Proceedings of ICEMGD 2025 Symposium: The 4th International Conference on Applied Economics and Policy Studies
DOI: 10.54254/2754-1169/2025.BJ24912

Machine Learning-Driven Prediction of Overnight Price Disparities in Dual-Listed Companies: A Cross-Market Analysis of Hong Kong and US Equities

Jiaxu Chen[1], Yifan Yin[2], Xulin Zhu[3]*

[1]*Antai College Economics Management, Shanghai Jiaotong University, Shanghai, China*
[2]*School of Mathematics, Shanghai University of Finance and Economics, Shanghai, China*
[3]*JinHe Center for Economic Research, Xi'an Jiaotong University, Xi'an, China*
**Corresponding Author. Email: zhuxulin@stu.xjtu.edu.cn*

Abstract. Within the context of globally integrated financial markets, this study addresses the limitations of conventional econometric models in capturing non-linear dynamics and the scarcity of research on opening price prediction within the realm of cross-listed stocks. We propose an MLP-based predictive framework that integrates information from both Hong Kong and U.S. dual-listed markets. Utilizing standardized trading data from nine dual-listed companies prior to 2022, encompassing fundamental data such as opening and closing prices, alongside 29 derived indicators (e.g., logarithmic returns, RSI, 5-day volatility), we employ overnight spread data to forecast the subsequent day's opening price. We construct and comparatively analyze benchmark, single-market, dual-market, and weighted-market models. Empirical results demonstrate that the dual-market model outperforms the benchmark model, exhibiting a 48% reduction in mean squared error (MSE) and a 30%-50% increase in R2. The model's efficacy is particularly pronounced in companies characterized by price stability and extended listing periods, surpassing the performance of single-market models. This underscores the unique advantage of cross-market information integration in capturing the overnight spread transmission mechanism.

Keywords: Dual-listed companies, Cross-market analysis, MLP, Opening price prediction, Overnight spread

1. Introduction

In the context of global financial market integration, the cross-market price linkage mechanisms of dual-listed companies have garnered significant academic interest. Prior literature, such as Li, Yi and Su [1], has demonstrated a notable bidirectional information spillover effect between the Hong Kong and U.S. stock markets, indicating complex interactive relationships in the price discovery process across these markets. However, existing research predominantly relies on traditional econometric models, including VAR and cointegration analysis, for statistical testing of historical price linkages. The application of machine learning in this domain remains limited. Consequently, this paper aims

DOI: 10.54254/2754-1169/2025.BJ24912

to employ machine learning techniques for stock price prediction in dual-listed companies. Furthermore, while numerous studies focus on predicting closing prices and stock price trends, there is a lack of systematic investigation into the critical issue of predicting opening prices.

The overnight information accumulation effect, as it pertains to trading, influences opening prices via cross-market spreads [2], yet the transmission mechanisms remain inadequately elucidated in existing literature. Cross-market arbitrage theory posits that price discrepancies arising from non-overlapping trading sessions encapsulate significant market information [3], but traditional linear models struggle to capture the inherent non-linear transmission characteristics. This limitation is currently being addressed through machine learning methodologies. Preliminary evidence, accumulated over the past several years, indicates that machine learning techniques can identify non-linear structures within financial market data [4], with ensemble algorithms demonstrating significant advantages in high-dimensional data processing. Machine learning methods present unique advantages in financial forecasting, offering novel perspectives on capturing complex cross-market dynamics.

Consequently, drawing upon existing literature, this study endeavors to construct a machine learning-based framework for forecasting overnight spread, integrating overnight price differentials of dual-listed companies in Hong Kong and the United States. The objective is to predict the opening price trend for the subsequent trading day, thereby providing theoretical support and empirical evidence for investors to optimize cross-market trading strategies.

This research analyzes nine companies dual-listed in Hong Kong and the U.S. markets prior to 2022. Building upon and refining the novel engineering and derived indices for stock price prediction proposed by Abolmakarem et al., 2024 [5], a Multi-Layer Perceptron (MLP) model is employed to forecast the opening price for the next trading day. Predictions are conducted separately for each company across the two markets, encompassing Benchmark, single-market, dual-market, and weighted-market scenarios. Comparative analysis reveals that leveraging dual-market information significantly enhances predictive accuracy within dual-listed enterprises.

2. Literature review

Given the proliferation of cross-listed firms in international markets, it is crucial to examine the locus of information incorporation into prices [6]. Alaganar and Bhar [7] investigated Australian-American dual-listed stocks, revealing unidirectional information flow from the U.S. to the Australian market. Eun and Sabherwal [6] found that for 62 Canadian multinational corporations listed on U.S. exchanges, prices exhibited co-integration and mutual adjustment between their home and host markets. Considering the increasing number of Chinese companies concurrently listed on both U.S. and Hong Kong exchanges, research [1] indicates significant bidirectional information spillover effects between the Hong Kong and U.S. markets.

