Robustness Investigation of ML Models for Stock Price Prediction During Market Volatility: A Case Study of the Japanese Stock Market Amidst the COVID-19 Pandemic

Zhuo Chen*

College of Letter and Science, University of Wisconsin Madison, Madison, United States of America

* Corresponding Author Email: zchen962@wisc.edu

Abstract. In the evolving realm of financial markets, precise stock price prediction remains a pivotal challenge. This paper investigates the efficacy of traditional machine learning models for stock price forecasting, particularly during the high volatility phase induced by the COVID-19 pandemic, using the Japanese stock market as a case study. This study utilized a dataset from the JPX Tokyo Stock Exchange Prediction competition, selecting two stocks at random to avoid bias. After standardizing the data for consistency, this study employed linear regression and decision trees to project stock prices, comparing their performance during stable periods and the economic upheaval of the pandemic. The analysis revealed that both models' predictive capabilities were compromised during the pandemic, with the decision tree model particularly underperforming. In normal times, linear regression showed improved accuracy for stock 1301 in 2021, while for stock 6758, the model's performance declined. These findings underscore the necessity for robust models that can withstand market turbulences like those observed during the COVID-19 crisis. The research contributes to both academic discussions on financial forecasting and practical strategies for market participants navigating complex stock environments. Future work will focus on developing more resilient forecasting methods to address these challenges.

Keywords: Machine learning, Linear regression, Decision trees, COVID-19 economic impact, Model prediction

1. Introduction

In the complex and changing financial market landscape, the ability to accurately forecast stock prices has long been one of the main research directions for investors, analysts, and economists. The core of stock price prediction is to predict the future value of a company's stock based on various factors such as market trends, economic indicators and company performance. This task is not only crucial for making informed investment decisions, but also plays a key role in guiding corporate strategy and economic policy making. The importance of accurate stock price forecasting cannot be overstated, as it enables market participants to measure potential investment returns, manage risks, and allocate resources more efficiently.

The development for forecasting stock price prediction approaches has seen significant progress, from classical financial models to sophisticated machine learning algorithms. Initially, financial analysts relied on fundamental and technical analysis to estimate stock prices. For instance, Islam used Artificial Neural Network (ANN) and Geometric Brownian Motion to predict the in his article Comparison of Financial Models for Stock Price Prediction next-day stock prices [1]. Mohan also mentioned the use of news articles related to companies to predict stock market prices in an article [2]. Fundamental analysis involves assessing a company's financial health and market position, while technical analysis focuses on patterns of stock price volatility and trading volume. The field has witnessed the integration of statistical models over time. For example, Khanderwal et al. used the Autoregressive Integrated Moving Average (ARIMA) model, which is more detailed in time series, to predict stock prices [3]. And the advent of machine learning has revolutionized stock price forecasting, providing new dimensions of analysis and accuracy. Machine learning models that can process large amounts of data and recognize complex patterns have shown promising results in this

area. These models range from linear regression, which is a straightforward and powerful tool for capturing linear relationships, to more complex nonlinear models such as decision trees and neural networks, which are adept at dealing with the complexity of financial data. They have demonstrated effectiveness in many tasks including medical prediction, education and cloud computing [4-8]. For instance, Sun et al. introduced a comprehensive framework designed for the autonomous activation of cells and the tailored allocation of physical resources based on machine learning methods [4]. Wu et al. proposed a Deep Learning-based BERT Model in Sentiment Analysis and achieved excellent performance [5]. In the domain of stock price prediction, Sunny et al. used Long Short-Term Memory Network (LSTM) and Bi-Directional LSTM Model to predict stock prices [9]. Wu et al. also proposed the CNN-LSTM model to predict the stock price [10]. Most research, however, has focused on stock price predictions in normal times. During this time, stock prices tend to stabilize and fluctuate less. During the epidemic, stock prices will fluctuate relatively large due to many factors, such as business shutdown and a decline in real economic activity. This article will focus on the performance of machine learning stock prediction methods in the case of high stock price volatility during the pandemic.

The COVID-19 crisis has had a profound impact on global financial markets, increasing volatility and uncertainty. This period provides a unique context for assessing the robustness and adaptability of forecasting models under stress. This research aims to do this by comparing the forecasting performance of linear and nonlinear machine learning models under different market conditions, especially during normal times and the unprecedented COVID-19 pandemic. This article will evaluate the performance of these models in predicting stock prices in stable and crisis periods using the Japanese stock market, which plays an important role in the global financial system. By combining linear regression and decision tree algorithms, this research navigates both simple and complex areas of model design.

