Forecasting Vietnam Bank Stock Price by applying Machine Learning and Statistical models.

NGO MANH KHUONG (21522241)[1], NGO CONG HUAN (21520881)[2], AND NGUYEN THI VAN ANH (21521835).[3]

[1]University of Information Technology, Ho Chi Minh city (email: 21522241@gm.uit.edu.vn)
[2]University of Information Technology, Ho Chi Minh city (email: 21520881@gm.uit.edu.vn)
[3]University of Information Technology, Ho Chi Minh city (email: 21521835@gm.uit.edu.vn)

ABSTRACT In the context of the stock market, a share signifies a fraction of ownership in a company that is publicly traded. When a company opts to go public to raise funds, it splits its ownership into small, equal parts known as shares, which are then sold to investors on the stock market. The creation of dependable prediction models for the stock market enables investors to make more informed decisions. This study seeks to forecast the future stock prices of some major banks (ACB, VCB, BIDV) in Vietnam. Moreover, Machine Learning, which entails the development of computer tasks that mimic human intelligence, is currently the most widely used technique. It is presently a potent analytical tool for effectively managing investments in financial markets. The extensive application of machine learning in the financial sector has led to the emergence of a revolutionary method that can assist investors in making superior investment and management decisions to boost the performance of their securities assets. In this study, we initially examine the share prices of the aforementioned companies on the Vietnamese stock exchanges over the past five years. Subsequently, we utilize data and integrate some machine learning algorithms to predict future stock prices.

INDEX TERMS Bank Stock Price Forecast, Linear Regression, ARIMA, SVR, Long Short-Term Memory (LSTM), Vector Autoregression (VAR), Random Forest, Seq2Seq, Fully Convolutional Neural Network (FCN)

I. INTRODUCTION

The accurate prediction of stock prices remains a critical challenge in financial markets, warranting continuous exploration and refinement of predictive models. In this paper, we undertake an in-depth analysis utilizing an array of forecasting models—ARIMA (AutoRegressive Integrated Moving Average), SVR (Support Vector Regression), LSTM (Long Short-Term Memory), Linear Regression, VAR (Vector Autoregression), Random Forest, Seq2Seq (Sequence-to-Sequence), FCN (Fully Convolutional Networks), alongside additional models—to forecast the stock prices of three prominent banks: ACB, VCB, and BIDV.

Stock markets exhibit a multifaceted nature, influenced by a multitude of internal and external factors, rendering the prediction of stock prices a complex endeavor. The inclusion of various models enables a holistic approach, accommodating diverse perspectives and capturing intricate patterns within the stock market data. Each model contributes distinct capabilities and intricacies, facilitating a comprehensive understanding of the underlying trends and dynamics governing the stock prices of these major banking institutions.

Throughout this study, we aim to not only assess the predictive accuracy but also to compare the robustness and reliability of these diverse forecasting models. Metrics such as Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Squared Logarithmic Error (MSLE) will be employed for a comprehensive evaluation of model performance. By examining these metrics, we seek to offer insights into the strengths and limitations of each model in the context of stock price prediction for ACB, VCB, and BIDV.

The significance of comprehending the nuances of stock price forecasting, coupled with the evaluation of multiple predictive models using diverse performance metrics, cannot be overstated. This paper strives to contribute to the ongoing discourse in financial market analysis by presenting a rigorous evaluation of various models, thus facilitating a deeper understanding of the complexities inherent in stock market dynamics.

Understanding the comparative performance of these models in terms of RMSE, MAPE, and MSLE metrics is instrumental for investors in making informed decisions and for researchers and practitioners in enhancing the precision and reliability of financial market predictions. By critically examining these models' performances, this study aims to provide valuable insights into optimizing investment strategies and fostering well-informed financial decision-making.

This paper serves as a comprehensive exploration of diverse predictive models applied to the stock prices of significant banking entities, paving the way for further advancements in the field of financial market analysis and forecasting.

II. RELATED WORK

Vaishnavi Gururaj and his team from Global Academy of Technology, Bengaluru, India apply Linear Regression (LR) to easily available sample data, followed by observations. Subsequently, Support Vector Regression (SVR) is employed, and observations and results are graphically plotted. Support Vector Machines (SVMs), with advanced features like high accuracy and predictability, are introduced for comparison. The survey of pros and cons for both techniques leads to the conclusive finding that SVM outperforms LR in predicting values. [5]

Prapanna Mondal , Labani Shit and Saptarsi Goswami conducted a study on fifty six stocks from seven sectors Accuracy of ARIMA model in predicting stock prices is above 85%, which indicates that ARIMA gives good accuracy of prediction. [10]

Bruno Miranda Henrique from The Journal of Finance and Data Science SVR to evaluate its performance on a range of Brazilian, American, and Chinese stocks with varying attributes, such as small-cap or blue-chip stocks. This research demonstrates that when utilizing a linear kernel with a fixed training set of daily prices, it is possible to get reduced prediction errors in the test set compared to the training set. Furthermore, with daily prices and fixed training models, this kernel performed better for price predictions than the radial and polynomial kernels. Nevertheless, employing a fixed training time decreased the model's predictive effectiveness when the price frequency was increased to minutes. Specifically, for nearly all stocks examined in real-time prices, SVR yielded less accurate predictions than a random walk model when employing fixed training. [6]

Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia developed LSTM program using Python and Tensorflow for stock prices prediction (accuracy 94.59% at 100 epoch). Comparing with other research for stock price forecasting, LSTM method is better (usually neural network method only about 90% accuracy). [3]

Phillip Hushani uses VAR model to make predictions using the NASDAQ closing price. Based on these experiments, a medium-term trading strategy has been developed. The analysis shows that the LSTM is the most accurate model. The VAR model can predict the trend more accurately than the other models. [7]

