Predicting Stock Market Trends in Vietnam Using SVM, XGBoost, and LSTM Architectures

Truong Dinh Tu[1], Bhagawan Nath[2]

[1]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam,
[2]Department of Information Technology, Greenwich University, Hanoi, Vietnam.

Correspondence: Truong Dinh Tu, truongdinhtu@tdtu.edu.vn
Communication: received 9 Oct 2024, revised 12 Dec 2024, accepted 16 Jan 2025
DOI: 10.32913/mic-ict-research.v2025.n1.1356

Abstract: **Predicting the behavior of a stock market with higher accuracy has been a critical task, however, application of various cutting-edge Machine Learning (ML) and Deep Learning (DL) architectures have shown better results compared to that of classical statistical models. This study explores the Vietnamese stock market with the aim of predicting closing price movements. Various machine learning models, including Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks, are employed for the analysis. The data for this research has been gathered from various companies listed on the Ho Chi Minh City Stock Exchange (HOSE) and Hanoi Stock Exchange (HNINDEX) in Vietnam. These datasets were systematically preprocessed and thoroughly analyzed to calculate performance metrics, including the Root Mean Squared Error (RMSE) for assessing prediction quality and Mean Absolute Percentage Error (MAPE) to gauge forecast accuracy. In this study, technical features such as Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Stochastic Oscillator (SO), among others, are employed. Additionally, Simple Moving Averages (SMA) and Weighted Moving Averages (WMA) are considered for various time frames. The significance of these features is systematically evaluated and determined. The experimental results indicate that the LSTM model outperforms both SVM and XGBoost in terms of prediction accuracy, as evidenced by its lower RMSE and MAPE values, as well as higher accuracy and F1-scores. However, the computational time of LSTM is significantly greater compared to that of SVM and XGBoost.**

Keywords: *Support vector machines, extreme gradient boosting, long short-term memory, relative strength index, stochastic oscillator, moving average convergence divergence.*

I. Introduction

Vietnam is classified as a lower middle-income country, characterized by a multi-sectoral market economy where the state sector plays a central role in guiding economic development, ultimately with the overarching aim of advancing socialism. As of 2022, Vietnam ranks as the 38th-largest economy globally in terms of nominal Gross Domestic Product (GDP) and the 26th-largest when measured by Purchasing Power Parity (PPP) [1]. The VNINDEX serves as a capitalization-weighted index encompassing all companies listed on both the HOSE and HNINDEX. Established on July 28, 2000, with a base index value of 100, this index provides a comprehensive overview of the stock market's performance. Presently, the Ho Chi Minh City Stock Exchange boasts 404 listed companies [2], while the Hanoi Stock Exchange has 341 listed companies [3].

The stock market functions as a platform where traders and investors engage in buying and selling shares of publicly traded companies. It holds a pivotal role in shaping the economic landscape of a country. The prediction of stock market behavior involves forecasting the future values of stocks of various companies and other financial instruments traded on an exchange. The primary goal of such predictions is to mitigate risks and maximize profits [4]. A stock market is mostly characterized by a combination of nonlinear behaviors such as demand-supply, fundamental components, government policies, political situations, inflation, economy etc., which are highly volatile and complex. From economic, political and investment decisions to unknown causes, in some respects, there are many difficulties in predicting the stock market [5]. Thus, although stock markets attract a lot of investors due to its high profitability, it also brings along a comparable size of risk factors.

Stock markets are known for their high volatility and unpredictability, making traditional statistical methods like Simple Moving Average, Weighted Moving Average, and Exponential Smoothing only moderately effective. The inherent linearity of these statistical approaches can result in suboptimal prediction performances, particularly when faced with abrupt fluctuations in stock prices, whether they experience a sudden rise or fall. Thus, sophisticated state-of-the-art approaches are required to deal with it more robustly [6, 7]. Recent advancement and application of technologies such as ML, DL, big data, cloud technologies etc. has helped significantly in developing higher performing algorithms in various sectors and it has largely drawn the attention of industries and academia as well [8]. Various ML and DL based algorithms such as Decision Trees, Logistic Regression, Support Vector Machines, Extended Gradient Boosting etc. have been applied in predicting stock market behaviors in research [9]. However, the Long Short-Term Memory (LSTM) network algorithms have been predominantly used to classify sequential data because of their ability to learn long-term dependencies between time steps of data and higher accuracy rates. The LSTM network, classified as a type of realizable recurrent neural network model, is equipped with selective memory and intra-temporal influence. This particular architecture proves highly suitable for analyzing stochastic non-stationary time series data, such as stock prices [10].

This paper uses LSTM, SVM and XGBoost network architectures to predict stock market behavior, especially the closing price prediction and this study is specific to Vietnam stock market. Datasets are collected from different sources, five company datasets (VRE, SSI, VCB, STB, and VNM) are extracted from the VN30 dataset. Three more datasets are collected from the Ho Chi Minh Stock Exchange (HOSE), HNINDEX and a combination of both, known as Vietnam Stock Exchange (VNINDEX) respectively. The time span of the datasets is mostly from either from 2006 or 2009, and until 2021. The datasets are preprocessed and certain technical features such as MACD, RSI, SO etc. are computed for a better understanding of the problem and to check their impacts in the analysis. Two moving averages (SMA and EMA) are calculated for different time intervals such as weekly, monthly, quarterly etc.

