Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 12, Issue. 5, May 2023, pg.17 – 31

Daily Stock Price Prediction: A Case Study in Vietnam

Ta Quang Chieu[1]; Vu Thi Kim Hang[2]

[1]Faculty of Computer Science and Engineering, Thuyloi University, Hanoi, Vietnam
[2]Vietnam Center of Research in Economics, Management and Environment (VCREME), Vietnam
[1] quangchieu.ta@tlu.edu.vn; [2] vuthikimhang7777@gmail.com

DOI: https://doi.org/10.47760/ijcsmc.2023.v12i05.004

Abstract:
Stock price prediction remains a topic of debate. The Efficient Market Hypothesis argues against the need for prediction, stating that markets already incorporate all relevant information. However, the emergence of Deep Learning in Machine Learning has led many to believe in the potential of sophisticated algorithms in uncertain and volatile stock markets. The outcome of this debate varies based on market characteristics. To analyse this discussion within the framework of emerging market like Vietnam, a study examines nine stocks listed on the Ho Chi Minh Stock Exchange (HOSE). Time-series data is visualized and analysed to test various hypotheses. Three algorithms Linear Regression, SARIMA, and LSTM are researched and applied to predict stock prices. The predictions are compared and benchmarked against the Naive approach. The study concludes that the Linear approach outperforms Deep Learning. This difference may be attributed to the characteristics of the emerging market, such as its young age, low regulation, and high volatility with outliers. Traditional Machine Learning and Deep Learning both surpass the Naive approach, rejecting the Efficient Market Hypothesis in Vietnam.
Keywords: Daily Stock Price Prediction; Quantitative Finance; Traditional Machine Learning; Deep Learning; Linear Regression; SARIMA; LSTM

I. INTRODUCTION

Practically, traders who have available cash are currently participating in the stock market with the expectation of making a profit. They aim to purchase stocks at a low cost and sell them at a higher price. Different trading strategies are based on traders' characteristics and attitudes towards the market. These strategies can be categorized as fundamental analysis, technical analysis, and quantitative trading.

Investors following a fundamental style prioritize investing in high-quality businesses that are undervalued in the market compared to their intrinsic value. Determining this intrinsic value often relies on fragmented information, leading investors to rely on their experience and knowledge when making decisions.

On the other hand, technical analysis focuses on the market price rather than the intrinsic value of stocks. Investors employing this strategy believe that historical prices reflect historical market behaviours, which are likely to repeat in the future due to consistent human reactions. Technical indicators and patterns are commonly used to identify these trends.

Quantitative trading, like technical analysis, relies on historical prices. However, it incorporates modern technology, scientific principles, and statistical methods. It may utilize Machine Learning, Deep Learning and optimization formulas to design portfolio strategies with high returns. Precise stock price prediction is a crucial aspect of quantitative trading, but forecasting stock prices is a debating topic. Some argue that volatility and numerous factors make it challenging to predict stock prices accurately, while others believe that stock price movements are essentially random, following the concept of the Efficient Market Hypothesis. To address this challenge, careful analysis of time series stock prices and the development of models are ongoing across different markets.

Stock price prediction is divided into short-term, mid-term, and long-term categories based on trading windows. Short-term prediction can range from minutes to daily intervals, while mid and long-term prediction can extend from one week to a year. Short-term prediction is more commonly discussed in this research.

In terms of prediction methodologies, stock price prediction can be classified into Traditional Machine Learning techniques and Deep Learning [11]. Traditional models encompass linear and non-linear methods. Linear approaches include Linear Regression, Exponential Smoothing Methods, and ARIMA, while non-linear methods include Random Forest (RF), Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbour (KNN), and Softmax. However, some Traditional Machine Learning methods have limitations. Linear approaches rely on assumptions about time series data that may not hold in real-world scenarios, while non-linear methods may struggle with predicting sequences when features exhibit smooth or nearly dependent covariates.

In recent years, Deep Learning has shown significant improvements in various domains, including forecasting. Deep learning excels at handling unstructured and complex data compared to Traditional Machine Learning approaches. Within Deep Learning, Recurrent Neural Networks (RNNs) are particularly suitable for time series datasets, with LSTM (Long Short-Term Memory) being a commonly used variant. Despite achieving promising results, it is still premature to conclude that Deep Learning outperforms traditional methods. Deep Learning's effectiveness is highly dependent on complex and large-volume training data. Thus, the question arises as to whether these conditions are met within the realm of stock price datasets.

Moving on to the Vietnam Stock Market, which has been operational since 2005, it encompasses three major stock exchanges: the Ho Chi Minh Stock Exchange (HOSE), the Ha Noi Stock Exchange (HNX), and the UPCOM. Among these exchanges, HOSE and HNX hold dominant positions, with approximately 402 and 359 listed companies, respectively. Conversely, despite being the youngest exchange, UPCOM boasts a higher stock volume, with nearly 900 listed companies. While HOSE and HNX exhibit similar functionalities and mechanisms that attract most investors, UPCOM operates with less stringent regulations, making it a riskier market that appeals more to speculators.

This study endeavours to investigate two key research questions concerning stock price forecasting in the emerging market of Vietnam:

- Does stock price forecasting outperform the performance of random walk (Naive model), and thus refute the Efficient Market Hypothesis?
- Does Deep Learning outperform traditional learning methods in the context of Vietnam's stock market?

To answer these questions, a list of randomly selected stocks from the Ho Chi Minh Stock Exchange (HOSE) is utilized, and features such as close, open, high, low, and trading volume are considered. Traditional approaches such as Linear Regression and SARIMA are compared with Deep Learning methods such as LSTM, and the accuracy of each model is assessed using the Mean Absolute Percentage Error (MAPE), and MAE (Mean Absolute Error) metric. Moreover, the findings are compared with the Naive model based on the Efficient Market Hypothesis to measure the improvement of the proposed models.