The results of this study revealed that the Random Forest (RF) algorithm has the best accuracy because it received a 94.12% accuracy ratio in this work. They worked on the choice of advantage, the advantages of classifying the basic advantage,and how to use new techniques to predict the best field in the stock market. [4]

Soonsung Hwang and his colleagues performed an experiment on Temperature Prediction in Firing Furnace Process. The proposed model shows very high accuracy in predicting the future temperature of the kiln, outperforming other baseline models. This shows that the Seq2Seq model has the ability to predict with high accuracy rate with time series data. [8]

Jizhong Wu researches a fully convolutional network (FCN)-based fault detection method to segment seismic images and identify faults. The results of the study show that the FCN model shows more accurate and effective error detection than traditional methods, with fast prediction time and high accuracy. [16]

III. MATERIALS AND METHODS

A. DATA COLLECTION

We use Vietnam bank stock price (ACB, BIDV, VCB) from investing.com website. Each dataset range from 22/12/2013 to 22/12/2023, and contains following columns:

- Date: stock trading opening day.
- Price (also known as Close Price): the last price at which a stock trades upon the end of the exchange.
- Open: the first price at which a stock opens for trading.
- High: highest stock price of the day.
- Low: lowest stock price of the day.
- Volume: the number of shares that the trader buys and sells.
- Change: today's change in stock price from the previous day is expressed as a percentage

B. DESCRIPTIVE STATISTIC

	ACB	VCB	BIDV
Count	2493	2496	2472
Median	11285.2	48528	26021.7
Mean	13078.93	50337.02	25255.92
Std	7898.92	26073.23	11804.1
Variance	62393053.17	679813431	139336897.2
Deviation Coefficient	0.603	0.517	0.467
Skewness	0.515	0.127	0.144
Kurtosis	-1.158	-1.366	-1.257
Min	3885.1	10477	8006.4
25%	5279.7	26196	12801.725
50%	11285.2	48528	26021.7
75%	21700	75174	34273.525
Max	30360	106500	49100

