Research Article

Pai-Chou Wang and Tram Thi Hoai Vo*

Stock price prediction based on dual important indicators using ARIMAX: A case study in Vietnam

https://doi.org/10.1515/jisys-2024-0101
received February 04, 2024; accepted August 12, 2024

Abstract: Vietnam's stock market is characterized as a frontier market and focuses on emerging market status by 2025. Tisco Advisory's report showed that Vietnam's stock market is expecting to draw 4 billion in foreign capital in 2024. Despite the appealing nature of the stock market, forecasting stock prices remains a complex endeavor owing to its fast-paced and fluctuating volatility. Effectively forecasting the fluctuation of stock prices has the potential to mitigate the risk associated with stock investments and enhance the overall investment yield. In this research, we combine the advantages of XGBoost for feature selection with the autoregressive integrated moving average (ARIMAX) time series model for forecasting to improve the accuracy of predicting next-day stock prices. A dual important features selection approach is proposed to extract key features for the ARIMAX model from a pool of 87 technical indicators. To demonstrate the effectiveness of this method, we compared it with four other methods – long-short term memory, genetics algorithms with long-short term memory, XGBoost, and Meta Prophet – in predicting the next day's closing price of the Vietnam stock index from January 2013 to April 2023. The results indicate that the performance of our method is better than others and suitable for traders to make stock investment decisions.

Keywords: stock price prediction, technical indicators, ARIMAX model, feature selection

1 Introduction

The stock market is attractive and has never ceased to be vibrant for decades. Apart from viable alternatives, one of the reasons is that stocks may offset the inflationary effect. Inflation may cause an increase in the cost of services and goods while lowering the buying power of money. Stocks reflect ownership in businesses, and their values are affected by many factors, such as the company's financial performance. Fama [1] showed that the future real activity eliminated the negative relationship between inflation rate and stock returns, which made unexpected inflation lose its explanatory power. Therefore, amidst the economic depression and inflation rate growth after the pandemic, various activities will be deployed by the government to control unexpected inflation, which will contribute to a more vibrant financial market.

When people participate in the stock market, the essential goal is to maximize the returns to achieve substantial profits in the short or long term. To achieve that goal, stock price forecast is inevitable. Due to the rapid and dynamic fluctuation of the market, predicting stock prices is regarded as a challenging and appealing task for investors and researchers to improve prediction accuracy. Vietnam's stock market is

*** Corresponding author: Tram Thi Hoai Vo,** College of Business, Southern Taiwan University of Science and Technology, Tainan, 710301, Taiwan, Republic of China, e-mail: db11g210@office.stust.edu.tw
Pai-Chou Wang: Department of Information Management, Southern Taiwan University of Science and Technology, Tainan, 710301, Taiwan, Republic of China, e-mail: pwang@stust.edu.tw

characterized as a frontier market and focuses on emerging market status by 2025. Tisco Advisory's report [2] showed that Vietnam's stock market is expecting to draw 4 billion in foreign capital in 2024. This paper selects the Vietnam stock market as the target to demonstrate the strength of our prediction model.

In terms of stock market prediction, researchers tend to achieve two kinds of outputs: future stock value and stock price movement. Predicting value is a regression task, whereas predicting movement (up or down) is a classification task. Due to the wide broadening of machine learning techniques and computer resources, copious statistical models, machine learning, and deep learning models are applied to the prediction field. These models can be categorized into two types: parametric and nonparametric models. Parametric models like linear regression, auto-regressive integrated moving average (ARIMA), autoregressive conditional heteroskedasticity (ARCH), and generalized autoregressive conditional heteroskedasticity were much more favorable for financial time series forecasting because of their interpretability. They were commonly employed for 1-day ahead stock price predictions [3,4] and the trend of stocks [5,6]. On the other hand, nonparametric models such as long–short-term memory (LSTM), artificial neural networks (ANN), k-nearest neighbors (KNN), and support vector machine (SVM) are favored for both forecasting stock prices [7,8] and predicting stock direction [9,10]. Nonparametric models have become more popular and leading algorithms owing to their remarkable capability of nonlinear mapping and fitting. Furthermore, researchers demonstrated that hybrid models achieved better performances, such as Jing et al. [11] and Tian et al. [12]. Predicting stock prices is inherently challenging due to the dynamic, complex, and noisy nature of stock price data. These data exhibit nonparametric and nonlinear patterns, making prediction tasks susceptible to volatility, irregularities, noise, and changing trends. Feature engineering plays a vital role in the success of prediction tasks [13]. Parametric models rely on predefined assumptions about market patterns and features, and their performances are often improved when these assumptions are met. Nonparametric models are often referred to as "black-box" models because they do not rely on explicit assumptions about the underlying data distribution. Instead, they learn the relationships directly from the data, which can make interpretation more challenging. The performance of nonparametric models is heavily dependent on the quality and relevance of the features used in the model.

Meanwhile, accurate predictions in the stock market also play an important role not only in evaluating a company's performance and producing trading profit for financial speculators and investors but also in making policies for government and financial institutions. They help policymakers to guide regulatory measures to the overall economy. In judging stock values, fundamental and technical analyses are usually applied. Fundamental analysis utilizes financial statements, industry information, and the general economy to predict the movement of the stock price. Morck et al. [14] showed that the stock market is not a sideshow, which indicates stock value can not only be judged by the stock fundamentals. People perceive information differently, and this causes stock values to deviate from the evaluation of fundamentals in the market. To observe the deviation of the market, technical analysis comes as a help.

The technical analysis applies indicators from past data to evaluate the trend or signal of stock prices, and it has been used by top traders and fund managers [15,16]. Many papers applied technical analysis as their superpower to predict stock prices. Fu et al. [17] applied the genetic algorithm (GA) to search the mix of seven commonly used technical indicators and showed that a combination of multiple technical indicators does better than using a single indicator. Dai et al. [18], Ma and Yan [19], and Neely et al. [20] showed that technical indicator-based stock return forecasts have statistically solid and economic significance for both in-sample and out-of-sample prediction accuracy. Chandar [21] used the Convolutional Neural Network (CNN) with 10 technical indicators to predict the 10 years of stock data of eight companies, namely AAPL, BAC, CTSH, GS, HAL, MSFT, OIS, and ORCL from NASDAQ and NYSE. He achieved better accuracy and F1 score than earlier deep learning models. Lee et al. [22] applied LSTM and attention-based BiLSTM with five technical indicators to predict the Taiwan stock market. Chourmouziadis et al. [23] used a fuzzy system with three trend indicators, and their findings demonstrated a markedly greater efficacy compared to the buy-and-hold (B&H) investment strategy as well as the returns accrued from traditional savings account deposits.

In terms of predicting stock prices, machine learning has been regarded as an effective tool for forecasting financial markets recently. The LSTM approach is capable of keeping prior inputs for an extended length of time, and this makes it ideal for dealing with long-term dependencies in sequences where earlier time steps might have a significant influence on subsequent time steps [24]. A combination of the LSTM model and genetic algorithm (GALSTM) increasingly exploits time series forecasting and has achieved remarkable results in recent years [25,26]. The XGBoost model is effective at finding patterns in data to make highly accurate predictions as long as the dataset contains sufficient temporal characteristics [27]. Besides machine learning approaches, Meta Prophet [28], an additive regression model specially carved for business forecast tasks, is also covered in this paper.

However, merely relying on common technical indicators or hybrid methods to enhance the strength of available models is insufficient. Additionally, selecting the appropriate features that support the stock price explanation can yield further benefits. Therefore, we make an effort to find the appropriate technical indicators for the Vietnam market, coupled with a robust model to achieve higher prediction accuracy for the Vietnamese stock market. This paper proposes an autoregressive integrated moving average (ARIMAX) model with dual important indicators to predict the Vietnam stock index (VNINDEX). ARIMAX model is a variant of the ARIMA model and X stands for incorporated exogenous variables in ARIMAX. The ARIMAX model offers a powerful framework for stock price prediction by incorporating relevant external variables into the forecasting process. The ARIMAX model was deployed in a variety of scientific fields with promising performance. Shilpa and Sheshadri [29] employed the ARIMAX algorithm to forecast short-term electrical load and achieved high accuracy with mean absolute percentage error (MAPE) = 2.86%. Anggraeni et al. [30] utilized ARIMAX and vector autoregressive (VAR) methods for forecasting exogenous variables in Indonesia. Variables encompassed consumer rice price, production, harvested area (LP), dry milled rice, as well as rice price in Thailand (HD). The ARIMAX algorithm demonstrated high accuracy in predicting the rice consumer price, achieving a MAPE value of 0.15%. It also provided a more accurate forecast compared to the VAR model. Sutthichaimethee and Ariyasajjakorn [31] found that with different periods of prediction time, either in the long or short term, ARIMAX operated effectively, reflected in a MAPE of less than 3%. Chiu et al. [32] exploited ARIMAX and ARIMA to forecast Greenhouse whitefly incidence; accordingly, the better model was ARIMAX, with temperature and humidity as exogenous factors. In financial sectors, ARIMAX was also used to predict inflation in Sierra in the study of Tamuke et al. [33] or prognosticated Nigeria's GDP in the study of Ugoh et al. [34] or prophesied currency flow at Bank Indonesia in Sulawesi region Suharsono et al. [35] and provided effective and efficient prediction results.

Technical indicators are dual important indicators if they are important in the past and the future. When an indicator has consistently proven to be significant in the past and is expected to remain relevant in the future, it provides validation of patterns observed in historical data. This consistency lends credibility to the indicator's predictive power and increases confidence in its ability to forecast future trends accurately. In this study, there are five dual important indicators – volume weighted average price (VMAP), weighted closing price (WCP), Fibonacci's weighted moving average (FWMA), exponential decay (Decay), and zero *lag* moving average (ZLMA) are selected. In conjunction with the ARIMAX model, mean absolute error (MAE), root-mean-square error (RMSE), mean absolute percentage error (MAPE), normalized root-mean-square error deviation (NRMSD), symmetric mean absolute percentage error (SMAPE), and R^2 are 1.1×10^{-12}, 1.35×10^{-12}, 9.05×10^{-16}, 2.19×10^{-15}, 9.05×10^{-14}, and 99.99%, respectively, using VNINDEX from January 2013 to April 2023. Results are compared to other models, including LSTM, GALSTM, XGBoost, and Meta Prophet. Two more experiments using the dataset from Pham and Ta [36] from 2010 to 2021 and the dataset from Do and Trang [37] from January 3, 2001, to August 30, 2019 are also conducted to evaluate the prediction accuracy of our ARIMAX model with five dual important indicators. Notably, our ARIMAX model also achieved greater accuracy in both datasets. This method helps to improve the accuracy of the previous results by up to 99.9%.

