Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.
Digital Object Identifier XX.XXXX/YYY.20XX.ZZZZZZZ

Forecasting the Vietnamese pharmacy companies' stock prices using Statistical, Machine Learning and Deep Learning Models.

Huu Quang Dang[1], Tan Phat Ly[2], and Dinh Vien Son Nguyen.[3].
[1]Vietname National University of Ho Chi Minh, University of Information Technology, Ho Chi Minh City (g-mail: 21522508@gm.uit.edu.vn)
[2]Vietname National University of Ho Chi Minh, University of Information Technology, Ho Chi Minh City (g-mail: 21522444@gm.uit.edu.vn)
[3]Vietname National University of Ho Chi Minh, University of Information Technology, Ho Chi Minh City (g-mail: 21522555@gm.uit.edu.vn)

ABSTRACT This paper investigates the application of various statistical models and machine learning algorithms in forecasting the stock prices of Vietnamese pharmaceutical companies. Leveraging a diverse array of methodologies, including Linear Regression (LR), Support Vector Regression (SVR), Long Short-Term Memory (LSTM) networks, Autoregressive Integrated Moving Average (ARIMA), Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX), K-Nearest Neighbors (KNN), and Boosting LSTM, our study aims to provide comprehensive insights into the predictive capabilities of these models within the dynamic and complex pharmaceutical stock market landscape. Additionally, the findings contribute to the ongoing discourse on applying statistical models and machine learning algorithms in predicting stock prices within emerging markets, particularly in the context of 3 Vietnamese Pharmaceutical Joint Stock Company (OPC, Vimedimex, and IMEXPHARM).

INDEX TERMS Stock price, forecasting, LR, SVR, LSTM, ARIMA, ARIMAX, KNN, Boosting LSTM, DLM, pharma companies, OPC, IMP, VMD.

I. INTRODUCTION

Maybe you still haven't forgotten how the emergence of the COVID-19 pandemic significantly altered the landscape of global health crises. Amidst the upheavals wrought by the pandemic, the pharmaceutical sector emerged as a focal point for investment consideration. The inherent resilience of pharmaceutical companies, coupled with increased attention to healthcare solutions, has sparked interest among investors. The potential for investing in pharmaceutical stocks lies in the innovation pipeline, regulatory developments, and the capacity of these companies to address evolving healthcare needs.

In light of the potential for investment in pharmaceutical stocks, this paper endeavors to leverage an array of sophisticated forecasting models, including Linear Regression (LR), Support Vector Regression (SVR), Long Short-Term Memory (LSTM) networks, Auto-regressive Integrated Moving Average (ARIMA), Auto-regressive Integrated Moving Average with Exogenous Variables (ARIMAX), K-Nearest Neighbors (KNN), and Boosting LSTM. These models will be employed to evaluate and forecast stock prices of 3 Vietnamese pharmaceutical companies (OPC, VMD, IMP).

II. RELATED WORK

Demirel, Çam, and Ünlü (2020) [1] focused on predicting opening and closing stock prices for 42 companies listed on the Istanbul Stock Exchange National 100 Index (ISE-100). They employed machine learning techniques and deep learning algorithms, specifically Multilayer Perceptions (MLP), Support Vector Machines (SVM), and Long Short-Term Memory (LSTM). The study spanned daily data from 2010 to 2019, totaling nine years. The dataset included 2249 records for each company's opening and closing stock prices, amounting to a total of 188,196 data points for analysis.

Ruochen Xiao, Yingying Feng, Lei Yan, and Yihan Ma [2] evaluated LSTM and ARIMA models' performance in predicting stock prices of 50 companies using MAE, MSE, and RMSE indicators. The dataset covered the highest transaction day prices from 2010,to 2018. The LSTM model was trained from 2010 to 2015, validated on 2016 to 2017, and tested on whole 2018. Conversely, the ARIMA model was trained on 2016 to 2017, and tested on whole 2018. Both models utilized 60 days of data to predict the subsequent day. LSTM demonstrating superior performance, particularly in capturing price fluctuations. However, ARIMA remains more

user-friendly for practical application.

Pawaskar, Shreya. [3] using machine learning finds applications in financial product recommendations and customer sentiment analysis. Predicting stock prices requires historical stock data. This research analyzes a data-set with seven attributes, using various regressors to forecast future prices. The study reveals that the Decision Tree Regressor model achieves the highest accuracy, exhibiting an R^2 error of 1.0 and an RMSE of 0.0. In summary, the Decision Tree Regressor outperforms other models, emerging as the most effective for stock price prediction.

Jolly Masih [4] conducted a study that have implemented algorithms such as SVM and LSTM on stock market data to see if major IT companies see a rise or fall during the COVID-19 pandemic. Authors have also used ARIMA forecasting method to predict the stocks of mentioned 4 companies Google, Microsoft, Apple and Amazon. Considering prediction of stock prices as of now, authors have got an accuracy of 75.95% using the SVM algorithm and ARIMA best fit forecast model. This drop in accuracy was intended, as described before due to increased volatility in the stock market due to COVID-19 outbreak.

Yara Kayyali Elalem et al. [5] introduced a sales prediction framework for newly launched products. They compared a traditional statistical method (ARIMAX) and three deep neural networks (LSTM, GRU, and CNN). Their study used Dell's data with weekly customer orders for 170 stock-keeping units (SKUs) from 2013 to 2016 and Retailer X's data comprising 843 complete product life cycles (PLCs). The analysis showed that among the machine learning methods, CNNs exhibited the least accuracy, with forecast errors (MASEs) notably higher by 15%, 578%, and 4291% compared to ARIMAX. Conversely, the three DNNs demonstrated higher robustness to data noise and were more suitable for accurately predicting sales of newly launched products.

Shiyue Kuang (2023) [6] used Linear Regression, LSTM and ARIMA to forecast HSBC's stock price. The results of the LSTM model are the best models to forecast stock prices. The LSTM model has the smallest RMSE, and the RMSE of ARIMA is nearly 40 times that of LSTM. Meanwhile, the Linear Regression model is ranked last, with an estimated error nearly 25% larger than ARIMA.

III. METHODOLOGY

Our team has decided to approach this problems by introducing to you 8 algorithms below includes Statistical Models, Machine Learning and Deep Learning Models.

A. LINEAR REGRESSION (LR)

Regression analysis is a tool for building mathematical and statistical models that characterize relationships between a dependent variable (which must be a ratio variable and not categorical) and one or more independent, or explanatory, variables, all of which are numerical (but may be either ratio or categorical).[7]

A Linear regression model that includes 2 regression, and a single independent variable is called simple regression and a regression model that involves two or more independent variables is called multiple regression.Simple linear regression is just a special case of multiple linear regression[8]. A multiple linear regression model has the form Eq.(20)

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + e. \quad (1)$$

where:

Y, is the dependent variable,

X1, . . . , Xk are the independent (explanatory) variables,

b0 is the intercept term,

b1, . . . , bk are coefficients for the independent variables,

e is the error term.

B. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

Models that describe such homogeneous non-stationary behavior can be obtained by assuming that some suitable difference of the process is stationary. The properties of the important class of models for which the dth difference of the series is a stationary mixed autoregressive moving average process. These models are called autoregressive integrated moving average (ARIMA) processes.[9].

- AutoRegressive - AR(p) is a regression model with lagged I, until p-th time in the past, as predictors. Eq.(20)

$$AR(p) = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \varepsilon. \quad (2)$$

- Integrated I(d) - The difference is taken d times until the original series becomes stationary. Eq.(20)

$$I(d) = \Delta y_t = y_t - y_{t-1}. \quad (3)$$

- Moving average MA(q) - A moving average model uses a regression-like model on past forecast errors.(20)

$$MA(q) = \alpha + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon. \quad (4)$$

- Regression equations for ARIMA(p,d,q): Eq.(20)

$$ARIMA(p,q,d) = AR(p) + I(d) + MA(q) \quad (5)$$

C. SUPPORT VECTOR REGRESSION (SVR)

Support Vector Regression (SVR) is a type of machine learning algorithm used for regression analysis. The goal of SVR is to find a function that approximates the relationship between the input variables and a continuous target variable, while minimizing the prediction error. Addition, SVR uses the same principle as Support Vector Machines (SVM), but for regression problems. [10]

SVR can handle non-linear relationships between the input variables and the target variable by using a kernel function to map the data to a higher-dimensional space. This makes it a powerful tool for regression tasks where there may be complex relationships between the input variables and the target variable. [10]

Polynomial is the kernel used in this study. In Support Vector Regression (SVR), can use various types of kernels to create different classification functions: Eq.(6),(7),(8),(9).

$$Linear : k\left(x, z\right) = x^T z \tag{6}$$

$$Polynomial : k(x, z) = (r + \gamma x T z)^d \tag{7}$$

$$RBF : k\left(x, z\right) = exp\left(\frac{-\gamma \left|\left|x - z\right|\right|^2}{2\sigma^2}\right), \ \gamma > 0 \tag{8}$$

$$Sigmoid : k(x, z) = tanh(\gamma x^T z + r) \tag{9}$$

D. LONG SHORT-TERM MEMORY (LSTM)

The clever idea of introducing self-loops to produce paths where the gradient can flow for long duration is a core contribution of the initial long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997). [11]

"Long Short-Term Memory" (LSTM), a novel recurrent network architecture in conjunction with an appropriate gradient-based learning algorithm. [12]

The below shows steps by steps walk through the LSTM fomular and architecture of the LSTM cell model [13]

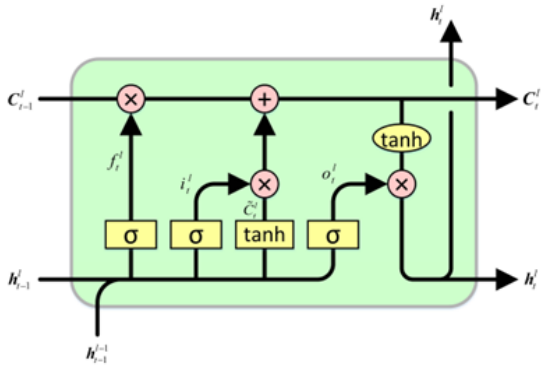


FIGURE 1: LSTM Cell Architecture

$$Forgetgate : f_t \ = \ \sigma(W_f \bullet \ [h_{t-1}, x_t] + b_f) \tag{10}$$

$$Inputgate : \ i_t \ = \ \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{11}$$

$$Outputgate : o_t \ = \ \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{12}$$

The input gate determines how much new information is added to the neuron state.

$$\hat{C}_t \ = \ tanh(W_C \bullet \ [h_{t-1}, x_t] + bC) \tag{13}$$

The output gate is used to control how many current neural unites state are filtered and how many controlling units state are filtered.

$$h_t \ = \ o_{ti} \bullet tanh(\hat{C}_t) \tag{14}$$

E. ARIMA WITH EXOGENOUS VARIABLES (ARIMAX)

ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) is an extension of the ARIMA model that uses the autoregressive (AR), difference (I), and moving average (MA) components of the ARIMA model and combines them with exogenous variables X.

Exogenous variables refer to independent variables that exist outside the scope of a given time series but possess the ability to impact the dependent variable within that time series. These variables are employed to account for the influence of external factors on the dependent variable.

Exogenous variables such as economic indicators that fluctuate over time, like GDP, interest rates, exchange rates, inflation rates, or other market-related factors such as oil prices or stock indexes. Additionally, they can include categorical variables that distinguish different days of the week.

The ARIMAX model was originally proposed by Box and Tiao (1975) for their study on the effect of gas input velocities on CO2 output concentrations. The equation of the ARIMAX is of the form: Eq.(20)

$$x_t = ARIMA\left(p, d, q\right) + b_1 X_1\left(t\right) + \ ... + \ b_k X_k\left(t\right) + \ \varepsilon \ \left(t\right) \tag{15}$$

F. K-NEAREST NEIGHBORS (KNN)

K-Nearest Neighbors (KNN) is one of the simpler prediction/classification techniques: there is no model to be fit (as in regression). This doesn't mean that using KNN is an automatic procedure. The prediction results depend on how the features are scaled, how similarity is measured, and how big K is set. Also, all predictors must be in numeric form.[14]

KNN is often used as a first stage in predictive modeling, and the predicted value is added back into the data as a predictor for second-stage (non-KNN) modeling. The number of nearest neighbors to compare a record to, K, is determined by how well the algorithm performs on training data, using different values for K. Typically, the predictor variables are standardized so that variables of large scale do not dominate the distance metric.

Similarity (distance) is determined by Euclidean distance or other related metrics. To calculate the k nearest neighbors for 2 data points, x and y, with k attributes: [14]

$$Euclidean : \sqrt{\sum_{i=1}^{k} \left(x_i - y_i\right)^2} \tag{16}$$

$$Manhattan : \sum_{i=1}^{k} \left|x_i - y_i\right| \tag{17}$$

$$Minkowski : (\sum_{i=1}^{k} \left|x_i - y_i\right|)^{\frac{1}{p}} \tag{18}$$

G. BOOSTING LSTM

The idea of ensemble methodology is to build a predictive model by integrating multiple models. It is well-known that

ensemble methods can be used for improving prediction performance.[15]

The most well known model-guided instance selection is boosting. Boosting (also known as arcing—Adaptive Resampling and Combining) is a general method for improving the performance of a weak learner (such as classification rules or decision trees). The method works by repeatedly running a weak learner, on various distributed training data. The classifiers produced by the weak learners are then combined into a single composite strong classifier in order to achieve a higher accuracy than the weak learner's classifiers would have had. [15]

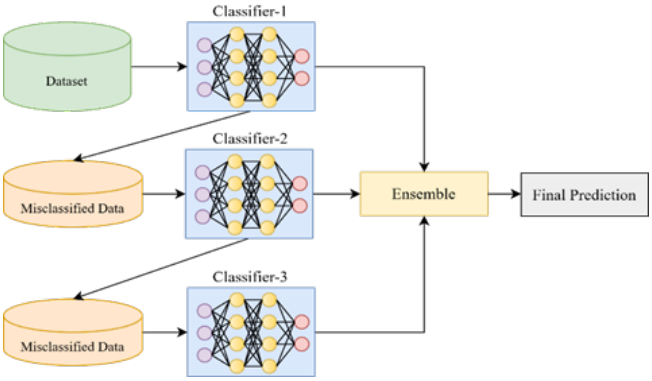


FIGURE 2: Boosting Iteration

The term "Boosting LSTM" not actually a comprehensive algorithm of any Models. It is refers to the combination of LSTM (Long Short-Term Memory) networks with Ensemble learning boosting algorithms. This combination leverages the strengths of both approaches to achieve improved performance, particularly on sequential data tasks.

