	86.3%
Proportion of samples with all 3 indicators overlapping	41.0%

Table 8 verifies the impact of different loss functions on the FA-LSTFormer backbone network. With cross-entropy loss, the model achieves a precision of 85.42% for high-risk identification, yet the recall rate is only 79.34%. This suggests that the model favors conservative predictions, at the expense of misclassifying a substantial proportion of high-risk companies. Weighted cross-entropy improves recall by reallocating class weights, but at the cost of sacrificing accuracy. Focal loss achieves a better balance between precision and recall, with a high-risk F1-score of 86.89%. Seesaw Loss further improves the F1-score to 89.22% and the Macro F1 to 90.36%. After replacing CIFL, the model’s F1-score for financial high-risk warning increased to 91.84%, and the overall F1-score also reached 92.48%, verifying the rationality and effectiveness of CIFL’s design concept of combining loss reweighting with data-level enhancement.

Table 8. Warning results of different loss functions for high-risk companies.

Loss Function	Precision	Recall	F1-Score	Macro F1
Cross-Entropy	85.42	79.34	82.27	85.19
Weighted Cross-Entropy	83.18	84.56	83.86	86.43
Focal Loss	86.07	87.73	86.89	88.71
Seesaw Loss	89.34	88.57	89.22	90.36
CIFL (Ours)	92.95	91.84	91.05	92.48

4.5. Discussion of Results

FA-LSTFormer consistently outperforms all compared methods. We attribute this to three perspectives. Its backbone captures both short-term continuity and long-term dependencies. LSTM preserves recent temporal order, while Transformer aggregates global context. This dual design suits financial risk evolution, where distress signals arise from both recent deterioration and accumulated imbalances. Second, the RSHA module adaptively focuses on the most relevant indicators for each company. High-risk companies receive higher attention on leverage and cash flow indicators, which aligns with financial theory. This improves both discriminative power and interpretability. Third, the CIFL function effectively mitigates class imbalance. CIFL combines focal modulation with effective-sample-number reweighting, substantially improving high-risk recall. Overall, the three components of FA-LSTFormer are individually effective and mutually complementary, jointly enabling more accurate, timely, and interpretable risk warnings.

5. Conclusions

Our FA-LSTFormer method solves the problem of underreporting caused by class imbalance in financial samples and insufficient interpretability of deep models, ultimately achieving a three-level early warning of company financial risks. FA-LSTFormer adopts LSTM–Transformer as its backbone, leveraging it as a temporal-encoding backbone to jointly characterize the local continuity and cross-period long-term dependencies of risk evolution. Then, through the RSHA module, sample-adaptive sparse attention was applied in the indicator dimension to focus on key risk-driving factors. And the CIFL function was employed to correct the class bias of high-risk samples. On the dataset of Chinese A-share manufacturing companies listed from 2015 to 2025, FA-LSTFormer achieved superior performance in all metrics, with the F1-score and AUC reaching 92.48% and 95.27%, respectively, which is an improvement of 1.67% and 3.31% compared to the suboptimal LTRNet. In the evaluation of enterprise risk decision-making, the ROC of high-risk warnings reached 0.954, and the cost-weighted total misjudgment decreased to 60.18. The second type of error rate was reduced to 9.82%, with an accuracy rate of 84.46% even after three years of early warning, and remained robust under different thresholds and company sizes. Ablation experiments further demonstrated the complementarity between RSHA and CIFL, which improved discrimination accuracy and high-risk recall, respectively. These results demonstrate that FA-LSTFormer not only achieves superior algorithmic performance but also exhibits lower decision costs, longer early-warning lead times, and auditable warning results in the cost-weighted misclassification simulation, indicating strong potential for practical application.

Our work has several limitations, which also point to promising directions for future research. (1) The samples only cover A-share manufacturing companies, and the risk labels rely on the special treatment system. Future work will extend the method to other industries, markets, and finer-grained risk grading to test its robustness. (2) The model currently only utilizes structured financial indicators and can integrate multiple heterogeneous sources of information, such as annual report text, management discussions, public opinion, and

macroeconomics, in the future to further enhance the foresight of early warning. (3) Causal inference and uncertainty quantification can be introduced on the basis of RSHA and CIFL, so that early-warning results can be explained while providing credibility intervals, thereby better serving the company's financial risk management and regulatory decision-making.

Author Contributions: Conceptualization, M.C. and N.W.; methodology, M.C.; software, M.C.; validation, M.C. and N.W.; formal analysis, M.C.; investigation, M.C.; resources, N.W.; data curation, M.C.; writing—original draft preparation, M.C.; writing—review and editing, N.W.; visualization, M.C.; supervision, N.W.; project administration, N.W.; funding acquisition, N.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Key Project of Humanities and Social Sciences Research in Anhui Higher Education Institutions, “Research on Management Innovation of Small and Medium-sized Enterprises under Blockchain Technology” (Grant No. SK2020A0614) and the “Research on Comprehensive Budget Management of Small and Medium-sized Enterprises Based on Machine Learning under the Smart Economy” (Grant No. 2024AH052440).

