Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Research on Stock Index Prediction Based on SARIMA-BP-KPCA+RF Model

Fengying Yan *

Brook Institution, Anhui University, Hefei, Anhui, China

* Corresponding Author Email: R22214019@stu.ahu.edu.cn

Abstract. The prediction of stock prices plays a vital role in guiding investors' decision - making processes and managing risks within the financial market. The stock price of ADANIPORTS is influenced by multidimensional data and exhibits nonlinear dynamic characteristics. This study is based on Yahoo Finance data, calculating 8 technical indicators including KDJ and RSI. After feature screening, seasonal autoregressive integral moving average model (SARIMA), backpropagation neural network model (BP neural network), and kernel principal component analysis random forest model (KPCA+RF) are constructed and their performance is compared. The experiment showed that the MAE of SARIMA was 8.4983 and the R^2 was 0.9668; The MAE of the BP neural network is 13.7540 and the R^2 is 0.9797; The MAE of KPCA+RF is 10.3501 and the R^2 is 0.9832, indicating the best prediction accuracy. The study has verified the effectiveness of ensemble learning and nonlinear dimensionality reduction, providing a multi model comparative analysis framework for stock prediction and having methodological reference value.

Keywords: Stock price prediction; SARIMA; BP neural network; KPCA+RF.

1. Introduction

In the realm of finance, the prediction of stock prices holds immense significance as it has a profound impact on investors' decision - making and risk management strategies [1, 2]. The price of ADANIPORTS stock is non - linear and dynamic, affected by various factors like opening price, trading volume, and technical indicators [3, 4]. Traditional single models, such as SARIMA [5], can capture linear trends and seasonal patterns, but face challenges in dealing with complex non - linear relationships. Machine learning models like the BP Neural Network have non - linear fitting capabilities but are prone to overfitting due to high - dimensional data redundancy [6, 7]. Ensemble learning models like Random Forest can enhance robustness but lack in - depth exploration of data structure.

This article focuses on comparing the performance of SARIMA, BP neural network, and KPCA + RF models [8] in predicting ADANIPORTS stock closing prices. It aims to verify the effectiveness of ensemble learning and non - linear dimensionality reduction in high - dimensional financial data. This paper obtained data from Yahoo Finance from 2022 - 2024, calculated 8 technical indicators including KDJ and RSI. Through correlation analysis and variance threshold method, 15 key features were selected and normalized. The SARIMA model determined stationarity via ADF testing and used SARIMAX (1,1,1) × (1,1,1,7) to capture weekly trends. The BP neural network with a 15 - 64 - 32 - 1 structure used ReLU and Adam optimizer for non - linear fitting. The KPCA + RF model mapped data to an 8 - dimensional space with an RBF kernel and used a 300 - tree random forest (max depth 15) for prediction.

The paper is organized into five chapters. The introduction presents the research background and emphasizes model comparison. Chapter 2 details data sources, feature calculation, preprocessing, and model principles. Chapter 3 shows model training and prediction results. Chapter 4 compares model performances, especially KPCA + RF's advantages in high - dimensional data. Chapter 5 concludes the model combination's effectiveness and proposes future directions like using dynamic weights or reinforcement learning for optimization.

2. Dataset and Methodology

2.1. Data Sources

The stock data in this study is sourced from Yahoo Finance. Yahoo Finance, as an important platform in the field of financial data, provides rich and comprehensive financial market data. This study focuses on ADANIPORTS stocks and obtained basic data including opening price, high price, low price, last price, close price, volume, turnover, etc. from the platform. Meanwhile, based on Python technology, various technical analysis indicators such as KDJ, Relative Strength Index (RSI), Moving Average (MA5 and MA10), MOM, MACD, BIAS, ROC, and investor sentiment indicators are calculated using relevant formulas. Table 1 shows the selection of indicators and factor explanations