H. HIDDEN MARKOV MODEL (HMM)

As per Raihan Tanvir and fellow researchers (2023) [16], the Hidden Markov Model (HMM) is a statistical tool utilized across various domains like natural language processing, speech recognition, DNA sequence analysis, data processing, and image analysis. Specifically, this model helps in anticipating concealed states by analyzing observations from real-life situations.

The model's observation is a three dimensional vector representing daily stock information,Eq.(20)

$$O_t = \left(\frac{close - open}{open}, \frac{high - open}{open}, \frac{open - low}{open}\right) \quad (19)$$

After training, an approximation of the Maximum a Posteriori (MAP) technique is employed to test it. When projecting future stock prices, we assume a d day latency. The observation vector O(d+1) is adjusted throughout the whole range of potential values. Because the denominator is invariant with regard to O(d+1) the MAP approximation reduces to (20)

$$(\hat{O}_{d+1}) = argmax\ P(O_1, O_2, \ldots, O_d, O_{d+1}|\lambda) \quad (20)$$

IV. EXPERIMENT

The data-set is crawled by using vnstock · PyPI (v0.2.8.6) from Pypi library through DNSC stock market in 10th Dec, 2023.

A. DATA DESCRIPTIVE

There are 3 data-set fetching 3 pharma company in Vietnam, includes OPC Pharmaceutical Joint Stock Company (OPC), 2469 rows; Vimedimex Pharmaceutical Joint Stock Company (VMD) 2290 rows and IMEXPHARM Pharmaceutical Joint Stock Company (IMP) with 2290 rows from 2nd Jan, 2014 to 8th Dec, 2023.

TABLE 1: Abbreviations and Acronyms

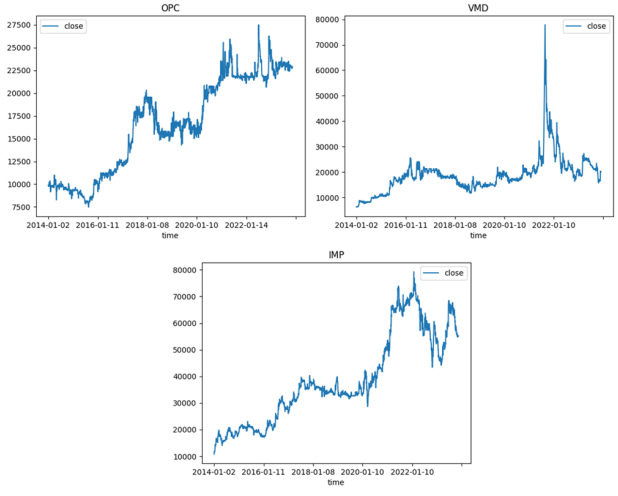
Attribute	Describe
Time	Stock Trading Day
Open	The Opening price of the stock at that time
High	The Highest price can be at that time
Low	The Lowest price can be in at that time
Close	The final price of at that time
Volume	The number of shares of traded at that time
Ticker	Company's stock Code

This paper will focus on the 'Close' price only so we would describe the data-set following the main target.

TABLE 2: Abbreviations and Acronyms

	Companies		
	OPC	VMD	IMP
Count	2,468	2,477	2,467
Mean	16,553	18,719	38,546
Standard deviation	5,180	7,165	16,621
Min	7,470	6,140	10,790
25%	11,172	14,950	23,280
50%	16,700	18,150	34,890
75%	21,760	21,000	52,075
Max	27,490	77,810	79,290
Variance	26,841,383	51,345,375	276,290,079
Skewness	-0.14240	2.37885	0.50166
Kurtosis	-1.35002	11.85794	-0.86655

