A Hybrid Investment Risk Prediction Framework Integrating ADASYN-RF, CS-SVM, PCA-BP, and ARIMA Models

Liucheng Zhang
Henan Quality Institute, Pingdingshan 467000, China
E-mail: zhc582@163.com

Keywords: Adaptive synthetic sampling approach, random forest, investment, risk prediction

Received: August 4, 2025

Investment risk is often the result of a long-term interplay among multiple factors, such as market volatility and business performance. Predicting investment risk in advance helps investors avoid losses and protect their assets. However, current prediction methods rely heavily on historical data, leading to poor timeliness and unreliable results. Therefore, this study proposes an investment risk prediction model based on the adaptive comprehensive sampling algorithm and the random forest algorithm to predict investment risks accurately. This model makes full use of the adaptive comprehensive sampling algorithm to balance the categories of risk data, and the random forest captures the risk characteristics to achieve investment risk prediction. CS-SVM is introduced to improve the prediction model and prevent overfitting. Meanwhile, principal component analysis and backpropagation networks are combined to solve the problems of unstructured data processing and time series dependency modeling. In the experiment, the study took historical investment data as the dataset and compared it with the investment risk prediction models constructed by three algorithms: CatBoost-GJO, TS-GA, and GAN-Stacking. Evaluate the prediction accuracy of each method in investment risk prediction, including the Sharpe ratio, value at risk, volatility, win rate and profit and loss. The results show that the prediction accuracy of this model reaches 99.0%, the Sharpe ratio is 1.9, and the maximum drawdown is 5.9%, all superior to the comparison models. Moreover, in the actual investment risk prediction, its volatility and value at risk are only 8.2% and 7.3% respectively, while the winning rate and profit and loss ratio can reach 88.9%, 6:1. These results indicate that the proposed model achieves high accuracy in investment risk prediction, effectively addressing the limitations of existing methods and providing a new approach to improve prediction performance. It contributes to the development of intelligent and efficient investment risk forecasting.

Povzetek: Študija predlaga model za napovedovanje investicijskega tveganja, ki z adaptivnim uravnoteženjem vzorcev, naključnim gozdom in CS-SVM ter s kombinacijo PCA in povratno-propagacijskih mrež za obdelavo ne-strukturiranih in časovnih podatkov izboljša pravočasnost in zanesljivost napovedi.

1 Introduction

In the context of deep integration and high-frequency fluctuations in global financial markets, asset price linkage has become significantly stronger. The risk exposure of a single investment can easily spread through financial chains and lead to systemic financial risks [1]. From the perspective of financial dynamics, the accumulation of investment risk results from the combination of multiple factors, including macroeconomic cycles and industry competition. Accurately identifying risk signals before a risk outbreak not only helps investors avoid losses but also prevents a chain reaction of market risks [2]. Therefore, building a scientific and efficient investment risk prediction mechanism is essential for maintaining financial stability and ensuring steady asset growth. Current prediction methods mainly include historical data analysis, econometric models, and artificial intelligence algorithms. However, these approaches often rely too much on historical data and fail to adapt to complex and changing market environments. As a result, there is a growing need for a flexible and accurate investment prediction method [3]. Investment risk prediction primarily focuses on balancing data distribution and modeling nonlinear relationships. The Adaptive Synthetic Sampling Approach (ADASYN) effectively addresses class imbalance in investment data, while the Random Forest (RF) algorithm excels at identifying complex nonlinear patterns in data [4, 5]. Therefore, this study develops a risk prediction model based on ADASYN and RF. The model aims to improve the accuracy of investment risk forecasting in rapidly changing financial markets, assisting investors in managing risks and maintaining asset stability and market order. By combining ADASYN and RF, the model provides a novel approach to address the limitations of traditional investment risk prediction methods. It provides a new analytical framework that integrates ADASYN and RF, filling the research gap in accurate risk forecasting under market volatility. This study aims to address the

insufficient accuracy and poor adaptability of current investment risk prediction methods and achieve high-precision risk prediction. The study uses the ADASYN algorithm to address class imbalance in investment data, thereby enhancing the model's ability to detect rare risk events. Combined with the powerful nonlinear modeling and feature selection capabilities of the RF algorithm, the generalization performance of the prediction model is further strengthened. The key parameters of RF are optimized using CS and Support Vector Machine (SVM) optimization algorithms to enhance the model's adaptability and prediction accuracy across different market environments. Finally, the model parameters are optimized by combining principal component analysis and backpropagation algorithm to solve the problems of unstructured data processing and time series dependency modeling.

2 Related works

ADASYN effectively addresses class imbalance in data, while the RF algorithm captures complex nonlinear relationships. Both methods have been widely applied to solve complex problems in various fields. Scholars from around the world have explored these approaches in depth [6]. To overcome the limitations of traditional methods in handling imbalanced data, Song et al. proposed a credit risk prediction model. This model used an improved ADASYN for sampling and trained the data with a lightweight gradient boosting machine. The results showed that the model outperformed other methods on loan datasets [7]. In response to security threats in the medical Internet of Things, Salehpour et al. developed a hybrid intrusion detection framework. The framework applied the Extreme Gradient Boosting algorithm to detect anomalies, used ADASYN for resampling, and adopted RF for intrusion detection. Experiments demonstrated that the framework achieved an accuracy of 92.23% in threat identification [8]. To evaluate pile performance in cone penetration tests, Dawei et al. proposed two models—one combining Particle Swarm Optimization with RF, and another integrating Harris Hawks Optimization with RF. A comparison of the two models showed that the one using Harris Hawks Optimization with RF delivered better performance [9]. Addressing the challenge of weighting influencing factors in fine particulate matter prediction, Ding et al. introduced a model combining weighted RF and a Long Short-Term Memory network. The RF algorithm was used to filter key factors, and the weighted data were input into the Long Short-Term Memory network for training. The results showed that the proposed model had stronger prediction capabilities [10].

Over time, the theoretical and practical applications of risk prediction methods have matured, and researchers in many countries have conducted in-depth studies [11]. Belhadi et al. proposed a hybrid ensemble learning method based on Rotation Forest to address credit risk in small and medium-sized enterprises. By analyzing data from 216 agricultural SMEs in Africa, the study confirmed that the model could provide reliable risk forecasts for small businesses [12]. Li et al. proposed a financial risk prediction model that optimizes the backpropagation neural network for the problem of enterprise financial risks. Analysis based on the financial data of listed companies from 2017 to 2020 as samples indicates that the model's accuracy rate in predicting the financial distress of normal companies exceeds 80% [13]. Doğru et al. addressed the inefficiency and low accuracy of diabetes risk prediction by proposing a new model based on super ensemble learning. The model used logistic regression and other methods as base learners and Support Vector Machine (SVM) as the meta-learner, achieving a prediction accuracy of 92 to 99.6% [14]. In response to limited data and privacy concerns in supply chain risk prediction, Zheng et al. put forward a method based on federated learning. Using a buyer-based prediction of supplier delivery delays as an example, the study explored the effects of data imbalance. Experimental results showed that this approach improved risk prediction capability [15].

In summary, although existing studies have advanced risk prediction, limitations remain in comprehensiveness and intelligence. In response to this, the study proposes a risk prediction model integrating ADASYN-RF, CS-SVM, PCA-BP, and ARIMA. The research expects that the proposed method can improve the efficiency and accuracy of risk prediction in the volatile financial market, so as to meet investors' demands for investment risk prediction. The comparison and summary of the relevant work are shown in Table 1.