TABLE 1. Descriptive Statistic

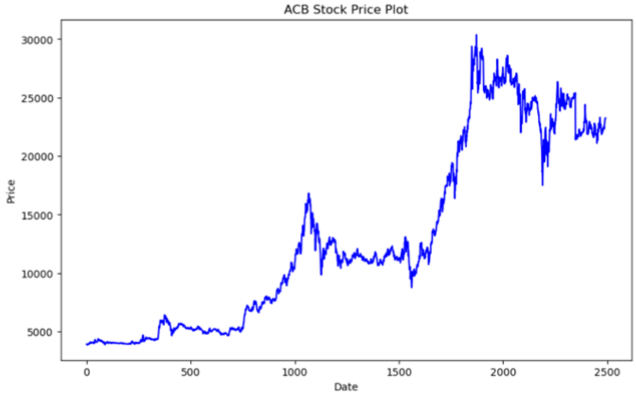

FIGURE 1. ACB stock price box plot

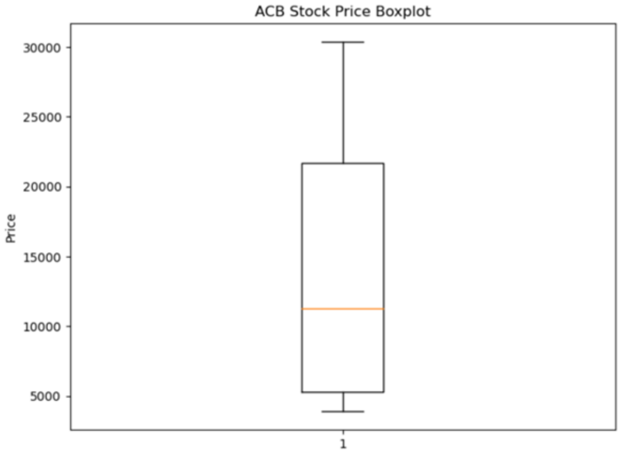

FIGURE 2. ACB stock price box plot

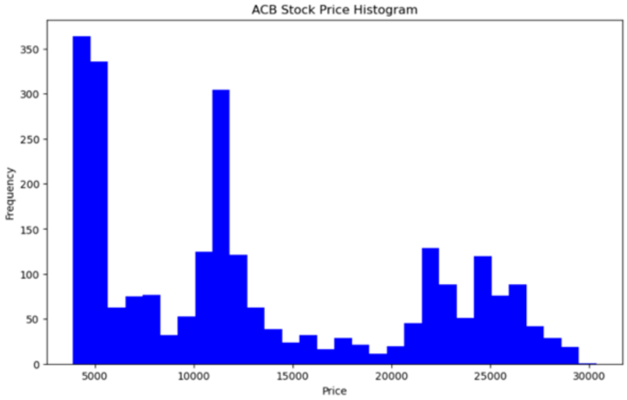

FIGURE 3. ACB stock price histogram plot

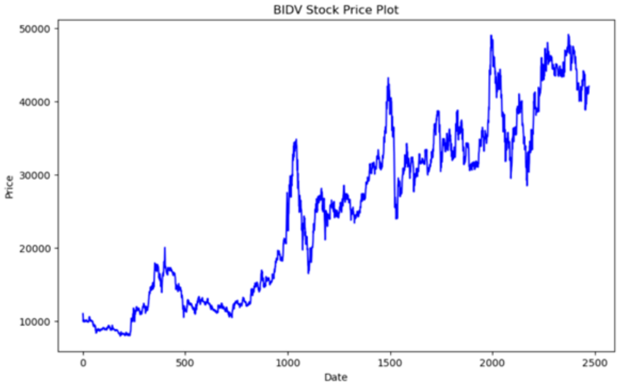

FIGURE 4. BIDV stock price plot

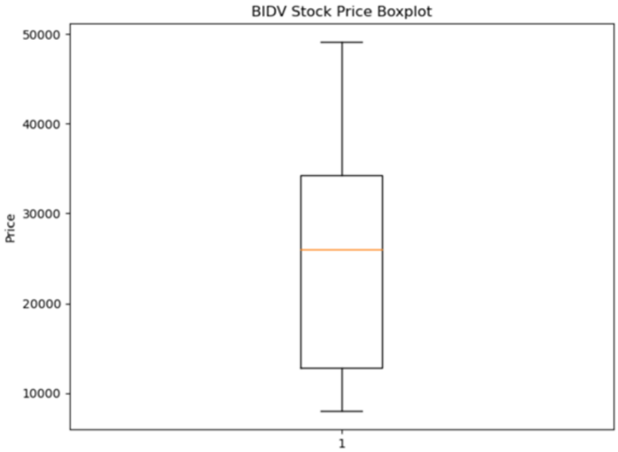

FIGURE 5. BIDV stock price box plot

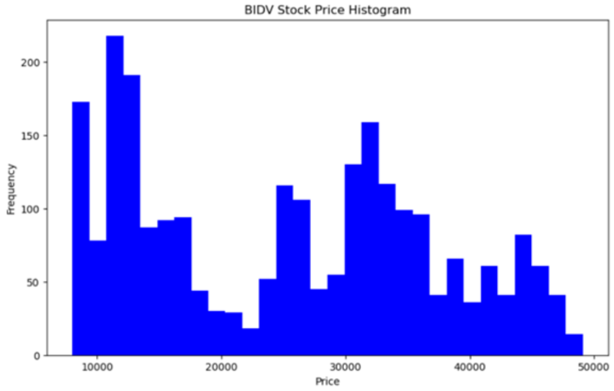

FIGURE 6. BIDV stock price histogram plot

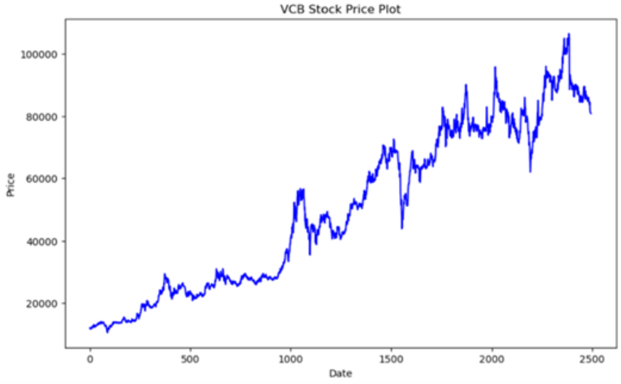

FIGURE 7. VCB stock price plot

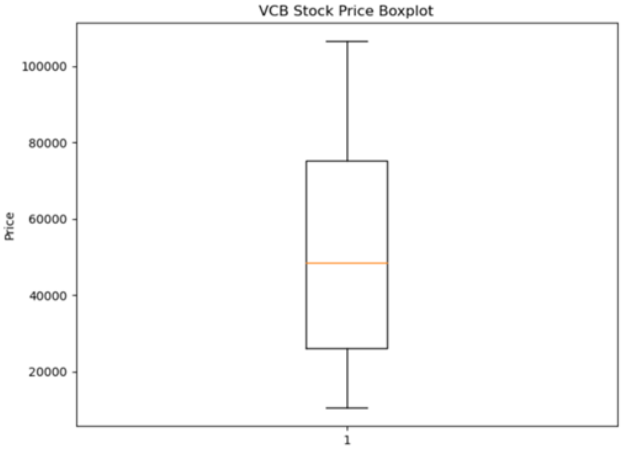

FIGURE 8. VCB stock price box plot

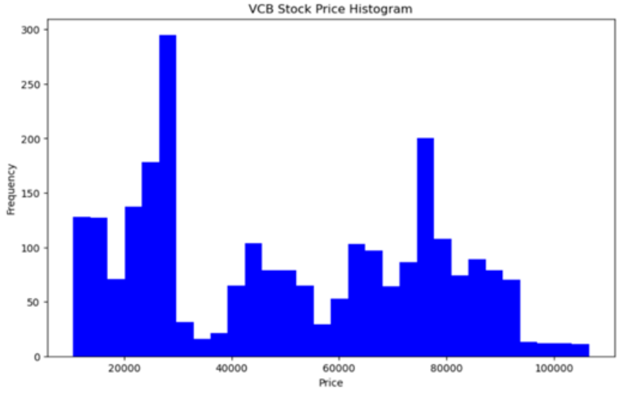

FIGURE 9. VCB stock price histogram plot

C. ALGORITHMS

The picture below shows how we perform algorithms in this report.

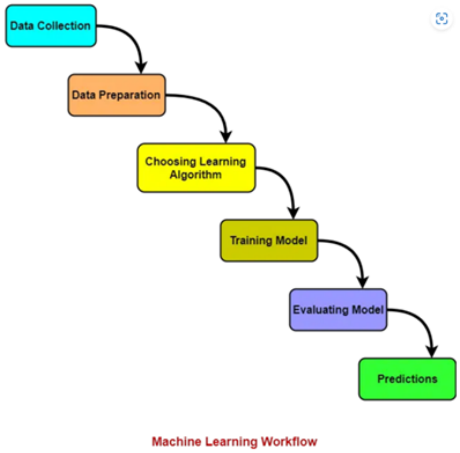

FIGURE 10. Workflow of performing Machine learning

1) Linear Regression

Linear Regression Model is a statistical model that describes the linear relationship between a response variable Y and multiple predictor variables $X_1, ..., X_p$., where Y is linearly dependent on X. [11]
Important terms:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + bpXp + e$$

where:

- **Dependent variable Y**: The variable that is being predicted. In other words, Y is the variable whose value we are trying to determine based on the value of X.
- **Independent variable** $X_1, X_2, ...X_p$: The variable that is used to predict Y. In other words, X is the variable whose value we know, and we are using that value to estimate the value of Y.
- **Intercept** b_0: The value of Y when X is 0. In other words, b0 is the value of Y when there is no relationship between X and Y.
- **Slope** $b1, b2, .., bp$: The rate of change of Y with respect to X. In other words, b1 indicates how much Y changes when X changes by one unit.