Finally, analysis is done using the algorithms to determine forecasting of the stock market. Several performance metrics, including MAPE, RMSE, accuracy score, F1-score, and computational time, are employed to assess the efficiency of the models. MAPE, widely utilized for assessing forecast accuracy, falls under percentage errors, making it scale-independent and suitable for comparing series on different scales.

The second metric considered is RMSE, a commonly used measure for evaluating prediction quality. It quantifies how far predictions deviate from measured true values, utilizing Euclidean distance. The accuracy score measures the models' performance by calculating the ratio of the sum of True Positives (TP) and True Negatives (TN) to all predictions made. F1-score, another crucial metric, gauges a model's accuracy by combining precision and recall scores, reflecting the frequency of correct predictions across an entire dataset. Accuracy is preferred when True Positives (TP) and True Negatives (TN) are of greater importance, whereas F1-score is prioritized when False Negatives (FN) and False Positives (FP) carry more significance.

The last metric, computational time, is used to determine the swiftness of the algorithms, which is equally important looking at the current scenario. Another crucial outcome of this paper is the determination of the importance of the technical features, that is done using the XGBoost architecture and presented in the paper.

The contributions of this study are:

- The study applies advanced machine learning (SVM, XGBoost) and deep learning (LSTM) architectures to predict stock market behavior in Vietnam, focusing on closing price prediction across multiple datasets, including company-specific and stock exchange datasets (HOSE, HNINDEX, VNINDEX).
- The study introduces a comprehensive analysis of technical indicators (e.g., MACD, RSI, SMA, EMA) and evaluates their impact on stock market forecasting for improved prediction accuracy.
- Experimental results demonstrate that LSTM outperforms SVM and XGBoost in prediction accuracy (e.g., RMSE, MAPE, F1-score), although it requires significantly higher computational time.
- The study focuses on the Vietnamese stock market, offering insights into an emerging market that has been underexplored in financial forecasting research. It encourages future research to develop more effective prediction models, offering valuable insights to investors, analysts, and other stakeholders in stock market forecasting.

The rest of the paper is organized as follows: Section II briefly describes the technical background, Section III discusses the related work done in the same domain, Section IV explains the methodology, Section V explores the results, and Section VI includes the conclusion and future work.

II. Technical Background

The ML and DL based algorithms and other technological components used in this investigation are briefly explained below.

1. Support Vector Machines (SVM)

SVMs are supervised ML algorithms mostly used for classification, regression, and outlier detection problems.

The basic idea of SVM is to create a hyperplane to separate the given data points into classes. The dimension of the hyperplane depends on the number of input features, for example, a set of 2-feature data points have a single line hyperplane, a 3-feature set will have a 2-D hyperplane and so on. A hyperplane is a decision boundary that helps to classify the data points. The data points that are closer to the hyperplane and influence the position and orientation of the hyperplane are known as the support vectors. These support vectors help in building the SVM with the basic idea of maximizing the margin between the data points and the hyperplane. SVMs can handle both classification and regression on linear and non-linear data, and are used in various applications such as malware detection, handwriting recognition, intrusion detection, face detection, email classification, gene classification etc.

2. Extended Gradient Boosting Algorithm (XGBoost)

XGBoost stands as a scalable and distributed Gradient-Boosted Decision Tree (GBDT) supervised machine learning technique, widely applied in tasks such as regression and classification. Its efficacy lies in its ability to handle large datasets and deliver robust performance across various domains. It is known as "extreme" because of its ability to do parallel computation on a single machine, which makes it more efficient than normal gradient boosting frameworks. XGBoost is an efficient implementation of gradient boosting algorithms, known for its scalability and high accuracy. It is specifically designed to boost the computing power of tree-based algorithms, primarily focused on enhancing both the performance of machine learning models and their computational speed [11]. XGBoost employs a parallel tree-building approach, distinguishing itself from GBDT, where tree construction is sequential. Its methodology adheres to a level-wise strategy, involving a thorough scan of gradient values. Utilizing these gradient values, XGBoost calculates partial sums to assess the quality of splits at each potential split point within the training set.

3. Long Short-Term Memory (LSTM)

LSTM, a deep learning-based recurrent neural network (RNN), excels in efficiently retaining long-term dependencies and is resilient against the vanishing gradient problem. To facilitate learning long-term dependencies, LSTMs leverage Gated Recurrent Units (GRUs). GRUs possess the capability to selectively forget information from the previous timestep and control the amount of information from the current timestep to be passed on to the next. Moreover, LSTMs offer versatility in training methods, including backpropagation through time and reinforcement learning. This adaptability makes LSTMs particularly well-suited for tasks requiring the retention and interpretation of information from extended sequences [12]. LSTMs find prominent application in learning, processing, classifying, and predicting based on time series data, especially in scenarios where there are lags of unknown duration between critical events [13]. They demonstrate effectiveness in complex problem domains such as machine translation, language modeling, speech recognition, and various other applications.