Table 1: A comparison and summary of related work

Number	Usage method	Accuracy (%)	Dataset	Restriction
[7]	Improve the ADASYN+ LGBM	-	Credit data of a certain commercial bank from 2015 to 2020	The influence of feature interaction was not taken into account
[8]	ExBoost +ADASYN	92.23	IoMT device network traffic log	The noise sample processing is insufficient
[9]	The Harris Eagle optimization algorithm combined with RF+ADASYN	-	On-site record of CPT static cone penetration test	It is prone to fall into local optimum
[10]	RF+ Long Short-	-	Data of PM2.5 monitoring	Insufficient consideration of spatial

	Term Memory network		stations in the Beijing-Tianjin-Hebei region from 2018 to 2022	heterogeneity and unsatisfactory optimization of the attention mechanism
[12]	Ensemble learning + Rotating Forest algorithms	-	Transaction records of agricultural supply chain finance in Southeast Asia from 2018 to 2023	Sample selection bias and the influence of weather dynamics lead to a decline in prediction accuracy
[13]	Generative Adversarial Network	92.0-99.6	Risk event data of a certain financial regulatory platform	Sample selection bias and the influence of weather dynamics lead to a decline in prediction accuracy
[14]	Logistic regression +SVM	-	Diabetes screening data in Turkey from 2021 to 2021	Insufficient fusion of multimodal data leads to high computational complexity
[15]	Federated learning	-	Supply chain platform data	The communication efficiency is low and the problem of independent and equally distributed data has not been solved

As shown in Table 1, the current technology is unable to effectively handle time-non-time mixed investment data. This is because traditional methods often only model a single type of data, making it difficult to simultaneously capture the complex correlations between the dynamic evolution characteristics of time series and the static attributes of non-time data. The AD-PB proposed in the research can solve this limitation. The AD-PB model integrates time series analysis and non-time feature processing, effectively modeling dynamic patterns and static attributes in investment risks. By using a variety of technical means such as the hybrid adaptive comprehensive sampling algorithm, random forest algorithm, principal component analysis, backpropagation network, and support vector machine, the model can adaptively capture key risk factors and simultaneously alleviate the information interference problem caused by data mixing, thereby significantly improving the prediction performance and applicability.

3 Hybrid model design for investment risk prediction

3.1 Risk prediction algorithm combining Adasyn-Rf and improved SVM

The ADASYN-RF algorithm integrates ADASYN and RF to address challenges in risk prediction effectively. ADASYN-RF first uses ADASYN to generate synthetic samples for the minority class, balancing the data distribution and improving the representation of high-risk samples. Then, RF processes the balanced data, reduces noise influence, and handles high-dimensional features, enhancing the accuracy and robustness of risk prediction [16]. The workflow of ADASYN-RF is shown in Figure 1.

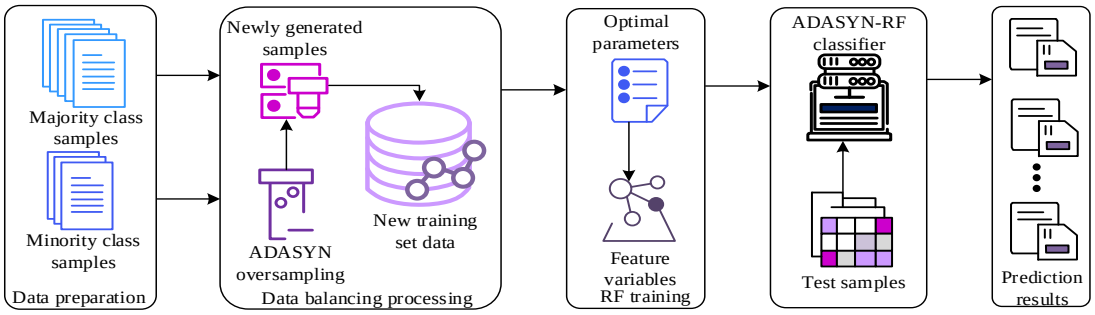


Figure 1: Workflow of the ADASYN-RF algorithm

As shown in Figure 1, the hybrid algorithm first collects the training dataset and divides it into minority and majority classes. ADASYN is then applied to oversample the minority class, and the generated samples are combined with the majority class to form a new training set, completing the data balancing process. Next, optimal parameters are selected, and feature variables are determined based on their importance to train the RF model. After training, the testing samples are input into the trained ADASYN-RF classifier. Finally, the classifier analyzes the testing data and outputs the prediction results, achieving the goal of risk prediction. The summary of the abbreviations and their full names that appear in the article is shown in Table 2.

Table 2: List of abbreviations and their full forms

Abbreviation	Full name
CS	Chaotic search
ADASYN	Adaptive synthetic sampling
BP	Back propagation
RF	Random forest
ARIMA	Autoregressive Integrated Moving Average model
PCA	Principal Components Analysis
SVM	Support vector machine
ADRS	ADASYN-RF-CS-SVM
AD-PB	ADRS-PCA-BP

The oversampling calculation in ADASYN is defined in Equation (1).

$$\beta_i = \frac{1 - D(x_i)}{\sum_{j-1}^{n_{\min}} (1 - D(x_j))} \tag{1}$$

In Equation (1), β_i represents the generation weight of the i -th minority sample, determining how many synthetic samples to generate. x_i denotes the i -th minority class sample. $D(x_i)$ indicates the distribution density of sample x_i , and $n_{\min}$ is the total number of minority class samples. The generation of new samples in ADASYN is further calculated as shown in Equation (2).

$$x_{new} = x_i + \alpha \times (x_{nn} - x_i) \tag{2}$$

In Equation (2), x_{nn} denotes the nearest neighbor of x_i among minority samples. α is the interpolation coefficient ranging within $[0,1]$, which controls the position of the new sample on the line between x_i and x_{nn} . x_{new} is the newly generated synthetic sample, belonging to the minority class and used for balancing the data distribution. The ADASYN-RF algorithm combines the advantages of both ADASYN and RF. ADASYN addresses data imbalance, while RF improves generalization. Together, they reduce false prediction rates in risk forecasting. However, ADASYN-RF has limitations in handling high-dimensional data efficiently. Its parameter tuning is complex, and it may suffer from overfitting. The CS-SVM algorithm, which combines CS and SVM, offers strong capabilities in high-dimensional space, uses simpler parameters, and mitigates overfitting through the hard or soft margin mechanism [17]. The CS algorithm conducts global search by introducing chaotic variables and can quickly converge to the optimal solution region in a complex parameter space. SVM, on the other hand, utilizes the principle of minimizing structural risk to construct the optimal classification hyperplane in a high-dimensional space. Combining the global optimization ability of CS and the good adaptability of SVM to high-dimensional data, CS-SVM shows higher stability and accuracy when solving the classification problem of imbalanced data. The inner core selects the RBF function, and the gamma value is optimized through the CS algorithm to avoid local optima. The value of C is set to a fixed value, which is 1. Therefore, this study integrates CS-SVM into ADASYN-RF for further improvement. The workflow of the CS-SVM algorithm is shown in Figure 2.

