Forecasting price trends of stock using forecasting models and machine learning

1st Do Dang Kien Nam  
IS403.N21.HTCL  
University of Information Technology[20521627@gm.uit.edu.vn](mailto:20521627@gm.uit.edu.vn)

2nd Bui Thi Thanh Ngan  
IS403.N21.HTCL  
University of Information Technology  
[20521643@gm.uit.edu.vn](mailto:20521643@gm.uit.edu.vn)

3rd Le Anh Thu  
IS403.N21.HTCL  
University of Information Technology  
[20521985@gm.uit.edu.vn](mailto:20521985@gm.uit.edu.vn)

***Abstract—*** *The stock market plays a critical role in the global economy, attracting immense interest from investors and researchers alike. Predicting stock market trends accurately has long been a challenging task due to the complex and dynamic nature of financial markets. This paper provides a brief overview of the techniques and methodologies employed for stock market prediction. It uses traditional statistical models and the emergence of machine learning algorithms for predicting stock market trends. , ARIMA, Linear Regression, LSTM, GRU, RNN, VAR, BVAR, RKN to forecast the trend of stocks from 2017 to 2023*

***Keywords—*** *Stock pricing, machine learning, ARIMA, Linear Regression, LSTM, GRU, RNN, VAR, BVAR, RKN.*

# INTRODUCTION

A stock is a security that represents a fractional ownership in a company. When you buy a company's stock, you're purchasing a small piece of that company, called a share. Investors purchase stocks in companies they think will go up in value. If that happens, the company's stock increases in value as well. The stock can then be sold for a profit.

Stock Price Prediction is the task of forecasting future stock prices based on historical data and various market indicators. The goal of stock price prediction is to help investors make informed investment decisions by providing a forecast of future stock prices.

In recent years, [artificial intelligence](https://www.sciencedirect.com/topics/computer-science/artificial-intelligence) (AI) has become more prevalent in the financial industry, including the stock market. AI algorithms have the capability to analyze vast amounts of data and make predictions or decisions based on that analysis. This can be useful for predicting stock prices, identifying trends, and making investment decisions. There are many algorithms and techniques that help us predict stock prices. In this paper, we will use statistical models and machine learning algorithms like Linear Regression, SVR, VAR, BVAR, ARIMA, LSTM, RNN, GRU, RKN models to predict the stock prices of some corporations in Vietnam like MWG, VIC and FTS for the next 30 days.

# RELATED WORK

There are lots of research work in stock market prediction. Almost every data mining and prediction techniques were applied for prediction of stock prices. Many different features and attributes were used for the same purpose. M. Almasarweh and S. A. Wadi analysis and research on stock price of bank by using ARIMA model [1]. Another point H. M, G. E.A., V. K. Menon, and S. K.P using four types of deep learning architectures i.e Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) for predicting the stock price of a company based on the historical prices available [2]. Another study B. M. Henrique, V. A. Sobreiro, and H. Kimura using Linear Regression, forecastTCS data set behavior and at the end they compared and evaluated the result of our proposed method with other approaches , this study uses a machine learning technique called Support Vector Regression (SVR) to predict stock prices for large and small capitalisations and in three different markets, employing prices with both daily and up-to-the-minute frequencies [3]. Y. Gao, R. Wang, and E. Zhou use sing depth learning LASSO and PCA approaches to predict stock price. In addition, a comparison of the performances of LSTM and GRU for stock market forecasting under various parameters was performed. This study show that both LSTM and GRU models can predict stock prices efficiently, not one better than the other [4]. E. Papapetrou using a multivariate vector-autoregression (VAR) approach, this paper attempts to shed light into the dynamic relationship among oil prices, real stock prices, interest rates, real economic activity and employment for Greece. The empirical evidence suggests that oil price changes affect real economic activity and employment. Oil prices are important in explaining stock price movements. [Stock returns](https://www.sciencedirect.com/topics/economics-econometrics-and-finance/capital-market-returns) do not lead to changes in real activity and employment [5].

