Forecasting stock price movement direction by machine learning algorithm

Bui Thanh Khoa[1], Tran Trong Huynh[2]
[1]Faculty of Commerce and Tourism, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
[2]Department of Mathematics, FPT University, Hanoi, Vietnam

Article Info

Article history:

Received Oct 17, 2021
Revised May 24, 2022
Accepted Jun 22, 2022

Keywords:

Artificial neural network
Forecasting
Logistic regression
Stock price movement direction
Support vector machine

ABSTRACT

Forecasting stock price movement direction (SPMD) is an essential issue for short-term investors and a hot topic for researchers. It is a real challenge concerning the efficient market hypothesis that historical data would not be helpful in forecasting because it is already reflected in prices. Some commonly-used classical methods are based on statistics and econometric models. However, forecasting becomes more complicated when the variables in the model are all nonstationary, and the relationships between the variables are sometimes very weak or simultaneous. The continuous development of powerful algorithms features in machine learning and artificial intelligence has opened a promising new direction. This study compares the predictive ability of three forecasting models, including support vector machine (SVM), artificial neural networks (ANN), and logistic regression. The data used is those of the stocks in the VN30 basket with a holding period of one day. With the rolling window method, this study got a highly predictive SVM with an average accuracy of 92.48%.

Corresponding Author:

Bui Thanh Khoa
Industrial University of Ho Chi Minh City
Ho Chi Minh City, Vietnam
Email: buithanhkhoa@iuh.edu.vn, khoadhcn@gmail.com

1. INTRODUCTION

The stock market, particularly in the short term, is described as time-varying, unpredictable, and non-linear. Financial forecasting of stock price movement direction (SPMD) is challenging for academics since several contributing elements include macroeconomic policy, investor mood, disclosure, and company-specific hazards. According to certain proponents of the efficient market theory, previous data are entirely represented in current prices; hence, utilizing historical data to forecast price direction [1]. However, specific tests have demonstrated that the market is occasionally inefficient, in which predicting SPMD is possible [2], [3]. On the other hand, creating the required and adequate circumstances for an efficient market is challenging, particularly in a developing market like Vietnam [4]–[6]. Thus, predicting SPMD is quite doable for the Vietnam market.

Technical analysis, time series forecasting, and machine learning-data mining are the three primary approaches for price predicting. As a result, technical analysts employ charts and indicators to forecast price behavior patterns and make purchases or sales [7], [8]. With the time series method, autoregressive integrated moving average (ARIMA), simple moving average (SMA), and moving average convergence/divergence (MACD) models have been adopted [9]. While the relationships between variables are challenging to observe, making econometric models less efficient, models using machine learning-data mining is more effective due to their learning ability. The emergence of machine learning algorithms that solve many fundamental problems in finance with high accuracy has created a breakthrough in the financial industry [10]. Logistic regression, neural networks, k-nearest neighbors (KNN), support vector machine (SVM), or random forest (RF) are some common algorithms for forecasting SPMD [11]–[15]. There are two methods to check the forecasting efficacy of the model. The first way is to split the data into two groups: training and testing. The training group is used to establish the ideal parameters in the algorithm, while the testing group is used to assess the prediction model's efficacy [16]. The second technique is the rolling method, in which the first observations are used to calculate the ideal parameter in the algorithm, further observations are predicted and then added to the data set, and subsequent observations are processed in the same manner.

Due to the time series' unique characteristics, the second technique was chosen for this investigation. Additionally, rather than forecasting all historical data, this study examines only a subset of the most recent observations due to the representativeness of the samples. To be more specific, the training data is limited to the latest 365 observations. Because the training data set is constantly updated, the starting parameters are adjusted, boosting the forecast's accuracy. Compared to the first technique, splitting the sample into two separate sets increases the sample's representativeness. Vijh *et al.* [17], for example, split the data set into two sections: training (4/6/2009-4/3/2017) and testing (4/4/2017-4/5/2019). The parameters produced by the training dataset are too old for predicting; it is debatable if forecasting for 2019 using data from 2017 is realistic. Cao and Tay [18], Ji *et al.* [19] divided the dataset into training, validation, and testing. While historical data addresses the reason, the performance will be much lower than the rolling window.

Machine learning and artificial intelligence developments have cleared the way for a potential new approach [20]. This research aimed to evaluate the predictive performance of three forecasting models: SVM, artificial neural networks (ANN), and logistic regression. The data utilized are those from the VN30 basket of equities with a one-day holding period. This work contributes both theoretically and practically. The study pointed out that SVM is a strong classification technique in machine learning that should be considered while addressing specific classification problems. Before employing an input variable, it should be analyzed using statistical and economic models to see whether or not it is related to the output variable. Furthermore, it suggests utilizing equation and logistic regression to explain the cause-effect relationship. Additionally, short-term investors may use it to forecast SPMD using the SVM and logistic regression. On the other hand, long-term investors should exercise greater caution, as this paradigm will no longer operate.