This article has two objectives: first, to evaluate the effectiveness of linear and nonlinear models for stock price forecasting under normal market conditions, and second, to evaluate their performance during the economic turmoil brought about by the COVID-19 pandemic. Through this comparative analysis, this study aims to reveal the adaptability and reliability of these models under different market dynamics. This effort not only contributes to the academic discourse on financial forecasting, but also provides practical implications for investors, analysts, and policy makers navigating the complexities of today's stock markets.

2. Method

2.1. Dataset Preparation

The dataset used came from a Kaggle competition called JPX Tokyo Stock Exchange Prediction [11]. The dataset contains identifiers for 1,865 stocks from January 2017 through December 2022, as well as weekday opening prices, closing prices, day highs and day lows, and day volume. 2 stocks were selected randomly for analysis. The two stock numbers chosen are 1301,6758. These two stocks were selected without knowing the company name, in order to ensure the randomness of the selection of experimental samples.

In a preliminary data check, it was found that data for a small number of trading days was missing. For these missing data, the forward fill method is used, that is, the data of the previous trading day is replaced, to ensure the continuity of the data. In addition, the dataset was examined for outliers, such as price jumps due to system errors, but no significant outliers were found. In order to reduce the influence of different dimensional features, all numerical features are standardized. Specifically, each feature is scaled using its mean and standard deviation to ensure that all features have the same scale. The final data set is divided into a training set and a test set. This experiment used the natural order of the time series, selecting the first 80% of the data as the training set for the training and validation of the model, and the last 20% as the test set for evaluating the predictive performance of the model.

In this experiment, the closing price, opening price, the highest price of the day and the lowest price of the day will be selected as feature variables to predict the opening price of the stock ten days later.

2.2. Machine Learning Models

2.2.1. Linear Regression

Linear regression is a statistical method used to establish a linear relationship between one or more independent variables (features) and dependent variables (targets). In the context of stock price forecasting, linear regression models attempt to find a linear relationship between features (such as historical prices, moving averages, trading volumes, etc.) and future stock prices. The linear regression model is based on an equation of the following form:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \in \quad (1)$$

Among them, y is the dependent variable, the stock price that this experiment wants to predict., $X_1, X_2$is the independent variable, that is, the factor that may affect the stock price. $\beta_0$is the intercept term. $\beta_1, \beta_2, \beta_n$ is the coefficient of each feature, indicating the impact of each feature on the stock price. ϵ is the error term, representing random factors that the model cannot account for. To construct the linear regression model in this experiment, the opening price, closing price and trading volume were selected as the selected relevant features. Mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) were used to evaluate the performance of the model.

2.1.2. Decision Tree

Decision trees are a popular and powerful machine learning algorithm used for classification and regression tasks. It solves prediction problems by splitting data through a series of rules. Decision trees mimic the "yes/no" problem in the human decision process by predicting the value of the target variable through a series of decisions from the root node to the leaf node. In this experiment, the relevant features needed to build the decision tree model include opening price, closing price, and trading volume. Decision trees construct tree structures by recursively splitting data sets. Each segment (or node) is based on features and thresholds that best reduce data uncertainty. For stock price forecasting, the goal is to build a tree model with middle nodes representing the predicted price range or specific value. The mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) were used to evaluate the performance of the model.

3. Results and Discussion

After applying linear regression and decision tree models to the datasets, it can be observed different performance metrics for the years 2019 and 2021 shown in Table 1, corresponding to stock numbers 1301 and 6758. In terms of the linear regression model, in 2019, the MSE for stock 1301 was 7919.21, and the MAE was 73.68, with an RMSE of 88.99. In contrast, for 2021, the MSE for the same stock number decreased to 4474.994, the MAE reduced to 51.94, and the RMSE lowered to 66.89. Fig. 1-Fig. 8 shows the predicted price curve compared to the actual price.

