ISSN 0305-0483
ELSEVIER
omega
The International Journal of Management Science
www.elsevier.com/locate/omega

A three-stage prediction model for firm default risk: An integration of text sentiment analysis

Xuejiao Ma , Tianqi Che , Qichuan Jiang

PII: S0305-0483(24)00171-3
DOI: https://doi.org/10.1016/j.omega.2024.103207
Reference: OME 103207

To appear in: *Omega*

Received date: 4 August 2024
Accepted date: 26 September 2024

Please cite this article as: Xuejiao Ma , Tianqi Che , Qichuan Jiang , A three-stage prediction model for firm default risk: An integration of text sentiment analysis, *Omega* (2024), doi: https://doi.org/10.1016/j.omega.2024.103207

Highlights

- A three-stage prediction model of the ARA-SVM-MPSO hybrid model for default risk prediction.
- Emphasize the importance of text information in predicting default risk.
- The association rule algorithm is used to select features to enhance the prediction accuracy and reduce the calculation burden.
- The MPSO is used to optimize SVM to reduce the influence of parameter sensitivity on prediction results.

A three-stage prediction model for firm default risk: An integration of text sentiment analysis

Xuejiao Ma[a], Tianqi Che[a], Qichuan Jiang[b,*]

[a]School of Economics and Management, Dalian University of Technology, Liaoning, 116024, China
[b]School of Finance, Dongbei University of Finance and Economics, Liaoning, 116023, China

[*]Corresponding author.
jiangqichuan123@163.com

Abstract: Predicting firm default risk is vital for financial institutions to avert significant economic losses, making the enhancement of its prediction precision both imperative and intricate. This research introduces a three-stage prediction model, including association rule algorithm (ARA), support vector machine (SVM) and modified particle swarm optimization algorithm (MPSO). Features selected by ARA are used as inputs for SVM, and penalty parameter and kernel parameter of SVM is optimized by MPSO that uses the adaptive inertia weight. The importance of text sentiment variables are emphasized to predict firm default risk. In the first stage, feature selection seeks to curtail the dimensions of both financial and non-financial variables. The empirical findings validate the efficacy of the ARA, revealing a strong correlation between text sentiment and default risk. The subsequent two stages deploy the SVM, refined by the MPSO, to predict the default risk. Compared with renowned models, the proposed model displays superior prediction precision and a reduced computational overhead. This research furnishes a potent instrument for regulators and firms alike, aiding in mitigating prospective default risks and forestalling broader economic upheavals.

Keywords

Text sentiment; Three-stage prediction model; Association rule algorithm; Support vector machine; Modified particle swarm optimization algorithm

1. Introduction

Listed firms hold a cornerstone position in the economic framework of nations globally, serving as vital economic linchpins (Li et al., 2023). Their salient role in economic and societal advancement cannot be overstated. Credit defaults by such firms not only jeopardize their capital procurement capabilities and reputation but also result in significant financial losses to stakeholders, further destabilizing the broader market economy (Xie et al., 2023). Hence, for financial entities and investors, the meticulous execution of credit risk management and predictive modeling of defaults becomes paramount. With these predictive insights, financial institutions are able to effectively assess the repayment ability of public companies and take advance steps to protect their loan principals from potential business fluctuations. Investors can mitigate potential financial setbacks by preemptively identifying stocks with credit risks. By accurately anticipating their developmental trajectory, listed firms can strategically navigate through financial landscapes, ensuring timely financing and borrowing to stave off credit calamities. Regulatory bodies can, in turn, streamline their oversight, concentrating on imminent default risk scenarios. Proactively addressing the default risks of listed firms can attenuate economic downturns, forestall fiscal crises, and bolster the robustness of the financial sector (Elhoseny et al., 2022).

A survey of extant literature reveals primary focus areas: the ramifications of macro-level (Shi et al., 2024); Fatouh and Giansante, 2023) and firm-level determinants (Cathcart et al., 2020; Korangi et al., 2023) on firm default risk. Early academic pursuits were predominantly centered on financial determinants of default risk, given their direct influence on firm performance (Altman, 1968). However, as economies evolved, scholarly emphasis shifted towards non-financial parameters. Contemporary research has delved into governance dynamics (Chen et al., 2017), corporate social responsibility (Yang et al., 2023) and other firms (Ren et al., 2023). However, these traditional indicators tend to contain only historical information and may not be able to predict corporate default risk in a timely manner. Textual metrics provide nuanced insights beyond historical narratives that may capture a company's overall performance. Therefore, incorporating textual information into risk projections is critical. Notably, the text sentiment offers a lucid reflection of a firm's outlook on macroeconomic landscapes and its future prospects, thereby influencing default likelihoods. Although the academic circles have explored text information (Bhatia, 2019; Huang et al., 2022; Xia et al., 2024), there is no consensus on the extraction of text emotion and its application in the process of default risk prediction. This research endeavors to embed text sentiment in default prediction, aiming to unearth incremental insights for enhanced predictive acuity.

Given the myriad factors influencing firm default risk, incorporating all into a singular model could exacerbate computational challenges and introduce overfitting concerns. Thus, prior to embarking on default prediction with the credit dataset, it is imperative to implement feature selection. This process curtails the dimensionality of data, emphasizing the salience of select features for subsequent analysis. Feature selection methodologies can be broadly categorized into three predominant classes. The first is filter method (Rabiee et al., 2024). This technique entails the sifting of sample features per established criteria, yielding a reduced feature subset. The filter method is characterized by its operational simplicity and rapid algorithm execution. A significant drawback, however, is its lack of model dependency, rendering it ineffective in procuring a feature system optimally tailored to a specific model. The second is wrapper method (Simsek et al., 2021). The ultimate model's performance is employed as the evaluative benchmark for the feature subset. While often yielding superior outcomes compared to the filter method, it is computationally demanding, entails protracted computation periods, and the resultant feature subset may exhibit reduced generalizability. The third is embedded method (Simumba et al., 2022), and it seamlessly integrates the feature selection protocol within the learning algorithm itself. While it assimilates the model-specific feature selection benefits of the wrapper method and the computational efficiencies of the filter method, it is less adept at discerning features with high correlations. The above three methods often require a trade-off between accuracy and efficiency, and may ignore hidden relationships between features (Simic et al., 2023; Lee et al., 2022; Zhu et al., 2023). Given the limitations inherent in the aforementioned methodologies, this study gravitates towards the association rule algorithm (ARA) for feature extraction. This algorithm excels in unearthing intriguing correlations amongst large data itemsets. By scrutinizing extant knowledge patterns, it facilitates the revelation of latent novel principles.

Regarding predictive modeling, seminal works have predominantly fixated on statistical models (Leow and Crook, 2016), econometric models (Zhang et al., 2022), and machine learning models (Shi et al., 2024). Traditional statistical and econometric models, despite their merits, often suffer from rigid assumptions, a predominantly linear scope, and a retroactive nature, making them less adept at prognosticating future firm default risks. In contrast, machine learning offers the prowess to discern intricate patterns within voluminous data sets. This study, therefore, champions the machine learning paradigm for risk prediction. Among the myriad of machine learning algorithms, the support vector machine (SVM) emerges as a formidable candidate owing to its aptitude for high-dimensional challenges, nonlinear feature interactions, and superior generalization capacities (Mahbobi et al., 2023). However, SVM's efficacy is significantly contingent on its parameters, bringing disparate predictive outcomes (Guerrero et al., 2024). Scholars have introduced various algorithms to optimize SVM parameters (Jiang et al., 2024). Complex algorithms, such as GA, may requires large parameter tuning and function evaluation. And some algorithms, such as FA and AFSA, may focus on local neighborhood solutions. Therefore, PSO is chosen to optimize the parameters of the SVM. To ameliorate this, this research introduces adaptive inertia weights into the Particle Swarm Optimization (PSO) algorithm, enhancing its global and local optimization capacities. Subsequently, the refined PSO algorithm fine-tunes the SVM's penalty and kernel function parameters, aiming to elevate the model's corporate default risk predictive performance.

This research endeavors to address several pivotal inquiries: Can text sentiment augment the precision of default risk prediction? What methodologies are most effective for feature selection to curtail variable dimensions? Which prediction models strike an optimal balance between accuracy and applicability in predicting default risk? To this end, we propose a three-stage prediction model, harnessing ARA, SVM and MPSO, encompassing feature selection, prediction, and optimization. The logic of the three-stage model is: the ARA is utilized to select the most relevant features to reduce dimension of variables, and variables selected by ARA are inputs of SVM. The MPSO with adaptive inertia weight aims to optimize the penalty parameter and kernel parameter of SVM, which can be used for firm default risk. The ARA-SVM-MPSO outperforms the single prediction model in the follow aspects: First, single SVM has low prediction accuracy since it needs to determine the penalty parameter and kernel parameter of SVM subjectively, which can loose the balance between accuracy and generalization. Second, the number of variables related to default risk is large, which may increase the calculation burden of the model and reduce accuracy. Therefore, the proposed ARA-SVM-MPSO hybrid model is able to improve the prediction accuracy of single models because traditional PSO is modified by using the adaptive inertia weight to enhance its ability in searching the global optimum and the parameters in the SVM are determined by the MPSO. Additionally, by combining with ARA, the proposed model can achieve the feature selection by using the most relevant variables to simplify the calculation burden of the model.

By applying the ARA-SVM-MPSO hybrid model, we obtain the following findings. First, the text sentiment variable plays a crucial role in predicting the firm default risk. Except the text sentiment variable, compared with non-financial indicators, financial indicators exert a more important influence on default risk. Profitability and cash flow rank first and two among financial indicators. Second, as for the proposed ARA-SVM-MPSO hybrid model, ARA has better performance in feature selection than the comparing methods, including principal component analysis, grey correlation, Lasso and entropy weight method. Third, SVM optimized by the MPSO can generate better prediction results than statistical models, machine learning methods and some deep learning models. The MPSO applied in this study is able to optimize the parameters in SVM accurately by searching the global optimum, and optimization algorithms based on swarm intelligence performs better than gradient-based algorithms. Results from the robustness tests confirm the effectiveness of the proposed model that when changing support and confidence degrees of ARA, and altering the population size, iteration number and learning factor of MPSO, the model can still have consistent prediction results.

Our findings culminate in three key contributions. First, we propose a three-stage prediction model of the ARA-SVM-MPSO hybrid model for default risk prediction. In consideration of the text sentiment, this model has better performance than only using traditional data used in previous literature (Zhang et al., 2023; Séverin and Veganzones, 2021; Albuquerque et al., 2019). Our research highlights the importance of text information in predicting default risk. Second, most of the current research adopts traditional feature selection methods including filter (Rabiee et al., 2024), wrapper (Simic et al., 2023), and embedded methods (Liu et al., 2021), which need to balance efficiency with accuracy. The ARA-SVM-MPSO hybrid model leverages the association rule algorithm, enhancing the prediction accuracy and reduce the calculation burden. By calibrating varying support and confidence thresholds, this algorithm adeptly identifies pertinent variables predicated not solely on direct correlations with credit default risk, but also on the synergistic effects of variable combinations. Third, this study uses MPSO to optimize SVM, aiming to reduce the influence of parameter sensitivity on prediction results. Compared with deep learning paradigms, it offers diminished prediction intricacies yet maintains equivalent accuracy levels. Simultaneously, its results surpass those of rudimentary regression models. In a departure from the prevailing trend of pursuing intricate models, our approach champions a confluence of accuracy and simplicity.

The rest of this paper is structured as follows: Section 2 delves into the literature review. Section 3 elucidates the three-stage prediction model and its evaluative techniques. Section 4 showcases our empirical findings, and Section 5 provides conclusions and policy implications.

2. Literature review

2.1 Influencing factors of firm default risk

The existing literature reveals the factors influencing corporate default risk mainly from a macroeconomic perspective and a firm-specific perspective, as summarized in **Table 1**. At the macro level, scholars have considered the impact of macroeconomic indicators on default risk. For example, Shi et al. (2024) provide an in-depth analysis of the fuzzy equivalence between local macroeconomic indicators and rated credit levels. Fatouh and Giansante (2023), on the other hand, focus on the impact of GDP volatility. In addition to economic factors, technological factors such as fintech advances are also explored by He et al. (2023). In addition, external shocks at the macro level will also affect default risk. At the policy level, Zhang et al. (2023) revealed that the institution of a carbon emissions trading framework effectively curtailed firm default vulnerabilities. As for financial level, financial network shocks (Ahn et al., 2024) as well as the financial crisis (Yfanti et al., 2023) will affect default risk, leading to credit risk contagion.

At the micro scale, pioneering research has predominantly focused on discerning the effects of financial metrics on a firm's susceptibility to default. Altman (1968) pioneered the development of the five-factor *Z*-score model, deploying a combination of critical financial indicators and serving as a proxy to quantify creditors' exposure to credit risk. In a more contemporary study, Cathcart et al. (2020) posited that financial leverage bears significant influence on the likelihood of firm default. Extending this line of inquiry, Ohlson (1980) deduced from his empirical analysis that three pivotal factors-the firm's size, its capital structure, and profitability level—exerted considerable influence on firm default risk. Séverin and Veganzones (2021) provided definitive findings about the capacity of earnings management–based variables to improve bankruptcy prediction models. Korangi et al. (2023) focused on the role of fundamental, market and pricing data on predicting default risk. In synthesizing this discourse, it becomes evident that financial metrics have not only historically served as pivotal indicators for assessing default risk but also continue to be instrumental control variables in contemporary risk-related scholarship.

Recent scholarship has broadened the analysis of corporate default risk by integrating non-financial determinants alongside traditional financial metrics. This shift is exemplified by several studies that examine diverse influences. For instance, Chen et al. (2017) highlighted how enhancing CEO competence and board recruitment strategies can act as safeguards against default. Another emerging area of research is the influence of corporate social responsibility and default risk (Albuquerque et al., 2019). For example, Yang et al. (2023) investigated the impact of environmental factors, such as pollution emissions, on default risk. Additionally, Fu and Trigilia (2024) found that the improved corporate disclosure could reduce firms' pay-for-performance sensitivity, thereby reducing default risk. Beyond the internal dynamics of individual firms, recent studies also consider inter-firm relationships. Ren et al. (2023) innovatively employed supply chain variables as key predictors of a firm’s credit rating, while Beaver et al. (2019) emphasized the significance of group information for predicting default within parent-subsidiary networks.