4. Moving Average Convergence Divergence (MACD)

The Moving Average Convergence Divergence is a momentum oscillator mainly used to deal with trends. It is an indicator that shows the relationship between two exponential moving averages (EMA). We obtain MACD by subtracting the longterm EMA (slow EMA) from the shortterm EMA (fast EMA). It is one of the most popular tools in technical analysis because it gives traders the ability to identify the short-term trend directions quickly and easily [14]. MACD is calculated using the formula:

$$MACD = fastEMA - slowEMA \tag{1}$$

5. Relative Strength Index (RSI)

Relative Strength Index (RSI) indicates magnitude of recent price changes. It can show that a stock is either overbought or oversold. Typically, RSI value ranges between 0 and 100, where 70 and above signal that a stock is becoming overbought/overvalued, whereas a value of 30 and less can mean that it is oversold. RSI can be calculated as:

$$RSI = 100 - [100/(1 + (AUPC/ADPC))] \tag{2}$$

where: AUPC = Average of Upward Price Change, and ADPC = Average of Down-ward Price Change.

6. Stochastic Oscillator (SO)

The Stochastic Oscillator serves as a momentum indicator, revealing the position of the closing price concerning the high-low range within a specified number of periods. In an uptrend, the closing price typically concludes near the high, while in a downtrend, it tends to approach the low. When the closing price deviates from the high or low, it signals a slowdown in momentum. Stochastics demonstrate optimal effectiveness in broader trading ranges or during periods of gradual trend movements.

$$\%K = [(C - L14)/(H14 - L14)] \times 100 \tag{3}$$

where: %K = The current value of the stochastic indicator, C = The most recent closing price, L14 = The lowest price traded of the 14 previous trading sessions, H14 = The highest price traded during the same 14-day period.

7. Moving Average (MA)

Moving average is used to smoothen stock prices on a chart by filtering out shortterm price fluctuations. Moving averages are calculated over a defined period, such as the last 7, 15, 30 or 200 days. There are two (most common) averages used in technical analysis.

a) Simple Moving Average (SMA)

A simple moving average is a straightforward average computed over a specified period. This technical indicator is useful for assessing whether the price of an asset will sustain its current trend or undergo a reversal in a bull or bear market.

$$SMA = (A1 + A2 + ... + An)/n \tag{4}$$

where: An is the price of an asset at period n, n is the number of total periods.

b) Exponential Moving Average (EMA)

The Exponential Moving Average (EMA) is a type of average that assigns higher weights to recent prices. Widely recognized as one of the most effective indicators for scalping, the EMA responds more swiftly to recent price fluctuations compared to older price changes. This responsiveness makes it particularly valuable for capturing short-term market movements and trends in scalping strategies.

$$EMA = Price(t) \times k + EMA(y) \times (1 - k) \tag{5}$$

where: t = current day / today, y = previous day / yesterday, k = 2 / (N + 1), and N = number of days in EMA.

III. Related Work

Yun et. al (2023), introduced an algorithm that utilizes two distinct input feature sets: internal technical indicators and external market prices. This algorithm undergoes a two-stage feature selection process, which includes feature set expansion, a hybridized approach integrating genetic algorithm-machine learning regressions for the selection of crucial features, and importance score filtering to identify the optimal features. Their effort aimed to select the most effective feature subset with a parsimoniously reduced number of features, prioritizing interpretability. The results demonstrated a 10.42% improvement in average forecasting RMSE for the optimal feature set and a 13.47% improvement for the best feature subset of the internal technical indicators. In their study, the researchers employed Savitzky–Golay smoothing as an integral component of piecewise optimal curve fitting to analyze various potential groupings of external features. The introduced local interpretability technique, incorporating piecewise optimal curve fitting and piecewise best feature subset, offered a more adaptable interpretation of stock price behavior. This method utilized a select few best features for different segments of the piecewise data, providing timely and flexible insights. In contrast to other feature-importance interpretability techniques that rely on either a singular data point or an entire data period, the proposed interpretability technique successfully overcame these limitations. It accurately reflected the primary characteristics inherent in time series data [15]. Houssein et. al (2022) implemented a hybrid model that integrated the Support Vector Regression (SVR) method with Equilibrium Optimizer (EO) to predict the closing prices of the Egyptian stock exchange (EGX). The study focused on three indices: EGX 30, EGX 30 capped, and EGX 50 EWI. Evaluation of the model was conducted using metrics such as mean absolute percentage error, average, standard deviation, best fit, worst fit, and CPU time. Furthermore, the researchers compared their model with recently developed metaheuristic optimization algorithms from the literature, including the whale optimization algorithm, salp swarm algorithm, Harris Hawks optimization, grey wolf optimizer, Henry gas solubility optimization, Barnacle's mating optimizer, Manta ray foraging optimization, and slime mould algorithm. The results demonstrated that the proposed EO-SVR model outperformed its counterparts. Notably, the study concluded that there was no discernible need to incorporate technical indicators and statistical measures, as their impact was not significant [16].