FIGURE 3: 3 Companies Stock's Prices



B. MODEL EXPERIMENT

1) Linear Regression

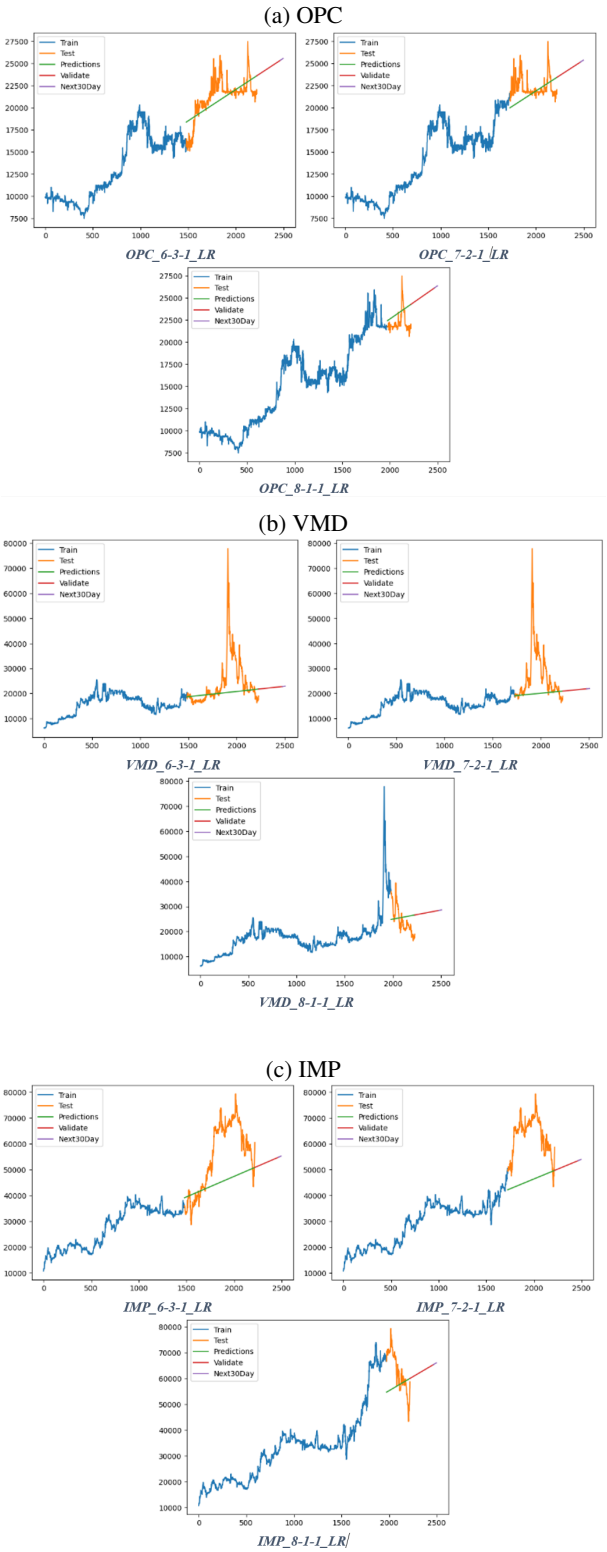


FIGURE 4: Linear Regreesion Model

2) ARIMA

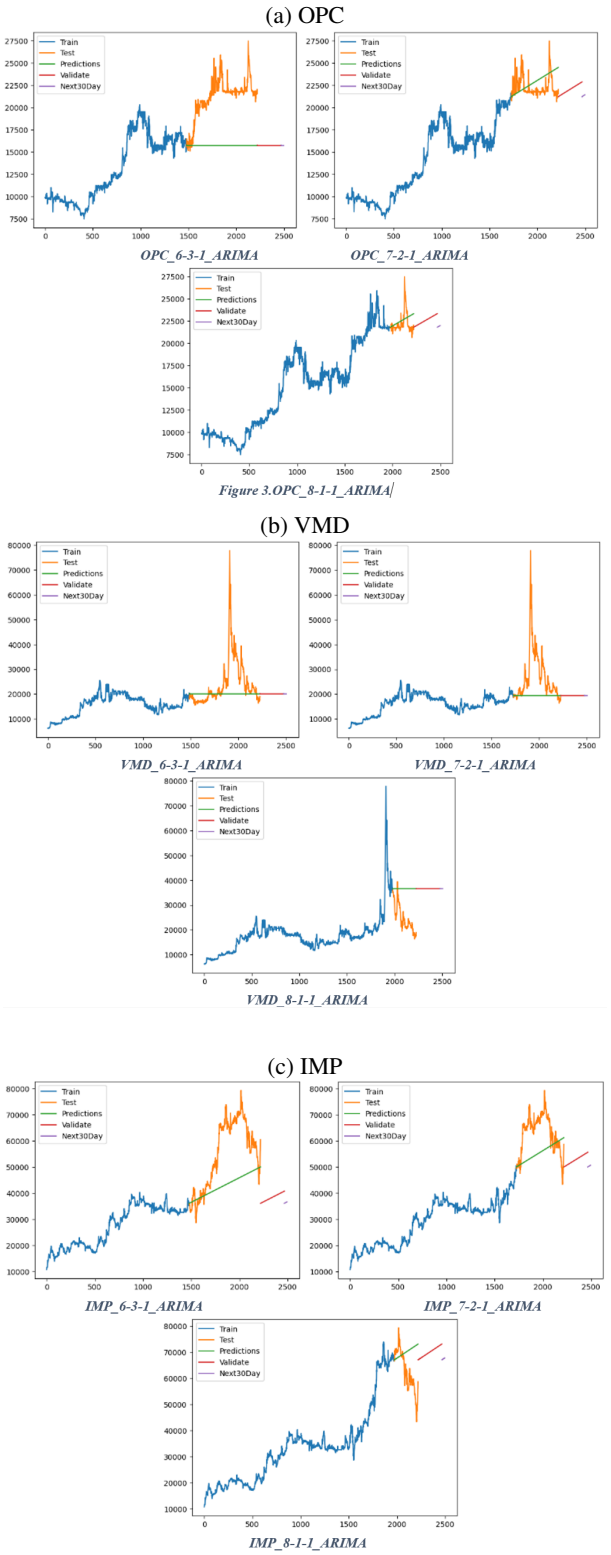
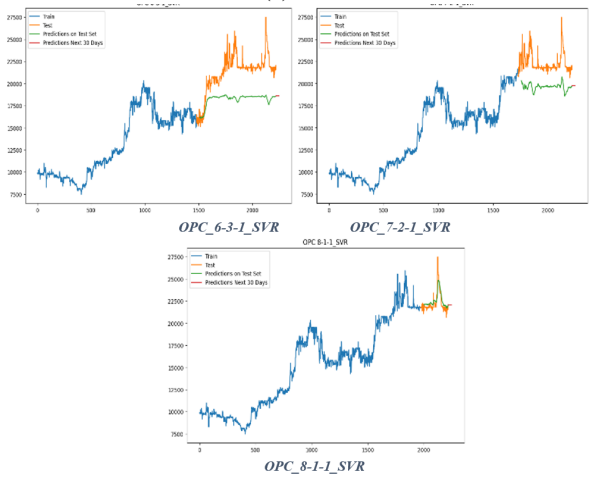


