



Comparative Analysis of Statistical, Machine Learning, and Deep Learning Models for Real Estate Stock Market Forecasting

PHAN CHI CUONG¹, NGUYEN LE KHANG², AND LE PHAM QUOC BAO³

¹Faculty of Information Systems, University of Information Technology, (e-mail: 21520673@gm.uit.edu.vn)

²Faculty of Information Systems, University of Information Technology, (e-mail: 21520960@gm.uit.edu.vn)

³Faculty of Information Systems, University of Information Technology, (e-mail: 21521849@gm.uit.edu.vn)

ABSTRACT This study shows a detailed comparative analysis of statistical models, machine learning, and deep learning models for predicting real estate stock prices. We check the predictive capabilities of different models like the Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX), Linear Regression, Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Random Forest (RF), Fast Fourier Transform (FFT), and TimesNet. Using a carefully prepared dataset of historical stock prices and technical indicators, we train and test these models with MAPE, RMSE and Mean Directional Accuracy (MDA). Our results show that deep learning models, like LSTM and GRU, perform better than traditional time series and machine learning models in catching the complex behaviors of real estate stock markets. Also, adding technical indicators as exogenous variables in ARIMAX, using frequency domain analysis with FFT, and applying Linear Regression all improve forecasting accuracy. This research gives useful insights for investors and financial analysts aiming to make informed decisions in the real estate stock market.

INDEX TERMS Real Estate Stock Prices, Statistical Models, Machine Learning, Deep Learning, Autoregressive Integrated Moving Average with Exogenous variable (ARIMAX), Linear Regression, Long Short-Term Memory (LSTM) Networks, Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Random Forest (RF), Fast Fourier Transform (FFT), TimesNet, Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE)

I. INTRODUCTION

The stock market forecasting of real estate companies is a difficult task due to the many economic, financial, and political factors that affect it. Accurate predictions in this area can lead to better investment strategies and more stable economic planning. Traditional methods like time series analysis and machine learning have been used for stock market predictions for a long time. But the rise of deep learning has brought new models that can capture more complex patterns and relationships in the data.

In this paper, we do a comparative analysis of different machine learning and deep learning models to predict the stock prices of real estate companies. The models we used include Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), Linear Regression, AutoRegressive Integrated Moving Average with Exogenous inputs (ARIMAX), Fast Fourier Transform (FFT), Random Forest, and TimesNet. Each of these models

has their own strengths and weaknesses, and we aim to find out which models perform best for this specific forecasting task.

Our study adds to the existing literature by providing a thorough comparison of these models in the context of real estate stock market forecasting. The results of this analysis can help investors and analysts to make more informed decisions based on the strengths of different predictive models.

II. RELATED WORKS

In recent years, many stock prediction models have been researched and many articles have been published, such as:

Hind Daori, Alanoud Alanazi, Manar Alharthi, Ghaida Alzahrani (2022) [1] used Artificial Neural Network (ANN), Random Forest Classifier, Logistic Regression, and then analyze and predict the patterns of previous stock prices

and the results showed that the models were efficient and produced better results.

Hugo Souto(2023) [2] has researched about TimesNet for Realized Volatility Prediction. Finally, they concluded that TimesNet stands out as a reliable and effective benchmark model for researching realized volatility. Although it may not always surpass NBEATSx and NHITS in every metric, its strong performance and consistency make it a valuable option, especially when compared to TFT. Overall, TimesNet presents a balanced and dependable choice that combines reliability with effectiveness, making it a suitable neural network model for researchers and practitioners in the field of realized volatility.

In another article by Bohumil Stádník, Jurgita Raudeliuniene, Vida Davidavičienė [3], they pointed out that the Fourier analysis may not be advantageous for investors forecasting stock market prices as it fails to detect existing predominant cycles. An attempt to identify significant periods in the US stock market data using FFT, a method of Fourier analysis, proved to be unacceptable. Similar failures can be expected with other liquid investment instruments or financial data series. Despite this, Fourier analysis is still used for forecasting in finance and its benefits are a topic of discussion among financial market practitioners and academicians.

Alex Sherstinsky (2020) [4] explores the complexities of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks in his work on their foundations emphasizing their structure roles and uses. The study emphasizes how important RNNs are for processing sequential data because of their innate capacity to preserve temporal dependencies via feedback loops. The vanishing gradient problem is addressed by LSTM networks an improved version of RNNs by adding gating mechanisms that regulate information flow and improve the models capacity to represent long-term dependencies. Sherstinsky talks about the gradient descent optimization techniques and the function of activation functions in the mathematical formulations that underpin these networks. The study also looks at how RNNs and LSTMs are used in a variety of contexts including speech recognition natural language processing and time-series prediction showcasing how well they work with complex sequential data.