II. STOCK PREDICTION TAXONOMY

For better understanding about the pros and cons of major techniques, it is briefly summarized in table 1 as below:

Regarding Traditional Machine Learning, this group includes Exponential Smoothing, Linear Regression, ARIMA, and SARIMA. These techniques are based on statistical analysis and are suitable for short-term forecasting with linear and stationary data. They require certain assumptions to be met, such as linearity and stationarity, and may not perform well with high seasonality or variation.

Reference to Non-Parametric Machine Learning, this group includes Support Vector Regression (SVR), Decision Tree, Random Forest, and K-Nearest Neighbours (KNN). These techniques are more flexible than Traditional Machine Learning techniques and can handle non-linear and non-stationary data. They do not require specific assumptions to be met and can handle large and complex datasets.

TABLE 1
TECHNIQUE COMPARISON

<table>
<tr><th>GROUP TECHNIQUE</th><th>TECHNIQUE</th><th>DATASET</th><th>MULTI-VARIABLE</th><th>FORECASTING TERM</th><th>ASSUMPTION</th><th>WEAKNESS</th></tr>
<tr><td rowspan="7">Traditional Machine Learning</td><td>Exponential Smoothing</td><td rowspan="7">No require</td><td rowspan="2">No</td><td rowspan="7">Better for Short - term</td><td rowspan="2">Linearity</td><td>Low performance with high seasonality/ trend
Low performance with high variation</td></tr>
<tr><td>Linear Regression</td><td>Sensitive with outliers
Problem of Multivariate</td></tr>
<tr><td>ARIMA, SARIMA</td><td rowspan="5">Yes</td><td>Linearity & Stationarity</td><td>Low performance with multi seasonal factors
Sensitive with outliers</td></tr>
<tr><td>SVR</td><td rowspan="4">No require</td><td rowspan="4">Low performance with highly linear data</td></tr>
<tr><td>Decision Tree</td></tr>
<tr><td>Random Forest</td></tr>
<tr><td>KNN</td></tr>
<tr><td>Deep Learning</td><td>CNN, ANN, RNN (LSTM)</td><td>Large & complex</td><td>Yes</td><td>No require</td><td>No require</td><td>Require complex model & Low performance with small dataset</td></tr>
</table>

Last but not least, Deep Learning includes Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks. These techniques are based on neural networks and can handle complex, non-linear, and non-stationary data. They can also handle large and complex datasets, but require a significant amount of data and computational power to train. They are typically used for long-term forecasting and can capture complex patterns in the data.

In the following section, we will acquire a more profound comprehension of the mechanisms underlying specific algorithms, namely Linear Regression, SARIMA, and LSTM, which will be expounded upon in subsequent sections.

A. Linear Regression

Linear regression is one of the simplest algorithms, with low requirement for running power, and no condition for hyperparameter tunning. Currently, it is also the most commonly use in various area [12].

There are many assumptions under this method including:

- The linear relationship between variables
- Independence: There is no correlation between residuals in time series data
- Homoscedasticity: The assurance that variances at every level of x remain constant
- Normality: The assurance that residual have normal distribution

This approach is to seek out the most efficient link between variables, which can result in the least errors. The formula of Linear Regression is described as

$$Y_i = \beta_0 + \beta_1 * X1_i + \cdots \beta_n * Xn_i$$

While:

- Y_i = Dependent variable
- β_0 = Constant
- $\beta_1, \ldots \beta_n$ = Coefficient
- $X1_i, \ldots Xn_n$ = Independent variable

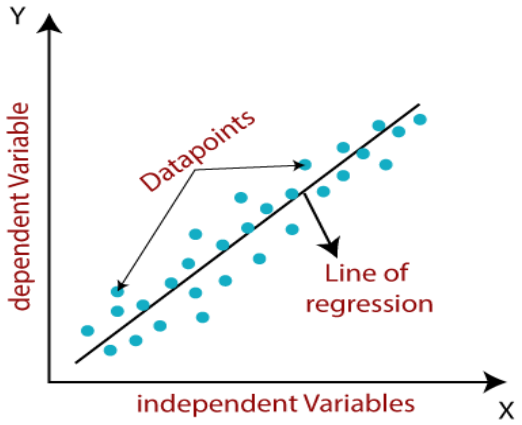

Fig. 1 Linear Regression [19]

B. ARIMA and SARIMA

1) Mechanism

ARIMA stands for Autoregressive Integrated Moving Average is one of statistical approaches on time series data for better analysis or forecasting the future value. Similar to Linear Regression, the model is based on linear relationship between current and previous time series data.

In comparison with Linear Regression, ARIMA is considered to be a better way for modelling data in time – series format. It belongs to univariate technique, however, if there is more than one variable that can explain the model, the Regression is better approach.

Other important premise for this model is that the series has to be strictly stationary. A stationary is the status that the data values is independent of the time in the series. In another word, a time series with seasonality, or trend are non – stationarity, which is much more complicated to be predicted under ARIMA mechanism [14].