The rest of this article is organized as follows. Section 2 discusses the literature reviews in predicting the stock price in the Vietnam market, including ARIMA, LSTM, gated recurrent unit (GRU), and ANN models. Section 3 presents the dual important indicators, and Section 4 shows the ARIMAX model with five dual important indicators and the empirical results. Results are also compared to LSTM, GALSTM, XGBoost, and Meta Prophet models.

2 Literature reviews

Vietnam Stock Market was founded in 2000 and has a relatively short history compared to established markets in the Western world. It comprises two primary trading platforms: Ho Chi Minh City Stock Exchange (HOSE) and Hanoi Stock Exchange (HNX). The growth and significance of the Vietnamese market within the region have attracted significant researchers' attention, inspiring numerous studies that enhance predictive accuracy for investors, policymakers, and academics. There are attempts to forecast Vietnamese stock prices as well as their fluctuations, and the commonly used methods are ARIMA, LSTM, GRU, and ANN.

The ARIMA model is the most popular one in time series analysis. The ARIMA model, developed in 1974 by Box and Jenkins, employs latency methods to give a decent prediction model. The future value is a linear mixture of past values and past errors. In summary, the ARIMA model is influenced by three distinct variables (p, d, q), where p represents the auto-regressive variables, d represents the number of differences, and q represents the moving average variable. The Akaike's information criteria (AIC) score was used to check the performance of the ARIMA model. The model exhibiting the lowest AIC score is deemed optimal, which strikes a balance between its capacity to capture the dataset and its ability to prevent overfitting accurately. Co et al. [38] applied the ARIMA model to the univariate dataset of VNINDEX in the period from 2000 to 2018. The lowest MSE was 200, and the results showed that the increased duration of days utilized for forecasting correlates with a diminished MSE achieved. Ngan [39] showed that the ARIMA model is suitable for estimating foreign exchange rates in Vietnam in a short period. Hằng and Dũng [40] researched Vietnam's GDP Forecasting from 2020 to 2025 using ARIMA model with the dataset from 1985 to 2019, and the smallest MAE attained is 1.3. Lee et al. [41] studied crude oil prediction in the US and European markets, the results indicated that ARIMA obtained better performance than SARIMA with the smallest MAPE value of 0.05 in the US market and 0.09 in the Europe market. Hybrid models of integrating the ARIMA and ANNs [42] in modeling the linear and nonlinear behaviors in the dataset were conducted in the context of the Vietnamese stock market. Results showed that at most 11.11% MAPE value was improved compared to the individual ARIMA or ANN models. One exceptional study from Bao and My [43] showed that the hybrid ARIMA model with the ANN model generates better prediction and estimation than other traditional methods, including ARIMA and ANN models, with 0.006 and 0.00005 for MAE and RMSE, respectively. The hybrid ARIMA and ANN study of Le and Nguyen [44] discovered that the hybrid of ARIMA and ANN was likely to outperform a single ARIMA model or ANNs as forecasting a multiple-step prediction of Vietnam CPI prediction, and the best MSE value was 0.2210.

Due to the efficiency of handling lengthy sequential data, the LSTM model has become well known in the field of time series prediction. LSTM is an extension of the recurrent neural network (RNN), which can prevent gradient problems from vanishing in RNN. LSTM networks use specialized memory cells and gating mechanisms to retain and discard data over lengthy periods selectively. Do and Trang [37] applied the hierarchical adaptive neuro-fuzzy inference systems (ANFIS) and LSTM models to forecast VNINDEX price one day ahead with a dataset from January 2001 to August 2019. LSTM had the highest prediction ability with a MAPE value of 0.009. Pham and Ta [36] proposed ARIMA, SARIMA, and LSTM methods to predict VNINDEX stock prices from 2010 to 2021. The LSTM model generated the best results (RMSE = 0.011, MAE = 0.007) compared to the other methods. Thu et al. [45] proposed a neighborhood LSTM model using a sliding window (LSTM–kNN) for stock price prediction for US and Vietnam stock market datasets. Results showed that the LSTM-kNN model had higher accuracy than other models such as CNN, LSTM, random forest, linear regression, and moving average. Yıldırım et al. [46] processed macroeconomic and technical indicators separately using LSTM, and combining the two results could achieve high accuracy performance. Aminimehr et al. [47] deployed the LSTM model to predict stock market trends in the Tehran stock exchange. Different feature selection strategies were applied to 78 technical indicators. LSTM with deep feature extraction method by wavelet yielded the highest accuracy in predicting the Tehran Stock Exchange index. The study of Zhuge et al. [48] demonstrated that the LSTM model could learn long-term dependence. The LSTM algorithm possessed the capability to significantly improve the precision of experimental outcomes when juxtaposed with the RNN and multilayer perceptron (MLP) models in the prediction of stock opening prices, utilizing 15 distinct variables derived from the Shanghai Composite Index alongside emotional data as input variables. Besides, LSTM was also known as an outstanding time series prediction model in various fields like weather, pandemic, vegetation dynamics, and pharmaceutical fields [49–51].

A GRU is a kind of RNN. It offers several advantages, such as faster convergence and fewer parameters compared to the LSTM model. Minh et al. [52] proposed a two-stream gated recurrent unit (TGRU) to forecast the price trends of VNINDEX using VietStock news and stock prices. Results were compared to GRU and LSTM algorithms using the dataset from October 2006 to November 2013. TGRU outperformed other models with 66.32% accuracy. Huynh et al. [53] proposed a bidirectional GRU with historical stock price and online financial news data from October 2006 to November 2013 to anticipate stock directions in the future. The result achieved nearly 60% correctness. Thanh and Meesad [54] combined linear support vector machine weight for feature selection and SVM classifier achieved up to 75% accuracy in comparison with ANN, KNN, and Naïve Bayes. Datasets were real-time news articles and daily stock index prices collected from Vietnamese websites over three years, from 2010 to 2012.

Methods like eXtreme gradient boosting (XGBoost), hybrid model integrating genetics algorithms with long-short term memory (GALSTM), and Meta Prophet models are also considered in this article. They have not received much attraction in predicting the Vietnam stock market, but they performed well in many areas. Boosting is an ensemble learning technique that increases prediction accuracy by aggregating weaker models to string one. XGBoost mainly relies on decision trees as base classifiers and enhances its performance by using second-order Taylor expansion to capture more complex information and relationships in the data. This combination of techniques has made XGBoost a highly effective algorithm for various machine-learning tasks [55]. XGBoost supports the parallel selection of split points, and the model training cost is much less. Chen and Fan [56] used the XGBoost algorithm and vehicle data to predict freeway travel time as well as investigate the effects of different parameters on model performance. Results showed that the XGBoost-based model produces better prediction accuracy and efficiency results compared to the gradient boosting model. Wang and Guo [27] showed that the classification and regression trees (CART) in XGBoost has a strong performance in nonlinear data prediction. It could be used to explain the nonlinear association among models and the dependency between variables in forecasting stock prices. Jeyakarthic et al. [57] combined XGBoost with the Bat algorithm to predict Apple and Facebook stocks and achieved a maximum accuracy of 96.42%. Furthermore, Budholiya et al. [58] assisted cardiologists in enhancing heart disease diagnosis by utilizing an optimized XGBoost model with Bayesian optimization, and the results reached a high prediction accuracy of 91.8%.

GA searches for near-optimal solutions based on an evolution process. In conjunction with the LSTM model, a GA is applied to filter the desirable features for the LSTM algorithm or optimize the hyperparameter configuration to enhance the model's performance. Bouktif et al. [59] employed a GA to determine the optimal time lags and the number of layers to improve the predictive performance of LSTM models. Stajkowski et al. [60] determined the optimal time window and number of units of the memory cell through GA feature selection to enhance the LSTM model. It outperformed the RNN model in water temperature prediction in Ontario. In the financial sector, Chen and Zhou [61] determined the importance of factor ranking using GA. Then, it combined the near-optimal factors and the LSTM algorithm to predict the stock price using the China Construction Bank and the CSI 300 stock datasets. The result showed an MSE value of 0.0039. He et al. [62] showed that the GALSTM model outperformed the LSTM algorithm and other machine learning models such as Random Forest, KNN, and SVM in forecasting the rising or falling of the stock.

Meta Prophet [28] is a procedure used by Facebook's Core Data Science team to predict time series data in 2017. Meta Prophet model may accurately predict without an excessive number of complex or multidimensional sources. At the same time, it also does well even when the data contains outliers and missing values. This model is an important tool for predicting and forecasting time series with trends, seasonality, and holidays. Sah et al. [63] and Satrio et al. [64] applied the Prophet to predict Coronavirus disease in Indonesia. Xie et al. [65] used the ARIMA and Prophet models to predict the incidence of Hand, foot, and mouth disease. The Prophet algorithm generated higher accuracy than the ARIMA model. In the financial sector, Weytjens et al. [66] forecasted accounts receivable cash flows. The results showed that Prophet was worse than MLP and the LSTM model, but it was better than ARIMA. Yenidoğan et al. [67] demonstrated the superiority of the Prophet model over the ARIMA model in the context of Bitcoin forecasting using a dataset from May 2016 to March 2018. Notably, the Prophet model exhibited remarkable accuracy that closely aligns with actual values with a precision rate of up to 94.5%, and the ARIMA model got a precision rate of only 68%. In predicting water quality using the water quality index dataset spanning from 2002 to 2007, Hemdan et al. [68] found that the error rate of the LSTM model was higher compared to the Prophet model.

As we know, technical indicators provide trends and signals of stocks in developing stock trading strategies, and previous studies also showed that technical analysis can help predict stock prices. To extract effective indicators to improve the prediction accuracy of the Vietnamese stock market, this paper proposes dual important indicators. Technical indicators are treated as dual important indicators if they are important to predict stock prices in the past and the future. After dual important indicators are extracted, an ARIMAX model is developed in conjunction with these dual important indicators. Results are compared to LSTM, XGBoost, GALSTM, and Meta Prophet algorithms. Before we present our ARIMAX model, dual important indicators are introduced.

3 Dual important indicators

Technical indicators are considered effective tools that provide an overall perspective and the ability to anticipate trends and signals in developing stock trading strategies. It would be great if we could identify the significance of indicators and show their relative contribution to the model's output. Past researchers used to apply predetermined indicators such as Vargas et al. [69], Dai et al. [18], Ma and Yan [19], and Neely et al. [20] or search for the best indicators such as Haq et al. [70] and Yuan et al. [71] of each day to predict the future price or the trend. Indicators obtained from closing data of each day stand for the past. Instead of merely applying indicators from the past to predict future prices or trends, this paper tries to incorporate indicators that are useful for the future, and we propose the dual important indicators.

Indicators are dual important if they are important to predict stock prices for the past and the future. At the beginning, *lag*-0 data are constructed using the closing price and 87 indicators from stock prices of each day. An XGBoost model is applied to find important indicators, and the top 10 indicators are extracted as important indicators for the past. XGBoost model is an ensemble machine learning algorithm based on the classification and regression trees. Features that contribute more to reducing the loss function are given higher importance. Feature importance in a decision tree involves assessing the impact of each attribute's split point on the overall performance measure of the tree. To achieve better performance for the XGBoost model, Randomized CV is employed for hyperparameter selection, and the best hyperparameters for the XGBoost model are shown in detail in Table 1.