The trajectory of research addressing determinants of firm default risk has evolved substantially, transitioning from an initial focus on purely financial variables to an inclusive examination of non-financial indicators. While traditional numerical metrics remain instrumental in predicting default tendencies, they largely encapsulate historical information. However, in the current era marked by unprecedented data diversity, textual indicators present an intriguing avenue, offering nuanced insights that extend beyond historical narratives. These insights potentially capture a holistic representation of a firm, particularly with respect to its prospective trajectories, thereby augmenting the empirical base for default prognosis. Historically, the predominant corpus of academic research was circumscribed by the constraints of numerical data modeling, rendering the invaluable contributions of textual data largely unexplored. However, the emergence of advanced computational techniques heralds renewed academic interest in the predictive potential of textual determinants in measuring default vulnerability. Existing studies have extracted word features (Katsafados et al., 2024), such as word distribution structures (Bhatia, 2019) and text sentiment (Huang et al., 2022), from texts including loan texts (Jiang et al., 2018), Q&A texts (Xia et al., 2024), analysts' reports (Roeder et al., 2022), and social media (Bartov et al., 2023).

Table 1. Summary of the influencing factors

Period	Region	Influencing factors			Ref.
		Macro level	Firm level		
			Financial	Non-financial	
1998-2013	China	√	√	√	Shi et al. (2024)
1948-2019	UK	√			Fatouh and Giansante (2023)
2011-2020	China	√	√		He et al.(2023)
2000-2018	17 countries	√	√		Zhang et al.(2023)
2011	German	√	√		Ahn et al. (2024)
2004-2020	EU and US	√			Yfanti et al. (2023)
1946-1965	US		√		Altaman(1968)
2005-2015	6 European countries		√		Cathcart et al.(2020)
1970-1976	US		√		Ohlson（1980）
2016-2017	French		√		Séverin and Veganzones (2021)
1990-2020	US		√		Korangi et al. (2023)
2001-2014	US		√	√	Chen et al.(2017)
2004-2015	US	√	√	√	Albuquerque et al. (2019)
2016	China		√	√	Yang et al. (2023)
-	-			√	Fu and Trigilia (2024)
2006-2020	North America		√	√	Ren et al. (2023)
2005-2012	Globe		√	√	Beaver et al. (2019)

2.2 Feature selection of firm default risk

Feature selection serves as a dimensionality reduction technique aimed at selecting a subset of features from the input data by removing irrelevant, redundant or noisy features while maintaining the model performance (Kozodoi et al., 2019). By adeptly filtering these features, one can achieve significant dimensional reduction without compromising the intrinsic information encapsulated within the original dataset. As shown in **Table 2**, contemporary methodologies predominantly revolve around three paradigms: filter, wrapper, and embedded techniques (Zhang et al., 2021).

The filter method accentuates the filtering of sample features in alignment with predetermined criteria, eventually yielding a feature subset. Fundamentally, this technique scores each feature based on specific evaluative benchmarks, thereby ascertaining their significance. An intriguing facet of the filter method is its prerogative to undertake feature selection prior to model training, remaining agnostic to the learning mechanism in place (Rabiee et al., 2024). Seminal contributions in this domain include Relief algorithm (Kira and Rendell, 1992), mutual information (Sun et al., 2019) and chi-square (Rezaei et al., 2022). In addition to the traditional methods, scholars have continuously improved the algorithms. For example, Zhu et al. (2023) proposed a fast unsupervised feature selection algorithm, named Compactness Score (CSUFS), to further improve the feature selection performance with lower computational complexity. Xu et al. (2024) proposed a feature filter and group evolution hybrid feature selection algorithm (FG-HFS) for high-dimensional gene expression data. In summation, the filter method's allure is manifold: it offers straightforward implementation, outpaces alternative methodologies in computational velocity, and remains unencumbered by the underlying learning model.

The wrapper method integrates distinct machine learning algorithms during feature selection, tailoring feature subsets to optimize specific learning algorithm outcomes (Simsek et al., 2021). The prime example for the wrapper method is recursive feature elimination, which selects a subset of the most relevant features to train the model and removes the weakest features until the specified number of features is reached (Simic et al., 2023). Yoganarasimhan (2020) used forward-selection wrapper to derive the set of optimal features. In a subsequent exploration, scholars have developed the wrapper method in conjunction with different algorithms, such as NSGA-II genetic algorithm (Kozodoi et al., 2019), optimum-seeking method (Kou et al., 2021) and Kolmogorov-Smirnov statistic (Lu et al., 2022). It's crucial to recognize that, in comparison to the filter method, the wrapper method, while offering potential accuracy advantages, often exhibits reduced efficiency. This stems from its intricate execution requirements and the necessity for iterative model training-a process that can be resource-intensive. Yet, despite these challenges, the subsets generated via the wrapper method often display heightened precision, showcasing its potential merit in critical applications.

The embedded method seamlessly integrates feature selection within the model learning process, obviating the need for evaluation of potential feature subsets (Simumba et al., 2022). The Lasso algorithm is a popular example for embedded method. It employs *L1* regularization during the compression of regression coefficients, results in certain regression coefficients equating to zero, leading to sparse regression solutions. Many scholars have improved this classical algorithm. For example, Jiang et al. (2021) proposed the square root fused LASSO, combining the square root loss function and joint penalty functions. Motivated by Lasso, Liu et al. (2021)

formulated the feature selection and acquisition problem as a nonlinear discrete optimization problem that minimizes prediction error subject to a budget constraint. Additionally, decision tree-based algorithms have become mainstay, which necessitate repeated node splits, gauging feature importance at each juncture. Features of heightened significance are selected as split nodes, continuing until a condition-satisfying tree emerges (Zhou et al., 2021a). It was then refined by scholars on the basis of cost. For example, Lee et al. (2022) proposed the BCP algorithm aims to achieve rapid convergence by considering only a limited number of variables in a solution search tree. Labbé et al. (2023) used a single linkage approach to obtain a tree diagram at the lowest possible total cost. Given its integrated approach, the embedded method augments the efficiency of learning algorithms since it does not need to be trained every time a feature subset is selected (Jiménez-Cordero et al., 2021).

While the above methodologies offer distinct advantages, each also has limitations. The filter method offers suboptimal learning and fitting capacities, largely due to its exclusion of subsequent learners during feature selection. The wrapper method's selected features often lack universality, necessitating reselection upon changing the learning algorithm, improving the computational intensity of this method. Embedded methods remain acutely sensitive to sample noises and demand considerable computational resources. Meanwhile, the above methods may ignore certain hidden information in the dataset. To circumvent these limitations, this study employs an association rule algorithm (ASA), a superior big data analytics technique, to extract relationships and dependencies between input features in the dataset (Sariyer et al., 2021).

Table 2. Summary of the feature selection methods

Methodology	Advantages	Disadvantages	Ref.
Filter	Fast and sample More independent	Less accurate	Rabiee et al. (2024), Kira and Rendell (1992), Sun et al. (2019), Rezaei et al. (2022), Zhu et al. (2023), Xu et al. (2024)
Wrapper	Higher accuracy	Risk of overfitting Less efficiency	Simsek et al. (2021), Simic et al. (2023), Yoganarasimhan (2020), Kozodoi et al. (2019), Kou et al. (2021), Lu et al. (2022)
Embedded	Less computationally intensive	Less interpretable	Simumba et al. (2022), Jiang et al. (2021), Liu et al. (2021), Zhou et al. (2021a), Lee et al. (2022), Labbé et al. (2023), Jiménez-Cordero et al. (2021)

2.3 Prediction models of firm default risk

Over the years, a plethora of models have been conceived and refined to predict firm default risk. Among these, statistical models, characterized by their transparent prediction processes and rapid prediction capabilities, have been prevalently adopted in the realm of default risk prediction. These models encompass techniques such as multivariate discriminant analysis, survival analysis, Bayesian model, and Merton model among others. In their seminal work, Leow and Crook (2016) delved into the stability of parameter estimations within discrete survival models. A more contemporary approach was adopted by Rozo et al. (2023), who incorporated web browsing metrics as predictors within a survival analysis framework to predict the probability of corporate defaults. Furthermore, Berloco et al. (2023) employed Bayesian spatio-temporal models with an objective to discern the contagion effects stemming from liquidity crises, subsequently aiming to predict short-term corporate defaults. Akyildirim et al. (2023) extended the Merton structural credit risk model for counterparty credit risk calculation in the context of calculating the credit value adjustment. However, it is crucial to acknowledge the limitations inherent in these models. Statistical models used in default risk prediction often require adherence to multiple assumptions and primarily focus on analyzing data and inferring from observed outcomes, highlighting the relationships within. This narrow focus can limit their effectiveness in making predictions about unseen data.

Econometric models primarily utilize regression analysis to discern the relationships between influential variables and default risk, facilitating the prediction of the latter. Notable examples among these are the logit model, probit model, and multivariate linear regression. Bonsall et al. (2017) formulated a logistic regression paradigm with the corporate credit rating as the dependent variable and managerial prowess as the independent variable. In a more granular approach, Zhang et al. (2022) segmented default instances into secondary and primary default events， choosing binomial logistic regression model, binomial logit model, and multinomial logit model to predict firm-specific default hazards. Expanding the methodological horizon, Jiang et al. (2021) integrated the LASSO penalty within the U-MIDAS logistic regression framework, crafting the U-MIDAS-Logit-GL model. Dumitrescu et al. (2022) improved logistic regression with non-linear decision-tree effects. However, Econometric models, much like statistical models, have inherent limitations. Besides, a Bayesian spatial hierarchical model is used by Calabrese (2023). Typically based on linear regression, these models require data to meet stringent conditions, including the presupposition of a normal data distribution—particularly crucial when testing hypotheses regarding regression coefficients. However, the actual collection of firm credit data might not adhere to this normal distribution, presenting a significant challenge. Additionally, before establishing a regression-based predictive model, it is necessary to ensure that the data are free from collinearity. In the current era of big data in credit analysis, where data features are characterized by their complexity and weak correlations, the linear approaches fundamental to econometric models often prove inadequate. This inadequacy primarily stems from their limited capability to accurately capture the nuanced relationships between credit attributes and associated risks, ultimately leading to reduced model efficacy.

The various complexities and the equifinality problem of Big Data make it difficult to apply traditional statistical techniques to creditworthiness evaluation, or credit scoring (Shi et al., 2024). Machine learning models, leveraging their ability to extract latent patterns and rules from extensive datasets, have been increasingly applied to challenges spanning classification, regression, and clustering domains (Chang et al., 2024). Particularly in default prediction, methodologies such as random forest, decision tree, neural networks, and support vector machines have gained considerable traction (Doumpos and Figueira, 2019; Liu et al., 2022). More recent algorithms include multi-layer perceptron artificial neural network (Wu et al., 2022), deep reinforcement learning (Zhang et al., 2024), Rotation Forest-Logit Boosting (Belhadi et al., 2020), LaGaBoost (Sigrist and Leuenberger, 2023) and . However, in the application of machine learning, there may be overfitting problems, and some models, such as neural networks, may need larger computing space and have poor interpretability. Within the ambit of machine learning, the SVM emerges as an exemplary archetype, primarily owing to its capacity to project data into higher-dimensional spaces, thereby facilitating data segregation within that space (Dutta, 2022). SVM can offer robustness, effectiveness in high-dimensional spaces, making it a powerful tool for predicting default risk. Empirical evidence suggests SVM's commendable accuracy benchmarks. For instance, Mahbobi et al. (2023) found that the SVM outperforms the other models in several performance metrics, and has the highest sensitivity with ALL-KNN sampling. Luo et al. (2020) proposed a new unsupervised kernel-free quadratic surface SVM (QSSVM) model to utilize the advantages of SVM and overcome the drawbacks of SVM-based clustering methods.

Above literature review is summarized in **Table 3**.

2.4 Research gaps

Upon reviewing the extant literature, it is evident that predicting default risk requires consideration of diverse factors, and various models have been employed to facilitate such prediction. Nevertheless, there exist discernible research gaps in this domain.

Primarily, financial indicators are widely recognized for their substantial influence on default risk (He et al., 2023; Korangi et al., 2023); however, non-financial factors also play crucial roles in risk assessment (Yang et al., 2023; Ren et al., 2023). Interestingly, while some research has explored using textual data to predict risk (Roeder et al., 2022), the results have not been definitively conclusive. To build on this, our study aims to refine the use of textual data by categorizing sentiments as either positive or negative, and then assessing the effectiveness of this approach through detailed comparative analysis. Secondly, as analyzed above, traditional feature selection methods, such as filter (Rabiee et al., 2024), wrapper (Simic et al., 2023), and embedded methods (Liu et al., 2021), often face trade-offs between efficiency and accuracy. Innovatively, this study adopts the association rule algorithm for feature selection, which is distinguished by its ability to identify not only individually significant features but also the combined effects of various factors on credit risk, thus revealing the underlying rules that govern inter-factor relationships. Finally, although numerous models have been developed for predicting default risk, there remains no consensus on which model is the most effective. As the amount of data increases, it is difficult for traditional models to obtain good performance (Shi et al., 2024). Some scholars have adopted machine learning methods such as neural networks and decision trees (Chang et al., 2024; Zhang et al., 2024), but there may be problems with overfitting. Although support vector machine performs well in high dimensional data space (Mahbobi et al., 2023), their dependence on parameters may affect the prediction results. This investigation introduces a three-stage hybrid model, leveraging machine learning algorithms to strike an optimal balance between prediction accuracy and computational efficiency.

