UNIVERSITY OF INFORMATION TECHNOLOGY
FACULTY OF INFORMATION SYSTEMS

Using Statistical Model And Machine Learning to Predict Stock Price

• NGUYỄN MINH THÀNH - 20521920
• NGUYỄN VĂN TÂN - 20521880
• NGUYỄN PHAN HIẾU THUẬN - 20521994

Using Statistical Model And Machine Learning

To Predict Stock Price

Nguyen M. Thanh
STAT3013.N11.CTTT-EN
University of Information Technology
20521920@m.uit.edu.vn

Nguyen V. Tan
STAT3013.N11. CTTT-EN
University of Information Technology
20521880@gm.uit.edu.vn

Nguyen P. H. Thuan
STAT3013.N11. CTTT-EN
University of Information Technology
20521994@gm.uit.edu.vn

Abstract— Recently, because of the market economy's ongoing growth, thus more and more people are participating in the stock market. Since the prediction of stock market trends is very complex, investors can decrease their losses and improve their profits by purchasing and selling stocks at a certain price with the use of precise forecasts and simulation tools. In this study, since the data used for prediction and analysis are nonlinear and have time-dependent issues, we will use nonlinear, linear, ARIMA, and machine learning regression as LSTM, Prophet to develop a stock price prediction model. Here, we use historical stock price data of Viet Nam companies. As a result of the LSTM model we built, in 8 - 2 train-test percent, we achieve the greatest accuracy of around MAPE 5.08% for forecasting HAG stock , 5.84% for HPG and MAPE in 7-3 train-test percent HAG – 6.17% and HPG – 6.42 % .

Keywords—LSTM, Deep Learning, stock price, regression

I. INTRODUCTION

Stocks are securities that confirm the owner's lawful rights and interests to part of the stock capital of the issuing organization. In 1602, the Dutch East India Company issued its first shares through the Amsterdam Stock Exchange, and it was the first company to issue stock and bonds.

The purpose of stock valuation is to determine the true value of a stock at a given point in time, find out the potential of the stock, and make relevant investment decisions. For businesses, stock valuation is one of the important steps of a joint stock company when it wants to offer shares, raise capital and increase its influence in the market. For investors, stock valuation helps investors know which stocks are worth buying and have the greatest return potential. An easy way to do that is to gauge how much the stock is worth. Then, we will proceed to buy the stock if the share price is lower than the value we value. Or sell the stock (if the investor owns the stock) if the share price is already higher than the valuation to make a profit.

There are many algorithms and techniques that help us predict prices. In this paper, we will use Linear Regression, Non-Linear Regression, ARIMA, LSTM, and Prophet models to predict the stock prices of some corporations in the next 30 days. From there, evaluate and compare the above 5 models.

II. RELATED WORK

Stock prediction is an interesting and necessary topic, so there have been many studies and each study give different results based on each model that the research builds.

To forecast stock index, Michael van Gysen et al. [3] employ both linear and non-linear models. They then

execute two models and assert that the nonlinear model is inferior to the linear one.

Machine learning techniques in this area have been shown to improve efficiency by 60-80% compared to previous methods [4]. The ARIMA model is widely used to predict linear time series data. ARIMA is integrated from other models such as Automatic Regression (AR), Integrated (I) and Moving Average (MA). The accuracy of the ARIMA model to predict stock prices and the accuracy of over 85%, we see that the ARIMA model is also a good model [5].

LSTM by default can retain information for a long period of time. it is used for processing, prediction and classification on time series database. [6] In 2015, Roondiwala e.t used LSTM to predict stock prices, and the results achieved the best results with training RMSE of 0.00983 and testing RMSE of 0.00859, so the efficiency of network LSTM is proved positive and better in forecasting with time series.

In another article, Mashtura[7] researched and compared Facebook Prophet with other machine learning models to find out the limitations as well as the good of the Prophet model and suggested that Prophet was given top priority because of its ease of use and its accurate.

III. METHOD

A. Data Collection

Investing Finance is the source of historical data for three of the largest businesses in Vietnam. The data set includes five years for Hoang Anh Gia Lai (HAGL), and Hoa Phat Group (HPG). from September 7, 2015 to November 30, 2022. The data will contain stock-related information like High, Low, Open, Price, Volume, and change%. Divide the data into training and test sets, and then divide the training set and test set according to one of the following ratios: 80-20%, and 70-30%. The model will be created using the training set, and its performance will be assessed using the test set.