Table 1: Hyperparameters tuning for XGBoost model

Parameters	Search space	Best value
learning_rate	(0.01, 0.5)	0.113970831
n_estimators	(100, 300)	290
max_depth	(3, 100)	6
min_child_weight	(1, 5)	2
Gamma	(0.01, 0.5)	0.222577937
reg_alpha	(0.01, 0.5)	0.397566412
reg_lambda	(0.01, 0.5)	0.479749471
random_state	(10, 100)	27

After we obtain important indicators for the past, the same operation is applied to *lag*-1 data, which is generated by shifting the *lag*-0 data by one timestamp for all indicators to extract important indicators for the future. After important indicators are extracted for the past and the future, the common indicators for the past and the future are called dual important indicators. The algorithm for extracting dual important indicators is shown, and the experiment dataset using the VNINDEX from January 2013 to April 2023 in the next section is applied. Table 2 lists 87 investigated technical indicators in this paper.

Table 2: Eighty-seven investigated technical indicators

No.	Name
1	Volume weighted average price (VWAP)
2	Weighted closing price (WCP)
3	Fibonacci's weighted moving average (FWMA)
4	Exponential moving average (EMA)
5	Accumulation/distribution oscillator (ADO)
6	Arnaud Legoux moving average (ALMA)
7	Volume weighted moving average (VWMA)
8	Variable index dynamic average (VIDYA)
9	Linear regression moving average (Linreg)
10	Zero *lag* moving average (ZLMA)
11	Tim Tillson's T3 moving average (T3)
12	Triple exponential moving average (TEMA)
13	Normalized average true range (NATR)
14	Holt-winter moving average (HWMA)
15	Chande momentum oscillator (CMO)
16	Kaufman's adaptive moving average (KAMA)
17	Triangular moving average (TRIMA)
18	Rolling mean absolute deviation (MAD)
19	Symmetric weighted moving average (SWMA)
20	Moving average convergence divergence (MACD)
21	Hull moving average (HMA)
22	Double exponential moving average (DEMA)
23	Pascal's weighted moving average (PWMA)
24	Weighted 10-day moving average (WMA)
25	Jurik moving average average (JMA)
26	Sine weighted moving average (SinWMA)
27	Entropy (Entropy)
28	Negative volume index (NVI)
29	McGinley dynamic indicator (MCGD)
30	Stochastic K% (K)
31	Mass index (MassI)
32	Positive volume index (PVI)
33	Price-volume (Pvol)
34	Rolling median (median)
35	Chande Forcast oscillator (CFO)
36	Coppock curve (Coppock)
37	Accumulation/distribution (AD)
38	On balance volume (OBV)
39	Log return (Log_return)
40	Larry William's R% (WR)
41	Detrend price oscillator (DPO)
42	Money flow index (MFI)
43	Ulcer index (UI)
44	Balance of power (BOP)
45	Rolling standard deviation (Stdev)
46	Momentum (MTM)
47	Bias (Bias)
48	Awesome oscillator (AO)
49	Rolling Z score (Zscore)
50	Rolling skew (skew)
51	Choppiness index (Chop)
52	Ease of movement (EOM)
53	Average true range (ATR)
54	Relative strength index (RSI)
55	Chaikin money flow (CMF)
56	Simple moving average (SMA)

(Continued)

Table 2: *Continued*

No.	Name
57	Vertical horizontal filter (VHF)
58	Rolling kurtosis (kurtosis)
59	Q Stick (Qstick)
60	Psychological line (PSL)
61	Commodity channel index (CCI)
62	Rate of change (ROC)
63	Price volume rank (PVR)
64	Price-volume trend (PVT)
65	Relative vigor index (RVI)
66	Correlation trend indicator (CTI)
67	Absolute price oscillator (APO)
68	Efficiency ratio (ER)
69	WildeR's moving average (RMA)
70	Price distance (Pdist)
71	Inertia (inertia)
72	Relative strength Xtra (RSX)
73	Exponential decay (decay)
74	True range (True_range)
75	Center of gravity (CG)
76	Ultimate oscillator (UO)
77	Slope (slope)
78	Stochastic D% (D)
79	Ehler's super smoother filter (SSF)
80	Even Better SineWave (EBSW)
81	Pretty good oscillator (PGO)
82	Elder's force index (EFI)
83	Decreasing (decreasing)
84	Increasing (increasing)
85	Rolling variance (variance)
86	Rolling quantile (quantile)
87	Percent return (Percent_return)

Algorithm of extracting dual important indicators using XGBoost

Input: 87 explanatory indicators matricies (X_0, X_1) and target attribute vectors (Y_0, Y_1) for *lag*-0 and *lag*-1 datasets.

Output: Dual Important Indicators

1. Split *lag*-0 dataset into 80% training and 20% testing data ($trainX_0$, $trainY_0$, $testX_0$, $testY_0$)
2. Create a parameter dictionary for Random Search and execute XGBoost using all features and Random Search to choose the optimal parameters
3. Rank feature relative importance and select the top 10 important features for *lag*-0 dataset
4. Split *lag*-1 dataset into 80% training and 20% testing data ($trainX_1$, $trainY_1$, $testX_1$, $testY_1$)
5. Create a parameter dictionary for Random Search and execute XGBoost using all features and Random Search to choose the optimal parameters
6. Rank feature relative importance and select the top 10 important features for *lag*-1 dataset
7. Return important indicators appear in both selected important features in *lag*-0 and *lag*-1 datasets.

Table 3a and b shows the top 10 important indicators obtained from *lag*-0 and *lag*-1 data, respectively. Five dual important indicators – VMAP, WCP, FWMA, exponential decay (Decay), and ZLMA, are found and shown in bold and italic. Formulas of selected dual important indicators are presented in Table 4. Data statistics and pairwise correlations are shown in Table 5 and Figure 1, respectively. Figure 1 shows a strong positive

Table 3: Top 10 important indicators extracted from *lag*-0 and *lag*-1 datasets using XGBoost

***a*. Important indicators from *lag*-0 dataset**		***b*. Important indicators from *lag*-1 dataset**	
Indicator	**Importance**	**Indicator**	**Importance**
VWAP	0.6803749	**WCP**	0.479226
WCP	0.2902989	JMA	0.243831
FWMA	0.0257779	**Decay**	0.160197
EMA	0.0021383	**VWAP**	0.103357
PVT	0.0004319	TEMA	0.005616
ALMA	0.0003241	**ZLMA**	0.002023
Decay	0.0002018	HMA	0.001285
VIDYA	0.0000820	KAMA	0.00122
OBV	0.0000437	DEMA	0.001055
ZLMA	0.0000366	**FWMA**	0.000644

Table 4: Formulas of selected five dual important indicators

Technical indicators	Equation	Description
Volume weighted average price (VWAP)	$\mathrm{VWAP} = \frac{\sum\left(\mathrm{Volume} \times \frac{\mathrm{High} + \mathrm{Low} + \mathrm{Close}}{3}\right)}{\sum \mathrm{Volume}}$ (1)	A critical metric of execution quality for big orders that institutional investors frequently utilize, and most institutions trade 50% of their transactions using VWAP [72]
WCP	$\mathrm{WCP} = \frac{(\mathrm{Close} \times 2) + \mathrm{High} + \mathrm{Low}}{4}$ (2)	It doubles the importance of the closing price, resulting in smoother data across the entire trading period [73]
Fibonacci's weighted moving average (FWMA)	$\mathrm{FWMA}_t = \frac{\mathrm{Close}_t \times \mu_1 + \mathrm{Close}_{t-1} \times \mu_2 + \ldots + \mathrm{Close}_{t-n+1} \times \mu_n}{\mu_1 + \mu_2 + \ldots + \mu_n}$ (3) where $\mu_n = \mu_{n-1} + \mu_{n-2}$; n is the length of the period	It is a moving average method that calculates multiple exponential moving averages using lookback periods based on the Fibonacci sequence. The goal of Fibonacci retracement is to identify potential support levels and subsequent resistance levels [74]
Exponential decay (Decay)	$\mathrm{Decay}_t = \frac{w_1\mathrm{Close}_t + w_2\mathrm{Close}_{t-1} + \ldots + w_n\mathrm{Close}_{t-n+1}}{w_1 + w_2 + \ldots + w_{t-n+1}}$ (4) n is the length of the period; $w_1 = 2$, $w_2 = \frac{w_1}{2}$, …, $w_n = \frac{w_{n-1}}{2}$	It is used to transfer signals from the past into the future, and it works well with algorithm trading and machine learning functions [75]
Zero *lag* moving average (ZLMA)	$\mathrm{ZLMA}_t = \alpha \times (2 \times \mathrm{Close}_t - \mathrm{Close}_{t-lag}) + (1-\alpha) \times \mathrm{ZLMA}_{t-1}$ (5)	A technical indicator that allows traders to see market trends more clearly by minimizing *lag* and smoothing out price swings [76]

relationship between dual important indicators and closing prices. After we extract dual important indicators, the dataset, which contains the closing price, and dual important indicators, are applied to all experiments in the next section.