Table 3. Summary of the prediction models

Period	Region	Methodology	Main conclusions	Ref.
2002-2011	UK	SA	The parameter estimates of survival models before and since the crisis are statistically different from each other.	Leow and Crook (2016)
2013-2017	UK	SA	The predictive accuracy of the probability of default model increases after adding the new web browsing.	Rozo et al. (2023)
2018-2019	Italian	BSTM	The model can detect risky situations that are not yet apparent, thus improving the effectiveness of short-term default prediction.	Berloco et al. (2023)
2013-2014	-	Merton	It extends the Merton structural credit risk model for counterparty credit risk calculation in the context of calculating the credit value adjustment mainly by estimating the probability of default.	Akyildirim et al. (2023)
1980-2010	-	LR	Management capacity improves credit ratings.	Bonsall et al. (2017)
1994-2014	China	LR	The model improves the accuracy of predicting firm default risk	Zhang et al.(2022)
2011-2018	China	LR	The model is effective in identifying important determinants from both high-frequency financial factors and low-frequency corporate governance profiles.	Jiang et al. (2021)
-	-	PLTR	The model performs better in out-of-sample than traditional linear and non-linear logistic regression, while being competitive relative to the random forest method.	Dumitrescu et al. (2022)
2016	London	BMSAR	The main advantages of the BMSAR model lie in its higher predictive accuracy and more conservative estimate of the expected shortfall compared to standard models available in the industry and in the literature.	Calabrese (2023)
1998-2013	China	NRS	This is the first time that NRSs have been developed to investigate fuzzy equifinality relationships so as to predict the creditworthiness of SMEs.	Shi et al. (2024)
2005	UK	GB, RF, LM,	This study identifies two effective algorithms, Random Forest and Gradient Boosting models, for credit risk detection.	Chang et al. (2024)
2012-2014	China	KNN, SVM, RF	Machine learning models are significantly better than that of logistic model.	Liu et al. (2022)
2016-2020	China	MLP ANN	The hybrid neural network model has a higher average correct classification rate	Wu et al. (2022)
2014-2018	China	DRL-Risk	The DRL-Risk approach can significantly improve the performance of credit risk prediction of SMEs in SCF compared with the baseline methods in recall, G-mean, and financial loss.	Zhang et al. (2024)
-	African	RotF-LB	the proposed RotF-LB ensemble approach has a good performance.	Belhadi et al. (2020)
1961-2020	US	LaGaBoost	The newly proposed "LaGaBoost frailty model" is overall large and exhibits strong variation over time.	Sigrist and Leuenberger (2023)
2005	US	SVM	Among the four models under this sampling technique, the SVM outperforms the other models in several performance metrics.	Mahbobi et al. (2023)
-	German, Australian, China	QSSVM	The proposed unsupervised QSSVM method outperforms well-known clustering methods, particularly in terms of classification accuracy.	Luo et al. (2020)

Note: *SA* is survival analysis; *BSTM* is Bayesian spatio-temporal models; *LR* is Logit regression; *PLTR* is penalised logistic tree regression; *BMSAR* is Bayesian spatial hierarchical model; NRS is neighborhood rough sets; *GB* is Gradient Boosting; *RF* is random forest; *KNN* is K-nearest neighbor; *SVM* is support vector machine; *MLPANN* is multi-layer perceptron artificial neural network; *DRL-Risk* is deep reinforcement learning; *RotF-LB* is Rotation Forest-Logit Boosting; *LaGaBoost* is latent Gaussian model boosting, *QSSVM* is quadratic surface SVM.

3. The three-stage model and model evaluation

3.1 The first stage: Association rule algorithm (ARA)

Features that are irrelevant or obscure the decision boundary can confuse the training process and lead to reduced performance. As the number of features increases, the computational cost becomes higher and the SVM may adapt to the noise in the training data instead of the signal in high dimensional cases. Therefore, we first utilize the association rule algorithm for feature selection before using SVM for default risk prediction. Originally introduced by Agarwal et al. (1993), the association rule algorithm stands as a cornerstone in data mining. This technique is adept at isolating and analyzing specific attributes, subsequently generating frequent patterns and associative models. The salient strength of association rules lies in their capacity to unearth significant associations or correlations from expansive datasets, which might be overlooked by conventional selection methods, thereby pinpointing recurrent sets of attribute values. Leveraging this method allows for the uncovering of latent knowledge and rules that govern the relationship between credit default risk and its determinants, resulting in more grounded and precise insights. The foundational definitions of association rules are delineated as follows:

Definition 1: Let $\boldsymbol{I=\left\{I_1,I_2,I_3\cdots I_n\right\}}$ be the set consisting of all items, and *D* is the set of database transactions, being a type of a non-empty subset of *I*. $\boldsymbol{T=\left\{T_1,T_2,T_3\cdots T_n\right\}}$ is the transactions in *D*, and each transaction *T* is a set of items, $\boldsymbol{T\subseteq I}$. An association rule is an equation of the form $X\rightarrow \boldsymbol{Y}$, where $\boldsymbol{X\Rightarrow Y}$, $\boldsymbol{Y\subset I}$ and $\boldsymbol{X\bigcap Y=\phi}$, and *X* does not intersect with *Y*. *X* is a precondition and *Y* is the corresponding association result.

Definition 2: Support degree. The support degree is the number of transactions containing *X* and *Y* in the transaction database as a percentage of all transactions, denoted as $\boldsymbol{Support(X\rightarrow Y)}$, as shown in ***Eq*** (1). The essence of the support degree is the probability of *X* and *Y* appearing in transaction *D* at the same time, which can measure the statistical importance of the association rule in the whole data set. The higher support degree indicates the higher relevance of the items in the sub-set is.

$$\boldsymbol{Support(X\rightarrow Y)}=\frac{\left\|\left\{\boldsymbol{T}\mid\left(\boldsymbol{X}\cup\boldsymbol{Y}\right)\subseteq\boldsymbol{T},\boldsymbol{T}\subseteq\boldsymbol{D}\right\}\right\|}{\left\|\boldsymbol{D}\right\|}\times 100\% \tag{1}$$

Definition 3: Confidence degree. For the association rule in the shape of $\boldsymbol{X\rightarrow Y}$, where *X* and *Y* are both item sets. Define the confidence level of the rule as the ratio of the number of transactions in the set of transactions *D* that contains both *X* and *Y* to the number of transactions in *D* that contain *X*, denoted as $\boldsymbol{Confidence(X\rightarrow Y)}$, as shown in ***Eq*** (2). The confidence level indicates the frequency of occurrence in the contained transactions, which can measure the reliability of reasoning through the rule. For a given rule, the higher the confidence level, the more likely it is that *Y* occurs in the transactions containing *X*.

$$\boldsymbol{Confidence(X\Rightarrow Y)}=\frac{\left\|\left\{\boldsymbol{T}\mid\left(\boldsymbol{X}\cup\boldsymbol{Y}\right)\subseteq\boldsymbol{T},\boldsymbol{T}\subseteq\boldsymbol{D}\right\}\right\|}{\left\|\left\{\boldsymbol{T}\mid\boldsymbol{X}\subseteq\boldsymbol{T},\boldsymbol{T}\subseteq\boldsymbol{D}\right\}\right\|}\times 100\% \tag{2}$$

Definition 4: Association rule discovery. The minimum support ($\boldsymbol{support_{\min}}$) and the minimum confidence ($\boldsymbol{confidence_{\min}}$) need to be set artificially according to the mining needs. The former describes the minimum importance of an association rule, while the latter specifies the minimum reliability that an association rule must satisfy. The set of support greater than the minimum support is called the frequent itemset, and the association rule is generated by calculating the confidence level based on the frequent itemset. If $\boldsymbol{Support}(X \rightarrow Y) \geq \boldsymbol{support}_{\min}$ and $\boldsymbol{Confidence}(X \rightarrow Y) \geq \boldsymbol{confidence}_{\min}$, the association rule is called as a strong rule, otherwise the association rule is called a weak rule.

Association rule algorithms have garnered considerable attention across myriad disciplines, primarily due to their unparalleled capacity to process voluminous data and their adeptness in unveiling patterns eluding conventional artificial intelligence and statistical techniques. Within the purview of this research, both the support and confidence degrees were calibrated at 50%. To rigorously ascertain the model's robustness, these thresholds were systematically adjusted to 60%, 70%, 80%, and 90%, respectively, and their validity assessed.

The Apriori algorithm is applied. The core of the Apriori algorithm is to utilize the a priori nature of the frequent itemset property, i.e., all subsets of a frequent itemset must also be frequent. The (*n*-1) itemsets are used to explore the *n*-item set, thus reducing the search space for association rules. The Apriori algorithm needs to be given a priori a support threshold as a criterion for determining the frequent itemsets. The transactional dataset is first scanned to produce the first candidate data itemset and that candidate dataset is a 1-item set collection. According to the pre-given support, the set that meets the threshold requirement is filtered as a 1-item frequent itemset, denoted as *S*_1. Then the 2-item set collection containing *S*_1 is searched again from the original dataset as the next candidate dataset, denoted as *S*_2. The same method is used until a frequent *n*-item set *S*_*n* is generated and the (*n*+1)-item set has failed to meet the constraints of the support and the confidence level. At this point, the work of extracting frequent itemsets from the original data has been completed. For these frequent itemsets, the extraction of association rules is completed by filtering the itemset rules with compound threshold conditions using the given confidence level.

Step 1. Scan the original transactional data and the candidate 1-item set can be extracted. According to the given minimum support *s*, scan the itemset frequency that does not less than the support *s* to form the frequent 1-item set.

Step 2. Scan the original transactional data and extract the set of candidate 2-item sets, and judge each 2-item set in the set, exclude the 2-item sets that do not contain frequent 1-item sets, forming a new set as a candidate set. The frequent 2-item sets can be filtered according to the support *s*.

Step 3. Repeat step 2 to get frequent 3-item sets, frequent 4-item sets and frequent *n*-item sets until none of the (*n*+1)-item sets satisfy the support threshold condition.

Step 4. Given the confidence level, determine whether each frequent item set rule is a valid and strong rule.

This a priori algorithm is applicable to Boolean data; therefore, it is necessary to convert numerical data to Boolean data. In this study, based on the *K*-mean clustering analysis method, each index is divided into two levels: high and low.

3.2 The second stage: Support vector machine (SVM)

SVM is particularly known for its robustness and effectiveness in binary classification problems where classes are well separated. Compared with models such as BPNN, CNN, or LSTM that might overfit complex patterns in the training data, SVM often shows better generalization abilities on unseen data. Fundamentally, SVM seeks to identify an optimal hyperplane that maximizes the separation between two classes of samples. This optimization ensures that the nearest samples from both categories to the hyperplane maintain the greatest possible distance from it, leading to an expansive classification margin. SVM accomplishes this by transmuting the input space into a higher-dimensional space through a nonlinear transformation governed by the inner product function. It is within this augmented space that SVM pursues the optimal classification boundary.

Let the classification hyperplane be:

$$\boldsymbol{w}*\varphi(\boldsymbol{x})+\boldsymbol{b}=0 \tag{3}$$

The decision function is:

$$\tilde{\boldsymbol{f}}(\boldsymbol{x})=\boldsymbol{sign}(\boldsymbol{w}*\varphi(\boldsymbol{x})+\boldsymbol{b}) \tag{4}$$

SVM can be expressed as a constrained optimization problem:

$$\mathbf{min}:\frac{1}{2}\|w\|^2+\boldsymbol{C}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}\xi_i \tag{5}$$

$$\boldsymbol{s.t.}\begin{cases} y_i\left(\boldsymbol{w}*\varphi(x_i)+\boldsymbol{b}\right)\geq 1-\xi_i \\ \xi_i\geq 0,\boldsymbol{i}=1,2,\cdots,\boldsymbol{n} \end{cases} \tag{6}$$

where *C* is the penalty parameter, which indicates the degree of tolerance to error; ξ_i is the introduced slack variable, which can be viewed as the bias generated by the sample by separating the hyperplane classifications.

The dyadic problem of the original problem can be derived by the Lagrangian function method:

$$\begin{gathered} \mathbf{max}_{\alpha}:\left[-\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{K}(x_i,x_j)+\sum_{i=1}^{n}\alpha_i\right] \\ \boldsymbol{s.t.}\begin{cases}\sum_{i=1}^{n}\alpha_i y_i=0,\boldsymbol{i}=1,2,\cdots,\boldsymbol{n} \\ 0\leq\alpha_i\leq\boldsymbol{C}\end{cases} \end{gathered} \tag{7}$$

where, $\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(x_i)\cdot\varphi(x_j)$ is the kernel function,; α_i and α_j is the Lagrange factor. This gives the classification function as:

$$\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{sign}(\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}}y_i\alpha_i\boldsymbol{K}(x_i,x)+\boldsymbol{b}) \tag{8}$$

The kernel function is one of the most important components of the support vector machine. Different kernel functions and their parameters can cause large differences in the performance of SVM. Commonly used kernel functions include linear kernel function, polynomial kernel function, Sigmoid kernel function and RBF kernel function. Since the radial basis function has certain advantages in analyzing high-dimensional data and has better anti-interference ability to the noise present in the data, the RBF kernel function is chosen to build the SVM classifier in this study.

$$\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{e}^{-\gamma(x-x_i)^2} \tag{9}$$

SVM has the advantages of simple structure, good generalization performance, fast learning speed and unique minima in the optimization solution. However, the classification effect of SVM is affected by its penalty parameter *C* and kernel parameter *γ* to a large extent, which has strong parameter sensitivity. Therefore, the next step is to optimize the parameters of the SVM.