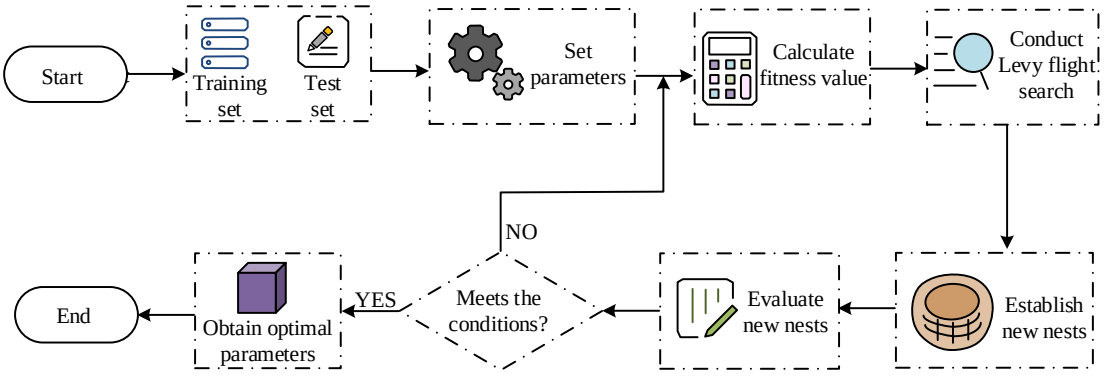


Figure 2: Workflow of the CS-SVM algorithm

As shown in Figure 2, the algorithm first divides the data into training and testing sets and normalizes them to unify the scale, improving processing efficiency and reducing the complexity of high-dimensional data. Then, parameters for SVM and CS are set. Bird nest positions are initialized randomly, and their fitness values are calculated. The CS algorithm updates positions using Levy flights to enhance the handling of high-dimensional data. To address class imbalance, poor-performing nests are discarded, and new nests are generated and evaluated

in each iteration. This process helps reduce false prediction rates. In multi-class scenarios, continuous parameter optimization enables the SVM model to better distinguish complex classes. Finally, the model is finalized based on the optimal parameters. The calculation is shown in Equation (3).

$$X_i^{t+1} = X_i^t + \alpha \otimes Levy(\lambda) \tag{3}$$

In Equation (3), X_i^t represents the position of the i-th nest in the t-th generation. α is the step size scaling factor, which controls the magnitude of the Levy step. $Levy(\lambda)$ is the Levy random path, a random step following the Levy distribution. λ is a parameter of the Levy distribution [18]. The new nests are generated based on the calculation shown in Equation (4).

$$X_{new} = X_j^t + rand \cdot (X_k^t - X_l^t) \tag{4}$$

In Equation (4), X_{new} is the new nest position vector. X_j^t is the current best-performing nest in the iteration. X_k^t and X_l^t represent randomly selected nest position vectors indexed by k and l. $rand$ is a random scalar within interval $[0,1]$. The hybrid algorithm combining ADASYN-RF and CS-SVM (ADRS) follows the workflow illustrated in Figure 3.

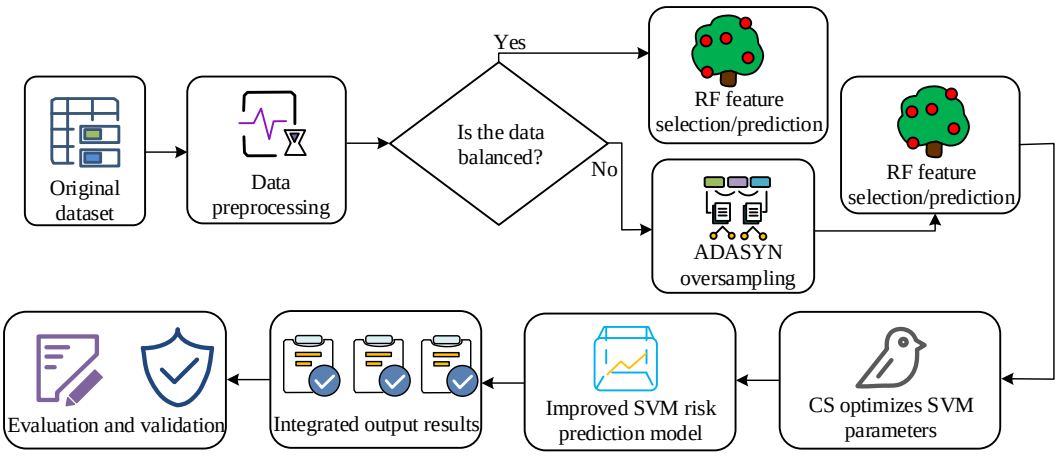


Figure 3: Workflow of the ADRS algorithm

As shown in Figure 3, ADRS first collects the original dataset and performs preprocessing, including filling missing values and applying standardization or normalization. Then, it checks for data balance. If the data is balanced, RF is directly applied for feature selection and prediction. If the data is imbalanced, ADASYN is used for oversampling, followed by RF for feature selection and prediction. Next, the CS algorithm is used to optimize SVM parameters and build an improved SVM risk prediction model. The results from RF and the improved SVM are integrated and output together. Finally, evaluation metrics such as the confusion matrix are used to assess the performance of the ensemble model and determine its risk prediction capability. The calculation of RF feature selection is shown in Equation (5).

$$Gini(X) = 1 - \sum_{k=1}^{K} p_k^2 \tag{5}$$

In Equation (5), $Gini(X)$ represents the Gini index of feature X, which measures the impurity of the dataset. K is the total number of classes, and p_k is the proportion of samples in class k. After completing these steps, SVM performance needs to be evaluated in preparation for optimization, as shown in Equation (6).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

In Equation (6), TP is the number of correctly predicted positive samples. TN is the number of correctly predicted negative samples. FP represents the number of false positives, and FN represents the number of false negatives.

3.2 Investment risk prediction model based on the ADRS hybrid algorithm

Although the ADRS algorithm achieves high prediction accuracy, real-world investment data exhibit three characteristics: unstructured information, strong temporal dependence, and sensitivity to unexpected events. These features result in dynamic and nonlinear structures in investment data. However, the ADRS algorithm struggles to capture such dynamic nonlinear relationships. The PCA-BP hybrid algorithm, combining Principal Component Analysis (PCA) and

Backpropagation (BP) neural network, effectively addresses this issue. It captures nonlinear temporal patterns and dynamically adapts to data changes through layered nonlinear transformation, while dimensionality reduction helps remove redundant features. This approach addresses the challenges of processing unstructured data and modeling temporal dependencies [19]. The workflow of PCA-BP is shown in Figure 4.

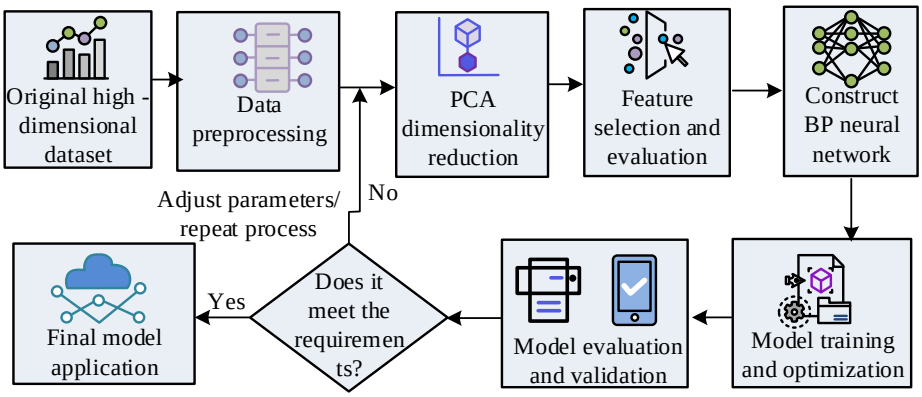


Figure 4: Workflow of the PCA-BP algorithm

In Figure 4, PCA-BP first loads the original high-dimensional investment dataset and performs data preprocessing. PCA is then applied for dimensionality reduction by calculating the mean vector and covariance matrix to select the top principal components, significantly reducing data dimensions. Next, it selects and evaluates the reduced features to ensure low correlation among them. A BP network is then constructed. The number of input nodes is reduced to the number of principal components, which decreases the number of connections and parameters. This simplifies the network structure, reduces computational cost, and lowers the risk of overfitting. The reduced input dimension also minimizes parameter redundancy and shortens training time, thus improving the BP network. The calculation of the PCA mean vector is shown in Equation (7).