In the domain of machine learning applications within financial markets, There have been many attempts to use machine learning techniques to model nonlinear relationships in financial time series. Artificial neural networks (ANN) and kernel support vector machines are widely used due to their ability in nonlinear mapping and generalization. Unlike econometric models, artificial neural networks do not have a strict model structure and a set of assumptions imposed. Ayyildiz and Iskenderoglu [8] conducted an assessment of various machine learning algorithms, including decision trees, random forests, k-nearest neighbors, Naive Bayes, logistic regression, support vector machines, and artificial neural networks, to ascertain their efficacy in forecasting the directional movements of stock market indices in developed economies. The findings indicated that artificial neural networks exhibited the highest average predictive performance, achieving accuracy rates

DOI: 10.54254/2754-1169/2025.BJ24912

exceeding 70% for significant indices such as the NYSE 100 and FTSE 100. Prior research by Ayyildiz [9] explored machine learning techniques for stock index movement prediction, emphasizing their potential to enhance predictive accuracy. Furthermore, Ayyildiz and Iskenderoglu [9] investigated the application of machine learning methodologies in forecasting stock index movements within developing countries, analyzing the effectiveness and applicability of different algorithms across diverse economic contexts, thereby broadening the scope of machine learning in stock market prediction. They also evaluated the performance of several machine learning models specifically for the BIST 100 index, which represents the top 100 companies listed on the Istanbul Stock Exchange.

Several studies have demonstrated the superiority of machine learning (ML) methods over classical approaches in financial time series forecasting [10]. Numerous researchers have posited that machine learning methods are better equipped to handle the complexities inherent in financial time series data, including non-linearity, high dimensionality, and noise. Baek & Kim [11] proposed their model significantly improves the prediction performance based on machine learning.

3. Indicators

Beyond the selection of effective models, the choice of appropriate indicators and inputs is also crucial. Most studies have utilized only raw price data (open, high, low, and close prices, along with trading volume) as inputs for predictive models [12]. Notably, Abolmakarem et al. [5] recognized the potential of engineered and derived indices. Their developed framework of engineered technical indicators demonstrated superior performance in single-market predictions. After understanding the characteristics of each indicator, we adopted some of their indicators and added three additional high-frequency market microstructure factors, resulting in a total of 29 factors for the MLP input layer.

The indicators are as follows:

- $r1 = \ln\left(\frac{\text{Close Price}_i}{\text{Close Price}_{i-1}}\right)$
- $r2 = \ln\left(\frac{\text{Close Price}_{i-1}}{\text{Close Price}_{i-2}}\right)$
- $r3 = \ln\left(\frac{\text{Close Price}_{i-2}}{\text{Close Price}_{i-3}}\right)$
- $r4 = \ln\left(\frac{\text{Close Price}_{i-3}}{\text{Close Price}_{i-4}}\right)$
- $r5 = \ln\left(\frac{\text{High Price}_i}{\text{Open Price}_i}\right)$
- $r6 = \ln\left(\frac{\text{High Price}_i}{\text{Open Price}_{i-1}}\right)$
- $r7 = \ln\left(\frac{\text{High Price}_i}{\text{Open Price}_{i-2}}\right)$
- $r8 = \ln\left(\frac{\text{High Price}_i}{\text{Open Price}_{i-3}}\right)$

DOI: 10.54254/2754-1169/2025.BJ24912

- $r9 = \ln\left(\frac{\text{High Price}_{i-1}}{\text{Open Price}_{i-1}}\right)$
- $r10 = \ln\left(\frac{\text{High Price}_{i-2}}{\text{Open Price}_{i-2}}\right)$
- $r11 = \ln\left(\frac{\text{High Price}_{i-3}}{\text{Open Price}_{i-3}}\right)$
- $r12 = \ln\left(\frac{\text{Low Price}_{i}}{\text{Open Price}_{i}}\right)$
- $r13 = \ln\left(\frac{\text{Low Price}_{i-1}}{\text{Open Price}_{i-1}}\right)$
- $r14 = \ln\left(\frac{\text{Low Price}_{i-2}}{\text{Open Price}_{i-2}}\right)$
- $r15 = \ln\left(\frac{\text{Low Price}_{i-3}}{\text{Open Price}_{i-3}}\right)$
- $\text{RSI (Relative Strength Index)} = 100 - \frac{100}{1+\left(\frac{\sum_{i=0}^{n-1} Up_{t-i}/n}{\sum_{i=0}^{n-1} Dw_{t-i}/n}\right)}$
- $\text{Momentum} = C_t - C_{t-n}$
- $\text{TR (True Range)} = \max\{\max_t C_{t-1}\} - \min\{\min_t C_{t-1}\}$
- $\text{ATR (Average True Range)} = \frac{1}{n}\sum_{i=1}^{n} TR_i$
- $\text{CCI (Commodity Channel Index)} = \frac{M_t - SM_t}{0.015 D_t}$
- $\text{SMA (Simple Moving Average)} = \frac{C_t + C_{t-1} + \cdots + C_{t-(n-1)}}{n}$, here n=14
- $\text{WMA (Weighted Moving Average)} = \frac{(n \times C_t + (n-1) \times C_{t-1} + \ldots + C_n)}{(n + (n-1) + \ldots + 1)}$
- $HMA\left(Hull\ Moving\ Average\right) = WMA\left(2 \times WMA\left(n/2\right) - WMA\left(n\right)\right)\sqrt{n}\Big)$
- $\text{William's \% R} = \frac{H_n - C_t}{H_n - L_n} \times 100$
- $\text{Stochastic \%K} = \left(\frac{C_t - LL_{t-n}}{HH_{t-n} - LL_{t-n}}\right) \times 100$
- $\text{\% D-slow} = \frac{\sum_{i=0}^{n-1} \%D_{t-i}}{n}$
- $\text{Intraday Volatility} = \frac{P_{\text{high}} - P_{\text{low}}}{P_{\text{open}}}$

DOI: 10.54254/2754-1169/2025.BJ24912

- 5-Day Volatility$_t = \sqrt{\frac{1}{5-1}\sum_{i=t-4}^{t}\left(r_i - \bar{r}\right)^2}$
- 5-Day Cumulative Log Return$_t = \sum_{i=t-4}^{t} r_i$

C_t is the closing price at time t, L_t is the low price at time t, H_t is the High price at time t, LL_t and HH_t mean lowest low and highest high in the last t, days t, Up_t means the upward price change, Dw_t means the downward price change at time t.