This study began by developing algorithms to serve as a basis for the research to accomplish the abovementioned aim. Following that, this research shared its methodology and data. The next part analyzed and discussed the research result. Finally, the study concluded with the study's contributions and further research prospects.

2. ALGORITHM

2.1. Logistic regression

Logistic regression is a statistical approach for describing the connection between two independent variables and a binary dependent variable (which can also be applicable for discrete dependent variables). Using this connection, logistic regression predicts the outcome based on a given set of input values. In logistic regression output forecasting, we compute the probability that the output will have the value 1 given the observation data, which means calculating *P(Y=1|X)*. Under the premise that the dependent variable has a binomial distribution, we define the odd ratio as (1).

$$G(X) = \frac{P(Y=1|X)}{P(Y=0|X)} = \frac{P(Y=1|X)}{1-P(Y=1|X)} \quad (1)$$

Take the log of both sides of (1) we have:

$$ln\, G\,(X) = ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = X\beta \quad (2)$$

where, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ are the estimated parameters. From (2), we make an equivalent transformation as (3).

$$P(Y = 1|X) = \frac{e^{X\beta}}{1+e^{X\beta}} \quad (3)$$

The maximum likelihood estimation (MLE) approach is often used to estimate β. The classification rule is determined by (3), which is:

$$y_i = \begin{cases} 1, P(y_i = 1|X) \geq 0.5 \\ 0, P(y_i = 1|X) < 0.5 \end{cases}$$

Logistic regression is commonly used for binary dependent variables in various applications. Some case studies in finance, such as Han *et al.* [21], anticipated challenging financial conditions; a sample of 76 enterprises and 32 characteristics linked to their financial measures were employed. In logistic regression, the authors utilized the backward stepwise technique. The result was reached with a high degree of accuracy (92.86 percent). Konglai and Jingjing [22] used logistic regression to assess the credit risk of Chinese listed businesses. The data set includes 130 firms with six dependent variables, separated into 90 for training and 40 for testing sets. The accuracy obtained in the training sample is 87.8%, and that of the testing set is 75%.

2.2. Support vector machine (SVM)

Vapnik and Lerner [23] proposed the SVM algorithm to solve the classification problem. SVM is a supervised mathematical approach used to categorize data across several areas. We have $wx^T + b = 0$, where w and b are the coefficients. The coefficients w and b should be chosen such that $wx^T + b \geq 1$ if vnic, vnipc related to the market portfolio are statistically significant at 0.05 [24]. The negative sign of the vnic estimator implies that the greater the market volatility, the lower the probability of an upward movement of the closing price is expected. This result can be explained as increasing market volatility will increase the market risks while the expected return remains unchanged; investors will postpone participating in the market or selling their holdings to reduce risks. This result leads to supply exceeding demand, which will escalate the downward pressure on prices. The variable vnipc, representing the return rate of the market portfolio, has a positive sign, implying that the larger the vnipc, the higher chance of a price increase of Mobile World Group, which have ticker as MWG. This result can be explained: investors are more optimistic when the market is favorable (return rate increases), making more investments. When analyzing univariate effects, one thing to consider is that we should implicitly assume that the remaining variables are fixed. and $wx^T + b \leq -1$ if $y_i = -1$. Use the training set to find w and b such that $||w||$ is minimized, and the vectors x_i, which $|y_i|(wx_i^T + b) = 1$ are called support vectors. A kernel function is used for mapping into a high-dimensional space where the data will be more clearly separated to improve classifier performance. The kernel function is defined by the dot product: $K(x, y) = \langle f(x), f(y) \rangle$. Standard kernel functions are linear, polynomial, radial basic function.

However, it is impossible to find a perfect classification hyperplane; Cortes and Vapnik [25] proposed adding soft margins, i.e., to accept some misclassified observations. The SVM algorithm is now minimized: $\min_{w,b,\xi} \left(\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \right)$ given that $y_i(w^T w \phi(x_i) + b) \geq 1 - \xi_i$. Where C is a hyperparameter, and ϕ is mapping to higher-dimensional space.

2.3. Artificial neural networks

ANN is first introduced by McCulloch and Pitts [26] as an algorithm that simulates the brain's neurons. Later, ANN became a widely used machine learning algorithm in various fields. An essential ANN consists of a linear combination of input variables that go through hidden layers and finally pass to the output. One of the crucial problems of ANN is training, which means optimization of input parameters [27]. In this case, the backpropagation algorithm (BP) is used. The BP is widely used for training neural networks. The data in the ANN is saved in connection weights, which may be thought of as system memory [28]. The goal of BP training is to modify the weights between neurons frequently to decrease the output error.