Table 1. The performance of different model in various time and stock

Method	Model							
	Linear Regression				Decision Tree			
	2019-1301	2021-1301	2019-6758	2021-6758	2019-1301	2021-1301	2019-6758	2021-6758
MSE	7919.21	4474.94	77749.69	612522.76	13672.19	3651.61	101570.38	989745.21
MAE	73.68	51.94	221.02	646.13	93.43	44.00	232.47	741.56
RMSE	88.99	66.89	278.83	782.64	116.93	60.43	318.70	994.86

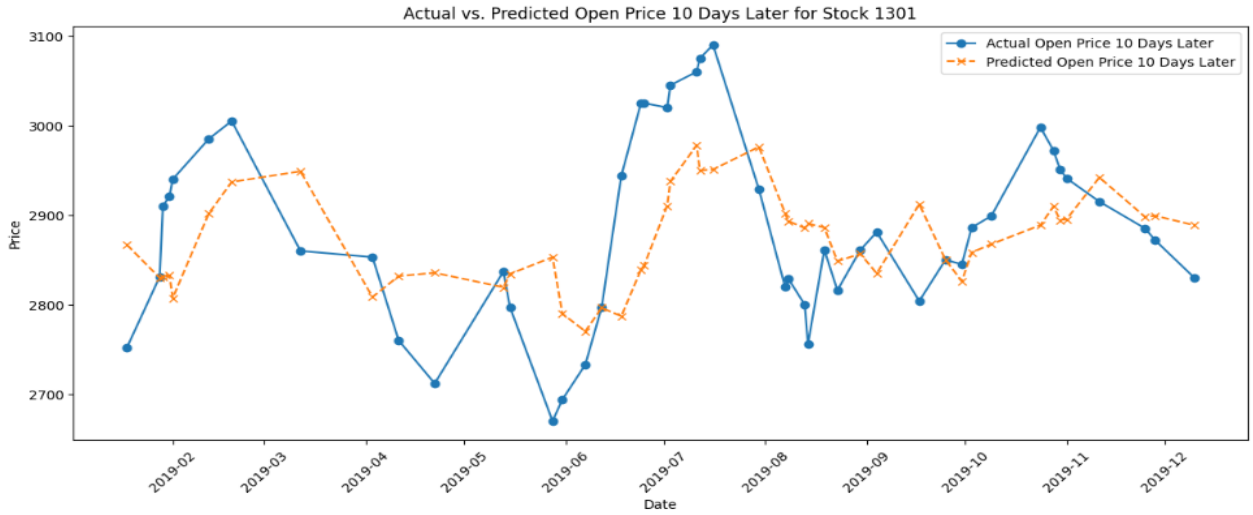

Fig. 1 Actual price vs predicted open prices for stock 1301 in 2019 using Linear regression (Photo/Picture credit : Original).

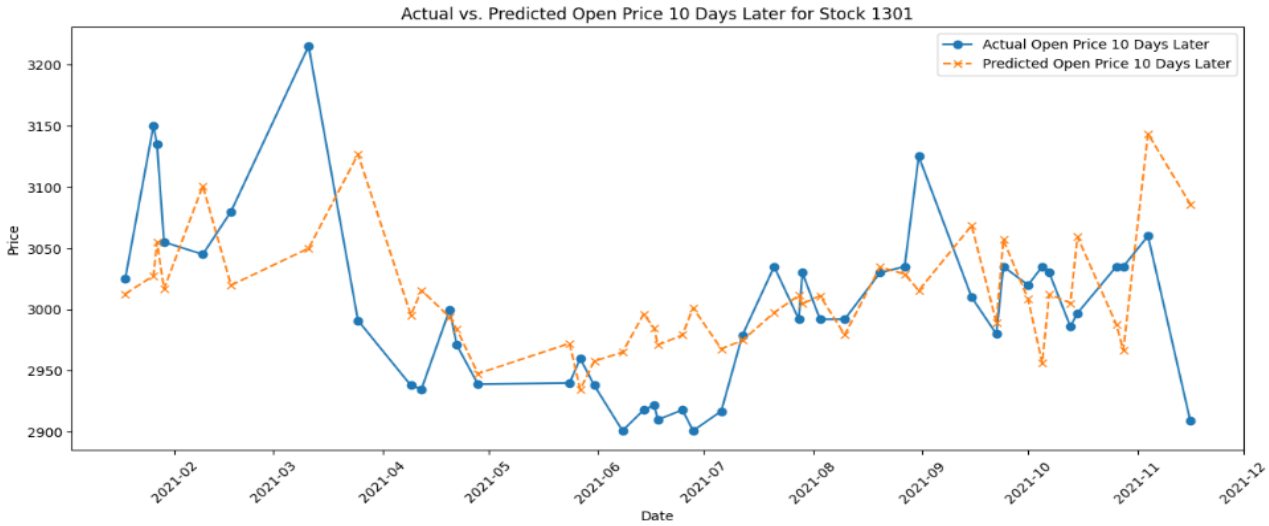

Fig. 2 Actual price vs predicted open prices for stock 1301 in 2021 using Linear regression (Photo/Picture credit : Original).