3.1. Indicator description

r1 through r15 are all calculated using natural logarithms. Due to the properties of logarithms, the volatility of stock price changes can be stabilized, and the returns over different time periods can be easily aggregated. This simplifies the calculation of total returns and facilitates interpretation using percentage changes, which is beneficial for financial analysis. In finance, logarithmic returns are used to measure price changes over a period.

1. r1, r2, r3, and r4 utilize logarithmic returns to assess the relative fluctuations between consecutive trading days. Given the daily data frequency, the sequential logarithmic returns over four days provide a direct indication of recent price trends and short-term upward or downward momentum.

2. r5–r8 and r9–r11 are indicators based on the high and opening prices.

Indicator r5 reflects the maximum intraday price increase, offering insights into short-term investor sentiment and the strength of bullish forces during the trading session.

r6–r8 compare the day's high with the previous day's or earlier opening prices, potentially signaling a breakout of prior price levels and reflecting the market's adjustment to previous valuations. r9, r10, and r11 respectively represent the maximum price increases over the preceding three days.

3. r12-r15 are indicators based on low and opening prices.

The lowest decline of the day is reflected by the log ratio of the low price to the opening price. This contrasts with positive comparison indicators such as r5 that help capture risk or sell-pressure signals. Similarly, calculating the minimum decline for four consecutive days provides further confirmation of whether there is persistent downward pressure or signs of increased volatility in the market.

The subsequent eleven indicators are frequently employed technical analysis tools:

4. Relative Strength Index (RSI) reflects the relative relationship between the average rise and average fall in a certain period, which is used to judge whether a stock is overbought or oversold, so as to predict the possible price reversal point.

5. Momentum measures the difference between the current price and the price at a point in the past and is often used to capture accelerated price movements. It can indicate trend continuation and signal potential reversals.

6. True Range (TR) takes into account the price difference between the current and previous trading day (such as the difference between high and low and the relationship between the previous closing price and the current high and low price). This is an important indicator to evaluate market volatility, especially when there is a short jump or large fluctuation.

7. Average True Range (ATR) is a smoothing result of TR, which is used to reflect the average volatility over a period of time and is often used to set stop losses and manage trading risk.

DOI: 10.54254/2754-1169/2025.BJ24912

8. The Commodity Channel Index (CCI) measures the dispersion of current prices from their moving averages, thereby identifying situations of unusually high or low prices, providing signals of a potential reversal or correction.

9. Simple Moving Average (SMA) is often used to build trend following models by calculating the average price value over a period of time to help identify major price trends and eliminate the interference of daily price fluctuations.

10. Weighted Moving Average (WMA) is similar to SMA, but it assigns different weights to prices in different periods (usually with more emphasis on recent data), making this indicator more sensitive to the latest price changes and helping to capture trend shifts more quickly.

11. Hull Moving Average (HMA) is used to track the price trend with lower lag and smooth out some noise at the same time, which can reflect the trend changes earlier.

12. William's %R, similar to RSI, measures the position of the price relative to the highest and lowest prices within a given period. Usually, when it is near the critical values of -20 or -80, it is regarded as an overbought or oversold signal.

13. %K is the fast line in the Stochastic Oscillator, capturing short-term price changes; %D-slow is its smoothed processing version, which is used to filter out noise and confirm signals. When used in combination, the momentum changes of the market can be judged and potential reversal points can be predicted.

Following are three indicators based on price range and volatility analysis:

14. Intraday Volatility measures the proportion of the daily price range relative to the opening price, which directly reflects the extent of intraday price fluctuations and reveals the market activity and short-term uncertainty.

15. 5-Day Volatility is measured by the standard deviation of log return of the last five trading days to reflect the intensity and risk level of market price fluctuations in the short term.

16. 5-Day Cumulative Log Return can demonstrate the overall trend and cumulative rise and fall within a certain period, making it suitable for capturing the overall strength of short-term trends, as well as possible reversals or persistence characteristics.

Logarithmic returns, Momentum, RSI, CCI, and cumulative returns all focus on revealing price trends and momentum information, helping to determine whether the current market is in an uptrend, downtrend, or showing reversal signals. TR, ATR, 5-day volatility, and Intraday Volatility are mainly used to assess market volatility, thereby providing a basis for risk management and position control. By analyzing the intraday high-low ratio, William's %R, and stochastic indicators (%K and %D-slow), investors can capture short-term price deviations from normal trajectories, thereby judging whether there are overbought, oversold, or market reversal signals. SMA, WMA, and HMA provide means of smoothing price signals, helping to extract long-term trends, filter short-term noise, and build more robust trend prediction models.

4. Methodology

4.1. Data collection

This study selected nine companies that completed dual primary listings in both Hong Kong and US markets prior to 2022 as analytical subjects. The data collection window spans from each company's dual-listing commencement date to December 31, 2024. Utilizing the professional iFinD financial data terminal, we systematically retrieved daily trading metrics for all constituent securities. The platform's currency conversion module was strategically deployed to normalize historical pricing data into USD equivalents using contemporaneous exchange rates, enabling standardized cross-

DOI: 10.54254/2754-1169/2025.BJ24912

market analysis. The finalized dataset encompasses five key metrics: trading date, opening price, daily high, daily low, and closing price.