Following the data collection activity, certain pre-processing techniques are used to enhance the data set's quality. The

knowledge discovery process contains several steps, including data cleansing, feature selection, data reduction, and data transformation.

In this study, we visualize the data of the data package about three Vietnamese enterprises by using the Matplotlib tool in Python. Here is a subchapter to create graphs that display it in the data. The closing price of a common stock determines the profit or calculates the loss for a day. So, plot the target variable to understand how it forms in the data (figure 1, figure 2).

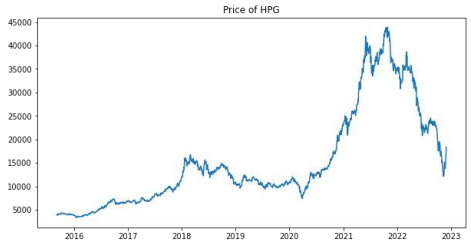

Figure 1 Visualize data of HPG

Figure 2 Visualize data of HAG

Mean, the standard deviation, maximum, and minimum of the data, as shown in (Table 1, Table 2)

Table 1 Descriptive statistics of HPG

	price	Open	High	Low
count	1812.00000	1812.000000	1812.000000	1812.000000
mean	15288.53819	15294.660375	15508.131623	15089.505408
std	10456.63874	10482.328760	10623.617303	10328.726382
min	3280.00000	3187.400000	3293.200000	3161.000000
25%	7739.47500	7728.450000	7808.575000	7660.925000
50%	11578.70000	11605.100000	11727.400000	11435.350000
75%	20906.90000	20704.500000	21422.550000	20470.600000
max	43895.80000	43895.800000	44198.500000	43517.400000

Table 2 Descriptive statistics of HAG

	price	Open	High	Low
count	1812.000000	1812.000000	1812.000000	1812.000000
mean	6989.701987	7000.524283	7141.986755	6866.313466
std	2831.815190	2840.975405	2915.588395	2769.333004
min	2550.000000	2450.000000	2630.000000	2400.000000
25%	5040.000000	5040.000000	5100.000000	4980.000000
50%	5870.000000	5900.000000	6005.000000	5765.000000
75%	8400.000000	8400.000000	8550.000000	8290.000000
max	15650.000000	15950.000000	16200.000000	15300.000000

B. Learning Algorithms

In this paper, we use 2 algorithms of machine learning (LSTM and Prophet) and 3 statistical models (Linear, Nonlinear regression, and ARIMA) for predicting stock price.

A regression model is an approach used to express the relationship between two variables; one is called independent variable X which can be more than one and the other scalar variable Y is dependent on X.

$$Y = f(X, b) \quad (1)$$

Linear and nonlinear regression are statistical models, and both assume that there exists a relationship between dependent and independent variables. Just making predictions about linear rise/fall cycles cannot be accurate because the relationship between the independent variable and the dependent variable becomes more complex. Non-linear models can be more accurate than linear models as the relationship between two variables becomes more complex, but they can also be more difficult to interpret and may require more data to train.

a) Linear regression

Linear regression is a method to examine the relationship between two variables, where the relationship is represented by the equation below this method.

$$Y = \beta_0 + \beta_1 * X_1 + \ldots + \beta_n * X_n + \varepsilon \quad (2)$$

Where

Y is a dependent variable

X_1 to Xn are independent variables

ε represents a random error

β_0 is the intercept

$\beta_{1 \text{ to }} \beta_{n \text{ is}}$ the slope of the line

A linear regression equation written in vector form [8] is

$$Y = a + \beta * X + \varepsilon \quad (3)$$

b) Non-Linear regression

Nonlinear regression is a form of regression analysis in which observed data is modeled by a function that is a non-linear combination of the model parameters and depends on one or more independent variables.

In nonlinear regression, a statistical model is in the form of:

$$y \sim f(x, \beta)$$

relative to a vector of independent variables, x, and it's associated observed dependent variables, y. The function f is nonlinear in the components of the vector of parameters β, but otherwise an arbitrary function.

c) ARIMA

ARIMA stands for Autoregressive Integrated MovingAverage [9]. The ARIMA model is based on the Box and Jenkins method of using three different concepts: autoregressive (AR) model, moving average (MA) model, and integration, together classified as an ARIMA (p, d, q). It is a quantitative forecasting model over time, the future value of the predictor variable will depend on the movement trend of that object in the past.

where

- p defines the AR
- d defines the differential
- q defines the MA.