# MATERIALS

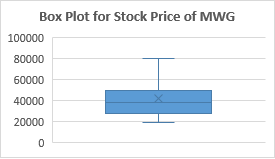
## Data soucre

Datasets is collected through stock trends of 3 companies (Mobile World Investment Corp, Vingroup Joint Stock Company, FPT) on Website Investing.com between January 2 , 2017 and June 16, 2023. Every dataset includes 1360 rows and 7 attributes in total. The data contains trading information on high price, low price, open price, close price, trading volume and change. These columns mean respectively the timeline of data representation, closing price of the stock, opening price of the stock, the highest closing price, the lowest price at which a stock trades over the course of a trading day, the number of shares traded in a given period.

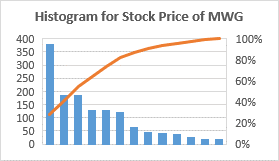
## Descriptive statistics

|  | MWG | FTS | VIC |
| --- | --- | --- | --- |
| Count | 1360 | 1360 | 1360 |
| Mean | 41602.58 | 16513.87 | 89532.10 |
| Std | 15199.76 | 11831.89 | 18628.47 |
| Min | 19198.00 | 5788.40 | 50500.00 |
| 25% | 28196.00 | 7726.32 | 78100.00 |
| 50% | 38008.00 | 9413.30 | 94133.00 |
| 75% | 48937.50 | 24685.30 | 103111.00 |
| Max | 79582.00 | 51873.90 | 128000.00 |

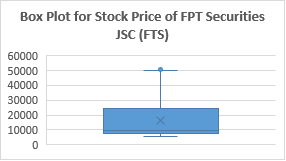
*Table 1: Descriptive Statistics*



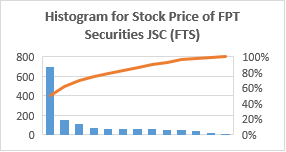
*Figure 1: Box plot for stock price of MWG*



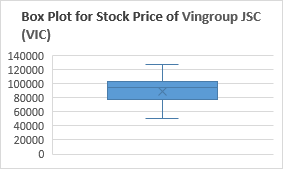
*Figure 2: Historgram for stock price of MWG*



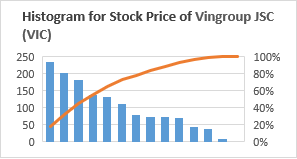
*Figure 3: Box plot for stock price of FPT Securities JSC (FTS)*



*Figure 4: Histogram for stock price of FPT Securities (FTS)*



*Figure 5: Box plot for stock price of Vingroup JSC (VIC)*



*Figure 6: Histogram for stock price of Vingroup JSC (VIC)*

Identify applicable funding agency here. If none, delete this text box.

# METHODOLOGY

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Linear Regression (LN)

Linear Regression is an algorithm of machine learning, based on supervised learning. This method is frequently used to forecast and determine the relationship between variables that cause and effect. The number of independent variables and the type of relationship between the independent and dependent variables are the primary differences between regression methods. The regression procedure enables you to confidently establish which elements are most important, which can be ignored, and how those factors interact with one another.

The formula for a simple linear regression is [7]

y = +

Where

* y is the predicted value of dependent variables (y) for any given of independent variables (x)
* is the intercept , the predicted value of y when x is 0
* is the regression coefficient – how much we expect y to change as x increases
* is the independent variable ( the variable we expect is influenced by y)
* is the error estimate, or how much variation there is in our estimate of the regression coefficient.

## Support Vector Regession (SVR)

Support Vector Regression as the name suggests is a regression algorithm that supports both linear and non-linear regressions. This method works on the principle of the Support Vector Machine. SVR differs from SVM in the way that[SVM is a classifier that is used for](https://www.educba.com/svm-algorithm/) predicting discrete categorical labels while SVR is a regressor that is used for predicting continuous ordered variables.[8]

SVM kernel function [9]

| Kernel | Function |
| --- | --- |
| Polynomial |  |
| RBF | Exp(-y||) |

*Table 2: SVM Kernel*

## ARIMA

ARIMA stands for Auto Regressive Integrated Moving Average.