2) ARIMA

ARIMA is commonly used for forecasting time series data that exhibit trends, seasonality, and other temporal patterns. They are versatile and can be applied to various domains such as finance, economics, weather forecasting, and more. ARIMA consists of three components:

- **AutoRegressive**: AR(p) is a regression model with lagged values of y, until p-th time in the past, as predictors. Here, p = the number of lagged observations in the model, ϵ is white noise at time t, c is a constant and ϕs are parameters.

$$y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

- **Integrated I(d)**: The difference is taken d times until the original series becomes stationary. A stationary time series is one whose properties do not depend on the time at which the series is observed.

- **Moving average MA(q)**: A moving average model uses a regression like model on past forecast errors. Here, ϵ is white noise at time t, c is a constant, and ϕs are parameters

$$y_t = c + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Combining all of the three types of models above gives the resulting ARIMA(p,d,q) model: [2]

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where:

$$\begin{aligned} Y_t &\text{ is the value of the time series at time } t, \\ c &\text{ is a constant term,} \\ \phi_1, .., \phi_p &\text{ are the autoregressive parameters,} \\ \varepsilon_t &\text{ is the white noise term at time } t, \\ \theta_1, .., \theta_q &\text{ are the moving average parameters,} \\ p &\text{ is the order of the autoregressive part, and} \\ q &\text{ is the order of the moving average part.} \end{aligned}$$

3) Support Vector Regression (SVR)

Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) technique. Researchers have found that SVM provides excellent performance in time series forecasting. SVM finds an optimal hyperplane to separate two classes of samples. SVR finds an optimal hyperplane in high-dimensional space to separate data samples and minimize prediction error.

The SVR optimization function formula, with w being the unknown parameter vector: [12]

$$\begin{aligned} \min_{w,b,\zeta,\zeta^*} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \\ \text{subject to} \quad & y_i - (w \cdot \phi(x_i) + b) \leq \varepsilon + \zeta_i, \quad i = 1, \ldots, n \\ & (w \cdot \phi(x_i) + b) - y_i \leq \varepsilon + \zeta_i^*, \quad i = 1, \ldots, n \\ & \zeta_i, \zeta_i^* \geq 0, \quad i = 1, \ldots, n \end{aligned} \tag{1}$$

Two common kernels used in SVR include: Radial Basis Function (RBF) Kernel:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Polynomial Kernel:

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d$$

4) Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is an improvement from the RNNs, which is able to solve the Gradient Problem. The LSTM models essentially extend the RNN's memory to enable them to keep and learn long-term dependencies of inputs. [15]. A typical LSTM block is configured mainly by memory cell state, forget gate, input gate, and output gate. The crucial element, memory cell state, runs down through the entire chain to selectively add or remove information to the cell state with the help of the three gates. [13]

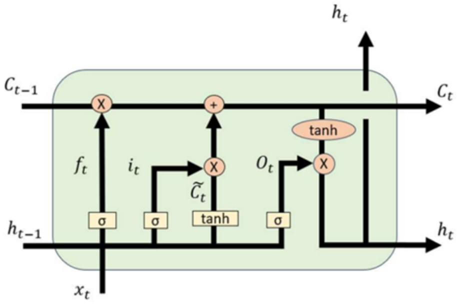

FIGURE 11. A single LSTM cell

Formulas for each LSTM cell:

$$\begin{aligned} i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\ f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\ o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\ g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned}$$

Where:

- x_t is the input at time t.
- h_{t-1} is the hidden state from the previous time step.
- i_t, f_t, o_t, and g_t are the input, forget, output, and cell input vectors, respectively.
- W and b are weight and bias matrices.
- σ is the sigmoid activation function.
- $\odot$ denotes element-wise multiplication.

5) Vector Autoregression (VAR)

Christopher Sims (1980) provided a new macroeconometric framework that held great promise: vector autoregressions (VARs). A univariate autoregression is a single-equation, single-variable linear model in which the current value of a variable is explained by its own lagged values. A VAR is an n-equation, n-variable linear model in which each variable is in turn explained by its own lagged values, plus current and past values of the remaining n-1 variables.

VAR is a forecasting algorithm that can be used when two or more time series influence each other, i.e. the relationship between the time series involved is bidirectional. [14]

The VAR(p) model of order p can be represented in the following formula: [1]

$$Y_1 = C + A_{1,1}Y_1^{(1)} + A_{1,2}Y_1^{(2)} + \ldots + A_{1,n}Y_1^{(n)} + E_1 \tag{2}$$

Where:

- Y_1 is the main variable.
- C is a constant.
- $A_{1,1}, A_{1,2}, \ldots, A_{1,n}$ are coefficients.

- $Y_1^{(1)}, Y_1^{(2)}, \ldots, Y_1^{(n)}$ are related variables.
- E_1 is an error term.

Steps that we need to follow to build the VAR model are:

1. Examine the Data
2. Test for stationarity
2.1 If the data is non-stationary, take the difference.
2.2 Repeat this process until you get the stationary data.
3. Train/Test Split
4. Grid search for order P
5. Apply the VAR model with order P
6. Forecast on new data.
7. If necessary, invert the earlier transformation.

6) Random Forest

Random forests are a machine learning model that combines multiple decision trees to produce more accurate and robust results than a single tree. They work by using the idea of "wisdom of the crowds", in which many simple models can make more accurate predictions than a more complex model. [11]

Random forests work in four steps:

1. Randomly select samples from the given dataset.
2. Build a decision tree for each sample and get the prediction results from each decision tree.
3. Take a vote for each prediction result.
4. Choose the most predicted result as the final prediction.