Three components/parameters: AR + I + MA

AR is denoted as p, where it shows the weighted linear of sum p values based on ARIMA (p, d, q) terminology. The p-value indicates the number of orders. The formula to denote this AR is shown (4)

$$Y_t = \beta 1 * Yt - 1 + .. + \beta p * Yt - p + \varepsilon = \varepsilon + \sum_{j=1}^{p} (\beta_j * Y_{t-j})$$

(4)

Where p is used to determine the number of orders of past values; t is the time series; β is the slope coefficient of the AR model; ε is the error term with mean zero and variance σ^2.

MA process is denoted by order q in the ARIMA (p, d, q) classification which shows an error value in (4), it also uses the number of orders in the past values, as denoted in (5)

$$Yt = \alpha_1 * \varepsilon_{t-1} + .. + \alpha_q * \varepsilon_{t-q} + \varepsilon_t = \varepsilon_t + \sum_{j=1}^{p} (\alpha_j * \varepsilon_{t-j})$$

(5)

Where t is the time series; α is the slope coefficient; q is the number of orders needed to identify the past values. To identify how many orders are in the calculation of AR, the parameter of q is used.

Integrated or differentiated versions are denoted as d in ARIMA (p, d, q), which is the number of time the time series got differenced .

$$I(1) = \nabla y_t = y_t - y_{t-1}$$

$$I(d) = \nabla^d (y_t) = \nabla (\nabla (..... \nabla ((y_t)))$$

Below are 2 expressions (7) ,(8) to show the difference when there is no difference and when there is no between ARMA(p,q),or can write ARMA(p,0,q) because difference is zero and ARIMA(pd, q)

$$Y_t = \varepsilon_t + \sum_{j=1}^{p} (\alpha_j * \varepsilon_{t-j}) + \sum_{j=1}^{p} (\beta_j * Y_{t-j})$$

Equation for an ARMA (p , q) model(7)

$$\tilde{N}Y_t = \varepsilon_t + \sum_{j=1}^{p} (\alpha_j * \varepsilon_{t-j}) + \sum_{j=1}^{p} (\beta_j * \nabla Y_{t-j})$$

Equation for an ARIMA (p , d , q) model(8)

d) LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that is well-suited for processing sequential data such as time series, natural language, and audio. LSTMs are able to learn long-term dependencies in data by using a special structure called a “memory cell”, which can store information over long periods of time.[6][11]

LSTM models are often used for time series prediction tasks, such as forecasting stock prices, weather, or energy consumption. In these tasks, the goal is to use past data to make predictions about future value.[12]

The structure of an LSTM cell is shown in the figure below. It consists of three interacting gates (input, output, and forget

gates) and a memory cell, which stores information. The gates control the flow of information into and out of the memory cell,

allowing the model to decide which information to keep and which to discard

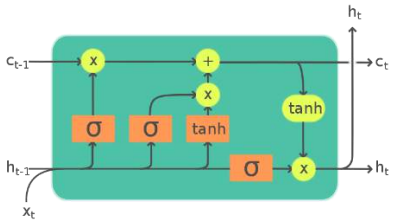

Figure 3 Structure of LSTM model

LSTMs can be trained to perform a wide range of tasks. They have proven to be highly effective for many tasks and continue to be a popular choice in the field of deep learning. However, they can be difficult to train on complex tasks and have a relatively high computational cost compared to other models.

e) Prophet

The prophet is an open-source time series forecasting library developed by Facebook that is designed to make it easy to forecast time series data using a simple, intuitive API. It is based on the idea of decomposing a time series into three components: trend, seasonality, and holidays, and then using these components to make predictions.[13]

The prophet is an additive model with the following components:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Prophet requires that the input data be in a specific format, with the time series data stored in a Pandas data frame with a column for the date with a “ds” (date stamp) and a column for the value with a “y” (value want to forecast). The dates should be stored as a DateTime object, and the values should be numeric.

IV. RESULT

ARIMA, LSTM, and Prophet. Stock price prediction of companies must follow Root Mean Square Error (RMSE) and mean absolute percentage error (MAPE) to find the most optimal error in price prediction.

RMSE measures the average difference between the predicted values and the true values, with lower values indicating better performance. MAPE is a measure of prediction accuracy for a forecasting method in statistics.

