systems

MDPI

Article

Stacked Ensemble Model with Enhanced TabNet for SME Supply Chain Financial Risk Prediction

Wenjie Shan and Benhe Gao *

Division of Logistics and Transportation, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518071, China; shanwj23@mails.tsinghua.edu.cn

* Correspondence: gaobh@sz.tsinghua.edu.cn; Tel.: +86-188-5209-6611

Abstract

Small and medium-sized enterprises (SMEs) chronically face financing frictions. While supply chain finance (SCF) can help, reliable credit risk assessment in SCF is hindered by redundant features, heterogeneous data sources, small samples, and class imbalance. Using 360 A-share–listed SMEs from 2019–2023, we build a 77-indicator, multidimensional system covering SME and core-firm financials, supply chain stability, and macroeconomic conditions. To reduce dimensionality and remove low-contribution variables, feature selection is performed via a genetic algorithm enhanced LightGBM (GA-LightGBM). To mitigate class imbalance, we employ TabDDPM for data augmentation, yielding consistent improvements in downstream performance. For modeling, we propose a two-stage predictive framework that integrates TabNet-based feature engineering with a stacking ensemble (TabNet-Stacking). In our experiments, TabNet-Stacking outperforms strong machine-learning baselines in accuracy, recall, F1 score, and AUC.

Keywords: small and medium-sized enterprises; supply chain finance; credit risk assessment; feature selection; data augmentation; TabNet; stacking ensemble

check for updates

Academic Editor: Wen-Chyuan Chiang

Received: 12 September 2025
Revised: 6 October 2025
Accepted: 9 October 2025
Published: 10 October 2025

Citation: Shan, W.; Gao, B. Stacked Ensemble Model with Enhanced TabNet for SME Supply Chain Financial Risk Prediction. *Systems* **2025**, *13*, 892. https://doi.org/10.3390/systems13100892

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Supply chain finance (SCF) leverages core-firm credit to optimize payment cycles and improve working-capital efficiency, thereby easing SMEs' financing frictions and stabilizing supply chains [1]. However, in the process of using supply chain financing, SMEs still face acute information asymmetries, persistent cash shortages, and dependence on core-enterprise payment terms or prepayments [2]. Traditional credit risk assessment methods, such as AHP and fuzzy comprehensive evaluation, are subjective, while machine learning and neural networks, though promising, face limitations in modeling complex nonlinearities and high-order interactions. In addition, in actual business, the number of credit risk defaulting companies is far lower than that of non-defaulting companies, and there is a data imbalance problem. Traditional methods include data processing, post-processing and cost-sensitive methods, but all have obvious limitations [3–5]. Faced with high-dimensional, redundant, and weakly informative financial covariates drawn from heterogeneous sources, machine learning and deep learning models can learn discriminative higher-order representations that capture nonlinear interactions and cross-source dependencies. However, the model opacity hampers the traceability of decision rationales, elevates compliance, audit, and communication burdens, and ultimately limits deployment in risk-sensitive settings.

Situated in the context of Chinese SMEs, this study makes the following contributions: (i) using SMEs listed on China's A-share market as the study sample, we construct a

multidimensional measurement system that comprehensively captures risks at the SME level, the core firm level, and the macroeconomic level, and for the supply chain as a whole, yielding broader dimensional coverage; (ii) given that SCF datasets are typically high dimensional, weakly correlated, and redundant, and that prior studies often proceed directly to machine learning prediction, we introduce a feature selection procedure based on GA-LightGBM and employ SHAP to interpret the influence of key factors; (iii) recognizing that SME risk assessment studies commonly involve small samples and class imbalance, and that traditional imbalance handling methods have notable limitations, we adopt a state-of-the-art deep learning model for tabular data augmentation, TabDDPM, to alleviate imbalance and enhance predictive performance; and (iv) to build a rigorous risk assessment model, we develop a two-stage deep learning predictive framework and leverage TabNet's sparse encoding mechanism to identify salient risk factors, thereby providing decision support for commercial banks, investors, and policymakers.

Our approach delivers strong performance in imbalanced SME credit-risk classification. On the held-out test set, the model records a Type I error rate of 0.3% and a Type II error rate of 12.5%. Trained on the augmented dataset, the model achieves 97.78% accuracy, an F1 score of 92.45%, an AUC of 97.60%, and a recall of 87.50% for the risky class. In 10-fold cross-validation, the F1 score reaches 93.05%, improving by 4.90 and 10.89 percentage points over versions trained without data augmentation and without preliminary feature screening, respectively. Concatenating TabNet outputs with the top-N raw features enhances discrimination, and TabNet's sparse masking highlights key drivers, including the debt to asset ratio, return on equity, current ratio, industry growth, customer and supplier concentration, etc. These results support banks' credit decisions and risk management in supply chain finance.

The paper is structured as follows: Section 2 reviews prior research on SCF credit risk assessment; Section 3 describes the risk-assessment framework, feature selection, data augmentation, and two-stage deep learning framework; Section 4 reports empirical results on predictive accuracy, robustness, and interpretability; and Section 5 concludes with key insights, limitations, and directions for future research.

2. Literature Review

2.1. Identification of Risk Factors in SCF

A relatively systematic theoretical framework has emerged in the study of credit risk in supply chain finance (SCF), with existing research primarily focusing on two key dimensions: the identification of risk sources and the innovation of assessment methodologies. Regarding risk sources, Chopra and Meindl pointed out that SCF models face complex risks, encompassing both systemic risks arising from external market fluctuations and idiosyncratic risks driven by firm-specific factors [6]. Leveraging big data analytics, Fan and Su further demonstrate the central role of the financing entity's credit risk within the broader SCF risk architecture [7].

Concurrently, scholars have advanced indicator frameworks that operationalize these risk sources for evaluation. For example, Shashank and Thomas highlight the importance of market conditions and industry characteristics and note that supply chain management capability moderates risk transmission [8]. Wang and Gao propose a three-dimensional evaluation framework spanning enterprise capability, organizational quality, and capital structure [9]. Abbasi and Wang integrate Internet-of-Things (IoT) information with traditional indicators to form a composite system covering financial and non-financial dimensions [10]. Complementary contributions include Li's systematic risk-assessment indicator set [11] and the four-dimensional risk-control framework of Li and Shi [12], which collectively facilitate the transition from theoretical constructs to implementable assessment schemes.

2.2. Risk Assessment Models in SCF

With the advancement of digital technologies, machine learning (ML) and deep learning (DL) have become key tools for enhancing risk assessment. Early ML approaches employed single classifiers such as SVM and BPNN [13], while ensemble methods achieved higher accuracy, e.g., the LR-ANN hybrid [14], RS-MultiBoosting [15], and GT-RotF-LB framework [16]. Optimization strategies further improved performance. Zhang et al. proposed a Firefly-Algorithm–optimized SVM (FA-SVM) [17], and Xu et al. employed an AdaBoost-IMPA-SVM framework to enhance classification accuracy [18].

DL techniques, capable of learning complex features, have advanced predictive performance in diverse contexts, including supply chain risk detection with DCNNs [19], distributed CNN-based fraud detection [20], and fuzzy logic-enhanced models [21]. Graph neural networks (GNNs) also introduced novel perspectives, with GCNs facilitating efficient credit propagation in supply chains [22].

More recently, Stacking has gained traction as a hierarchical ensemble strategy that integrates heterogeneous models to improve generalization. Applications span multiple domains, including small-sample disease recognition [23], lung cancer risk prediction [24], and credit default evaluation in P2P lending [25], where it achieved the lowest error rates.

2.3. Data Augmentation Model

In supply chain finance (SCF) credit risk research, class imbalance and sample scarcity are common challenges. Early work primarily employed random oversampling/undersampling and their variants. For example, Shen et al. combined an improved SMOTE with LSTM and AdaBoost, enhancing prediction under small-sample conditions [26]. With the development of generative adversarial networks (GANs), researchers began using synthetic data generation to mitigate minority-class shortages. Liu et al. employed GANs with stacked autoencoders to improve performance in high-dimensional and small-sample contexts [27], while Li et al. integrated kernel density estimation with GANs to balance distributions and improve robustness [28].