Decision trees are simple machine learning models that can be used for classification or forecasting tasks. They work by splitting the data into branches based on certain features. Random forests use multiple decision trees to produce more accurate results.

1) For $k = 1$ to K:
   - Sample a bootstrap sample from the training data.
   - Train a decision tree T_k on the bootstrap sample.
2) For a new input X:
   - For each tree T_k, obtain the prediction Y_k.
   - Aggregate predictions: $\hat{Y} = \frac{1}{K}\sum_{k=1}^{K} Y_k$ (for regression) or $\hat{Y} = \text{mode}(Y_1, Y_2, \ldots, Y_K)$ (for classification).

7) Seq2Seq

The seq2seq model is a deep learning neural network architecture used in natural language processing and tasks involving sequential data. It consists of two main parts are encoder and decoder.

Encoder:

- **Input:** Input sequence $x = (x_1, x_2, \ldots, x_t)$
- **Context representation:** The hidden state of the LSTM is computed as $h_t = \text{LSTM}(x_t, h_{t-1})$

Decoder:

- **Output:** Output sequence $y = (y_1, y_2, \ldots, y_t')$
- **Initial state:** (s_0) is initialized from h_t
- **Update hidden state:** $s_t = \text{LSTM}(y_t, s_{t-1})$
- **Output prediction:** $P(y_t|y_{<t}, x) = \text{softmax}(W_{\text{out}} s_t + b_{\text{out}})$

Training:

- **Loss function:** $L(y, y') = -\frac{1}{T'}\sum_{t=1}^{T} \log P(y_t|y_{<t'}, x)$
- Parameters are updated using backpropagation and optimization algorithms (e.g., SGD, Adam)

8) Fully Convolutional Neural Network (FCN)

FCNs (Fully Convolutional Neural Networks) were introduced by Wang et al. (2017b) for classifying time series. They lack local pooling layers, maintaining the time series length during convolutions. FCNs use a Global Average Pooling (GAP) layer instead of a final FC layer, reducing parameters and enabling the identification of crucial segments contributing to classifications via Class Activation Mapping (CAM). The architecture includes three convolutional blocks with batch normalization and ReLU activation. FCNs exhibit parameter invariance in most layers across time series lengths, facilitating transfer learning. [9]

$$conv(i, j) = R(\sum_{u=0}^{M-1}\sum_{v=0}^{M-1} w_{u,v} x_{i+u,j+v} + b)$$

Where:

- $\text{conv}(i, j)$ represents the convolution operation at position (i, j).
- R is an activation function (not specified in the formula).
- M is the size of the convolutional filter.
- $w_{u,v}$ are the weights of the convolutional filter.
- $x_{i+u,j+v}$ are the input values at positions $(i + u, j + v)$.
- b is the bias term.

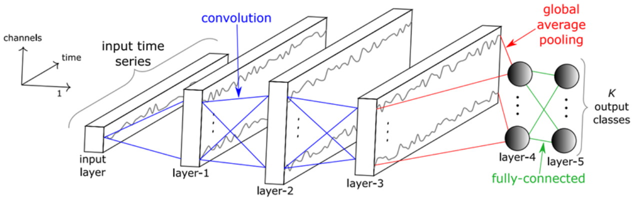

FIGURE 12. FCN Architecture

IV. EVALUATION

To evaluate the accuracy of the models, we use three parameters which are Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Squared Logarithmic Error (MSLE). The algorithm with the lowest value of those three parameters has the best performance. Below is the formula for RMSE, MAPE, and MSLE.

1) **Root Mean Squared Error (RMSE):**

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

2) **Mean Absolute Percentage Error (MAPE):**

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

3) Mean Squared Logarithmic Error (MSLE):

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(1+y_i) - \log(1+\hat{y}_i))^2$$

V. RESULT

In this section, we use the best performance (lowest error) model for each train/test ratio of each dataset to predict the next 30 days.

A. ACB

1) 7:3 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	7914.5637	29.493	0.1508
ARIMA	5241.16	16.975	0.0549
SVR	12453.958	47.253	0.5152
LSTM	665.57	2.186	0.0007
VAR	4531.319	14.287	0.0392
Random Forrest	6273.3679	23.4715	0.0829
Seq2Seq	736.9556	2.19	0.000961
FCN	1206.976	3.91	0.00234

TABLE 2. 7:3 ACB Ratio

The best-suited model: LSTM model for 7:3 ACB

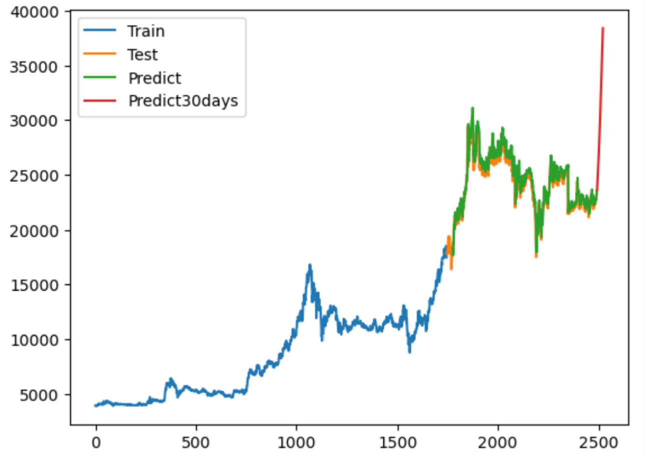

FIGURE 13. LSTM model

2) 8:2 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	3245.4505	10.863	0.01932
ARIMA	5944.317	22.883	0.0526
SVR	453.4623	1.265	0.00038
LSTM	500.7	1.464	0.0004
VAR	6337.585	24.557	0.0588
Random Forrest	858.3885	2.837	0.001329
Seq2Seq	726.2946	2.35	0.001004
FCN	1241.7439	4.45	0.002867