Table 1. Indicator Selection and Factor Explanation

Indicator Name (Full)	Formula (Simplified)	Explanation
Stochastic Indicator (KDJ)	$K = \frac{100\times(Close-Min)}{Max-Min}$, $D = MA(K)$, $J = 3D - 2K$	Measures trend strength and trading signals. K and D reflect price momentum.
Relative Strength Index (RSI)	$RSI = 100 - \frac{100}{1+RS}$, $RS = \frac{AvgGain}{AvgLoss}$	Indicates overbought (high) or oversold (low) conditions (0–100 scale).
Moving Average (MA5/MA10)	$MA = \frac{\sum_{i=1}^{n} Close_i}{n}$	Smooths price trends; short-term (MA5) and long-term (MA10) averages.
Momentum Indicator (MOM)	$MOM = Close - Close_n$	Tracks price change speed/direction; positive/negative values indicate momentum.
Moving Average Convergence Divergence (MACD)	$DIF = EMA_{12} - EMA_{26}$, $DEA = MA(DIF)$, $MACD = 2 \times (DIF - DEA)$	Analyzes trend and momentum via EMA crossovers.
Bias Ratio (BIAS)	$BIAS = \frac{Close - MA(N)}{MA(N)} \times 100\%$	Measures deviation from MA; extreme values signal reversals.
Rate of Change (ROC)	$ROC = \frac{Close - Close_n}{Close_n} \times 100\%$	Calculates percentage price change over n periods.
Sentiment Indicator (AR/BR)	$AR = \frac{\sum(High-Open)}{\sum(Open-Low)} \times 100$, $BR = \frac{\sum(High - PrevClose)}{\sum(PrevClose - Low)} \times 100$	AR reflects market sentiment; BR gauges trading willingness.

2.2. Data Preprocessing

Calculate the correlation matrix between the selected indicators and the closing price to analyze their linear correlation. Calculate the variance of each feature and use Variance Threshold (with a threshold set to 0.01) to remove low - variance features and reduce data redundancy. The results of the correlation analysis are shown in Figure 1.

