Thailand and The World Economy
P-ISSN: 2630-0931
E-ISSN: 2651-0529
Homepage : *https://www.tci-thaijo.org/index.php/TER*
Vol. 39, No.2, May - August 2021 **Page [56-64]**

Stock market index prediction using machine learning: evidence from leading Southeast Asian countries

Supakorn Chaengkham
Strategic Mutual Fund Department, Country Group Securities Public Company Limited, Thailand

Suthin Wianwiwat[*]
Faculty of Economics,Khon Kaen University, Thailand

Received 13 January 2021, Received in revised form 29 April 2021,
Accepted 10 May 2021, Available online 7 June 2021

Abstract

Financial investment in stock markets in emerging economies has played an important role in wealth management. Thus, this study aimed to present the application of the Support Vector Machine (SVM) with using the Toda-Yamamoto causality test to select economic indicators for predicting the movement of one month ahead of the stock market index in four leading Southeast Asian (ASEAN) nations: Indonesia, Malaysia, Singapore, and Thailand. Monthly data were sampled, ranging from January 2002 to December 2019. The linear kernel SVM provided useable results with accuracy ranging from 58.14 % to 65.12 %, which performed better than the sigmoid kernel SVM. According to the efficient-market hypothesis (EMH), the stock markets of Singapore and Malaysia were the most efficient among the four stock markets, whereas investors could strategically utilise this SVM algorithm to gain more returns from stock markets in Thailand and Indonesia.

Keywords: Stock market; Support Vector Machine; SVM; Toda-Yamamoto causality
JEL Classifications: C53, G17

[*]**Corresponding author**: Khon Kaen University, Khon Kaen, Thailand 40002. E-mail: wsuthi@kku.ac.th.

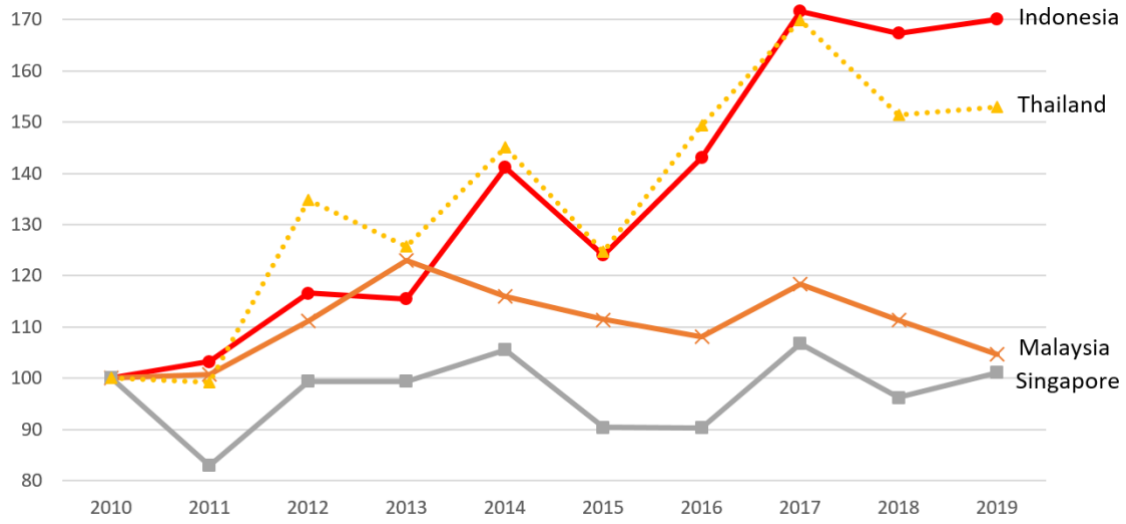
1. Introduction

Financial investment in stock markets in emerging economies has played a more role in wealth management. Chien et al. (2015) pointed out that the stock markets in the Association of Southeast Asian Nations (ASEAN) have become one of the attractive choices for investment. However, stock exchange indices usually go up and down, as examples in major ASEAN stock markets shown in Table 1. Also, different economies have different key economic drivers. Thus, each stock market index may be mainly determined by different factors. Therefore, it would be useful for professionals in the capital market, such as fund managers and investment analysts, to select key drivers of stock market indices and use these factors to predict stock exchange indices more accurately.