More recently, diffusion models have emerged as a powerful alternative for data augmentation. By iteratively adding and removing noise, these models learn to generate realistic samples from simple distributions. Ho et al. [29] established their theoretical foundation, and Nichol and Dhariwal [30] demonstrated their superiority over GANs in quality and diversity. For tabular data, Kotelnikov et al. [31] proposed TabDDPM, which jointly models continuous and categorical variables and consistently outperforms GANs and VAEs. Building on this, Ceritli et al. applied TabDDPM to electronic health records, showing improved data quality and utility [32].

Diffusion models represent a class of generative modeling paradigms based on Markov chains. The core idea is to transform a simple initial distribution—typically a standard Gaussian—into a complex target data distribution through a two-phase process: forward diffusion (progressive noise addition) and reverse denoising. A deep neural network is trained to learn this reverse diffusion process, gradually refining noisy inputs to generate samples resembling the target data distribution. TabDDPM is a diffusion model tailored for tabular data. It supports both numerical and categorical features and is broadly applicable to various tabular datasets. The overall architecture is shown in Figure 1.

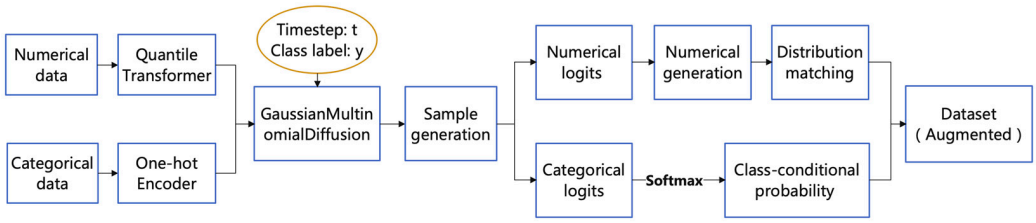


Figure 1. TabDDPM data augmentation framework for classification.

2.4. TabNet Model

Traditional deep learning models, such as CNNs and GCNs, often encounter difficulties in handling tabular data due to sparsity and mixed attribute types. To address this limitation, Arik et al. proposed TabNet, a specialized architecture that incorporates a sparse feature selection mechanism to simulate tree-based learning and improve interpretability [33]. TabNet has since shown strong performance across classification tasks. For example, Chowdhury et al. found it outperformed traditional ensemble and MLP models [34]; Naseer et al. used its attention mechanism for effective URL feature extraction [35]; and Khademi et al. applied TabNet in a multi-stage framework that significantly improved minority class prediction [36].

We propose a Bayesian-optimized TabNet for feature extraction and fusion. TabNet uses a differentiable attention mechanism for automated feature selection and nonlinear relationship discovery. Its learnable feature mask selects the most informative features, enabling end-to-end feature selection and dimensionality reduction. The model's predicted probabilities, combined with the top-N important features identified through the sparsemax attention mechanism, form an enhanced feature vector for classification (see Figure 2 for the TabNet feature engineering framework based on Bayesian optimization).

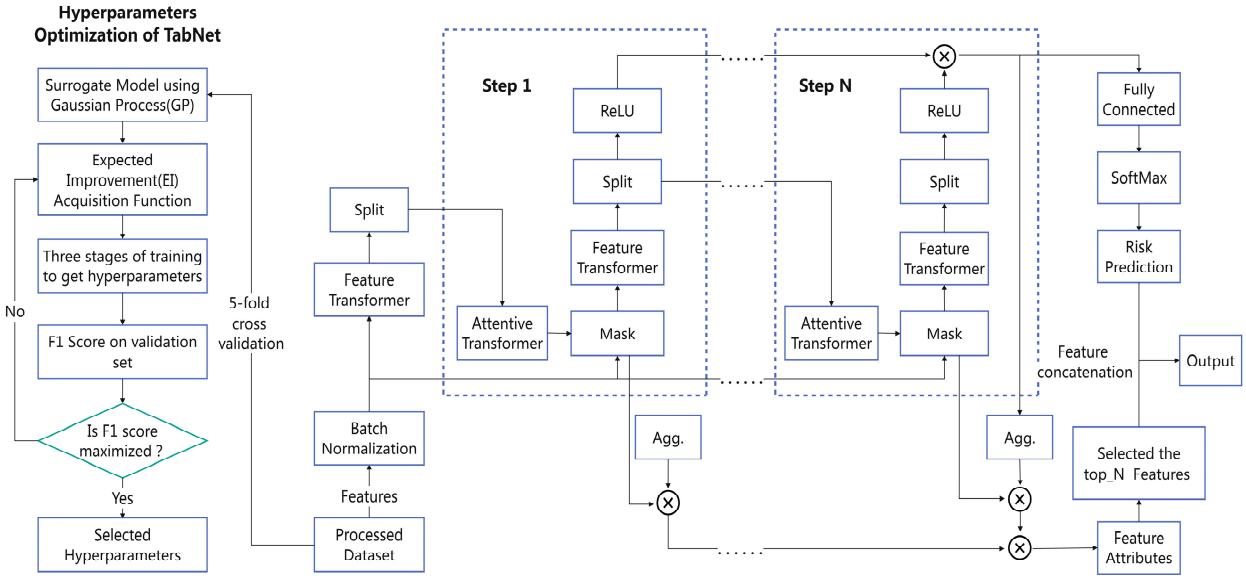


Figure 2. Improved TabNet feature engineering framework based on Bayesian optimization.

3. Materials and Methods

3.1. Dataset Description and Preparation

Based on a structured review and synthesis of the literature [37–39], candidate indicators were identified and organized into a hierarchical framework comprising four primary dimensions—overall status of SMEs, overall status of core enterprises, supply chain conditions, and the macroeconomic environment—further refined into secondary and tertiary

levels. The resulting instrument comprises 77 indicators and is intended to provide a comprehensive characterization of SME credit risk under SCF (refer to Appendix A Table A1). Enterprises are labeled "risky" if at least one of the following holds: (i) an SME's interest-bearing debt ratio (IBDR) exceeds the lower bound of the industry-specific SME IBDR reported for the corresponding statistical year [40], or (ii) within that year, the enterprise exhibits any default event—such as past-due acceptance of commercial bills, overdue debt, or bond default. Otherwise, the enterprise is labeled "non-risky". We assemble a sample of 360 SMEs listed on China's A-share market spanning the machinery, pharmaceuticals, electronics, automotive, and telecommunications industries during 2019–2023. Firm-level and supply chain data are obtained from Wind, iFinD, and CSMAR.

By consolidating over one thousand supply chain records involving the 360 SMEs, a dataset of 1800 samples with 77 credit risk evaluation indicators was constructed, comprising 309 risk-enterprise samples and 1491 non-risk-enterprise samples. Given the limited sample size, the preprocessed dataset was randomly partitioned into training and testing sets in an 8:2 ratio to facilitate robust model training.

Our feature space contains three categories: binary indicators for whether an attribute is present, one-hot representations of categorical fields, and real-valued variables. The preprocessing procedures also imputed missing values: numerical features used median imputation (see Equation (1)), and categorical features used mode imputation.

The data $= \{X, y\}$, with $X \in R^{n \times m}$ and $y \in \{0,1\}^n$. The integers n and m denote the sample and feature counts, respectively. The label y_i indicates the presence (1) or absence (0) of represents a risky enterprise. Continuous features were scaled to [0,1] using min–max normalization (see Equation (2)).

$$x_{ij} = \begin{cases} X_{ij}, & \text{if } X_{ij} \neq \text{NaN} \\ \text{median}(X_j), & \text{if } X_{ij} = \text{NaN} \end{cases} \quad (1)$$

$$x_{ij}^{'} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (2)$$

3.2. Feature Selection Model Based on GA-LightGBM

To address the weak correlations between most features and labels, as well as the redundancy inherent in high-dimensional data, we propose a feature selection framework that integrates the global search capability of the Genetic Algorithm (GA) with the efficient evaluation of LightGBM. LightGBM provides rapid and accurate subset assessment to guide GA evolution, thereby enabling a scalable solution for high-dimensional analysis.