For better understanding, the time – series is decomposed into 3 factors as follow:

- Auto regression (AR): Autoregressive is the model under the hypothesis that the previous data is highly correlated with current value. In an auto regression, the forecast is based on the linear of p (past values). The term of regression is illustrated as:

$$y_t = \beta_0 + \beta_1 * y_{t-1} + \cdots \beta_p * y_{t-p}$$

- Integrated (I): Integrated indicates the difference of raw observations to allow the time series to become stationary
- Moving Average (MA): Moving Average is a regression – like model, which use q previous forecasting errors as independent variable:

$$y_t = c + \alpha_1 * z_{t-1} + \cdots \alpha_p * z_{t-p}$$

While z = past forecast error

2) Non-seasonal ARIMA

The combination of 3 previous components represents non – season ARIMA model:

$$y'_t = c + \beta_1 * y'_{t-1} + \cdots \beta_p * y'_{t-p} + \alpha_1 * z_{t-1} + \cdots \alpha_p * z_{t-p}$$

While

- y'_t = the differenced series

ARIMA (p, q, d) are important hyperparameter of the model, while

- p = order of auto regressive part
- d = degree of first differencing involved
- q = order of moving average part

3) Seasonal ARIMA

One of the restrictions of ARIMA is that it doesn't include seasonality in the models. SARIMA is an extension of non – seasonal approach with the existence of seasonal terms, it is written as follow:

Non-seasonal elements:

- p = order of auto regressive part
- d = degree of first differencing involved
- q = order of moving average part

Seasonal elements:

- P = seasonal autoregressive order
- D = seasonal difference order
- Q = seasonal moving average order
- m = the number of time steps for a single seasonal period

C. Long short term memory (LSTM)

1) RNN

In traditional neural networks, input data is typically processed in a layered manner, with all input features being considered at the same time. However, in many real-world scenarios such as sequential data (e.g., text, audio, or time series), the order in which data is processed can significantly impact the results. To address this issue, a more sophisticated approach called Recurrent Neural Network (RNN) has been introduced. RNN allows for the processing of sequential data by incorporating a feedback loop, which allows the network to use previous output as input to the current layer. This way, the model can leverage information from previous inputs and generate better predictions for sequential data.

Figure 2 describes the overall mechanism of RNN

While:

- x = representation of sequential inputs (separated by times)
- x_t represents the time step at t, and y_t is the output of the step

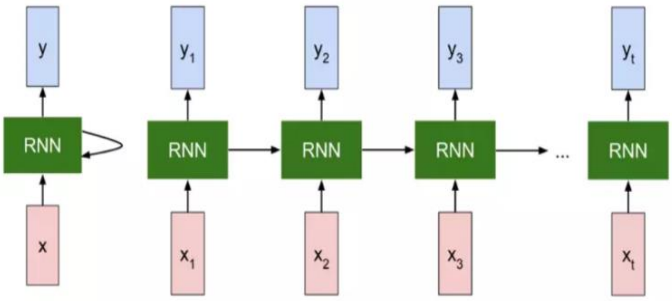

Fig. 2 RNN Mechanism [20]

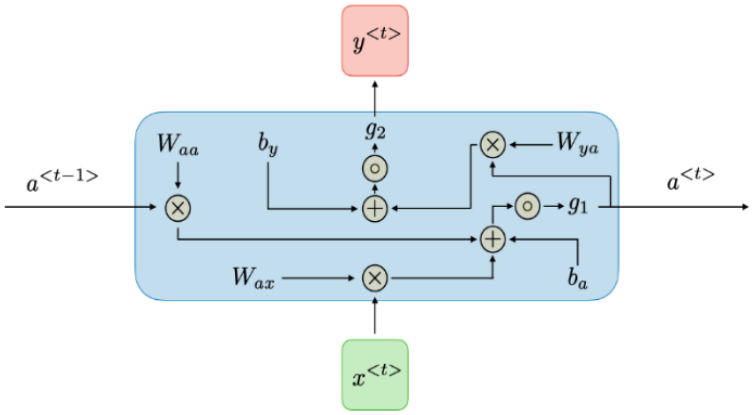

Fig. 3 What happen in each RNN step [20]

Figure 3 gives a clear explanation about what happens in each steps, which includes

- Hidden state at: is the memory of the neural network, which includes the previous memory at t-1 and the input as step t ($x^{<t>}$).
- Activation function is the function to defines the outputs of the step given the input. This function is usually under Tanh or ReLu.
- Output of the steps $y^{<t>}$). is the summary of information, and continue to be the input of following steps.

2) LSTM and Multivariate LSTM

LSTM is a type of RNN that shares many similarities with the latter. However, LSTM is superior in its ability to tackle the Vanishing Gradient problem, achieved by discarding irrelevant information. Unlike the simple repeating module in RNN that employs only a single tanh function, the repeating module in LSTM comprises four interacting layers, enabling better retention or forgetting of information in the long run (refer to figure 4, figure 5)

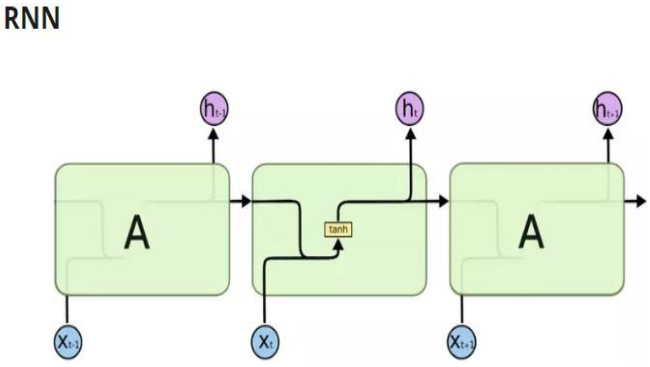

Fig. 4 RNN [16]

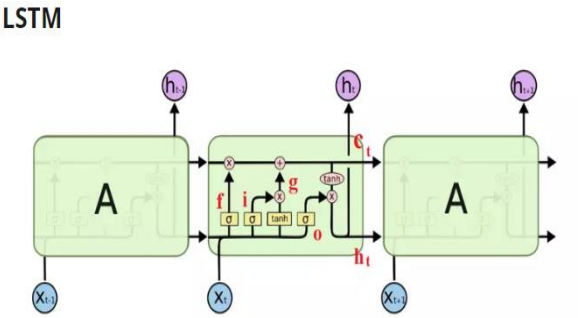

Fig. 5 LSTM [16]

In figure 5, each neuron contains one memory cell and three gates: input, output and forget. The functions of these gates are to safeguard the information by stopping or allowing the flow of it:

- **Input gate (i)** to determine which information from the input to be stored in the cell and return **g**
- **Output gate (o)** to determine the amount of output information and return y_{t}
- **Forget gate (f)** is useful to forget some prior values, i.e., it controls the extent to which a value remains in the cells due to some future works.