Some typical studies using ANN have obtained favorable results. In predicting SPMD, Leung *et al.* [29] examined several prediction models based on multivariate classification approaches and compared them to many parametric and non-parametric models. Empirical evidence suggests that layered models (quadratic discriminant analysis (QDA), logit, probit, and ANN) outperform traditional econometric estimation methods (adaptive exponential smoothing, vector autoregression) in forecasting stock market movement and maximizing profits from investment transactions. Tsai and Wang [30] used the ANN algorithm and decision tree (DT) to build a model to forecast SPMD in the Taiwan industry. Its accuracy in the ANN model is 59.016%. Patel *et al.* [31] combined the ANN classification algorithm with photo news to predict the opening price of securities and obtained an accuracy of 70%.

3. METHOD

As demonstrated in Table 1, research data contains 30 firms from the VN30 list (unadjusted pricing) and the closing VN-Index in a one-day period. The data collecting period was from July 28, 2000, to July 30, 2021, during which certain firms were newly created, and there were some days off, resulting in a variation in the number of observations for these companies. The information was gathered from the website cafef.vn.

Table 1. Observations and tickers in the VN30 list

Ticker	Observations	Ticker	Observations	Ticker	Observations
BID	1840	MWG	1731	TCB	762
BVH	2989	NVL	1112	TCH	1173
CTG	2975	PDR	2573	TPB	789
FPT	3611	PLX	1039	VCB	2987
GAS	2265	PNJ	3043	VHM	772
HDB	857	POW	605	VIC	3424
HPG	3381	REE	5050	VJC	1076
KDH	2833	SBT	3319	VNM	3838
MBB	2401	SSI	3600	VPB	957
MSN	2896	STB	3720	VRE	898

Date, ticker, closing price, opening price, highest price, lowest price, and trading volume were all included in each observation. The variables in the research are listed in Table 2. This research focuses on three models: logistic regression, SVM, and ANN. Assuming that the historical data has a maximum value of one year, the research will employ 365 observations of fixed training data to create predictions using the rolling window approach. As illustrated in Figure 1, algorithms are employed to identify the ideal parameters, beginning with the first 365 observations, forecasting the 366th observation, and continue until the final observation.

Table 2. Variable description

Variable	Formular	Description
$close_t$		The closing price is at date t.
$foredir_t$	$foredir_t = \begin{cases} 1, close_t \geq close_{t-1} \\ 0, close_t < close_{t-1} \end{cases}$	The SPMD. foredir=1 indicates that the closing price is higher than before.
HL_t	$High_t$-Low_t	The range of the price within a trading day.
LO_t	Low_t-$Open_t$	The range of the lowest price in comparison to the opening price.
$variation_t$	$Close_t$–$Close_{t-1}$	The difference in closing prices between two consecutive days.
$ma7_t$	$\frac{1}{7}\sum_{i=0}^{6} close_{t-i}$	The average closing price for 7 consecutive trading sessions.
$ma14_t$	$\frac{1}{14}\sum_{i=0}^{6} close_{t-i}$	The average closing price throughout 14 consecutive trading sessions.
$ma21_t$	$\frac{1}{21}\sum_{i=0}^{6} close_{t-i}$	The average closing price throughout 21 consecutive trading sessions.
$sd7_t$	$\sqrt{var(close_t, close_{t-1}, \ldots, close_{t-6})}$	The standard deviation of closing price of 7 consecutive trading sessions
$vnic_t$	$vnindex_t - vnindex_{t-1}$	The difference in the value of the VN-index between two consecutive trading sessions.
$vnipc_t$	$\frac{vnindex_t - vnindex_{t-1}}{vnindex_{t-1}} \times 100$	The VN-index portfolio's return rate
$insec_t$		The variable of a time trend. The default origin is January 1, 1970.

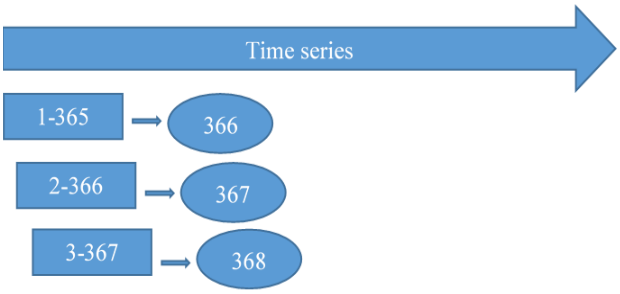

Figure 1. Rolling window method

4. RESULTS AND DISCUSSION

4.1. Result

Because of the large number of tickers in the VN30 list, we will choose a particular ticker, MWG (Mobile World Investment Joint Stock Company), to present descriptive statistics of variables and perform a forecasting SPMD; the same approach will be applied to the remaining tickers. The fundamental statistical values of the MWG are presented in Table 3.