Table 5: Data descriptive statistics

	Close	VWAP	WCP	FWMA	Decay	ZLMA
Count	3,757	3,757	3,757	3,757	3,757	3,757
Mean	856.40	856.11	856.18	856.16	858.66	856.34
Std	284.12	283.88	283.93	284.08	285.07	284.15
Min	442.85	443.88	443.75	446.68	443.37	446.02
Q1	589.81	590.69	590.47	589.73	591.51	590.73
Q2	854.78	855.08	854.30	854.65	856.13	854.28
Q3	1028.01	1030.32	1029.12	1030.94	1035.91	1029.44
Max	1528.57	1529.32	1529.11	1526.57	1528.57	1531.26

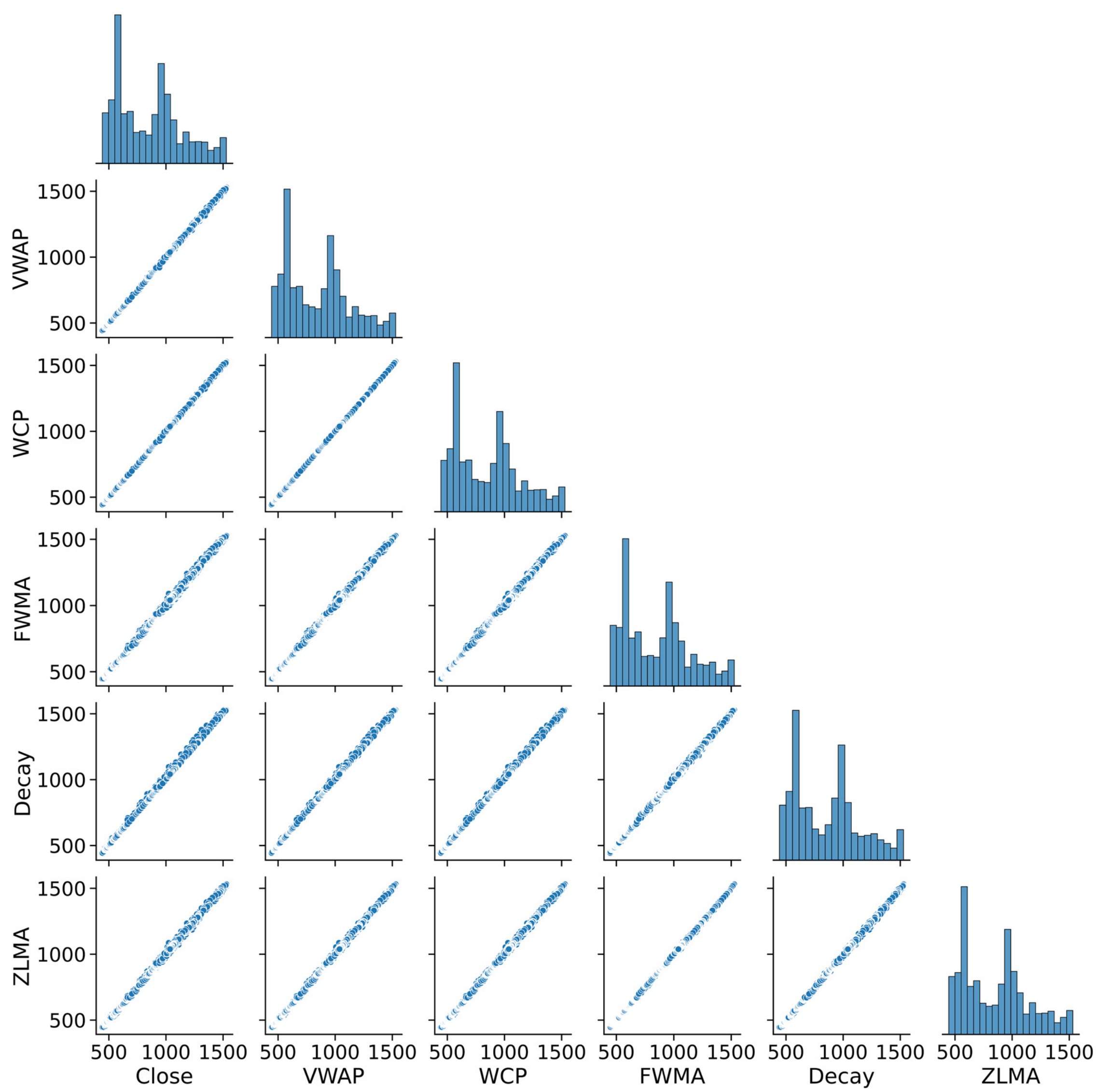

Figure 1: Selected dual important indicators pairwise correlation. Source: Created by the author.

4 Experiments

Before we present our ARIMAX model using five dual important indicators found in the previous section, Figure 2 illustrates the comprehensive workflow of all experiments. The dataset comprises closing, opening, high, low prices, and volume data for the VNINDEX. Vietnam stock Index or VNINDEX is a capitalization-weighted index of all the companies listed on the HOSE. The VNINDEX dataset was sourced from the Investing.com website from January 2, 2013, to April 25, 2023. At the preprocessing stage, missing data are filled in gaps with the previous value which is based on the understanding that stock exchanges remain closed during holidays or weekends, resulting in unchanged prices [77]. Subsequent to the preprocessing stage, 87 technical indicators are computed, and dual important indicators are extracted, which is covered in the previous section. After we found dual important indicators, the dataset is subsequently partitioned into training and testing sets, with 80% allocated for the training phase and the remaining 20% for testing. The transformation of the dataset is contingent upon the specific requirements of each testing model, and detailed descriptions of these transformations will be covered in the respective model experiments. To show the strength of our ARIMAX model with five dual important indicators, two more experiments are conducted using the first dataset from Pham and Ta [36] from 2010 to 2021 and the second dataset from Do and Trang [37] from January 3, 2001, to August 30, 2019. Results for these two additional experiments are shown at the end of this section.

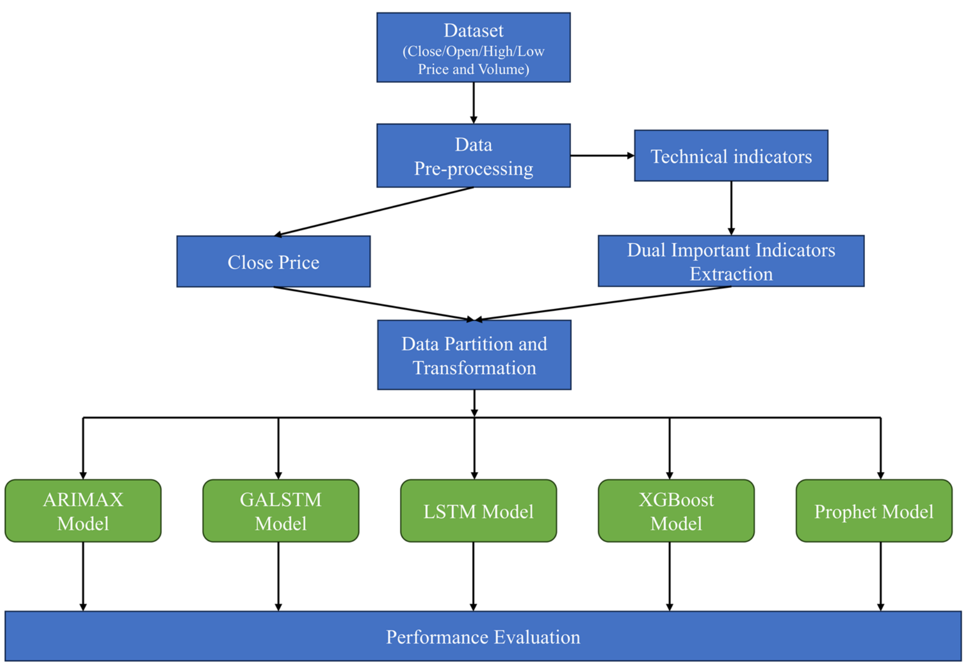

Figure 2: The empirical processes. Source: Created by the author.

After we illustrate the workflow of all experiments, our ARIMAX model with dual important indicators is presented. The ARIMAX model is an extension of the ARIMA model. Unlike ARIMA, which is exclusively employed for univariate time series, ARIMAX excels in explaining exploratory variables. The foundation of ARIMA rests upon the concept of autocorrelation. The ARIMA can be employed to model time series data that is either stationary or nonstationary, with the potential for achieving stationarity through differentiation. ARIMA model denoted by ARIMA (p, d, q) is a combination model of AR (autoregressive) and MA (Moving Average) who have undergone a differencing process, p lags for the AR part, and q lags for the MA part. The auto-regressive part establishes a linear connection between the predicted value at time t and the previous values of the time series. The moving average part establishes a linear association between the predicted value at time t and a Gaussian distribution of independent and identically distributed (i.i.d) samples [78]. The ARIMAX is a variation of ARIMA and X stands for exogenous variables. In the study by Sharma et al. [79], the ARIMAX model was used to predict Indian stock price with/without considering sentiment analysis, and the best performance metrics belong to the ARIMAX model aligned with sentiment analysis with the smallest RMSE value of 18.3. Abd et al. [80] utilized the ARIMAX model to forecast the Dow Jones index, incorporating daily prices of Brent crude as an external factor. Their study achieved the best performance with an RMSE value of 485. Serafini et al. [81] compared the performance between ARIMAX and RNN in predicting Bitcoin time series. The findings indicated that RNN demonstrated superior performance with an MSE value of 0.00100892, compared to ARIMAX which achieved an MSE value of 0.00011893. Patil [82] conducted a study to investigate the Indian economy through forecasting the stock prices from various industrial sectors. ARIMAX, ARIMA, and LSTM are the three selected models used in the study. ARIMAX outperformed the others with an RMSE value of 165.85. In this study, five dual important indicators are exogenous variables and ARIMAX's equation is demonstrated as equation (6).

$$\Delta^d Y_t = \sum_{i=1}^{p} \varphi_i \Delta^d \varepsilon_{t-i} + \sum_{j=1}^{q} \theta_j \omega_{t-j} + \sum_{m=1}^{M} \beta_m a_t^m + \omega_t, \tag{6}$$

where Y_t is the time series under consideration for forecasting; φ is the autoregressive coefficient, θ is the moving average coefficient; ω_t is the residual process at time t; M is the number of exogenous variables; β is the exogenous inputs coefficient; a_t^m is the exogenous variable m-th influencing Y_t; Δ^d indicates the d-th order difference of the modeled time series. A first-order differenced series is computed as the discrepancy between consecutive observations in the original time series: $\Delta^1 Y_t = Y_t - Y_{t-1}$.

When experimenting with the ARIMAX model, three main steps should be gone through:

(1) Data stationarity: In the ARIMAX model, data stationarity preprocessing is necessary. If the original dataset lacks stationarity, it must undergo differencing to achieve greater stationarity before applying the ARIMAX model.

(2) Model selection: Compare the performance of different ARIMAX models using appropriate criteria such as AIC (Akaike information criterion) or BIC (Bayesian information criterion). Choose the model with the lowest information criterion value, indicating the best trade-off between model complexity and goodness-of-fit.

(3) Model diagnostic checking: Evaluate the goodness of fit of the ARIMAX model by examining diagnostic plots such as the residuals plot and autocorrelation of residuals. Ensure that the residuals are uncorrelated and normally distributed.

To check out the stationary of time series dataset, the unit root test is conducted through the augmented Dickey-Fuller (ADF) test with the null hypothesis of a unit root presence. If the null hypothesis is accepted, we assume there is a unit root in time series data and a first-differencing technique should be employed before running a model. Analogously, if the null hypothesis is rejected, the data are stationary and can be used without differencing. The ADF test results are demonstrated in Table 6. The original dataset is nonstationary, yet after the first differencing, all the variables are stationary when p-value <0.05 and $|t\text{-statistic}| > |\text{critical value (1\%)}|$.

Table 6: The ADF test results

	Original time series		After 1st differencing	
Time series	***t*-statistic**	***p*-value**	***t*-statistic**	***p*-value**
Close	−1.456	0.554	−27.635	0.000
VWAP	−1.396	0.584	−13.448	0.000
WCP	−1.434	0.566	−30.352	0.000
FWMA	−1.479	0.544	−21.656	0.000
Decay	−1.355	0.603	−11.307	0.000
ZLMA	−1.375	0.594	−13.559	0.000

Note: Critical values at 1%, 5%, and 10% levels are −3.432, −2.862, and −2.567, respectively.