In the initialization phase, the population is randomly generated, with each individual encoded as a binary feature mask representing a feature subset, while ensuring at least one feature is retained. Key GA parameters, including population size, number of iterations, and mutation probability, are predefined. During fitness evaluation, LightGBM with five-fold cross-validation is employed, and the mean F1-score of each subset is adopted as the fitness metric. To reduce computational overhead, a caching mechanism is incorporated to avoid redundant evaluations.

For genetic operations, tournament selection is used to preserve high-performing individuals, multi-point crossover is applied to recombine subsets, and mutation is introduced with a predefined probability to maintain population diversity. To mitigate premature convergence, an early-stopping and restart mechanism is implemented: when no improvement is observed over several iterations, the mutation rate is increased, and the population is reinitialized, while elite individuals are retained to preserve evolutionary direction. The pseudocode of the proposed model is summarized as follows Algorithm 1:

Algorithm 1: The pseudocode of the proposed model

Input:
X: Feature matrix
y: Target variable
pop_size: Population size
max_iter: Maximum number of generations
mutation_rate: Initial mutation rate
early_stop_rounds: Early stopping rounds for no improvement
Initialization:
Randomly generate initial population (binary masks)
Set best_mask, best_score, fitness_cache, no_improve_counter
Evolution:
For epoch in 1 to max_iter:
1. Evaluate fitness of all individuals (LightGBM + CV, with caching)
2. Keep elite: If improved, update best_mask and best_score.
3. If no improvement for early_stop_rounds:
- Restart population (retain elite)
- Increase mutation_rate
4. Select parents via tournament selection
5. Apply multi-point crossover to produce offspring
6. Mutate offspring (ensure at least one feature selected)
7. Replace current population with offspring
Final:
Train LightGBM on features selected by best_mask
Compute feature importance (gain) and SHAP values
Output:
best_mask, selected_indices, best_score, shap_importance

3.3. Risk Assessment Model

3.3.1. Data Enhancement Based on TabDDPM

The ratio of risk to non-risk samples is approximately 1:5 in the dataset, leading to a moderate class imbalance problem. To address this issue, we employ the TabDDPM model to generate synthetic samples exclusively for the minority (high-risk) class, thereby alleviating the imbalance. The procedure begins with a label distribution analysis on the pre-filtered training set to identify the minority class, followed by bootstrap resampling with replacement to construct training subsets. Next, a Gaussian–multinomial diffusion framework based on a conditional multilayer perceptron (MLP) is constructed. The diffusion process is governed by a cosine annealing schedule for noise control, while the AdamW optimizer is employed to minimize the variational lower bound loss during training. During inference, synthetic samples are generated using repeated independent sampling ($K = 5$), dropout-based randomization, and sample concatenation techniques. The generated data are then calibrated via affine transformation, and a Euclidean-distance-based filtering mechanism is introduced to retain only high-quality synthetic samples. Considering the limited quantity and sparse distribution of minority instances, excessive oversampling may lead to unrealistic boundaries or distorted feature distributions. Therefore, the final strategy restricts augmentation to no more than twice the number of original minority samples, ensuring a balanced yet reliable enhancement.

3.3.2. Feature Engineering Based on TabNet

Previously, we proposed a Bayesian-optimized TabNet for feature extraction and fusion. We now introduce an ensemble learning framework based on TabNet, which leverages its differentiable attention mechanism for automatic feature selection and modeling of higher-order feature interactions. The newly extracted features are then input into a two-stage ensemble learning model. The TabNet model is implemented in PyTorch (2.1.0), with hyperparameters optimized via Bayesian Optimization (BO) using a Gaussian Process (GP) surrogate, the F1-score as the objective, and the Expected Improvement (EI) strategy. The search is conducted over five hyperparameters: network width (n_d), number of decision steps (n_{steps}), γ, sparse regularization coefficient (λ_{sparse}), and learning rate (lr). To prevent overfitting, the training is limited to 50 epochs with an early stopping criterion: training terminates if validation loss shows no improvement for ten consecutive epochs.

This study adopts an adaptive top-N feature truncation to set the feature budget. Concretely, We perform stratified K-fold cross-validation and, within each fold, train a TabNet model only on the training split to obtain a fold-specific global importance ranking of the original features. Let be a grid of candidate subset sizes. For each, we take the top-N features under this ranking and concatenate the fold-specific TabNet class probability as an additional feature. For a given random seed s, we average F1 across folds and compute its standard error. Using the 1-SE rule, we then pick the smallest N whose mean F1 is within one SE of the seed's best mean F1; call it N_s^*. We then refit TabNet on the full training set to extract the corresponding top-N_s^* feature set and assess stability across seeds via pairwise Jaccard similarity of these sets. The final budget N^* is the median of $\{N_s^*\}$ over seeds. With this fixed, TabNet is refit on all training data to produce class probabilities, which are concatenated with the top-N^* raw features to train the final stacking classifier; the same composition is used at test time.

We use the two-stage deep learning ensemble model on the feature-selected dataset without augmentation (Selected Dataset) as an example. Across seven random seeds, the resulting distribution of N^* (see Figure 3) has a median of 32; for the decisive seed, the corresponding within-fold mean F1 curve is shown in Figure 4. Accordingly, we set the feature budget to N = 32. The mean pairwise Jaccard similarity among the seed-specific top-N^* feature sets is 0.764 (see Figure 5), indicating good cross-seed stability. These results suggest that a moderately sized subset retains sufficient discriminative power and generalization capacity, whereas adding too many low-importance features introduces noise, dilutes signal, and reduces model stability. Additionally, to support practical deployment in risk warning and financial control systems, a dynamic threshold optimization strategy is applied. Among all candidate thresholds that ensure a recall of at least 0.80 for high-risk cases, the one yielding the highest F1-score is selected as the final decision threshold and is adopted in the stacking ensemble model.

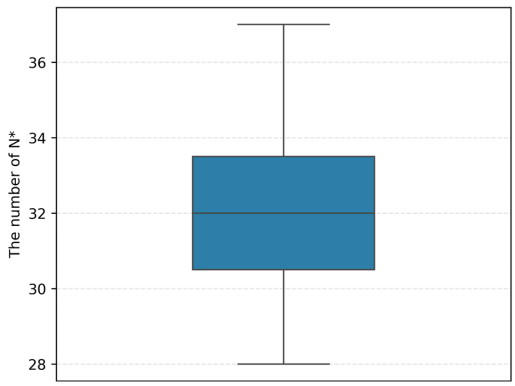


Figure 3. Distribution of selected N* across random seeds (Median = 32).

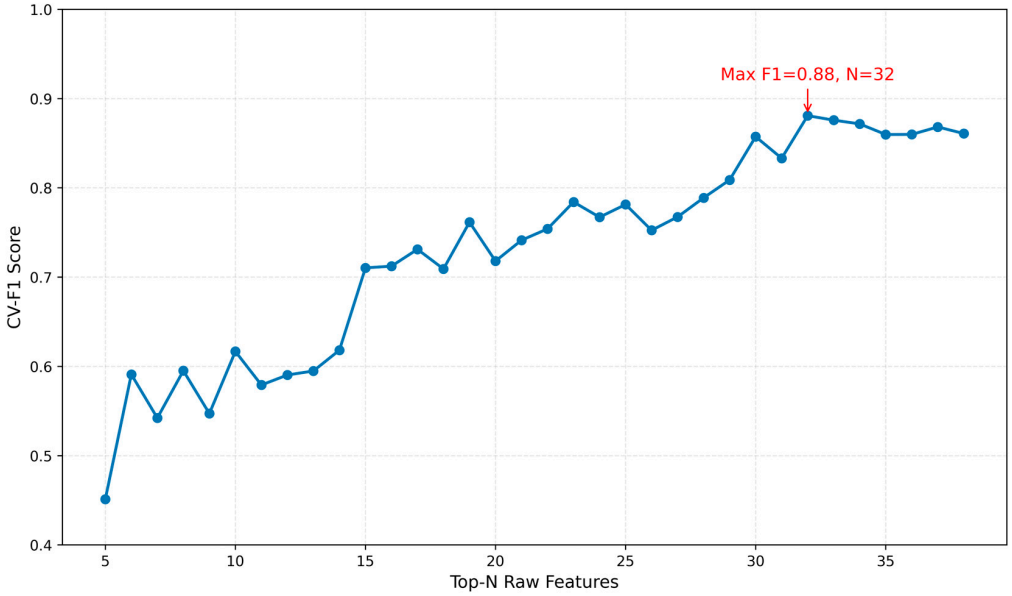


Figure 4. F1-score trend across different numbers of TabNet-selected raw features.