4.2. Data preprocess

4.2.1. Standardization of single market data

To analyze the cross-market dynamics of companies dual-listed on the Hong Kong and US stock markets, it is first necessary to standardize single market data to establish a uniform time frame and ensure data integrity.

The complete daily timeline from 2021 to 2024 is reconstructed to process market data by first inserting missing dates to ensure that the calendar is continuous and non-trading days fill zeros to distinguish between real market activity and missing data, then generating binary trading day identifiers and detecting non-zero values in the identifiers to identify valid trading periods. The integrity of the time series is preserved after the trading day fill and zero fill cover, thus laying a solid foundation for subsequent cross-market analysis.

4.2.2. Cross-market data integration

After standardizing single-market data, the next step is the convergence of cross-market data, that is, aligning Hong Kong and US trading hours to capture the interdependence of the cross-border market.

Initially, the trading day data in the Hong Kong and US data sets are filtered with the trading day mark, and then the trading data of the two markets are closely aligned according to the "Date" column. The internal link only retains the date data when both markets have trading activities, and adds market-specific suffixes such as "_HK" and "_US" to the financial characteristics to build a unified feature space. This fusion method can not only preserve the unique features of each market but also facilitate direct comparison, which can filter unilateral transaction noise well and make the analysis effect of synchronous price dynamics stronger.

4.3. Benchmark

To facilitate comparative analysis, we establish a benchmark model. The daily log return rate is obtained by first calculating the logarithm of the ratio between the current opening price and the previous closing price, and then the forecast value is generated by applying the historical extended average return rate to the previous closing price. The estimation formula of the next opening price is as follows:

$$\text{Predicted Open}_t = \text{Close}_{t-1} \times \exp\left(\text{Average Historical Return}_{t-1}\right),$$

where

$$\text{Average Historical Return}_{t-1} = \frac{1}{t-1} \sum_{i=1}^{t-1} \text{Return}_i.$$

This method is simple and efficient to calculate, but it also has many significant limitations. It ignores the common volatility aggregation mode in financial markets, cannot timely respond to

DOI: 10.54254/2754-1169/2025.BJ24912

sudden market shocks, and relies on the cumulative average, which delays the response to recent trends, so it is not suitable for accurate and timely prediction in dynamic market environment.

4.4. MLP

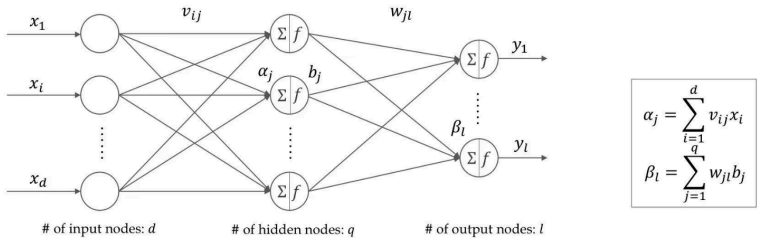


Figure 1. A schematic representation of the MLP architecture

The application of Artificial Neural Networks (ANNs), particularly Multi-Layer Perceptrons (MLPs), is prevalent in time series forecasting [5]. The selection of MLPs for this study is primarily justified by the following rationales:

(1) Complex nonlinear relationships in financial time series data can be effectively captured by multi-layer perceptrons, because accurate prediction of price trends in stock price prediction is dependent on these nonlinear relationships. Ayyild1z [9] emphasized that multi-layer perceptrons perform well in predicting stock market trends.

(2) The multi-layer perceptron is flexible. Engineered and derived indicators can be used as input features, which makes it more applicable and helps to analyze market dynamics more deeply. Technical indicators such as moving average and relative strength index, as well as fundamental factors such as earnings per share, can be integrated into the multi-layer perceptron model as additional features.

(3) Compared with other machine learning models such as logistic regression and support vector machine, the effectiveness of multi-layer perceptron is verified, which comprehensively evaluates the prediction ability of different methods and further shows that multi-layer perceptron is suitable for the task of stock price prediction.

(4) Compared with more complex machine learning models such as convolutional neural network and long and short-term memory network, multi-layer perceptron has lower computational cost, which shortens training time and reduces resource requirements, making it especially suitable for real-time application in trading environment.

The structure of a neuron consists of inputs $x_1, x_2, \ldots, x_n$, $weight\ sw_1, w_2, \ldots, w_n$, bias b, transfer function f, and output a is presented in Figure 2.

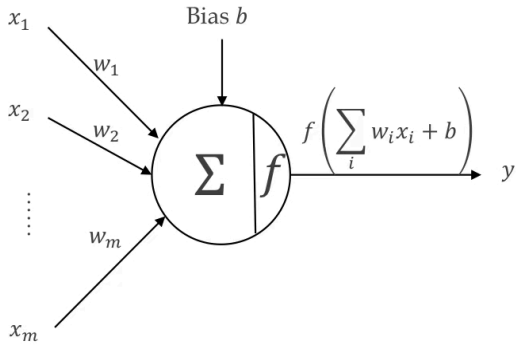


Figure 2. The structure of neuron

DOI: 10.54254/2754-1169/2025.BJ24912

The neuronal input originates from other neurons within the preceding layer. The weighted inputs, along with the bias, are summed, and the resultant value is then processed through a transfer function, denoted as "f" in the provided figure, to generate an output. This output subsequently serves as an input for other neurons in the subsequent layer.