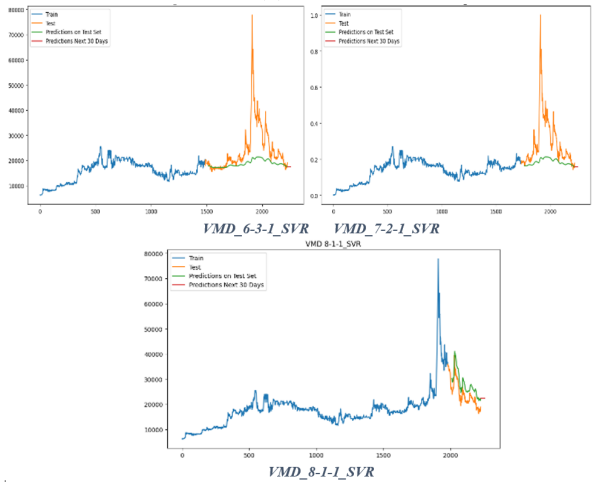
FIGURE 5: ARIMA Model

3) SVR

(a) OPC



(b) VMD



(c) IMP

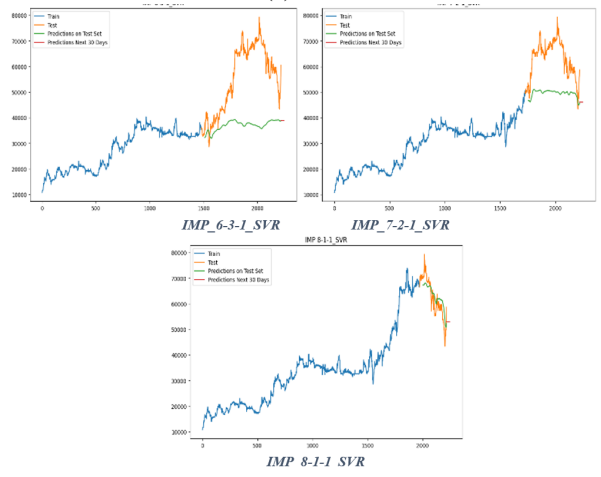
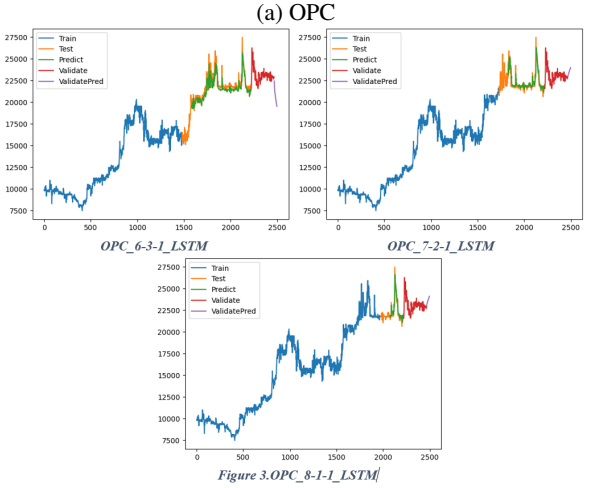