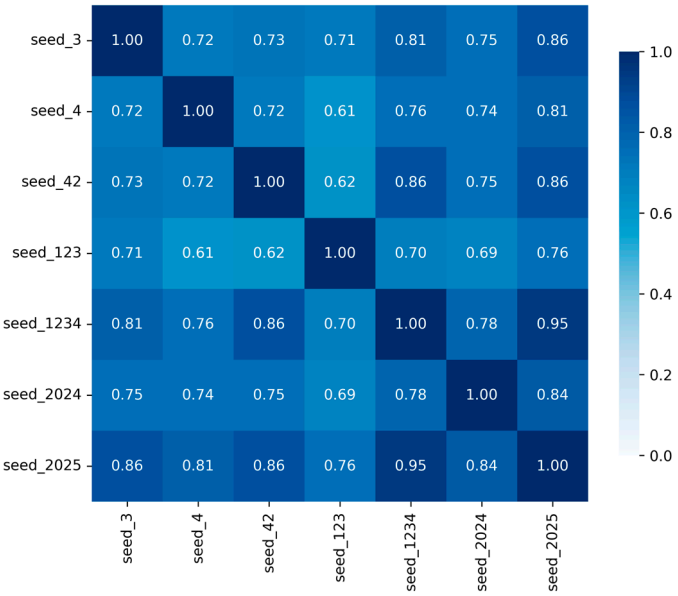


Figure 5. Jaccard heatmap of top-N feature overlap across seeds (Mean = 0.764).

3.3.3. Stacking Classifier

To enhance generalization and feature modeling in structured data classification, we propose a classification framework that integrates TabNet-based feature augmentation with a stacking ensemble strategy. A two-layer stacking architecture is employed to improve robustness. In the first layer, base learners are selected to emphasize algorithmic heterogeneity and complementarity. Three models with distinct inductive biases are adopted—KNN, XGBoost, and CatBoost—which collectively capture local structure, nonlinear trends, and categorical attributes. In the second layer, LightGBM serves as the meta-learner for final integration.

The entire training process is conducted with stratified K-fold cross-validation to preserve class distribution across training and validation sets. Each base learner is trained independently on different folds to avoid information leakage and enhance ensemble diversity. Their output probability features are then passed to the LightGBM meta-learner,

which efficiently aggregates predictive information and optimizes decision-making. The overall framework is illustrated in Figure 6.

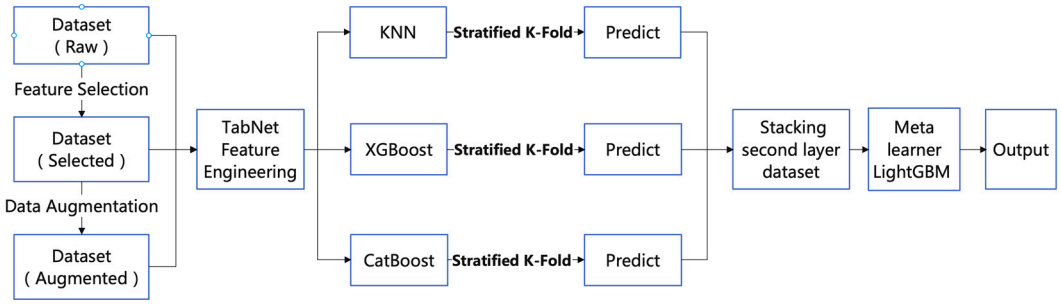


Figure 6. Two-stage ensemble learning framework of the stacking model improved by TabNet.

To validate the effectiveness of the proposed approach, multiple comparative experiments are conducted, including single-model baselines versus the stacking ensemble, as well as analyses with and without feature selection and data augmentation. For a comprehensive evaluation of model performance in supply chain financial credit risk prediction, seven metrics are adopted: Accuracy, F1-score, Area Under the ROC Curve (AUC), Precision (risk class), Recall (risk class), Precision (non-risk class), and Recall (non-risk class).

4. Results

The results of the experiments evaluating the proposed methodology are presented and discussed in this section. In terms of the performance measures, the findings and specific values are provided and discussed.

4.1. Feature Analysis Based on GA-LightGBM Preselection

The experimental results indicate that after 20 iterations (sensitivity analysis on iteration count refer to Appendix A Table A2), the performance of the GA-LightGBM feature selection model stabilized, yielding a consistent subset of 38 raw features for subsequent analysis. To elucidate the role of these features in model decision-making, we further employed SHAP analysis with the LightGBM model to provide a quantitative assessment of feature importance and contribution direction.

As illustrated in Figure 7, features with positive SHAP values are identified as risk-driving factors; larger values correspond to stronger contributions toward high-risk predictions. Representative risk-driving factors include the debt to asset ratio, cash to total debt ratio, inventory, accounts receivable proportion, supplier concentration, number of employees, and the debt to asset ratio of core enterprises, underscoring the influence of financial pressure and supply chain structural risks. Conversely, features with negative SHAP values are regarded as stability factors; more negative values indicate a stronger mitigating effect against high-risk classification. Stability factors primarily include operating profit margin, the net profit margin of core enterprises, net profit, operating revenue, and return on assets (ROA), all reflecting sound profitability and operational efficiency. Furthermore, the SHAP value associated with information synergy is also negative, suggesting that effective information sharing and collaboration within the supply chain enhance systemic stability and mitigate potential risks.

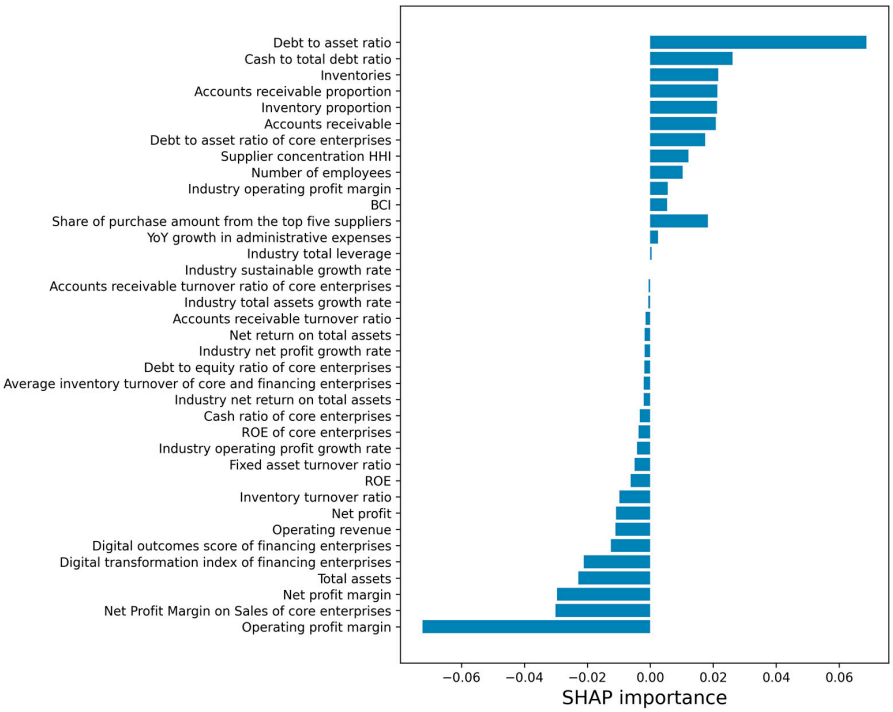


Figure 7. SHAP importance analysis of selected features (Including feature impact direction).