The basic model in the time series analysis is the ARIMA model. It is a combination of two processes – autoregressive (AR) and moving average (MA), which is weighted delayed random components. The letter I in the model name indicates the level of integration of the analyzed variable. Integrated variables are variables that can become stationary through differentiation. The structure of ARIMA is based on the phenomenon of autocorrelation. ARIMA can be used for modeling stationary time series or non-stationary time series that can become stationary through differentiation. [10]

There are seasonal and Non-seasonal ARIMA models that can be used for forecasting:

* Non-Seasonal ARIMA model.
* Seasonal ARIMA (SARIMA) models.

Non-Seasonal ARIMA(p,d,q)

Where:

* p : Periods to lag for eg: (if P= 3 then we will use the three previous periods of our time series in the autoregressive portion of the calculation) P helps adjust the line that is being fitted to forecast the series
* q : This variable denotes the lag of the error component, where error component is a part of the time series not explained by trend or seasonality.
* d : In an ARIMA model we transform a time series into stationary one(series without trend or seasonality) using differencing. D refers to the number of differencing transformations required by the time series to get stationary.

Stationary time series is when the mean and variance are constant over time. It is easier to predict when the series is stationary.

Differencing is a method of transforming a non-stationary time series into a stationary one. This is an important step in preparing data to be used in an ARIMA model. [11]

## VAR

Vector Autoregression (VAR) is a forecasting algorithm that can be used when two or more time series influence each other. That is, the relationship between the time series involved is bi-directional.

The basic requirements in order to use VAR are:

- At least two time series (variables).

- The time series should influence each other.

VAR is considered as an Autoregressive model because, each variable (Time Series) is modeled as a function of the past values, that is the predictors are nothing but the lags (time delayed value) of the series.

Where

* α is the intercept
* a is constant
* β1, β2 till βp are the coefficients of the lags of Y till order p

Order ‘p’ means, up to p-lags of Y is used and they are the predictors in the equation. The ε\_{t} is the error, which is considered as white noise.

In the VAR model, each variable is modeled as a \*\*linear combination of past values of itself and the past values of other variables in the system\*\*. Since you have multiple time series that influence each other, it is modeled as a system of equations with one equation per variable (time series). [12]

## BVAR

Bayesian Vector Autoregression (BVAR) models have the same mathematical form as any other VAR model.

Where

* yt is a K×1vector of endogenous variables in period t
* yt is a vector of m economic variables at time t
* Ai is the cofficient matrix corresponding to the i

th lag of yt

* c is a constant deterministic term
* ϵ is an error term with zero mean and variance-covariance Σ

The only difference between usual VAR models and BVAR models is the way parameter estimates are obtained and interpreted.

Bayesian estimators are slightly more complicated and more burdensome in terms of algebra and calculation power. [13]

.

## GRU

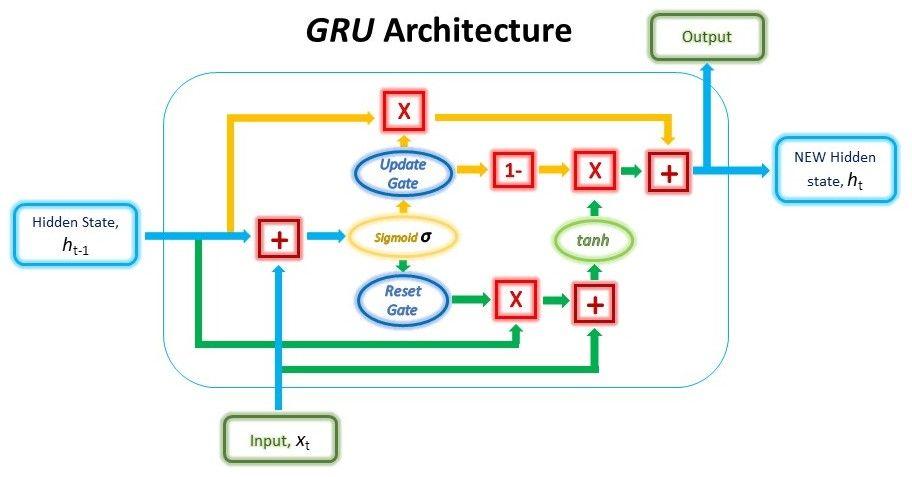
A Gated Recurrent Unit (GRU) is a variant of the RNN architecture, and uses gating mechanisms to control and manage the flow of information between cells in the neural network. GRUs can be considered a relatively new architecture, especially when compared to the widely-adopted LSTM.