Fig. 3 Actual price vs predicted open prices for stock 6758 in 2019 using Linear regression (Photo/Picture credit : Original).

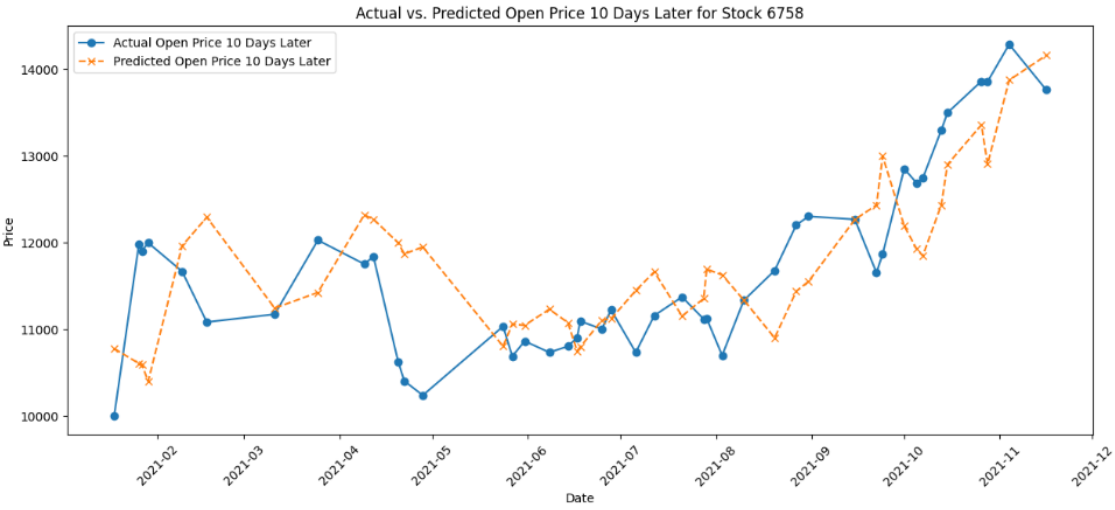

Fig. 4 Actual price vs predicted open prices for stock 6758 in 2021 using Linear regression (Photo/Picture credit : Original).

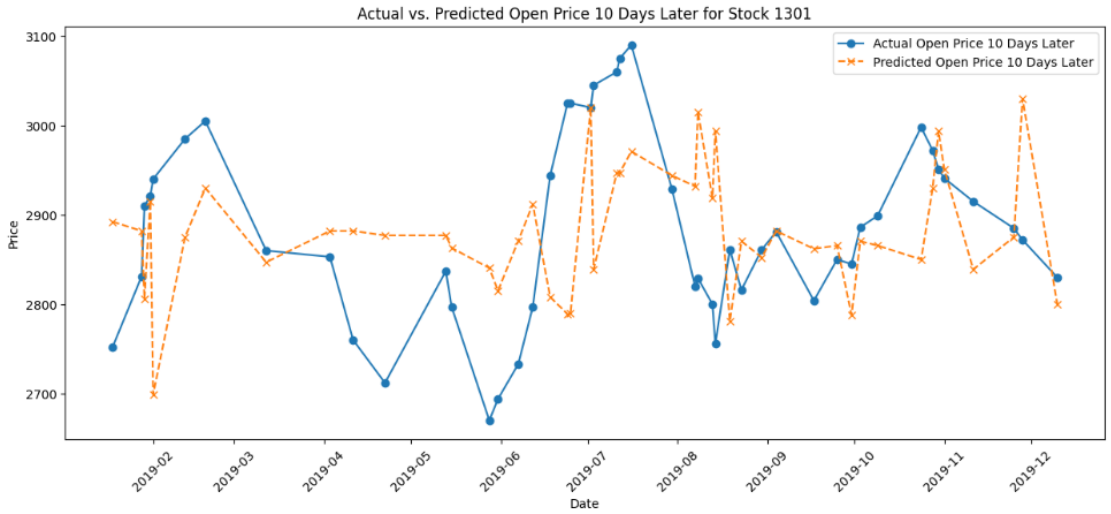

Fig. 5 Actual price vs predicted open prices for stock 1301 in 2019 using Decision tree (Photo/Picture credit : Original).

