A Multi-Level LSTM-K-Means Deep Learning Framework for Robust Stock Prediction and Risk-Controlled Quantitative Investment

Sangbing Tsai[1*], Jaheer Mukthar.KP [2]

[1]International Engineering and Technology Institute, Hong Kong; klj0418@gmail.com

[2]Kristu Jayanti College Autonomous Bengaluru, India; jaheermukthar@gmail.com

*Corresponding Author: klj0418@gmail.com

DOI: https://doi.org/10.30211/JIC.202503.013

Submitted: Aug. 03, 2025 Accepted: Sep. 23, 2025

ABSTRACT

In recent years, the stock market has attracted increasing attention. The inherent volatility of stock prices, often influenced by national and social policies, poses significant challenges for investors seeking profitable returns. With the rapid development of artificial intelligence, computers have demonstrated outstanding capabilities in handling complex mathematical problems. Consequently, efforts to leverage computational power to analyze and predict stock market trends have been growing. However, existing methods suffer from limited long-term sequence modeling capabilities and struggle to select candidate factors that align with individual investment preferences from a vast array of features. To address these issues, this paper proposes a deep learning factor-based comprehensive prediction model combining LSTM and K-Means. The multi-level LSTM-K-Means integrated prediction approach overcomes traditional neural networks' shortcomings in processing long sequences and nonlinear data by incorporating stock returns and volatility to accurately identify potential high-quality stocks. Furthermore, a multi-factor scoring stock selection strategy, coupled with a fixed-percentage stop-profit and stop-loss mechanism, is designed to effectively control trading risks and enhance the robustness and profitability of quantitative investment. Experimental results demonstrate that the proposed method alleviates gradient vanishing problems, optimizes stock selection and risk management processes, and provides strong support for investors to achieve excess returns.

Keywords: Stock prediction; K-means clustering; Quantitative investment

1. Introduction

Stock price prediction has long been a research hotspot in the field of finance. With the changes in the international landscape and the increasing complexity of the global financial environment, accurately predicting stock prices is critical for personal financial management and national economic stability[1], [2], [3]. However, the stock market is influenced by a variety of complex factors, including economic conditions, political events, and market sentiment. Quantitative investment has

gained widespread attention and application in global financial markets. This trend is driven by the rapid development of big data technologies and the continuous improvement of algorithms and hardware performance, which enable efficient collection, storage, and processing of vast amounts of data generated in financial markets[4], [5], [6]. The core of quantitative investment lies in combining theory with practical trading. By constructing mathematical models and applying statistical analysis methods, quantitative investors are able to identify potential patterns in the market and convert them into actual trading strategies, thereby optimizing investment returns. The key to successful trading is establishing clear execution logic, using computer programs and algorithmic trading technologies to automatically execute pre-set trading decisions, avoiding emotional fluctuations and human interference, and thus enhancing the accuracy and efficiency of trading.

The central idea behind portfolio optimization is how to achieve high returns while minimizing risk in the stock market, which is the primary goal of every investor. The method to achieve portfolio optimization is by grouping stocks and employing different capital investment proportions among different stocks to maximize profit and minimize risk. In 1952, Markowitz proposed a mean-variance model to optimize portfolios. Investors can use this model to obtain the expected return at the minimum risk[7], [8]. Since its introduction, this model has attracted widespread attention in both academia and the financial industry and has been extensively applied in financial research. However, the Markowitz theory also has some limitations, such as its high sensitivity to historical prices and its inability to incorporate subjective opinions of investors. To overcome these problems, Fama proposed the Efficient Market Hypothesis[9], which argues that it is impossible to predict future stock prices and beat the market because stock prices fully reflect all relevant information. However, many scholars have challenged this view, suggesting that stock prices are partially predictable and began using algorithms capable of modeling the stock market.

In the financial market, the goal of investors is to achieve profit. If private or institutional investors can accurately predict market behavior, they will be able to consistently obtain higher risk-adjusted returns than the market[10], [11]. With the continuous development of machine learning-based forecasting methods, an increasing number of financial problems have been effectively solved. In machine learning, research is generally divided into shallow learning and deep learning. Shallow learning originated in the 1920s with the introduction of the Back-propagation algorithm for artificial neural networks, which facilitated the widespread application of statistical-based machine learning algorithms[12]. Although early artificial neural networks were called Multi-Layer Perceptrons[13], due to the difficulty in training multi-layer networks, shallow models with only one hidden layer were typically constructed. In 2006, Hinton proposed deep learning algorithms, significantly enhancing the capabilities of neural networks and marking the rise of deep learning in both academia and industry[14]. This development has led to the adoption of deep learning and other computational intelligence methods to create accurate stock market prediction models. Currently, nonlinear prediction models primarily include deep learning-based artificial neural networks, machine learning's Random Forest model, the Prophet model, and Generalized Additive Models[15] in statistical learning. As a typical nonlinear method, artificial neural networks excel at processing

nonlinear, discontinuous, and high-frequency multi-dimensional data, making them widely used in financial forecasting. However, traditional shallow artificial neural networks face several limitations when applied to financial forecasting, such as susceptibility to overfitting, which leads to poor performance on out-of-sample data, potential issues like vanishing or exploding gradients during optimization that limit the network's learning capacity, and difficulty in finding global optima, affecting overall model performance. In recent years, researchers have focused on improving the structure and training algorithms of artificial neural networks, employing more complex network structures, introducing regularization techniques, and optimizing algorithms to effectively address challenges such as overfitting, vanishing gradients, and global optimization, thereby improving the prediction accuracy and stability of the models.

As a classical method in deep learning, Long Short-Term Memory networks (LSTM)[16] have demonstrated remarkable capability in capturing long-term dependencies in sequential data. The core advantage of LSTM lies in its ability to effectively model nonlinear features and complex interactions in financial time-series data, fully exploiting sequential information and enabling predictions on high-dimensional, non-stationary data that are challenging for traditional statistical models and shallow machine learning methods[17]. Compared with conventional artificial neural networks, LSTM offers several notable benefits: first, its unsupervised layer-wise feature extraction enhances feature representation, allowing the model to capture more complex functional mappings and nonlinear relationships; second, LSTM exhibits strong generalization ability, improving prediction accuracy on the training set while maintaining robustness and adaptability on out-of-sample data. Nevertheless, financial markets encompass numerous potential influencing factors, which often exhibit highly nonlinear coupling and dynamic interactions, making the identification and selection of effective factors a challenging task. Although existing methods can accomplish basic forecasting to some extent, they struggle to fully capture the intrinsic complexity and long-term dependency structures of large-scale, multidimensional, and multi-frequency financial data, thereby limiting prediction accuracy and robustness. In recent years, researchers have increasingly explored integrating LSTM with attention mechanisms, graph neural networks, and hybrid deep models, aiming to enhance the modeling of long-term dependencies and complex nonlinear patterns in financial time series, and provide more reliable tools for precise forecasting and intelligent investment decision-making.