Adil Moghar, Mhamed Hamiche (2020) [5] research about The use of Long Short-Term Memory (LSTM) networks for stock market price prediction is investigated in the paper Stock Market Prediction Using LSTM Recurrent Neural Network by Adil Moghar and Mhamed Hamiche. By using historical data to create an LSTM model the authors are able to predict future stock values with greater accuracy than they could with more conventional techniques. According to their research LSTM networks significantly increase

prediction accuracy by efficiently capturing the intricate temporal dependencies present in stock price fluctuations.

Amit Gupta and T.J. Nagalakshmi (2019) [6], in their paper "Stock Price Prediction using Linear Regression in Machine Learning" (October 2019, International Journal of Innovative Technology and Exploring Engineering), explore the use of linear regression for stock price forecasting. They detail the process of gathering historical data, selecting key features, and constructing a linear regression model. Their findings suggest that linear regression, despite its simplicity, achieves reasonable accuracy in predicting stock prices, making it a valuable tool for financial analysis.

K. Ritwik Reddy B. Tarun Kumar V. Along, Rohit Ganesh Polisetty Swetha and Prakash Kumar Sarangi(2022) [7] have looked into the use of alternative machine learning techniques and linear regression in the prediction of stock prices. Their research demonstrates how well linear regression models historical data including trading volumes and technical indicators to predict stock prices. There are trade-offs between predictive accuracy and model complexity when compared to neural networks support vector machines and ensemble methods. More recently efforts have been made to improve interpretability in dynamic market conditions and strengthen the models resilience against overfitting. In order to address persistent issues in financial forecasting future research directions include investigating hybrid models that combine linear regression with other methodologies. Sudhakar Kalva, s.Naganjaneyulu (2020) [8] investigated stock price prediction by combining LASSO regression ARIMAX and Natural Language Processing (NLP) with finance-related news. They showed that by taking historical data into account and accounting for news impact and market sentiment combining textual analysis of news with conventional statistical models could improve the prediction accuracy of stock prices.

Miaomiao Liu, Dan Yao, Zhigang Liu, Jingfeng Guo and Jing Che (2023) [9] using adaptive coefficients and composite gradients the suggested algorithm improves on the Adam optimization technique. They also makes use of randomized block coordinate descent to address problems with slow convergence missing global optima and managing high-dimensional vectors. To improve global optimization it predicts gradients uses adaptive coefficients to adjust the gradient deviation and joins current gradients with first-order momentum to form a composite gradient. By using the random block coordinate method computational overhead is decreased. Through simulations on benchmark datasets the new algorithm demonstrates a significant reduction in CPU and memory usage while outperforming six other gradient descent methods in terms of convergence speed and accuracy. The suggested technique for reservoir porosity prediction using BP neural networks shows improved stability accuracy and reduced system overhead with over 86%

of absolute error data falling within 0 percent demonstrating its efficacy.

Huizi Qian evaluates the effectiveness of LSTM and ARIMA models for stock price prediction. The study, using historical closing prices from Yahoo Finance, compares the models through metrics like MSE, MAE, and RMSE. Results indicate that LSTM outperforms ARIMA, offering greater prediction accuracy and proving more beneficial for investors forecasting future stock prices.

III. MATERIALS

A. DATASET

The dataset comprises historical daily closing stock prices (in Vietnamese Dong - VND) for three prominent Vietnamese real estate companies:

- Quoc Cuong Gia Lai Joint Stock Company (QCG)
- Dat Xanh Group Joint Stock Company (DXG)
- Vinhomes Joint Stock Company (VHM)

The data spans a five-year period from March 1, 2019, to March 1, 2024. While the raw data includes additional attributes such as opening price, high, low, volume, and change, this study focuses solely on the "Close" price to develop predictive models for future closing price movements.

B. DESCRIPTIVE STATISTICS

TABLE 1. QCG, VHM, DXG's Descriptive Statistics

	DXG	VHM	QCG
Observations	1919	1919	1919
Mean	17471.9	61135	7870.88
Median	15624	61179	7330
Std	7659	12375.84	3343.56
Min	6739	38450	3320
Max	46750	88722	23200
25%	12260	51500	5050
50%	15624	61179	7330
75%	20262	70756	9500
Skewness	1.49	-0.026	1.02
Kurtosis	2.37	-0.96	0.87