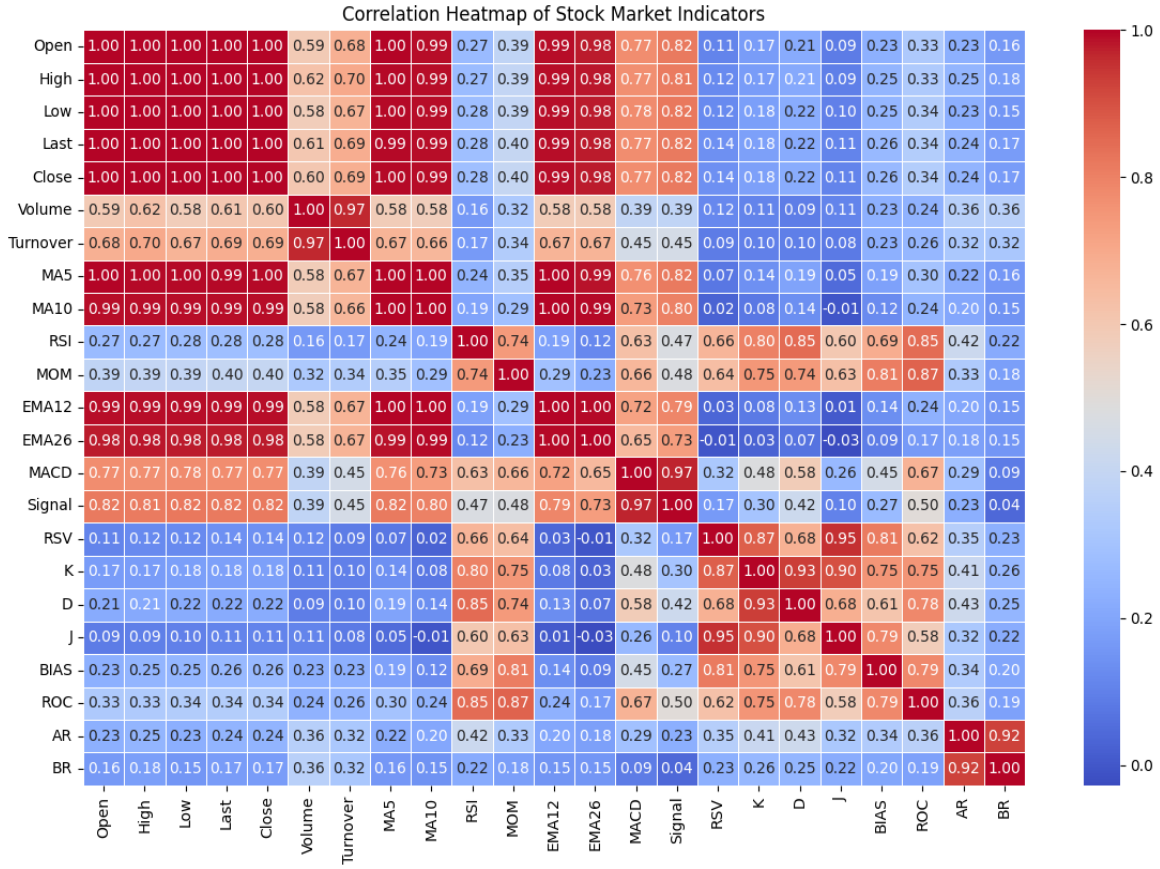

Figure 1. Correlation Heatmap of Stock Market Indicators (Picture credit: Original)

Therefore, the input variables of the model are selected as Close, Open, Volume, RSI, MA5, MOM, AR, BR, MACD, Turnover, EMA12, EMA26, ROC, and BIAS. Before using the data, a series of preprocessing operations are performed on it, including missing value processing, Box - Cox transformation, and normalization.

2.3. Methods

2.3.1. SARIMA.

The SARIMA model [9] is developed based on the ARIMA model. It incorporates seasonal autoregressive (SAR), differencing (SD), and moving average (SMA) components to deal with time - series data presenting seasonal features. The expression of the SARIMA $(p, d, q)(P, D, Q)_s$ model is:

$$\left(\Phi(B^s)\phi(B)(1-B)^d(1-B^s)^D y_t = c + \Theta(B^s)\theta(B)\epsilon_t\right) \tag{1}$$

$$\Phi(B^s) = 1 - \Phi_{1B^s} - \cdots - \Phi_{PB^{Ps}} \tag{2}$$

$$\Theta(B^s) = 1 + \Theta_{1B^s} + \cdots + \Theta_{QB^{Qs}} \tag{3}$$

Where (P, D, Q) are the orders of seasonal autoregression, seasonal differencing, and seasonal moving average respectively, and s is the seasonal period.

2.3.2. BP Neural Network Model.

Based on the error backpropagation algorithm, the BP neural network [10] is a multi - layer feed - forward neural network that is extensively applied in data prediction and various other fields. It constructs a multi-layer network structure consisting of an input layer, hidden layers, and an output layer. The network weights are iteratively optimized through the backpropagation algorithm to model

complex non-linear relationships. The number of neurons in the input layer is consistent with the dimension of the input features, and it is responsible for receiving the original data features. In this study, two hidden layers are constructed. The first hidden layer has 64 neurons, and the second has 32 neurons. The Rectified Linear Unit (ReLU) activation function is used to extract and transform the input data features, effectively solving the gradient vanishing problem and improving the network training efficiency. The number of neurons in the output layer is consistent with the dimension of the target variable, and it is used to output the prediction results.

The key technologies of the network include forward propagation, backpropagation, and weight update. During forward propagation, the input z_j of neuron j in the hidden and output layers is the weighted sum of the outputs of the neurons in the previous layer, expressed as

$$z_j = \sum_i w_{ij} x_i + b_j \tag{4}$$

After being processed by the activation function f, the output a_j of neuron j is obtained as

$$a_j = f(z_j) \tag{5}$$

Finally, the prediction value is generated by the output layer. The backpropagation algorithm calculates the error between the predicted and true values and propagates the error backward through the network to update the weights. First, the loss function is calculated. Then, based on the chain rule, the gradient of the loss function with respect to the weights of each layer, $\frac{\partial L}{\partial w}$, is calculated starting from the output layer to determine the direction and magnitude of the weight adjustment. In this study, the Adam optimizer is used for weight update. It adaptively adjusts the learning rate, accelerating convergence and avoiding getting stuck in local optima. Additionally, to prevent overfitting, the Dropout regularization technique is introduced. During training, hidden layer neurons are randomly dropped with a probability of 20%. The model training adopts a strategy of fixed epochs and batch size, and the model performance is monitored in real-time on the validation set.