To address these issues, we propose a deep learning factor integration prediction model based on LSTM-K-Means, aimed at improving the accuracy of stock price forecasting and investment returns. By combining multiple LSTM layers, we are able to extract features from different levels and learn the temporal relationships, thus enhancing the model's ability to express stock price sequences, especially in capturing potential nonlinear patterns. To further improve prediction accuracy, we also introduce the K-means clustering algorithm as an auxiliary analysis tool to cluster the data processed by the LSTM model and classify stocks based on characteristics such as stock returns and volatility. This strategy helps us identify stocks with strong upward momentum, thus optimizing the investment portfolio. In addition, we combine a multi-factor scoring stock selection model with a fixed-percentage stop-loss and take-profit strategy to propose a comprehensive quantitative investment

method. In this method, the multi-factor scoring stock selection model identifies the most promising investment factors by comparing multiple financial indicators, while the fixed-percentage stop-loss and take-profit strategy helps us control risks and protect the portfolio from significant losses. By combining these two strategies with the LSTM model, we not only enhance prediction accuracy but also provide investors with effective risk control and return growth strategies, aiming to achieve long-term "excess returns."

Our research contributions are mainly reflected in the following aspects: first, the multi-level LSTM model enhances the learning ability of stock time-series data; second, the combination of the K-means clustering algorithm optimizes the stock selection process; third, the proposed comprehensive quantitative investment method effectively improves the accuracy of investment decisions through multi-factor scoring stock selection and stop-loss/take-profit strategies; finally, by combining the LSTM model with quantitative strategies, we provide an effective risk control and return growth solution, aimed at enabling investors to achieve long-term excess returns.

2. Literature Review

2.1 LSTM-based Stock Prediction Models

Stock price prediction remains a critical and challenging task in financial research due to the complex and nonlinear nature of market dynamics. Among various deep learning architectures, Long Short-Term Memory (LSTM) networks have gained substantial attention for their ability to capture temporal dependencies in sequential data. Numerous hybrid models combining LSTM with other techniques have been proposed to improve prediction accuracy and robustness. For instance, SACLSTM [18] integrates Convolutional Neural Networks (CNN) with LSTM by constructing sequential arrays of historical data and indicators, where CNN extracts features fed into LSTM for time series forecasting. Gao et al. [19] developed a multifactor model utilizing technical indicators, investor sentiment, and financial data, employing dimensionality reduction methods such as LASSO and PCA to validate the effectiveness of LSTM and GRU networks. Bhandari et al. [20] compared single-layer and multi-layer LSTM models for predicting next-day closing prices of the S&P 500 index, finding superior performance with the single-layer architecture. AMV-LSTM [21] enhances stability and generalization by optimizing gating structures and incorporating attention mechanisms alongside Adam optimization to mitigate overfitting and instability issues. Prabakar et al. [22] proposed a hybrid model combining feature selection and LSTM to improve prediction accuracy by reducing dimensionality with a selected set of 15 indicators. WCN-LSTM [23] integrates market, industry, and stock-related news classification with weighted sentiment analysis to strengthen sequence learning capability. Burak et al. [24] introduced an LSTM model optimized via Artificial Rabbit Optimization (ARO), demonstrating improved accuracy over other neural networks on DJIA data. Baek et al. [25] employed a Genetic Algorithm (GA) to optimize a CNN-LSTM hybrid model for next-day stock price prediction, utilizing 20 days of historical price and technical data. Yong et al. [26] combined Graph Convolutional Networks (GCN) with LSTM to leverage capital flow features and graph-structured stock relationships for more precise price trend predictions. Muhammad

et al. [27] proposed a hybrid approach integrating improved Empirical Mode Decomposition (EMD) with LSTM, using Akima spline interpolation to decompose noisy stock data into intrinsic mode functions (IMFs) for enhanced nonlinear volatility prediction. Lastly, LSTM-BO-LightGBM [28] synergizes multi-layer bidirectional LSTM with Bayesian-optimized LightGBM, achieving superior time series feature extraction and parameter tuning, resulting in highly accurate stock price fluctuation forecasts across multiple assets.

Despite recent advances in LSTM-based hybrid forecasting models, several critical challenges remain. First, many models require substantial computational resources due to complex network architectures or multi-step processing pipelines, limiting their applicability in high-frequency or real-time trading scenarios. Second, although algorithmic integration can improve prediction accuracy, such hybrid strategies often lack interpretability, making it difficult for investors or decision-makers to understand the underlying prediction logic and risk sources, thereby reducing model transparency and trustworthiness. Third, existing approaches predominantly focus on historical prices and technical indicators, while underutilizing alternative or unstructured data sources such as macroeconomic variables, social sentiment, and financial news, which often carry additional informational value that can significantly enhance predictive comprehensiveness and robustness. Furthermore, risk management mechanisms and adaptive strategies are seldom embedded within these forecasting frameworks, constraining their practical utility in highly dynamic and uncertain financial markets. Future research should prioritize the development of lightweight, efficient, and interpretable LSTM-based hybrid forecasting systems, integrating multi-source heterogeneous data, risk control mechanisms, and adaptive learning strategies to improve stability, generalization, and practical applicability in complex financial environments.

2.2 ARIMA in Stock Market

Stock price prediction has been a critical area of research in financial time series analysis, with the ARIMA (AutoRegressive Integrated Moving Average) model being one of the most widely applied statistical tools. Several studies have employed ARIMA and its variants to forecast stock prices across different markets and sectors. Meher et al.[29] applied the ARIMA model to pharmaceutical stocks in India's NIFTY100 index, first confirming data stationarity using the Augmented Dickey-Fuller (ADF) test, then selecting optimal AR and MA terms based on ACF and PACF plots, and finally choosing best-fit models according to volatility, adjusted R^2, and AIC criteria. Similarly, Mashadihasanli et al.[30] utilized ARIMA to predict monthly stock indices on the Istanbul Stock Exchange, confirming stationarity before testing multiple AR and MA parameter combinations to identify the best model via goodness-of-fit and prediction errors. Dadhich et al.[31] focused on India's BSE and NSE indices, employing ADF tests for stationarity and ACF/PACF analyses to determine candidate ARIMA parameters. Ashok et al.[32] combined ARIMA modeling with LSTM networks using Tata Global Beverages stock data, demonstrating ARIMA's effectiveness in short-term forecasting but showing LSTM's superior accuracy overall. Kobiela et al.[33] compared ARIMA and LSTM models on NASDAQ stock prices, revealing that ARIMA outperforms LSTM when relying solely on historical prices, especially for long-term forecasts. Other advancements

include hybrid approaches such as the ARIMA T model by Pokou et al.[34], which accounts for fat-tailed financial data distributions, ensemble ARIMA-LSTM models by Verma et al.[35], and adaptive wavelet-based hybrid models integrating LSTM and ARIMAX-GARCH components proposed by Zolfaghari et al.[36], enhancing multi-scale volatility and price prediction accuracy.

However, ARIMA-based models still face several inherent limitations. First, ARIMA assumes linearity and stationarity in the data, which often fails to capture the nonlinear, complex dynamics and structural breaks inherent in financial markets. In addition, the requirement for data stationarity typically necessitates differencing or transformation procedures, which may result in the loss of valuable information. ARIMA models also struggle to incorporate external factors or adapt to changing market regimes, often requiring substantial adjustments to function effectively across different market conditions. Although hybrid models that combine ARIMA with machine learning or deep learning techniques partially mitigate these issues, challenges remain in optimizing model complexity, enhancing robustness under diverse market conditions, and maintaining interpretability. Therefore, there is an urgent need to develop more flexible, adaptive, and high-accuracy forecasting frameworks capable of better capturing the complex behavior patterns of stock price movements.