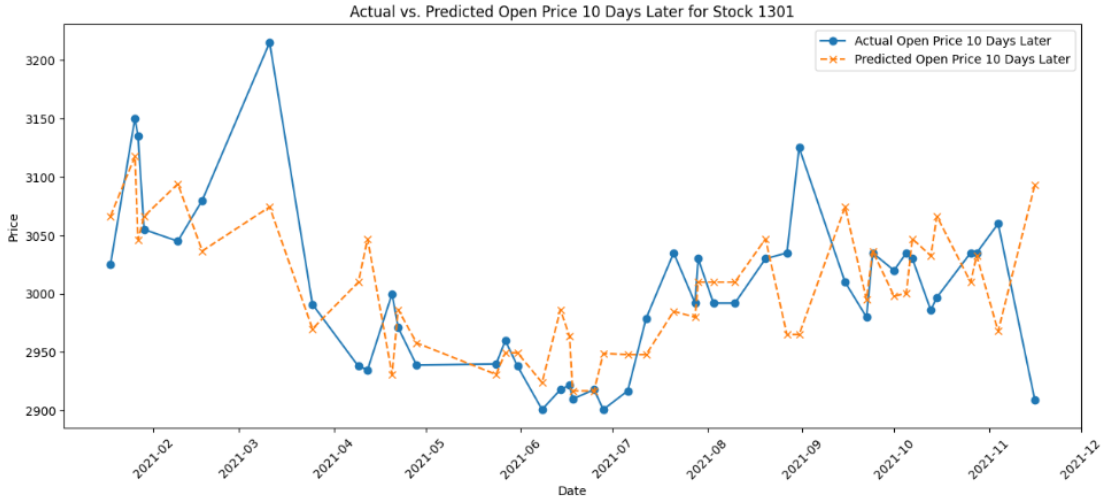

Fig. 6 Actual price vs predicted open prices for stock 1301 in 2021 using Decision tree (Photo/Picture credit : Original).

Fig. 7 Actual price vs predicted open prices for stock 6758 in 2019 using Decision tree (Photo/Picture credit : Original).

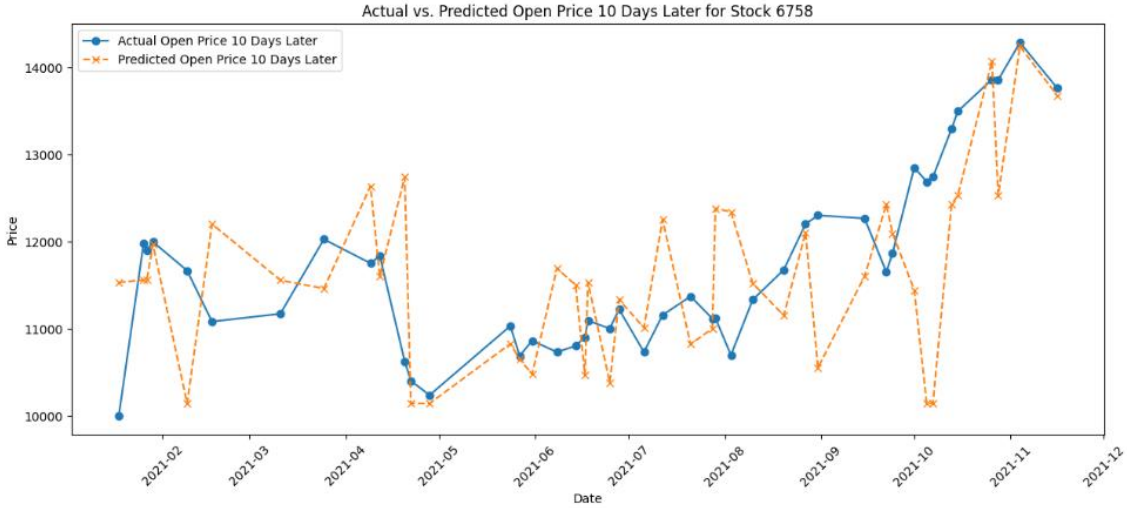

Fig. 8 Actual price vs predicted open prices for stock 6758 in 2021 using Decision tree (Photo/Picture credit : Original).