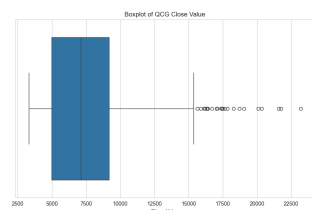


FIGURE 1. QCG stock price's boxplot

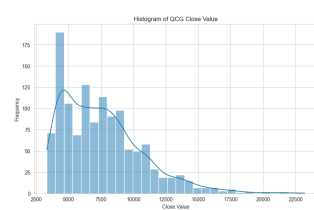


FIGURE 2. QCG stock price's histogram

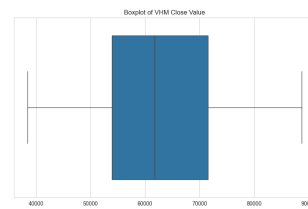


FIGURE 3. VHM stock price's boxplot

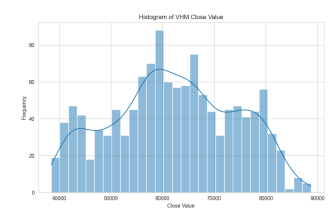


FIGURE 4. VHM stock price's histogram

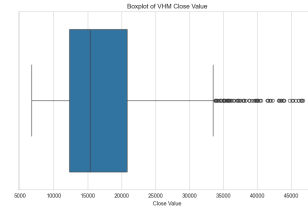


FIGURE 5. DXG stock price's boxplot

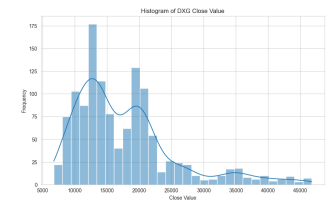


FIGURE 6. DXG stock price's histogram

IV. METHODOLOGY

A. DATA PREPROCESSING

The initial dataset of daily closing stock prices was incomplete, lacking data for weekends and potentially other non-trading days, resulting in a non-consecutive time series. Recognizing the importance of a continuous time series for accurate analysis, we took steps to fill these gaps. We assumed that the market doesn't experience significant changes over non-trading days and used the closing price of the preceding Friday to fill the missing values for weekends and holidays.

Furthermore, to enhance the predictive power of our models, we calculated several technical indicators from the closing prices. These indicators included Simple Moving Average (SMA), Exponential Moving Average (EMA), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Bollinger Bands (BB), Average True Range (ATR), and On-Balance Volume (OBV). These widely-used indicators provide valuable information about market trends, momentum, and volatility, serving as potential predictors in our linear regression model.

By addressing the missing data and incorporating technical indicators, we created a more comprehensive and informative dataset for our subsequent analysis. This enhanced dataset enabled us to explore the relationships between stock prices and various market factors, ultimately contributing to the development of more accurate predictive models.

B. LINEAR REGRESSION

A linear regression model was employed to analyze the relationship between the closing price of real estate company stocks and various technical indicators. Linear regression is a statistical method that models the linear relationship between a dependent variable and one or more independent variables. In this context, the closing price of real

estate stocks was chosen as the dependent variable, while several technical indicators derived from the stock's price and volume data were considered as potential independent variables.

The mathematical representation of a multiple linear regression model is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Where:

- Y is the predicted closing price of the real estate stock.
- X_1, X_2, \dots, X_k are the independent (explanatory) variables.
- β_0 is the intercept term.
- β_1, \dots, β_k are the regression coefficients for the independent variables.
- ε is the error term.

The dataset used for this analysis included stock price data for various real estate companies, spanning a specific time period. The dataset included various technical indicators such as Simple Moving Average (SMA), Exponential Moving Average (EMA), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Bollinger Bands (BB_High, BB_Middle, BB_Low), Average True Range (ATR), and On-Balance Volume (OBV). These indicators were selected as potential independent variables due to their established relevance in technical stock analysis.

C. RANDOM FOREST

Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result.

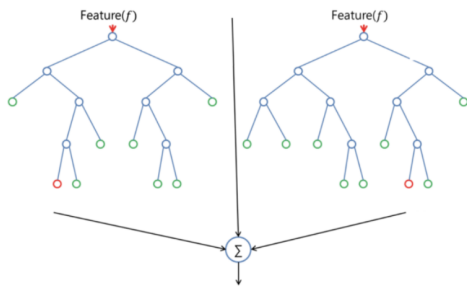


FIGURE 7. Random forest models

Despite its complexity and computational intensity, Random Forest effectively reduces overfitting and enhances prediction performance, making it a powerful and versatile machine learning algorithm.

D. GRU

GRU is a simplified version of LSTM (Long Short-Term Memory) and has fewer parameters, which helps reduce the time and computational resources required during model

training. The structure of GRU consists of two main gates:

- **Update Gate:** Controls the amount of information from the previous hidden state that needs to be carried over to the current state.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$$

- **Reset Gate:** Decides how much of the past information to forget.