Table 3. Descriptive statistics table of MWG

statistics	close	HL	LO	variation	ma7	ma14	ma21	sd7	vnic	vnipc
median	109	2.2	-1	0	109	109.7	110.4	1.8	0.7	0.1
min	58.9	0	-13.5	-90.3	60.2	61.1	61.3	0.2	-73.2	-8
max	183.4	16	5.1	11	179	174.9	171.8	45.9	76.2	7.4
mean	110.5	2.7	-1.3	0	110.4	110.4	110.3	2.5	0.3	0.02
sd	27.8	1.8	1.5	3.6	27.5	27.1	26.8	3.2	9.7	1.3

The statistics table shows that the price fluctuates from 58.9 (thousand dongs/share) to 183.4 (thousand dongs/share); the average price is 110.5 (thousand dongs/share). The foredir ends up with 910 observations that predict a decrease in the closing price compared to the previous day, and the remaining 821 observations of the closing price were not decreased. The most robust closing price movement, -90.3 (thousand dongs/share), was since MWG paid cash dividends and bonus shares on May 22, 2017, and its issuance of shares to raise more capital. The standard deviation (sd) of the MWG variation is much lower than the standard deviation of the vnic, which proves that the price movement between 2 consecutive sessions of MWG is lower than that of the market. This result implies that the overall risk of MWG is lower than the market risks. The following chart will better visualize the price movement of the VN-index (red line) and MWG (black line) shown in Figure 2.

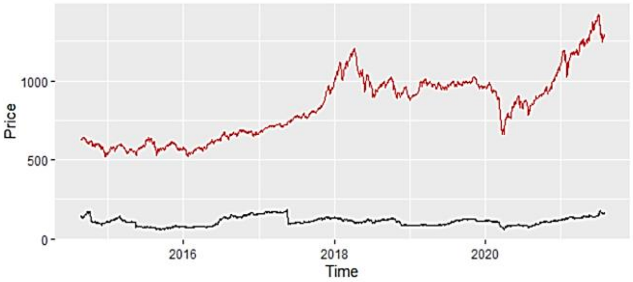

Figure 2. Closing prices of VN-index and MWG

This study performed logistic regression for all observations of the MWG data to summarize the significant level of the independent variables. The results are summarized in Table 4. The variables close, HL, LO, vnic, vnipc, insec, ma7, sd7 have the lowest statistical significance of 0.1. The closing price has a negative (-) sign, implying that today’s high closing price will lower the possibility of a price increase for the next session. The variables HL, LO, which measure stock volatility, have positive (+) signs, implying that today’s significant volatility will enhance the probability of the next session’s price increase. The time trend variable insec is also statistically significant at 0.001, showing that MWG’s prices tend to increase over time. However, the estimated coefficient has a low value of only 0.00024, which makes little impact on changes in the price movement. Among the trend indicator variables ma7, ma14, ma21, only ma7 has statistical significance and a positive sign proving that the average price increase in the short term (7 days) is expected to boost a price increase. However, the variable ma21 is statistically significant at 0.12 and has a negative estimator, indicating the SPMD in the longer term (21 days). Specifically, if the variable ma21 increases, we expect to lower the possibility of a price increase.

The variables vnic, vnipc related to the market portfolio are statistically significant (at 0.05), as shown in Table 4. The negative sign of the vnic estimator implies that the greater the market volatility, the lower the probability of an upward movement of the closing price is expected. This result can be explained as increasing market volatility will increase the market risks while the expected return remains unchanged; investors will postpone participating in the market or selling their holdings to reduce risks. This result leads to supply exceeding demand, which will escalate the downward pressure on prices. The variable vnipc, representing the return rate of the market portfolio, has a positive sign, implying that the larger the vnipc, the higher chance of a price increase of MWG. This result can be explained: investors are more optimistic when the market is favorable (return rate increases), making more investments. When analyzing univariate effects, one thing to consider is that we should implicitly assume that the remaining variables are fixed. MWG data includes 1,731 observations; the first 365 observations are used to estimate the parameters in the model using the rolling window method. The forecast results of the logistic regression model, SVM, and ANN, are:

Table 4. The logistic regression results of MWG

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.39400	1.19600	-3.675	0.000237***
close	-0.04529	0.01769	-2.561	0.010444*
HL	0.06806	0.03951	1.723	0.084966.
LO	0.10140	0.04677	2.168	0.030125*
variation	-0.00151	0.02233	-0.068	0.946027
vnic	-0.04484	0.02008	-2.233	0.025545*
vnipc	0.36400	0.17910	2.032	0.042133*
insec	0.00024	0.00007	3.543	0.000395***
ma7	0.05613	0.02954	1.9	0.057436.
ma14	0.02365	0.03616	0.654	0.51306
ma21	-0.03321	0.02142	-1.551	0.121009
sd7	-0.05048	0.02260	-2.234	0.025510*

*Significant level: '***': 0.001; '**': 0.01; '*': 0.05: '.': 0.1; ' ':1*

4.1.1. Logistic regression

Using the sigmoid function and the maximum likelihood estimation method to estimate the parameters. The logistic regression model achieves relatively low accuracy, only 53.84%, as shown in Table 5. The model correctly predicted a price increase in 92 observations and a price decline in 644 observations while incorrectly predicted a price increase in 594 observations.