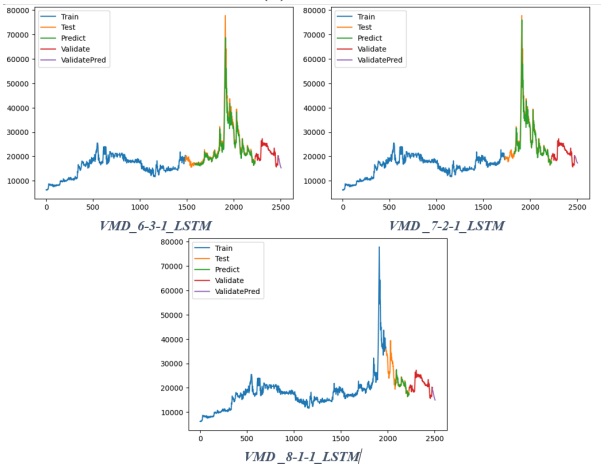
FIGURE 6: SVR Model

4) LSTM

(a) OPC



(b) VMD



(c) IMP

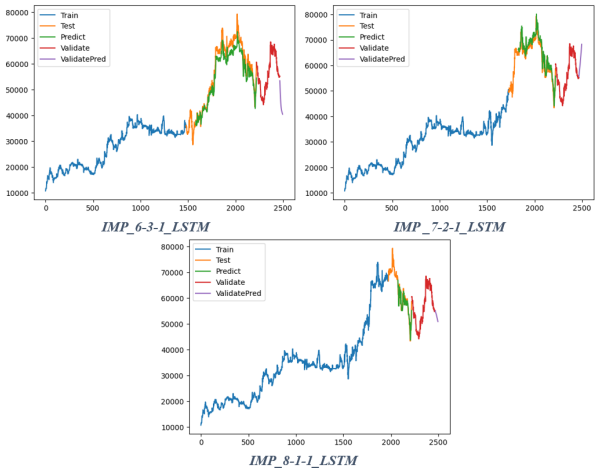


FIGURE 7: LSTM Models

5) ARIMAX

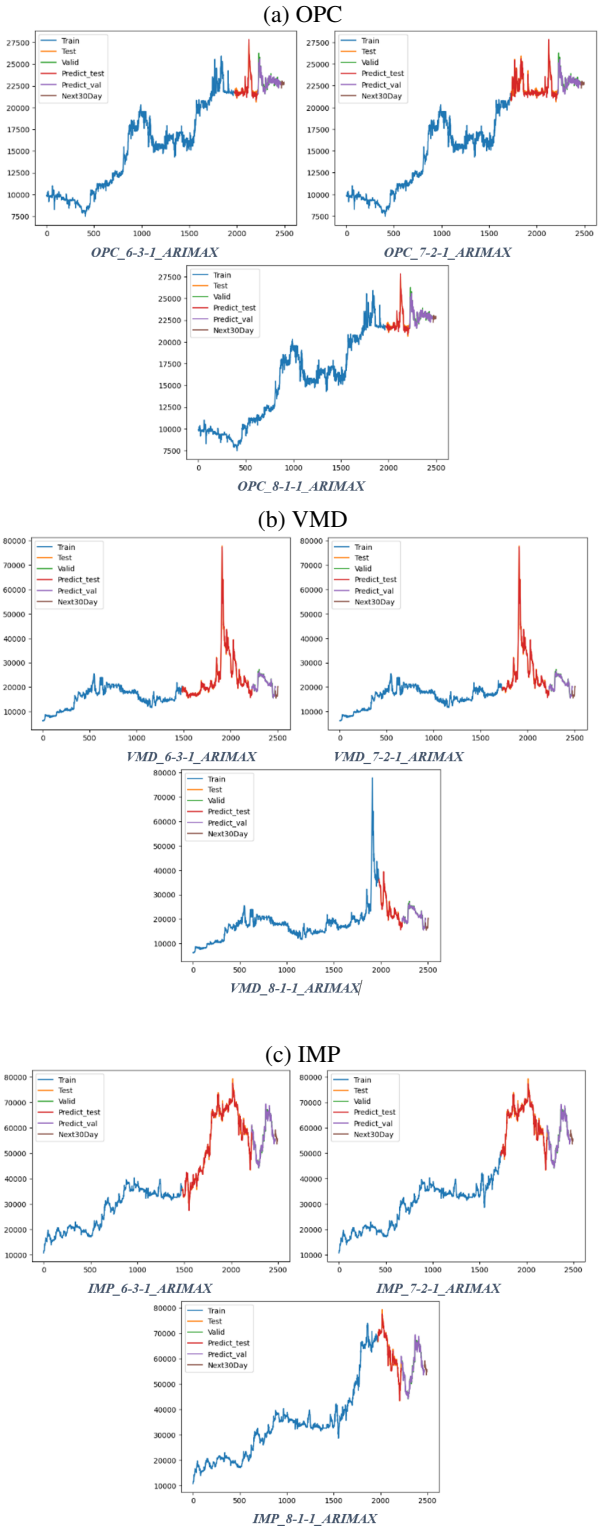


FIGURE 8: ARIMAX Model

6) Boosting LSTM

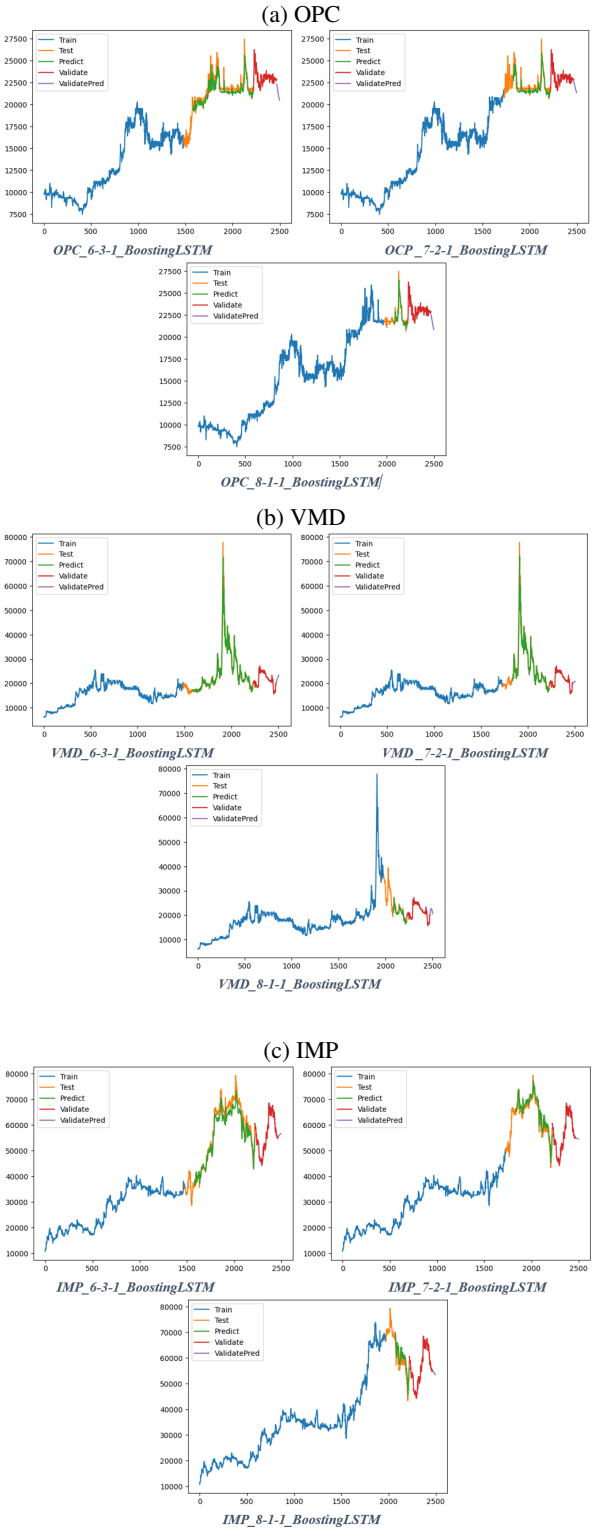
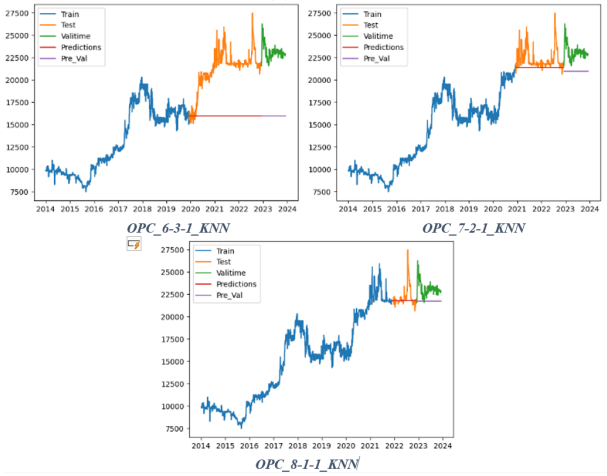


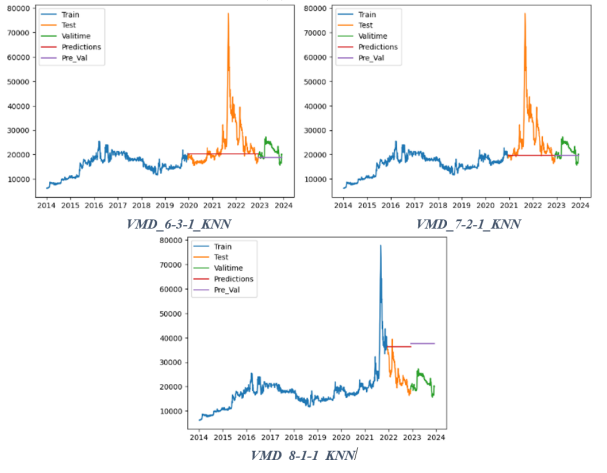
FIGURE 9: Boosting LSTM Models

7) KNN

(a) OPC



(b) VMD



(c) IMP

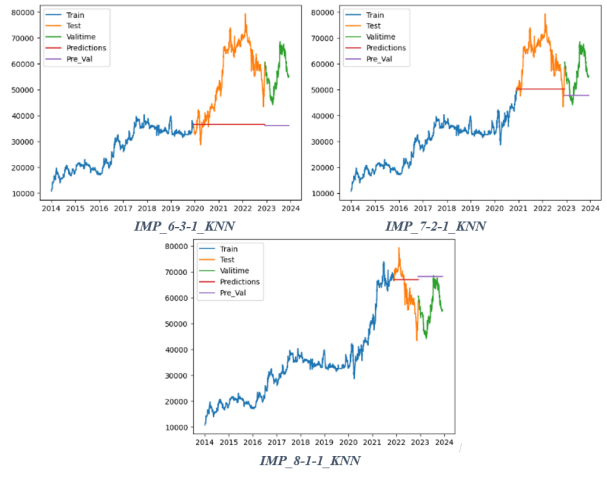
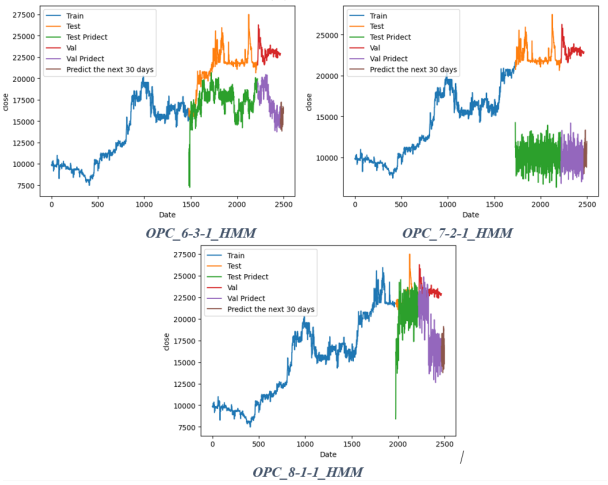