In addition, one of the significant advantages of LSTM network is their ability to handle multiple input variables seamlessly. This feature proves to be particularly beneficial in time series forecasting tasks, where conventional linear techniques such as ARIMA and SARIMA may struggle to accommodate multivariate or multiple input forecasting problems [16].

III. METHODOLOGY

A well-designed methodology is crucial for achieving the research objectives outlined in the introduction. Thus, in this study, a meticulous methodology has been developed to ensure reliable results. The methodology comprises several steps. Firstly, extensive research was conducted on the Ho Chi Minh Stock Exchange to identify a suitable list of stocks for later forecasting. Secondly, relevant data, including basic stock price indicators such as close price, high price, low price, open price, and trading volume, was collected for analysis. These indicators were also utilized as features for some of the multivariate models in the subsequent steps. The third step involved time-series analysis, including validating necessary hypotheses such as white noise, random walk, and stationarity, and data preprocessing, which included missing value detection and handling, data scaling, among others. The preprocessed data was then split into training and testing sets. Next, multiple hyperparameter tuning techniques were applied to each model to ensure optimal results. Three different models, namely Linear Regression, ARIMA, and LSTM, along with the benchmark of Naive Estimation, were then executed. Finally, the forecast of daily close price was validated based on R-squared (R2) and Mean Absolute Percentage Error (MAPE), and the best approach was identified.

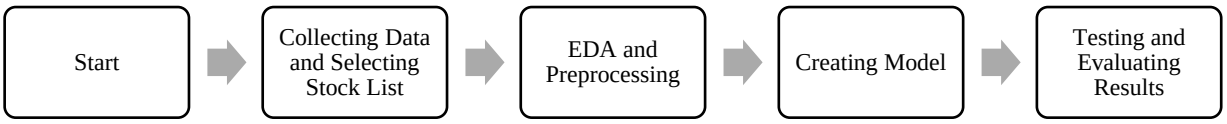

Fig. 6 Methodology Followed

A. Collecting Data and Selecting Stock List

To begin the study, a comprehensive analysis of the Ho Chi Minh Stock Exchange was conducted, consisting of 402 stocks during the period from January 2010 to March 2023, with approximately 3,280 trading days for each stock. Basic features including date (trading day), stock code, open, high, low prices, and trading volume were collected. A Naive Approach was then employed to estimate the stock price at time t based on the previous stock price (t-1). This approach served to test the Efficient Market Hypothesis (as the higher the performance of Naive Approach, the higher the capability of EMH) , provide a benchmark for future algorithms, and randomly select a sample of stocks for prediction based on the ranking of their Naive scores to ensure we select a sample that correctly reflects the market. Particularly, based on the MAPE of Naive Approach, we categorize stock market as 3 groups:

- Low: low estimation by Naive method with MAPE > 10%
- Medium: medium estimation by Naive method with MAPE 5% - 10%
- High: high estimation by Naive method with MAPE < 5%

The contribution for each group is described by figure 7

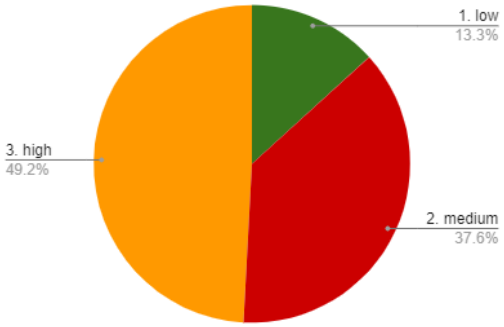

Fig. 7 Contribution of Stock Group

After that, nine companies were selected for the final list, with three randomly selected from each of the three categorized groups based on their MAPE scores using the Naive Approach: low (MAPE > 10%), medium (MAPE between 5% and 10%), and high (MAPE < 5%). The selected companies, including APG, VIP, TMT, DTL, TTE, PET, PHR, BCM, and NSC, were chosen to represent a diverse range of industries, price levels, and market capitalizations, all of the information is presented in table 2

TABLE 2
STOCK INFORMATION

STOCK GROUP	STOCK	INDUSTRY	MARKET CAPITAL (VND)	PRICE LEVEL (VND)
Low	APG	finance	1,000 billion	7,000
	VIP	energy	787 billion	11,000
	TMT	transportation	693 billion	18,000
Medium	DTL	manufacturing	1,970 billion	32,000
	TTE	energy	1,568 billion	12,000
	PET	energy	2,490 billion	25,000
High	PHR	manufacturing	5,514 billion	40,000
	BCM	construction	82,000 billion	79,000
	NSC	manufacturing	1,237 billion	70,000

B. Data Exploratory and Preprocessing

Data exploratory and pre-processing are critical components of time series analysis and forecasting. Exploratory data analysis, such as visualization, provides a means to identify data characteristics, validate hypotheses, and discover patterns that may inform the selection of appropriate forecasting techniques. Following exploratory data analysis, data pre-processing techniques are applied to enhance data quality, including the transformation of data to a suitable form when hypotheses are not passed, detection and handling of missing values, scaling of data to reduce the impact of outliers, and the normalization of different unit ranges among features. These pre-processing techniques are essential for ensuring the accuracy and reliability of the forecasting models used to analyse the time series data.