3.3 The third stage: Modified particle swarm optimization algorithm (MPSO)

While SVM undoubtedly boasts features that cater to non-linear interactions, eliminate local minima issues, possess impressive generalization capabilities, and offer independence from comprehensive data reliance, its inherent sensitivity to kernel function initialization and associated parameters poses challenges. Such sensitivities drastically influence the ultimate predictive accuracy. The penalty parameter *C* which interplays between SVM's learning and generalizing capabilities, further complicates the scenario. An excessively large *C* value might impeccably fit training data, yet compromise the model's generalization prowess. Addressing this parameter selection quandary has been a pivotal challenge. Scholars have progressively integrated intelligent optimization algorithms—ranging from the genetic algorithm (GA), artificial fish swarm algorithm (AFSA), ant colony algorithm (ACA), artificial hummingbird optimization algorithm (AHOA), to the PSO—all aimed at refining SVM's penalty factor and kernel function parameters (Jiang et al., 2024). Compared with complex algorithms, such as GA and ACA, PSO requires less parameter tuning and function evaluation, and is therefore easier and faster to implement. Moreover, its dynamic tuning of particle velocities gives it a good trade-off between local and global search capabilities, while some other algorithms, such as FA and AFSA may focus on local neighborhood solutions. Therefore, PSO is chosen to optimize the parameters of the SVM.

The PSO algorithm, a computational method inspired by the collective behavior observed in bird foraging. Within the problem's solution space, multiple food sources can be envisioned. The preeminent food source parallels the global optimum of the solution space, while the ancillary sources correlate with local optima. The avian collective's mission revolves around the identification of this global optimum. During this search, individual birds exchange positional data, leading the entire flock closer to the most abundant food source, effectively locating the global optimal solution. Within the PSO framework, each potential solution in the space is termed a 'particle', endowed with two fundamental attributes: the position vector and the velocity vector. The former facilitates the calculation of a particle's fitness value, reflecting its proximity to the global optimum, while the latter prescribes the particle's subsequent movement direction and magnitude. Through the iterative optimization process, these particles recalibrate their trajectories based on individual and collective experiences, steering the entire swarm towards the global optimum. The steps of PSO algorithm are as follows:

Step 1: Population initialization. The population size, the position and velocity of the initial particles are all initialized. The fitness value of each particle to determine the individual optimal particle and the global optimal particle can be determined. Assume that in a *D*-dimensional search space, there are *N* particles to form a colony, where the *i*th particle is represented as a *D*-dimensional vector as shown in ***Eq***. (10) and the velocity of the ith particle is shown as ***Eq***. (11):

$$\boldsymbol{X_i} = \left\{ \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \boldsymbol{x}_{i3} \cdots \boldsymbol{x}_{iDn} \right\}, \boldsymbol{i} = 1, 2, \cdots, \boldsymbol{N} \tag{10}$$

$$\boldsymbol{V_i} = \left\{ \boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \boldsymbol{v}_{i3} \cdots \boldsymbol{v}_{iDn} \right\}, \boldsymbol{i} = 1, 2, \cdots, \boldsymbol{N} \tag{11}$$

The optimal solution for each particle and the optimal solution for the entire population are recorded $\boldsymbol{p_{best}}$ and $\boldsymbol{g_{best}}$ simultaneously.

Step 2: Update each particle's own velocity and position and calculate the fitness value. If the fitness value of each particle is better than the historical fitness value, update the position and fitness value of the particle. If the optimal fitness value in the whole population is better than the current global optimal fitness value, update the position and fitness value of the global optimal particle. The velocity and position of the *i*th particle are updated as shown in ***Eq.*** (12). and ***Eq.*** (13).

$$\boldsymbol{v}_{(i+1)d} = \boldsymbol{w}\boldsymbol{v}_{id} + \boldsymbol{c}_1\boldsymbol{r}_1(\boldsymbol{p}_{id} - \boldsymbol{x}_{id}) + \boldsymbol{c}_2\boldsymbol{r}_2(\boldsymbol{p}_{gd} - \boldsymbol{x}_{id}) \tag{12}$$

$$\boldsymbol{x}_{(i+1)d} = \boldsymbol{x}_{id} + \boldsymbol{v}_{(i+1)d} \tag{13}$$

where $\boldsymbol{p}_{id}$ is the individual known optimal solution; $\boldsymbol{p}_{gd}$ is the population known optimal solution; $\boldsymbol{w}$ is the inertia weight; $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are the learning factor; $\boldsymbol{r}_1$ and are a random number in the range [0, 1].

Step 3: Judge whether the convergence condition of the algorithm is satisfied. If so, stop the iteration and output the result; otherwise, skip to ***step 2*** to continue the execution.

In the PSO algorithm, the inertia weight factor represents the ability of the particle to inherit the speed of the previous iteration, and the setting of the inertia factor parameter can seriously affect the global and local search ability of the PSO algorithm. Since the inertia factor is fixed in the PSO algorithm, it can lead to slow convergence and easy to fall into local optimization. At the early stage of search, a larger inertia weight factor is needed for global optimization; while at the later stage of search, a smaller inertia weight is needed to strengthen the local search ability. To better balance the global and local search ability of the algorithm, this study sets an adaptive inertia weight factor *w* as ***Eq***. (14), and this parameter can be changed with the increase of iteration number. The modified algorithm is called MPSO.

$$\boldsymbol{w}(\boldsymbol{t}) = (\boldsymbol{w}_{\max} - \boldsymbol{w}_{\min}) / (1 + \mathbf{exp}(\frac{15\boldsymbol{t}}{\boldsymbol{T}} - 8)) + \boldsymbol{w}_{\min} \tag{14}$$

where $\boldsymbol{w}_{\mathbf{min}}$ denotes the initial inertia weight value; $\boldsymbol{w}_{\mathbf{max}}$ denotes the inertia weight value at the maximum number of iterations; $\boldsymbol{T}$ denotes the maximum number of iterations, and $\boldsymbol{t}$ denotes the current number of iterations.

3.4 ARA-SVM-MSPO hybrid model

The three stage is combined by the following logic. ARA is used to select features that are most relevant to the credit default risk of firms, to reduce the dimension of variables, and the selected variables are the inputs of the prediction model of SVM. By inputting the variables in SVM, the prediction results can be obtained. However, the prediction accuracy of single SVM may be low since some parameters should be determined subjectively. The MPSO is employed to search the optimum of kernel parameter and penalty parameters of SVM, and this process will not stop until the results with the smallest errors are obtained or the maximum iteration number is reached. The pseudo code of the proposed three-stage prediction model of ARA-SVM-MPSO is shown in the **Table A3.** And the framework of the three-stage prediction model is displayed in **Fig. 1**.

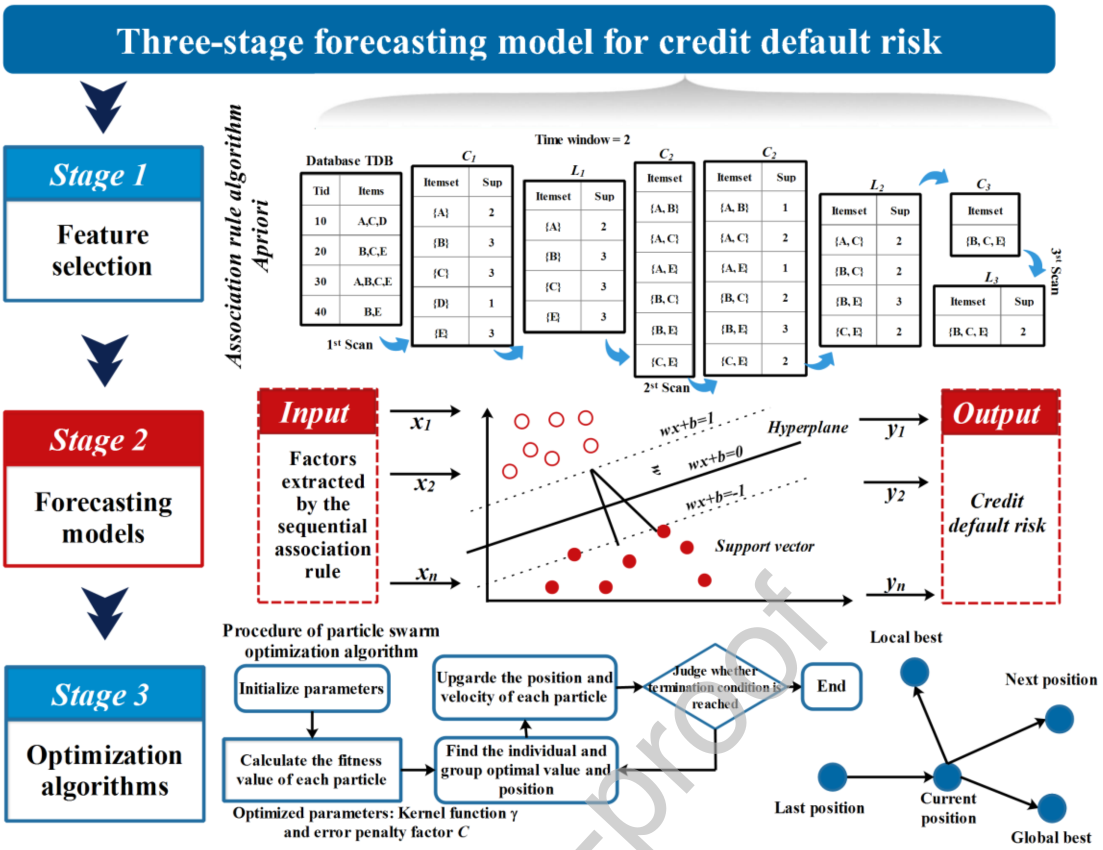

Fig. 1. Framework of the three-stage prediction model

3.4 Model evaluation

This study primarily employs the Area Under the Curve (AUC) and the Geometric Mean (G-Mean) as principal evaluative metrics. Recall and Specificity are also used as secondary metrics. The AUC, representing the area beneath the Receiver Operating Characteristic (ROC) curve, offers a quantifiable measure of the model's discriminative prowess. The G-Mean, constituting the geometric mean of Recall and Specificity, assesses the model's holistic predictive accuracy, . Recall quantifies the ratio of accurately predicted credit default risks to the overall instances of such risks. Conversely, Specificity delineates the model's aptitude in identifying normative samples. Comprehensive descriptions of these metrics are tabulated in **Table 4**.

Table 4. Evaluation metrics of model performance

Metrics	Meanings	Equation	Value range
AUC	Measure model classification effects	Area under the ROC curve	Value range is 0 to 1, and larger value means better prediction effectiveness.
G-Mean	Collectively evaluate the prediction effectiveness of the model	$\sqrt{\boldsymbol{Recall} \times \boldsymbol{Specificity}}$	Value range is 0 to 1, and larger value means prediction effectiveness.
Recall	Measure the prediction accuracy of samples with credit default risk	$\frac{\boldsymbol{TP}}{(\boldsymbol{TP}+\boldsymbol{FN})}$	Value range is 0 to 1, and larger value means lower underreporting rate.
Specificity	Measure the prediction accuracy of normal samples	$\frac{\boldsymbol{TN}}{(\boldsymbol{FP}+\boldsymbol{TN})}$	Value range is 0 to 1, and larger value means lower false positive rate.

For illustrative purposes, consider the dataset bifurcated into the credit default risk category (termed "Positive") and the regular category (termed "Negative"). Potential outcomes encompass True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN), as shown in **Table 5**. The True Positive Rate (TPR) elucidates the fraction of accurate classifications within the credit default risk samples relative to the entirety of such classifications. A heightened TPR insinuates an augmented classification precision, as explicated in ***Eq***. (15). On the contrary, the False Positive Rate (FPR) signifies the ratio of misclassifications in the credit default risk to all classifications within the normative category. An escalating FPR suggests diminishing accuracy in credit default risk classifications, indicating the model's suboptimal efficacy in such classifications. FPR is formally encapsulated in ***Eq***. (16). The ROC curve is plotted with FPR and TPR on the *x* and *y* axes, respectively. The AUC, crucially, represents the area subsumed by this ROC curve.

$$\boldsymbol{TPR} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP}+\boldsymbol{FN})} \tag{15}$$

$$\boldsymbol{FPR} = \frac{\boldsymbol{FP}}{(\boldsymbol{FP}+\boldsymbol{TN})} \tag{16}$$

Table 5. Results of classification

		Predicted results	
		Positive	Negative
Actual result	Positive	*TP*	*FN*
	Negative	*FP*	*TN*

4. Empirical results

4.1 Variables, samples and data source

4.1.1 Variables

Drawing upon the extant literature, particularly Jiang et al. (2021), this investigation delineates both financial and non-financial indicators for the appraisal of firm credit default risk. In consonance with the frameworks posited by Sun and Li (2011) and Wu et al. (2022), we systematically categorize financial indicators into six discernible dimensions: *Debt-paying ability, Operating ability, Profitability, Cash flow, Risk management ability,* and *Growth potential*. A detailed tabulation of these metrics can be found in **Table A1**. The emergence of firm credit default risk emanates from an intricate interplay of multifarious factors. Beyond purely financial considerations, the significance of non-financial indicators in shaping such risk cannot be understated. Drawing from the foundational works of Liu et al. (2023), this study discerningly classifies non-financial indicators into four distinct dimensions: *Management attributes*, *Governance attributes*, *Macroeconomic factors*, and *Text sentiments*. A comprehensive representation of these indicators is delineated in **Table A2**.

Serving as a novel dimension in the realm of non-financial indicators, *Text sentiment*—both of a positive and negative ilk—is meticulously assessed. The specific steps for obtaining the textual sentiment are as follows. First, before extracting the text sentiment, it is necessary to preprocess the crawled text data. According to Zhou et al. (2021b), we construct the deactivation thesaurus[1] and delete the deactivated words in the text. Secondly, grounded in the lexical repositories curated by Loughran and McDonald (2011) and Henry (2008), affective tonal words are assiduously culled from annual reports across a spectrum of firms, which can better reflect tone in financial text. Specifically, the word list contains negative words, such as “loss”, “impairment”, “claims”, “against”, “adverse” and “restated”, and positive words, such as “achievein”, “attain”, “efficient”, “improve”, “profitable”, “upturn”. Finally, using python crawling technique, we analyze the text of corporate annual reports to obtain the number of positive words, negative words, and total words, and calculate the text sentiment accordingly (Hsu and Wang, 2013). The metrics for these text sentiments are elucidated in ***Eq***. (17) and ***Eq***. (18). To check the validity of our text sentiment measures, we identify an independent random sample of 200 annual reports to manually determine whether the selected report has a negative or a positive tone. We then compare textual sentiment to manual ratings[2] and find that the measurement of text sentiment is reasonable.

$$\boldsymbol{Positive = \frac{Positivewords}{Totalwords}} \tag{17}$$

$$\boldsymbol{Negative = \frac{Negativewords}{Totalwords}} \tag{18}$$

Where *Positive* represents the positive text sentiment, *Positivewords* represents the number of positive words; *Negative* denotes the negative text sentiment; *Negativewords* denotes the number of negative words; *Totalwords* denotes the total number of the text.