Following the ADF test, it was established that the dataset achieves stationarity after first differencing, setting d equal to 1 in this instance. The selection of the best ARIMA model is based on the AIC value; the smaller, the better. The ARIMAX (0,1,0) model, with its lowest AIC value of −63636.673 (shown in Table 7), emerges as the optimal choice among the ARIMAX models considered. The five dual important indicators are used to make regression analysis, and the regression equation is expressed as equation (7).

Table 7: ARIMAX model selection using AIC criteria

Model	**AIC**
ARIMAX (1,1,1)	−63632.625
ARIMAX (0,1,0)	−63634.673
ARIMAX (1,1,0)	−63633.863
ARIMAX (0,1,1)	−63634.636
ARIMAX (0,1,0)	−63636.673

$$\sum_{m=1}^{M} \beta_m a_t^m = -3\text{VWAP} + 4\text{WCP} - 1.096 \times 10^{-15}\text{FWMA} - 1.26 \times 10^{-15}\text{Decay} - 1.01 \times 10^{-15}\text{ZLMA}. \tag{7}$$

The residual diagnostics are presented in Figure 3 to evaluate the performance of ARIMAX model. The residuals in the normal Q–Q test consistently cluster around the reference line, and the correlogram reveals that most blue dots fall comfortably within the blue band. These findings underscore the ARIMAX model's exceptional capability to capture underlying patterns within the Vietnam stock price dataset and make accurate predictions.

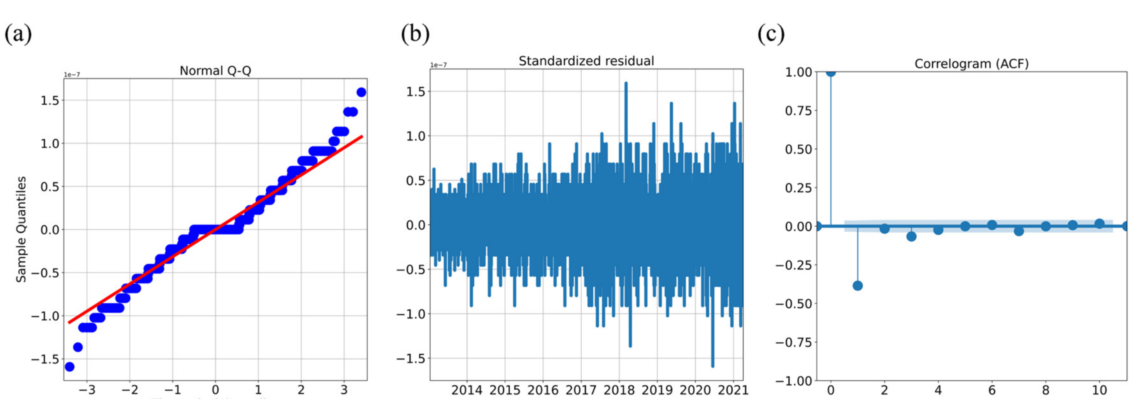

Figure 3: ARIMAX residual diagnostics plots: (a) Q-Q plot, (b) residuals over time, (c) Autocorrelation. Source: Created by the author.

In this study, apart from our proposed method, there are four other experiments conducted with LSTM, GALSTM, XGBoost, and Meta Prophet models to compare to our ARIMAX model. The assessment of output from each model has been quantified using six widely recognized statistical metrics: MAPE, MAE, NRMSD, and RMSE, SMAPE, and R^2 as shown in equations (8)–(13).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \check{y}_i)}, \tag{8}$$

$$\text{MAPE} = \frac{|y_m - \check{y}_m|}{y_m} \times 100, \tag{9}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \check{y}_i|, \tag{10}$$

$$\text{NRMSD} = \frac{1}{y_{\max} - y_{\min}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \check{y}_i)^2}, \tag{11}$$

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \check{y}_i|}{(|y_i| + |\check{y}_i|)/2}, \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \check{y}_i)^2}{\sum_{i=1}^{n}(\bar{\check{y}}_i - \check{y}_i)^2}, \tag{13}$$

where y_i and $\check{y}_i$ signify the actual value as well as the corresponding forecasted value of the variable for the i-th observation; y_m represents the observed series' mean value; $\check{y}_m$ is the predicted series' mean value; $y_{\max}$ is the maximum value in the observed series; $y_{\min}$ implies the minimum value as well as n indicates the total number of observations; and $\bar{\check{y}}_i$ is the average value. The diminished values of RMSE, MAPE, MAE, NRMSD, and SMAPE correlate with an increase in predictive accuracy. R^2 values exhibit a range from 0 to 1. A value that approaches 1 signifies an enhancement in performance.

In LSTM, the dataset containing the closing price and five dual important indicators is transformed on a scale of −1 and 1 by the MinmaxScaler technique. The MSE is used as the loss function. Random values are assigned as the network's initial weights, and the weights are adjusted by the *adam* optimizer. One-day window length and two hidden layers are applied to enhance efficiency. Moreover, an exhaustive search technique is applied to select the right batch size and number of epochs to enhance the performance of the LSTM model. To search for the optimum batch size, Kandel and Castelli [83] recommended starting with a smaller batch size, usually 32 or 64. In the XGBoost model, a random search technique was employed to explore the best values for parameters. Results of tuning parameters for LSTM and XGBoost models are shown in Tables 1 and 8, respectively. No data transformation is necessary when inputting data into XGBoost and Meta Prophet models. The parameter of interval_width is set at 0.95 for the Prophet model. The interval_width parameter in the Prophet model controls the width of the uncertainty intervals around the predicted values, providing a measure of the uncertainty associated with the predictions. Setting interval_width = 0.95 indicates that the uncertainty interval will cover 95% of the possible outcomes, leaving 5% of the outcomes outside the interval.

Table 8: Hyperparameters tuning for LSTM model

Parameters	Search space	Best value
Batch size	32, 64, 128	32
Epoch	100, 200, 300	200

The GALSTM model is a hybrid model of GA and LSTM. In terms of the GALSTM model, the best values for batch sizes and epochs obtained from the LSTM model will be directly applied to the GALSTM model without using an exhaustive search technique. The GA searches near optimal feature sets, and LSTM is trained using these feature sets. The parameters set for the GA process are illustrated in Table 9.

Table 9: Parameters of GA model

Parameter name	Parameter value	Parameter name	Parameter value
num_generations	5	num_genes	5
parent_selection_type	rws	mutation_type	Inversion
num_parents_mating	2	mutation_percent_genes	10
crossover_type	single_point	keep_elitism	2
sol_per_pop	20	gene_space	[0, 1]

All empirical results are shown in Table 10. The RMSE parameters of the five algorithms vary from 1.35×10^{-12} to 139.738, and the MAPE values fluctuate from 9.05×10^{-16} to 0.079. More importantly, the ARIMAX model's performance achieves the best accuracy in forecasting the stock price in Vietnam, followed by the GALSTM, Prophet, LSTM, and XGBoost models. The ARIMAX model achieves the best performance metrics compared to other models, because it achieves the smallest MAPE, RMSE, NRMSD, SMAPE, and MAE and the highest R^2. Compared to GALSTM, Prophet, LSTM, and XGBoost models, the proposed method utilizing ARIMAX with five dual important indicators demonstrates a significant enhancement in prediction accuracy, achieving an accuracy rate of 99.99%. The ARIMAX model achieves an R^2 value of 99.99%, indicating that the proposed method effectively explains 99.99% of stock price movements in the Vietnam market. This represents a notable improvement compared to GALSTM by 0.95%, Prophet model by 1.43%, LSTM model by 1.79%, and XGBoost model by 81.39%. Notably, in Figure 4(a), the predicted values generated by the ARIMAX model closely coincide with the actual stock prices in the testing segment of the dataset.

Table 10: Empirical results

	ARIMAX	GALSTM	Prophet	LSTM	XGBoost
MAPE	9.05×10^{-16}	0.010	0.014	0.014	0.079
SMAPE	9.05×10^{-14}	0.95	1.37	1.31	8.03
MAE	1.10×10^{-12}	12.164	17.812	18.319	96.006
RMSE	1.35×10^{-12}	16.200	19.116	22.304	139.738
NRMSD	2.19×10^{-15}	0.028	0.031	0.036	0.461
R^2	99.99%	99.04%	98.56%	98.20%	18.60%

The performance of the GALSTM model achieves the second. GAs demonstrate their capability to enhance the performance of the LSTM model. Following the GA process, two near-optimal features – WCP and Decay – are selected. This optimization reduces forecasting errors compared to the LSTM model only: a decrease of approximately 28.6% in MAPE, nearly 27.4% in RMSE, around 33.6% in MAE, 30.65% in SMAPE, 0.48% in R^2, and about 22.2% in NRMSD. The predicted stock price in the testing stage from the GALSTM model is pointed out in Figure 4(b). The prophet model is third. The MAPE values of Prophet and LSTM are similar to each other, yet there are discrepancies among MAE, RMSE, SMAPE, R^2, and NRMSD values. The Prophet model performs better than the LSTM model in predicting Vietnamese stock prices using five dual important indicators. The Prophet model is normally used for univariate time series data. In this study, we base regression foundation to add more variables into the main model. The relationships between the closing price and its variables are linear, as shown in Section 3. That may be the reason why the Prophet model can achieve better than LSTM. The