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7}$$

In Equation (7), $\bar{X}$ represents the mean vector of the data, which is composed of the average values of all features. n is the number of samples, and X_i is the raw data of the i-th sample, represented as a row vector of dimension $1\times m$ [20]. After this, the covariance matrix is calculated as shown in Equation (8).

$$C = \frac{1}{n-1} X_{centered}^{T} X_{centered} \tag{8}$$

In Equation (8), $X_{centered}$ denotes the centralized data matrix, which eliminates the influence of the mean. $X_{centered}^{T}$ is the transpose of the centralized matrix, and C is the covariance matrix. While PCA-BP reduces dimensionality and redundancy via PCA and improves prediction accuracy and efficiency via BP, it still struggles to model dynamic trends and periodic fluctuations in time series data. It also cannot effectively utilize sequence autocorrelation. The Autoregressive Integrated Moving Average (ARIMA) model, specifically designed for time series, stabilizes data through differencing and models trends and seasonality using autoregression and moving average techniques. This enhances PCA-BP's ability to make dynamic predictions. Therefore, the study integrates PCA-BP with ARIMA to form the PBA algorithm. The workflow of PBA is shown in Figure 5.

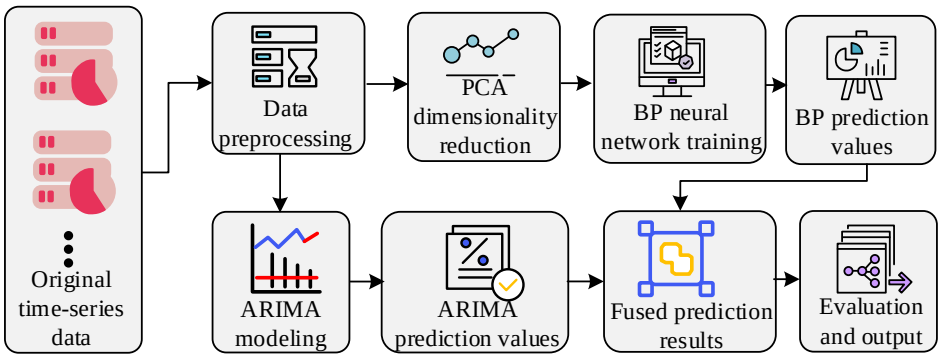


Figure 5: Workflow of the PBA algorithm

As shown in Figure 5, this method first inputs the original time series data and performs preprocessing. One portion of the data is used for ARIMA modeling, which establishes an autoregressive moving average model by differencing the data and performing stationarity analysis. This allows the model to capture sequence autocorrelation and model dynamic trends and seasonal fluctuations. Meanwhile, another portion of the data undergoes PCA-based dimensionality reduction and is input into the BP network for training. Finally, the predictions from ARIMA and BP are integrated to compensate for PCA-BP's weaknesses in handling dynamic patterns. This produces more accurate results, which are then evaluated. In the fusion mechanism, the combination of ARIMA and BP prediction results is not a simple average. The study introduces the idea of combining static weights and dynamic adaptation of learning weights. The static weights are artificially set based on the historical performance of the model, while the learning weights are automatically learned from the data through optimization algorithms. The study first applies the sliding window method to dynamically evaluate historical errors and, based on this, calculates the real-time weight distribution of the ARIMA and BP neural network models. Through the rolling optimization mechanism, the weights are updated with the prediction cycle, thereby adapting to the evolution of the characteristics of time series data. In addition, to prevent a single weight from dominating the fusion result, a weight smoothing constraint term is introduced to ensure the stability of weight distribution and a smooth transition. The ARIMA modeling process is shown in Equation (9).

$$(1-\sum_{i=1}^{p}\phi_i B^i)(1-B)^d Y_t = (1+\sum_{i=1}^{q}\theta_i B^i)\varepsilon_t \quad (9)$$

In Equation (9), Y_t represents the observed value of the time series at time t. B is the lag operator. ϕ_i is the parameter for the autoregressive part. d is the number of differences used to stabilize the series. θ_i represents the parameter of the moving average part. ε_t is the white noise error term. Meanwhile, the other portion of the data is sent to the BP network for training. The forward computation from the input layer to the hidden layer is shown in Equation (10).

$$net_h = \sum_{i=1}^{n} v_{ih} x_i - \Upsilon_h \quad (10)$$

In Equation (10), net_h and Υ_h denote the net input and threshold of the h-th neuron in the hidden layer. n is the number of neurons in the input layer. v_{ih} is the weight connecting the i-th neuron in the input layer to the h-th neuron in the hidden layer. x_i is the input value from the i-th neuron in the input layer. Since the ARIMA and BP prediction models are processed separately, in order to align the range of their prediction results and solve the problems of time lag or errors between the two paths, it is necessary to normalize the two sets of prediction results. The study first scales the predicted values to the interval [0, 1] through the minimum-maximum normalization method. To avoid the influence of time lag on the fusion results, the study adopted the sliding window correlation analysis method to time align the two groups of sequences. The study takes the output of the ARIMA model as the benchmark, calculates the Pearson correlation coefficients between it and the output of the BP model at different time offsets, and selects the offset corresponding to the maximum correlation coefficient as the optimal time compensation parameter. The investment risk prediction model combining ADRS and PBA is named AD-PB. Its workflow is shown in Figure 6.

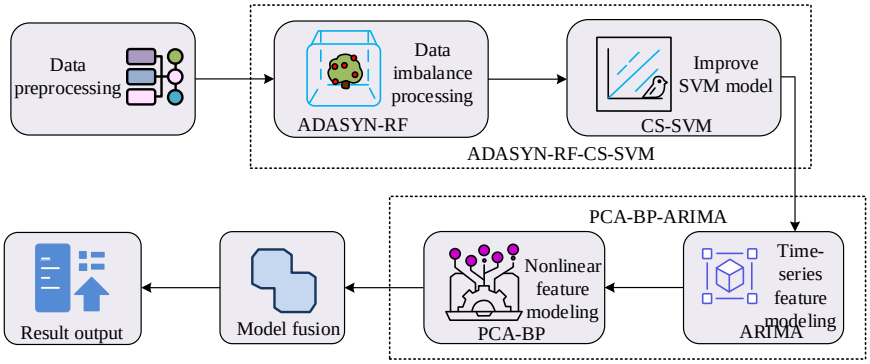


Figure 6: Workflow of the AD-PB algorithm

As shown in Figure 6, this model first takes the raw investment data and performs data cleaning and standardization. The data is then split into two parts. One part is used to build the ADASYN-RF-CS-SVM model. ADASYN balances the risk categories in the training data, RF extracts features, and CS optimizes SVM hyperparameters. The optimized SVM is then integrated with RF for classification modeling. The other part is used to construct the PCA-BP-ARIMA model. Stationarity tests and differencing are applied to the time series data for ARIMA modeling and prediction. Non-time-series data undergoes PCA for dimensionality reduction. The ARIMA predictions are then combined with the PCA-reduced features and input into the BP network for

training. Finally, the outputs from the two models are fused at both the feature and result levels. Techniques such as weighted averaging are used to produce the final investment risk prediction. The forward propagation in BP network training is calculated as shown in Equation (11).

$$a_j = f\left(\sum_{i=1}^{m} w_{ji} x_i + b_j\right) \tag{11}$$

In Equation (11), w_{ji} represents the weight from the i-th input neuron to the j-th neuron in the current layer. b_j and a_j are the bias term and output of the j-th neuron in the current layer. After this computation, BP also performs backpropagation, which is calculated as shown in Equation (12).