Description: The vertical axis lists the features, with the most influential ones displayed at the extremes. The horizontal axis represents SHAP values, which quantify the contribution of each feature to the model output and are ranked by predictive impact.

4.2. Visual Assessment of TabDDPM Data Augmentation

To assess the quality of synthetic samples and their relationship with real data, supervised UMAP was applied for dimensionality reduction and visualization. As shown in Figure 8, the scatter distribution demonstrates clear separation among real majority samples (gray), real risk samples (purple), and synthetic risk samples (yellow) in the embedded space, indicating that the proposed model effectively captures the intrinsic structural characteristics of different classes.

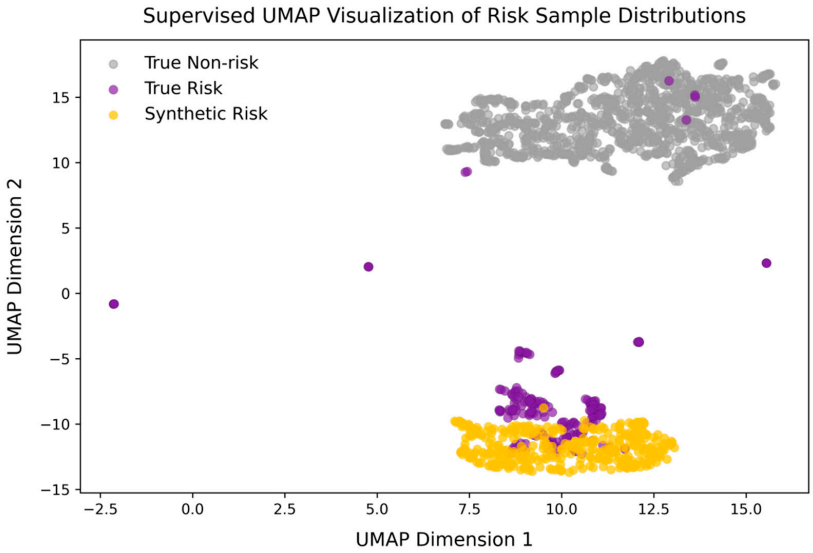


Figure 8. Visualization of data generated by TabDDPM (Reduced with Supervised UMAP).

Although slight discrepancies between real and synthetic samples remain, such differences are often beneficial in practice, as they may reveal latent features not fully covered by the original data, thereby enhancing model generalization. In addition, a Kolmogorov–Smirnov (KS) test was performed to evaluate distributional consistency between original minority and synthetic samples. The results show that KS statistics across all features are below 0.015 with $p = 1.0$, confirming that the generated samples preserve distributional fidelity and provide reliable augmentation.

4.3. Models' Performance on the Test Dataset

TabDDPM augmentation generates quality-controlled synthetic samples by simulating the underlying feature distribution, showing good application potential and robustness in improving the performance of imbalanced classification. The experimental results (see Table 1) indicate that the Stacking model improved by TabNet (TabNet-Stacking), when applied to the dataset augmented with TabDDPM, consistently outperforms the non-augmented dataset. Specifically, the augmented model achieves an Accuracy of 97.78%, an F1-score of 92.45%, and an AUC of 97.60%, compared to 96.39%, 87.50%, and 98.14%, respectively, for the baseline. Furthermore, the model maintains stable Precision and Recall for the non-risk class, demonstrating enhanced consistency and robustness.

Table 1. Summary of Performance Metrics for the TabNet-Stacking Model and Baseline Models.

Model	Dataset	Accuracy	F1 Score	AUC	Precision (Y = 1)	Recall (Y = 1)	Precision (Y = 0)	Recall (Y = 0)
XGBoost	Raw	0.9306	0.7899	0.9596	0.7705	0.8103	0.9632	0.9536
LightGBM	Raw	0.9389	0.8429	0.9666	0.8551	0.8310	0.9588	0.9654
KNN	Raw	0.8556	0.6486	0.9203	0.5161	0.8727	0.9738	0.8525
CatBoost	Raw	0.9472	0.8288	0.9612	0.8364	0.8214	0.9672	0.9704
Stacking	Raw	0.9417	0.8531	0.9686	0.8133	0.8971	0.9754	0.9521
TabNet-Stacking	Raw	0.9500	0.8333	0.9628	0.8654	0.8036	0.9643	0.9770
XGBoost	Selected	0.9417	0.8205	0.9723	0.7869	0.8571	0.9732	0.9572
LightGBM	Selected	0.9583	0.8718	0.9752	0.8361	0.9107	0.9833	0.9671
KNN	Selected	0.9306	0.7934	0.9457	0.7742	0.8136	0.9631	0.9535
CatBoost	Selected	0.9528	0.8545	0.9784	0.8704	0.8393	0.9706	0.9770
Stacking	Selected	0.9639	0.8829	0.9836	0.8909	0.8750	0.9770	0.9803
TabNet-Stacking	Selected	0.9722	0.9057	0.9762	0.96	0.8571	0.9742	0.9934
XGBoost	Augmented	0.9444	0.8246	0.9633	0.8103	0.8393	0.9702	0.9638
LightGBM	Augmented	0.9639	0.8850	0.9799	0.8772	0.8929	0.9802	0.9770
KNN	Augmented	0.9472	0.8257	0.9463	0.8491	0.8036	0.9642	0.9737
CatBoost	Augmented	0.9556	0.8519	0.9727	0.8846	0.8214	0.9675	0.9803
Stacking	Augmented	0.9667	0.8909	0.9818	0.9074	0.8750	0.9771	0.9836
TabNet-Stacking	Augmented	0.9778	0.9245	0.9760	0.9800	0.8750	0.9774	0.9967

Description: The table reports the predictive performance of TabNet-Stacking and baseline models on the test set for risk prediction. "Raw" denotes the preprocessed dataset, "Selected" refers to the dataset after feature selection based on GA-LightGBM, and "Augmented" corresponds to the dataset augmented with TabDDPM.

From a methodological perspective, concatenating the outputs of TabNet with the top-N raw features enables the joint exploitation of raw feature representations and deep abstract features, thereby strengthening discriminative capability. On the test set, the final TabNet-Stacking model achieves an F1-score of 90.57% (see Table 1), representing a 2.3 percentage point improvement over the ensemble model without TabNet integration. These results substantiate the synergistic advantage of embedding the improved TabNet model within the stacking ensemble framework.

Among all evaluated models, the Stacking and TabNet-Stacking variants deliver the highest overall performance, particularly in Recall (Y = 1) and F1-score, where they consistently outperform alternative approaches. This superiority arises from the ensemble's ability to integrate the responses of heterogeneous learners to diverse feature representations,

thereby enhancing its capacity to capture complex decision boundaries and heterogeneous structures. As illustrated in the confusion matrix (see Figure 9), the TabNet-Stacking model applied to the dataset augmented with TabDDPM achieves a type I error rate of 0.3% and a type II error rate of 12.5% on the test set. These results indicate that the model not only maintains high precision in identifying risk enterprises but also effectively controls both false positives and false negatives. Nevertheless, recall performance for risk cases still presents room for further improvement.