2.3 CNN in Stock Market Prediction

Accurate stock price prediction remains a critical and challenging area of research within financial markets. Traditional time series models like ARIMA have been widely used but often struggle with the nonlinear, high-frequency, and multifactorial nature of stock price movements. Recent advances in deep learning have led to the development of hybrid models that integrate ARIMA with neural networks and other machine learning techniques to improve predictive accuracy and robustness. For instance, CAGTRADE [37] combines convolutional neural networks (CNN), attention mechanisms, and gated recurrent units (GRU) to dynamically weight input sequences for multi-index short-term trend forecasting. Das et al. [38] integrate ensemble empirical mode decomposition (EEMD), ensemble CNN, and Twitter sentiment analysis to robustly predict stock prices by decomposing signals and fusing multiple predictive features. You et al. [39] propose a CNN-GRU framework for market sentiment analysis and risk warning using extensive web text data, capturing local patterns and temporal emotional dynamics. Jagadesh et al. [40] utilize wavelet transform preprocessing, dandelion optimization algorithm (DOA) for feature selection, and a 3D-CNN-GRU hybrid model optimized by blood coagulation algorithm (BCA) for spatial-temporal stock prediction. CLATT[41] incorporates CNN, bidirectional LSTM, and attention mechanisms to model short-term stock correlations with dynamic weighting of temporal features. Woojung et al. [42] combine TimeGAN for time series data augmentation and 3D-CNN for capturing multidimensional temporal-spatial dependencies in the futures market. Somkunwar et al. [43] develop a CNN and multivariate linear regression hybrid for accurate stock valuation on the NSE NIFTY index. Khattak et al.[44] enhance cryptocurrency trend prediction by integrating Fibonacci technical indicators with CNN-LSTM hybrid networks, implementing a six-stage prediction and evaluation system with multi-class trend intensity classification. [45-48]

Despite significant progress, existing ARIMA-based hybrid models continue to exhibit several critical limitations. These models often rely on extensive data preprocessing and intricate feature engineering, which constrains scalability and reduces adaptability in rapidly evolving market environments. While hybridization improves the extraction of nonlinear patterns, many models remain limited in their capacity to integrate heterogeneous data sources, including market sentiment, macroeconomic indicators, and alternative unstructured data, which are increasingly recognized as essential for robust financial forecasting. Furthermore, most approaches emphasize short-term prediction accuracy, providing limited capability to capture long-term dependencies, structural breaks, or regime shifts inherent in financial markets. A persistent trade-off between model complexity and interpretability further hampers real-time deployment and actionable decision-making. Future research should focus on designing flexible, interpretable, and multi-modal predictive frameworks that leverage heterogeneous data, incorporate adaptive learning mechanisms, and effectively capture both short-term dynamics and long-term dependencies, thereby enhancing robustness and practical utility in complex, volatile financial environments.

3. Methodology

Considering the different analysis requirements of stock data, this paper proposes a deep learning factor integration prediction model based on LSTM-K-Means. The model combines the LSTM network with the K-Means algorithm, aiming to predict the closing price of the target stock while also using data clustering to intuitively discover the stock's upward trend and purchase recommendation. The overall framework of the model is shown in Figure 1.

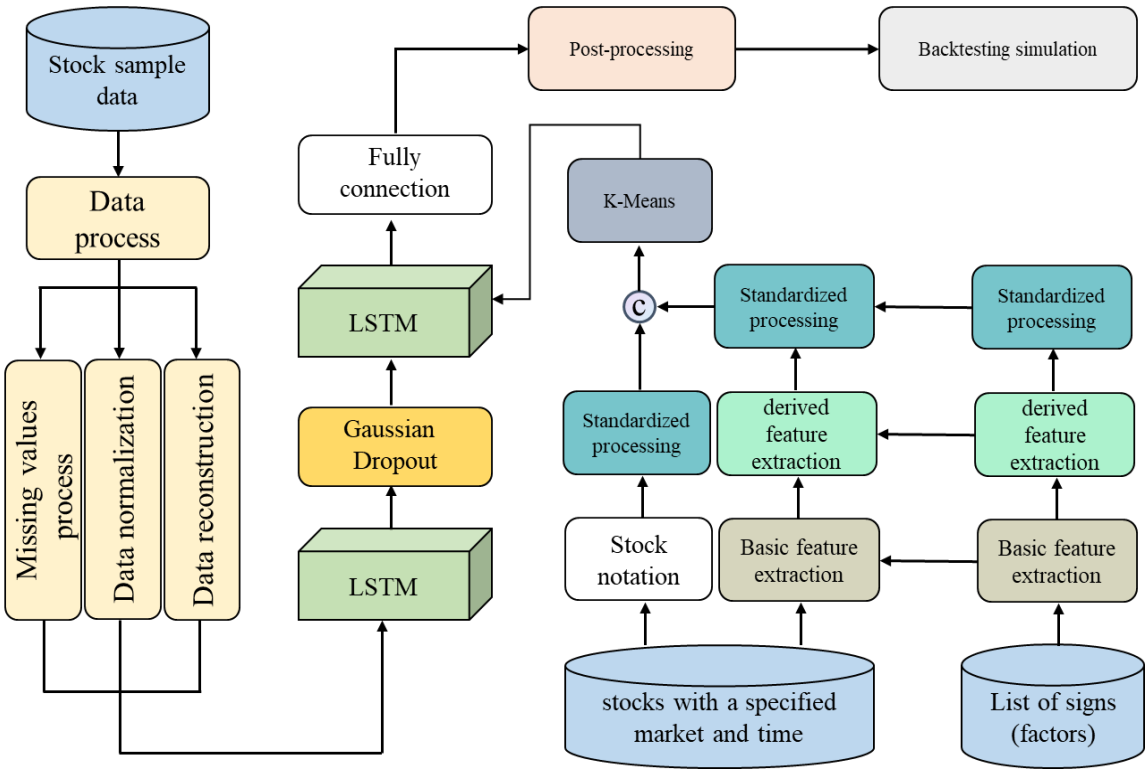

Figure 1. Overall framework diagram of the LSTM-K-Means network

The deep learning factor integration prediction model based on LSTM-K-Means combines LSTM and K-Means algorithms to achieve accurate predictions of future stock movements through deep learning methods while also assisting investors in identifying potential investment opportunities through clustering analysis. The model begins by preprocessing the stock data, including missing value handling, outlier correction, and data normalization, to ensure the quality and consistency of the data. The preprocessed data includes fundamental and technical factors. The fundamental factors, based on monthly data, include earnings per share, operating income per share, operating profit per share, retained earnings per share, and net profit growth rate, which are used in multi-factor regression stock selection. The technical factors, based on weekly data, include the highest price per share and the closing price, which are used as input to the LSTM model for stock price prediction. The LSTM model captures long-term dependencies in time-series data, improving prediction accuracy, especially when handling nonlinear and long-span stock data. At the same time, the K-Means clustering algorithm partitions the data into multiple clusters, helping to identify stocks with similar trends and characteristics, thereby revealing market behavior patterns and providing guidance for portfolio optimization. Ultimately, the prediction results from the LSTM model show the upward trend and expected returns of the target stock, while the K-Means clustering analysis reveals the stock's recommendation level and similarity, helping investors make more informed decisions and maximize their investment portfolio.