4.1.2. SVM

Using radial kernel function with $cost = 100$ and $\gamma = 0.1$. The SVM model correctly predicted a price increase in 617 observations, a price decrease in 634 observations, and falsely predicted 116 observations as Table 6. The accuracy achieved is very high, up to 91.51%, much higher than the accuracy of the logistic regression model as the SVM model.

Table 5. Summary of the forecast results using logistic regression

		Actual	
		Positive-1	Negative-0
Predicted	Positive-1	92	37
	Negative-0	594	644
Accuracy		53.84%	

Table 6. Summary of the forecast results using SVM

		Actual	
		Positive-1	Negative-0
Predicted	Positive-1	617	69
	Negative-0	47	634
Accuracy		91.51%	

4.1.3. ANN

Using a hidden layer with four nodes, the Tanh function, and the Backpropagation algorithm. The parameter estimation results are illustrated in Figure 3, and the forecast results are summarized in Table 7. The accuracy obtained is 58.81%, higher than the logistic regression model but lower than the SVM model.

All 30 tickers are handled in the same manner as MWG. The forecast's accuracy is reflected in Table 8. The average accuracies in predicting 30 tickers using logistic regression, SVM, and ANN are 58.93%, 92.48%, and 60.03%, respectively. The SVM model is more efficient than the other two models.

4.2. Discussion

First, when examining the statistical relationship between the independent and dependent variables, the regression results for MWG in Table 4 show that most of the variables are statistically significant at 0.1, except the variation, ma14, and ma21. This result implies that the included variables exerted their significant impact. However, the low magnitude of the estimated coefficients is the reason for the limited predictiveness of the logistic regression model, with an average accuracy of 58.93%. This conclusion is similar to the previous studies by Jabbarzadeh *et al.* [32] with an accuracy of 60.73%, Jiao and Jakubowicz [33] with an average area under the receiver operating characteristic curve (AUC) of 0.6942. We cannot compare our study with Jiao and Jakubowicz [33] because of the difference in research measures. On the other hand, for the total of 30 tickers, the forecast accuracy of the logistic regression model is relatively stable, ranging from the lowest at 53.48% to the highest at 63.32% in Table 9.

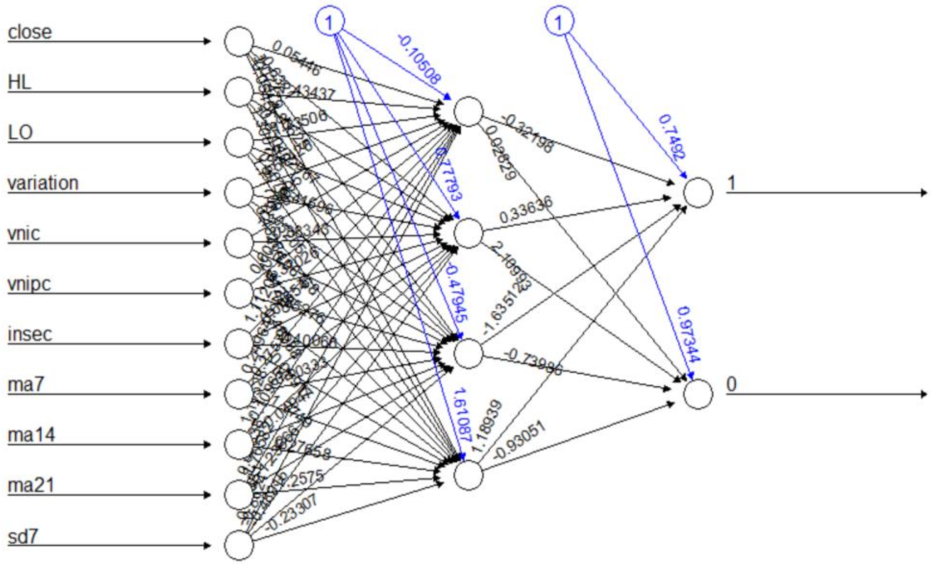

Figure 3. Diagram of a single-hidden-layer ANN model with four nodes

Table 7. Summary of the forecast results using ANN

		Actual	
		Positive-1	Negative-0
Predicted	Positive-1	158	523
	Negative-0	40	646
Accuracy		53.84%	