Data Availability Statement: Our dataset is collected from relevant data disclosed by Chinese-listed companies, and we are willing to share the dataset after the paper is published.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RSHA	Risk-Sensitive Hierarchical Indicator Attention
CIFL	Class-Imbalance-Aware Focal Loss

References

- Altman, E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **1968**, *23*, 589–609. [\[CrossRef\]](#)
- Beaver, W.H. Financial ratios as predictors of failure. *J. Account. Res.* **1966**, *4*, 71–111. [\[CrossRef\]](#)
- Ohlson, J.A. Financial ratios and the probabilistic prediction of bankruptcy. *J. Account. Res.* **1980**, *18*, 109–131. [\[CrossRef\]](#)
- Wu, G.; Bradic, J.; Tan, K.; Zhou, W. Linear Hypothesis Testing in High-Dimensional Expected Shortfall Regression with Heavy-Tailed Errors. *J. Mach. Learn. Res.* **2025**, *26*, 1–54.
- Elhoseny, M.; Metawa, N.; Sztano, G.; El-hasnony, I.M. Deep learning-based model for financial distress prediction. *Ann. Oper. Res.* **2025**, *345*, 885–907. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhang, Z.; Zhu, Z.; Hua, Y. Research on the financial early warning models based on ensemble learning algorithms: Introducing MD&A and stock forum comments textual indicators. *PLoS ONE* **2025**, *20*, e0323737. [\[CrossRef\]](#) [\[PubMed\]](#)
- Sun, J.; Fujita, H.; Zheng, Y.; Ai, W. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Inf. Fusion* **2020**, *54*, 128–144. [\[CrossRef\]](#)
- Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)
- Wei, G.; Han, G.S.; Lang, X. Fire risk assessment using machine learning techniques: A case study of Jinan City, China. *Sci. Rep.* **2026**, *16*, 6410. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE Computer Society: Washington, DC, USA, 2017; pp. 2999–3007. [\[CrossRef\]](#)
- Paul, C.; Fred, L.; Robert, P. Multifactor Timing with Deep Learning. *J. Financ. Econom.* **2026**, *24*, nbag006. [\[CrossRef\]](#)
- Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence 2021*; AAAI Press: Menlo Park, CA, USA, 2021; Volume 35, pp. 11106–11115. [\[CrossRef\]](#)
- Lim, B.; Arik, S.O.; Loeff, N.; Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [\[CrossRef\]](#)
- Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430.

15. Cheng, X. An attention embedded DUAL-LSTM method for financial risk early warning of the three new board-listed companies. *PeerJ Comput. Sci.* **2023**, *9*, e1271. [[CrossRef](#)] [[PubMed](#)]
16. Mai, F.; Tian, S.; Lee, C.; Ma, L. Deep learning models for bankruptcy prediction using textual disclosures. *Eur. J. Oper. Res.* **2019**, *274*, 743–758. [[CrossRef](#)]
17. Hosaka, T. Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Syst. Appl.* **2019**, *117*, 287–299. [[CrossRef](#)]
18. Liu, S. LTR-Net: A deep learning-based approach for financial data prediction and risk evaluation in enterprises. *PLoS ONE* **2025**, *20*, e0328013. [[CrossRef](#)] [[PubMed](#)]
19. Che, W.; Wang, Z. Predicting financial distress using multimodal data: An attentive and regularized deep learning method. *Inf. Process. Manag.* **2024**, *61*, 103703. [[CrossRef](#)]
20. Mohamed, A.; Mohamed, M.S. Deep learning-based financial risk early warning model for listed companies: A multi-dimensional analysis approach. *Expert Syst. Appl.* **2025**, *283*, 127331. [[CrossRef](#)]
21. Yang, A. Big data-driven corporate financial forecasting and decision support: A study of CNN-LSTM machine learning models. *Front. Appl. Math. Stat.* **2025**, *11*, 1566078. [[CrossRef](#)]
22. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2024**, arXiv:2312.00752.
23. Ahamed, M.A.; Cheng, Q. TSCMamba: Mamba meets multi-view learning for time series classification. *Inf. Fusion* **2025**, *120*, 103079. [[CrossRef](#)] [[PubMed](#)]
24. Jiang, J.; Zhuang, C.; Lei, F.; Chen, Z.; Xiao, L.; Liang, M.; Peng, J. A novel Mamba-hypergraph enhanced time-frequency fusion network for multivariate time series classification. *Complex Intell. Syst.* **2025**, *11*, 380. [[CrossRef](#)]
25. Jung, Y.M.; Kim, L. MambaSL: Exploring single-layer Mamba for time series classification. In Proceedings of the International Conference on Learning Representations (ICLR, 2026), Singapore, 24 April–2 May 2026.
26. CSMAR. *China Stock Market & Accounting Research Database*; GTA Data Service Co., Ltd.: Shenzhen, China, 2025. Available online: <https://www.csmar.com> (accessed on 10 March 2026).
27. Wind Information Co., Ltd. *Wind Financial Terminal*; Wind Information: Shanghai, China, 2025. Available online: <https://www.wind.com.cn> (accessed on 27 March 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.