2.3.3. KPCA + Random Forest.

In this paper, KPCA and RF are combined [11]. Through non - linear dimensionality reduction and integrated learning strategies, efficient feature extraction and accurate prediction of complex data are achieved. This model shows significant advantages in financial market prediction, pattern recognition, and other fields.

In the kernel space, KPCA represents a non - linear expansion of Principal Component Analysis (PCA). Its purpose is to address the shortcomings of traditional PCA when dealing with non - linear data. Its core idea is to map the original data to a high - dimensional feature space through a kernel function and perform linear PCA operations in this space, thus achieving non - linear data dimensionality reduction.

The Random Forest is an ensemble learning algorithm based on decision trees. By constructing multiple decision trees and aggregating their prediction results, it effectively reduces the model variance and improves the generalization ability. Its main structure is divided into two parts: decision tree construction and integrated prediction. In the decision tree construction process, the Random Forest uses the Bootstrap Sampling (Bagging) method to randomly sample multiple subsets from the original training set with replacement, and each subset is used to train an independent decision tree. During the node splitting process of the tree, a part of the features (feature subspace) is randomly selected to find the optimal splitting point, further enhancing the model diversity. In the integrated prediction process, for regression tasks, the Random Forest generates the final output by averaging the prediction results of all decision trees. Let the prediction values of T decision trees be y_1, y_2, ..., y_T, then the prediction value y of the Random Forest is:

$$y = \frac{1}{T}\sum_{t=1}^{T} y_t \quad (6)$$

The combination of KPCA and the Random Forest realizes a two - step optimization of "dimensionality reduction first, then prediction": KPCA explores the internal structure of the data through non - linear mapping and reduces the feature dimension; the Random Forest uses integrated learning to make robust predictions on the reduced - dimensional data. In addition, in this study, Grid Search is used in the code to optimize the parameters of the Random Forest, and the optimal model configuration is found by exhaustively searching parameter combinations.

3. Experiments

3.1. Experimental Results

The experimental results of this study are presented in Table 2. The SARIMA model exhibits a relatively low MAE, showing good performance in mean absolute deviation. However, its high RMSE implies inaccurate handling of data fluctuations. It assumes linear data changes and has a limited ability in dealing with complex non - linear relationships, being sensitive to outliers. The BP neural network model possesses a high R^2, demonstrating a favorable data - fitting effect and robust non - linear fitting capabilities. However, its relatively high MAE and MAPE suggest room for improvement in prediction deviation control, and it may suffer from overfitting during training, with long training times. The KPCA + RF model has relatively optimal values for MAE, RMSE, and MAPE, and the highest R^2, indicating the best overall prediction accuracy and data - fitting effect. It combines the dimensionality reduction advantage of KPCA with the non - linear fitting and anti - interference capabilities of the random forest, adapting well to data, performing well in high - dimensional data processing, and having relatively fast model training and prediction speeds.

Table 2. Comparison of Performance Metrics Among Different Models

Model Name	MAE	RMSE	MAPE	R^2
SARIMA	8.4983	21.1528	2.1543%	0.9668
BP Neural Network	13.7540	18.7350	3.51%	0.9797
KPCA + RF	10.3501	17.0328	2.44%	0.9832

3.2. Experimental Analysis

3.2.1. Prediction Analysis of the SARIMA Model.

In this study, the SARIMA model is utilized to predict the closing prices of ADANIPORTS stock. The data used ranges from February 5, 2022 to April 30, 2024. The modeling process began with an Augmented Dickey-Fuller (ADF) test, which identified non-stationarity in the original price series. After applying first-order differencing to achieve stationarity, ACF/PACF analysis combined with AIC optimization led to the selection of the SARIMAX (1,1,1) × (1,1,1,7) model, effectively capturing both trend and weekly seasonal patterns. Diagnostic tests revealed no significant residual autocorrelation, though some heteroskedasticity and non-normal distribution were present. The model demonstrated strong performance with an R^2 of 0.97, indicating excellent fit for long-term trends and seasonal fluctuations. However, the relatively high RMSE suggests limitations in predicting short-term price movements, highlighting the model's challenges with nonlinear dynamics. These findings suggest that while SARIMA performs well in trend-dominated markets, future research could enhance prediction accuracy by incorporating nonlinear modeling approaches. Figure 2 shows the comparison chart of SARIMA predicted values and actual values.