Table 3 Evaluation results for the five models on the time series prediction task for three companies

Stock	Model	Train-test	RMSE	MAPE
HAG	Linear	10 – 0	2816.24	34.49%
	Non-linear	10 - 0	1415.10	15.76%
	ARIMA	7 - 3	4638.5	33.05%
		8 - 2	4418.2	34.39%
	LSTM	7 - 3	665.72	6.17%
		8 - 2	644.44	5.08%
	Prophet	7 - 3	1383.82	14.82%
		8 – 2	1381.93	14.21%
HPG	Linear	10 – 0	6368.27	31.21%
	Non-linear	10 - 0	2505.02	12.34%
	ARIMA	7 - 3	16502.57	44.71%
		8 - 2	11843.59	41.76%
	LSTM	7 - 3	2408.96	6.42%
		8 - 2	2012.21	5.84 %
	Prophet	7 - 3	1846.16	9.40%
		8 – 2	1860.46	9.28%

Figure 4 Predictive results of the Linear model with the non-rate (%)

Figure 5 Predictive results of the Nonlinear model with the non-rate (%)

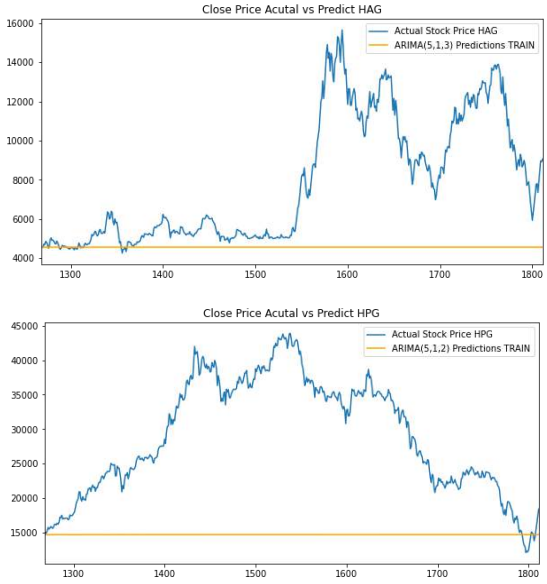

Figure 6 Predictive results of the ARIMA model with the rate of 7-3(%)

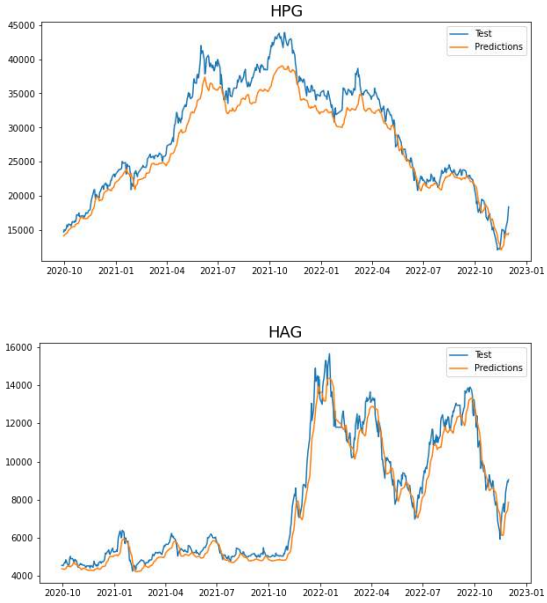

Figure 8 Predictive results of the LSTM model with the rate of 7-3 (%)

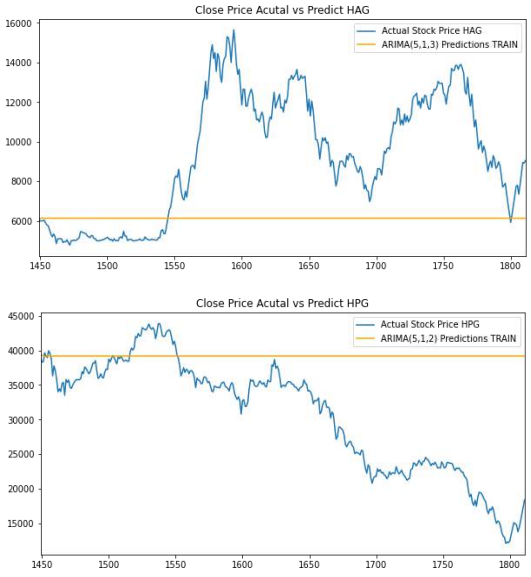

Figure 7 Predictive results of the ARIMA model with the rate of 8-2(%)