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$$

- **Current memory content :** determines the potential contribution to the updated hidden state.

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h)$$

- **Final memory at current time step :** is the updated hidden state that combines the previous hidden state and the new candidate hidden state based on the update gate's decision. This updated hidden state effectively balances retaining information from the past and incorporating new information from the current time step.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

Where:

- h_t is the final hidden state at time step t .
- z_t is the update gate vector at time step t .
- h_{t-1} is the hidden state from the previous time step $t - 1$.
- \tilde{h}_t is the candidate hidden state at time step t .
- \odot represents element-wise multiplication. This operation is applied element-wise to vectors or matrices.
- $1 - z_t$ is the complement of the update gate vector.

FIGURE 8. Correlation Matrix of Filtered Data

The model was then trained using the preprocessed dataset, and its performance was evaluated using appropriate metrics such as Mean Squared Error (MSE), R-squared, and adjusted R-squared.

The choice of independent variables for the final model was guided by the correlation matrix (Figure 8), which revealed the strength and direction of linear relationships between the closing price and each indicator. Variables exhibiting higher correlation with the closing price were considered more influential and were prioritized for inclusion in the model.

By analyzing the estimated coefficients ($\beta_1, \beta_2, \dots, \beta_n$) of the linear regression model, we can quantify the impact of each technical indicator on the predicted closing price of real estate stocks. This analysis provides valuable insights into the factors that drive the stock's price movements and can inform investment decisions.

E. ARIMAX

Stock prices, much like weather patterns, exhibit both inherent trends and reactions to external forces. To capture this duality, we employed an Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) model. Building upon the established ARIMA framework, ARIMAX allows us to weave external factors, or "exogenous variables," into our forecasting tapestry. Mathematically, the ARIMAX model is expressed as:

$$y(t) = \alpha + \sum_{i=1}^p \beta_i y(t-i) + \sum_{j=1}^q \phi_j \varepsilon(t-j) + \sum_{k=1}^r \gamma_k x_k(t) + \varepsilon(t)$$

In our context, $y(t)$ represents the real estate stock price at time (t). The terms $(\sum_{i=1}^p \beta_i y(t-i))$ and $(\sum_{j=1}^q \phi_j \varepsilon(t-j))$ capture the autoregressive (AR) and moving average (MA) components, respectively, similar to ARIMA. The added dimension lies in $(\sum_{k=1}^r \gamma_k x_k(t))$, where $x_k(t)$ are our carefully curated exogenous variables, namely the technical indicators derived from our preprocessed dataset.

The selection of the ARIMAX model's parameters – the number of AR and MA terms, the order of differencing, and the specific exogenous variables – was a meticulous process. It involved scrutinizing autocorrelation and partial autocorrelation plots, along with consulting information criteria like AIC and BIC.

With the model's architecture finalized, we trained it on our refined dataset, where the closing prices served as the main melody, and the technical indicators provided the counterpoint. We assessed the model's performance using various metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), seeking to minimize these measures of forecasting error.

This ARIMAX model, enriched by the inclusion of exogenous variables, allowed us to not only decipher the inherent rhythm of real estate stock prices but also to anticipate their movements based on the external influences captured by the technical indicators. It represents a step towards a more holistic understanding of the stock market's complex dance.

F. RNN

Recurrent Neural Networks (RNNs) are a class of artificial neural networks designed to recognize patterns in sequences of data, such as text, time series data, and speech. Unlike traditional feedforward neural networks, RNNs have connections that form directed cycles, allowing information to persist. This characteristic enables RNNs to exhibit temporal dynamic behavior, making them suitable for tasks where context and sequential data play a critical role. The core concept behind Recurrent Neural Networks

(RNNs) is the introduction of a hidden state that captures information about previous inputs. This hidden state is updated at each time step as new inputs are processed. Mathematically, this process can be represented as follows:

$$h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t + b)$$

where:

- h_t is the hidden state at time t .
- W_h and W_x are weight matrices.
- x_t is the input at time t .
- b is the bias term.
- σ is a non-linear activation function (e.g., tanh or ReLU).

G. LONG SHORT TERM MEMORY (LSTM)

Long Short Term Memory networks (LSTM), often known as LSTMs, are a special type of recurrent neural network (RNN) with the ability to learn and remember long-term dependencies. LSTMs were introduced by Hochreiter and Schmidhuber in 1997, and have since been refined and developed further by many researchers and experts in the field. Thanks to their exceptional performance on various tasks, LSTMs have become increasingly popular.