3.1 Data Processing

For the LSTM model, data processing involves several essential steps. In the first step, missing values in the stock data are handled. These missing values usually arise due to market closures during weekends or holidays, causing data gaps. To improve prediction accuracy, it is necessary to handle these missing values by either deleting them or performing imputation, depending on the specific situation. The second step focuses on data normalization. Given the differences in the magnitudes of the features, and the fact that the speed of gradient descent during training is proportional to the magnitude of the features, Min-Max normalization is applied to scale the data to the range [0, 1]. This transformation not only accelerates the model's convergence speed but also improves prediction accuracy. The normalization formula is given by:

$$\text{norm}(X) = \frac{X - \min(X)}{\max(X) - \min(X)}$$

The third step involves data reconstruction. The dataset is first divided into the input (x) and output (y), where the input x is used to predict y. This process transforms the data into a time-series format and reframes it as a supervised learning problem. Subsequently, the list-type data is converted into array format, and the two-dimensional data is transformed into a three-dimensional structure. The input for the LSTM model requires a three-dimensional array of the form (samples, timesteps, features), where "samples" represent the number of data instances, "timesteps" refer to the time steps in the sequence, and "features" denote the number of variables. At this point, both x_train and x_test

are still two-dimensional arrays of shape (samples, timesteps=60), and since only the closing prices are analyzed, an additional feature dimension is added.

For the K-Means data processing, the data is first formatted by applying the normalization procedure described earlier and converting it into a NumPy array format for compatibility with the K-Means algorithm. In addition, any NaN values in the data are removed and replaced with zero, which ensures that the clustering process is not negatively affected by missing data points.

3.2 Two-Layer LSTM Architecture

We designed a two-layer LSTM structure, which provides a higher degree of fit while maintaining a reasonable balance between model complexity and training difficulty. The main structure is shown in Figure 2.

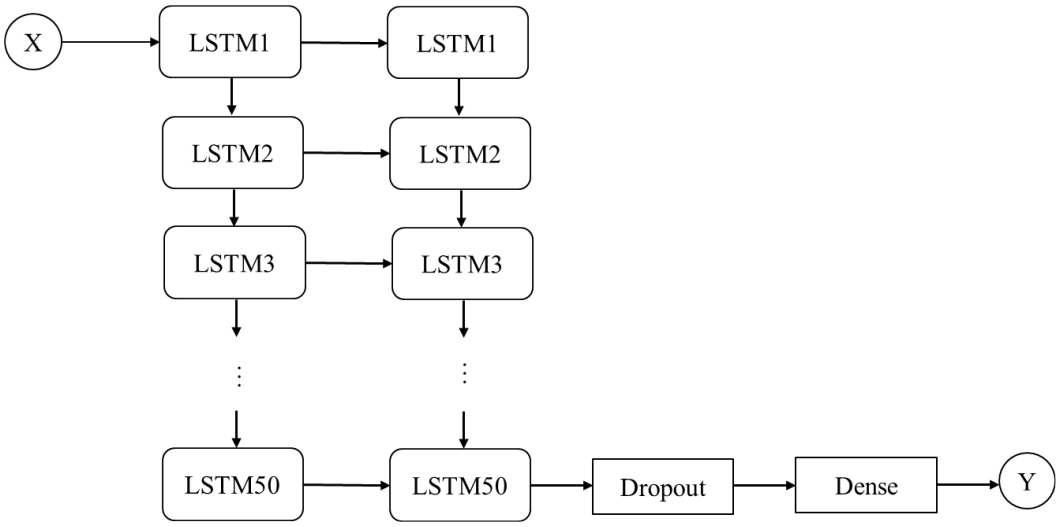

Figure 2. Two layer LSTM model structure

After the data is input into the two-layer LSTM model, it passes sequentially through the pre-set LSTM neurons, outputting vectors of the corresponding dimensions. After the processing in the first layer is completed, the data is passed through a Dropout layer that randomly discards 20\% of the data to prevent overfitting, and then the remaining data is input into the second LSTM layer. Both layers of the LSTM have the same number of neurons. The output of each layer is passed through a fully connected layer with a single neuron, generating the output used to plot the target stock price chart. To optimize the model's performance, the Adam optimizer is used, and the root mean square error is employed as the loss function to evaluate the model's prediction accuracy.

3.3 Effective Factor Selection and Model Optimization

In quantitative investing, the effectiveness of factors is critical for constructing robust investment models and achieving superior returns. Effective factors typically refer to variables that are statistically significant and economically meaningful in relation to stock returns, offering valuable insights into market behavior or asset characteristics. These factors can stem from various dimensions, such as fundamental metrics (e.g., price-to-earnings ratio, price-to-book ratio, earnings growth),

technical indicators (e.g., moving averages, momentum indicators, volatility), macroeconomic factors (e.g., market return, interest rates), and style factors (e.g., value, momentum, quality). In this study, we consider factors' stability, implementability, and interpretability when selecting them, building a candidate factor pool and employing a systematic process for screening and evaluating these factors.

To improve the predictive performance of the LSTM model on stock returns, the preliminary factor screening combines both statistical analysis and machine learning methods. Statistical techniques, such as linear regression, information coefficient (IC), and information ratio (IR), are employed to quantify the relationship between factors and returns. Machine learning algorithms, such as decision trees, random forests, and Lasso regression, are applied for feature selection in large factor sets. This paper selects 30 candidate factors, including both financial and technical factors, and conducts empirical tests on a subset of these factors. Financial factors, extracted on a monthly basis from the cross-sectional perspective, include earnings per share, operating income per share, profit margins, and return on equity. Technical factors are derived from the time-series perspective on a weekly basis, such as closing prices and high prices over certain periods.

To further evaluate the selected factors, a range of performance metrics is introduced, including cumulative returns at the lowest and highest quantile, returns over the past month and year, IC mean, IR mean, factor congestion, and factor valuation. The lowest (highest) quantile cumulative returns measure the performance difference in low (high) sorted portfolios, assessing the factor's effectiveness in stratified investing. The IC mean reflects the strength of the linear relationship between factor signals and future returns, while the IR value quantifies the risk-adjusted return by considering the stability and volatility of returns. Factor congestion measures the extent to which a factor is widely applied in the market, helping to avoid risks associated with strategy homogeneity. Factor valuation reveals the median price-to-book ratio across factor groups, aiding in the identification of whether the factor is overvalued or undervalued, thereby assisting in asset allocation decisions.

3.4 Risk Control Strategy

The fixed percentage take-profit and stop-loss risk control strategy used in this paper is a widely adopted approach in practical trading. When the stock price rises and reaches or exceeds the set take-profit percentage, the stock is sold to realize the profit. This helps lock in some of the gains and prevents missed profit opportunities. The selection of the take-profit percentage should be based on expectations of the stock's potential growth and an assessment of market conditions. Conversely, when the stock price falls and reaches or drops below the set stop-loss percentage, the stock is sold to limit further losses. This helps protect the portfolio from substantial losses and assists in managing investment risk. The choice of the stop-loss percentage should take into account the investor's risk tolerance and the market's volatility.

The fixed percentage take-profit and stop-loss strategy is characterized by its simple and clear rules, helping investors avoid emotional decision-making and holding losing positions for extended periods. However, it is important to note that this strategy may sometimes lead to premature take-profit or stop-loss actions because it relies solely on fixed percentage levels, without considering the

specific conditions of the stock or market changes. In this paper, based on market response, the take-profit and stop-loss percentages are set within the range of 5% to 10%.

4. Experiment description and results

4.1 Data Collection

The NVIDIA stock data used in this study is sourced from the Yahoo Finance platform. This platform provides rich financial market data, covering historical prices, trading volumes, financial indicators, and other information for numerous companies worldwide. The data variables used in this paper are listed in Table 1.