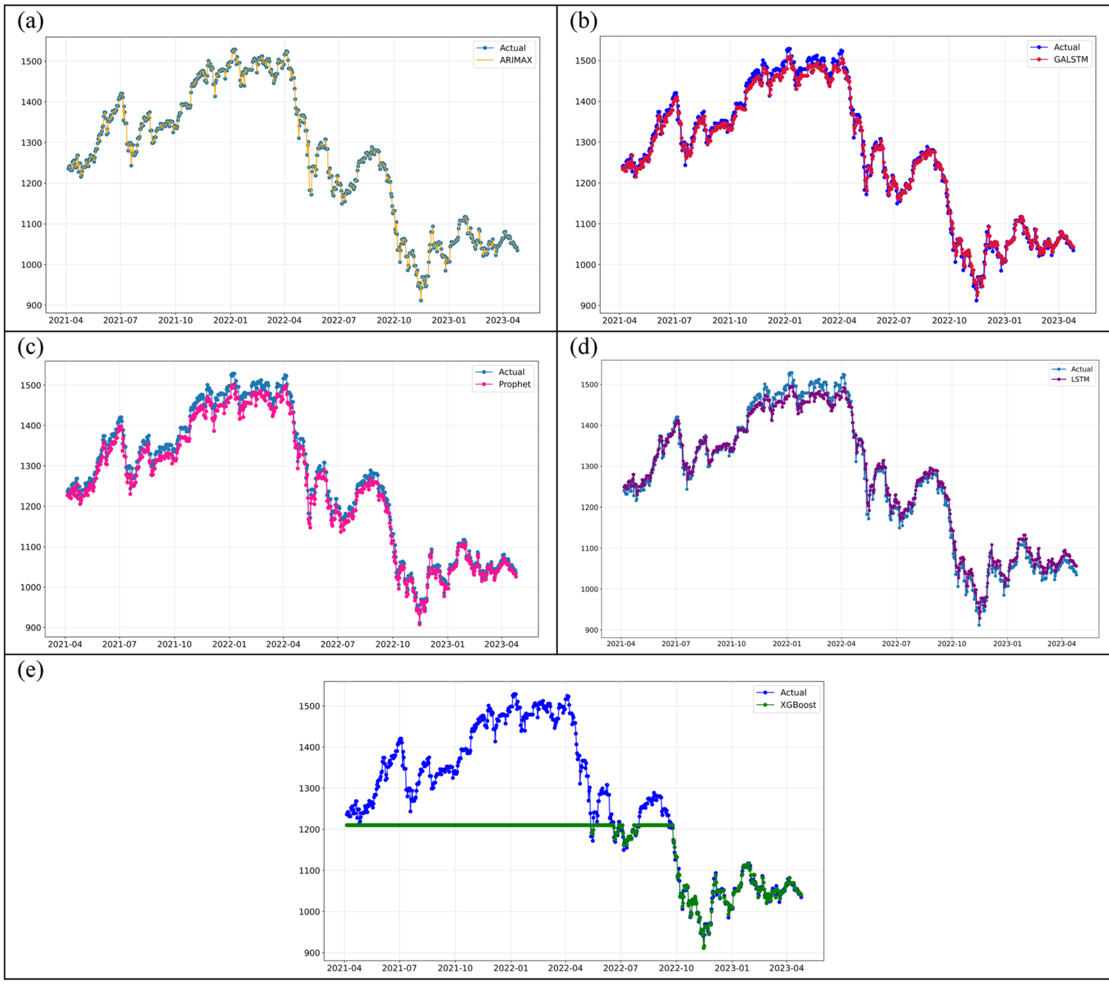

Figure 4: Predictive and actual values in the testing stage for: (a) ARIMAX model, (b) GALSTM model, (c) Meta Prophet model, (d) LSTM model, (e) XGBoost model. Source: Created by the author.

predicted stock price in the testing stage using the Prophet model is shown in Figure 4(c). Figure 4(d) shows that the LSTM model struggles between October 2021 and April 2022; however, the LSTM captures and follows the trends presented in the data during that period very well. The result of the XGBoost model is shown in Figure 4(e). It does not perform well right from the start of the testing phase and continues to underperform until almost October 2022. This hampers its ability to make accurate predictions.

To visually compare the performance of all models, Figure 5 is presented to illustrate the plots of residuals for predicting the testing phase across all models. A notable observation is that the residuals of the ARIMAX model appear to be quite small, with deviations that are generally very close to the zero line. This pattern suggests that the predictions generated by the ARIMAX model exhibit a high degree of accuracy, as the residuals – the differences between the actual and predicted values – are minimal. The residuals from the GALSTM model are spread around both sides of the zero line, indicating a balanced distribution of over-estimations and underestimations. In contrast, the residuals from the Prophet model tend to lie predominantly above the zero line, suggesting a tendency towards overestimation. In the case of the LSTM model, its residuals exhibit a distribution that is significantly displaced from the zero line. The residuals of the XGBoost model appear to be the farthest away from the zero line, indicating the largest deviations between predicted and actual values. This observation suggests that the XGBoost model yields the least accurate predictions among the five models evaluated.

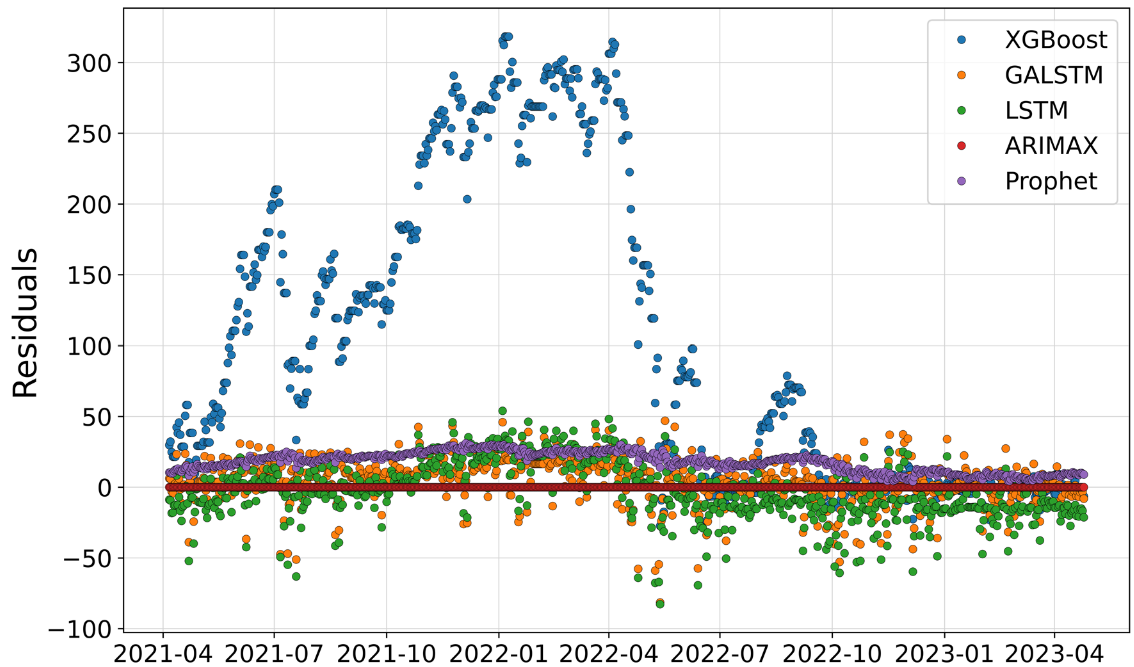

Figure 5: Residuals plot in testing part. Source: Created by the author.

To validate the prediction accuracy of the ARIMAX model paired with five dual important indicators in this study, we performed two more experiments using the dataset from Pham and Ta [36] from 2010 to 2021 and the dataset from Do and Trang [37] from January 3, 2001, to August 30, 2019. VNINDEX is the only stock price used in these two studies. Referring to the literature review, both studies employed distinct models to predict the univariate time series of VNINDEX stock prices. The results of two additional experiments are detailed in Table 11. Notably, the ARIMAX model in conjunction with five dual important indicators consistently achieved greater accuracy in both datasets. This method helps to improve the accuracy of the previous results by up to 99.9%.

Table 11: Two more experiment results

Experiment on the first dataset from Pham and Ta [36]			
Performance metrics	**The original findings**		**ARIMAX + 5 dual important indicators**
	LSTM	**HANFIS**	
MAPE	0.009	0.010	4.925×10^{-15}
RMSE	9.957	12.569	5.373×10^{-12}
MAE	7.081	7.953	4.512×10^{-12}

Experiment on the second dataset from Do and Trang [37]				
Performance metrics	**The original findings**			**ARIMAX + 5 dual important indicators**
	LSTM	**ARIMA**	**SARIMA**	
RMSE	0.0113	0.0115	0.0119	2.729×10^{-5}
MAE	0.0079	0.0083	0.0086	1.957×10^{-5}

5 Conclusion and discussion

This study proposes dual important indicators, which are technical indicators that are important to predict stock prices in the past and the future of the VNINDEX. The VNINDEX dataset was sourced from the investing.com website from January 2, 2013, to April 25, 2023. Five dual important indicators – VMAP, WCP, FWMA, exponential decay (Decay), and ZLMA, are found. In conjunction with the ARIMAX model, the empirical result achieves an R^2 value of 99.99%, which is significantly better than popular models like LSTM, GALSTM, XGBoost, and Meta Prophet. To validate our finding, two more experiments using the dataset from Pham and Ta [36] from 2010 to 2021 and the dataset from Do and Trang [37] from January 3, 2001, to August 30, 2019, are applied to evaluate the prediction accuracy of the ARIMAX model with five dual important indicators. The results show that the ARIMAX model with five dual important indicators significantly enhanced the predictive precision in different periods of time within the Vietnam stock market context. This finding suggests that these indicators play a crucial role in refining the ARIMAX model's accuracy and its ability to forecast stock prices more effectively. By leveraging this distinct set of indicators, the ARIMAX model becomes more adept at capturing and understanding the intricate patterns and behaviors prevalent within the Vietnamese stock market.

The proposed model in this study serves as an effective tool for financial investors, brokers, policymakers, and other stakeholders, providing them with a reliable means to evaluate future stock values and make informed decisions. With its high accuracy in predicting stock price movements, this model offers valuable insights that can aid in strategic decision-making processes within the financial market. For example, the individual can use predictions to determine the entry and exit points to maximize returns or minimize losses, and financial advisors can provide better guidance to their clients to apply a suitable strategy, such as hedging strategies or diversified portfolio strategy to protect their assets against adverse market movements, policymakers can understand potential economic trends and market conditions, aiding in the formulation of economic policies that support market stability and growth, regulators can monitor market trends to identify periods of potential volatility or instability, allowing them to take preemptive measures to protect investors and maintain market integrity. The findings from this study achieve better results than some recent studies. For instance, the study from Lu et al. [84] created a new method that combines CNN, BiLSTM, and attention mechanism to predict the closing price of the Shanghai Composite Index for the next day over 1,000 trading days, and the result gained highest R^2 value of 0.9804, MAE value of 21.952, and RMSE value of 31.694. The study of Yu and Yan [85] proposed a deep neural networks-based prediction model designed based on the time series phase-space reconstruction method to predict stock indices of the S&P 500, the Dow Jones industrial average (DJIA). The proposed method yields the best results, on average, for the S&P 500 with an RMSE value of 7.92 and an MAPE value of 1.3. For the DJIA index, the method achieves an RMSE value of 5.88 and a MAPE value of 0.57. In the case of the N225 index, the RMSE value is 5.6, and the MAPE value is 0.83. For the HSI index, the RMSE value is 55.25, while the MAPE value is 0.48. The CSI index exhibits an RMSE value of 5.92 and a MAPE value of 0.93. Lastly, for the ChiNext index, the RMSE value is 4.15, and the MAPE value is 1.28. Therefore, this study contributes by proposing a new way to select relevant indicators to perform the prediction, and it achieves better results than current commonly used methods.