LSTMs are designed to address the problem of long-term dependencies. Retaining information over extended periods is an inherent characteristic of LSTMs, requiring no special training to achieve this capability. In other words, the ability to remember long-term information is built into LSTMs.

Unlike traditional RNNs, which have a simple structure with a single tanh activation layer, LSTMs have a more complex chain-like structure, with modules that contain up to four layers interacting in a special way.

In the t -th state of the LSTM model:

Output: c_t, h_t , where c is the cell state, and h is the hidden state.

Input: c_{t-1}, h_{t-1}, x_t , where x_t is the input at state t of the model, and c_{t-1} and h_{t-1} are the outputs from the previous layer. The hidden state h is similar to s in RNN, while c is the unique aspect of LSTM.

Reading the diagram: The symbols σ and tanh indicate that the step uses the sigmoid and tanh activation functions, respectively. The multiplication is element-wise, and the addition is matrix addition.

Gates: f_t, i_t, o_t correspond to the forget gate, input gate, and output gate, respectively.

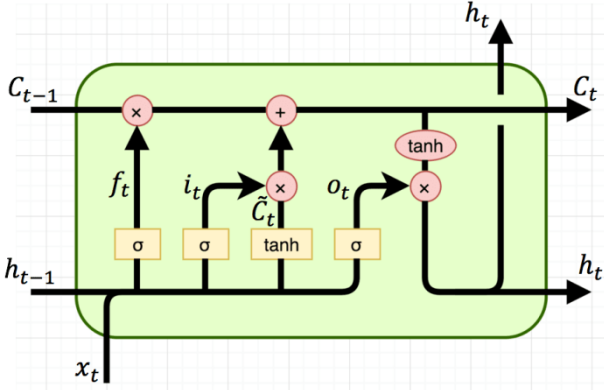


FIGURE 9. LSTM Model

• **Forget gate:**

$$f_t = \sigma(U_f \cdot x_t + W_f \cdot h_{t-1} + b_f)$$

• **Input gate:**

$$i_t = \sigma(U_i \cdot x_t + W_i \cdot h_{t-1} + b_i)$$

• **Output gate:**

$$o_t = \sigma(U_o \cdot x_t + W_o \cdot h_{t-1} + b_o)$$

Thus, the expressions for each gate of the LSTM illustrate how each gate manages the information flowing in and out of the model's states.

H. TIMESNET

TimesNet is a recent advancement in machine learning designed specifically for time series analysis tasks. Developed by researchers at Shanghai Jiao Tong University, it tackles the core challenge of capturing temporal variations within data.

Core Idea: Transforming 1D Time Series to 2D Space

Traditional time series models handle data as a single-dimensional sequence. TimesNet breaks this mold by introducing a transformation step. It converts the 1D data into a 2D tensor. This allows the model to capture complex relationships between data points, including both cyclical and inter-cyclical variations.

The Building Block: TimesBlock

The core building block of TimesNet is the TimesBlock. It is a module that performs the following:

Inception Block: This block, inspired by the Inception architecture used in image recognition, efficiently extracts features from the transformed 2D tensor. It uses multiple filter sizes within a single layer, allowing it to capture patterns at different scales within the data.

Adaptive Multi-Periodicity Discovery: TimesNet does not require pre-defining the cyclical patterns in the data. The TimesBlock can automatically discover these periodicities within the 2D representation. This makes TimesNet adaptable to various time series data with different inherent periodicities.

I. FAST FOURIER TRANSFORM FORECASTING MODEL (FFT)

The fast Fourier transform (FFT) is a computational tool that transforms time-domain data into the frequency domain by deconstructing the signal into its individual parts: sine and cosine waves. This computation allows engineers to observe the signal's frequency components rather than the sum of those components.

The FFT forecasting model leverages the fact that any periodic time series can be represented as a sum of sinusoidal functions (sines and cosines) of different frequencies. By transforming the time series data into the frequency domain, we can isolate significant frequencies that capture the underlying periodic patterns.

Given a time series $x(t)$, the Fast Fourier Transform (FFT) converts it into the frequency domain $X(f)$:

$$X(f) = \sum_{t=0}^{N-1} x(t) e^{-i2\pi ft/N}$$

where:

- N is the number of data points.
- f represents different frequency components.

To reconstruct the time series from significant frequencies:

$$x(t) = \sum_{f \in F} X(f) e^{i2\pi ft/N}$$

where F is the set of significant frequencies.

The Fast Fourier Transform Forecasting Model is a powerful tool for analyzing and forecasting time series data with periodic components. By transforming data into the frequency domain, it enables the identification of significant patterns and trends, offering an efficient and effective approach to time series forecasting.