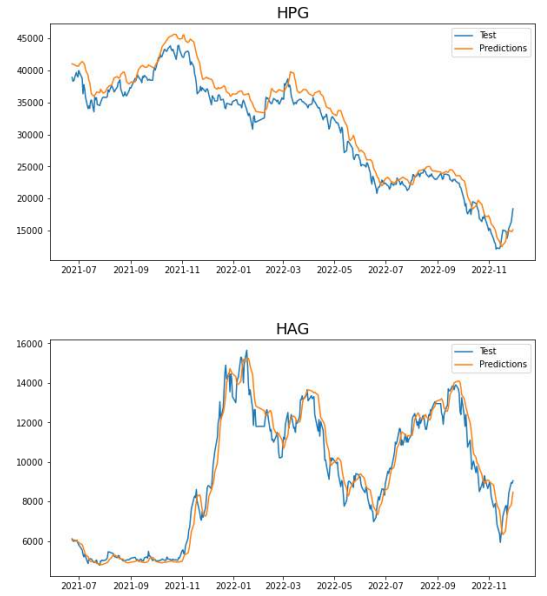

Figure 9 Predictive results of the LSTM model with the rate of 8-2 (%)

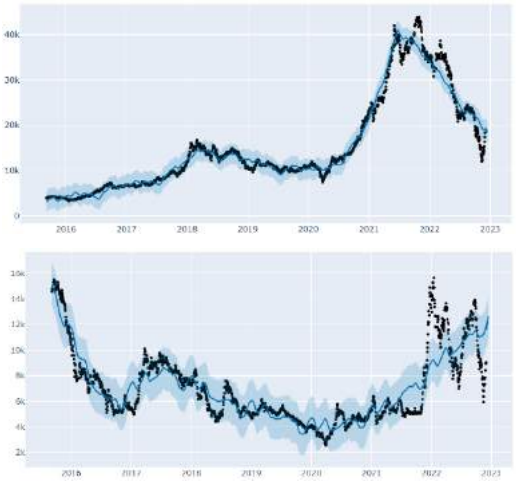

Figure 10 Predictive results of the Prophet model with the rate of 7-3 (%)

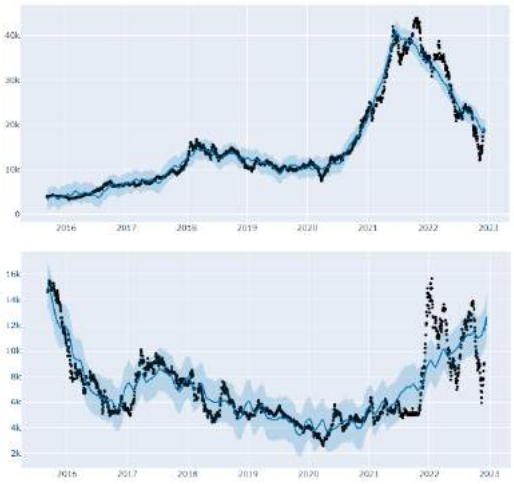

Figure 11 Predictive results of the Prophet model with the rate of 8-2 (%)

Following testing and evaluation of the outcomes for the test set with three rates, respectively, 10 - 0 for two models Regression models that are linear, and non-linear for two models. Table 3 displays 3 models with 2 ratios of 7-3 and 8-2 for 2 stocks, HPG and HAG, respectively: ARIMA, LSTM, and Prophet.

For HAG's closing price historical data set, with the results obtained from table 4, the LSTM model has the lowest RMSE value with a value of 644.44 when dividing the data with a ratio of 8-2 for train- test and 665.72 when scaling 7 - 3, with other models such as Prophet, ARIMA all having higher RMSE indexes, respectively, with a ratio of 7 - 3 1383.82, 6517.96; with a ratio of 8 - 2 1381.93, 4418.2 . Particularly, Linear and Non-linear models are the only two models that do not scale with the results 2816.24 and 1415.10.

In terms of MAPE value, the LSTM model has the lowest MAPE value with a value of 5.08%when dividing the data at the rate of 8-2 for the train test and 6.17%when dividing the scale from 7 to 3, with other models. Like Prophet, ARIMA both have higher MAPE with a ratio of 7 - 3, 14.82%, and 33.05%, respectively; with a rate of 8 - 2, 14.21%, and 34.39%. Particularly, Linear and Non-linear models are the only two models that do not scale with the results 34.49% and 15.76%.