[1] According to Zhou et al. (2021b), we sort out the stop word lists such as “Harbin Institute of Technology Stopword Thesaurus”, “Sichuan University Machine Learning Intelligent Lab Stopword Thesaurus”, and “Baidu Stopword Thesaurus”, and then build a stop word list.

[2] If the measure is reasonable, the proportion of positive words should be higher than negative words for annual reports that are manually perceived as positive emotions, and the gap between them will increase with positive emotions.

4.1.2 Samples and data source

To ensure both the relevance and availability of data, this research meticulously analyzes data spanning two decades (2000-2020) from 2,253 publicly listed firms. The reasons for choosing the sample period are as follows: due to the small amount of data on listed firms and the poor standardization of corporate annual reports before 2000, the sample period of this paper starts from 2000 limited by the data availability. Secondly, since the China Securities Regulatory Commission (CSRC) revised the guidelines on the format of annual reports of listed companies in 2021, which may influence the text data, the sample period ends in 2020. The reason for choosing the listed firms is that they are important subjects in bank lending market compared to unlisted firms and their data are well available.

These firms are bifurcated into two distinct categories: defaulted and non-defaulted entities. First, for defaulted firms, these are entities that have received the 'ST' (special treatment) designation by the CSRC. This study incorporates all firms assigned the 'ST' designation within the specified timeframe. For firms receiving this designation multiple times, only data from their initial 'ST' designation year is incorporated. Consequently, 372 firms, with their earliest 'ST' designation, form the "defaulted" subset. It's pertinent to note that the prediction for these entities harnesses data from two years prior (t-2) to their respective 'ST' designations. For non-defaulted firms: these firms have remained untouched by the 'ST' designation by the CSRC within the examined window. Our subset comprises 1,881 such entities, all of which have evaded the 'ST' designation. Given the stipulation that an 'ST' designation signals two successive years of financial underperformance, the prediction model is constructed to utilize data from two years prior to the 'ST' designation year. To elucidate, a firm assigned 'ST' in 2020, indicating losses for both 2019 and 2020, will have its 2018 (or earlier) data employed for prediction purposes.

Upon aggregating both subsets (defaulted and non-defaulted), the dataset undergoes meticulous refinement: indicators with less than 95% completeness are omitted, and any extant gaps are populated through interpolation. Subsequently, this consolidated dataset is partitioned into training and testing subsets, adhering to a 7:3 ratio. As for data sources, financial metrics are retrieved from the Wind database, while non-financial metrics are culled from CSMAR. Macroeconomic data is sourced from both the National Bureau of Statistics (NBS) and the China Economic and Social Development Statistics Database .

4.2 Prediction results of the proposed model

To streamline the computational demands of the model, the association rule algorithm is first employed to discern the most pertinent financial and non-financial indicators. The temporal window for each variable is fixed at two units. **Table 6** shows the outcomes of the association rule algorithm. Of the 48 variables assessed, 21 rules are derived for singular indicators[3]. The association rules underscore both support and confidence degrees surpassing 0.5, signifying their salient association with firm default risk.

First, holding paramount significance in default risk determination, profitability indicators exhibit both support and confidence degrees exceeding 0.8. Firms demonstrating robust profitability typically exhibit diminished credit

[3] It is worth noting that over 100 rules would have been delineated when examining the combined effects of multifarious variables. However, for the sake of brevity and given that the default risk prediction operationally harnesses only singular indices, the conglomerate variable rules are omitted in **Table 6**.

default risks, given their imperative to cultivate a favorable public image and their intrinsic market competitiveness. Conversely, firms grappling with diminished profitability encounter heightened predispositions towards credit default risks. Second, as illuminated by *Rules 20* and *21*, both positive and negative text sentiments impart profound impacts on default risk, with their metrics ranking second, exceeding a threshold of 0.76. Such results underscore the paramountcy of text sentiment. Challenging situations faced by firms often resonate within disclosed textual content. This can manifest either as overt negative word usage or subtly through ambiguous language, which can be discernible by astute investors. Third, both of debt-paying ability and cash flow have direct implications on credit risk, given their intuitive relationship with a firm's financial health. Other variables, including management risk, growth potential, and macroeconomic factors, also exhibit significant ties with default risk, as evinced by support and confidence degrees surpassing 0.6. While default risk fundamentally denotes the propensity of a firm to experience payment impediments or outright defaults, the financial leverage metric further amplifies this relationship. This ratio epitomizes the extent of a firm's reliance on debt capital, with escalating values signaling heightened firm risk. Indicators echoing profit growth rates correlate intensely with default risk, underscoring profitability as a cardinal gauge for firm developmental potential. Among macroeconomic factors, those epitomizing regional economic development are especially salient in affecting firm credit risk. Operating ability, managerial attributes, and corporate governance dynamics assume relatively peripheral roles. Although these metrics record lesser support and confidence degrees, values remain above the 0.5 threshold, cementing their inclusion in the prediction model.

In summation, financial variables maintain a profound association with default risk and, as such, their integration enhances prediction precision. Non-financial variables, while generally secondary in importance, see exceptions like text sentiments, which ranks second only to profitability. Firms' credit risks emanate from an amalgamation of financial and non-financial drivers, with some entities establishing direct causal links and others merely correlating. Model efficiency negates the need for comprehensive indicator inclusion, with the association rule algorithm emerging as an efficacious method to discern the most pivotal indicators.

Table 6. Results of association rule

No.	Rules	Support degree	Confidence degree	No.	Rules	Support degree	Confidence degree
1	$X_{1}\rightarrow RISK$	0.7033	0.7124	12	$X_{25}\rightarrow RISK$	0.6431	0.6226
2	$X_{8}\rightarrow RISK$	0.7628	0.7731	13	$X_{26}\rightarrow RISK$	0.6970	0.6831
3	$X_{10}\rightarrow RISK$	0.5522	0.5318	14	$X_{27}\rightarrow RISK$	0.6334	0.6495
4	$X_{14}\rightarrow RISK$	0.8122	0.8169	15	$X_{32}\rightarrow RISK$	0.5039	0.5146
5	$X_{15}\rightarrow RISK$	0.8240	0.8337	16	$X_{35}\rightarrow RISK$	0.5637	0.5425
6	$X_{16}\rightarrow RISK$	0.8382	0.8469	17	$X_{42}\rightarrow RISK$	0.6338	0.6581
7	$X_{17}\rightarrow RISK$	0.8422	0.8379	18	$X_{45}\rightarrow RISK$	0.6429	0.6655
8	$X_{19}\rightarrow RISK$	0.7719	0.7453	19	$X_{46}\rightarrow RISK$	0.6213	0.6430
9	$X_{20}\rightarrow RISK$	0.7142	0.7023	20	$X_{48}\rightarrow RISK$	0.7625	0.7748
10	$X_{21}\rightarrow RISK$	0.6215	0.6543	21	$X_{47}\rightarrow RISK$	0.7869	0.7934
11	$X_{24}\rightarrow RISK$	0.6749	0.6860				

Upon the extraction of salient features via the association rule algorithm, a SVM—refined by the modified PSO—is employed for default risk prediction. Employing temporal windows ranging from 2 to 4, the results data are tabulated in **Table 7**. The ensuing observations can be distilled: First, among the time windows considered, a span of 2 emerges as the most efficacious, as evinced by its superlative AUC, G-mean, Recall, and Specificity across all model iterations. This superiority can be attributed to the rapid evolution in technological and economic landscapes, which in turn subject firms to an array of dynamic influences. Contemporary firms, besieged by these rapid alterations, often necessitate expeditious decision-making. Consequently, both financial and non-financial indices experience swift transitions, rendering shorter-term prognostications more reliable and accurate than their long-term counterparts. Second, the proposed model outperforms its contemporaries in prediction prowess, adeptly discerning between true and false samples. When the time window of 2 is considered, models eschewing text sentiment yield an AUC, G-Mean, Recall, and Specificity of 0.926, 0.915, 0.913, and 0.918, respectively. These figures, albeit commendable, are overshadowed by models incorporating text sentiment indices, thereby underscoring the latter's significance. While PSOSVM and SVM models might demonstrate marginally diminished effectiveness relative to the previously mentioned models, their AUC metrics still comfortably exceed 0.9. This robust performance accentuates the intrinsic efficacy of the SVM.

Table 7. Prediction results of the proposed model

Model	Time window	AUC	G-Mean	Recall	Specificity
ARA-SVM-MPSO hybrid model	2	0.958	0.937	0.931	0.944
	3	0.921	0.910	0.915	0.906
	4	0.903	0.889	0.892	0.886
	5	0.864	0.853	0.857	0.849
APAPSOSVM without text sentiment	2	0.926	0.915	0.913	0.918
	3	0.907	0.886	0.893	0.880
	4	0.875	0.861	0.864	0.858
	5	0.842	0.828	0.836	0.821
PSO+SVM	2	0.912	0.899	0.905	0.894
	3	0.870	0.860	0.861	0.859
	4	0.852	0.839	0.842	0.837
	5	0.835	0.822	0.826	0.818
SVM	2	0.903	0.888	0.892	0.885
	3	0.884	0.857	0.860	0.854
	4	0.850	0.831	0.837	0.825
	5	0.826	0.807	0.813	0.802

4.3 Comparisons

In an endeavor to authenticate the efficacy of the proposed model for predicting a firm's default risk, a tripartite comparative analysis was conducted, as delineated in **Fig. 2**.

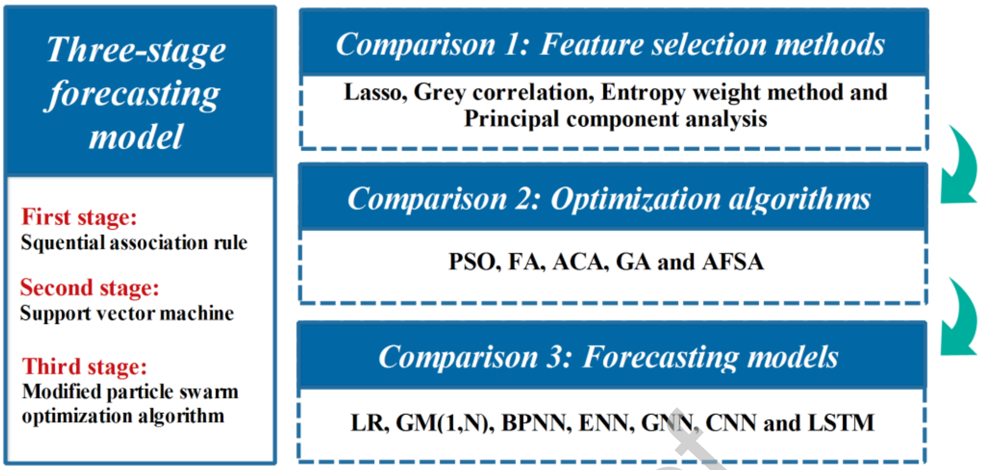
Three-stage forecasting model
First stage:
Squential association rule
Second stage:
Support vector machine
Third stage:
Modified particle swarm optimization algorithm
Comparison 1: Feature selection methods
Lasso, Grey correlation, Entropy weight method and Principal component analysis
Comparison 2: Optimization algorithms
PSO, FA, ACA, GA and AFSA
Comparison 3: Forecasting models
LR, GM(1,N), BPNN, ENN, GNN, CNN and LSTM

Fig. 2. Comparison settings of this study

4.3.1 Feature selection methods

In this section, the association rule algorithm was compared with various alternative feature selection methodologies, including principal component analysis, Lasso, grey correlation, and the entropy weight method. While each approach distills the key factors affecting default risk, their underpinning extraction paradigms differ markedly (Palma et al., 2023; Bucci et al., 2023; Bai et al., 2022). Notably, both the principal component analysis and Lasso method are adept at isolating variables intimately tied to credit risk, while the grey correlation and entropy weight method spawn correlation coefficients, selecting variables exhibiting coefficients or weights exceeding 0.5.

Table 8 delineates the outcomes, highlighting the unparalleled performance of the proposed model in metrics such as AUC, G-mean, Recall, and Specificity. The following is the entropy weight method with comprehensive index inclusion, which is lack of the ability to address substantial correlations between indicators and a susceptibility to skewed weight allocations. The grey correlation and Lasso, notwithstanding their extraction of 17 and 14 variables respectively, manifest AUC and G-mean values languishing below 0.8. The grey correlation's partiality to average values compromises its objectivity, whereas Lasso grapples with the quandary of optimal regularization coefficient selection-erring on excess may induce model underfitting, while parsimony may nullify model constraints. Alarmingly, the principal component analysis-based model exhibits subpar outcomes, potentially an offshoot of inherent limitations imposed by sample size and variable quantity.

In summation, of the feature selection methodologies under scrutiny, the association rule algorithm emerges as the most efficacious, courtesy of its agnosticism towards variable quantity, rendering it adept at discerning latent rules within expansive databases-a prowess particularly invaluable for credit risk feature selection. A salient observation meriting attention is the unanimous inclusion of text sentiment indices across all methodologies, reinforcing the pivotal role of text sentiment in the realm of default risk prediction.