V. RESULT

Placeholder line

A. EVALUATION METHODS

Mean Percentage Absolute Error (MAPE): is the average percentage error in a set of predicted values.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n |y_i - \hat{y}_i| = 1$$

Root Mean Squared Error (RMSE): is the square root of average value of squared error in a set of predicted values.

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

Mean Absolute Error (MAE): is the relative difference



between the log-transformed actual and predicted values.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

where:

- n is the number of observations.
- \hat{y}_i represents the predicted values.
- y_i represents the actual values.

B. DXG DATASET

DXG Dataset's Evaluation				
Model	Training:Testing	RMSE	MAPE (%)	MAE
LR	7:3	1023.75	5.06	748.41
	8:2	950.44	4.06	707.96
	9:1	843.98	3.25	579.74
ARIMAX	7:3	9354	53.17	8742.21
	8:2	4348.41	20.75	3860.00
	9:1	2355.16	11.38	2006.81
GRU	7:3	715.7	1.015	513.7
	8:2	759.27	1.013	534.78
	9:1	743.16	0.969	507.51
RNN	7:	719.14	0.99	500.41
	8:2	1159.64	1.765	932.25
	9:1	663.95	0.886	464.22
RF	7:3	743	1.06	537.83
	8:2	823.04	1.164	612.98
	9:1	794	1.071	562.85
LSTM	7:3	289.96	2.81	215.99
	8:2	330.30	3.15	244.18
	9:1	317.71	3.08	239.83
TimesNet	7:3	4344.14	25.61	3567.59
	8:2	3370.33	18.06	2613.58
	9:1	1668.46	6.20	1231.19
FFT	7:3	6346.25	32.63	5436.1
	8:2	5927.6	28.66	5259.65
	9:1	4936.33	23.9	4433.31

TABLE 2. DXG Dataset's Evaluation

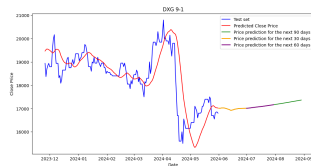


FIGURE 10. LR model's result with 9:1 splitting proportion

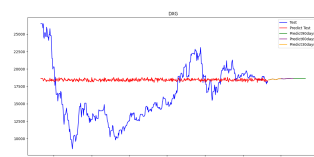


FIGURE 11. TimesNet model's result with 7:3 splitting proportion



FIGURE 12. ARIMAX model's result with 9:1 splitting proportion

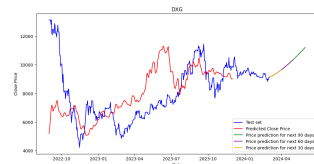


FIGURE 13. LSTM model's result with 7:3 splitting proportion



FIGURE 14. GRU model's result with 7:3 splitting proportion



FIGURE 15. RNN model's result with 9:1 splitting proportion

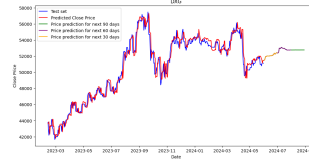


FIGURE 16. RF model's result with 7:3 splitting proportion

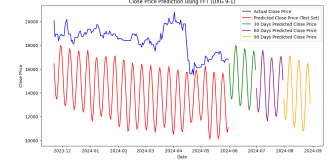


FIGURE 17. FFT model's result with 9:1 splitting proportion

C. VHM DATASET

VHM Dataset's Evaluation				
Model	Training:Testing	RMSE	MAPE (%)	MAE
LR	7:3	2291.01	3.71	1763.66
	8:2	1815.25	3.04	1414.67
	9:1	1220.21	2.36	988.38
ARIMAX	7:3	7732.84	11	5724.74
	8:2	9199.75	19.04	8278.95
	9:1	9199.75	19.04	8278.95
GRU	7:3	1199.35	1.726	816.59
	8:2	1383.71	1.59	678.3
	9:1	1412.4	1.54	437.64
RNN	7:3	1433.18	2.451	1123.99
	8:2	1027.73	1.638	712.2
	9:1	538.78	0.868	362.53
RF	7:3	1751.89	5.53	416.56
	8:2	1460.66	2.67	1128.53
	9:1	879.48	1.82	759.41
LSTM	7:3	669.02	9.18	517.60
	8:2	703.27	10.10	558.62
	9:1	619.56	7.93	480.60
TimesNet	7:3	8752.95	13.01	6901.65
	8:2	8731.15	12.68	6713.77
	9:1	4092.95	6.36	2913.35
FFT	7:3	6581.06	11.46	5510.62
	8:2	9263.1	18.83	8251.64
	9:1	16528.65	38.38	15958.64