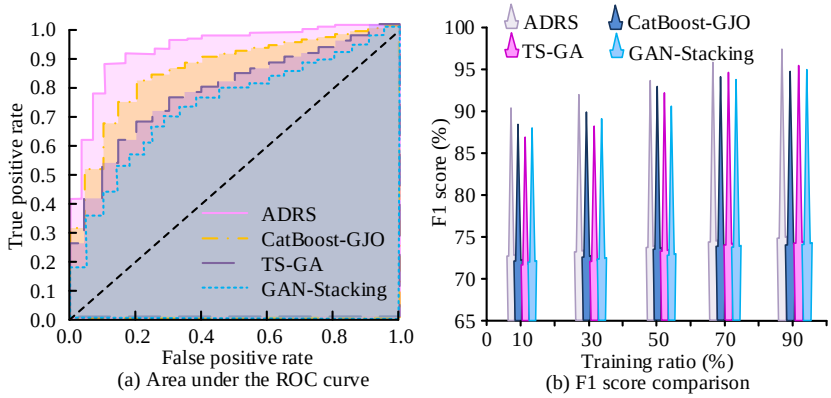
$$\begin{cases} \Delta w_{ji} = -\eta \dfrac{\vartheta E}{\vartheta w_{ji}} \\ \Delta b_j = -\eta \dfrac{\vartheta E}{\vartheta b_j} \end{cases} \tag{12}$$

In Equation (12), Δw_{ji} and Δb_j are the update values for weight w_{ji} and bias b_j, respectively. η is the learning rate. E is the error function. $\dfrac{\vartheta E}{\vartheta w_{ji}}$ and $\dfrac{\vartheta E}{\vartheta b_j}$ are the partial derivatives of the error function with respect to weight w_{ji} and bias b_j, respectively.

4 Performance analysis of the investment risk prediction model based on ADASYN-RF

4.1 Validation of the effectiveness of the ADRS algorithm

To verify the superiority of the ADRS risk prediction algorithm, this study compared it with three other algorithms: CatBoost-GJO (a combination of Categorical Boosting and Golden Jackal Optimization), TS-GA (a combination of Tabu Search and Genetic Algorithm), and GAN-Stacking (a combination of Generative Adversarial Network and Stacked Generalization). Risk scenarios were simulated through data injection, and the dataset used for risk prediction was the S&P 500 historical trading dataset, which is widely recognized and representative in the field of financial risk research. The investment time frame in the experiment was set from January 2020 to December 2024, covering asset classes such as stocks, bonds, and derivatives. Market conditions included bull and bear market transitions as well as extreme volatility scenarios. The study first preprocesses the original data using missing value imputation, standardization, and sliding window feature construction, then extracts key features and divides them into training and test sets. Price data includes the opening price, closing price, highest price and lowest price. The return data includes daily return rate and weekly return rate. Macroeconomic variables include the inflation rate, GDP growth rate, changes in interest rates, etc. The experiments were implemented using Python as the programming environment, with a high-performance CPU as the hardware platform, and TensorFlow as the deep learning framework. Historical investment data was used as the dataset. First, the study compared the Area under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve and the F1 scores for the four algorithms. The experimental results are shown in Figure 7.



(a) Area under the ROC curve (b) F1 score comparison

Figure 7: Comparison results of ROC curve and F1 score

As shown in Figure 7 (a), the ROC curve of the ADRS algorithm was closest to the top-left corner. Its AUC value reached 0.897, the highest among all algorithms, significantly outperforming the comparison methods with AUC values of 0.857, 0.793, and 0.781. According to Figure 7 (b), the F1 scores of the ADRS algorithm were all above 90.0%, with the lowest and highest values being 91.1% and 97.5%, respectively. These results were higher than those of the other algorithms, further demonstrating the classification performance of ADRS. In summary, the

ADRS algorithm outperformed the comparison algorithms in both classification and recall. To further evaluate the prediction performance of each model, the study compared the predicted values with the actual values. The results are shown in Figure 8.

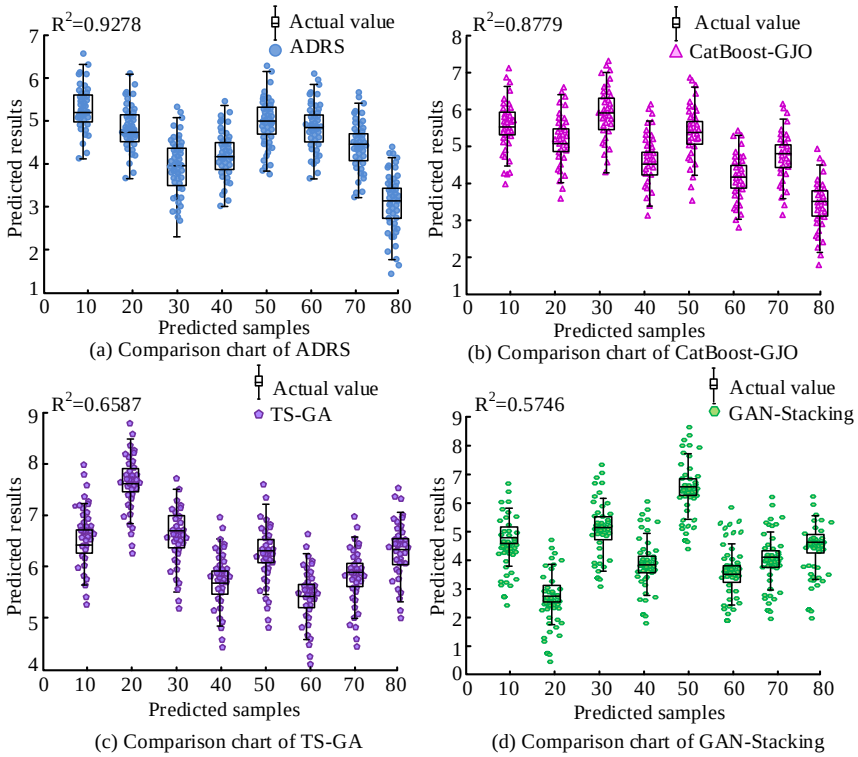


Figure 8: Comparison results of predicted values and actual values

As shown in Figure 8 (a), the predicted values of the ADRS algorithm closely matched the actual values, with a Coefficient of Determination (R^2) of 0.9278. According to Figure 8 (b), CatBoost-GJO also produced predictions that closely matched the actual values, with an R^2 of 0.8779. In contrast, TS-GA and GAN-Stacking showed lower agreement between predicted and actual values, with more outliers. Their R^2 values were 0.6587 and 0.5746, respectively. Overall, ADRS exhibited the smallest prediction error and the best fit among all models. To evaluate the prediction efficiency of the four algorithms, the study also compared their memory usage and response time. The experimental results are shown in Figure 9.