TABLE 3. 8:2 ACB Ratio

The best-suited model: SVR model for 8:2 ACB

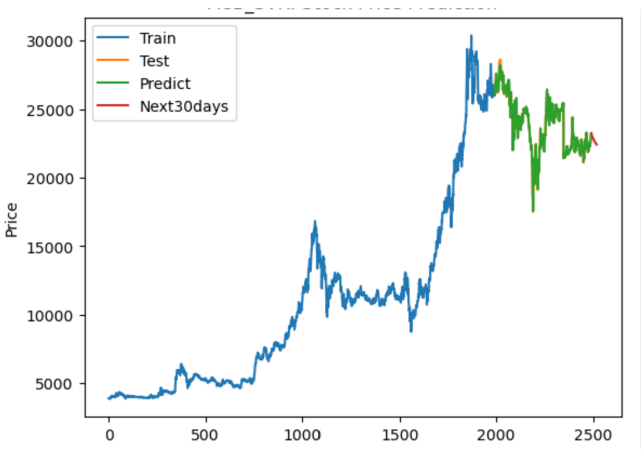

FIGURE 14. SVR model

3) 9:1 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	2609.8566	9.719	0.01197
ARIMA	1354.823	4.851	0.0033
SVR	402.7491	1.00598	0.000292
LSTM	522.41	1.723	0.0005
VAR	1910.577	7.798	0.0066
Random Forrest	454.8253	1.289836	0.000385
Seq2Seq	563.8249	1.71	0.00059
FCN	1022.7937	3.50	0.001943

TABLE 4. 9:1 ACB Ratio

The best-suited model: SVR model for 9:1 ACB

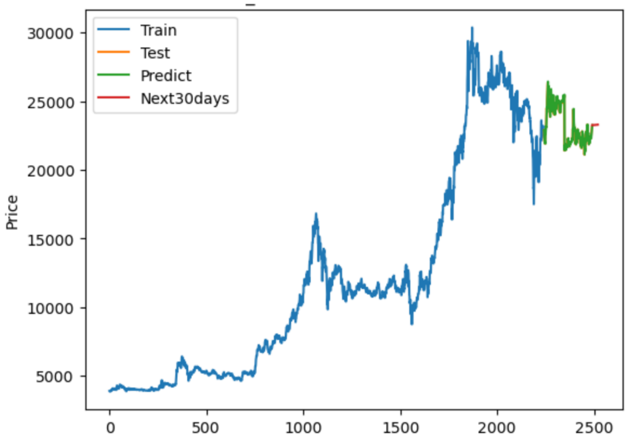

FIGURE 15. SVR model

B. BIDV

1) 7:3 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	3881.874	8.3104	0.01029
ARIMA	5273.644	12.038	0.0184
SVR	5500.5356	7.7549	0.01883
LSTM	1039.54	2.066	0.0007
VAR	6386.942	15.203	0.027
Random Forrest	2586.5432	4.3943	0.00368
Seq2Seq	857.092	1.62	0.000506
FCN	1812.1182	3.80	0.00224

TABLE 5. 7:3 BIDV Ratio

The best-suited model: Seq2Seq model for 7:3 BIDV

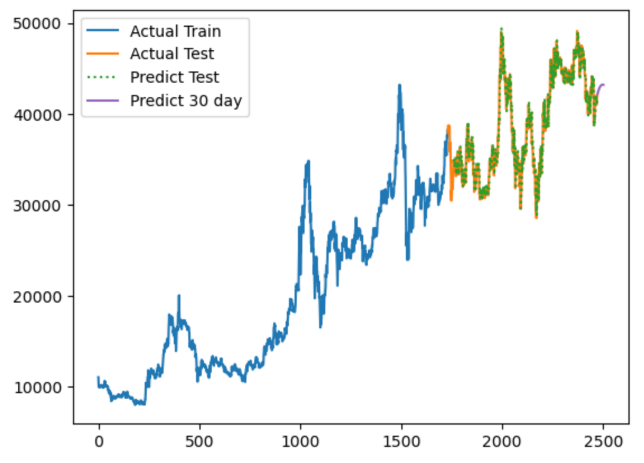

FIGURE 16. Seq2Seq model

2) 8:2 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	4379.2149	9.0322	0.01221
ARIMA	7547.189	15.115	0.0361
SVR	6203.97	10.201	0.02334
LSTM	927.69	1.746	0.0005
VAR	4998.594	9.926	0.0153
Random Forrest	1127.7548	1.91	0.000632
Seq2Seq	905.8123	1.66	0.000554
FCN	1293.7144	2.56	0.001106

TABLE 6. 8:2 BIDV Ratio

The best-suited model: Seq2Seq model for 8:2 BIDV

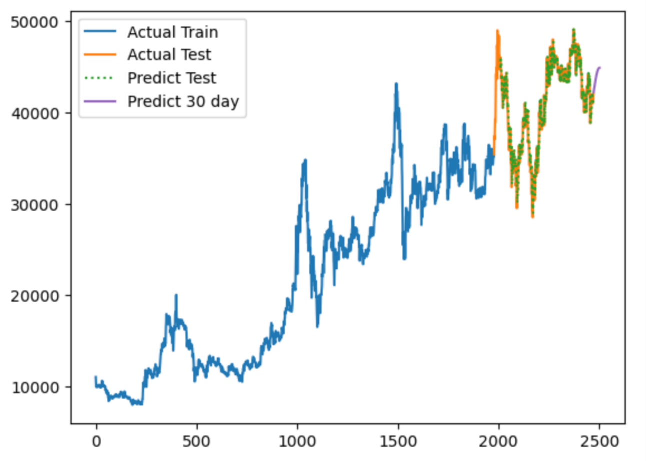

FIGURE 17. Seq2Seq model

3) 9:1 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	3572.8447	7.1005	0.00674
ARIMA	5854.336	12.034	0.0193
SVR	1493.611	2.44154	0.00109
LSTM	1429.35	2.841	0.001
VAR	3705.591	7.2904	0.0072
Random Forrest	1199.4247	2.0366	0.000711
Seq2Seq	739.4482	1.17	0.000284
FCN	964.1519	1.62	0.000489