TABLE 3. VHM Dataset's Evaluation

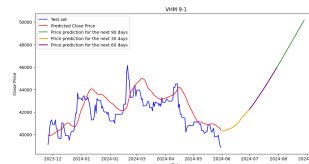


FIGURE 18. LR model's result with 9:1 splitting proportion

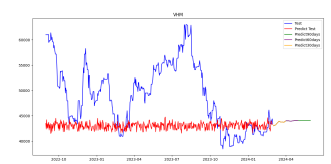


FIGURE 19. TimesNet model's result with 7:3 splitting proportion

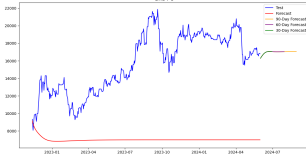


FIGURE 20. ARIMAX model's result with 7:3 splitting proportion

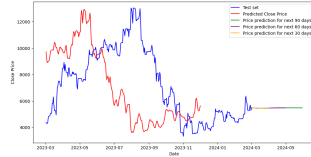


FIGURE 21. LSTM model's result with 8:2 splitting proportion

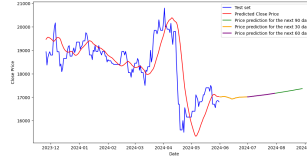


FIGURE 26. LR model's result with 9:1 splitting proportion



FIGURE 27. ARIMAX model's result with 9:1 splitting proportion



FIGURE 22. GRU model's result with 7:3 splitting proportion

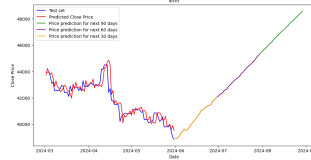


FIGURE 23. RNN model's result with 9:1 splitting proportion



FIGURE 28. GRU model's result with 9:1 splitting proportion

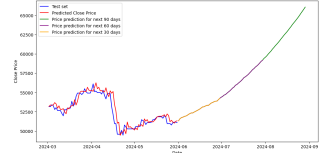


FIGURE 29. RNN model's result with 9:1 splitting proportion



FIGURE 24. RF model's result with 9:1 splitting proportion

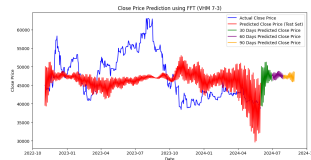


FIGURE 25. FFT model's result with 7:3 splitting proportion

D. QCG DATASET

QCG Dataset's Evaluation				
Model	Training:Testing	RMSE	MAPE (%)	MAE
LR	7:3	1023.74	5.05	748.41
	8:2	1036.14	7.10	781.55
	9:1	794.24	4.27	536.38
ARIMAX	7:3	7070.94	59.03	5922.51
	8:2	6176.84	48.68	5649.47
	9:1	2803.88	20.30	2384.16
GRU	7:3	708.29	0.974	492.8
	8:2	786.29	1.023	539.05
	9:1	665	0.835	437.64
RNN	7:3	777.04	1.113	561.96
	8:2	757.27	0.998	527.08
	9:1	698.83	10043	527.0
RF	7:3	461.24	1.698	983.65
	8:2	1479.93	1.709	1046.63
	9:1	2378.81	2.51	1686.66
LSTM	7:3	379.56	3.95	298.40
	8:2	314.71	2.84	216.48
	9:1	315.31	2.41	200.57
TimesNet	7:3	3259.52	51.75	2808.82
	8:2	2938.94	37.76	2455.60
	9:1	2488.24	17.49	2024.84
FFT	7:3	5854.61	46	4724.21
	8:2	6772.32	53.6	6173.79
	9:1	6855.68	48.97	6041.34

TABLE 4. QCG Dataset's Evaluation

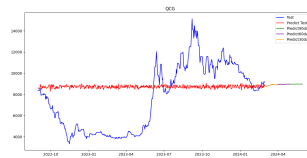


FIGURE 30. TimesNet model's result with 7:3 splitting proportion



FIGURE 31. LSTM model's result with 7:3 splitting proportion

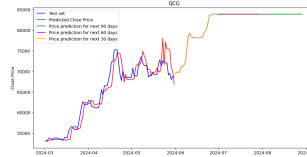


FIGURE 32. RF model's result with 9:1 splitting proportion

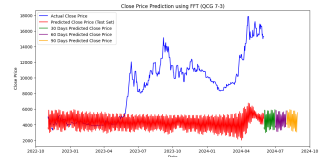


FIGURE 33. FFT model's result with 7:3 splitting proportion

VI. CONCLUSION

A. SUMMARY

In the pursuit of forecasting stock prices, various methodologies have been explored, ranging from traditional statistical models to advanced machine learning algorithms. Among the examined models, Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), Random Forest, Fast Fourier Transform (FFT), TimesNet, and Auto Regressive Integrated Moving Average (ARIMA) stand out. Notably, RNN, GRU, LSTM, and Random Forest have emerged as the most promising and effective models for predicting stock prices.