The key advantage of our method is its feature selection process powered by the XGBoost model. This process extracts important indicators that are significant both in the past and future. These features encapsulate crucial information regarding the relationship between input variables and the target variable. By selecting the most important features, a substantial amount of significant information from the original dataset can be effectively summarized. Moreover, ARIMAX stands as a classic statistical model specifically tailored to handle time series data. Hence, the combination of both types of models enables the exploitation of their respective strengths. However, this method also carries the risk of potential bias, as it relies on the manual selection of dual important indicators from XGBoost *lag*-0 and *lag*-1 models rather than depending on specific techniques. While this research achieves better accuracy using XGBoost to select dual important indicators, future studies need to check out different methods. We need to compare them and figure out which one makes the models perform even better. Therefore, the future study can employ some advanced techniques for dual import indicators selection which can enhance results by effectively screening and narrowing down the number of indicators. This study focuses on the Vietnamese stock market. To extend the generalizability of our findings, other markets can be applied to test the stability of dual important indicators. Macroscopic factors, fundamental indicators, or alternative feature-selecting methods can be adopted to improve the accuracy of the model. By doing so, we can develop a more comprehensive understanding of how these indicators influence different markets, thus enhancing the broader applicability of our methodology. In the near future, it would be beneficial to carry out additional experiments in different markets using our proposed method. This will help determine if the ARIMAX model, along with these dual important indicators, can deliver exceptional results in other markets as well.

Every stock market possesses unique characteristics. This study serves as a contribution to understanding the specific nuances of the Vietnam market. It aims to identify five dual important indicators – VMAP, WCP, FWMA, Exponential Decay (Decay), and ZLMA, for observing stock prices in Vietnam. Subsequently, a robust model will be employed to predict stock prices, leveraging the insights gained from this exploration. Based on our findings, practitioners can make more informed decisions regarding investment strategies, risk management, and trading activities. Researchers can use our method to test and validate hypotheses about market behavior and the effectiveness of various trading strategies. Moreover, accurate predictions and advanced analytical tools can confer a competitive advantage in the financial markets. This allows practitioners to stay ahead of the competition and adapt to changing market conditions more effectively.

Acknowledgments: The authors appreciate the support from the Southern Taiwan University of Science and Technology, Taiwan.

Funding information: The authors received no specific funding for this article.

Author contributions: Pai-Chou Wang: writing – review and editing, software, methodology, formal analysis. Tram Thi Hoai Vo: writing – original draft, review and editing, software, methodology, formal analysis, data curation, conceptualization.

Conflict of interest: The authors have declared that no conflict of interest exists.

Data availability statement: The Vietnamese stock prices are available and can be downloaded from https://www.investing.com/indices/vn. The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Fama EF. Stock returns, real activity, inflation, and money. Am Econ Rev. 1981;71(4):545–65.

[2] Polkuamdee N. Stock investors urged to monitor Vietnam [Internet]. Thailand: Bangkok Post; 2023. [cited 2024 Jan 04] Available from https://www.bangkokpost.com/business/general/2674169/stock-investors-urged-to-monitor-vietnam.

[3] Mondal P, Shit L, Goswami S. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. Int J Comput Sci Eng Appl. 2014;4(2):13. doi: 10.5121/ijcsea.2014.4202.

[4] Babu CN, Reddy BE. Prediction of selected Indian stock using a partitioning–interpolation based ARIMA–GARCH model. Appl Comput Inf. 2015;11(2):130–43. doi: 10.1016/j.aci.2014.09.002.

[5] Naik N, Mohan BR, Jha RA. GARCH model identification for stock crises events. Procedia Comput Sci. 2020;171:1742–9. doi: 10.1016/j.procs.2020.04.187.

[6] Devi BU, Sundar D, Alli P. An effective time series analysis for stock trend prediction using ARIMA model for nifty midcap-50. Int J Data Min Knowl Manag Process. 2013;3(1):65. doi: 10.5121/ijdkp.2013.3106.

[7] Ince H, Trafalis TB. Short term forecasting with support vector machines and application to stock price prediction. Int J Gen Syst. 2008;37(6):677–87. doi: 10.1080/03081070601068595.

[8] Pawar K, Jalem RS, Tiwari V. Stock market price prediction using LSTM RNN. Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS. 2019, Springer; 2018. doi: 10.1007/978-981-13-2285-3_58.

[9] Nelson DM, Pereira AC, De Oliveira RA. Stock market's price movement prediction with LSTM neural networks. 2017 International joint conference on neural networks (IJCNN). IEEE; 2017. doi: 10.1109/IJCNN.2017.7966019.
[10] Huang W, Nakamori Y, Wang S-Y. Forecasting stock market movement direction with support vector machine. Comput Oper Res. 2005;32(10):2513–22. doi: 10.1016/j.cor.2004.03.016.
[11] Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Syst Appl. 2021;178:115019. doi: 10.1016/j.eswa.2021.115019.
[12] Tian L, Feng L, Yang L, Guo Y. Stock price prediction based on LSTM and LightGBM hybrid model. J Supercomput. 2022;78(9):11768–93. doi: 10.1007/s11227-022-04326-5.
[13] Long W, Lu Z, Cui L. Deep learning-based feature engineering for stock price movement prediction. Knowl Syst. 2019;164:163–73. doi: 10.1016/j.knosys.2018.10.034.
[14] Morck R, Shleifer A, Vishny RW, Shapiro M, Poterba JM. The stock market and investment: is the market a sideshow? Brook Pap Econ Act. 1990;1990(2):157–215. doi: 10.2307/2534506.
[15] Lo AW, Hasanhodzic J. The heretics of finance: Conversations with leading practitioners of technical analysis. Canada: John Wiley and Sons; 2010.
[16] Covel M. Trend following: how great traders make millions in up or down markets. New expanded ed. New Jersey: Financial Times Prentice Hall; 2006.
[17] Fu T-C, Chung C-P, Chung F-L. Adopting genetic algorithms for technical analysis and portfolio management. Comput Math Appl. 2013;66(10):1743–57. doi: 10.1016/j.camwa.2013.08.012.
[18] Dai Z, Zhu H, Kang J. New technical indicators and stock returns predictability. Int Rev Econ Finance. 2021;71:127–42. doi: 10.1016/j.iref.2020.09.006.
[19] Ma C, Yan S. Deep learning in the Chinese stock market: The role of technical indicators. Finance Res Lett. 2022;49:103025. doi: 10.1016/j.frl.2022.103025.
[20] Neely CJ, Rapach DE, Tu J, Zhou G. Forecasting the equity risk premium: the role of technical indicators. Manage Sci. 2014;60(7):1772–91. doi: 10.1287/mnsc.2013.1838.
[21] Chandar SK. Convolutional neural network for stock trading using technical indicators. Autom Softw Eng. 2022;29:1–14. doi: 10.1007/s10515-021-00303-z.
[22] Lee M-C, Chang J-W, Yeh S-C, Chia T-L, Liao J-S, Chen X-M. Applying attention-based BiLSTM and technical indicators in the design and performance analysis of stock trading strategies. Neural Comput Appl. 2022;34(16):13267–79. doi: 10.1007/s00521-021-06828-4.
[23] Chourmouziadis K, Chourmouziadou DK, Chatzoglou PD. Embedding four medium-term technical indicators to an intelligent stock trading fuzzy system for predicting: a portfolio management approach. Comput Econ. 2021;57:1183–216. doi: 10.1007/s10614-020-10016-2.
[24] Sahoo BB, Jha R, Singh A, Kumar D. Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophys. 2019;67(5):1471–81. doi: 10.1007/s11600-019-00330-1.
[25] Al Ali A, Khedr AM, El Bannany M, Kanakkayil S. GALSTM-FDP: A Time-series modeling approach using hybrid GA and LSTM for financial distress prediction. Int J Financ Stud. 2023;11(1):38. doi: 10.3390/ijfs11010038.
[26] Jin J, Wang B, Zhang Y, Yu M, Zheng X, eds. Short-Term power load forecasting based on HFEMD and GALSTM. 2021 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia). Chengdu, China: IEEE; 2021. doi: 10.1109/ICPSAsia52756.2021.9621579.
[27] Wang Y, Guo Y. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. China Commun. 2020;17(3):205–21. doi: 10.23919/JCC.2020.03.017.
[28] Cheng J, Tiwari S, Khaled D, Mahendru M, Shahzad U. Forecasting Bitcoin prices using artificial intelligence: Combination of ML, SARIMA, and Facebook Prophet models. Technol Forecast Soc Change. 2024;198:122938. doi: 10.1016/j.techfore.2023.122938.
[29] Shilpa G, Sheshadri G. ARIMAX model for short-term electrical load forecasting. Int J Recent Technol Eng. 2019;8(4):2786–90. doi: 10.35940/ijrte.D7950.118419.
[30] Anggraeni W, Andri KB, Mahananto F. The performance of ARIMAX model and Vector Autoregressive (VAR) model in forecasting strategic commodity price in Indonesia. Procedia Comput Sci. 2017;124:189–96. doi: 10.1016/j.procs.2017.12.146.
[31] Sutthichaimethee P, Ariyasajjakorn D. Forecasting energy consumption in short-term and long-term period by using arimax model in the construction and materials sector in Thailand. J Ecol Eng. 2017;18(4):52–9. doi: 10.12911/22998993/74396.
[32] Chiu L-Y, Rustia DJA, Lu C-Y, Lin T-T. Modelling and forecasting of greenhouse whitefly incidence using time-series and ARIMAX analysis. IFAC-PapersOnLine. 2019;52(30):196–201. doi: 10.1016/j.ifacol.2019.12.521.
[33] Tamuke E, Jackson EA, Sillah A. Forecasting inflation in Sierra Leone using ARIMA and ARIMAX: A comparative evaluation. Model building and analysis team. Theor Pract Res Econ Fields. 2018;9(1):63–74. doi: 10.14505/tpref.v9.1(17).07.
[34] Ugoh CI, Uzuke CA, Ugoh DO. Application of ARIMAX Model on forecasting Nigeria's GDP. Am J Theor Appl Stat. 2021;10(5):216. doi: 10.11648/j.ajtas.20211005.12.
[35] Suharsono A, Suhartono S, Masyitha A, Anuravega A. Time series regression and ARIMAX for forecasting currency flow at Bank Indonesia in Sulawesi region. AIP Conference Proceedings. AIP Publishing; 2015. doi: 10.1063/1.4937107.
[36] Pham PM, Ta QC. Stacked Long Short-Term Memory for Vietnamese Stock Market Returns Prediction; Int J Appl Sci Curr Future Res Trends. 2022;13(1):193–208.
[37] Do Q, Trang T. Forecasting Vietnamese stock index: A comparison of hierarchical ANFIS and LSTM. Decis Sci Lett. 2020;9(2):193–206. doi: 10.5267/j.dsl.2019.11.002.