Table 8. Summary of the accuracy of forecast results of VN30 tickers

Ticker	Logistic	SVM	ANN	ticker	Logistic	SVM	ANN
BID	59.49	92.55	62.94	POW	56.85	93.36	87.14
BVH	58.4	92.91	59.31	REE	62.21	92.53	44.58
CTG	61.93	93.53	62.43	SBT	62.2	93.6	61.08
FPT	58.67	90.64	47.89	SSI	58.9	91.75	48.58
GAS	57.55	92.9	61.49	STB	63.32	92.19	43.3
HDB	57.61	93.71	81.34	TCB	55.53	95.73	76.88
HPG	58.93	91.22	48.53	TCH	54.88	91.97	62.18
KDH	61.89	93.36	49.98	TPB	58.82	96.94	81.65
MBB	62.15	93.67	46.39	VCB	58.6	93.25	50.59
MSN	62.48	93.05	45.81	VHM	57.84	94.85	81.13
MWG	53.84	91.51	58.81	VIC	59.87	91.57	48.82
NVL	58.29	93.72	65.11	VJC	58.57	86.66	73.17
PDR	59.98	92.76	48.48	VNM	61.83	91.34	45.6
PLX	56.3	91.41	70.22	VPB	57.5	88.03	68.8
PNJ	57.86	91.64	48.71	VRE	55.62	91.95	77.53
Average Accuracy		Logistic 58.93		SVM 92.48		ANN 60.03	

Table 9. Summary of the descriptive statistics on the accuracy

Statistics	min	Q1	median	mean	Q3	max
Logistic	53.84	57.56	58.63	58.93	61.37	63.32
SVM	86.66	91.59	92.65	92.48	93.49	96.94
ANN	43.30	48.54	60.20	60.28	69.87	87.14

SVM model proves superior accuracy and stability, both MWG and all the rest of the tickers in the research data, as shown in Tables 8 and 9. SVM model correctly classifies with average accuracy up to 92.48% and over 90%, except for the ticker of VietJet Aviation Joint Stock Company (VJC) and Vietnam Prosperity Joint Stock Commercial Bank (VPB), with 88.66% and 88.03%, respectively. Moreover, the lowest accuracy is 86.66%, and the highest is 96.94% (the ticker TPB). This result is better than previous similar studies such as Kim [34], Kara *et al.* [35], Patel *et al.* [36], Duong *et al.* [37].

While both the logistic regression and SVM are pretty stable, the accuracy of the ANN model highly fluctuates among different stocks, ranging from 43.3% to 87.14% and reaching an average of 60.03%. On the other hand, in Tables 8 and 9 out of 30 tickers with less than 50% (even worse than the random selection method). The immense volatility among tickers implies a wrong choice of the number of hidden layers, nodes, and the active function. Achieving more than 25% of the tickers having an accuracy of 69.87% or higher is good for improving the model. The Boxplot graph in Figure 4 gives us a better visualization.

The model estimating approach is responsible for the SVM model's success. The rolling window approach is more efficient than other methods due to the time series continuity, which makes the input parameters more precise. Furthermore, a 365-day cycle is a good decision; if it had been longer, the data would have gotten too old, and if it had been shorter, it would not have been a suitable picture of the totality.

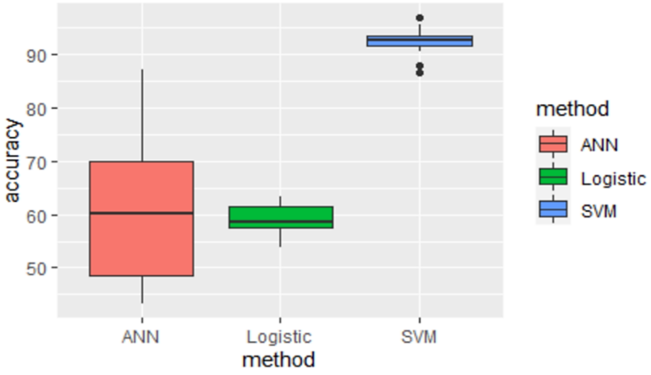

Figure 4. Accuracy of the three forecasting methods

5. CONCLUSION

This study aims to forecast the SPMD for the VN30 list. The forecasting performance of the three models (logistic regression, SVM, and ANN) is compared based on historical data of stocks in the VN30 list for July 28, 2000, to July 30, 2021. This research used the latest 1-year data to estimate the input parameters and combine it with the rolling window method to forecast the successive observations.

Logistic regression is critical for deriving an explanation for the statistical connection between independent and dependent variables. The logistic regression model's accuracy is constant at an average of 58.93 percent. Although the ANN model has an average accuracy of 60.28%, higher than that of the logistic regression model, it fluctuates significantly. The SVM model is the most practical forecasting model compared to the other two models, achieving a very high average accuracy of 92.48%. This result implies that short-term investors can use SPMD to make short-term investments to maximize profits. Moreover, this study has some theoretical and practical contributions. To begin, SVM is a robust classification method in machine learning that should be considered while tackling specific classification issues. Before utilizing an input variable, it should be evaluated to see whether or not it is connected to the output variable using statistical and economic models. Second, to ascertain the variables influencing SPMD, this study proposed (4) as a theoretical model and logistic regression to explain the cause-effect connection. Furthermore, through the SVM and logistic regression, short-term investors can use it to predict SPMD. However, long-term investors should be more cautious because this model will not work anymore.