Table 1. Variable Description

Variable Name	Variable Description	Data Type
Date	Specific stock trading date	Date
Closing Price	Stock price at the end of each trading day	Numeric
Opening Price	Stock price at the beginning of each trading day	Numeric
High	Highest price reached during a trading day	Numeric
Low	Lowest price reached during a trading day	Numeric
Trading Volume	Number of shares traded during a trading day	Numeric
Price Change (Daily Return)	The amplitude of stock price change for the day	Numeric
Rolling Volatility	Average trend of stock price volatility	Numeric
Realized Volatility Mean	More accurate measure of stock price volatility	Numeric
Rolling Mean	Average trend of stock price	Numeric
Upper Band	Calculated based on statistical features of stock price, serving as a resistance reference level	Numeric
Lower Band	Calculated based on statistical features of stock price, serving as a support reference level	Numeric
Realized Volatility	Actual range of stock price volatility within a trading day	Numeric

Missing Value Detection Method: The dataset used in this experiment has been checked and confirmed to be complete, with no missing values. Data normalization was performed using the MinMaxScaler function. Data reconstruction was performed as follows:

1. Separate the input (x) and output (y).
2. Convert list-type data to array data.
3. Convert two-dimensional data into three-dimensional data structure.

4.2 Experimental Settings

The proposed framework integrates a Long Short-Term Memory (LSTM) network with a subsequent K-Means clustering module. The LSTM component comprises two hidden layers: the input layer receives sequential data with dimensions determined by the number of time steps and feature size, followed by an LSTM layer with 50 units and a fully connected (Dense) hidden layer. Each hidden layer incorporates a dropout mechanism with a rate of 0.2, effectively mitigating overfitting by randomly deactivating 20% of neurons during training. The LSTM layer produces a 50-dimensional output vector, resulting in a two-dimensional output shape of (batch size, number of units). The layer employs the hyperbolic tangent (tanh) activation function, while the recurrent activation uses the hard sigmoid function. The recurrent kernel is initialized using the orthogonal method, the kernel weights follow a Glorot Uniform distribution, and biases are initialized to zero. The final Dense output layer contains a single neuron, generating a scalar prediction. Model training is conducted using the Adam optimizer, ensuring efficient and stable parameter updates.

4.3 Experimental Settings

In this study, we employ eleven evaluation metrics to comprehensively assess the performance of the proposed strategy:

1. **Return Rate** (R): The return rate measures the percentage gain or loss of a stock investment over a holding period. A positive return indicates profit, while a negative return indicates a loss. The higher the return percentage, the greater the investor's gains.

2. **Annualized Return** (R_{ann}): This metric standardizes the return to an annual basis, facilitating comparison of long-term performance across different investments. It is computed by scaling the cumulative return over the holding period to a yearly rate.

3. **Benchmark Return** (R_{bench}): The benchmark return serves as a reference for performance evaluation, representing the average return of a specific market, sector, or portfolio.

4. **Alpha** (α): Alpha quantifies the excess return of an investment relative to the expected return predicted by asset pricing models, such as the Capital Asset Pricing Model (CAPM). It is defined as:

$$\alpha = R_p - \left(R_f + \beta(R_m - R_f)\right)$$

where R_p is the portfolio return, R_f is the risk-free rate, R_m is the market return, and β is the portfolio beta.

5. **Beta** (β): Beta measures the sensitivity of the stock's returns to market fluctuations, reflecting systematic risk. Specifically,

$$\beta = \frac{\mathrm{Cov}(R_i, R_m)}{\mathrm{Var}(R_m)}$$

where R_i is the stock return and R_m is the market return.

- β=1: stock volatility matches the market.
- β>1: stock is more volatile than the market.
- β<1: stock is less volatile than the market.

6. **Sharpe Ratio** (S): The Sharpe ratio evaluates risk-adjusted return by comparing excess return over the risk-free rate to the standard deviation of returns:

$$S = \frac{R_p - R_f}{\sigma_p}$$

where σ_p is the standard deviation of portfolio returns.

7. **Win Rate** (W): The win rate denotes the proportion of profitable trades within a given period. A higher win rate indicates greater success in achieving profitable transactions.

8. **Profit-Loss Ratio** (PLR): This ratio compares the average profit of winning trades to the average loss of losing trades, reflecting the efficiency of the trading strategy.

9. **Return Volatility** (σ): Volatility measures the dispersion of returns, commonly expressed as the standard deviation, and indicates the investment's risk level.

10. **Information Ratio** (IR): The information ratio assesses active management ability by comparing excess returns over a benchmark to the tracking error:

$$IR = \frac{R_p - R_{bench}}{\sigma_{TE}}$$

Where σ_{TE} is the standard deviation of the active returns.

11. **Maximum Drawdown** (MDD): Maximum drawdown quantifies the largest peak-to-trough decline in the portfolio value over a specified period, representing the worst potential loss:

$$MDD = \max_{t\in[0,T]}\left(\frac{P_{peak} - P_t}{P_{peak}}\right)$$

where P_{peak} is the highest portfolio value before time t.

These metrics together provide a robust framework to evaluate both the profitability and risk characteristics of the investment strategy.

4.4 Data Visualization Analysis

In the study of NVIDIA stock data, the visualization of closing price trends serves as a crucial step for understanding the patterns and behaviors of stock price fluctuations. By processing the raw data and applying diverse plotting techniques, we can intuitively present the temporal trajectory of NVIDIA's closing prices, thereby gaining a clear comprehension of its market performance. Such

visualization not only reveals key characteristics of price movements but also provides a solid foundation for subsequent research and informed decision-making.

As illustrated in Figure 3, the closing price of NVIDIA stock from 2014 to 2025 exhibits an overall upward trend, reflecting the company's sustained growth and market confidence over this period. Notably, starting from 2024, the rate of increase accelerates significantly, indicating a phase of rapid appreciation possibly driven by strong financial results, market expansions, or technological breakthroughs. However, at the beginning of 2025, the stock price shows signs of a downturn, marking a shift in market dynamics that may be influenced by broader economic conditions, industry factors, or investor sentiment.

Figure 3. Closing price trend

Such visual trend analysis is indispensable for understanding the stock's historical performance and volatility, enabling stakeholders to better anticipate future price movements and optimize investment strategies accordingly.

The closing price represents the final trading price of a stock at the end of each trading day and serves as a reliable indicator of market conditions. We analyzed the relationships between the closing price and other indicators, such as average prices and volatility measures computed over different time periods, and found significant correlations: when these indicators increase, the closing price tends to rise as well. To visualize these relationships more intuitively, we employed a color-coded heatmap, where deeper red indicates stronger positive correlation (changes in the same direction) and deeper blue indicates stronger negative correlation (changes in opposite directions). As shown in Figure 3, the results show that the closing price is highly correlated with the 7-day, 14-day, and 30-day average prices and volatility, suggesting that these indicators can effectively assist in predicting price trends and provide valuable insights for investment analysis.