[38] Co NT, Son HH, Hoang NT, Lien TTP, Ngoc TM. Comparison between ARIMA and LSTM-RNN for VN-index prediction. Intelligent Human Systems Integration 2020. Proceedings of the 3rd International Conference on Intelligent Human Systems Integration (IHSI 2020): Integrating People and Intelligent Systems. Modena, Italy: Springer; 2020. doi: 10.1007/978-3-030-39512-4_168.
[39] Ngan TMU. Forecasting foreign exchange rate by using ARIMA model: A case of VND/USD exchange rate. Methodology. 2013;2014:2015.
[40] Hằng LTT, Dũng NX. ARIMA model–vietnam's GDP forecasting. In: Ngoc Thach N, Ha DT, Trung ND, Kreinovich V, editors. Prediction and Causality in Econometrics and Related Topics. Cham: Springer International Publishing; 2022. p. 145–51. doi: 10.1007/978-3-030-77094-5_14.
[41] Lee J-Y, Nguyen T-T, Nguyen H-G, Lee J-Y. Towards predictive crude oil purchase: a case study in the USA and Europe. Energies. 2022;15(11):4003. doi: 10.3390/en15114003.
[42] Wang L, Zou H, Su J, Li L, Chaudhry S. An ARIMA-ANN hybrid model for time series forecasting. Syst Res Behav Sci. 2013;30(3):244–59. doi: 10.1002/sres.2179.
[43] Bao TQ, My BTT. Forecasting stock index based on hybrid artificial neural network models. VNUHCM J Econ Bus Law. 2019;3(1):52–7. doi: 10.32508/stdjelm.v3i1.540.
[44] Le TTH, Nguyen TN. A hybrid model based on ARIMA and artificial neural network to forecast consumer price index: the case of Vietnam. In: Ngoc Thach N, Kreinovich V, Ha DT, Trung ND, editors. Optimal Transport Statistics for Economics and Related Topics. Cham: Springer Nature Switzerland; 2024. p. 449–64. doi: 10.1007/978-3-031-35763-3_32.
[45] Thu HGT, Thanh TN, Le Quy T. A neighborhood deep neural network model using sliding window for stock price prediction. 2021 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE; 2021. doi: 10.1109/BigComp51126.2021.00022.
[46] Yıldırım DC, Toroslu IH, Fiore U. Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators. Financ Innov. 2021;7:1–36. doi: 10.1186/s40854-020-00220-2.
[47] Aminimehr A, Raoofi A, Aminimehr A, Aminimehr A. The role of feature engineering in prediction of tehran stock exchange index based on LSTM. Iran. J Econ Stud. 2020;9(2):527–48. doi: 10.22099/ijes.2021.39877.1739.
[48] Zhuge Q, Xu L, Zhang G. LSTM neural network with emotional analysis for prediction of stock price. Eng Lett. 2017;25:2.
[49] Yan B, Aasma M. A novel deep learning framework: Prediction and analysis of financial time series using CEEMD and LSTM. Expert Syst Appl. 2020;159:113609. doi: 10.1016/j.eswa.2020.113609.
[50] Reddy DS, Prasad PRC. Prediction of vegetation dynamics using NDVI time series data and LSTM. Model Earth Syst Environ. 2018;4:409–19. doi: 10.1007/s40808-018-0431-3.
[51] Karevan Z, Suykens JA. Transductive LSTM for time-series prediction: An application to weather forecasting. Neural Network. 2020;125:1–9. doi: 10.1016/j.neunet.2019.12.030.
[52] Minh DL, Sadeghi-Niaraki A, Huy HD, Min K, Moon H. Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. IEEE Access. 2018;6:55392–404. doi: 10.1109/ACCESS.2018.2868970.
[53] Huynh HD, Dang LM, Duong D. A new model for stock price movements prediction using deep neural network. Proceedings of the 8th International Symposium on Information and Communication Technology; 2017. doi: 10.1145/3155133.3155202.
[54] Thanh HT, Meesad P. Stock market trend prediction based on text mining of corporate web and time series data. J Adv Comput Intell Intell Informat. 2014;18(1):22–31. doi: 10.20965/jaciii.2014.p0022.
[55] Monika MK, Kumar M. eds. XGBoost: 2D-object recognition using shape descriptors and extreme gradient boosting classifier. Computational Methods and Data Engineering: Proceedings of ICMDE 2020. Vol. 1, Springer; 2021, doi: 10.1007/978-981-15-6876-3_16.
[56] Chen Z, Fan W. A freeway travel time prediction method based on an XGBoost model. Sustainability. 2021;13(15):8577. doi: 10.3390/su13158577.
[57] Jeyakarthic M, Punitha S, Punitha S. Hybridization of Bat algorithm with XGBOOST model for precise prediction of stock market directions. Int J Eng Adv Technol. 2020;9(3):3375–82. doi: 10.35940/ijeat.C5535.029320.
[58] Budholiya K, Shrivastava SK, Sharma V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. J King Saud Univ-Comput Inf Sci. 2022;34(7):4514–23. doi: 10.1016/j.jksuci.2020.10.013.
[59] Bouktif S, Fiaz A, Ouni A, Serhani MA. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. Energies. 2018;11(7):1636. doi: 10.3390/en11071636.
[60] Stajkowski S, Kumar D, Samui P, Bonakdari H, Gharabaghi B. Genetic-algorithm-optimized sequential model for water temperature prediction. Sustainability. 2020;12(13):5374. doi: 10.3390/su12135374.
[61] Chen S, Zhou C. Stock prediction based on genetic algorithm feature selection and long short-term memory neural network. IEEE Access. 2020;9:9066–72. doi: 10.1109/ACCESS.2020.3047109.
[62] He Y, Zeng X, Li H, Wei W. Application of LSTM model optimized by individual-ordering-based adaptive genetic algorithm in stock forecasting. Int J Intell Comput Cybern. 2023;16(2):277–94. doi: 10.1108/IJICC-04-2022-0104.
[63] Sah S, Surendiran B, Dhanalakshmi R, Mohanty SN, Alenezi F, Polat K. Forecasting COVID-19 pandemic using Prophet, ARIMA, and hybrid stacked LSTM-GRU models in India. Comput Math Methods Med. 2022;2022(1):1556025. doi: 10.1155/2022/1556025.
[64] Satrio CBA, Darmawan W, Nadia BU, Hanafiah N. Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. Procedia Comput Sci. 2021;179:524–32. doi: 10.1016/j.procs.2021.01.036.
[65] Xie C, Wen H, Yang W, Cai J, Zhang P, Wu R, et al. Trend analysis and forecast of daily reported incidence of hand, foot and mouth disease in Hubei, China by Prophet model. Sci Rep. 2021;11(1):1445. doi: 10.1038/s41598-021-81100-2.
[66] Weytjens H, Lohmann E, Kleinsteuber M. Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet. Electron Commer Res. 2021;21(2):371–91. doi: 10.1007/s10660-019-09362-7.

[67] Yenidoğan I, Çayir A, Kozan O, Dağ T, Arslan Ç, ed. Bitcoin forecasting using ARIMA and PROPHET. 2018 3rd International Conference on Computer Science and Engineering (UBMK). IEEE; 2018. doi: 10.1109/UBMK.2018.8566476.
[68] Hemdan EE-D, Essa YM, Shouman M, El-Sayed A, Moustafa AN. An efficient IoT based smart water quality monitoring system. Multimed Tools Appl. 2023;82(19):28827–51. doi: 10.1007/s11042-023-14504-z.
[69] Vargas MR, Dos Anjos CE, Bichara GL, Evsukoff AG. Deep leaming for stock market prediction using technical indicators and financial news articles. 2018 International Joint Conference on Neural Networks (IJCNN). IEEE; 2018. doi: 10.1109/IJCNN.2018.8489208.
[70] Haq AU, Zeb A, Lei Z, Zhang D. Forecasting daily stock trend using multi-filter feature selection and deep learning. Expert Syst Appl. 2021;168:114444. doi: 10.1016/j.eswa.2020.114444.
[71] Yuan X, Yuan J, Jiang T, Ain QU. Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market. IEEE Access. 2020;8:22672–85. doi: 10.1109/ACCESS.2020.2969293.
[72] Mitchell D, Białkowski J, Tompaidis S. Volume-weighted average price tracking: A theoretical and empirical study. IISE Trans. 2020;52(8):864–89. doi: 10.1080/24725854.2019.1688896.
[73] Sarkar A. Weighted Close Indicator Strategy with Examples [Internet]. India: Stock Maniacs; 2023. [cited 2024 Jan 04] https://www.stockmaniacs.net/weighted-close-indicator/.
[74] Chen T-L, Cheng C-H, Teoh HJ. Fuzzy time-series based on Fibonacci sequence for stock price forecasting. Phys A. 2007;380:377–90. doi: 10.1016/j.physa.2007.02.084.
[75] Pedersen KI, Mogensen PE, Fleury BH. A stochastic model of the temporal and azimuthal dispersion seen at the base station in outdoor propagation environments. IEEE Trans Veh Technol. 2000;49(2):437–47. doi: 10.1109/25.832975.
[76] Zakamulin V, Zakamulin V. Anatomy of trading rules. market timing with moving averages: the anatomy and performance of trading rules. Cham: Springer International Publishing; 2017. p. 71–101. doi: 10.1007/978-3-319-60970-6_5.
[77] Konety K. Real-Time Stock Market Recommendation & Prediction using Multi Source Data [Dissertation]. Ireland: Technological University Dublin; 2022.
[78] Kadri F, Harrou F, Chaabane S, Tahon C. Time series modelling and forecasting of emergency department overcrowding. J Med Syst. 2014;38:1–20. doi: 10.1007/s10916-014-0107-0.
[79] Sharma A, Tiwari P, Gupta A, Garg P, Use of LSTM and ARIMAX algorithms to analyze impact of sentiment analysis in stock market prediction. Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI. Vol. 2021, Springer; 2020. doi: 10.1007/978-981-15-9509-7_32.
[80] Abd HT, Essa AK, Jassim FM. Analyzing the relationship between the dow jones index and oil prices using the ARIMAX model. Int J Adv Sci Eng Inf Technol. 2021;11(2):465–73. doi: 10.18517/ijaseit.11.2.14080.
[81] Serafini G, Yi P, Zhang Q, Brambilla M, Wang J, Hu Y, et al. Sentiment-driven price prediction of the bitcoin based on statistical and deep learning approaches. 2020 International Joint Conference on Neural Networks (IJCNN). IEEE; 2020. doi: 10.1109/IJCNN48605.2020.9206704.
[82] Patil R. Time series analysis and stock price forecasting using machine learning techniques. [MSc Research Project report]. Ireland: National College of Ireland; 2021. doi: 10.1994/Rajat/AI.
[83] Kandel I, Castelli M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. ICT Express. 2020;6(4):312–5. doi: 10.1016/j.icte.2020.04.010.
[84] Lu W, Li J, Wang J, Qin L. A CNN-BiLSTM-AM method for stock price prediction. Neural Comput Appl. 2021;33(10):4741–53. doi: 10.1007/s00521-020-05532-z.
[85] Yu P, Yan X. Stock price prediction based on deep neural networks. Neural Comput Appl. 2020;32(6):1609–28. doi: 10.1007/s00521-019-04212-x.