Apart from the benefits of this study, several drawbacks may represent possibilities for further research. Firstly, the study scope remained limited, encompassing only 30 tickers in Ho Chi Minh City Stock Exchange (HoSE). Further research should broaden the scope of the study to include more stock exchanges to guarantee reproducibility. Other classification methods not utilized in machine learning include KNN, naive Bayes, deep learning, and decision trees. A further study on SPMD will compare the algorithms' efficacy to find the best successful forecast model.

REFERENCES

[1] B. G. Malkiel and E. F. Fama, "Efficient capital markets: a review of theory and empirical work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, May 1970, doi: 10.1111/j.1540-6261.1970.tb00518.x.

[2] G. Dutta, P. Jha, A. K. Laha, and N. Mohan, "Artificial neural network models for forecasting stock price index in the bombay stock exchange," *Journal of Emerging Market Finance*, vol. 5, no. 3, pp. 283–295, Dec. 2006, doi: 10.1177/097265270600500305.

[3] Z. Hu, J. Zhu, and K. Tse, "Stock market prediction using support vector machine," in *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, 2013, pp. 115–118.

[4] P. D. Khanh and P. T. Dat, "Efficient market hypothesis and calendar effects: Empirical evidences from the Vietnam stock markets," *Accounting*, pp. 893–898, 2020, doi: 10.5267/j.ac.2020.5.005.

[5] B. T. Khoa and D. T. Thai, "Capital structure and trade-off theory: evidence from Vietnam," *The Journal of Asian Finance, Economics and Business*, vol. 8, no. 1, pp. 45–52, 2021.

[6] B. T. Khoa and T. T. Huynh, "Predicting exchange rate under UIRP framework with support vector regression," *Emerging Science Journal*, vol. 6, no. 3, pp. 619–630, Apr. 2022, doi: 10.28991/ESJ-2022-06-03-014.

[7] H. Li, W. W. Y. Ng, J. W. T. Lee, B. Sun, and D. S. Yeung, "Quantitative study on candlestick pattern for Shenzhen stock market," in *2008 IEEE International Conference on Systems, Man and Cybernetics*, Oct. 2008, pp. 54–59, doi: 10.1109/ICSMC.2008.4811250.

[8] B. T. Khoa and T. T. Huynh, "Support vector regression algorithm under in the CAPM framework," in *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, Oct. 2021, pp. 186–190, doi: 10.1109/ICDABI53623.2021.9655797.

[9] A. S. Ahmar, "Sutte Indicator: an approach to predict the direction of stock market movements," *arXiv preprint arXiv:1903.11642*, 2019

[10] P. H. D. Abd Samad, S. Mutalib, and S. Abdul-Rahman, "Analytics of stock market prices based on machine learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 16, no. 2, pp. 1050–1058, Nov. 2019, doi: 10.11591/ijeecs.v16.i2.pp1050-1058.

[11] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Systems with Applications*, vol. 42, no. 20, pp. 7046–7056, Nov. 2015, doi: 10.1016/j.eswa.2015.05.013.

[12] E. Schöneburg, "Stock price prediction using neural networks: A project report," *Neurocomputing*, vol. 2, no. 1, pp. 17–27, Jun. 1990, doi: 10.1016/0925-2312(90)90013-H.

[13] H. Bessembinder and K. Chan, "The profitability of technical trading rules in the Asian stock markets," *Pacific-Basin Finance Journal*, vol. 3, no. 2–3, pp. 257–284, Jul. 1995, doi: 10.1016/0927-538X(95)00002-3.

[14] P. N. Rodriguez and A. Rodriguez, "Predicting stock market indices movements," *WIT Transactions on Modelling and Simulation*, vol. 38, 2004

[15] B. T. Khoa and T. T. Huynh, "Factors affecting forecast accuracy of individual stocks: SVR algorithm under CAPM framework," in *2022 International Conference for Advancement in Technology (ICONAT)*, Jan. 2022, pp. 1–6. doi: 10.1109/ICONAT53423.2022.9725916.

[16] K. Kumar and D. P. Gandhmal, "An intelligent indian stock market forecasting system using LSTM deep learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, pp. 1082–1089, Feb. 2021, doi: 10.11591/ijeecs.v21.i2.pp1082-1089.

[17] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock closing price prediction using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 599–606, 2020, doi: 10.1016/j.procs.2020.03.326.

[18] L. J. Cao and F. E. H. Tay, "Support vector machine with adaptive parameters in financial time series forecasting," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1506–1518, Nov. 2003, doi: 10.1109/TNN.2003.820556.

[19] X. Ji, J. Wang, and Z. Yan, "A stock price prediction method based on deep learning technology," *International Journal of Crowd Science*, vol. 5, no. 1, pp. 55–72, Apr. 2021, doi: 10.1108/IJCS-05-2020-0012.