Hidden layers receive inputs from the preceding layer and subsequently provide outputs to either another hidden layer or the output layer. Various non-linear activation functions, including ReLU, Sigmoid, and Tanh, can be employed within the hidden layers of an Artificial Neural Network (ANN). The selection of the activation function in the output layer is contingent upon the nature of the target variable. In the context of this study, which focuses on regression prediction, a linear output function is utilized in the final layer.

4.4.1. The settings of MLP

In this study, the proposed Multi-Layer Perceptron (MLP) is implemented in Python. The dataset for each selected stock varies in size, yet is consistently partitioned into three distinct subsets: 64% allocated for training, 16% for validation, and 20% for testing. The validation dataset, comprising approximately 160 data points, is utilized for hyperparameter tuning during the training phase. The final MLP parameter configurations, post-testing, are detailed in the accompanying table 1.

Table 1. The MLP model parameters setting

Parameter	Value
Number of the input layer neurons	29 or 58
Number of the input layer neurons	15
Number of epochs	100
Performance Goal	1e-6
Maximum Validation Checks	20
Hidden layer function	Relu
Regularization	L2($\lambda = 0.01$)
Performance function	MSE

4.5. Employing overnight spread dynamics to forecast the opening price

The methodology employed herein utilizes overnight spread forecasting to predict the opening price, predicated on two primary rationales.

In the context of non-linear model forecasting, the opening price frequently exhibits higher volatility compared to the price differential, which is defined as the variance between the opening price and the previous day's closing price. Consequently, an approximation of the opening price can be derived by summing the price differential with the preceding closing price.

Furthermore, within the scope of this analysis, the Mean Squared Error (MSE) for both predicted price spreads and predicted opening prices is equivalent. This observation suggests that forecasting price spreads serves as a viable proxy for directly predicting opening prices, thereby offering a more robust and reliable predictive methodology. The following is the validation process.

Original MSE Formula

DOI: 10.54254/2754-1169/2025.BJ24912

$$\text{MSE} = \sum \Big(P_i - O_i \Big)^2$$

$(\text{Where} P_i = \text{Predict}(i), O_i = \text{Open}(i), C_{i-1} = \text{Close}(i-1))$
Reparameterized Spread Form

$$P_i - O_i = \Big(P_i - C_{i-1} \Big) - \Big(O_i - C_{i-1} \Big) = \Delta_i^{\text{predicted}} - \Delta_i^{\text{true}}$$

Final MSE Expression

$$\text{MSE} = \sum \left(\Delta_i^{\text{predicted}} - \Delta_i^{\text{true}} \right)^2$$

For each dual-listed company, three predictive models were implemented: a single-market prediction (forecasting the U.S. stock price using only U.S. market indicators), a dual-market prediction (forecasting the U.S. stock price using both U.S. and Hong Kong market indicators), and a weighted-market prediction (generating a new variable for prediction by assigning weights to indicators from both markets). The efficacy of the dual-market prediction will be demonstrated through the application of these three models to the selected companies.

5. Empirical Results and Analysis

5.1. Evaluation Metrics

To assess the accuracy and effectiveness of the constructed model, the following three evaluation metrics are employed: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of determination(R^2). The definitions of these metrics are as follows:

Mean Squared Error (MSE) quantifies the average squared difference between the model's predicted values and the actual values, directly reflecting the absolute magnitude of the prediction error.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

Root Mean Squared Error (RMSE) is similar to MSE, but its units match the target variable, providing a more intuitive measure of prediction error.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}$$

R^2 gauges the model's capacity to explain the variability of the target variable, reflecting the extent to which the model can account for fluctuations within the data, thereby indicating the degree of fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}$$

DOI: 10.54254/2754-1169/2025.BJ24912

5.2. Empirical Results

This study encompasses 18 stocks from 9 companies listed on both the Hong Kong Stock Exchange and the U.S. stock market prior to 2022. To validate the rationality and efficacy of the model designed in this research, we employ four models to forecast the overnight spread for each stock: the benchmark model, the single-market model, the weighted-market model, and the dual-market model. The weighted-market model and the dual-market model both represent the theory that the dual-market information is more effective, as proposed in this study. Specifically, the benchmark model directly forecasts the opening price of the stock market the following day (Predicted Open_t). Single-market models, weighted market models, and dual-market models predict overnight spreads ($\Delta_t^{\text{Predicted}}$). The subsequent day's opening price is indirectly forecasted via a linear transformation ($Predicted\ Open_t = Closing\ price_{t-1} + \Delta_t^{Predicted}$), ensuring that the model's goodness-of-fit metrics remain constant throughout this transformation. The model's validity is established if its predictive performance surpasses the benchmark model. Furthermore, the model's efficacy is demonstrated when its predictions outperform a single-market model, indicating that incorporating information from another market's equities enhances the accuracy of the price forecasts. Given that the benchmark is a linear model, while the MLP is non-linear, a direct comparison of their goodness-of-fit is not statistically meaningful. The comparison of the other three models against the benchmark is solely presented using MSE.

The fitting results of Two market model for all models, employing the dual-market data, are presented in Figure 3. The blue line represents the real stock price fluctuations, and the yellow line indicates that the model fits the historical stock value.