TABLE 7. 9:1 BIDV Ratio

The best-suited model: Seq2Seq model for 9:1 BIDV

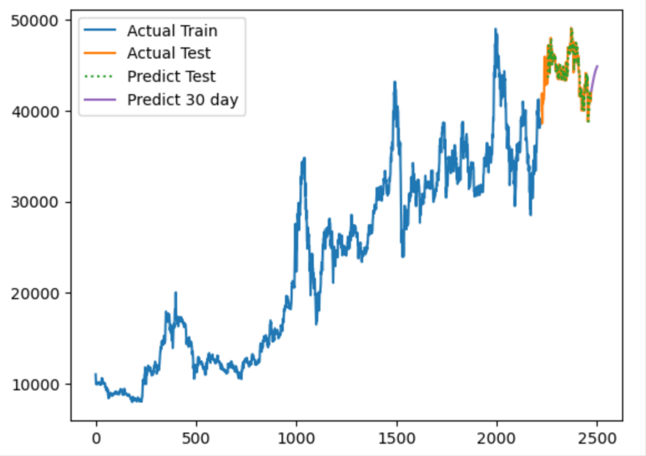

FIGURE 18. Seq2Seq model

C. VCB

1) 7:3 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	7481.893	7.4121	0.00864
ARIMA	9723.279	9.89	0.013
SVR	24611.323	21.92	0.1225
LSTM	2283.62	2.327	0.0007
VAR	9684.298	9.829	0.0129
Random Forrest	10069.8906	8.2158	0.01429
Seq2Seq	1841.9644	1.64	0.000496
FCN	2817.4579	2.0	0.001133

TABLE 8. 7:3 VCB Ratio

The best-suited model: Seq2Seq model for 7:3 VCB

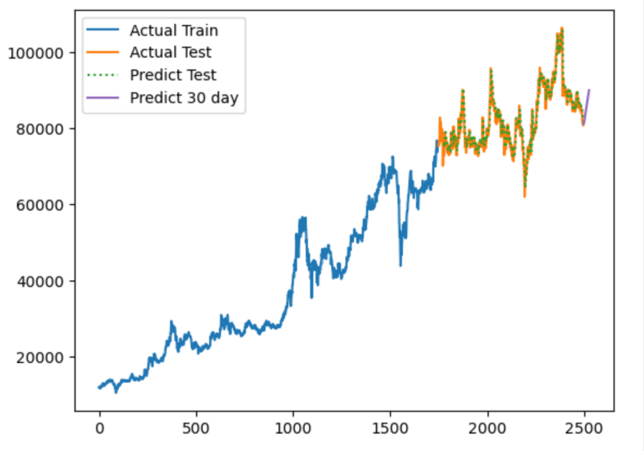

FIGURE 19. Seq2Seq modelmodel

2) 8:2 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	7562.7874	7.5426	0.00818
ARIMA	7467.38	6.853	0.0076
SVR	13616.361	9.815	0.02771
LSTM	1642.38	1.3701	0.0003
VAR	7200.996	6.963	0.0073
Random Forrest	4751.289	3.149	0.00265
Seq2Seq	2068.8175	1.77	0.000606
FCN	2728.555	2.47	0.001055

TABLE 9. 8:2 VCB Ratio

The best-suited model: LSTM model for 8:2 VCB

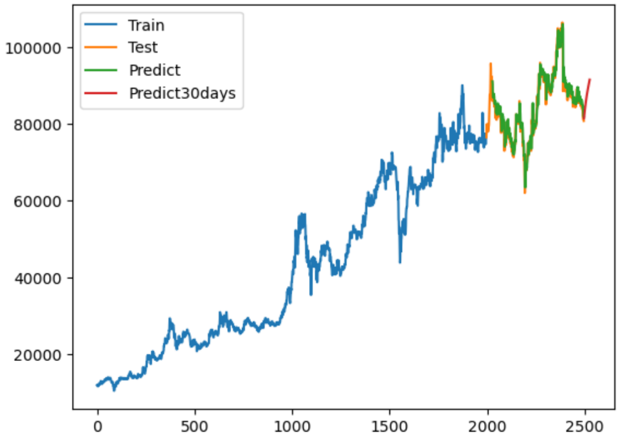

FIGURE 20. LSTM modelmodel

3) 9:1 Train/Test ratio

Model	**RMSE**	**MAPE**	**MSLE**
Linear Regression	7562.7874	7.5426	0.00818
ARIMA	9968.438	8.484	0.0124
SVR	11241.226	8.32024	0.016302
LSTM	1615.53	1.099	0.0002
VAR	10112.09	8.643	0.0128
Random Forrest	5611.5296	3.98	0.003508
Seq2Seq	2061.942	1.45	0.000481
FCN	2689.939	2.13	0.000819

TABLE 10. 9:1 VCB Ratio

The best-suited model: LSTM model for 9:1 VCB

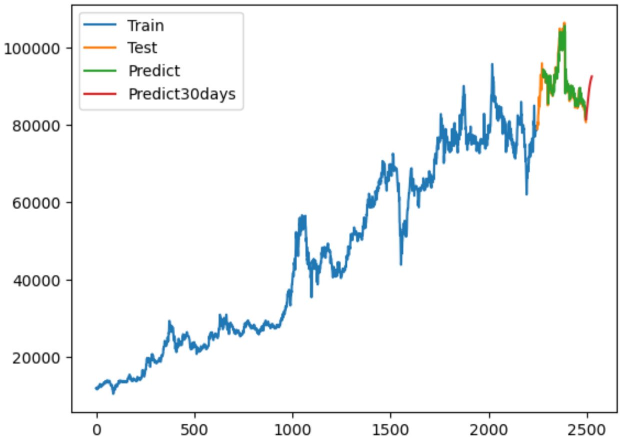

FIGURE 21. LSTM model

VI. CONCLUSION

The utilization of LSTM and seq2seq models to predict the stock prices of major banks in Vietnam such as ACB, VCB, and BIDV has demonstrated substantial potential in learning and generating forecasts based on the learned models alongside the other models used in this project (Linear Regression, ARIMA, SVR, LSTM, VAR, Random Forest, Seq2Seq, FCN). Both of these models exhibit the capability to handle time series data and learn intricate relationships within the sequences.