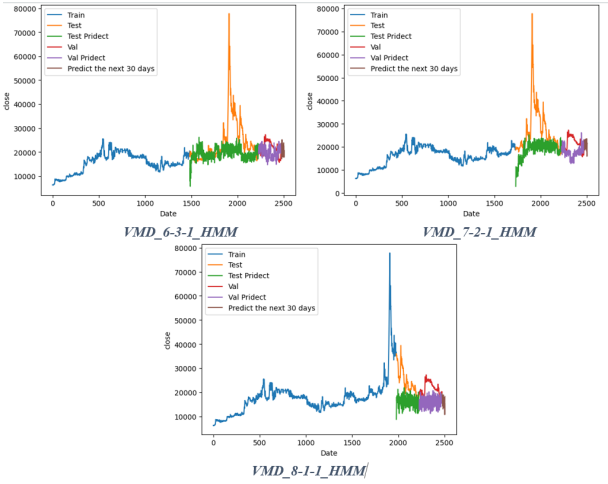
FIGURE 10: KNN model

8) HMM

(a) OPC



(b) VMD



(c) IMP

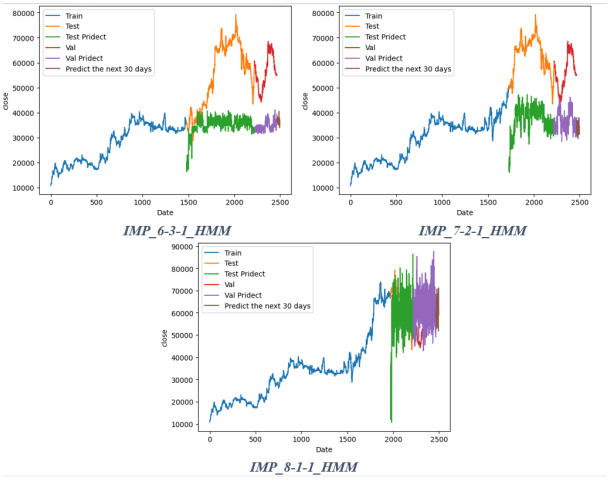


FIGURE 11: HMM model

TABLE 3: OPC Evaluate Models

	OPC			
Model	Ratio	RMSE	MAPE	MAE
LR	6 – 3 – 1	1651.359	6.179	1618.75
	7 – 2 – 1	1700.792	5.453	5.453
	8 – 1 – 1	1603.972	6.365	2218.027
ARIMA	6 – 3 – 1	5903.303	25.004	7342.278
	7 – 2 – 1	1623.326	6.185	1026.099
	8 – 1 – 1	1175.016	3.855	702.727
SVR	6 – 3 – 1	2487.187	9.799	2164.073
	7 – 2 – 1	2284.680	9.038	2053.724
	8 – 1 – 1	1167.387	3.858	887.750
LSTM	6 – 3 – 1	571.944	1.926	486.944
	7 – 2 – 1	403.939	1.154	197.606
	8 – 1 – 1	410.994	1.248	209.331
ARIMAX	**6 – 3 – 1**	**200.118**	**0.641**	**142.276**
	7 – 2 – 1	243.033	0.738	165.235
	8 – 1 – 1	200.118	0.641	142.276
KNN	6 – 3 – 1	5691.622	24.026	7113.816
	7 – 2 – 1	1461.386	4.252	2096.266
	8 – 1 – 1	1192.508	2.444	1348.323
Boosting LSTM	6 – 3 – 1	724.186	2.229	554.786
	7 – 2 – 1	592.866	1.544	339.519
	8 – 1 – 1	737.7	1.878	359.912
HMM	6 – 3 – 1	4387.945	17.882	3899.961
	7 – 2 – 1	12242.473	54.110	12122.201
	8 – 1 – 1	3054.832	9.839	2213.876

TABLE 4: VMD Evaluate Models

	VMD			
Model	Ratio	RMSE	MAPE	MAE
LR	6 – 3 – 1	9649.451	16.289	2418.075
	7 – 2 – 1	12120.135	21.137	2283.258
	8 – 1 – 1	5714.464	21.078	5916.466
ARIMA	6 – 3 – 1	9937.794	18.275	2548.507
	7 – 2 – 1	12457.71	22.682	2883.917
	8 – 1 – 1	13172.876	55.797	15002.287
SVR	6 – 3 – 1	14479.187	13.528	5291.337
	7 – 2 – 1	16930.87	18.189	7341.4975
	8 – 1 – 1	2006.211	4.244	1206.469
LSTM	6 – 3 – 1	1893.718	3.675	474.926
	7 – 2 – 1	1766.212	3.334	333.847
	8 – 1 – 1	706.622	2.307	308.219
ARIMAX	**6 – 3 – 1**	**406.139**	**0.961**	**239.240**
	7 – 2 – 1	474.067	1.172	309.271
	8 – 1 – 1	484.918	1.514	356.366
KNN	6 – 3 – 1	9837.156	18.329	3282.339
	7 – 2 – 1	12302.448	22.017	2744.855
	8 – 1 – 1	12951.406	54.788	16010.962
Boosting LSTM	6 – 3 – 1	3295.73	5.526	894.255
	7 – 2 – 1	4454.588	6.321	614.9697
	8 – 1 – 1	767.626	2.476	505.563
HMM	6 – 3 – 1	9964.968	20.851	6136.134
	7 – 2 – 1	12427.107	26.853	8515.060
	8 – 1 – 1	9884.588	31.466	8311.402

TABLE 5: IMP Evaluate Models

	IMP			
Model	Ratio	RMSE	MAPE	MAE
LR	6 – 3 – 1	14575.349	19.263	5784.508
	7 – 2 – 1	18488.891	26.043	6404.986
	8 – 1 – 1	10409.816	12.933	7164.764
ARIMA	6 – 3 – 1	15721.151	20.404	17852.991
	7 – 2 – 1	10833.288	14.187	5648.963
	8 – 1 – 1	11875.223	16.744	13740.165
SVR	6 – 3 – 1	20897.694	26.534	16737.284
	7 – 2 – 1	23058.191	33.035	21447.705
	8 – 1 – 1	6817.402	8.126	5450.938
LSTM	6 – 3 – 1	3007.534	4.116	2573.013
	7 – 2 – 1	1809.724	2.337	1054.136
	8 – 1 – 1	1472.944	1.93	940.072
ARIMAX	**6 – 3 – 1**	**566.148**	**0.723**	**386.874**
	7 – 2 – 1	637.844	0.713	441.870
	8 – 1 – 1	745.529	0.794	489.264
KNN	6 – 3 – 1	22355.075	30.649	20068.646
	7 – 2 – 1	14449.104	19.35	9161.79
	8 – 1 – 1	8747.178	12.626	11940.73
Boosting LSTM	6 – 3 – 1	2761.74	3.872	1693.198
	7 – 2 – 1	2501.154	3.167	1511.618
	8 – 1 – 1	2377.513	3.279	1337.611
HMM	6 – 3 – 1	23104.035	33.584	20096.218
	7 – 2 – 1	26204.521	40.751	25651.172
	8 – 1 – 1	14171.327	16.793	10559.083

C. INSIGHT

As you can see in table 3,4 and 5 ARIMAX has most less difference between actual value and predict (RMSE: 200.118vnd, MAPE: 0.641%, MAE: 142.276) value at ratio (6-3-1) for both 3 companies. LSTM Model and Boosting LSTM model rank 2nd(Ratio: 8-1-1, RMSE: 767.626vnd, MAPE: 2.307%, MAE: 380.219vnd) and 3rd(Ratio: 8-1-1, RMSE: 767.626vnd, MAPE: 2.4%, MAE: 505.563vnd) respectively. While ARIMA has big gap between 2 compared values (approximately 13k vnd and percent error mostly 56%).