The intricacies of stock price forecasting, rooted in the complexity and unpredictability of financial markets, demand models capable of capturing nuanced patterns and relationships within the data. Recurrent Neural Networks (RNN) excel at handling sequential data, offering robust predictions. Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) models, with their enhanced ability

to capture sequential dependencies, exhibit notable performance in stock price forecasting. Random Forest, with its ensemble learning approach, further refines predictive capabilities by combining multiple decision trees to provide collective insights

As evidenced by evaluation metrics such as RMSE, MAPE, and MAE, the RNN, GRU, LSTM, and Random Forest models consistently demonstrate superior performance across various aspects of forecasting accuracy. Their adaptability in managing the inherent uncertainties of stock markets positions them as formidable tools for investors and analysts seeking reliable predictions.

B. FUTURE CONSIDERATIONS

The exploration of forecasting stock prices using Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), Random Forest, Fast Fourier Transform (FFT), TimesNet, and AutoRegressive Integrated Moving Average (ARIMA) models has given some initial insights, yet the accuracy of these models is still a challenge. To improve the predictive performance and usefulness of these models, several future considerations are needed:

- Integrating more diverse and high-frequency data sources like social media sentiment, macroeconomic indicators, and real-time news feeds could make the dataset richer. This would allow models to capture a more comprehensive range of market dynamics and improve forecasting accuracy.
- Using and possibly developing more comprehensive evaluation metrics beyond RMSE, MAPE, and MAE to assess model performance can give deeper insights into their effectiveness and areas needing improvement.
- Creating hybrid models that combine the strengths of different approaches could lead to better performance. For example, integrating the sequential learning capabilities of RNNs with the ensemble approach of Random Forests could result in models that better capture both short-term and long-term market patterns.
- These future considerations aim to address the current limitations and explore new ways to improve the effectiveness of stock price forecasting models.

ACKNOWLEDGMENT

First and foremost, we would like to express our sincere gratitude to **Assoc. Prof. Dr. Nguyen Dinh Thuan**, **Mr. Nguyen Minh Nhut** and **Ms. Dang Vu Phuong Uyen** for their exceptional guidance, expertise, and invaluable feedback throughout the research process. Their mentorship and unwavering support have been instrumental in shaping the direction and quality of this study. Their profound knowledge, critical insights, and attention to detail have significantly contributed to the success of this research. This research would not have been possible without the support and contributions of our mentors. We would like to extend our heartfelt thanks to everyone involved for

their invaluable assistance, encouragement, and belief in our research. Thank you all for your invaluable assistance and encouragement

REFERENCES

- [1] Hind Daori, Alanoud Alanazi, Manar Alharthi, Ghaida Alzahrani , 2022. Predicting Stock Prices Using the Random ForestClassifier
- [2] Hugo Souto, 2023. TimesNet for Realized Volatility Prediction.
- [3] Bohumil Stádník, Jurgita Raudeliuniene, Vida Davidavičienė , 2016. Fourier Analysis For Stock Price Forecasting: Assumption And Evidence .
- [4] Alex Sherstinsky, 2020, Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network.
- [5] Adil Moghar, Mhamed , 2020, Stock Market Prediction Using LSTM Recurrent Neural Network
- [6] Amit Gupta and T.J. Nagalakshmi, 2019, Stock Price Prediction using Linear Regression in Machine Learning.
- [7] K. Ritwik Reddy B. Tarun Kumar V. Along, Rohit Ganesh Polisetty Swetha and Prakash Kumar Sarang, 2022. Stock Market Prediction Using Recurrent Neural Network.
- [8] Sudhakar Kalva, s.Naganjaneyulu, 2020. STOCK PRICE PREDICTION BASED ON FINANCE RELATED NEWS USING NLP, LASSO AND ARIMAX.
- [9] Miaomiao Liu, Dan Yao, Zhigang Liu, Jingfeng Guo and, Jing Che, 2023, An Improved Adam Optimization Algorithm Combining Adaptive Coefficients and Composite Gradients Based on Randomized Block Coordinate Descent.
- [10] Huizi Qian, 2022, Stock Predicting based on LSTM and ARIMA