Assessing the stock price predictions for the upcoming 30 days using these models, both LSTM and seq2seq indicate a slight upward trend in the future stock prices of these banks. However, evaluating the "suitability" of a model for predicting an upward trend relies not only on its predictive capability but also on a profound understanding of the financial market and the fine-tuning and optimization of the model.

Hence, it's crucial to note that the accuracy of predictions does not solely stem from the model but also hinges on the harmonious integration of the model with domain expertise and the process of adjusting the model for effective real-world applications.

VII. ORIENTATION

Our current models may not be entirely accurate, and it is essential to allocate time for validating real-world predictions made by these models. Errors within the models might arise from research limitations or inappropriate model selection concerning the datasets used. Hence, there is room for further improvement in our models for the future.

For instance, enhancing existing models could involve integrating attention mechanisms into Seq2Seq models. This addition enables a focus on crucial parts of the input sequence, thereby enhancing the transformation from input to output data. Additionally, combining LSTM with other architectures like Convolutional Neural Networks (CNNs) or Attention Mechanisms could yield more robust models.

Furthermore, we plan to explore other models such as Transformers and various Gradient Boosting methods to compare their performance with our current models in practical scenarios. This comparative analysis aims to identify the most suitable model for our purposes.

In conclusion, our strategy involves both enhancing existing models and exploring new ones to improve accuracy and efficacy in real-world applications.

ACKNOWLEDGMENT

We extend our heartfelt appreciation to Assoc. Prof. Dr. Nguyen Dinh Thuan and TA. Nguyen Minh Nhut for their invaluable expertise and wholehearted guidance throughout this project. Without your passionate supervision, our group's report would have been incredibly challenging to complete.

This project has provided an excellent opportunity for each team member to collaborate, enhance their cooperative skills, exchange knowledge, and, significantly, put theoretical learning into practical application.

Throughout the project's execution, the team effectively utilized the taught knowledge while embracing new concepts, aspiring to achieve perfection in our work. Nevertheless, due to constraints such as time, limited knowledge, and experience, shortcomings are inevitable. Hence, the group eagerly anticipates receiving your valuable suggestions. These insights will aid in augmenting our understanding and skills, enabling us to contribute better to future projects and real-world scenarios.

Lastly, our team wishes you continued good health to persist in your noble endeavor of imparting knowledge to the upcoming generations.

REFERENCES

[1] Lemya Taha Abdullah. Forecasting time series using vector autoregressive model. International Journal of Nonlinear Analysis and Applications, 13(1):499–511, 2022.

[2] Neha BORA. Understanding arima models for machine learning, 2021.

[3] Widodo Budiharto. Data science approach to stock prices forecasting in indonesia during covid-19 using long short-term memory (lstm). Journal of big data, 8:1–9, 2021.

[4] Hind Daori, MANAR ALHARTHI, ALANOUD ALANAZI, GHAIDA ALZAHRANI, MAJED ABOROKBAH, and Nojood Aljehane. Predicting stock prices using the random forest classifier. 2022.

[5] Vaishnavi Gururaj, VR Shriya, and K Ashwini. Stock market prediction using linear regression and support vector machines. Int J Appl Eng Res, 14(8):1931–1934, 2019.

[6] Bruno Miranda Henrique, Vinicius Amorim Sobreiro, and Herbert Kimura. Stock price prediction using support vector regression on daily and up to the minute prices. The Journal of finance and data science, 4(3):183–201, 2018.

[7] Phillip Hushani. Using autoregressive modelling and machine learning for stock market prediction and trading. In Third International Congress on Information and Communication Technology: ICICT 2018, London, pages 767–774. Springer, 2019.

[8] Soonsung Hwang, Gunwoo Jeon, Jongpil Jeong, and JunYoul Lee. A novel time series based seq2seq model for temperature prediction in firing furnace process. Procedia Computer Science, 155:19–26, 2019.

[9] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. Data mining and knowledge discovery, 33(4):917–963, 2019.

[10] Prapanna Mondal, Labani Shit, and Saptarsi Goswami. Study of effectiveness of time series modeling (arima) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications, 4(2):13, 2014.

[11] Aileen Nielsen. Practical time series analysis: Prediction with statistics and machine learning. O'Reilly Media, 2019.

[12] A Le My Phung and K James. Comparison of support vector regression and neural networks. Trabalho para obtenção parcial do grau de Mestre em Ciência). Universidade do Minnesota Duluth, Duluth, Estados Unidos da América, 2016.

[13] Xuanyi Song, Yuetian Liu, Liang Xue, Jun Wang, Jingzhe Zhang, Junqiang Wang, Long Jiang, and Ziyan Cheng. Time-series well performance prediction based on long short-term memory (lstm) neural network model. Journal of Petroleum Science and Engineering, 186:106682, 2020.

[14] James H Stock and Mark W Watson. Vector autoregressions. Journal of Economic perspectives, 15(4):101–115, 2001.

[15] Savvas Varsamopoulos, Koen Bertels, and Carmen Garcia Almudever. Comparing neural network based decoders for the surface code. IEEE Transactions on Computers, 69(2):300–311, 2019.

[16] Jizhong Wu, Bo Liu, Hao Zhang, Shumei He, and Qianqian Yang. Fault detection based on fully convolutional networks (fcn). Journal of Marine Science and Engineering, 9(3):259, 2021.