1) Hypothesis about White Noise

White noise is a concept stating that “the time series is generated randomly, then there is zero correlation between the data points. Therefore, it is impossible to make a prediction based on previous prices” [17]. An appearance of white noise is confirmed if:

- All variables have the same variance throughout a specific time dimension
- It is difficult to find correlation with all other values in the series

Visualization reveals that variance of stock price varies based on monthly basis (figure 8, figure 9, figure 10). In conclusion, these illustrations undermine the hypothesis of white noise.

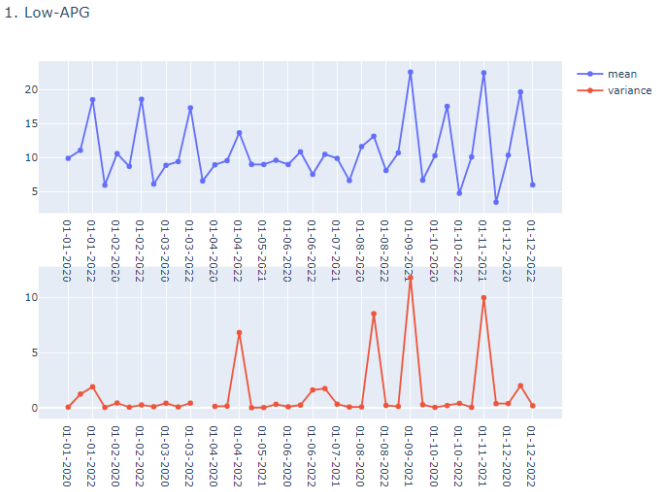

Fig. 8 Mean and Variance of Stock in Low group (APG)

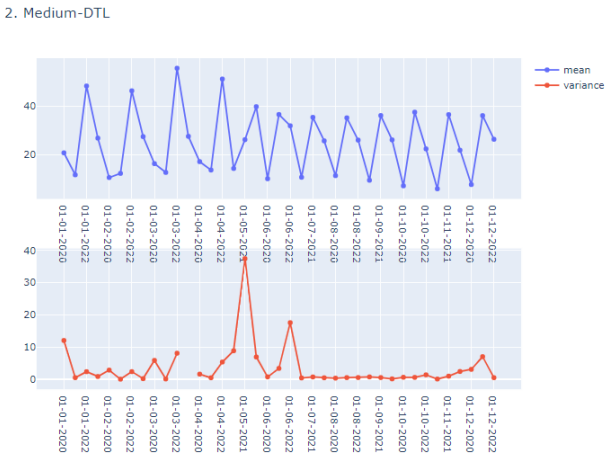

Fig. 9 Mean and Variance of Stock in Medium group (DTL)

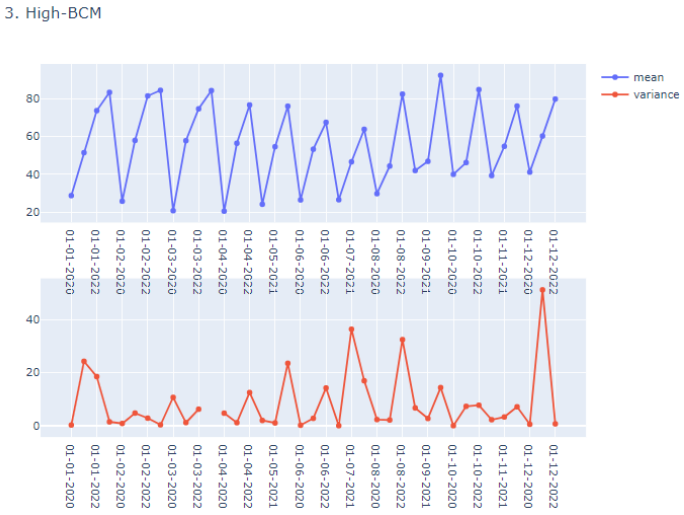

Fig. 10 Mean and Variance of Stock in High group (BCM)

2) Hypothesis about Random Walk

The term Random Walk was initially introduced by Karl Person in 1905. The random walk is different from random number (white noise) as the following value is a modification of previous value by a random function [17]. If a stock is proven to have a random walk, most of the information is currently reflected in the market at t - 1, and the best model is Naive forecasting (which we have introduced previously). Therefore, complicated machine learning is unnecessary in this case.

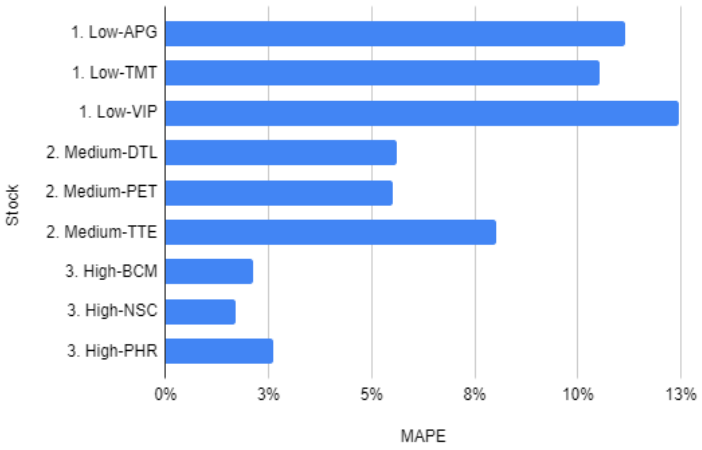

Fig. 11 Naive Ranking by Stocks

3) Hypothesis about Stationarity

If white noise and random walk are hypotheses to detect whether data is "predictable", stationarity is an essential hypothesis to be checked for ARIMA learning.