For the historical data set of closing prices of HPG, with the results obtained from Table 4, the Prophet model has the lowest RMSE value with a value of 1860.46 when dividing the data with a ratio of 8-2 to train- test and 1846.16 when scaling 7 - 3, with other models such as LSTM, ARIMA all have higher RMSE indexes with the ratio 7 - 3 2408.96, 4638.54 respectively; with a ratio of 8 - 2 2012.21 , 4418.2 . Particularly, Linear and Non-linear models are the only two models that do not scale with the results 6368.27 and 2505.02.

In terms of MAPE value, the LSTM model has the lowest MAPE value with a value of 5.84% when dividing the data at the rate of 8-2 for the train test and 6.42% when dividing the ratio of 7 - 3, with other models. Like Prophet, ARIMA both have higher MAPE with the rate of 7 - 3, 9.4%, and 44.71%, respectively; with the rate of 8 - 2, 9.28%, and 41.76%. Particularly, Linear and Non-linear models are 2 models. the only shape that doesn't scale with 31.21% and 12.34% results.

In general, the LSTM model is the best of all the 5 models with the RMSE smaller than almost and smallest MAPE. the remaining models like Prophet, ARIMA, Linear, and Non-linear have higher results but they will be useful in certain cases. From the conclusions on the model, LSTM is a suitable model for predicting future prices in the next 30 days of 2 stocks HPG and HAG.

V. Discussion

According to the experimental comparison of the predicted values of the HPG and HAG stock portfolios from December 1 to December 30, 2022, we obtained the predicted results for the next 30 days of trading, shown in Figure 12 and Figure 13 below.

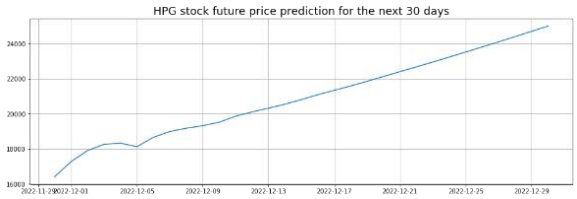

Figure 12 HPG stock future price prediction for the next 30 days

Table 4 Survey of 22 real trading days of HPG

date	price	future price	Change %
12/1/2022	18200	17278.8	5.06%
12/2/2022	19450	17910.36	7.92%
12/5/2022	20000	18126.54	9.37%
12/6/2022	18600	18664.62	0.35%
12/7/2022	18250	18992.74	4.07%
12/8/2022	18900	19182.47	1.49%
12/9/2022	19200	19326.22	0.66%
12/12/2022	18600	20110.78	8.12%
12/13/2022	19000	20323.58	6.97%
12/14/2022	19200	20552.56	7.04%
12/15/2022	19350	20815.12	7.57%
12/16/2022	20400	21096.15	3.41%
12/19/2022	20000	21860.37	9.30%
12/20/2022	19000	22135.75	16.50%
12/21/2022	18900	22411.19	18.58%
12/22/2022	18900	22681.29	20.01%
12/23/2022	18350	22958.16	25.11%
12/26/2022	17100	23819.29	39.29%
12/27/2022	18250	24110.79	32.11%
12/28/2022	18200	24407.71	34.11%
12/29/2022	18000	24709.01	37.27%
12/30/2022	18000	25013.42	38.96%

For the HPG dataset, the LSTM model was able to predict the trade price of 22 sessions from December 1, 2022 to December 30, 2022, as shown in Table 4. According to the calculation, the percentage error tends to increase from 5.06% for the first trade session and 12.85% for 20 trade sessions, and we believe that this percentage error will continue to increase due to the model's tendency to increase, as shown in Figure 12. However, overall, the model's prediction results for the first 7 days are closest to the actual price as of December 6, 2022 with an error of only 0.35%,

and the highest is 9.35% on December 5, 2022. In general, the model has an average percentage difference of 5.06%, 5.35%, 4.13% for the first 1, 5, and 7 days of trade prediction, respectively.