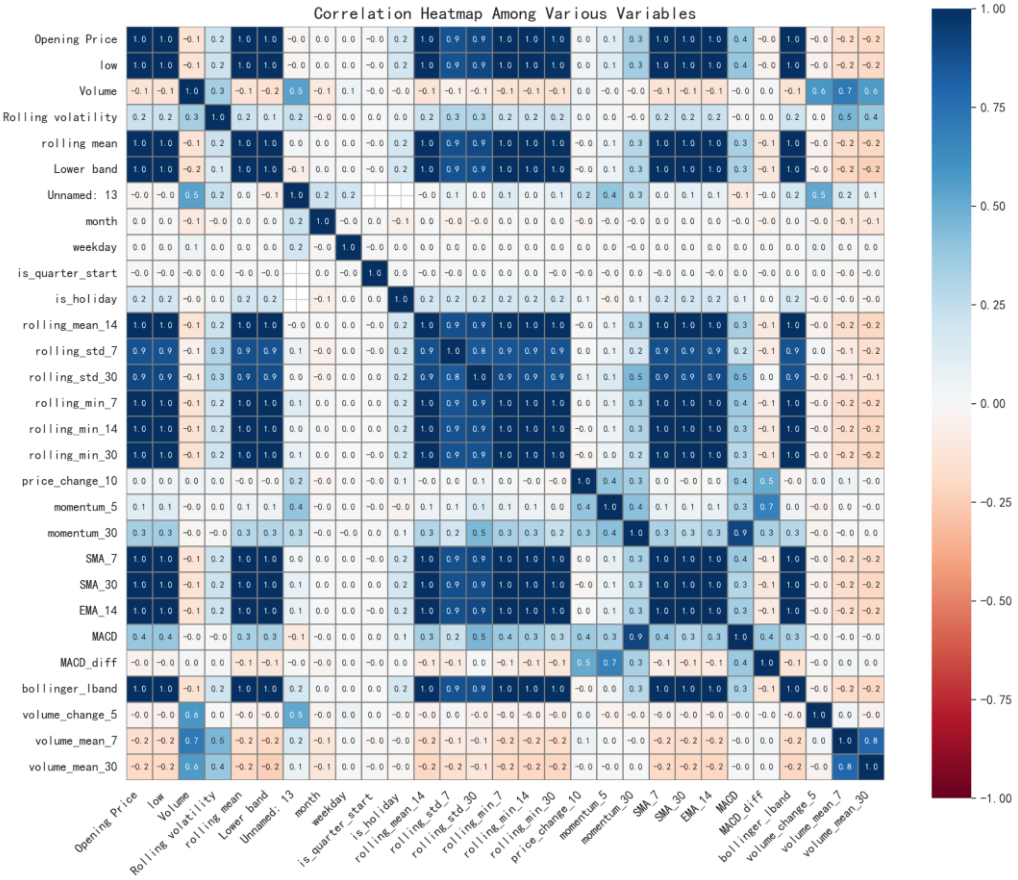

Figure 4. Correlation Heatmap Among Various Variables

This heatmap intuitively reveals the intrinsic relationships among various technical indicators for NVIDIA stock. It is evident that the closing price exhibits a strong positive correlation with moving averages of different periods, especially the 7-day, 14-day, and 30-day moving averages, indicating that the moving average system effectively tracks the price trend. Additionally, significant correlations among moving averages across different time frames suggest consistency between short-term and long-term price trends. In contrast, volume-related indicators appear relatively independent, with short-term volume fluctuations showing weak association with the moving average system. Notably, the MACD indicator demonstrates a close relationship with short-term moving averages, confirming its practical value in assessing price momentum. Overall, these findings provide critical insights for constructing quantitative trading strategies, particularly emphasizing the combined use of moving average systems and momentum indicators.

4.5 Method Comparison

To provide a more intuitive demonstration of the forecasting performance, this experiment employs both the RNN and ARIMA models to predict the same experimental dataset. The visualization of the prediction results is presented in the corresponding Figure 5.

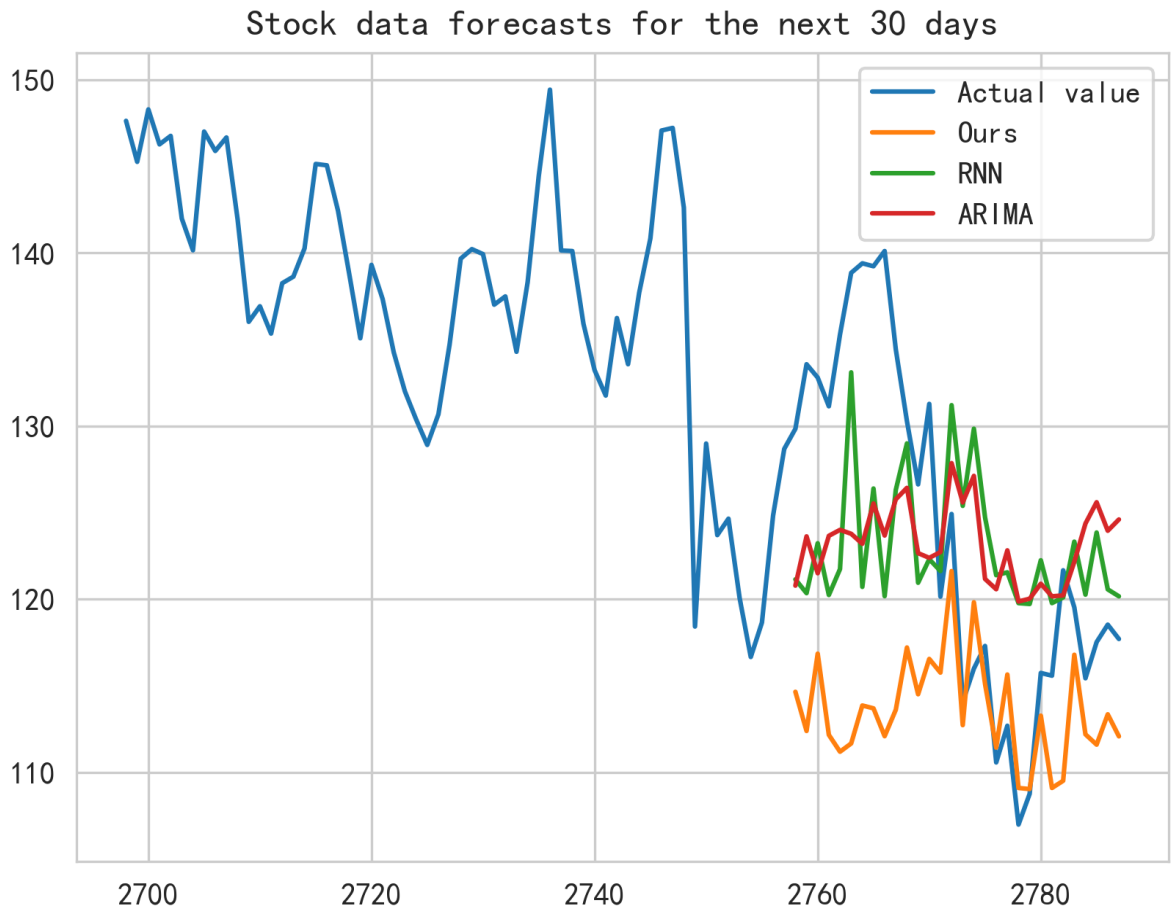

Figure 5. The visualization of the prediction results

When utilizing data within the same temporal range for stock price forecasting, the LSTM model demonstrates superior predictive performance particularly in the latter period, characterized by a smoother and more stable forecast curve. In contrast, the RNN model exhibits greater volatility in its predictions, with numerous extreme points, which may result in shorter holding periods and comparatively lower returns in practical trading scenarios. Meanwhile, the ARIMA model, as a traditional time series approach, underperforms when dealing with long-span data, often producing predictions that approximate a flat line, thereby failing to capture the intricate fluctuations inherent in stock prices. In summary, due to its higher accuracy and stability, the LSTM model proves to be a more reliable tool for stock price prediction tasks. These findings highlight the critical importance of selecting appropriate forecasting models for effective investment decision-making and risk management.