In case of stationarity, the data is one whose statistical properties such as mean, variance, covariance, and standard deviation do not vary with time [17]. As another explanation, time series does not include any "trend" or "seasonality".

The reason is that ARIMA only decomposes characteristics of previous data (Auto Regressive and Moving Average) but not the overall trend or seasonality, so that it is only suitable with simple sequence known as stationarity. In other words, the time series which is much more complicated (non - stationarity) will need many preprocessing skills to decompose the trend for better learning.

One of the most common techniques to recognize non - stationarity data is using hypothesis testing, specifically, Augmented Dickey-Fuller test is a type of one.

ADF is based on the following conclusion:

- Null Hypothesis (H0): Series is non - stationary
- Alternative Hypothesis (Ha): Series is stationary

If the Null Hypothesis is failed to reject (test statistics < critical value and p-value < 0.05), we reject that data is non - stationary, which means it is stationary

TABLE 3
ADF TEST FOR CLOSE VARIABLE

STOCK	ADF STATISTIC	P VALUE
1. Low-APG	-3.60	0.6%
1. Low-TMT	-2.78	6.1%
1. Low-VIP	-3.08	2.8%
2. Medium-DTL	-2.16	22.0%

2. Medium-PET	-2.35	15.7%
2. Medium-TTE	-3.75	0.3%
3. High-BCM	-0.71	84.3%
3. High-NSC	-1.68	44.1%
3. High-PHR	-2.06	26.2%

ADF Testing reveals that data is non - stationarity (p-value > 5%) for most of the stock (TMT, DTL, PET, BCM, NSC, PHR). Thereby, To decompose the season or trend, an easy approach of differentiating price (t) with the previous price (t-1) is applied, new variable names close_diff_1. Then close_diff_1 is examined with ADF (table 4) and ensures that the difference handles non - stationarity efficiently.

TABLE 4
ADF TEST FOR CLOSE_DIFF_1_VARIABLE

STOCK	ADF STATISTIC	P VALUE
1. Low-APG	-14.6	0.00%
1. Low-TMT	-12.5	0.00%
1. Low-VIP	-15.0	0.00%
2. Medium-DTL	-9.6	0.00%
2. Medium-PET	-11.4	0.00%
2. Medium-TTE	-19.2	0.00%
3. High-BCM	-8.0	0.00%
3. High-NSC	-33.9	0.00%
3. High-PHR	-16.3	0.00%

4) Data Preprocess

The data used in this study covers the period from January 2010 to March 2023, comprising approximately 3,280 trading days for each of the stocks considered. To address the issue of missing stock prices on weekends and holidays, the time series is pre-filled with the previous day's stock price. Additionally, to eliminate the potential impact of variables with different unit ranges in the model, the volume feature is scaled using the Min-Max scaling technique. Subsequently, the preprocessed data is split into training and testing sets in an 80:20 ratio for model optimization, learning, and prediction purposes.

C. Creating Model

1) Linear Regression

The application is Ridge Regression, which is an extension of Linear Regression by adding a penalty term equivalent to the square of magnitude of coefficient.

Loss function = OLS + alpha*summation (squared coefficient values)

According to this equation, alpha will reduce the complexity of the model by shrinking OLS parameters. Therefore, it is mostly used to prevent multicollinearity.

In term of Ridge Regression hyperparameter, a specific stock (APG) is selected for tunning process. Then a space (list of scenarios) of alpha is identified. After that, we use grid search cv to simulate these situations and find the best parameter. Finally, the best parameter (alpha) is applied to all of the other stocks in list.

2) SARIMA

In the context of SARIMA modelling, we utilized the "pmdarima.arima.auto_arima" pre-defined function in Python for the purposes of hyperparameter tuning and model training. To elaborate further, we executed the

optimization and learning processes on each stock in our dataset. Specifically, we provided a range of values for the Auto Regressive (p), Moving Average (q), and seasonal (seasonality) parameters, which also included P and Q factors. The function then automatically generated and trained the model based on the optimal scenario determined by the hyperparameter tuning process.

3) LSTM & Multivariate LSTM

With reference to LSTM, Univariate and Multivariate are both applied. The best scenario is tested based on a list of parameters described below:

- Number of n steps is the number of lagged features used to make prediction
- Number of training epochs is the number of times the entire data set has to be worked through the learning algorithm. The higher the number, the more complex the model, as well as the capability of learning better on training dataset. However, this situation will also increase the possibility of overfitting.
- Batch size is a number of samples processed before gradient descent update (its loss function).
  Larger batch size will result in faster training time. However, it can lead to local minimum problems or not converge at fast.

D. Testing and Evaluating Results

This is the final stage of the process. All of the results from 3 models Ridge Regression, SARIMA, and LSTM will be evaluated by MAPE (Mean Absolute Percentage Errors) and MAE (Mean Square Error).

Indeed, MAPE is a widely used metric to evaluate the accuracy of a stock price regression model. It measures the percentage difference between the actual and predicted values of the stock prices. MAPE is calculated as the average of the absolute percentage errors between the actual and predicted values, which makes it more interpretable and helps to compare across different stock pricing units.

However, the con of MAPE is that it tends to heavily penalize large forecast errors. This is because the metric is calculated as a percentage, so a large absolute error on a small actual value will result in a larger MAPE than the same absolute error on a larger actual value. To cover this disadvantage, a combination with MAE is suitable. Particularly, the fact that MAE is calculated based on volume errors will eliminate the risk that overweight the errors of really small actual value we have mentioned previously.