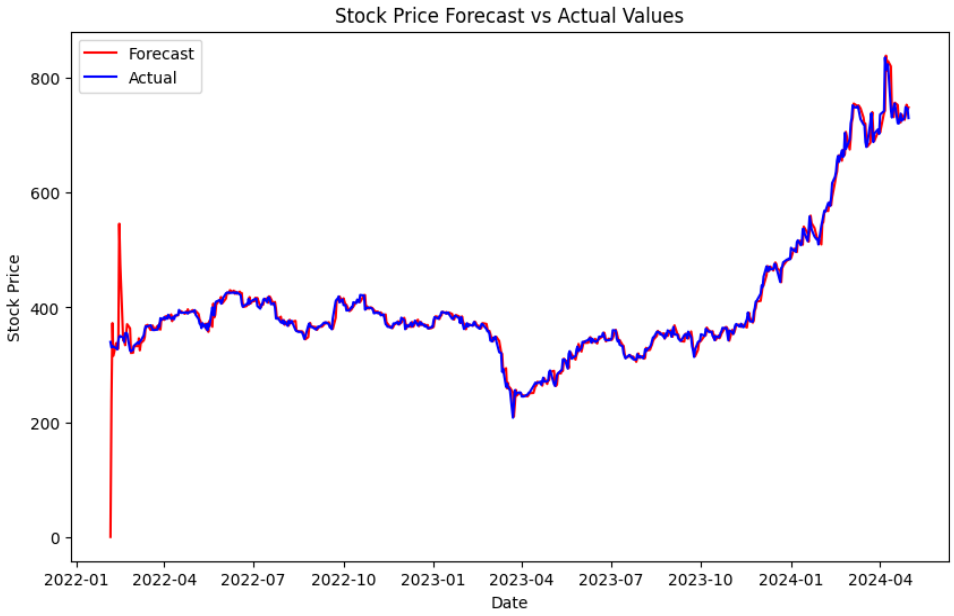

Figure 2. SARIMA Predicted Values and Actual Values (Picture credit: Original)

3.2.2. Training and Prediction of the BP Neural Network Model.

The study utilized 15 selected technical indicators - including closing price, trading volume, and RSI - as input features to predict closing prices. After preprocessing, the data was split into 80% training and 20% testing sets. A BP neural network architecture was implemented with 15 input nodes, two hidden layers (64 and 32 neurons respectively), and a single output node. The model was trained over 100 epochs with a batch size of 32, using the test set for validation. Training dynamics showed rapid initial loss reduction as key patterns were learned, followed by gradual convergence. While the decelerating loss reduction suggested approaching optimal performance, it also indicated potential overfitting risks that warrant monitoring in future applications. This architecture effectively captured the complex relationships in financial data while maintaining computational efficiency. Figure 3 shows the neural network model training.

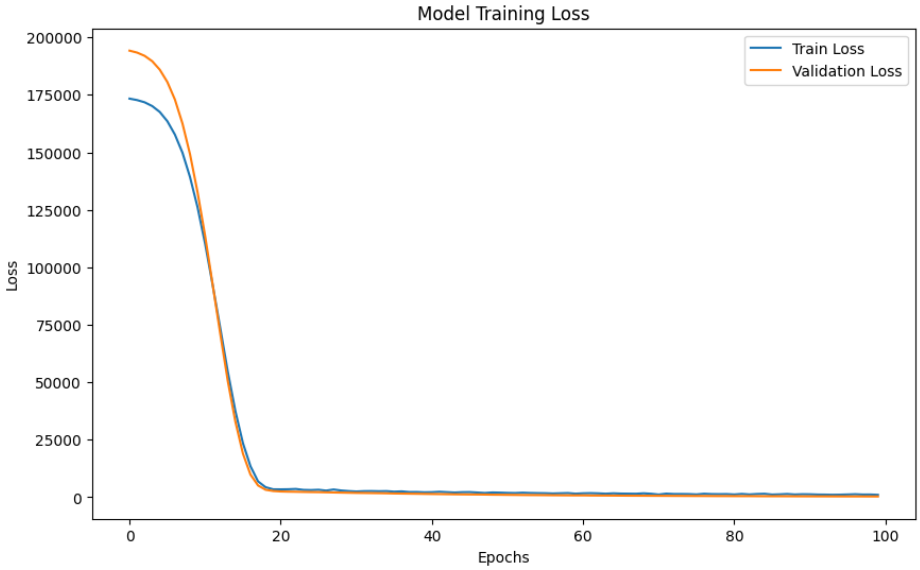

Figure 3. Neural Network Model Training (Picture credit: Original)