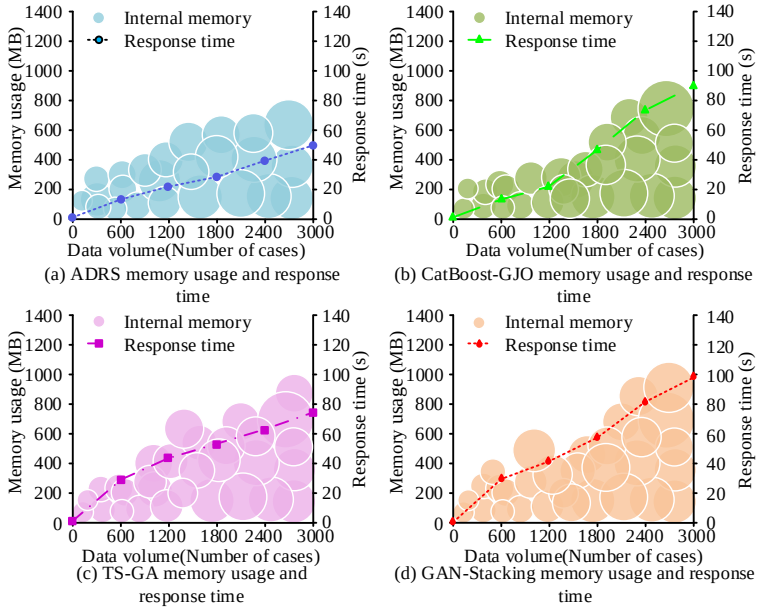


Figure 9: Comparison results of memory usage and response time

As shown in Figure 9 (a), the memory usage of the ADRS algorithm increased slowly, with a maximum memory consumption of 795 MB. When the dataset size reached 3000, its response time was 50 seconds. According to Figure 9 (b), CatBoost-GJO had a maximum memory usage of 950 MB and the highest response time of 89 seconds. From Figure 9 (c) and 9 (d), the maximum memory usage of TS-GA and GAN-Stacking was 1010 MB and 1080 MB, respectively. Their highest response times were 74 seconds and 99 seconds. In summary, compared with the other algorithms, ADRS demonstrated more stable and efficient data processing performance.

4.2 Application evaluation of the improved ADASYN-RF investment risk prediction model

To further evaluate the performance of the AD-PB investment risk prediction model based on the ADASYN-RF algorithm, the study compared it with three other investment risk prediction models constructed using CatBoost-GJO, TS-GA, and GAN-Stacking. The experimental environment consisted of a Python-based data analysis platform, a risk assessment algorithm library, and data visualization tools. The research selected key time points in different market cycles for stress tests, including the sharp decline at the beginning of the 2020 COVID-19 pandemic, the technology stock boom in the first quarter of 2021, and the market adjustment triggered by the Federal Reserve's interest rate hikes in 2022. Each round of testing involves 1,000 simulated transactions to assess the model's robustness and generalization ability in actual investment scenarios. To avoid overfitting, a five-fold cross-validation mechanism was introduced during the training process, and the Monte Carlo method was adopted to randomly sample and optimize the parameter space. The data segmentation, hyperparameter tuning range, and repeatable protocol are as follows: The data segmentation adopts the time series partitioning method, with the ratio of the training set to the test set being 8:2, ensuring the continuity of the time series. The range of hyperparameter tuning covers decision tree depth of 3 to 12, learning rate of 0.01 to 0.3, and feature sampling rate of 0.6 to 1.0. The experimental process followed the reproducibility protocol, with all random seeds fixed and the data preprocessing steps standardized. The study used AD-PB, CatBoost-GJO, TS-GA, and GAN-Stacking to predict four types of investment risks: macroeconomic conditions, industry development trends, corporate financial performance, and market price fluctuations. The prediction accuracy results are shown in Figure 10.

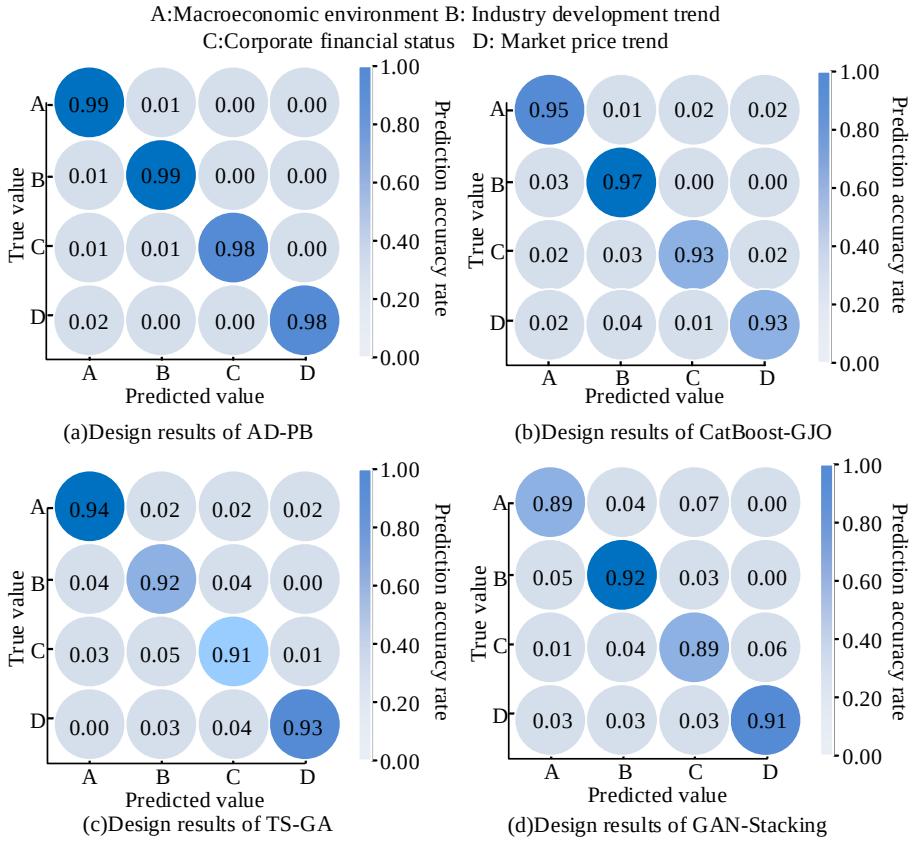


Figure 10: Comparison results of classification prediction accuracy

As shown in Figure 10 (a), the AD-PB model achieved a minimum prediction accuracy of 98.0% for corporate financial performance and market price fluctuations. Its maximum prediction accuracy reached 99.0% for macroeconomic conditions and industry development trends. According to Figure 10 (b), the CatBoost-GJO model reached a maximum prediction accuracy of 97.0%. The TS-GA and GAN-Stacking

models performed less accurately, with maximum prediction accuracies of 94.0% and 92.0%, respectively. Overall, the AD-PB model demonstrated the highest prediction accuracy among all models. To further evaluate the risk assessment capabilities of the four models, the study compared their Sharpe ratios and maximum drawdowns. The results are shown in Figure 11.