Hanh, Le Hong et al. (2022), used about 69,950 articles from four Vietnam based reputable online newspapers as their input data, which were divided into 8 stock market trend related parts, and they were collected between 2001 and 2021. The data were analyzed using Machine Learning architectures such as decision trees, random forests, K-Nearest Neighbors (KNNs) and Support Vector Machines (SVMs). They found SVM as the best performing model and the best dataset being Vietstock, digging deeper and refining the findings helped them to raise the accuracy up to 60.1% [17].

Hung et al. (2021) conducted a study examining the repercussions of COVID-19 on the performance of the Vietnamese Stock Market. Their analysis was based on panel data encompassing stock returns of 733 listed companies on both the Ho Chi Minh Stock Exchange (HOSE) and the Hanoi Stock Exchange (HNX) during the period from

January 2, 2020, to December 13, 2020. Employing the Random-Effect Model (REM), the researchers observed that the impacts were more pronounced during the pre-lockdown and second-wave periods, in comparison to the impact during the lockdown period. Furthermore, the study revealed variations in the impacts across sectors, with the financial sector experiencing the most significant effects [18].

Nti et. al (2020) conducted a comprehensive comparative analysis of ensemble techniques, including boosting, bagging, blending, and super learners (stacking). The study utilized Decision Trees (DT), Support Vector Machine (SVM), and Neural Network (NN) to construct twenty-five different ensembled regressors and classifiers. The evaluation encompassed execution times, accuracy, and error metrics using stock data from the Ghana Stock Exchange (GSE), Johannesburg Stock Exchange (JSE), Bombay Stock Exchange (BSE-SENSEX), and New York Stock Exchange (NYSE) spanning from January 2012 to December 2018. The findings indicated that stacking and blending ensemble techniques exhibited higher prediction accuracies, ranging from 90% to 100% and 85.7% to 100%, respectively, in comparison to bagging (53% to 97.78%) and boosting (52.7% to 96.32%). Additionally, the study recorded the root mean squared error (RMSE) for stacking within the range of 0.0001 to 0.001, and for blending, it ranged from 0.002 to 0.01 [19].

Hiransha et. al (2018) In 2018, Hiransha and collaborators employed deep learning architectures, including Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN), to analyze two distinct stock markets: the National Stock Exchange (NSE) of India and the New York Stock Exchange (NYSE). The models were trained using the stock prices of a single company from NSE and then applied to predict stock prices for five different companies from both NSE and NYSE. Comparing the results with the Autoregressive Integrated Moving Average (ARIMA) model, the study revealed that the neural networks, including MLP, RNN, LSTM, and CNN, consistently outperformed the existing linear model (ARIMA) in terms of predictive accuracy [20].

Phuoc, Tran, et. al (2024) [21] proposed a model to predict stock price trends in an emerging economy, utilizing the LSTM algorithm combined with technical indicators predict such as Simple Moving Average (SMA), Moving Average Convergence Divergence (MACD), and Relative Strength Index (RSI). Their study, based on VN-Index and VN-30 stocks, demonstrated a high prediction accuracy, highlighting the effectiveness of LSTM in analyzing and forecasting stock price movements. Unlike their work, our study employs a broader comparison by using multiple models (SVM, XGBoost, and LSTM) and evaluates multiple performance metrics, including RMSE, MAPE, F1-score, and computational time of each algorithm, for a comprehensive assessment of each algorithm. Additionally, our work incorporates an assessment of the importance of technical features on datasets, which was not emphasized in [21].

IV. Methodology

The investigation is done using three ML and DL based architectures - SVM, LGBoost and LSTM on the datasets - VRE, SSI, VCB, STB, VNM, HOSE, HNINDEX and VNINDEX to predict the stock market behavior of Vietnam. The datasets are pre- processed and various technical features such as MACD, SO, RSI, SMA and EMA are calculated after checking the suitability of different time frames. The following algorithm (Algorithm-1) is used to calculate the MACD.

Algorithm 1 Computation MACD

(1) Compute_MACD(dataset, slowEMA, fastEMA)
(2) slow ← compute EMA for slowEMA
(3) fast ← computione EMA for fastEMA
(4) df['MACD'] ← fastEMA - slowEMA
(5) df['MACD signal'] ← Moving Average of the MACD (line)
(6) end

The algorithm below (Algorithm-2) is used to calculate the SO.

Algorithm 2 SO Calculation

(1) Compute_SO(dataset, k value, d value)
(2) low min ← Minimum of df['low'] based on window size = k value
(3) high min ← Maximum of df['high'] based on window size = k value
(4) df['stock k'] ← 100 * (df['Close'] - low min) / (high max - low min)
(5) df['stock d'] ← Mean of df['stock k'] with window size = d
(6) end

RSI is used to determine the magnitude of recent price changes, which is calculated using the Algorithm-3, given below.