The prediction results show close alignment between predicted and actual values in terms of overall trends, though minor discrepancies exist due to complex market influences like macroeconomic conditions, company fundamentals, and industry competition. While no model can achieve perfect accuracy, this one effectively captures key price trends, offering investors valuable insights. Figure 4 shows the prediction results of the BP neural network model.

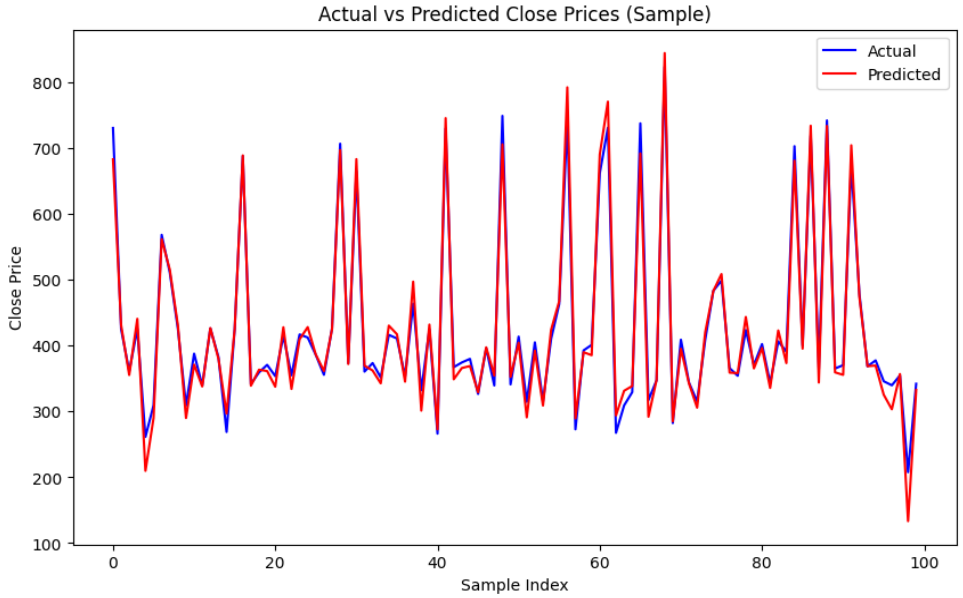

Figure 4. Prediction Results of the BP Neural Network Model (Picture credit: Original)

3.2.3. Training and Prediction of the KPCA + RF Model.

In this study, the same dataset and target variable (Close price) as those in the BP neural network model are used. KPCA is applied for dimensionality reduction, with 8 principal components obtained using an RBF kernel (gamma = 0.01). This preprocessing effectively reduces data dimensionality while preserving key features, minimizing redundancy and noise. The processed data was split into 80% training and 20% testing sets. The random forest model was configured with 300 decision trees (max depth=15, min samples split=5, min leaf samples=3) and parallel computing for efficient training. Results demonstrate strong alignment between predicted and actual price trends, though minor deviations persist due to inherent market complexities like macroeconomic fluctuations. Despite these challenges, the KPCA-RF combination reliably captures overall price movements, offering valuable predictive insights for investors. Figure 5 shows the prediction results of the KPCA + RF model.