4.6 Comparison of Results from Quantitative Investment Methods

We focus on analyzing the performance variations of the LSTM model when trained with datasets of different lengths. By comparing key financial metrics over two periods, 2020-2022 and 2021-2022, we gain a clearer understanding of how training data size affects model efficacy, as well as the model's robustness and risk-return profile in practical investment decision-making, as shown in Table 2.

Table 2. Performance Results of Quantitative Methods Based on a Single Model with Different Training Periods

Year	2020-2022	2021-2022
Return	-8.74%	-17.94%
Annualized Return	-6.53%	-13.59%
Benchmark Return	-21.63%	-22.94%
Alpha	0.07	-0.01
Beta	0.69	0.72
Sharpe Ratio	-0.34	-0.75
Win Rate	0.47	0.51
Profit-Loss Ratio	1.18	0.94
Return Volatility	21.52%	20.64%
Information Ratio	0.05	0.02
Max Drawdown	22.81%	24.8%

The first set of experiments reveals that reducing the training period from two years to one year results in noticeable changes in the performance of the LSTM-based quantitative investment strategy. Specifically, the alpha value decreases from 0.09 to 0.02, indicating a reduction in the model's excess returns relative to the overall market. The beta value remains relatively stable, suggesting that the selected stocks maintain consistent market volatility. The Sharpe ratio declines from a positive 0.22 to a negative -0.08, reflecting diminished risk-adjusted returns. Furthermore, the decrease in the information ratio indicates a reduction in excess returns per unit of risk. These metric changes further corroborate the superiority of the LSTM model when handling longer time-series data compared to alternative approaches. The LSTM-based quantitative investment model effectively integrates stock selection and risk management strategies, demonstrating strong robustness and achieving higher excess returns at comparable risk levels. This implies that investors utilizing this model can attain improved investment outcomes under controlled risk conditions.

The candidate stock pool is constructed by selecting stocks predicted to increase by the LSTM model, which demonstrates a noticeable reduction in both risk control and stock selection performance compared to the integrated quantitative approach. The experimental results are summarized in the Table 3.

Table 3. Performance Results of Quantitative Methods Based on LSTM Model with Different Training Pe-riods

Year	2020-2022	2021-2022
Return	13.05%	-1.88%
Annualized Return	5.63%	-0.82%
Benchmark Return	-21.63%	-22.94%
Alpha	0.09	0.02
Beta	0.25	0.23
Sharpe Ratio	0.22	-0.08
Win Rate	0.22	0.23
Profit-Loss Ratio	0.94	0.88
Return Volatility	20.84%	20.81%
Information Ratio	0.05	0.03
Max Drawdown	19.35%	19.34%

As observed from the Table 3, the returns under the same dataset conditions decreased by approximately 20%. The alpha value of 0.07 on the 2020-2022 sample set indicates that the LSTM model's stock selection can still generate excess returns above the market benchmark. The beta values in both periods are below 1, suggesting that the stocks selected by this strategy exhibit lower volatility relative to the overall market. Overall, despite some decline in risk control and stock selection efficacy compared to the comprehensive quantitative method, the LSTM-based stock selection strategy demonstrates certain advantages. Therefore, incorporating the LSTM model's stock selection outcomes into investment decision-making processes could potentially achieve market outperformance.

5. Conclusions

This study addresses the inherent limitations of conventional neural networks in capturing long-term dependencies and nonlinear dynamics in financial time series, as well as the inefficacy of single-model approaches in effectively ranking and recommending target stocks. We propose an integrated deep learning framework that synergistically combines Long Short-Term Memory (LSTM) networks with K-Means clustering to enhance both stock price prediction and portfolio construction. The hierarchical LSTM architecture is designed to model complex temporal dependencies and nonlinear patterns in stock price movements, while the K-Means clustering module facilitates the identification of stocks with similar return and volatility profiles, thereby enabling more precise stock selection. The framework further incorporates a multi-factor scoring mechanism to filter candidate stocks, which are then subjected to rigorous training and forecasting. Investment decisions are informed by model predictions and complemented with a fixed-percentage stop-loss and take-profit risk management strategy, mitigating downside risks and preserving portfolio gains. Empirical results demonstrate that this integrated quantitative investment approach outperforms traditional models,

delivering superior excess returns and enhanced stability under comparable risk exposures. The methodology effectively balances predictive accuracy, risk control, and portfolio optimization, providing a robust tool for investors seeking consistent abnormal returns in dynamic equity markets.

Future research directions aim to further strengthen the proposed quantitative investment framework. First, integrating alternative deep learning architectures, such as Transformer-based models, may improve the modeling of long-range dependencies and complex nonlinearities in financial time series. Second, expanding the factor selection process to incorporate alternative data sources, including sentiment information extracted from news articles and social media, could enrich predictive capabilities and enhance portfolio optimization. Third, developing adaptive risk management strategies that dynamically adjust stop-loss and take-profit thresholds according to market volatility and regime shifts may enhance strategy robustness. Finally, extending the framework to multi-asset portfolios and exploring cross-asset correlations could improve diversification benefits and risk-adjusted returns. Collectively, these advancements will contribute to the development of more flexible, accurate, and practical quantitative investment systems capable of navigating increasingly complex financial markets.

Acknowledgements

This article received no financial or funding support.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Lu, W., Li, J., Wang, J. and Qin, L. A CNN-BiLSTM-AM method for stock price prediction. Neural Computing and Applications, 2021, 33(10), 4741–4753.

[2] Ariyo, A.A., Adewumi, A.O. and Ayo, C.K. Stock price prediction using the ARIMA model. In: 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, IEEE, 2014, 106–112.

[3] Soni, P., Tewari, Y. and Krishnan, D. Machine learning approaches in stock price prediction: a systematic review. In: Journal of Physics: Conference Series, IOP Publishing, 2022, 012065.

[4] Schöneburg, E. Stock price prediction using neural networks: A project report. Neurocomputing, 1990, 2(1), 17–27.

[5] Hu, Z., Zhao, Y. and Khushi, M. A survey of forex and stock price prediction using deep learning. Applied System Innovation, 2021, 4, 9.

[6] Yu, P. and Yan, X. Stock price prediction based on deep neural networks. Neural Computing and Applications, 2020, 32(6), 1609–1628.

[7] Leung, C.K.-S., MacKinnon, R.K. and Wang, Y. A machine learning approach for stock price prediction. In: Proceedings of the 18th International Database Engineering & Applications Symposium, 2014, 274–277.

[8] Obthong, M., Tantisantiwong, N., Jeamwatthanachai, W. and Wills, G. A survey on machine learning for stock price

prediction: algorithms and techniques. 2020.

[9] Sewell, M. History of the efficient market hypothesis. Rn, 2011, 11(4), 04.

[10] Ghallabi, F., Souissi, B., Du, A.M. and Ali, S. ESG stock markets and clean energy prices prediction: insights from advanced machine learning. International Review of Financial Analysis, 2025, 97, 103889.

[11] Alshater, M.M., Kampouris, I., Marashdeh, H., Atayah, O.F. and Banna, H. Early warning system to predict energy prices: the role of artificial intelligence and machine learning. Annals of Operations Research, 2025, 345(2), 1297–1333.

[12] Yao, Y. Stock price prediction using an improved transformer model: capturing temporal dependencies and multi-dimensional features. Journal of Computer Science and Software Applications, 2025, 5(2).