A picture containing text, screenshot, diagram, font

Description automatically generated

*Figure 7: Overview of GRU*

The structure of the GRU allows it to adaptively capture dependencies from large sequences of data without discarding information from earlier parts of the sequence. This solve the vanishing/exploding gradient problem of traditional RNNs.



*Figure 8: GRU Architecture*

The ability of the GRU to hold on to long-term dependencies or memory stems from the computations within the GRU cell to produce the hidden state. While LSTMs have two different states passed between the cells — the cell state and hidden state, which carry the long and short-term memory, respectively — GRUs only have one hidden state transferred between time steps. This hidden state is able to hold both the long-term and short-term dependencies at the same time due to the gating mechanisms and computations that the hidden state and input data go through.

Reset Gate:

Update Gate:

Combining the outputs:

[14]

## RNN

A diagram of a machine

Description automatically generated with low confidence

*Figure 9: RNN Architecture*

A Recurrent Neural Networks is a more flexible model, since it encodes the temporal context in its feedback connections, which are capable of capturing the time varying dynamics of the underlying system RNNs are learning machines that recursively compute new states by applying transfer functions to previous states and inputs. Typical transfer functions are composed by an affine transformation followed by a nonlinear function, which are chosen depending on the nature of the particular problem at hand.

An activation function determines whether a neuron should be activated. The nonlinear functions typically convert the output of a given neuron to a value between 0 and 1 or -1 and 1. [15]

A picture containing text, diagram, line, font

Description automatically generated

*Figure 10: GRU activation function*

*::*

## LSTM

RNN is a time-series neural network. The interconnection structure between the hidden layers reflects the interaction between time series. However, there are vital problems exist in RNN: the fast gradient descent problem and nonconvergent problem. Fortunately, the bi-directional LSTM model can solve the gradient problem of RNN network by adding gates and using the context relation of forward and backward time directions in time series, improving the prediction accuracy subsequently.

In 1997, Hochreiter and Schmidhuber proposed LSTM, which had achieved surprising performance in the NLP field. LSTM aims at resolving long-term dependence problem based on improved RNN (Annotation) neural network. Keeping information in mind for a long time is an inherent characteristic of LSTM. All RNN models have a chain form of repetitive neural network modules. [16]

Similarly, as a variant of RNN, LSTM also has this chain module structure, but with different repetitive modules and layers. LSTM has three more gates than RNN with only the tanh layer.

A picture containing screenshot, circle, line, design

Description automatically generated

*Figure 11: LSTM Architecture*

LSTM equations:

Forget Gate:

Input Gate:

Cell Gate:

Output Gate:

Cell State: [17]

Bidirectional long-short term memory (Bi-LSTM) is a technique that allows any neural network to store sequence information both forward and backward. Bi-LSTM allows input flow in both directions, whereas normal LSTM only allows input flow in one direction

A picture containing diagram, text, plan, screenshot

Description automatically generated

*Figure 12: Bidirectional LSTM*

## RKN

The prediction and observation updates result in a new type

of recurrent neural network, that we call Recurrent Kalman

Network, which allows working in high dimensional state

spaces while keeping numerical stability, computational

efficiency and (relatively) low memory consumption.

The RKN provides a principled method to deal

with absent inputs by just omitting the update step and

setting the posterior to the priority.

A picture containing text, diagram, line, font

Description automatically generated

*Figure 13: Reccurrent neural network*

# RESULT

### Performance measure

To assess the predictive power of our proposed models, we use two performance measures: the root means square error (RMSE) and the MAPE. When we train models, we use RMSE as a loss function, and MAPE is a statistical measure of prediction accuracy. The following are the equations:

Where:

* f = forecasts (expected values or unknown results),
* o = observed values (known results).