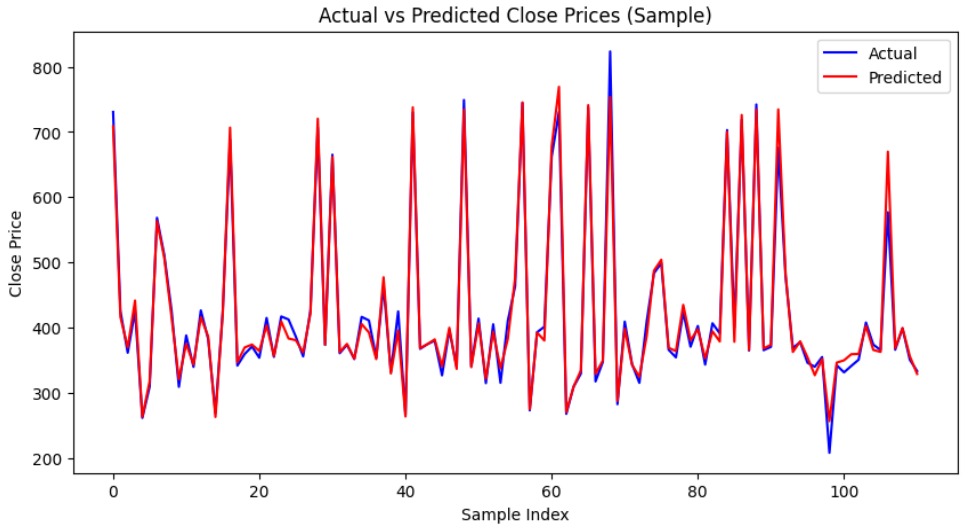

Figure 5. Prediction Results of the KPCA + RF Model (Picture credit: Original)

4. Conclusion

In this research, SARIMA, BP neural network, and KPCA + RF models are established to predict the closing prices of ADANIPORTS stock. Moreover, the performances of traditional time - series, machine learning, and integrated learning models in high - dimensional financial data are compared

and analyzed. The experiments show that the SARIMA model accurately describes linear trends and weekly seasonality (MAE = 8.4983, R^2 = 0.9668), but has insufficient non - linear fitting (RMSE = 21.1528); the BP neural network has a strong non - linear mapping ability (MAE = 13.7540, R^2 = 0.9797), but is affected by high - dimensional feature redundancy (MAPE = 3.51%); the KPCA + RF model reduces the dimension to 8 through kernel principal component analysis and combines the advantages of the random forest integration. It has the best error control and fitting effect (MAE = 10.3501, R^2 = 0.9832), verifying the effectiveness of non - linear dimensionality reduction and integrated learning.

Limitations exist in this study. For example, the selection of kernel functions depends on experience, and external variables are not incorporated. In the future, optimization algorithms can be introduced to dynamically adjust parameters, and macro - economic and public opinion data can be integrated to construct hybrid models. This study provides a cross - model comparative analysis framework for stock prediction, has practical value for financial institutions' strategy optimization and investors' risk management, and expands the methodology of high - dimensional financial data modeling.

References

[1] Wang R. Research on stock price prediction based on similarity measurement and dual-channel multi-scale hybrid neural network. Shandong University of Technology, 2024.

[2] Guan M. Pre-investment enterprise financial crisis identification scheme based on KPCA-XGBoost. Shanghai Normal University, 2021.

[3] Zhang Y, Li L. RF-MIP-LSTM stock price prediction model. Comput Eng Appl, 2024, 60(17): 272 – 281.

[4] Fu W. Research on stock price prediction based on ARIMA-RF combination model. Sci Technol Innov, 2023, (08): 40 – 43.

[5] Yang B. SARIMA stock price index prediction modeling based on neural network. Entrepreneur World, 2010, (04): 109 – 110.

[6] Zhao H. Research on stock price prediction method based on ARIMA-BP neural network and sentiment analysis. Shenyang Industry University, 2021.

[7] Ma G. Research on stock index prediction based on ARIMA-LSTM-BP combination model. Kashgar University, 2024.

[8] Li J. Stock index futures price prediction and quantitative strategy construction based on KPCA-gcForest. Northwest University, 2021.

[9] Wu C, Xu F, Duan M, Zhang L. Comparative study on the prediction of tuberculosis incidence in China based on SARIMA model and LSTM neural network model. China Port Sci Technol, 2025, 7(03): 4 – 12.

[10] Ren D, Xing B, Tian Y, Liu W. A PM2.5 prediction model based on BP neural network. China Sci Technol Inf, 2025, (08): 70 – 72.

[11] Xu X, Pan T. Research on wind farm power prediction method based on KPCA-RF. Renew Energy, 2018, 36(09): 1323 – 1327.