Table 8. Comparison with other feature selection methods

Model	Text sentiment index	Number of index	AUC	G-Mean	Recall	Specificity
ARA-SVM-MPSO hybrid model	Yes	21	0.958	0.937	0.931	0.944
Entropy weight method	Yes	31	0.927	0.918	0.914	0.923
Grey correlation	Yes	17	0.793	0.758	0.767	0750
Lasso	Yes	14	0.762	0.751	0.755	0.748
Principal component analysis	Yes	26	0.751	0.742	0.746	0.739

4.3.2 Optimization algorithm

The proposed model integrates a modified PSO algorithm to refine the selection of the parameter *g* and error penalty factor *C* within the support vector machine, as delineated by Chen et al. (2019). To ascertain its proficiency, an array of established optimization algorithms are enlisted for comparative analysis. These include the conventional PSO, firefly algorithm (FA), genetic algorithm (GA), ant colony algorithm (ACA), and the artificial fish swarm algorithm (AFSA). The above algorithms are widely used by scholars to solve the parameter sensitivity problem of machine learning model (Wang et al., 2023), and cover different aspects of feature selection, including data degradation and regularization, handling of uncertain information and information quantization, which have a significant impact on the performance of the model. As expounded upon by Jiang (2021), each of these algorithms is configured with a maximum iteration of 300 and a population size of 100. For the PSO, the respective learning and inertia weight factors are set at 2 and 1.2, with velocity parameters spanning a range from -1 to 1. The FA's configuration stipulates a step-length factor of 0.5 and a light intensity absorption coefficient of 0.1. The GA deploys a crossover probability of 0.5 and a mutation probability of 0.001. Within the ACA, heuristic, expected heuristic, and pheromone volatile factors are defined as 1.5, 5, and 0.5 respectively. Lastly, the AFSA utilizes a uniform value of 1 for both the step-length and crowding degree factors.

Table 9 shows the outcomes derived from these juxtaposed optimization algorithms. Models incorporating the text sentiment variable invariably outperform those bereft of it, further underscoring the prowess of the modified PSO algorithm. While the SVM manifests commendable efficacy in adjudicating default risk, its performance is intrinsically tethered to its kernel and penalty parameters. The modified PSO algorithm distinguishes itself with superior capabilities in both global optimization and local refinement. Among the examined algorithms, the AUC values for AFSA, ACA, AFA, and PSO are tabulated as 0.961, 0.956, 0.946, and 0.944, respectively, outshining the performance of the GA. This differential can be attributed to the fact that, barring the GA, which operates on gradient principles, the remaining algorithms are anchored in swarm intelligence optimization. These probabilistic search-based methods, characterized by parallel distribution and devoid of centralized control constraints, remain impervious to individual perturbations. Given their inherent flexibility concerning problem continuity, these algorithms typically deliver augmented optimization outcomes.

The AFSA exhibits superior optimization performance, registering an AUC of 0.961 and a G-Mean of 0.937. Notably, the temporal improvement associated with this algorithm stands at a pronounced -93.06%. In the realm of prediction, timeliness is an indispensable consideration. Given that our study's sample size is relatively modest, the issues pertaining to timeliness appear to be mitigated. Nonetheless, as sample sizes expand, the optimization velocity of AFSA is anticipated to decelerate, potentially undermining prediction efficacy. The ACA demonstrates comparable metrics, with an AUC of 0.956 and a G-Mean of 0.937, akin to the modified PSO. However, it shares the time improvement shortcoming with AFSA, evidenced by a -92.59% metric. An escalation in the count of artificial fishes necessitates augmented storage, leading to a substantial computation surge. Central to ACA's operations is the pheromonal feedback mechanism; excessive positive feedback, however, may culminate in a languished convergence rate. Prediction derived from the FA and the conventional PSO trail those facilitated by the modified PSO. The FA is contingent upon the presence of an optimal individual within its sensing ambit to proffer pivotal data. Absent this, the individual ceases its explorative endeavors. This dependency on optimal individuals curtails convergence velocity. Furthermore, when an individual approaches the apogee, if the step size surpasses the interstitial distance, oscillations around the zenith may ensue. The traditional PSO grapples with challenges when confronted with discrete and combinatorial optimization conundrums. The GA's AUC and G-Mean metrics stand at 0.918 and 0.903, respectively. This algorithm is encumbered by a myriad of training parameters, necessitating subjective determination, inclusive of factors like crossover and mutation rates. In summation, of the optimization algorithm under scrutiny, the modified PSO algorithm has superior capabilities in both global optimization and local refinement, because of its dynamic tuning of particle velocities. Since it requires few parameters to adjust, it is simple to implement and greatly reduce the training time.

Table 9. Comparison with other optimization algorithms

Text sentiment index	Algorithm	Improvement of training time	AUC	G-Mean	Recall	Specificity
With text sentiment index	ARA-SVM-MPSO hybrid model	-	0.958	0.937	0.931	0.944
	PSO	-28.57%	0.944	0.926	0.925	0.927
	FA	-80.77%	0.946	0.924	0.927	0.921
	ACA	-92.59%	0.956	0.937	0.931	0.943
	GA	-60.00%	0.918	0.903	0.904	0.903
	AFSA	-93.06%	0.961	0.937	0.930	0.945
Without text sentiment index	Proposed model	-	0.926	0.915	0.913	0.918
	PSO	-30.77%	0.917	0.905	0.905	0.906
	FA	-82.00%	0.912	0.887	0.899	0.876
	ACA	-93.08%	0.920	0.906	0.908	0.904
	GA	-62.50%	0.908	0.894	0.901	0.888
	AFSA	-93.53%	0.927	0.912	0.910	0.915

4.3.3 Forecasting models

In our concluding comparative analysis, we compare our proposed model with a series of established models. This array includes logistic regression (LR), grey model (GM), back-propagation neural network (BPNN), Elman neural network (ENN), generalized neural network (GNN), convolutional neural network (CNN), and long short-term memory neural network (LSTM). Such models are widely used in predicting default risk and represent both parametric and non-parametric methods, covering a broad spectrum of statistical techniques (Zhang et al., 2022; Mahbobi et al., 2023). Both LR and GM are categorized under statistical models. The GM, a representative of statistical models, has its developmental grey coefficient set at 0.5. In contrast, the remaining models are typified as neural networks, all with a prescribed maximum iteration of 300. The BPNN employs the Traingd training algorithm with stipulated parameters including a minimum training accuracy of 0.0001, a minimum training rate of 0.9, and a learning rate of 0.01. The ENN utilizes the Traingd training algorithm in tandem with the Learngdm learning algorithm, anchored by the mean squared error as its performance function. Parameters for the GNN encompass a minimum training rate of 0.9, a dynamic of 0.7, and a sigmoid parameter fixed at 1. Both CNN and LSTM, as deep learning paradigms, possess intricate network architectures. Their shared parameters are a dropout rate of 0.5, a learning rate of 0.001, and the ReLU activation function.

Data presented in **Table 10** underscores the superior prediction efficacy of our proposed model, especially when assessed on metrics such as AUC and G-Mean. Among the models under scrutiny, the statistical frameworks exhibited sub-optimal prediction outcomes. The proposed model's AUC improvement rate over logistic regression and GM(1,N) is documented at 11.06% and 10.86% respectively. The foundational tenets of statistical models, rooted in parameter estimation, could be predisposed to substantial biases. Pertaining to more rudimentary prediction methodologies, the proposed model's enhancement over BPNN, ENN, and GNN is registered at 4.28%, 3.86%, and 3.13% respectively. While neural networks commendably approximate intricate non-linear relationships, bolstered by their robustness, their need for extensive parameters to buttress their complex topological structures can impinge upon prediction outcomes. In contrast, SVM excels in processing high-dimensional data, marked by its commendable generalization prowess and robustness. SVM's efficacy is magnified by its ability to transmute a non-linear challenge into a linear one through judicious kernel function selection. Both CNN and LSTM proffer results that bear semblance to the proposed model, given that deep learning frameworks possess enhanced capacities to grapple with multifaceted non-linear correlations. Nevertheless, they are not without their limitations. The stringent prerequisites concerning data volume and quality, combined with the network's intricate parameters, substantially amplify computational overheads. Thus, in pragmatic applications, while deep learning frameworks may indeed offer high prediction precision, our proposed model, lauded for its relative simplicity, emerges as a preferable contender.

Table 10. Comparison with other prediction models

Text sentiment index	Model	AUC	Improvement rate	G-Mean	Improvement rate	Recall	Improvement rate	Specificity	Improvement rate	DM test
With text sentiment index	ARA-SVM-MPSO hybrid model	0.958	-	0.937	-	0.931	-	0.944	-	-
	LR	0.852	11.06%	0.839	10.46%	0.841	9.67%	0.837	11.33%	0.000^{***}
	GM(1,*N*)	0.854	10.86%	0.843	9.98%	0.846	9.13%	0.841	10.91%	0.000^{***}
	BPNN	0.917	4.28%	0.910	2.88%	0.908	2.47%	0.912	3.39%	0.000^{***}
	ENN	0.921	3.86%	0.910	2.83%	0.915	1.72%	0.906	4.03%	0.000^{***}
	GNN	0.928	3.13%	0.912	2.62%	0.916	1.61%	0.909	3.71%	0.000^{***}
	CNN	0.952	0.63%	0.937	0.00%	0.930	0.11%	0.944	0.00%	3.334
	LSTM	0.951	0.73%	0.940	-0.37%	0.938	-0.75%	0.943	0.11%	3.226
Without text sentiment index	ARA-SVM-MPSO hybrid model	0.926	-	0.915	-	0.913	-	0.918	-	-
	Logistic regression	0.831	10.26%	0.822	10.16%	0.825	9.64%	0.819	10.78%	0.000^{***}
	GM(1,*N*)	0.836	9.72%	0.823	10.00%	0.824	9.75%	0.823	10.35%	0.000^{***}
	BPNN	0.894	3.46%	0.880	3.77%	0.886	2.96%	0.875	4.68%	0.000^{***}
	ENN	0.899	2.92%	0.877	4.16%	0.883	3.29%	0.871	5.12%	0.000^{***}
	GNN	0.905	2.27%	0.895	2.13%	0.899	1.53%	0.892	2.83%	0.000^{***}
	CNN	0.923	0.32%	0.912	0.33%	0.916	-0.33%	0.908	1.09%	3.659
	LSTM	0.922	0.43%	0.906	0.93%	0.910	0.33%	0.903	1.63%	3.772

4.4 Robustness check

The preceding comparative assessments substantiate the efficacy of our proposed model by other renowned methodologies. This section delves deeper, seeking to ascertain the model's robustness through varied parameter settings. Given that the parameters of the SVM are fine-tuned via the modified PSO algorithm, our focus primarily shifts to the parameter configurations of both the association rule algorithm and the PSO. First, changing support and confidence helps to determine how sensitive the association rule algorithm is to the prevalence and reliability of the identified rules. **Fig. 3(I)** elucidates the outcomes stemming from variations in support and confidence degrees, ranging from 50% to 90%. Second, the population size of the modified PSO algorithm will directly affect the optimization process. A larger population might enhance the exploration capabilities, potentially leading to better global optimization but at increased computational cost, whereas a smaller population might speed up the computations but risk premature convergence to local optima. **Fig. 3(II)** portrays the ramifications of altering the population size of the modified PSO algorithm, with variations spanning from 100 to 500. Third, modifying the iteration count explores the balance between computational efficiency and the depth of search. More iterations allow for more thorough exploration of the solution space, potentially leading to a more optimal setting of SVM parameters but requiring more time and computational resources. In **Fig. 3(III)**, we adjust the iteration count, setting it in a range between 100 and 500. Finally, the learning factor in the modified PSO algorithm plays a critical role in how individual particles in the swarm adjust their position in the search space, ultimately influencing the optimization process and the conclusions derived from it. The learning factor of the PSO is modulated between 1 and 2, with the resultant observations depicted in **Fig. 3(IV)**. Upon scrutiny, one discerns that metrics such as AUC, G-Mean, Recall, and Specificity manifest commendable stability across the diverse parameter configurations. This consistent performance serves as a testament to the robust nature of the model we propose.