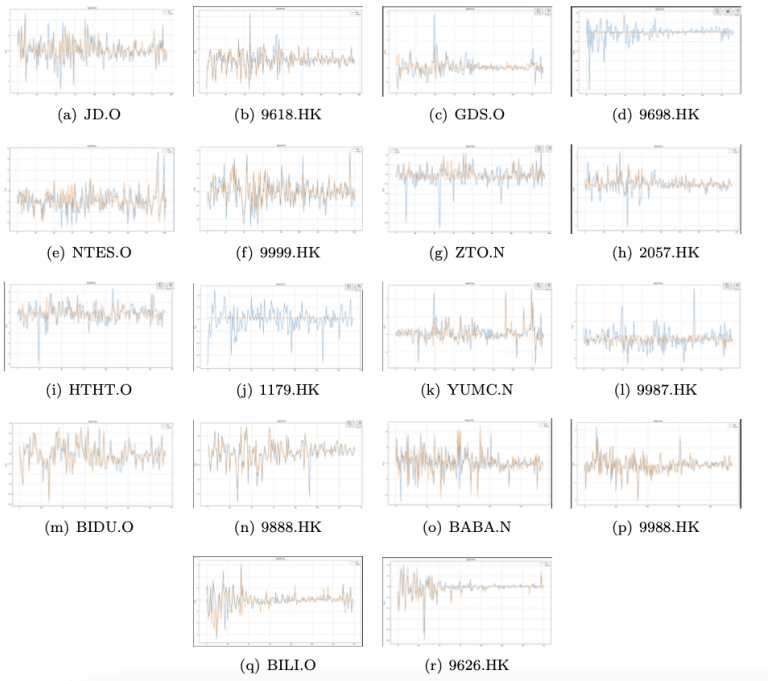


Figure 3. The fitting effect of the Two market model of 18 stocks

The research findings for various models concerning the stocks of ZTO Express, Yum China, and Bilibili are presented in Table 2. The results unequivocally demonstrate that the two-market model outperforms the others across all evaluation metrics. Specifically, it exhibits the lowest Mean

DOI: 10.54254/2754-1169/2025.BJ24912

Squared Error (MSE) and Root Mean Squared Error (RMSE) values, alongside a closer proximity to 1 in R-squared .

Table 2. The result of prediction of opening price on ZTO Express, Yum China, and Bilibili

		Evaluation Metrics		
Stock	Model	MSE	RMSE	R^2
ZTO.N	Benchmark	0.2282	0.4779	0.0345
	Single Market	0.2297	0.4807	0.0249
	Weighted Market	0.1548	0.3936	0.3257
	Two Market	0.1284	0.3589	0.4907
2057.HK	Benchmark	0.2379	0.4887	0.0757
	Single Market	0.1352	0.3678	0.2155
	Weighted Market	0.0983	0.3134	0.3993
	Two Market	0.0767	0.2770	0.5227
YUMC.N	Benchmark	0.2282	0.4779	0.0645
	Single Market	0.6438	0.8016	0.0123
	Weighted Market	0.7817	0.9367	0.0101
	Two Market	0.9133	0.9550	0.0084
9887.HK	Benchmark	0.1938	0.4403	0.2538
	Single Market	0.7304	0.8552	0.0131
	Weighted Market	0.9510	0.9752	0.0100
	Two Market	1.2843	1.1331	0.0072
BILI.O	Benchmark	0.2513	0.5013	0.0670
	Single Market	0.6703	0.8190	0.0124
	Weighted Market	0.8286	0.9091	0.0108
	Two Market	1.6900	1.3000	0.0035
9626.HK	Benchmark	0.3466	0.5893	0.1274
	Single Market	0.7402	0.8605	0.0101
	Weighted Market	1.1900	1.0910	0.0088
	Two Market	1.6900	1.3000	0.0033

Table 3 presents the stock research findings for Alibaba Group, NetEase, and Baidu. The results, evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared , indicate that, with the exception of BABA.N, two market models demonstrate superior performance across the evaluation metrics. The only notable difference is observed in the price predictions for BABA.N, where the weighted market model achieves the highest value at 0.5804, slightly exceeding the two-market model's 0.5716. Furthermore, in terms of predictive accuracy and the ability to explain sample volatility, models incorporating two-market information significantly outperform those utilizing single-market and benchmark models.

DOI: 10.54254/2754-1169/2025.BJ24912

Table 3. The result of prediction of opening price on Alibaba Group, NetEase, and Baidu

		Evaluation Metrics		
Stock	Model	MSE	RMSE	R^2
BABA.N	Benchmark	10.406		
	Single Market	4.9667	2.2286	0.2853
	Weighted Market	4.5223	2.1266	0.5804
	Two Market	4.1788	2.0442	0.5716
9988.HK	Benchmark	0.1799		
	Single Market	0.0580	0.2408	0.5971
	Weighted Market	0.0433	0.2080	0.7957
	Two Market	0.0184	0.1358	0.9222
NTES.O	Benchmark	4.8759		0.0001
	Single Market	5.8514	2.4190	0.0001
	Weighted Market	3.6151	1.9013	0.0731
	Two Market	1.8041	1.3432	0.5432
9999.HK	Benchmark	0.1815		0.0064
	Single Market	0.2004	0.4477	0.0064
	Weighted Market	0.2083	0.4564	0.0001
	Two Market	0.0703	0.2652	0.6679
BIDU.O	Benchmark	8.6564		0.0536
	Single Market	7.6895	2.7730	0.0536
	Weighted Market	4.3042	2.0747	0.4479
	Two Market	1.9615	1.4005	0.7522
9888.HK	Benchmark	0.1692		
	Single Market	0.1741	0.4172	0.0019
	Weighted Market	0.0587	0.2424	0.6466
	Two Market	0.0298	0.1727	0.7832

Table 4 presents the findings for Huazhu Group, GDS Holdings, and JD. No single model demonstrates a definitive advantage; however, the single-market model exhibits lower MSE and RMSE exclusively when forecasting GDS.O. In most scenarios, incorporating alternative market data yields more precise and robust simulation outcomes.