The EMA and SMA are calculated for different time spans, such as 7, 9, 10, 15, 30, and 90 days respectively. These technical features are added to the datasets and the datasets are normalized for the ease of analysis. The

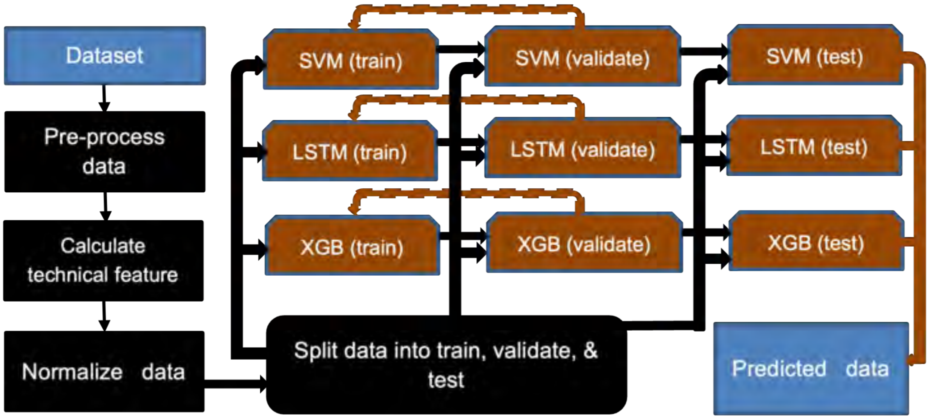


Figure 1. Flow diagram of the methodology.

Algorithm 3 RSI Calculation

(1) Compute_RSI(dataset, k value, d value)
(2) rollUp ← Average of Upward Price Change
(3) rollDown ← Average of Downward Price Change
(4) RS← rollUp / rollDown
(5) RSI←100-(100/(1+RS)
(6) end

datasets are split into three parts -train at 70%, test as 15% and validation with the rest 15%, and the architectures are then trained and validated. The algorithms are fine-tuned by resetting certain hyperparameters to achieve a better performance and finally they are tested. Various performance metrics such as RMSE, MAPE, accuracy score and F1-score are computed to see the effectiveness of the models on each dataset. The study has also considered the computational time as a final performance metric to justify the swiftness of the algorithms. The figure 1 represents the general methodology used in the study.

The study also investigated and determined the technical features that are crucial in the prediction of the stock market behavior. The RGBoost network architecture is specifically used to do that, and the results have been presented in the later section.

Table I
DATASET INFORMATION

Datasets	Start Date	End Date	Size
SSI	15/12/2009	25/06/2021	3606
VCB	30/06/2009	25/06/2021	2991
STB	02/07/2006	25/06/2021	3727
VNM	19/01/2006	25/06/2021	3842
VRE	28/07/2000	25/06/2021	5056
HOSE	15/07/2010	25/06/2021	65535
HNINDEX	24/05/2006	14/10/2019	3180
VNINDEX	07/08/2000	14/10/2019	4550

V. EXPERIMENTAL

The results obtained by using SVM, RGBoost and LSTM networks to predict the stock market behavior of Vietnam have been divided into different sections and presented here. The first two sections represent raw data and different technical features that are used in the experiment. The third section represents the confusion matrices of the algorithms on different datasets, and the fourth section depicts the importance of various technical features used in the study. The fifth section represents the closing price prediction of the datasets, and the final section shows various performance metrics scores obtained by the algorithms for individual datasets and a summary of it.

1. Dataset

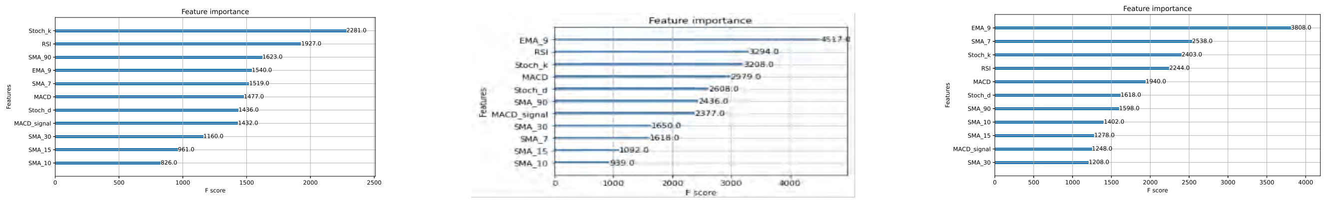
The objective of this study is specific to Vietnam, the datasets being used are collected from the two stock exchanges of HOSE and HNINDEX. A total of eight datasets are used in this study, five of them are from the HOSE based companies extracted from the VN30 dataset, the other two datasets are from the individual stock exchanges and the last one is the VNINDEX. HOSE, HNINDEX and VNINDEX are collected from different sources. The time span and number of records vary from dataset to dataset, however, most of them are from either 2006 or 2009, and until 2021. A Table 1 representation of the time and size of the datasets is given below.