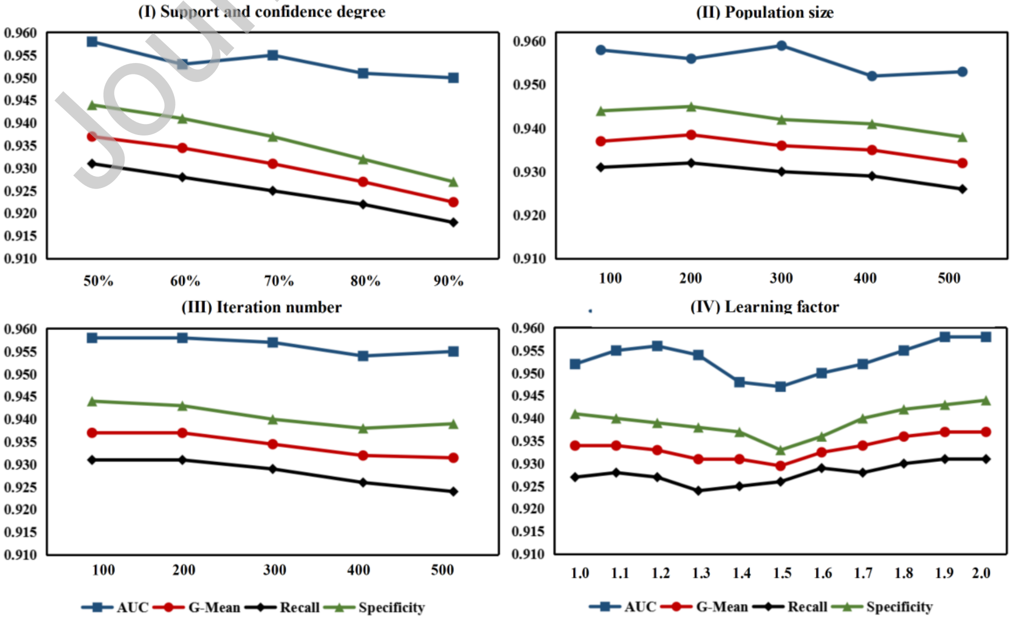

Fig. 3. Results of different robustness checks

5. Conclusions and policy implications

5.1 Conclusions

Drawing from the detailed discourse presented, several salient points emerge regarding the superior performance of our three-stage model in prediction accuracy when compared with other extant models. First, in the initial stage of feature selection, the variable text sentiment emerges as quintessential to accurate prediction of default risk. Comparative analyses, both in the absence and presence of this variable, underscore its significance. Predicts that incorporate both positive and negative text sentiments yield superior accuracy. Beyond the realm of text sentiment, financial indicators too exert significant influence on default risk. Notably, profitability and cash flow stand out as pivotal determinants. When appraising the hierarchy of influence, text sentiment secures the third position, superseded only by the aforementioned financial metrics. Second, the association rule algorithm, with a support degree and confidence degree both set at 50%, demonstrates efficacy in feature selection amidst diverse influencing factors. This algorithm not only streamlines the variable count but also astutely pinpoints those most germane to default risk. This outstrips the capabilities of alternative methodologies like the entropy weight method, grey correlation, Lasso, and principal component analysis. Third, in the stage of optimization spotlights the commendable capabilities of the modified PSO algorithm in optimizing parameters for the SVM. Evidently, algorithms grounded in swarm intelligence exhibit superiority over their gradient-based counterparts. While certain algorithms-like the ACA and AFSA-mirror the results of the modified PSO, they unfortunately entail a more onerous computational load. Finally, the SVM, when fine-tuned using the modified PSO, outperforms an array of models, ranging from logistic regression and grey models to BPNN, ENN, and GNN. Intriguingly, while the prediction acumen of our proposed model aligns with some deep learning paradigms, its inherently simpler architecture renders it more congruous for real-world applications.

5.2 Managerial implications

Our study provides important managerial implications. Firms' defaults can lead to serious consequences, jeopardizing the interests of their stakeholders and even disrupting financial markets. Therefore, it is crucial to predict and control default risk. Our study proposes a three-stage prediction model to predict default risk and emphasizes the role of text sentiment variables. Although macro, financial, and managerial information also plays an important role in predicting default risk of listed firms, the absence of textual information will reduce the accuracy of prediction and is not conducive to risk management. Therefore, banks and other financial institutions need to consider textual information such as textual sentiment of listed companies when granting loans. Our study provides a useful reference for banks to predict default risk, emphasizing the importance of textual information.

Second, at the firm level, firms ought to elevate their intrinsic risk management protocols. The bespoke nature of firm default risk-underscored by a pronounced moral risk and governance deficits-mandates the establishment of rigorous firm governance frameworks. Firms must cultivate and perpetuate risk cognizance, attune themselves to the prevailing developmental milieu, and remain astutely aware of their salient challenges. They should formulate long-term development strategies and improve their risk response mechanisms so that they can adopt disciplined processes in dealing with challenges. The active disclosure of their responses can also send positive signals to their stakeholders and change their expectations. Furthermore, there's an imperative for firms to perpetually enhance their information disclosure processes, ensuring data's timeliness, relevance, and veracity. High-quality textual data not only sends positive signals to stakeholders, but also helps to predict corporate risks and can effectively manage default risks. Such steps are indispensable in sidestepping pitfalls like adverse selection and moral risk.

5.3 Limitations

While the introduced three-stage prediction model provides substantial utility in practical applications, there are also some limitations. Our proposed model is restricted to a binary classification framework, only discerning the likelihood of a firm being exposed to credit default risk. In subsequent research endeavors, it would be advantageous to delve into the nuanced impacts of varied information milieus on the textual tone within firms' presentations, particularly in relation to their commercial performance metrics. Moreover, an examination of the interplay between the tone of presentation text and fluctuations in a listed firm's market capitalization or stock price volatility could furnish intriguing insights. It is also salient to note that textual datasets are replete with features transcending mere sentiment predilections. Metrics like textual readability, granularity of responses, among others, encapsulate significant value and merit further academic exploration.

Author Statement

Xuejiao Ma:Conceptualization, Visualization, Methodology, Writing – original draft.
Tianqi Che: Methodology, Data curation, Writing – original draft.
Qichuan Jiang: Conceptualization, Methodology, Supervision, Writing - review and editing.

Acknowledgement

This research was funded by National Social Science Foundation of China (23BTJ039; 23CTJ023).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Table A1. Financial indicators

First-level indicator	Second-level indicator	Signal	Explanation
Debt-paying ability	Current ratio	X_1	Current assets divided by current liabilities
	Quick ratio	X_2	(Current assets - Inventories)/Current liabilities
	Cash ratio	X_3	Closing cash balance divided by current liabilities
	Working capital to borrowings ratio	X_4	(Current assets - Current liabilities)/Borrowings
	Interest coverage multiple	X_5	EBIT divided by finance costs
	Interest coverage multiple on cash flows	X_6	Net cash flows from operating activities divided by finance costs
	Ratio of net cash flows from operating activities to interest-bearing debt	X_7	Average balance of net cash flows from operating activities/(Non-current liabilities + Short-term borrowings + Non-current liabilities due within one year)
	Debt asset ratio	X_8	Total liabilities divided by total assets
Operating ability	Accounts receivable turnover	X_9	Operating income divided by accounts receivable balance
	Current asset turnover	X_{10}	Operating income divided by current asset balance
	Fixed asset turnover	X_{11}	Operating income divided by net fixed assets
	Total asset turnover	X_{12}	Operating income total assets
Profitability	Return on assets	X_{13}	(Total profit + Finance costs)/Total assets
	Return on equity	X_{14}	Net profits divided by net assets
	Net sales margin	X_{15}	Net profit divided by sales revenue
	Operating profit margin	X_{16}	Operating profit divided by operating income
	Gross profit margin	X_{17}	(Sales revenue - Cost of sales)/Sales revenue
	Expense Margin for the period of the sale	X_{18}	(Selling expenses + Administrative expenses + Financial expenses) / (Operating income)
Cash flow	Working capital	X_{19}	Current assets minus current liabilities
	Growth rate of net cash flows from operating activities	X_{20}	(Cash flows-Cash flows in the last year)/Cash flows in the last year
Risk management ability	Financial leverage	X_{21}	(Net profit + Income tax + Finance costs) / (Net profit + Income tax)
	Operating leverage	X_{22}	(Net profit + income tax + Finance costs + Depreciation + Amortization of intangible assets + Amortization of long-term amortization expenses) / (Net profit + Income tax + Finance costs)
	Consolidated leverage	X_{23}	Financial leverage multiplied by operating leverage
Growth potential	Basic EPS growth rate	X_{24}	EPS growth divided by EPS in the last year
	Operating profit growth rate	X_{25}	Operating profit growth divided by operating profit in the last year
	Total profit growth rate	X_{26}	Total profit growth divided by total profit in the last year
	Net profit growth rate	X_{27}	Net profit growth divided by net profit in the last year
	Total assets growth rate	X_{28}	Total assets growth divided by total assets in the last year

	Total operating income growth rate	X_{29}	Total operating income growth divided by total operating income in the last year
	Net assets growth rate	X_{30}	Net assets growth divided by net assets in the last year

Note:Debt-paying ability assesses a firm's capacity to marshal its existing assets to service both long-term and short-term liabilities. A robust ability to manage and liquidate debt is pivotal to ensuring a firm's sustained, stable, and prosperous trajectory. Operating ability provides a holistic assessment of a firm's operational, production, and managerial endeavors. It epitomizes the enterprise's strategic foresight and decision-making aptitude, thereby influencing its potential and future prospects. Notably, the efficacy of the firm's operations bears direct implications on its profitability. Profitability quantifies a firm's prowess in accruing profits via its managerial and operational activities. Profits not only assuage creditors by facilitating the repayment of principal and interest but also offer investors a return on their investments. Thus, it stands as a linchpin in the realm of firm operation and management. Cash flow is a firm's dexterity in settling short-term obligations. It offers a lucid snapshot of the enterprise's cash flow health, ensuring its readiness to service impending debts. Risk management underscores a firm's proficiency in navigating operational risks. By proactively gauging risks, an organization's financial and operational resilience can be bolstered. Often, leverage metrics serve as an indicative barometer of a firm's risk management acumen. Growth potential is a testament to a firm's prospective expansion trajectory, and this dimension is gleaned from the growth patterns evident in its financial indicators. Such trends offer insights into the firm's prospective developmental vigor.The occurrence of firm credit default risk is caused by multiple reasons, and the influence of non-financial indicators needs to be considered when making indicator selection.

Table A2. Non-financial indicators

First-level indicator	Second-level indicator	Signal	Explanation
Management attribute	Percentage of male executives	$\boldsymbol{X}_{31}$	Number of male executives divided by that of all executives
	Percentage of executives with master's degree or higher	$\boldsymbol{X}_{32}$	Number of executives with master's degree or higher divided by that of all executives
	Average age of executives	$\boldsymbol{X}_{33}$	Total age of executives divided by number of all executives
	Number of executives	$\boldsymbol{X}_{34}$	Number of all executives
Governance attribute	Shareholding concentration	$\boldsymbol{X}_{35}$	Shareholding of the largest shareholder divided by total shares of the enterprise
	Ownership percentage of real controller	$\boldsymbol{X}_{36}$	Ownership as a right to cash flow
	Control percentage of real controller	$\boldsymbol{X}_{37}$	Control is the right to vote
	Separating extent of ownership and control	$\boldsymbol{X}_{38}$	Difference between ownership and control
	Percentage of shares held by directly controlling shareholders	$\boldsymbol{X}_{39}$	Proportion of shares of listed companies directly held by controlling shareholders
	Duality	$\boldsymbol{X}_{40}$	Whether the chairman and CEO is the same
	Relationship with Top 10 Shareholders	$\boldsymbol{X}_{41}$	Whether the top ten shareholders are related
Macroeconomic factor	GDP growth rate	$\boldsymbol{X}_{42}$	GDP growth divided by GDP in the last year
	CPI growth rate	$\boldsymbol{X}_{43}$	CPI growth divided by CPI in the last year
	RPI growth rate	$\boldsymbol{X}_{44}$	RPI growth divided by RPI in the last year
	PPI growth rate	$\boldsymbol{X}_{45}$	PPI growth divided by PPI in the last year
	Regional fixed-asset investment growth rate	$\boldsymbol{X}_{46}$	Regional fixed-asset investment growth divided by regional fixed-asset investment in the last year
Text sentiment	Positive text sentiment	$\boldsymbol{X}_{47}$	Extracted by text analysis
	Negative text sentiment	$\boldsymbol{X}_{48}$	Extracted by text analysis

Note:Management Attributes is rooted in the tenets of principal-agent theory, and it's posited that the idiosyncrasies inherent in management—both in terms of persona and performance—bear consequential ramifications for a firm's operational image and trajectory. Intrinsically, the caliber of management stands as a pivotal determinant of a firm's developmental prospects. The next is Governance attributes. Pertaining to this dimension, equity concentration serves a dual purpose. On one hand, it empowers predominant shareholders to oversee and regulate the actions of their smaller counterparts, thereby curtailing opportunistic free-riding tendencies. Conversely, in situations characterized by default risk, it may engender detrimental repercussions for the rights and privileges of minor shareholders. In essence, the equity landscape fundamentally informs governance mechanisms and, by extension, the inherent default risk. For Macroeconomic factors, building upon the credit rationing paradigm, it's contended that economic downturns or macroeconomic volatilities often precipitate banks into adopting a more circumspect lending stance. This risk-averse behavior can inadvertently hamstring firms' loan acquisition efforts, amplifying the specter of credit default.

Table A3. Pseudo Code

Pseudo Code 1. ***Three-stage prediction model of ARA-SVM-MPSO***

```
Feature selection: Association rule algorithm- Apriori
INPUT: D: datasets $I = \{I_1, I_2, I_3 \cdots I_n\}$; N: the length of maximum frequent set; Support: support threshold; Confidence: confidence threshold
OUTPUT: extracted Rules with single relationship or joint relationships among variables
Scan D, n=1, calculate candidate itemset: S_1;
Calculate frequent 1-itemset: F_1;
WHILE n<N DO
    n+=1;
    Calculate candidate n itemsets;
    Calculate frequent n itemsets;
END WHILE
Prediction: Support vector machine
INPUT: Variables extracted from ARA and the default risk
OUTPUT: Prediction results for firm default risk
Initialize penalty parameter, kernel parameter, Gamma, etc.
For n=1,...,j
    Set the classification hyperplane: $w*\varphi(x)+b=0$
    Determine the decision function: $\tilde{f}(x) = sign(w*\varphi(x)+b)$
    Transform the decision function: $f(x) = sign(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b)$
    Determine the kernel function: $K(x_i, x_j) = e^{-\gamma(x-x_i)^2}$
END FOR
Optimization: Modified particle swarm optimization algorithm

INPUT: Parameters of SVM: penalty C and kernel parameter γ
OUTPUT: Best penalty C and kernel parameter γ
FOR each particle i
    Initialize velocity V_i and position X_i for particle i;
    Evaluate particle i and set pBest_i=X_i
END FOR
gBest = min{pBest_i}
While not step
    FOR i=1 to N
        Update the velocity V: $v_{(i+1)d} = wv_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id})$
        Update the position X: $x_{(i+1)d} = x_{id} + v_{(i+1)d}$
        Modify the inertia weight: $w(t) = (w_{\max} - w_{\min}) / (1 + \exp(\frac{15t}{T} - 8)) + w_{\min}$
        Evaluate particle i
        IF fit(X_i)<fit(gBest)
        gBest = X_i;
        IF fit(pBest_i)<fit(gBest)
        gBest = pBest_i
    END FOR
END WHILE
OUTPUT gBest
```

References

1. Acheampong, A., & Elshandidy, T. (2021). Does soft information determine credit risk? Text-based evidence from European banks. Journal of International Financial Markets, Institutions and Money, 75, 101303.