Figure 13 HAG stock future prediction for the next 30 days

Table 5 Survey of 22 real trading days of HAG

date	price	future price	Change %
12/1/2022	8810	8677.54	1.3246
12/2/2022	9110	8541.51	5.6849
12/5/2022	8900	8164.83	7.3517
12/6/2022	9520	8047.51	14.7249
12/7/2022	9180	7935.83	12.4417
12/8/2022	9200	7828.21	13.7179
12/9/2022	9410	7724.53	16.8547
12/12/2022	9010	7435.07	15.7493
12/13/2022	9400	7345.33	20.5467
12/14/2022	9250	7258.71	19.9129
12/15/2022	9130	7175.09	19.5491
12/16/2022	9250	7094.33	21.5567
12/19/2022	8950	6868.1	20.819
12/20/2022	8330	6797.69	15.3231
12/21/2022	8910	6729.6	21.804
12/22/2022	8900	6663.76	22.3624
12/23/2022	8620	6600.07	20.1993
12/26/2022	8030	6421.15	16.0885
12/27/2022	8590	6365.3	22.247
12/28/2022	8600	6311.24	22.8876
12/29/2022	8900	6258.89	26.4111
12/30/2022	9160	6208.2	29.518

For the HAG data set, the LSTM model was able to accurately predict the trading price of 22 sessions from December 1, 2022, to December 30, 2022, as shown in Table 4. According to the calculations, the percentage of error increased from 1.32% for the first trading session 16.55% for the 20th trading session, and we believe that this percentage will continue to increase due to the model's learning trend decreasing, which is shown in Figure 13. However, overall, the model's prediction results in the first three days were the closest to the actual price on December 1, 2022, with an error of only 1.3246%, and the highest at 7.35% on December 5, 2022. Overall, the model's average error rate was 1.3246%, 8.305%, and 10.3% for the first, fifth, and seventh days of trading prediction, respectively.

We have noticed that for models with an increasing trend, the percentage of error for the first 15 predicted sessions is usually more accurate or close to the actual value compared to models with a decreasing trend, which may be because the model learned the trend from past data of the stock.

VI. Conclusion

The results of this study demonstrate that out of the five models tested (Linear, Non-linear, ARIMA, LSTM, and

Prophet), the most suitable for predicting the future price of HPG and HAG stocks in the resulting time series was LSTM model. The other models, including the Linear, Non-lienar, ARIMA, Prophet, did not perform as well. This study highlights the importance of considering a variety of modeling approaches in financial analysis, and the potiential value of using LSTM model for predicting stock pirces in the future. Futher research could be conducted to verify the results of this study and to investigate the performance of the other models on different types stock price prediction tasks.

VII. REFERENCES

[1] Masoud, Najeb MH. (2017) "The impact of stock market performance upon economic growth." International Journal of Economics and Financial Issues 3 (4): 788–798.

[2] Murkute, Amod, and Tanuja Sarode. (2015) "Forecasting market price of stock using artificial neural network." International Journal of Computer Applications 124 (12): 11-15

[3] Michael Van Gysen, Chun-Sung Huang, Ryan Kruger, The Performance Of Linear Versus Non-Linear Models In Forecasting Returns On The Johannesburg Stock Exchange, International Business & Economics Research Journal (IBER), 2013

[4] Prapanna Mondal, Labani Shit1 and Saptarsi Goswami "STUDY OF EFFECTIVENESS OF TIME SERIES MODELING (ARIMA) IN FORECASTING STOCK PRICES"

[5] Li, Lei, Yabin Wu, Yihang Ou, Qi Li, Yanquan Zhou, and Daoxin Chen. (2017) "Research on machine learning algorithms and feature extraction for time series." IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC): 1-5.

[6] "Predicting Stock Prices Using LSTM" Murtaza Roondiwala, Harshal Patel, Shraddha Varma

[7] "Stock Price Prediction Using Time Series Data" Mashtura Mazed, BRAC University in Computer Science on August 28, 2019.

[8] Freedman, David (2005) Statistical Models: Theory and Practice, Cambridge University Press.

[9] C. Chatfield, "Time-series forecasting," New York, USA: Chapman & Hall/CRC, 2005.

[10] R. Nau, "Introduction to ARIMA models," Fuqua School of Business, Duke University, 2014. [Online]. Available:https://people.duke.edu/ rnau/Slides on ARIMA models–Robert Nau.pdf, Accessed on: Mar. 21, 2018.

[11] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[12]https://www.tutorialspoint.com/time_series/time_series_lstm_model.htm

[13] https://facebook.github.io/prophet/