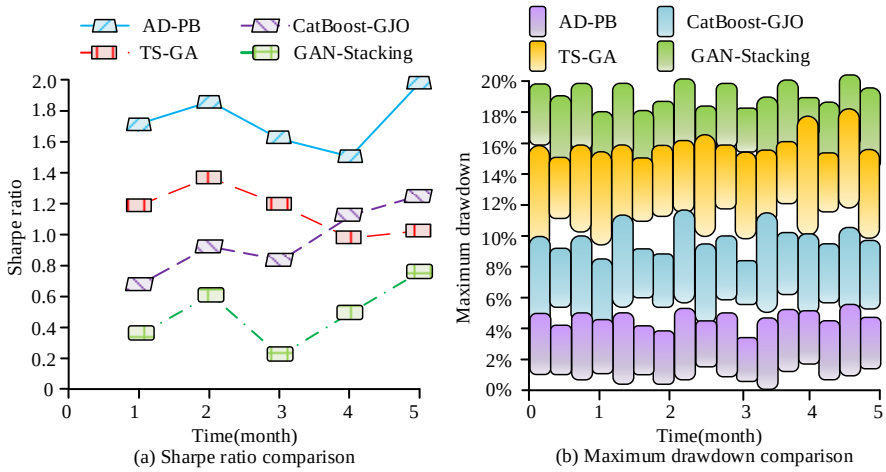


Figure 11: Comparison results of Sharpe ratio and maximum drawdown

As shown in Figure 11 (a), the AD-PB model consistently achieved the highest Sharpe ratio, with a peak value of 1.9. Among the comparison models, TS-GA performed relatively well, with a maximum Sharpe ratio of 1.4. According to Figure 11 (b), the AD-PB model had the smallest maximum drawdown overall, with a maximum value of 5.9%. In contrast, the GAN-Stacking model exhibited the largest drawdown, reaching 19.1%. In conclusion, the AD-PB model demonstrated strong risk resilience. Subsequently, in order to comprehensively verify the risk prediction performance of the four models, the study compared volatility, value at risk, win rate and profit-to-loss ratio. The profit-loss ratio is calculated as the absolute value of the average profit divided by the average loss, i.e., profit-loss ratio = average profit / average loss. In addition, the study introduced three indicators, namely Precision, Recall and Specificity, to further evaluate the predictive efficacy of the model. The value range of Specificity is between 0 and 1. The larger the value, the stronger the model's ability to identify non-risk events. The results are shown in Table 3.

Table 3: Results of volatility, value at risk, win rate, and profit-loss ratio

Model	Volatility (%)	Value at risk (%)	Win rate (%)	Profit-loss ratio	Precision	Recall	Specificity
AD-PB	8.2±1.5@#*	7.3±1.2@#*	88.9±8.4@#*	6:1	0.96±0.04@#*	0.94±0.03@#*	0.93±0.05@#*
CatBoost-GJO	15.7±2.4	8.9±2.0	79.6±7.6	9:2	0.91±0.05	0.90±0.04	0.90±0.06
TS-GA	17.6±2.8	16.8±2.5	71.4±6.9	7:2	0.89±0.04	0.89±0.06	0.88±0.05
GAN-Stacking	19.8±3.0	20.4±3.7	65.8±6.5	3:1	0.86±0.06	0.85±0.05	0.86±0.07

*Note: In Table 3, @ indicates that the value of the AD-PB index is significantly different from that of CatBoost-GJO, $p<0.05$; # indicates that the difference between the AD-PB index value and TS-GA is significant, $p<0.05$; * Indicates that the difference in the AD-PB index value compared with GAN-Stacking is significant, $p<0.05$*

Table 3 shows that the AD-PB model achieves the highest levels of Precision, Recall, and Specificity—0.96, 0.94, and 0.93, respectively—indicating superior accuracy and stability in identifying risk events. In contrast, the three indicators of other models were all lower than those of the AD-PB model. Especially, the GAN-Stacking model had the lowest values, which were 0.86, 0.85 and 0.86 respectively, indicating that it was relatively weak in identifying risk events. The AD-PB model achieved a volatility of 8.2%, a value at risk of 7.3%, a win rate of 88.9%, and a profit-loss ratio of 6:1. The CatBoost-GJO model yielded a volatility of 15.7%, a value at risk of 8.9%, a win rate of 79.6%, and a profit-loss ratio of 9:2. Among the TS-GA and GAN-Stacking models, TS-GA demonstrated relatively better overall performance, with a volatility of 17.6%, a value at risk of 16.8%, a win rate of 71.4%, and a profit-loss ratio of 7:2. In summary, the AD-PB model performed well in risk control, profitability, and return efficiency.

The features selected based on RF and SVM components include trading frequency (Feature 1), holding time (Feature 2), stop-loss range (Feature 3),

profit target (Feature 4), capital utilization rate (Feature 5), account balance volatility (Feature 6), order interval time (Feature 7), maximum drawdown range (Feature 8), average position ratio (Feature 9), and risk preference coefficient (Feature 10). The importance weights of each feature are shown in Figure 12.

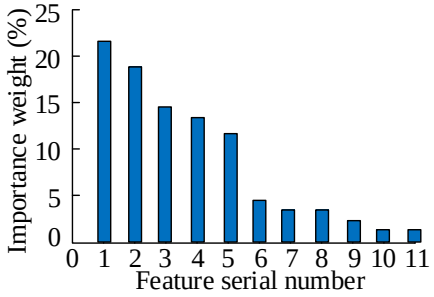


Figure 12: The importance weights of each feature

As can be seen from Figure 12, the feature weight distribution screened out by the RF and SVM models is relatively concentrated, among which the cumulative weight of the top five features exceeds 60%. The weight proportions of Feature 1 and Feature 2 are the highest, reaching 22% and 19% respectively, indicating that the strategy is highly sensitive to market response speed and holding period. The weights of Feature 3 and Feature 4 are 14% and 13% respectively, indicating that risk control and profit setting have a relatively high priority in model decision-making. Feature 5 accounts for 12%, reflecting the trading strategy's emphasis on capital efficiency. The proportion of other features does not exceed 10%.

To test the complexity of the proposed method in the research and explore its sensitivity to the severity of data imbalance and parameter changes, an experiment was designed to analyze its robustness and time complexity. Specifically, the study tests the model's performance under different data distributions by adjusting the ratio of positive to negative samples in the training set. The data distribution includes multiple gradient settings, with the proportion of positive samples ranging from 10% to 90%. The comparison indicators include accuracy rate, F1 score, AUC value and model training time. The results are shown in Table 4.

Table 4: The performance of the model under different data distributions

Proportion of positive samples	Accuracy rate (%)	F1	AUC value	Model training time (s)
10%	92.7	0.90	0.93	41.20
20%	92.9	0.91	0.94	38.77
30%	93.0	0.92	0.95	39.68
40%	93.5	0.92	0.96	37.94
50%	94.3	0.93	0.97	37.48
60%	94.0	0.92	0.96	37.99
70%	93.8	0.91	0.96	38.56
80%	93.2	0.90	0.95	39.77
90%	92.7	0.90	0.94	40.54

As shown in Table 4, as the proportion of positive samples gradually increases from 10% to 90%, the overall accuracy of the model shows a trend of first rising and then falling, reaching a peak of 94.3% when the proportion of positive samples is 50%. The F1 score and AUC value also exhibited a similar trend, reaching optimal values of 0.93 and 0.97, respectively, when the proportion was 50%, indicating that the model achieved the highest recognition ability for positive and negative samples under this condition. The training time remained basically stable, averaging approximately 39.10 s, demonstrating that the algorithm has good adaptability and stability to changes in data distribution.

5 Conclusion

To address the issues of existing investment risk prediction methods, such as reliance on historical data leading to poor adaptability, insufficient capture of complex market dynamics, and low prediction accuracy, this study innovatively proposed an investment risk prediction model based on the ADASYN-RF algorithm. The model balances the data using ADASYN, simplifies features through PCA, captures temporal patterns with ARIMA, and uses RF and CS-SVM for classification and prediction, enabling precise and efficient investment risk forecasting. The experimental results show that the AD-PB prediction model based on the ADASYN-RF algorithm achieved a maximum prediction accuracy of 99.0% for different risk types, with the lowest accuracy reaching 98.0%. Additionally, its Sharpe ratio and maximum drawdown were 1.9 and 5.9%, respectively. In terms of risk assessment performance, the AD-PB model outperformed the comparison models with a volatility of 8.2%, a value at risk of 7.3%, a win rate of 88.9%, and a profit-loss ratio of 6:1. Overall, the ADASYN-RF-based investment risk prediction model demonstrates strong risk resistance and prediction capabilities, meeting the needs of enterprises for investment risk forecasting. Although the proposed model performed excellently in terms of both performance and application, it has not yet addressed policy risks and irrational user behavior during the prediction process. Therefore, the accuracy and universality of the predictions still need to be improved. In the future, further efforts will be made to separate these influencing factors and continuously enhance the model's prediction accuracy.