The MAPE is defined as:

$$MAPE = \frac{1}{n}\sum |1 - \frac{F_t}{A_t}|$$

While:

- A_t = Actual value
- F_t = Forecast value

The MAE is defined as:

$$MAE = \frac{1}{n}\sum |F_t - A_t|$$

IV. CONCLUSION

The result shows that traditional approaches with linear assumptions (Linear regression and SARIMA) are still the best forecast method while LSTM (univariate LSTM) underperforms other methods. In term of percentage error, MAPE for testing set varies from 1.2% - 1.9% for linear approaches while this performance is from 1.9% - 3.4% for LSTM algorithms. All of the algorithms are 2-3 time better than Naive estimation by reducing the MAPE from 6.4% to 1.4% (reference to linear method), and 2.3% (reference to LSTM).

TABLE 5
MAPE RESULT

VALIDATION	MAPE (MEAN ABSOLUTE PERCENTAGE ERRORS)				
STOCK	NAIVE	LINEAR REGRESSION	SARIMA	UNIVARIATE LSTM	MULTIVARIATE LSTM
1. Low-APG	10.90%	1.90%	1.80%	3.40%	2.30%
1. Low-TMT	10.50%	1.90%	1.80%	2.30%	2.20%
1. Low-VIP	12.20%	1.40%	1.50%	2.20%	2.00%
2. Medium-DTL	5.50%	1.20%	1.10%	1.80%	3.40%
2. Medium-PET	5.20%	1.20%	2.00%	2.40%	1.70%
2. Medium-TTE	7.90%	1.60%	0.80%	3.60%	3.00%
3. High-BCM	1.90%	1.50%	1.20%	3.00%	1.70%
3. High-NSC	1.60%	1.10%	0.70%	3.10%	1.70%
3. High-PHR	2.30%	1.20%	1.30%	1.90%	2.80%
Average	**6.40%**	**1.40%**	**1.40%**	**2.60%**	**2.30%**

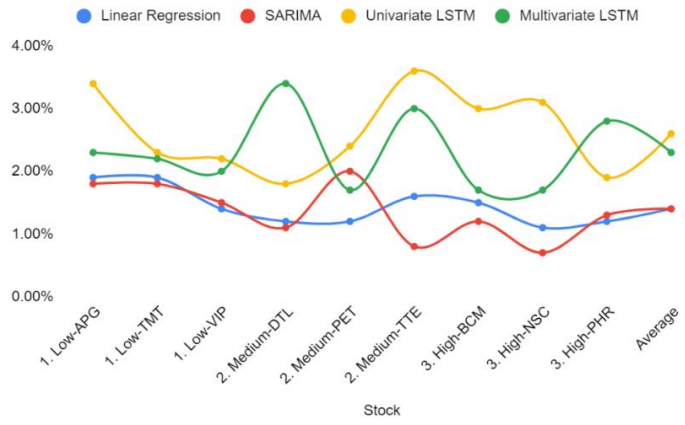

Fig. 12 MAPE Result

Indeed, measurement by MAE (Mean Absolute Error) reveals that Linear Regression is that best method with 0.34 error while this figure was 0.57 for LSTM. Furthermore, Linear approach still have a better result in comparison with LSTM calculating.

TABLE 6
MAE RESULT

VALIDATION	MAE (MEAN ABSOLUTE ERRORS)				
STOCK	NAIVE	LINEAR REGRESSION	SARIMA	UNIVARIATE LSTM	MULTIVARIATE LSTM
1. Low-APG	1.01	0.15	0.21	0.23	0.17
1. Low-TMT	1.02	0.28	0.22	0.38	0.32
1. Low-VIP	1.00	0.13	0.13	0.23	0.19
2. Medium-DTL	1.07	0.23	0.28	0.41	0.75
2. Medium-PET	1.15	0.24	0.55	0.49	0.32
2. Medium-TTE	0.98	0.20	0.11	0.45	0.38
3. High-BCM	1.48	0.59	0.93	1.13	0.70

3. High-NSC	1.23	0.78	0.59	2.10	1.16
3. High-PHR	1.31	0.46	0.77	0.89	1.11
Average	**1.14**	**0.34**	**0.42**	**0.70**	**0.57**

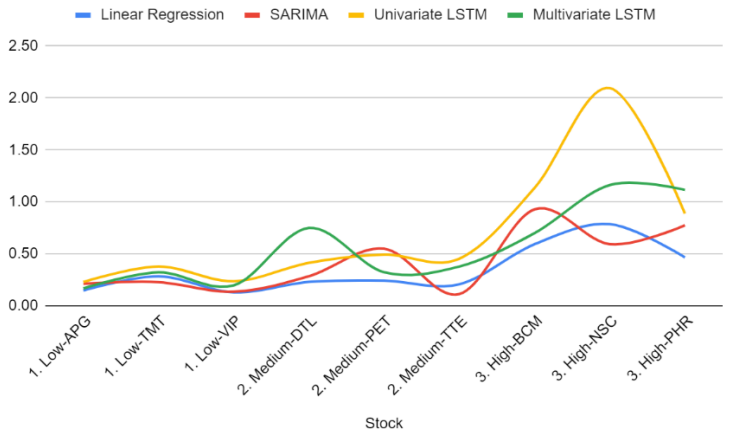

Fig. 13 MAE Result

Overall, our study provides many evidences. Firstly, the disapproval of Efficient Market Hypothesis in HOSE by demonstrating that both Traditional Machine Learning and Deep Learning can outperform the Naive forecast. Secondly, Traditional Machine Learning methods like Linear Regression and SARIMA perform well, Deep Learning (LSTM) fails to improve upon them. This may be due to the young age of the Vietnamese stock market and its relatively small volume of daily price data, which may not provide sufficient data for LSTM to train effectively. Additionally, the unregulated nature of the Vietnamese stock market can result in speculation and high volatility, which can lead to noise in the data. While traditional methods may be better suited to handle this noise, the sophisticated mechanism of LSTM may not be able to filter it out effectively.