2. Ahn, D., Chen, N., & Kim, K. (2024). Robust risk quantification via shock propagation in financial networks. Operations Research, 72(1), 1-18.

3. Akyildirim, E., Hekimoglu, A. A., Sensoy, A., et al. (2023). Extending the Merton model with applications to credit value adjustment. Annals of Operations Research, 326, 27–65.

4. Albuquerque, R., Koskinen, Y., & Zhang, C. (2019). Corporate social responsibility and firm risk: Theory and empirical evidence. Management Science, 65(10), 4451-4469.

5. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23, 589-609.

6. Bai, L., Zheng, K., Wang, Z., & Liu, J. (2022). Service provider portfolio selection for project management using a BP neural network. Annals of Operations Research, 308, 41–62.

7. Baradaran Rezaei, H., Amjadian, A., & Sebt, M. V. et al. (2023). An ensemble method of the machine learning to prognosticate the gastric cancer. Annals of Operations Research, 328, 151–192.

8. Bartov, E., Faurel, L., & Mohanram, P. (2023). The role of social media in the corporate bond market: Evidence from Twitter. Management Science, 69(9), 5638-5667.

9. Beaver, W. H., Cascino, S., Correia, M., & McNichols, F. M. (2019). Group affiliation and default prediction. Management Science, 65(8), 3559-3584.

10. Belhadi, A., Kamble, S. S., Mani, V., et al. (2021). An ensemble machine learning approach for forecasting credit risk of agricultural SMEs' investments in agriculture 4.0 through supply chain finance. Annals of Operations Research.

11. Berloco, C., Argiento, R., & Montagna, S. (2023). Forecasting short-term defaults of firms in a commercial network via Bayesian spatial and spatio-temporal methods. International Journal of Forecasting, 39(3), 1065-1077.

12. Bhatia, S. (2019). Predicting risk perception: New insights from data science. Management Science, 65(8), 3800-3823.

13. Bonsall, S. B., Holzman, E. R., & Miller, B. P. (2017). Managerial ability and credit risk assessment. Management Science, 63(5), 1425-1449.

14. Bucci, A., He, L., & Liu, Z. (2023). Combining dimensionality reduction methods with neural networks for realized volatility forecasting. Annals of Operations Research. Available online.

15. Calabrese, R. (2023). Contagion effects of UK small business failures: A spatial hierarchical autoregressive model for binary data. European Journal of Operational Research, 305(2), 989-997.

16. Casey, C. J., & Bartczak, N. J. (1985). Using operating cash flow data to predict financial distress: Some extensions. Journal of Accounting Research, 23, 384-401.

17. Cathcart, L., Dufour, A., Rossi, L., & Varotto, S. (2020). The differential impact of leverage on the default risk of small and large firms. Journal of Corporate Finance, 60, 101541.

18. Chang, V., Xu, Q. A., Akinloye, S. H., et al. (2024). Prediction of bank credit worthiness through credit risk analysis: An explainable machine learning study. Annals of Operations Research.

19. Chen, T. K., Liao, H. H., & Chen, W. H. (2017). CEO ability heterogeneity, board's recruiting ability and credit risk. Review of Quantitative Finance and Accounting, 49, 1005–1039.

20. Doumpos, M., & Figueira, J. R. (2019). A multicriteria outranking approach for modeling corporate credit ratings: An application of the Electre Tri-nC method. Omega, 82, 166-180.

21. Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 297(3), 1178-1192.

22. Dutta, H. (2022). A consensus algorithm for linear support vector machines. Management Science, 68(5), 3703-3725.

23. Elhoseny, M., Metawa, N., Sztano, G., & El-hasnony I. M. (2022). Deep learning-based model for financial distress prediction. Annals of Operations Research.

24. Fatouh, M., & Giansante, S. (2023). The cyclicality of bank credit losses and capital ratios under the expected loss model. Annals

of Operations Research, 330, 807–840.

25. Fu, S., & Trigilia, G. (2024). Voluntary disclosure, moral hazard, and default risk. Management Science, 70(6), 3447-3469.

26. Giesecke, K., Longstaff, F. A., Schaefer, S., & Strebulaev, I. (2011). Corporate bond default risk: A 150-year perspective. Journal of Financial Economics, 102(2), 233-250.

27. Guerrero, N. M., Moragues, R., Aparicio, J., & Valero-Carreras, D. (2024). Support vector frontiers with kernel splines. Omega, 128, 103130.

28. Guha, D., & Hiris, L. (2002). The aggregate credit spread and the business cycle. International Review of Financial Analysis, 11(2), 219-227.

29. He, C., Geng, X., Tan, C., & Guo, R. (2023). Fintech and corporate debt default risk: Influencing mechanisms and heterogeneity. Journal of Business Research, 164, 113923.

30. Hsu, A. W., & Wang, T. (2013). Does the market value corporate response to climate change? Omega, 41(2), 195-206.

31. Huang, B., Yao, X., Luo, Y., & Li, J. (2022). Improving financial distress prediction using textual sentiment of annual reports. Annals of Operations Research.

32. Jiang, C., Wang, Z., Wang, R., et al. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Annals of Operations Research, 266, 511–529.

33. Jiang, C., Xiong, W., Xu, Q., & Liu, Y. (2021). Predicting default of listed companies in mainland China via U-MIDAS Logit model with group lasso penalty. Finance Research Letters, 38, 101487.

34. Jiang, H., Luo, S., & Dong, Y. (2021). Simultaneous feature selection and clustering based on square root optimization. European Journal of Operational Research, 289(1), 214-231.

35. Jiang, P., Liu, Z., Abedin, M. Z., Wang, J., Yang, W., Dong, Q. (2024). Profit-driven weighted classifier with interpretable ability for customer churn prediction. Omega, 125, 103034.

36. Jiménez-Cordero, A., Morales, J. M., & Pineda, S. (2021). A novel embedded min-max approach for feature selection in nonlinear support vector machine classification. European Journal of Operational Research, 293(1), 24-35.

37. Katsafados, A. G., Leledakis, G. N., Pyrgiotakis, E. G., Androutsopoulos, I., & Fergadiotis, M. (2024). Machine learning in bank merger prediction: A text-based approach. European Journal of Operational Research, 312(2), 783-797.

38. Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. European Journal of Operational Research, 201(2), 838-846.

39. Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Proceedings of the Ninth International Workshop on Machine Learning (pp. 249-256).

40. Korangi, K., Mues, C., & Bravo, C. (2023). A transformer-based model for default prediction in mid-cap corporate markets. European Journal of Operational Research, 308(1), 306-320.

41. Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., & Kou, S. (2021). Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decision Support Systems, 140, 113429.

42. Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y., & Baesens, B. (2019). A multi-objective approach for profit-driven feature selection in credit scoring. Decision Support Systems, 120, 106-117.

43. Labbé, M., Landete, M., & Leal, M. (2023). Dendrograms, minimum spanning trees, and feature selection. European Journal of Operational Research, 308(2), 555-567.

44. Lee, I. G., Yoon, S. W., & Won, D. (2022). A mixed integer linear programming support vector machine for cost-effective group feature selection: Branch-cut-and-price approach. European Journal of Operational Research, 299(3), 1055-1068.

45. Leow, M., & Crook, J. (2016). The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. European Journal of Operational Research, 249(2), 457-464.

46. Li, J., Cheng, W., Song, X., & Zheng, Y. (2023). The value of buyer financing with a minimum quantity commitment in pull supply chains. Omega, 121, 102938.

47. Liu, X., Li, X., & Sarkar, S. (2023). Cost-restricted feature selection for data acquisition. Management Science, 69(7), 3976-3992.

48. Liu, Y., Yang, M., Wang, Y., Li, Y., Xiong, T., & Li, A. (2022). Applying machine learning algorithms to predict default

probability in the online credit market: Evidence from China. International Review of Financial Analysis, 79, 101971.

49. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66, 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

50. Lu, Y., Yang, L., Shi, B., Li, J., & Abedin, M. Z. (2022). A novel framework of credit risk feature selection for SMEs during industry 4.0. Annals of Operations Research.

51. Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2023). Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural networks. Annals of Operations Research, 330, 609–637.

52. Matin, R., Hansen, C., Hansen, C., & Mølgaard, P. (2019). Predicting distresses using deep learning of text segments in annual reports. Expert Systems with Applications, 132, 199-208.

53. Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 18, 109-131.

54. Palma, M., De Iaco, S., Cappello, C., & Distefano, V. (2023). Tourism composite spatial indicators through variography and geographically weighted principal components analysis. Annals of Operations Research. Available online.

55. Premachandra, I. M., Chen, Y., & Watson, J. (2011). DEA as a tool for predicting corporate failure and success: A case of bankruptcy assessment. Omega, 39(6), 620-626.

56. Ren, L., Cong, S., Xue, X., et al. (2023). Credit rating prediction with supply chain information: A machine learning perspective. Annals of Operations Research.

57. Roeder, J., Palmer, M., & Muntermann, J. (2022). Data-driven decision-making in credit risk management: The information value of analyst reports. Decision Support Systems, 158, 113770.

58. Rozo, B. J. G., Crook, J., & Andreeva, G. (2023). The role of web browsing in credit risk prediction. Decision Support Systems, 164, 113879.

59. Sariyer, G., Mangla, S. K., Kazancoglu, Y., et al. (2021). Data analytics for quality management in Industry 4.0 from a MSME perspective. Annals of Operations Research.

60. Séverin, E., & Veganzones, D. (2021). Can earnings management information improve bankruptcy prediction models? Annals of Operations Research, 306, 247–272.

61. Shi, B., Bai, C., & Dong, Y. (2024). A big data analytics method for assessing creditworthiness of SMEs: Fuzzy equifinality relationships analysis. Annals of Operations Research.

62. Sigrist, F., & Leuenberger, N. (2023). Machine learning for corporate default risk: Multi-period prediction, frailty correlation, loan portfolios, and tail probabilities. European Journal of Operational Research, 305(3), 1390-1406.

63. Simic, V., Ebadi Torkayesh, A., & Ijadi Maghsoodi, A. (2023). Locating a disinfection facility for hazardous healthcare waste in the COVID-19 era: A novel approach based on Fermatean fuzzy ITARA-MARCOS and random forest recursive feature elimination algorithm. Annals of Operations Research, 328, 1105–1150.

64. Simsek, S., Dag, A., Tiahrt, T., & Oztekin, A. (2021). A Bayesian belief network-based probabilistic mechanism to determine patient no-show risk categories. Omega, 100, 102296.

65. Simumba, N., Okami, S., Kodaka, A., & Kohtake, N. (2022). Multiple objective metaheuristics for feature selection based on stakeholder requirements in credit scoring. Decision Support Systems, 155, 113714.

66. Sun, J., Xiao, K., Liu, C., Zhou, W., & Xiong, H. (2019). Exploiting intra-day patterns for market shock prediction: A machine learning approach. Expert Systems with Applications, 127, 272-281.

67. Wang, J., Zhuang, Z., & Gao, D. (2023). An enhanced hybrid model based on multiple influencing factors and divide-conquer strategy for carbon price prediction. Omega, 120, 102922.

68. Wu, D., Ma, X., & Olson, D. L. (2022). Financial distress prediction using integrated Z-score and multilayer perceptron neural networks. Decision Support Systems, 159, 113814.

69. Xia, H., Liu, J., & Zhang, Z. J. (2024). Identifying fintech risk through machine learning: Analyzing the Q&A text of an online loan investment platform. Annals of Operations Research, 333, 579–599.

70. Xie, X., Shi, X., Gu, J., Xu, X. (2023). Examining the contagion effect of credit risk in a supply chain under trade credit and bank loan offering. Omega, 115, 102751.

71. Xu, Y., Kou, G., Peng, Y., Ding, K., Ergu, D., Alotaibi, F. S. (2024). Profit- and risk-driven credit scoring under parameter uncertainty: A multiobjective approach. Omega, 125, 103004.

72. Xu, Z., Yang, F., Tang, C., Wang, H., Wang, S., Sun, J., & Zhang, Y. (2024). FG-HFS: A feature filter and group evolution hybrid feature selection algorithm for high-dimensional gene expression data. Expert Systems with Applications, 245.

73. Yang, J., Li, Y., Fang, L. (2023). Financing capacity planning with environmental considerations: A non-parametric analysis. Omega, 118, 102866.

74. Yfanti, S., Karanasos, M., Zopounidis, C., & Christopoulos, A. (2023). Corporate credit risk counter-cyclical interdependence: A systematic analysis of cross-border and cross-sector correlation dynamics. European Journal of Operational Research, 304(2), 813-831.

75. Yoganarasimhan, H. (2020). Search personalization using machine learning. Management Science, 66(3), 1045-1070.

76. Zhang, W., Yan, S., Li, J., et al. (2024). Deep reinforcement learning imbalanced credit risk of SMEs in supply chain finance. Annals of Operations Research.

77. Zhang, X., Zhao, Y.,& Yao, X. (2022). Forecasting corporate default risk in China. International Journal of Forecasting,38(3),1054-1070.

78. Zhang, Y., Zhu, R., Chen, Z., Gao, J., & Xia, D. (2021). Evaluating and selecting features via information theoretic lower bounds of feature inner correlations for high-dimensional data. European Journal of Operational Research, 290(1), 235-247.

79. Zhou, H., Zhang, J., Zhou, Y., Guo, X.,& Ma, Y. (2021a). A feature selection algorithm of decision tree based on feature weight. Expert Systems with Applications,164,113842.

80. Zhou, Z., Gao, M., Xiao, H., Wang, R., & Liu, W. (2021). Big data and portfolio optimization: A novel approach integrating DEA with multiple data sources. Omega, 104, 102479.

81. Zhu, P., Hou, X., Tang, K., et al. (2023). Compactness score: A fast filter method for unsupervised feature selection. Annals of Operations Research.