V. CONCLUSION

A. CONCLUSION

Overall, Stocks can be great for investing, but they can also be risky because their value can change a lot. Luckily, new ways of analyzing data using statistics, machine learning, and deep learning have led to making better predictions about stocks.

When we tested different ways of splitting our data into sets, we found that deep learning methods like ARIMAX, LSTM, and Boosting LSTM were better at predicting stock prices than other methods.

So, using these deep learning methods could really help us predict what might happen to stock prices in the future. But remember, stock prices can change because of lots of different things, not just news. To make predictions even better, we need to do more research and consider all these other things that affect stock prices.

B. CHALLENGING

While working on our project to predict Stock Prices in Vietnam, we faced some tough challenges:

Learning New Methods and Creating Reliable Models: It was our first time dealing with prediction models and time series problems. Understanding these new concepts took a lot of time, especially because some models didn't have available information for us to learn from. Additionally, Making prediction models for the stock market is hard. It needed a deep understanding of how the stock market works.

Checking How Good Our Models Were: We used different measures to see how well our models were working. Sadly, the results showed that our models weren't accurate enough, which made it tough to know if they were really effective.

C. FUTURE INTENTION

To tackle these challenges and boost stock price prediction accuracy, we'll implement the following strategies:

1. Using more advanced prediction models: We'll try out newer techniques like Holt-Winters or EST. Holt-Winters focuses on trends and seasonality, excelling in clear patterns, while ETS offers flexibility for diverse or uncertain data structures. Choose Holt-Winters for clear trends and ETS for versatile pattern handling based on the data's clarity and complexity for accurate predictions.

2. Improving models: Reinforcement Learning such as Combining ensemble learning like HMM-LSTM models or XGBoost-LSTM. In order to improve the drawbacks of HMM models or can increase the accuracy of LSTM by using Ensemble Learning like XGBoost model.

3. Improving how we evaluate models: We'll look into and adopt the latest and widely accepted measures used in the stock price prediction field. Measures such as Mean Absolute Scaled Error (MASE), or Symmetric Mean Absolute Percentage Error (SMAPE) will allow us to thoroughly assess the performance of our prediction models.

ACKNOWLEDGMENT

We deeply appreciate the invaluable expertise and enthusiastic guidance provided by Prof. Assoc. Prof. Dr. Nguyen Dinh Thuan and TA. Nguyen Minh Nhut throughout this project. Their unwavering support was instrumental in navigating the challenges we faced in completing this group report.

This project offered us a unique opportunity to collaborate, enhance our teamwork, share knowledge, and importantly, apply theoretical concepts to practical situations. Throughout the project, we applied our existing knowledge and explored novel concepts to achieve optimal results. However, recognizing that imperfections are inevitable, we eagerly await your valuable feedback. Your insights will undoubtedly contribute to our ongoing learning process and aid us in enhancing our future endeavors.

We'd also like to express gratitude to our team members and friends whose mutual support and collective effort significantly contributed to our shared understanding and successful project completion.

REFERENCES

[1] Demirel, Uğur & Çam, Handan & Ünlü, Ramazan. (2020). Predicting Stock Prices Using Machine Learning Methods and Deep Learning Algorithms: The Sample of the Istanbul Stock Exchange. GAZI UNIVERSITY JOURNAL OF SCIENCE. 34. 10.35378/gujs.679103 (accessed Dec, 10th,2023).

[2] Ruochen Xiao, Yingying Feng, Lei Yan, Yihan Ma. (2022) "PREDICT STOCK PRICES WITH ARIMA AND LSTM" arXiv:2209.02407 (accessed Dec, 10th,2023).

[3] Pawaskar, Shreya. (2023). Stock Price Prediction using Machine Learning Algorithms. 10. 2321-9653. 10.22214/ijraset.2022.39891.

[4] Masih, J., Rajasekaran, R., Saini, N., Kaur, D. (2021). Comparative Analysis of Machine Learning Algorithms for Stock Market Prediction During COVID-19 Outbreak. In: Musleh Al-Sartawi, A.M., Razzaque, A., Kamal, M.M. (eds) Artificial Intelligence Systems and the Internet of Things in the Digital Era. EAMMIS 2021. Lecture Notes in Networks and Systems, vol 239. Springer, Cham. (accessed Dec, 10th,2023).

[5] Yara Kayyali Elalem, Sebastian Maier, Ralf W. Seifert, A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks, International Journal of Forecasting,Volume 39, Issue 4, 2023,Pages 1874-1894,ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2022.09.005. (accessed Dec, 10th,2023).

[6] Kuang, Shiyue. (2023). A Comparison of Linear Regression, LSTM model and ARIMA model in Predicting Stock Price A Case Study: HSBC's Stock Price. BCP Business & Management. 44. 478-488. 10.54691/bcpbm.v44i.4858. (accessed December 9, 2023)

[7] James R. Evans, University of Cincinnati, "Business Analytics Methods, Models, and Decisions" -SECOND EDITION,Chapter 8 Trendlines and Regression Analysis, page 238. (accessed December 9, 2023)

[8] James R. Evans, University of Cincinnati, "Business Analytics Methods, Models, and Decisions" -SECOND EDITION,Chapter 8 Trendlines and Regression Analysis, page 250. (accessed December 9, 2023)

[9] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung ,"Time Series Analysis: Forecasting and Control, Fifth Edition". page 88. (accessed December 10, 2023) .

[10] Alakh Sethi, "Support Vector Regression Tutorial for Machine Learning," Analytics Vidhya, September 14th, 2023.Support Vector Regression In Machine Learning (analyticsvidhya.com) (accessed December 9, 2023)

[11] Ian Goodfellow, Yoshua Bengio, Aaron Courville, "Deep Learning", CHAPTER 10. SEQUENCE MODELING: RECURRENT AND RECURSIVE NETS, page 410.

[12] Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation.

[13] Understanding LSTM Networks – colah's blog (accessed Dec. 09, 2023)

[14] Peter Bruce, Andrew Bruce, and Peter Gedeck, "Practical Statistics for Data Scientists", page 238- 329 (accessed December 10,2023)

[15] Rokach, L. Ensemble-based classifiers. Artif Intell Rev 33, 1–39 (2010). https://doi.org/10.1007/s10462-009-9124-7 (accessed December 11, 2023)

[16] Raihan Tanvir , Md Tanvir Rouf Shawon and Md. Golam Rabiul Alam. (2023, 26 January). DSE Stock Price Prediction using Hidden Markov Model. Ahsanullah University of Science and Technology, Dhaka, Bangladesh. (accessed December 12, 2023)