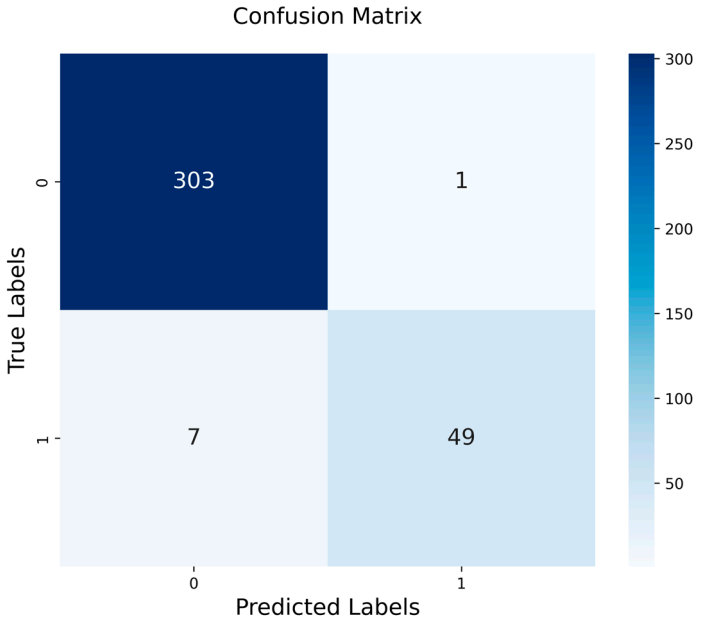


Figure 9. Confusion matrix of the TabNet-Stacking model on the augmented dataset.

To further validate the results, the two-stage ensemble learning model was evaluated on three datasets (Raw, Selected, and Augmented) using ten-fold cross-validation, and the classification outcomes were analyzed with a Wilcoxon signed-rank test. The results show that the performance difference between the models before and after augmentation is statistically significant (see Table 2). The model applied to the augmented dataset achieved an F1-score of 93.05%, corresponding to improvements of 4.90 and 10.89 percentage points compared to the Selected and Raw datasets, respectively. The non-parametric test further confirms significant differences in mean F1-scores across the three datasets ($p < 0.05$), demonstrating the effectiveness of TabNet feature modeling and fusion, as well as the synergistic advantage of the proposed strategy in enhancing classification performance. This also shows that GA-lightGBM feature selection can effectively filter out redundant features, improve model performance, and reduce computing resource consumption.

Table 2. Performance comparison of the TabNet-Stacking model across different datasets using the Wilcoxon Signed-Rank Test.

Data	Mean F1	SD	Compared To	*p*-Value	Mean Diff	95% CI
Augmented	0.9305	0.0172	-	-	0.00	-
Selected	0.8815	0.0305	Augmented	0.002	0.0490	[0.0268, 0.0712]
Raw	0.8216	0.0498	Augmented	0.002	0.1089	[0.0761, 0.1417]

Figure 10 presents the top 20 features identified by the TabNet model through its sparse feature masking mechanism as the most influential in prediction. Among these, the debt-to-asset ratio, return on equity, and current ratio exhibit markedly higher scores, highlighting the central role of capital structure and profitability indicators in driving model decisions.

Elevated leverage compresses covenant headroom and heightens the sensitivity of default risk to earnings volatility; strong profitability provides a loss-absorption buffer; and the current ratio proxies the firm's capacity to service short-term obligations without resorting to distress asset sales. Several industry-level attributes, including the industry net profit growth rate and industry operating profit growth rate, also rank prominently, underscoring the influence of external market conditions on firm-level classification outcomes. By shifting external demand conditions and credit availability, these industry-level indicators influence SMEs' cash flows and the collect-ability of receivables from core enterprises. Furthermore, customer concentration and supply chain indicators, such as the customer concentration Herfindahl–Hirschman Index (HHI) and supplier concentration, are assigned substantial weights, reflecting that dependence on a narrow set of counterparties increases vulnerability to idiosyncratic shocks and operational disruptions that propagate along the chain. By leveraging the sparse feature masking mechanism, TabNet effectively suppresses redundant information, thereby providing interpretable, data-driven support for financial decision-making, risk monitoring, and corporate credit evaluation.

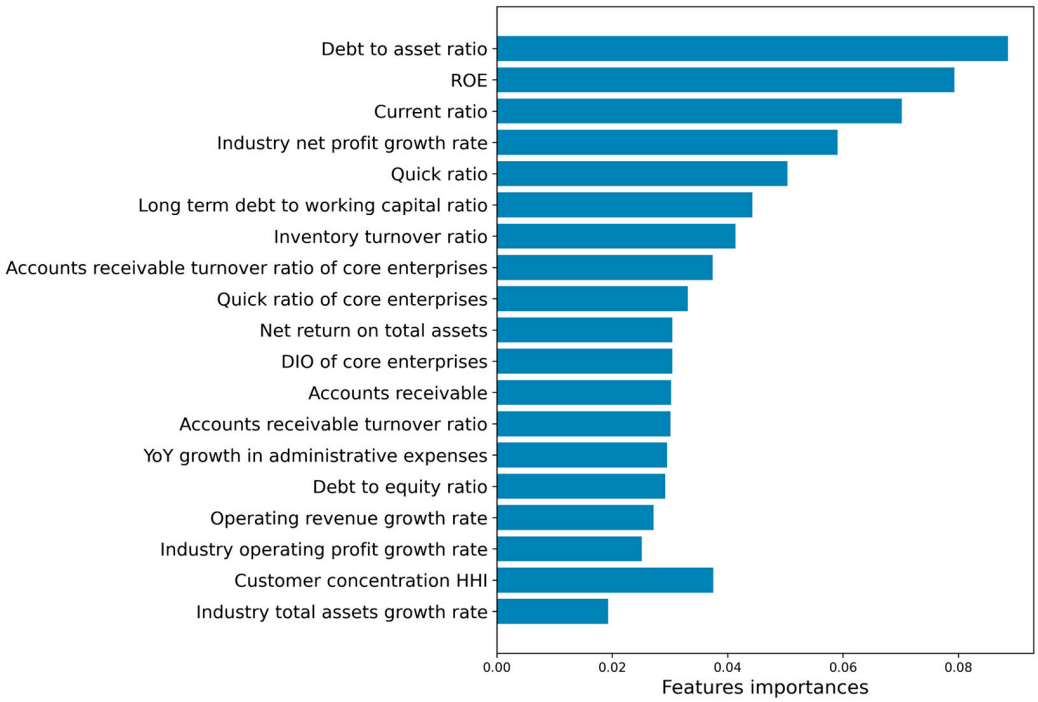


Figure 10. Top 20 features identified by the sparse feature mask mechanism of TabNet based on contribution.

5. Discussion

This paper proposes a framework for predicting credit risk in supply chain finance for small and medium-sized enterprises (SMEs), combining TabNet deep feature extraction with a stacking ensemble learning approach. Through a multi-stage optimization strategy, the framework effectively improves prediction performance and model generalization, especially under imbalanced data conditions. The main innovations and conclusions are summarized as follows:

1. A comprehensive credit risk assessment system was developed, incorporating 77 indicators that span financial metrics, supply chain stability, and macroeconomic factors. Using the GA-LightGBM feature selection method, 38 key features were identified, and SHAP value analysis revealed that debt-related metrics, including debt-to-asset ratio, cash-to-debt ratio, and supplier concentration, are critical drivers of risk pre-

diction. Profitability indicators, such as operating profit margin and ROA, were also found to be essential indicators of business stability.

2. A data augmentation technique using TabDDPM is applied to the dataset to mitigate the imbalance between risk and non-risk samples. The results from the Wilcoxon signed-rank test show that the augmented data significantly improves the optimal model performance, with the F1-score increasing by 4.9% and prediction variance reducing by 1.3%.
3. This paper proposes a two-stage ensemble model framework that integrates TabNet feature engineering with a Stacking fusion strategy. In the first stage, a Bayesian-optimized TabNet model is employed to evaluate the importance of features using its sparse attention mask mechanism. The adaptively selected Top-N raw features are then combined with the softmax probabilities output by TabNet's fully connected layer. This enables secondary feature selection and enhancement, creating an enriched hybrid feature space. TabNet's learnable feature mask mechanism not only selects the most informative features but also reduces dimensionality, contributing to a more efficient and interpretable model. In the second stage, base learners such as XGBoost, CatBoost, and KNN are used, with LightGBM serving as the meta-learner to build the stacking ensemble model. Experimental results show that under augmented data conditions, the model achieves an F1-score of 92.45%, a 2.4% improvement over the ensemble model without TabNet-based feature fusion and selection. The TabNet-Stacking ensemble model effectively combines local and global feature representations, significantly outperforming traditional machine learning models.