[20] B. T. Khoa, P. T. Son, and T. T. Huynh, "The relationship between the rate of return and risk in fama-french five-factor model: a machine learning algorithms approach," *Journal of System and Management Sciences*, vol. 11, no. 4, pp. 47–64, Dec. 2021, doi: 10.33168/JSMS.2021.0403.

[21] D. Han, L. Ma, and C. Yu, "Financial prediction: application of logistic regression with factor analysis," in *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, Oct. 2008, pp. 1–4, doi: 10.1109/WiCom.2008.2308.

[22] Z. Konglai and L. Jingjing, "Studies of discriminant analysis and logistic regression model application in credit risk for China's listed companies," *Management Science and Engineering*, vol. 4, no. 4, pp. 24–32, 2011.

[23] V. Vapnik and A. Y. Lerner, "Recognition of patterns with help of generalized portraits," *Avtomat. i Telemekh*, vol. 24, no. 6, pp. 774–780, 1963

[24] B. T. Khoa and T. T. Huynh, "Is it possible to earn abnormal return in an inefficient market? an approach based on machine learning in stock trading," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–14, Dec. 2021, doi: 10.1155/2021/2917577.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[26] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943, doi: 10.1007/BF02478259.

[27] M. R. Pahlawan, E. Riksakomara, R. Tyasnurita, A. Muklason, F. Mahananto, and R. A. Vinarti, "Stock price forecast of macro-economic factor using recurrent neural network," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 74–83, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp74-83.

[28] K. J. Kumar and K. Kumar, "Prediction of future stock close price using proposed hybrid ANN model of functional link fuzzy logic neural model (FLFNM)," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 1, no. 1, Mar. 2012, doi: 10.11591/ij-ai.v1i1.362.

[29] M. T. Leung, H. Daouk, and A.-S. Chen, "Forecasting stock indices: a comparison of classification and level estimation models," *International Journal of Forecasting*, vol. 16, no. 2, pp. 173–190, Apr. 2000, doi: 10.1016/S0169-2070(99)00048-5.

[30] C. F. Tsai and S. P. Wang, "Stock price forecasting by hybrid machine learning techniques," in *Proceedings of the international multiconference of engineers and computer scientists*, 2009, vol. 1, no. 755.

[31] H. R. Patel, S. M. Parikh, and D. N. Darji, "Prediction model for stock market using news based different classification, regression and statistical techniques: (PMSMN)," in *2016 International Conference on ICT in Business Industry and Government (ICTBIG)*, 2016, pp. 1–5, doi: 10.1109/ICTBIG.2016.7892636.

[32] A. Jabbarzadeh, S. Shavvalpour, H. Khanjarpanah, and D. Dourvash, "A multiple-criteria approach for forecasting stock price direction: Nonlinear probability models with application in S&P 500 Index," *International Journal of Applied Engineering Research*, vol. 11, no. 6, pp. 3870–3878, 2016

[33] Y. Jiao and J. Jakubowicz, "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks," in *2017 IEEE International Conference on Big Data (Big Data)*, Dec. 2017, pp. 4705–4713, doi: 10.1109/BigData.2017.8258518.

[34] K. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1–2, pp. 307–319, Sep.

2003, doi: 10.1016/S0925-2312(03)00372-2.

[35] Y. Kara, M. Acar Boyacioglu, and Ö. K. Baykan, “Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange,” *Expert Systems with Applications*, vol. 38, no. 5, pp. 5311–5319, May 2011, doi: 10.1016/j.eswa.2010.10.027.

[36] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, “Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques,” *Expert Systems with Applications*, vol. 42, no. 1, pp. 259–268, Jan. 2015, doi: 10.1016/j.eswa.2014.07.040.

[37] D. Duong, T. Nguyen, and M. Dang, “Stock market prediction using financial news articles on ho chi minh stock exchange,” in *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, Jan. 2016, pp. 1–6, doi: 10.1145/2857546.2857619.

BIOGRAPHIES OF AUTHORS

Bui Thanh Khoa received his Master’s degree in Business Economics from Université Toulouse 1 Capitole in France in 2012 and his doctorate in Business Administration from Ho Chi Minh City Open University in Vietnam in 2020. He has numerous papers in the SCOPUS and ISI databases. In addition to serving as a reviewer for numerous prestigious journals, he is a member of the Advisory International Editorial Board of Journal the Messenger, an ISI system journal; as well as a member of the Editorial Board of Journal of System and Management Sciences, Advances in Operations Research, Scopus indexed journals; and International Journal of Technology Transfer and Commercialization from Inderscience Publisher. His research interests are methodology, electronic commerce, organizational behavior, and consumer behavior. He can be contacted at email: khoadhcn@gmail.com.

Tran Trong Huynh is a lecturer at FPT University, He got a Master’s degree in Mathematics in 2013 at Ho Chi Minh City University of Education and finance in 2020 at the University of Economics Ho Chi Minh City. His current research interests include finance, applied mathematics, data science, econometrics, and machine learning. He can be contacted at email: huynhtt4@fe.edu.vn.