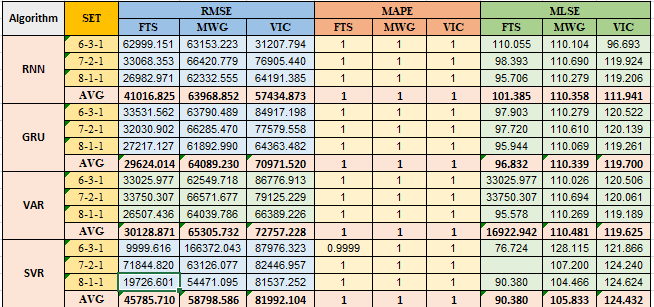
Where:

* n is the number of fitted points,
* At is the actual value,
* Ft is the forecast value.
* Σ is summation notation (the absolute value is summed for every forecasted point in time).

MSLE is the relative difference between the log-transformed actual and predicted values.

The formula of the MSLE is :

Where:



A picture containing text, screenshot, number, font

Description automatically generated

*Table 3: Performance measure*

1. Predictting next 30 days

* Linear Regression (LN)

|  |  |
| --- | --- |

* Support Vector Regession (SVR)

|  |  |
| --- | --- |

* Auto Regressive Integrated Moving Average (ARIMA)

|  |  |
| --- | --- |

* Vector Autoregression (VAR)

|  |  |
| --- | --- |

* Bayesian Vector Autoregression (BVAR)

| A blue line graph on a white background  Description automatically generated with medium confidence | A picture containing text, screenshot, plot, line  Description automatically generated |
| --- | --- |

* A Gated Recurrent Unit (GRU)

|  |  |
| --- | --- |

* Recurrent Neural Network (RNN)

|  |  |
| --- | --- |

* Long short-term memory (LSTM)

|  |  |
| --- | --- |

* A bidirectional LSTM (BiLSTM)

|  |  |
| --- | --- |

* Recurrent Kalman Network (RKN)

| A blue line graph on a white background  Description automatically generated with medium confidence | A picture containing text, screenshot, plot, line  Description automatically generated |
| --- | --- |

# CONCLUSION

The following table are the experimental results of six models for training and testing, this study show that the BiLSTM model was the best for predicting the future price of the FTS, MWG, and VIC stocks in the resultant time series out of the nine models tested (LN, SVR, ARIMA, VAR, BVAR, RNN, BiLSTM, GRU, LSTM). The performance of the other models, such as the SVR and RNN, was subpar. This study emphasizes the significance of taking into account a range of modeling strategies in financial analysis, as well as the potential benefit of employing the BiLSTM model to forecast future stock prices. Future research could be done to confirm the findings of this study and to examine how well the other models perform on various stock price prediction tasks.

##### Acknowledgment

I want to start by expressing my sincere gratitude to ***Assoc. Prof. Dr. Nguyen Dinh Thuan*** for always looking out for our team during this course. I value your support and sharing your invaluable experience with us because it will enable us to complete our project. I want to extend my sincere gratitude to ***Mr. Nguyen Minh Nhut*** and ***Ms. Nguyen Viet Huong***, who has taught and helped our team throughout the course.

I want to take this chance to thank ***Assoc. Prof. Dr. Nguyen Dinh Thuan*** for letting us work on this project. Finally, I want to say how grateful I am that our team was able to attend your session and learn from you.

Many thanks to ***Assoc. Prof. Dr. Nguyen Dinh Thuan, Mr. Nguyen Minh Nhut*** and ***Ms. Nguyen Viet Huong.*** We sincerely appreciate your tireless efforts in leading our team to success and motivating us to keep moving forward. Your tireless efforts have been instrumental in ensuring that our project was a resounding success. We would also like to extend our sincere gratitude to all of our classmates who generously volunteered their time and support to help us bring our project to fruition. Without your contributions, we would not have been able to achieve such an outstanding result.

##### References

[1] M. Almasarweh and S. A. Wadi, “ARIMA Model in Predicting Banking Stock Market Data,” *Mod. Appl. Sci.*, vol. 12, no. 11, p. 309, Oct. 2018, doi: 10.5539/mas.v12n11p309.

[2] H. M, G. E.A., V. K. Menon, and S. K.P., “NSE Stock Market Prediction Using Deep-Learning Models,” *Procedia Comput. Sci.*, vol. 132, pp. 1351–1362, 2018, doi: 10.1016/j.procs.2018.05.050.