DOI: 10.54254/2754-1169/2025.BJ24912

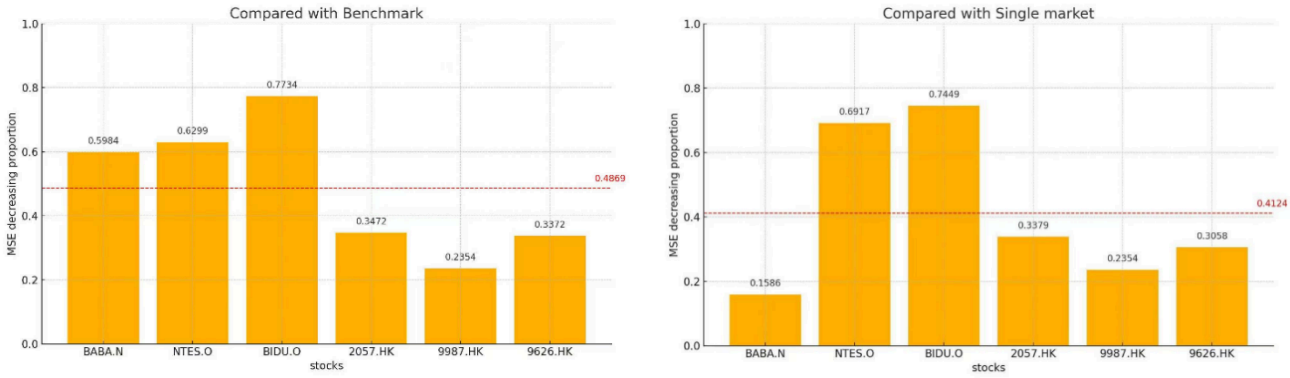
Table 4. The result of prediction of opening price on Huazhu Group, GDS Holdings, and JD

		Evaluation Metrics		
Stock	Model	MSE	RMSE	R^2
HTHT.O	Benchmark	0.8093		
	Single Market	1.0687	1.0289	0.1599
	Weighted Market	0.5558	0.7455	0.2165
	Two Market	0.5167	0.7188	0.2909
1179.HK	Benchmark	0.0090		
	Single Market	0.0165	0.1285	0.0001
	Weighted Market	0.0086	0.0928	0.0012
	Two Market	0.0086	0.0929	0.0131
GDS.O	Benchmark	0.9558		0.0010
	Single Market	0.6280	0.7925	0.0010
	Weighted Market	0.9240	0.8584	0.0751
	Two Market	0.7369	0.9612	0.2460
9698.HK	Benchmark	0.0243		0.0011
	Single Market	0.0260	0.1611	0.0011
	Weighted Market	0.0240	0.1548	0.0441
	Two Market	0.0238	0.1544	0.0155
JD.O	Benchmark	2.3158		0.0780
	Single Market	3.3052	1.8180	0.0780
	Weighted Market	0.4894	0.6996	0.2172
	Two Market	0.8715	0.9336	0.5711
9618.HK	Benchmark	0.5450		
	Single Market	0.6337	0.7960	0.0090
	Weighted Market	0.4870	0.6979	0.2578
	Two Market	0.2262	0.4756	0.6260

5.3. Analysis

We selected the three stocks with the highest individual share prices from nine companies in the U.S. stock market, which are: BABA.N, NTES.O, BIDU.O, 2057.HK, 9987.HK and 9626.HK. Siultaneously, we selected the three stocks with the highest individual share prices in the Hong Kong stock market. We observed that the models incorporating data from both markets outperformed single-market models and the benchmark, with all achieving an R-squared value above 0.3. These data indicate a significant positive impact of the other market's data on the price prediction of these stocks. We calculated the Mean Squared Error (MSE) reduction for the models representing dual-market information for these six stocks, and the results are shown in Figure 4.

DOI: 10.54254/2754-1169/2025.BJ24912



(a) Compared with Benchmark (b) Compared with Single market

Figure 4. Dual market information enhancement magnitude

The Mean Squared Error (MSE) demonstrates a substantial reduction, exceeding 40% on average, when compared to both the baseline and single-market models. Specifically, the average MSE reduction reaches 48% relative to the baseline, effectively halving the MSE value. Furthermore, the single-market model also exhibits an average MSE decrease exceeding 40%, with the minimum reduction observed being 15%.

Among all model prediction outcomes, three stocks (1179.HK, 9698.HK, GDS.O) exhibit R-squared close to zero. This indicates poor model fitting and difficulty in explaining data fluctuations. As illustrated in Figure 4, the average price calculation across 18 stocks clearly demonstrates that these specific stocks are trading at low price levels. Considering the study's reliance on the MLP model, the input features of low-mean companies (e.g. stock price, trading volume, bid-ask spread) typically have narrow numerical ranges and small absolute values. Without adequate standardization, these features may experience signal attenuation through the multiple linear transformations within the MLP. The deep structure of the MLP further amplifies this attenuation effect, hindering deep-layer neurons from capturing effective non-linear patterns. Simultaneously, during the initial training phase, the model tends to minimize the loss function (e.g. MSE) rapidly by predicting the mean. The subtle feature differences in low-mean data are insufficient to drive the network beyond this local optimum. Furthermore, when the correlation between features and the target variable is weak, the MLP may excessively rely on the bias term to compensate for the missing signal, ultimately leading to predictions "anchored" near the data mean. This phenomenon is particularly pronounced in high-noise, low signal-to-noise ratio data from low-priced stocks. Therefore, the results for these companies can be considered outliers in this experiment, reflecting certain limitations in the model's applicability.