2. Raw data Visualization

The figure 2 (a) is the representation of the raw closing price plot of the STB dataset. The figure 2 (b) shows the raw OLHC (Open, High, Low, Close) graph of the VCB dataset. The figure 2(c) depicts the raw closing price data of VNINDEX.

(a) STB Closing Stock Price History (b) VCB dataset (c) The closing price data of VNINDEX

Figure 2. Raw data Visualization



(a) Features ranking in SSI dataset (b) Features ranking in VCB dataset (c) Technical features ranking in HtNINDEX dataset

Figure 3. The importance of technical features on SSI, VCB and HNINDEX datasets.

(a) VRE prediction graph using LSTM (b) VCB prediction graph using LSTM

Figure 4. The prediction of the stock market behavior using the LSTM model.

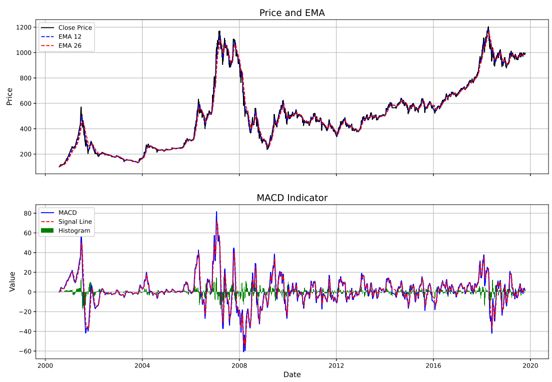


Figure 5. MACD of VNINDEX dataset.

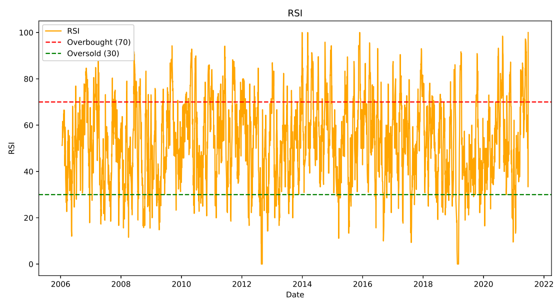


Figure 6. RSI of VNM dataset.

3. Technical features visualization

The visuals of different technical features used in this experiment are presented here. The trade trends of the VNINDEX as MACD graph is shown in figure 5. The Figure 6 represents the speed and change of price move-

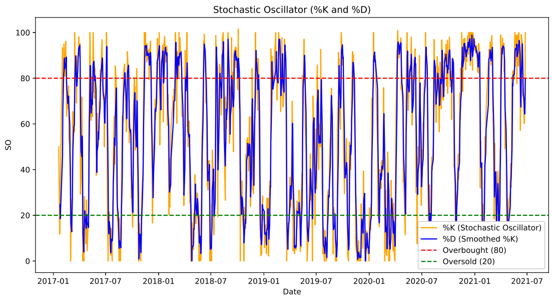


Figure 7. SO of SSI dataset.

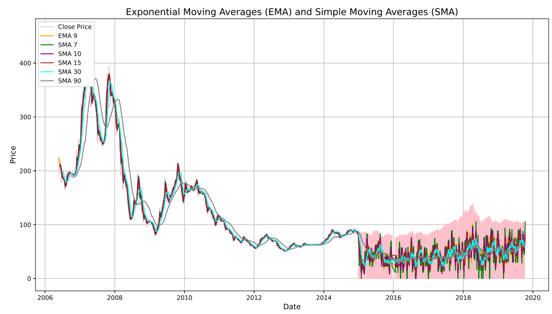


Figure 8. MA of HNINDEX dataset.

ments of the VNM dataset as an RSI graph. The Figure 7 shows the location of the closing price relative to the high-low range over a set number of periods as the stochastic oscillation of the SSI dataset. The moving averages (EMA and SMA) of HNINDEX dataset are presented in the Figure 8.

4. Confusion Matrices

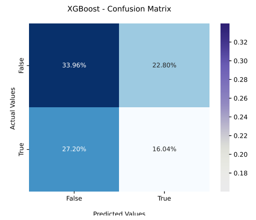
Confusion matrices of the algorithms are computed on each dataset and four of them are presented in their normalized form Figure 9.

5. Importance of technical features

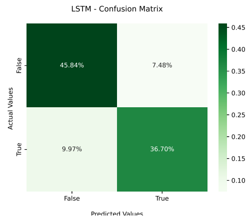
The technical features such as MACD, SO, RSI, EMA, and SMA are used in the investigation and their importances are determined using the XGBoost algorithm. Figures 3(a), 3(b) and 3(c) represent the importance of those technical features on SSI, VCB and HNINDEX datasets.

6. Stock Market Prediction Graphs

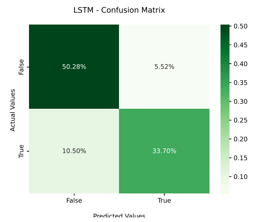
Below are the diagrams showing the prediction of the stock market behavior using the LSTM model. Figure 4(a) shows the VRE closing stock price prediction and Figure 4(b) represents the closing stock price prediction of VCB.