[3] B. M. Henrique, V. A. Sobreiro, and H. Kimura, “Stock price prediction using support vector regression on daily and up to the minute prices,” *J. Finance Data Sci.*, vol. 4, no. 3, pp. 183–201, Sep. 2018, doi: 10.1016/j.jfds.2018.04.003.

[4] Y. Gao, R. Wang, and E. Zhou, “Stock Prediction Based on Optimized LSTM and GRU Models,” *Sci. Program.*, vol. 2021, p. e4055281, Sep. 2021, doi: 10.1155/2021/4055281.

[5] E. Papapetrou, “Oil price shocks, stock market, economic activity and employment in Greece,” *Energy Econ.*, vol. 23, no. 5, pp. 511–532, Sep. 2001, doi: 10.1016/S0140-9883(01)00078-0.

[6] E. W. Saad, D. V. Prokhorov, and D. C. Wunsch, “Advanced neural network training methods for low false alarm stock trend prediction,” in *Proceedings of International Conference on Neural Networks (ICNN’96)*, Washington, DC, USA: IEEE, 1996, pp. 2021–2026. doi: 10.1109/ICNN.1996.549212.

[7] R. Bevans, “Simple Linear Regression | An Easy Introduction & Examples,” *Scribbr*, Feb. 19, 2020. https://www.scribbr.com/statistics/simple-linear-regression/ (accessed Jun. 17, 2023).

[8] “Support Vector Regression (SVR) using Linear and Non-Linear Kernels in Scikit Learn,” *GeeksforGeeks*, Jan. 25, 2023. https://www.geeksforgeeks.org/support-vector-regression-svr-using-linear-and-non-linear-kernels-in-scikit-learn/ (accessed Jun. 17, 2023).

[9] D. Team, “Kernel Functions-Introduction to SVM Kernel & Examples,” *DataFlair*, Aug. 12, 2017. https://data-flair.training/blogs/svm-kernel-functions/ (accessed Jun. 17, 2023).

[10] D. Kobiela, D. Krefta, W. Król, and P. Weichbroth, “ARIMA vs LSTM on NASDAQ stock exchange data,” *Procedia Comput. Sci.*, vol. 207, pp. 3836–3845, Jan. 2022, doi: 10.1016/j.procs.2022.09.445.

[11] Sangarshanan, “Time series Forecasting — ARIMA models,” *Medium*, Apr. 07, 2019. https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06 (accessed Jun. 17, 2023).

[12] “Vector Autoregression (VAR) - Comprehensive Guide with Examples in Python - Machine Learning Plus.” https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/ (accessed Jun. 17, 2023).

[13] G. Koop, “Bayesian Multivariate Time Series Methods for Empirical Macroeconomics,” *Found. Trends® Econom.*, vol. 3, no. 4, pp. 267–358, 2009, doi: 10.1561/0800000013.

[14] P. Dua and S. C. Ray, “A BVAR model for the connecticut economy,” *J. Forecast.*, vol. 14, no. 3, pp. 167–180, May 1995, doi: 10.1002/for.3980140303.

[15] “Gated Recurrent Unit (GRU) With PyTorch,” *FloydHub Blog*, Jul. 22, 2019. https://blog.floydhub.com/gru-with-pytorch/ (accessed Jun. 17, 2023).

[16] F. M. Bianchi, E. Maiorino, M. C. Kampffmeyer, A. Rizzi, and R. Jenssen, *An overview and comparative analysis of Recurrent Neural Networks for Short Term Load Forecasting*. 2017. doi: 10.1007/978-3-319-70338-1.

[17] “Impact of News on the Trend of Stock Price Change: an Analysis based on the Deep Bidirectiona LSTM Model,” *Procedia Comput. Sci.*, vol. 174, pp. 128–140, Jan. 2020, doi: 10.1016/j.procs.2020.06.068.

[18] “Deep Learning | Introduction to Long Short Term Memory,” *GeeksforGeeks*, Jan. 16, 2019. https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/ (accessed Jun. 17, 2023).