6 Discussion

Compared with CatBoost-GJO, TS-GA, and GAN-Stacking, the prediction accuracy of AD-PB increased by 1.2%, 3.5%, and 4.7%, respectively, demonstrating stronger stability and generalization ability. Furthermore, when dealing with a highly volatile market environment, the volatility of AD-PB is only 8.2%, significantly lower than the 15.7%, 17.6% and higher levels of other models, indicating that it has a stronger risk control ability in extreme market conditions. Meanwhile, the win rate of AD-PB is 88.9%, and the profit-to-loss ratio is 6:1, demonstrating high profitability and stability. These

characteristics enable AD-PB to more effectively identify and avoid risks when facing a complex and volatile market environment, thereby enhancing the reliability and return performance of investment decisions. This is because the AD-PB model combines the oversampling advantage of ADASYN with the nonlinear classification ability of RF. Meanwhile, it effectively reduces feature redundancy through PCA, and ARIMA enhances the capture of time series trends. This multi-stage collaborative mechanism not only enhances the model's ability to identify complex risk patterns but also improves the robustness and adaptability of predictions, increasing its applicability in diverse market environments. In addition, AD-PB performs well in terms of memory efficiency, F1 score, Sharpe ratio, and drawdown reduction. This is closely related to its architectural design. The ADASYN component effectively alleviates the problem of data imbalance, enabling the model to still have reliable predictive capabilities in small sample risk categories. Meanwhile, the random feature selection mechanism of RF further optimizes the generalization performance of the model. Meanwhile, the introduction of PCA not only enhances computational efficiency but also provides a guarantee for the compact expression of the feature space. ARIMA, on the other hand, effectively compensates for the dynamic changes in market trends. These designs enable AD-PB to perform exceptionally well in multi-dimensional evaluations and possess stronger practicality and promotion potential.

7 Model replication

To facilitate the reproduction and application of the AD-PB model by other researchers or practitioners, this section will detail the construction steps and key parameter Settings of the model. Firstly, in the data preprocessing stage, the ADASYN algorithm is used to oversample the imbalanced samples in the training set, enhancing the model's recognition ability for minority risk samples. Subsequently, the features are reduced in dimensionality through PCA, retaining principal components with a cumulative contribution exceeding 90%, thereby reducing feature redundancy and improving computational efficiency. In the model training section, the number of decision trees of the Random Forest (RF) classifier is set to 200, with a maximum depth of 10, and the Gini index is adopted as the splitting criterion. For the ARIMA component, the differential method is used to perform stationarity processing on the target sequence, and the optimal (p, d, q) parameter combination is determined based on the AIC criterion. Finally, the prediction results of RF are used as input, and the trend prediction output of ARIMA is fused via linear weighting to form the final AD-PB prediction model.

References

[1] Qin G, Juan M, Rui M H. Iot-based intelligent power supply management using ensemble learning for seismic observation stations. Informatica, 2025, 49(8). 1-18. DOI: 10.31449/inf.v49i8.6502

[2] Mahmood S A, Hamadi S S. Ensemble machine learning algorithms for predicting thyroid disorders in diabetic patients: a comparative analysis. Informatica, 2025, 49(24). 7-25. DOI: 10.31449/inf.v49i24.8373

[3] Ileberi E, Sun Y. Advancing model performance With ADASYN and recurrent feature elimination and cross-validation in machine learning-assisted credit card fraud detection: A comparative analysis. IEEE access, 2024. 12(1): 133315-133327. DOI: 10.1109/ACCESS.2024.3457922

[4] Aish M A. Predictive modeling of cerebral strokes: an ADASYN-RF approach for imbalanced data. VFAST Transactions on Software Engineering, 2024, 12(4): 12-26. DOI: 10.21015/vtse.v12i4.1932

[5] Zhang M. Enhanced load forecasting for electric vehicle charging stations using a hybrid random forest-convolutional neural network algorithm. Informatica, 2024, 48(19): 89-102. DOI: 10.31449/inf.v48i19.6649

[6] Saiyed A. AI-Driven Innovations in Fintech: Applications, challenges, and future trends. International Journal of Electrical and Computer Engineering Research, 2025, 5(1): 8-15. DOI: 10.53375/ijecer.2025.437

[7] Song M, Ma H, Zhu Y, Zhang M. Credit risk prediction based on improved ADASYN sampling and optimized LightGBM. Journal of Social Computing, 2024, 5(3): 232-241. DOI: 10.23919/JSC.2024.0019.

[8] Salehpour A, Norouzi M, Balafar M A, SamadZamini K. A cloud-based hybrid intrusion detection framework using XGBoost and ADASYN-Augmented random forest for IoMT. IET Communications, 2024, 18(19): 1371-1390. DOI: 10.1049/cmu2.12833

[9] Dawei Y, Bing Z, Bingbing G, Xibo G, Razzaghzadeh B. Predicting the CPT-based pile set-up parameters using HHO-RF and PSO-RF hybrid models. Structural Engineering and Mechanics, 2023, 86(5): 673-686. DOI: 10.12989/sem.2023.86.5.673

[10] Ding W, Sun H. Prediction of PM2. 5 concentrations based on the weighted RF-LSTM model. Earth Science Informatics, 2023, 16(4): 3023-3037. DOI: 10.1007/s12145-023-01111-7

[11] Jiao Z. Dynamic Financial Distress Prediction Using Combined LASSO and GBDT Algorithms. Informatica, 2024, 48(17): 139-152. DOI: 10.31449/inf.v48il7.6493

[12] Belhadi A, Kamble S S, Mani V, Benkhati I, Touriki F E. An ensemble machine learning approach for forecasting credit risk of agricultural SMEs' investments in agriculture 4.0 through supply chain finance. Annals of Operations Research, 2025, 345(2): 779-807. DOI: 10.1007/s10479-021-04366-9

[13] Li X, Wang J, Yang C. Risk prediction in financial management of listed companies based on optimized BP neural network under digital economy. Neural Computing and Applications, 2023, 35(3): 2045-

2058. DOI: 10.1007/s00521-022-07377-0

[14] Doğru A, Buyrukoğlu S, Arı M. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Medical & Biological Engineering & Computing, 2023, 61(3): 785-797. DOI: 10.1007/s11517-022-02749-z

[15] Zheng G, Kong L, Brintrup A. Federated machine learning for privacy preserving, collective supply chain risk prediction. International Journal of Production Research, 2023, 61(23): 8115-8132. DOI: 10.1080/00207543.2022.2164628

[16] Rhodes J S, Cutler A, Moon K R. Geometry-and accuracy-preserving random forest proximities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10947-10959. DOI: 10.1109/TPAMI.2023.3263774

[17] Kurani A, Doshi P, Vakharia A, Shah M. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, 2023, 10(1): 183-208. DOI: 10.1007/s40745-021-00344-x

[18] Cheng J, Xiong Y. Multi-strategy adaptive cuckoo search algorithm for numerical optimization. Artificial Intelligence Review, 2023, 56(3): 2031-2055. DOI: 10.1007/s10462-022-10222-4

[19] Fei R, Guo Y, Li J, Hu B, Yang L. An improved BPNN method based on probability density for indoor location. IEICE TRANSACTIONS on Information and Systems, 2023, 106(5): 773-785. DOI: 10.1587/transinf.2022DLP0073

[20] Marukatat S. Tutorial on PCA and approximate PCA and approximate kernel PCA. Artificial Intelligence Review, 2023, 56(6): 5445-5477. DOI: 10.1007/s10462-022-10297-z