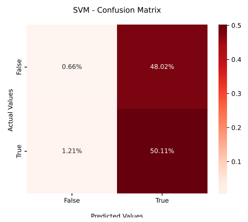
(a) XGBoost on HNINDEX



(b) LSTM on SSI



(c) LSTM on VRE



(d) SVM on VNINDEX

Figure 9. Confusion matrices on each dataset

7. Performance metrics

The Table 2 shows the RMSE, MAPE and computational time of each algorithm on different datasets. Accuracy score and F1-score of the architectures are computed and represented in the Table 3. The Table 4 represents a summary of all four performance metrics scores by the algorithms.

VI. Conclusion

The experimental findings indicate that, across all performance metrics, LSTM outperforms the other two architectures. However, it's noteworthy that LSTM comes with a trade-off, as it consumes a significantly higher amount of computational time compared to the alternative architectures. XGBoost, on the other hand, has a lower F1-score than the SVM, but it performs better than the SVM in other performance metrics. It is also observed that the technical features such as exponential moving average (EMA9), relative strength index (RSI), simple moving averages (SMA7, SMA90), stochastic oscillator (stochK) etc. are crucial in determining the stock market behavior and may be used as strong indicators in the same domain. The findings of the investigation are completely specific to the datasets used in the study and it is, therefore, tricky to generalize them. However, this will encourage researchers to use novel approaches to develop more effective prediction methods, and thereby assisting investors, analysts and other stakeholders involved in the stock market by providing more definite and accurate information of the future stock market.

The proposed models can assist investors and financial analysts in making more informed trading decisions by

Table II
PERFORMANCE COMPARISON BETWEEN LSTM, SVM, AND XGB MODELS ON VARIOUS STOCKS.

Models	Metrics	SSI	VCB	STB	VNM	VRE	HOSE	HNINDEX	VNINDEX	Average
LSTM	RMSE	0.036	0.744	0.122	0.700	0.023	0.186	0.628	0.383	0.353
	MAPE	16.604	13.790	14.719	15.461	17.136	15.659	17.544	19.672	16.323
	TIME	69.799	18.832	36.368	76.301	47.920	567.473	21.220	21.681	107.449
SVM	RMSE	0.422	0.134	0.831	0.247	0.271	0.370	0.432	0.468	0.397
	MAPE	19.541	17.390	13.761	12.090	19.325	23.211	19.932	16.721	17.746
	TIME	6.249	5.659	8.879	7.280	3.580	368.566	9.564	4.307	51.761
XGB	RMSE	0.273	0.830	0.670	0.026	0.057	0.317	0.676	0.081	0.366
	MAPE	17.755	15.811	15.904	13.641	20.954	20.871	18.431	15.596	17.370
	TIME	9.230	8.769	11.172	9.114	1.141	17.771	31.911	2.240	11.419

Table III
ACCURACY SCORE AND F1 SCORE

Datasets	Metrics	SVM	LSTM	XGBoost
SSI	Accuracy Score	0.822	0.832	0.805
	F1-score	0.802	0.808	0.799
VCB	Accuracy Score	0.806	0.808	0.793
	F1-score	0.714	0.704	0.713
STB	Accuracy Score	0.823	0.840	0.879
	F1-score	0.782	0.788	0.784
VNM	Accuracy Score	0.799	0.868	0.860
	F1-score	0.809	0.858	0.797
VRE	Accuracy Score	0.707	0.735	0.851
	F1-score	0.496	0.525	0.730
HNINDEX	Accuracy Score	0.460	0.761	0.475
	F1-score	0.427	0.658	0.459
VNINDEX	Accuracy Score	0.679	0.695	0.475
	F1-score	0.671	0.681	0.454
HOSE	Accuracy Score	0.348	0.695	0.393
	F1-score	0.322	0.691	0.388

Table IV
SUMMARY OF PERFORMANCE METRICS

Metrics	SVM	LSTM	XGBoost
Accuracy Score	0.605	0.679	0.629
F1 Score	0.662	0.732	0.597
RMSE	0.397	0.353	0.366
MAPE	17.746	16.323	17.370
Compu. Time	51.761	107.449	11.419

providing insights into stock price trends. The integration of these predictive models into automated trading systems or financial advisory tools can help optimize portfolio management and risk mitigation strategies.

Additionally, our analysis of technical indicators can aid financial professionals in refining their feature selection for stock forecasting in similar emerging markets.

REFERENCES

[1] Economy of Vietnam. In *Wikipedia*, retrieved from https://en.wikipedia.org/wiki/Economy_of_Vietnam

[2] Hanoi Stock Exchange. "In Sustainable Stock Exchanges (SSE) initiative", available at: https://sseinitiative.org/stock-exchange/hnx/

[3] Ho Chi Minh Stock Exchange, "In Sustainable Stock Exchanges (SSE) initiative," available at: https://sseinitiative.org/stock-exchange/hsx/

[4] Biswas, M., Shome, A., Islam, M. A., Nova, A. J., & Ahmed, S. ", Predicting stock market price: A logical strategy using deep learning, " *In 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 218-223). IEEE. 2021. doi: 10.1109/ISCAIE51753.2021.9431817.