The model's interpretable output provides a valuable reference for decision-making by lenders and investors. Characteristics with positive values, such as debt-to-asset ratio, cash-to-debt coverage ratio, accounts receivable ratio, supplier/customer concentration, and core enterprise leverage ratio, are generally associated with increased risk and can be managed as risk drivers in credit approval, risk pricing, term agreements, and risk early warning mechanisms. Furthermore, capital structure, liquidity indicators, and industry growth have a significant impact on the probability of default and are recommended for inclusion in industry credit limit setting and scenario-based stress testing frameworks. For regulators and policymakers, the stabilizing effect of information collaboration provides a rationale for promoting investment in electronic invoicing infrastructure, accounts receivable registration and title confirmation systems, and core enterprise information disclosure standards, thereby helping to reduce information friction within the supply chain.

The main limitations of this research stem from situational and time constraints. The study's sample was limited to small and medium-sized enterprises (SMEs) in China, which may not fully represent the diversity of global SMEs. Additionally, the research did not explore features related to the temporal aspect of changes in enterprise credit, such as time series data, which could further enhance the predictive power of the model. Furthermore, the credit risk assessment system could be expanded to include additional dimensions, such as dynamic economic factors and the information flow between enterprises within the supply chain.

Future research can strengthen predictive performance along several fronts. First, we will broaden the dataset across industries to improve generalization and robustness. Recent research shows that large language models can extract predictive signals from unstructured text [41,42]. We will also use a large language model (e.g., ChatGPT 5) to derive sentiment and event indicators from annual reports, MD&A, supply chain announcements, and news about core firms and counterparties, and integrate these signals into the pipeline as text augmented features. Second, improve synthetic sample generation and class-imbalance handling—e.g., via adversarial augmentation and principled resampling—to increase sam-

ple diversity and representativeness. Third, move beyond our current cross-sectional evidence by explicitly modeling temporal dynamics. We will build a real-time, incremental multimodal pipeline that fuses supply chain transaction records and corporate sentiment with tabular features, and adopt temporal architectures (Temporal Fusion Transformer, temporal CNNs, time-stratified TabNet) enriched with macroeconomic and policy covariates; where appropriate, state-space or regime-switching specifications will be considered. Evaluation will rely on expanding-window backtests and strict out-of-time validation. For deployment, we will implement drift monitoring of features and scores with threshold recalibration or Bayesian parameter updating, and apply time-series–aware conformal prediction to ensure calibrated coverage. Additionally, Domain Adaptation techniques could be employed to develop industry-specific feature libraries for sectors such as manufacturing and biopharmaceuticals, addressing cross-industry data distribution discrepancies. The integration of counterfactual explanations and rule extraction techniques in real-world corporate cases could further clarify the pathways for risk conversion, providing financial institutions with precise and actionable risk management strategies and decision support.

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

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Framework for the Supply Chain Finance Risk Assessment Indicator System for SMEs.

Primary	Secondary	Tertiary	Primary	Secondary	Tertiary
Overall situation of SMEs	Basic information of the company	Number of employees	Overall situation of core enterprises		Revenue growth rate
		Total assets			Net return on total assets
		Operating revenue		Debt-paying ability	Current ratio
		Net profit			Quick ratio (acid-test ratio)
		Inventories			Cash ratio
		Accounts receivable		Operational capabilities	Debt to assets ratio
	Innovation and governance capabilities	General and administrative expenses (G&A)			Debt to equity ratio
		Year over year growth in administrative expenses			Accounts receivable turnover ratio
		Research and development expenses (R&D)			Days inventory outstanding (DIO)
		Year over year growth in R&D expenses			Total asset turnover ratio
		ESG composite score	Supply chain status	Information coordination	Average inventory turnover of core and financing enterprises

Table A1. *Cont.*

Primary	Secondary	Tertiary	Primary	Secondary	Tertiary
Overall situation of SMEs	Profitability	Return on equity (ROE)	Supply chain status	Level of digitalization	Digital outcomes score of financing enterprises
		Return on assets (ROA)			Digital application score of financing enterprises
		Net return on total assets			Digital transformation index of financing enterprises
		Return on invested capital (ROIC)		Customer concentration	Share of sales revenue from the top five customers
		Net profit margin			Customer concentration Herfindahl-Hirschman Index (HHI)
		Operating profit margin		Supplier concentration	Share of purchase amount from the top five suppliers
	Debt-paying ability	Quick ratio			Supplier concentration Herfindahl-Hirschman Index (HHI)
		Current ratio		Cash flow coordination ability	Accounts receivable proportion
		Debt to asset ratio			Inventory proportion
		Interest coverage ratio		Years of cooperation	Years of trading relationship between SMEs and core enterprises
		Debt to equity ratio	Macro environment	Overall economic environment	GDP growth rate
		Long term debt to working capital ratio			Local GDP growth rate
		Cash to total debt ratio			Business climate index (BCI)
	Operational capabilities	Days inventory outstanding (DIO)		Industrial Policy	Number of regional supply chain finance policies
		Days sales outstanding (DSO)		Overall industry situation	Industry return on assets (ROA)
		Inventory turnover ratio			Industry net return on total assets
		Accounts receivable turnover ratio			Industry return on equity (ROE)
		Current asset turnover ratio			Industry operating profit margin
		Fixed asset turnover ratio			Industry total assets growth rate
		Total asset turnover ratio			Industry ROE growth rate
		Operating profit growth rate			Industry net profit growth rate
		Operating revenue growth rate			Industry operating profit growth rate
		Total operating revenue growth rate			Industry sustainable growth rate
	Credit history	External guarantees			Industry financial leverage
		Credit history			Industry operating leverage
	Profitability	Return on equity (ROE)			Industry total leverage
		Net return on total assets			Industry concentration
		Net Profit Margin on Sales			

Table A2. Sensitivity of the Iteration Cap in GA-LightGBM Feature Selection (CV-F1, Subset Size, Jaccard).

Max_Iter	Mean_CV_F1	Std_CV_F1	Mean_n_Features	Std_n_Features	Mean_Jaccard
10	0.7368	0.0084	41.0000	3.0912	0.3978
20	0.7376	0.0078	38.0000	3.7417	0.3592
30	0.7376	0.0078	38.0000	3.7417	0.3592