[13] Ghritlahre, H.K. and Verma, M. Solar air heaters performance prediction using multi-layer perceptron neural network – a systematic review. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2025, 47(1), 7682–7699.

[14] LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. Nature, 2015, 521(7553), 436–444.

[15] Hastie, T.J. Generalized additive models. Statistical Models in S, 2017, 249–307.

[16] Graves, A. Long short-term memory. Supervised Sequence Labelling with Recurrent Neural Networks, 2012, 37–45.

[17] Hu, Z., Shen, B., Hu, Y. and Zhao, C. Research on stock price forecast of general electric based on mixed CNN-LSTM model. arXiv preprint arXiv:2501.08539, 2025.

[18] Wu, J.M.-T., Li, Z., Herencsar, N., Vo, B. and Lin, J.C.-W. A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. Multimedia Systems, 2023, 29(3), 1751–1770.

[19] Gao, Y., Wang, R. and Zhou, E. Stock prediction based on optimized LSTM and GRU models. Scientific Programming, 2021, 2021(1), 4055281.

[20] Bhandari, H.N., Rimal, B., Pokhrel, N.R., Rimal, R., Dahal, K.R. and Khatri, R.K. Predicting stock market index using LSTM. Machine Learning with Applications, 2022, 9, 100320.

[21] Sang, S. and Li, L. A novel variant of LSTM stock prediction method incorporating attention mechanism. Mathematics, 2024, 12(7), 945.

[22] Prabakar, S., Kumar, P., Banu, N. and Raj, A. Strategic integration for future selection-LSTM stock prediction algorithm based on the Internet of Things (IoT). In: 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET), IEEE, 2024, 1–6.

[23] Usmani, S. and Shamsi, J.A. LSTM based stock prediction using weighted and categorized financial news. PLoS One, 2023, 18(3), e0282234.

[24] Gülmez, B. Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. Expert Systems with Applications, 2023, 227, 120346.

[25] Baek, H. A CNN-LSTM stock prediction model based on genetic algorithm optimization. Asia-Pacific Financial Markets, 2024, 31(2), 205–220.

[26] Shi, Y., Wang, Y., Qu, Y. and Chen, Z. Integrated GCN-LSTM stock prices movement prediction based on knowledge-incorporated graphs construction. International Journal of Machine Learning and Cybernetics, 2024, 15(1), 161–176.

[27] Ali, M., Khan, D.M., Alshanbari, H.M. and El-Bagoury, A.A.-A.H. Prediction of complex stock market data using an improved hybrid EMD-LSTM model. Applied Sciences, 2023, 13(3), 1429.

[28] Tian, L., Feng, L., Yang, L. and Guo, Y. Stock price prediction based on LSTM and LightGBM hybrid model. The Journal of Supercomputing, 2022, 78(9), 11768–11793.

[29] Meher, B.K., Hawaldar, I.T., Spulbar, C.M. and Birau, F.R. Forecasting stock market prices using mixed ARIMA model: a case study of Indian pharmaceutical companies. Investment Management and Financial Innovations, 2021, 18(1), 42–54.

[30] Mashadihasanli, T. Stock market price forecasting using the ARIMA model: an application to Istanbul, Turkiye. İktisat Politikası Araştırmaları Dergisi, 2022, 9(2), 439–454.

[31] Dadhich, M., Pahwa, M.S., Jain, V. and Doshi, R. Predictive models for stock market index using stochastic time series ARIMA modeling in emerging economy. In: Advances in Mechanical Engineering: Select Proceedings of CAMSE 2020, Springer, 2021, 281–290.

[32] Ashok, A. and Prathibhamol, C. Improved analysis of stock market prediction: (ARIMA-LSTM-SMP). In: 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), IEEE, 2021, 1–5.

[33] Kobiela, D., Krefta, D., Król, W. and Weichbroth, P. ARIMA vs LSTM on NASDAQ stock exchange data. Procedia Computer Science, 2022, 207, 3836–3845.

[34] Pokou, F., Sadefo Kamdem, J. and Benhmad, F. Hybridization of ARIMA with learning models for forecasting of stock market time series. Computational Economics, 2024, 63(4), 1349–1399.

[35] Verma, S., Sahu, S.P. and Sahu, T.P. Ensemble approach for stock market forecasting using ARIMA and LSTM model. In: Proceedings of Third International Conference on Intelligent Computing, Information and Control Systems: ICICCS 2021, Springer, 2022, 65–80.

[36] Zolfaghari, M. and Gholami, S. A hybrid approach of adaptive wavelet transform, long short-term memory and ARIMA-GARCH family models for the stock index prediction. Expert Systems with Applications, 2021, 182, 115149.

[37] Friday, I.K., Pati, S.P., Mishra, D., Mallick, P.K. and Kumar, S. CAGTRADE: predicting stock market price movement with a CNN-attention-GRU model. Asia-Pacific Financial Markets, 2024, 1–26.

[38] Das, N., Sadhukhan, B., Bhakta, S.S. and Chakrabarti, S. Integrating EEMD and ensemble CNN with X (Twitter) sentiment for enhanced stock price predictions. Social Network Analysis and Mining, 2024, 14(1), 29.

[39] Wu, Y., Sun, M., Zheng, H., Hu, J., Liang, Y. and Lin, Z. Integrative analysis of financial market sentiment using CNN and GRU for risk prediction and alert systems. In: 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS), IEEE, 2024, 410–415.

[40] Jagadesh, B., Ramesh, V., Mohan, S. and Kumar, R. Enhanced stock market forecasting using dandelion optimization-driven 3D-CNN-GRU classification. Scientific Reports, 2024, 14(1), 20908.

[41] Luo, A., Zhong, L., Wang, J., Wang, Y., Li, S. and Tai, W. Short-term stock correlation forecasting based on CNN-BiLSTM enhanced by attention mechanism. IEEE Access, 2024, 12, 29617–29632.

[42] Kim, W., Park, J., Lee, S., Cho, H. and Han, J. Prediction of index futures movement using TimeGAN and 3D-CNN: empirical evidence from Korea and the United States. Applied Soft Computing, 2025, 171, 112748.

[43] Somkunwar, R., Rao, J. and Varvante, N. Stock value prediction accuracy enhancement using CNN and multiple linear regression for NIFTY. In: 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), IEEE, 2024, 1–7.

[44] Khattak, B.H.A., Shafi, I., Rashid, C.H., Safran, M., Alfarhood, S. and Ashraf, I. Profitability trend prediction in crypto financial markets using Fibonacci technical indicator and hybrid CNN model. Journal of Big Data, 2024,

11(1), 58.

[45] Wu, C.H. Eco-economic predictions: applying QPSO-BiLSTM and attention mechanisms for accurate renewable energy forecasting. Journal of Management Science and Operations, 2024, 2(4), 28–47. DOI: 10.30210/JMSO.202402.011.

[46] Tsai, S. Advancing credit card fraud detection: a deep learning approach for improved risk management. International Journal of Management and Organization, 2025, 3(1), 1–15.

[47] Shu, M. Utilizing transfer learning for deep learning based image classification. Journal of Intelligence Technology and Innovation, 2025, 3(1), 58–73.

[48] Wang, S., Jiang, R., Wang, Z. and Zhou, Y. Deep learning-based anomaly detection and log analysis for computer networks. Journal of Information and Computing, 2024, 2(2), 34–63. DOI: 10.30211/JIC.202402.005.