These findings suggest that there is still much room for research on stock price prediction in Vietnamese market, and that the choice of model should be carefully evaluated based on the characteristics of the data. Particularly, researchers should consider the age and volume of the stock market, as well as the level of regulation and volatility when selecting a model for prediction.

V. FUTURE SCOPE

The increasing demand for the application of machine learning in finance, particularly in stock trading, has led to an interest in time series data analysis in the Vietnamese market. This study has contributed valuable insights into the efficacy of traditional modeling techniques, which have proven to be superior to more complex models such as LSTM. Despite this, there remain several limitations that future research should aim to address. Firstly, the short-term forecasting windows used in this study may not be suitable for real-world trading, as many investment strategies operate on longer time frames, such as weekly, monthly or yearly. Longer forecasting windows would provide a more comprehensive picture of the prediction problem, albeit with increased complexity. Secondly, while forecasting prices is a useful exercise, it is not practically applicable in the context of real-world trading. Future work could combine prediction and portfolio optimization problems to develop efficient investment strategies. Thirdly, this study focuses solely on the HOSE exchange. Future research should aim to expand these results to other exchanges in Vietnam, such as HNX or UPCOM, for more comprehensive insights into the behavior of the Vietnamese stock market. Finally, this study only considers historical stock price forecasting. Future research could expand on this by exploring other sources of data, such as technical indicators or sentiment features, and testing various models, including other traditional methods like SVR or decision trees, and more complex models such as RNNs, CNNs, GRUs, or even hybrid models, in order to further improve forecasting accuracy.

ACKNOWLEDGEMENT

The authors would like to thank to Thuyloi University (TLU) and Vietnam Center of Research in Economics, Management and Environment (VCREME).

REFERENCES

[1]. Malini, H., 2019. Efficient market hypothesis and market anomalies of LQ 45 index in Indonesia stock exchange. *Sriwijaya International Journal of Dynamic Economics and Business*, *3*(2), pp.107-121.

[2]. Mubarok, F. and Fadhli, M.M., 2020. Efficient Market Hypothesis and Forecasting in the Industrial Sector on the Indonesia Stock Exchange. *Journal of Economics, Business, & Accountancy Ventura*, *23*(2), pp.160-168.

[3]. Jebran, K., Chen, S., Ullah, I. and Mirza, S.S., 2017. Does volatility spillover among stock markets varies from normal to turbulent periods? Evidence from emerging markets of Asia. *The Journal of Finance and Data Science*, *3*(1-4), pp.20-30.

[4]. Henrique, B.M., Sobreiro, V.A. and Kimura, H., 2018. Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of finance and data science*, *4*(3), pp.183-201.

[5]. Vijh, M., Chandola, D., Tikkiwal, V.A. and Kumar, A., 2020. Stock closing price prediction using machine learning techniques. *Procedia computer science*, *167*, pp.599-606

[6]. Alkhatib, K., Khazaleh, H., Alkhazaleh, H.A., Alsoud, A.R. and Abualigah, L., 2022. A new stock price forecasting method using active deep learning approach. *Journal of Open Innovation: Technology, Market, and Complexity*, *8*(2), p.96.

[7]. Nahil, A. and Lyhyaoui, A., 2018. Short-term stock price forecasting using kernel principal component analysis and support vector machines: the case of Casablanca stock exchange. Procedia Computer Science, 127, pp.161-169.

[8]. Zhang, L., Aggarwal, C. and Qi, G.J., 2017, August. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2141-2149).

[9]. Zhang, L., Aggarwal, C. and Qi, G.J., 2017, August. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2141-2149).

[10]. Matsuba, I., 1991, November. Application of neural sequential associator to long-term stock price prediction. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks* (pp. 1196-1201). IEEE.

[11]. Soni, P., Tewari, Y. and Krishnan, D., 2022. Machine Learning approaches in stock price prediction: A systematic review. In *Journal of Physics: Conference Series* (Vol. 2161, No. 1, p. 012065). IOP Publishing.

[12]. Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021. *Introduction to linear regression analysis*. John Wiley & Sons.

[13]. Singh, S., Rehan, S. and Kumar, V., 2022. Stock Price Prediction Using Linear Regression, LSTM and Decision Tree. *Easy Chair Preprint*, (7805).

[14]. The otexts website. [Online]. Available: https://otexts.com/fpp2/AR.html

[15]. The otexts website. [Online]. Available: https://otexts.com/fpp2/seasonal-arima.html

[16]. Zhang, R., Guo, Z., Meng, Y., Wang, S., Li, S., Niu, R., Wang, Y., Guo, Q. and Li, Y., 2021. Comparison of ARIMA and LSTM in forecasting the incidence of HFMD combined and uncombined with exogenous meteorological variables in Ningbo, China. *International journal of environmental research and public health*, *18*(11), p.6174.

[17]. Brownlee, J., 2017. *Introduction to time series forecasting with python: how to prepare data and develop models to predict the future*. Machine Learning Mastery.

[18]. Phuong Mai Pham, Quang Chieu Ta, 2022. Stacked Long Short-Term Memory for Vietnamese Stock Market Returns Prediction. International Journal of Applied Sciences: Current and Future Search Trends

[19]. The javatpoint website. [Online]. Available: https://www.javatpoint.com/linear-regression-in-machine-learning

[20]. The analyticsvidhya website. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/