References

1. Zhu, Y.; Jia, R.; Wang, G.J.; Xie, C. A review of supply chain finance risk assessment research: Based on knowledge graph technology. *Syst. Eng.-Theory Pract.* **2023**, *43*, 795–812. [CrossRef]
2. Zhao, J.F.; Li, B. Credit risk assessment of small and medium-sized enterprises in supply chain finance based on SVM and BP neural network. *Neural Comput. Appl.* **2022**, *34*, 12467–12478. [CrossRef]
3. Yu, H.; Sun, C.; Yang, X.; Yang, W.; Shen, J.; Qi, Y. ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data. *Knowl.-Based Syst.* **2016**, *92*, 55–70. [CrossRef]
4. Akila, S.; Reddy, U.S. Cost-sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for credit card fraud detection. *J. Comput. Sci.* **2018**, *27*, 247–254. [CrossRef]
5. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Trans. Syst. Man Cybern. Part B* **2009**, *39*, 539–550.
6. Chopra, S.; Meindl, P. *Supply Chain Management: Strategy, Planning, and Operation*, 7th ed.; Pearson Education: Boston, MA, USA, 2001.
7. Fan, Z.F.; Su, G.Q.; Wang, X.Y. Research on Credit Risk Evaluation and Risk Management of Small and Medium Enterprises under Supply Chain Financial Model. *J. Cent. Univ. Financ. Econ.* **2017**, *12*, 10.
8. Rao, S.; Goldsby, T.J. Supply chain risks: A review and typology. *Int. J. Logist. Manag.* **2009**, *20*, 97–123. [CrossRef]
9. Wang, C.; Gao, G.J. Research on corporate credit risk assessment in supply chain finance business based on LSSVM. *Credit. Ref.* **2015**, *6*, 4.
10. Abbasi, W.A.; Wang, Z.; Zhou, Y.; Shahzad, H. Research on measurement of supply chain finance credit risk based on Internet of Things. *Int. J. Distrib. Sens. Netw.* **2019**, *15*, 1550147719874002. [CrossRef]
11. Li, Y.X. Supply Chain Financial Risk Assessment. *J. Cent. Univ. Financ. Econ.* **2011**, *10*, 6.
12. Li, Z.H.; Shi, J.Z. Risk identification and control of supply chain finance-comparison based on online and offline models. *J. Commer. Econ.* **2015**, *8*, 3.
13. Sang, B. Application of genetic algorithm and BP neural network in supply chain finance under information sharing. *J. Comput. Appl. Math.* **2020**, *384*, 113170. [CrossRef]
14. Zhu, Y.; Xie, C.; Sun, B.; Wang, G.J.; Yan, X.G. Predicting China's SME Credit Risk in Supply Chain Financing by Logistic Regression, Artificial Neural Network and Hybrid Models. *Sustainability* **2016**, *8*, 433. [CrossRef]
15. Zhu, Y.; Zhou, L.; Xie, C.; Wang, G.J.; Nguyen, T.V. Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *Int. J. Prod. Econ.* **2019**, *211*, 22–33. [CrossRef]
16. Belhadi, A.; Kamble, S.S.; Mani, V.; Benkhati, I.; Touriki, F.E. An ensemble machine learning approach for forecasting credit risk of agricultural SMEs' investments in agriculture 4.0 through supply chain finance. *Ann. Oper. Res.* **2025**, *345*, 779–807. [CrossRef]
17. Zhang, H.; Shi, Y.; Yang, X.; Zhou, R. A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance. *Res. Int. Bus. Financ.* **2021**, *58*, 101482. [CrossRef]
18. Xu, Z.H.; Rao, Z.Y.; Huang, X.D. Research on credit risk evaluation of small and medium-sized enterprises in rare earth supply chain finance using hybrid enhanced machine learning algorithm. *Rare Met. Cem. Carbides* **2024**, *52*, 94–102.
19. Tang, Q.; Lu, Y.; Wang, B.; Li, Z. A Deep Convolutional Neural Network Based Risk Identification Method for E-Commerce Supply Chain Finance. *Sci. Program.* **2022**, *2022*, 6298248. [CrossRef]
20. Zhou, H.; Sun, G.; Fu, S.; Fan, X.; Jiang, W.; Hu, S.; Li, L. A Distributed Approach of Big Data Mining for Financial Fraud Detection in a Supply Chain. *Comput. Mater. Contin.* **2020**, *64*, 1091. [CrossRef]
21. Hu, Z. Statistical optimization of supply chain financial credit based on deep learning and fuzzy algorithm. *J. Intell. Fuzzy Syst.* **2020**, *38*, 7191–7202. [CrossRef]
22. Zhao, M.; Zhang, X.N.; Dong, P.P. Credit risk prediction for small and medium-sized enterprises based on graph convolutional networks. *Finance* **2024**, *14*, 575–588. [CrossRef]
23. Nagro, S. A stacked ensemble approach for symptom-based monkeypox diagnosis. *Comput. Biol. Med.* **2025**, *191*, 110140. [CrossRef]
24. Alonso, E.; Calle, X.; Gurrutxaga, I.; Beristain, A. Survival Stacking Ensemble Model for Lung Cancer Risk Prediction. *Stud. Health Technol. Inform.* **2024**, *321*, 155–159. [CrossRef] [PubMed]

25. Yin, W.; Kirkulak-Uludag, B.; Zhu, D.M.; Zhou, Z.X. Stacking ensemble method for personal credit risk assessment in Peer-to-Peer lending. *Appl. Soft Comput.* **2023**, *142*, 110302. [CrossRef]
26. Shen, F.; Zhao, X.C.; Kou, G.; Alsaadi, F.E. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Appl. Soft Comput.* **2021**, *98*, 106852. [CrossRef]
27. Liu, Y.; Li, S.; Yu, C.; Lv, M. Research on green supply chain finance risk identification based on two-stage deep learning. *Oper. Res. Perspect.* **2024**, *13*, 100311. [CrossRef]
28. Li, L.S.; Zhang, H.F.; Song, R.Z. GKN-Stack: An Ensemble Deep Learning Framework for Blood Glucose Forecasting Based on Medical Examination Data. *IEEE Access* **2024**, *12*, 178089–178103. [CrossRef]
29. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239. [CrossRef]
30. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
31. Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; Babenko, A. TabDDPM: Modelling Tabular Data with Diffusion Models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; Volume 202.
32. Ceritli, T.; Ghosheh, G.O.; Chauhan, V.K.; Zhu, T.; Creagh, A.P.; Clifton, D.A. Synthesizing Mixed-type Electronic Health Records using Diffusion Models. *arXiv* **2023**, arXiv:2302.14679. [CrossRef]
33. Arik, S.O.; Pfister, T. TabNet: Attentive Interpretable Tabular Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021.
34. Chowdhury, M.N.H.; Reaz, M.B.I.; Ali, S.H.M.; Crespo, M.L.; Ahmad, S.; Salim, G.M.; Haque, F.; Ordóez, L.G.G.; Islam, M.J.; Mahdee, T.M. Deep learning for early detection of chronic kidney disease stages in diabetes patients: A TabNet approach. *Artif. Intell. Med.* **2025**, *166*, 103153. [CrossRef]
35. Naseer, M.; Ullah, F.; Saeed, S.; Algarni, F.; Zhao, Y. Explainable TabNet ensemble model for identification of obfuscated URLs with features selection to ensure secure web browsing. *Sci. Rep.* **2025**, *15*, 9496. [CrossRef]
36. Khademi, S.; Hajiakhondi, Z.; Vaseghi, G.; Sarrafzadegan, N.; Mohammadi, A. FH-TabNet: Multi-Class Familial Hypercholesterolemia Detection via a Multi-Stage Tabular Deep Learning. *arXiv* **2024**, arXiv:2403.11032.
37. Liu, J.; Liu, S.M.; Li, J.; Li, J.Z. Financial credit risk assessment of online supply chain in construction industry with a hybrid model chain. *Int. J. Intell. Syst.* **2022**, *37*, 8790–8813. [CrossRef]
38. Hu, H.Q.; Zhang, L.; Zhang, D.H. Research on Credit Risk Assessment of Small and Medium Enterprises from the Perspective of Supply Chain Finance-A Comparative Study Based on SVM and BP Neural Network. *Manag. Rev.* **2012**, *24*, 11.
39. Yuwono, M.A.; Rachmawati, D. The Role of Risk Management in Improving the Operational Effectiveness of SMEs: A Systematic Review. *JISR Manag. Soc. Sci. Econ.* **2024**, *22*, 1–25. [CrossRef]
40. Kuang, H.B.; Du, H.; Feng, H.Y. Construction of credit risk indicator system for small and medium-sized enterprises under supply chain finance. *Sci. Res. Manag.* **2020**, *41*, 209–219.
41. LoGrasso, M. Could ChatGPT have earned abnormal returns? A retrospective test from the U.S. stock market. *Mod. Financ.* **2025**, *3*, 112–132. [CrossRef]
42. Lopez-Lira, A.; Tang, Y. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv* **2023**, arXiv:2304.07619. [CrossRef]

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.