6. Conclusion and Suggestions for Future Research

Predicting the future opening price of stocks holds considerable importance within the realm of financial economics. More precise stock price predictions not only serve as a reflection of market sentiment but also facilitate the identification of arbitrage opportunities. In this study, we have selected companies listed on both the Hong Kong and U.S. stock exchanges, employing an artificial neural network model to forecast the opening prices of their respective stocks across both markets. To mitigate biases arising from limited data and to exclude the influence of volatile companies, we have focused on nine companies, representing eighteen stocks, that were dual-listed prior to 2022.

DOI: 10.54254/2754-1169/2025.BJ24912

Furthermore, we have adopted an indirect approach to predict opening prices by forecasting overnight price gaps. Compared to directly predicting opening prices, overnight price gaps exhibit lower volatility and greater data stability, which is expected to enhance the model's fitting performance. Based on the MLP framework, we have constructed a weighted market model and a dual-market model to capture the benefits of incorporating dual-market information. Additionally, we have developed single-market models and benchmark models to assess the predictive efficacy of the aforementioned models. The single-market, weighted market, and dual-market models all outperformed the benchmark model. Compared to the single-market model, the weighted market model and the dual-market model, which incorporate information from both markets, generally performed better in most cases, with a significant improvement in the accuracy of stock price predictions, particularly for stable stocks with higher prices and longer listing periods. Overall, incorporating information from another market led to a certain degree of improvement in the prediction accuracy as reflected by the backtesting results based on historical data. Simultaneously, the model's ability to explain the fluctuations in the original sample also improved, which is reflected in the overall enhancement compared to using only single-market information.

This study is subject to certain limitations, primarily stemming from the inherent constraints of artificial neural networks. Specifically, the model's performance is less robust for stocks characterized by low input features, such as low price, low trading volume, and narrow bid-ask spreads, due to the limitations of the artificial neural network architecture. Furthermore, this research did not incorporate pre-market trading data. It is evident that pre-market trading activity can influence the opening price of a stock. However, this paper posits that the opening price data utilized in the model fitting already encapsulates the effects of pre-market trading, as these activities are reflected in the price. Therefore, no additional adjustments were deemed necessary.

Recommendations for future research: The prediction of opening prices for dual-listed stocks can potentially generate arbitrage opportunities, necessitating the development of specific arbitrage strategies that more closely reflect the realities of stock market arbitrage. Furthermore, the information influencing the opening prices of dual-listed stocks across both markets could be expanded to include factors such as macroeconomic conditions in the other stock market and Federal Reserve policies. Consequently, incorporating these additional factors into the optimization model may yield a more accurate model, which could then be applied to future real-world arbitrage scenarios.

References

[1] Bijing Li, Ronghua Yi, and Roger Su. Spillover effect of chinese cross-listed companies across shanghai, hong kong and us markets. International Journal of Economics and Finance, 3(6):135, 2011.

[2] Shane A Corwin and Paul Schultz. A simple way to estimate bid-ask spreads from daily high and low prices. The journal of finance, 67(2):719–760, 2012.

[3] Louis Gagnon and G Andrew Karolyi. Multi-market trading and arbitrage. Journal of Financial Economics, 97(1):53–80, 2010.

[4] Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. European journal of operational research, 270(2):654–669, 2018.

[5] Shaghayegh Abolmakarem, Farshid Abdi, Kaveh Khalili-Damghani, and Hosein Didehkhani. A multi-stage machine learning approach for stock price prediction: Engineered and derivative indices. Intelligent Systems with Applications, 24:200449, 2024.

[6] Cheol S Eun and Sanjiv Sabherwal. Cross-border listings and price discovery: Evidence from us-listed canadian stocks. The Journal of Finance, 58(2):549–575, 2003.

[7] Vaira T Alaganar and Ramaprasad Bhar. Information and volatility linkage under external shocks: Evidence from dually listed australian stocks. International review of financial analysis, 11(1):59–71, 2002.

DOI: 10.54254/2754-1169/2025.BJ24912

[8] Nazif Ayyildiz and Omer Iskenderoglu. How effective is machine learning in stock market predictions? Heliyon, 10(2), 2024.
[9] Nazif AYYILDIZ. Prediction of Stock Market Index Movements with Machine Learning. Ozgur Publications, 2023.
[10] Fabio D Freitas, Alberto F De Souza, and Ailson R De Almeida. Prediction-based portfolio optimization model using neural networks. Neurocomputing, 72(10-12):2155–2170, 2009.
[11] Yujin Baek and Ha Young Kim. Modaugnet: A new forecasting framework for stock market index value with an overfitting prevention lstm module and a prediction lstm module. Expert Systems with Applications, 113:457–480, 2018.
[12] Kai Chen, Yi Zhou, and Fangyan Dai. A lstm-based method for stock returns prediction: A case study of china stock market. In 2015 IEEE international conference on big data (big data), pages 2823–2824. IEEE, 2015.