[5] Sharma, A., Bhuriya, D., & Singh, U., "Survey of stock market prediction using machine learning approach, " *In 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA)*, Vol. 2, pp. 506-509. IEEE, 2017. doi: 10.1109/ICECA.2017.8212715.

[6] Bhattacharjee, I., & Bhattacharja, P.,"Stock Price Prediction: A Comparative Study between Traditional Statistical Approach and Machine Learning Approach," *In 2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, (pp. 1-6), 2019, IEEE. doi: 10.1109/EICT48899.2019.9068850.

[7] Koukaras, P., Nousi, C., & Tjortjis, C., "Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning," *Telecom*, 3(2):358-378., 2022. https://doi.org/10.3390/telecom3020019

[8] Zhang, R., "LSTM-based Stock Prediction Modeling and Analysis," *Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)*, pp. 2537-2542, 2022. doi: 10.2991/aebmr.k.220307.414

[9] Obthong, M., Tantisantiwong, N., Jeamwatthanachai, W., & Wills, G., "A Survey on Machine Learning for Stock Price Prediction: Algorithms and Techniques," 2020. doi: 10.5220/0009340700630071.

[10] Khang, P. Q., Kaczmarczyk, K., Tutak, P., Golec, P., Kuziak, K., Depczyński, R., ... & Rot, "Machine learning for liquidity prediction on Vietnamese stock market, " *Procedia Computer Science*, 192, 3590-3597, 2021. doi: 10.1016/j.procs.2021.09.132

[11] Gumelar, A. B., Setyorini, H., Adi, D. P., Nilowardono, S., Widodo, A., Wibowo, A. T., ... & Christine, E., "Boosting the accuracy of stock market prediction using xgboost and long short-term memory, " *In 2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, (pp. 609-613), 2020. doi: 10.1109/iSemantic50169.2020.9234256.

[12] Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., Salwana, E., & Shahab, S., "Deep Learning for Stock Market Prediction, " *Entropy* 22(8), 840, 2020. doi: 10.3390/e22080840

[13] Moghar, A., & Hamiche, M, "Stock Market Prediction Using LSTM Recurrent Neural Network, " *Procedia Computer Science*, 170, 1168-1173, 2020. doi: 10.1016/j.procs.2020.03.049

[14] Mokhtari, S., Yen, K. K., & Liu, J., "Effectiveness of

artificial intelligence in stock market prediction based on machine learning, " arXiv preprint arXiv:2107.01031, 2021. https://doi.org/10.48550/arXiv.2107.01031
[15] Yun, K. K., Yoon, S. W., & Won, D., "Interpretable stock price forecasting model using genetic algorithm-machine learning regressions and best feature subset selection, " *Expert Systems with Applications*, 213, 2023. doi: 10.1016/j.eswa.2022.118803
[16] Houssein, E.H., Dirar, M., Abualigah, L., "An efficient equilibrium optimizer with support vector regression for stock market prediction, " *Neural Comput & Applic*, 34, 3165–3200, 2022. doi: 10.1007/s00521-021-06580-9.
[17] Hanh, L. H., "Stock Market Prediction: The Application of Text-Mining in Vietnam, " *VNU Journal of Economics and Business*, 2(2), 2021. doi: 10.57110/jeb.v2i2.4715.
[18] Hung, D., Hue, N., & Duong, V., "The Impact of COVID-19 on Stock Market Returns in Vietnam," *Journal of Risk and Financial Management*, 2021.
[19] Nti, I.K., Adekoya, A.F. & Weyori, B.A, "A comprehensive evaluation of ensemble learning for stock-market prediction, " *Journal of Big Data*, 7, 20, 2020. doi: 10.1186/s40537-020-00299-5.
[20] Hiransha, M., Gopalakrishnan, E.A., Vijay Krishna Menon, Soman, K.P, "NSE Stock Market Prediction Using Deep-Learning Models, " *Procedia Computer Science*, 132, 1351-1362, 2018. doi: 10.1016/j.procs.2018.05.050.
[21] Phuoc, T., Anh, P. T. K., Tam, P. H., & Nguyen, C. V, "Applying machine learning algorithms to predict the stock price trend in the stock market–The case of Vietnam, "*Humanities and Social Sciences Communications*, 11(1), 1-18, 2024.

Truong Dinh Tu received his Ph.D. degree from Southeast University, Nanjing, China, in 2016. He is currently a lecturer and researcher at the Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam. His research interests include Cybersecurity, Data Analysis, Machine Learning, and Artificial Intelligence. Email: truongdinhtu@tdtu.edu.vn

Bhagawan Nath Bhagawan Nath is currently pursuing a Ph.D. program at the University of Jyväskylä, Finland. Since May 2024, he has been a lecturer at the Faculty of Information Technology, Greenwich University, Hanoi, Vietnam. His research interests include cybersecurity and data analysis using Machine Learning, Deep Learning, and other advanced technologies. Email: jitu.nath.vn@gmail.com