When examining the decision tree model, the MSE for stock 1301 in 2019 was significantly higher at 13672.19, with an MAE of 93.43 and an RMSE of 116.93. However, in 2021, there was a dramatic increase in MSE to 3651.61, while MAE decreased to 44.00, and RMSE to 60.43, indicating a varied model performance across the two years. For stock number 6758, both the linear regression and decision tree models yielded higher error metrics in 2021, suggesting challenges in model predictions for that year.

It is noteworthy that the decision tree model for stock 6758 in 2021 resulted in an exceptionally high MSE of 989745.21, an MAE of 741.56, and an RMSE of 994.86, indicating that the decision tree model's predictions were significantly off from the actual values for this particular stock in that year. The results demonstrate the importance of model selection and validation for accurate stock price prediction, as the chosen model can significantly affect the outcome, as reflected by these error metrics.

For stock 1301, the Linear Regression model's performance improved in 2021 as seen by lower MSE, MAE, and RMSE values compared to 2019. This suggests a better fit of the model to the data or possibly less volatility or more predictable behavior in the stock price movements during that year. For stock 6758, however, the error metrics in 2021 increased significantly for the Linear Regression model, which indicates a decline in predictive performance. This difference could be due to the volume of stock 1301 is lesser than the stock 6758.

Turning to the Decision Tree model, there's a different pattern. For stock 1301, the Decision Tree has lower error metrics in 2021 than in 2019, indicating a more accurate fit for that period. However, for stock 6758 in 2021, the error metrics are substantially higher than in 2019, suggesting severe overfitting or an inability of the model to generalize from its training data to the actual price movements. This suggests that under the influence of COVID-19, the traditional training model's performance in forecasting is far worse than that without the influence of COVID-19.

4. Conclusion

This study used traditional machine learning models to predict stock prices with and without the impact of COVID-19. In this experiment, Linear regression and decision tree were used to try to compare the difference with or without influence, and two stocks were randomly selected for experimental evaluation. The experimental results demonstrate that the prediction performance of the two traditional models under the influence of COVID-19 is lower than that without the influence, especially the decision tree model. In the future, further study plans to develop better models for stock forecasting to solve this problem once and for all.

References

[1] Islam MR, Nguyen N. Comparison of Financial Models for Stock Price Prediction. Journal of Risk and Financial Management, 2020, 13(8):181.

[2] Mohan S, Mullapudi S, Sammeta S, Vijayvergia P, Anastasiu DC. Stock Price Prediction Using News Sentiment Analysis. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 2019, pp. 205-208.

[3] Khanderwal S, Mohanty D. Stock Price Prediction Using ARIMA Model. International Journal of Marketing & Human Resource Research, 2021, 2(2): 98-107.

[4] Sun G, Zhan T, Owusu BG, Daniel AM, Liu G, Jiang W. Revised reinforcement learning based on anchor graph hashing for autonomous cell activation in cloud-RANs. Future Generation Computer Systems. 2020 Mar 1;104:60-73.

[5] Wu Y, Jin Z, Shi C, Liang P, Zhan T. Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis. arXiv preprint arXiv:2403.08217. 2024 Mar 13.

[6] Qiu Y, Wang J, Jin Z, Chen H, Zhang M, Guo L. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control. 2022 Feb 1;72:103323.

[7] Qiu Y, Chang CS, Yan JL, Ko L, Chang TS. Semantic segmentation of intracranial hemorrhages in head CT scans. In2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) 2019 Oct 18 (pp. 112-115). IEEE.

[8] Zhao F, Yu F, Trull T, Shang Y. A new method using LLMs for keypoints generation in qualitative data analysis. In2023 IEEE Conference on Artificial Intelligence (CAI) 2023 Jun 5 (pp. 333-334). IEEE.

[9] Sunny MAI, Maswood MMS, Alharbi AG. Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2020, pp. 87-92.

[10] Wu JMT, Li Z, Herencsar N et al. A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. Multimedia Systems, 2023, 29: 1751–1770.

[11] Kaggle. JPX Tokyo Stock Exchange Prediction. https://www.kaggle.com/competitions/jpx-tokyo-stock-exchange-prediction, 2023.