Fortunately, besides econometric approaches, there has been a significant development of machine learning techniques to forecast dependent variables. In financial applications, the Support Vector Machine (SVM) and the Artificial Neuron Network (ANN) have been powerful techniques for stock index prediction. Several studies proved their predictive ability. For example, Kara et al. (2011) indicated that the average accuracy of ANN and SVM models were greater than 70% in predicting the direction of the stock price index. Shrimalve and Talekar (2018) findings also confirmed that SVM and ANN models provided meaningful performance results for stock price movement prediction.

However, several studies indicated that SVM seems to outperform ANN. For example, SVM can find a global optimum, whereas ANN may only reach a local optimum (Arora et al., 2010). In addition, SVM is less sensitive to the training ratios (Ren, 2012). Furthermore, Arora et al. (2010) and Emir et al. (2012) pointed out that models based on fundamental and technical variables using the SVM provided more reliable predictions than using the ANN.

Figure 1: Normalized stock exchange indices in the ASEAN-4 countries (2010 = 100)



Source: CEIC Data

Besides those mentioned above, there have been no other studies using the SVM with causality testing to investigate and compare stock market predictions among leading Southeast Asian countries. Hence, this study chose the SVM technique and also applied causality testing to select appropriate economic variables to predict the stock market movement of the ASEAN-4: Indonesia, Malaysia, Singapore, and Thailand. Consequently,

if this SVM approach could provide robust results, i.e. above 60% accuracy, investors would be able to use this algorithm to predict the trends (up and down) of one month ahead of the stock markets and gain excess returns in the stock market.

2. Literature review

The SVM algorithm was introduced by Vapnik and Chervonenkis (1964). This approach was categorised as supervised learning, i.e., some of the data are used as a training data set, whereas the rest of the information is employed to test the algorithm's accuracy. The SVM algorithm is to predict categorical class labels, i.e., discrete values, not continuous values. Also, relationships among variables are likely to be nonlinear. Hence, the Kernel function was incorporated in SVM to improve performance in predicting nonlinear relationships by Boser et al. (1992).

There have been several empirical studies about using SVM to predict stock index movement. For instance, Anwar and Ismal (2011) and Usmani et al. (2016) compared the accuracy between the SVM and the ANN using economic variables to predict the movement of stock exchange indices. They found that the ANN and the SVM provided useful results. However, these studies had no process of selecting explanatory variables. Some researchers enhanced machine learning by adding a process of selecting variables. For example, Wang (2014) employed the principal component analysis (PCA) to select independent variables for predicting the stock market indices of Korea and Hong Kong and found that the PCA helped his SVM and ANN models improve the accuracy. Furthermore, Grigoryan (2016) used the independent component analysis (ICA) to select technical variables to forecast the Bucharest stock exchange index. He found that the ICA helped the SVM model reduce the root mean square error (RMSE).

In addition, Lahmiri (2011) initiated the Granger (1969) causality test to select economic and technical variables. His findings confirmed that the SVM model based on economic information could predict the S&P 500 index movement with 64% accuracy, even though some previous studies indicated that most stock prices follow a random walk process. However, the Granger testing method might encounter spurious regression and lead to wrong conclusions; hence Dritsaki (2017) suggested the Toda-Yamamoto causality test examine the candidates of variables. The Toda-Yamamoto causality test could also address a problem that variables are of different orders of integration (Dutta, 2018), i.e., a mixture of I(0), I(I), and I(2).

Most commonly used economic and financial indicators in testing relationship with stock market indices were exchange rate (Khan et al., 2017), interest rate (Chen et al., 2005), inflation rate (Hsing, 2011; Hussainey and Ngoc, 2009), money supply (Forson and Janrattanagul, 2014; ChaengKham and Wianwiwat, 2021), industrial production index (Rasiah and Ratneswary, 2010), oil and gold prices (Arfaoui and Rejeb, 2017; Gokmenoglu and Fazlollahi, 2015), and price to earnings ratio (Aras and Yilmaz, 2008; Sadeghzadeh, 2018). However, tourism and automobile industries, which have played a significant role in economic growth, have not been found in their relationships with the stock market index. Thus, this kind of study should also include the two sectors' indicators as candidates for explanatory variables for predicting the stock market movement.

3. Methodology

3.1 Variables and Data

In this study, stock market indices (SI) of the ASEAN-4 nations comprising Thailand, Singapore, Malaysia, and Indonesia were the dependent variables. On the other hand, potential economic and financial variables selected were set to be explanatory

variables. The 11 candidates of explanatory variables discussed in the literature review included manufacturing production index (MPI), the number of tourists (NT), motor sales (MS), export value (EX), import value (IM), broad money (BM), price to earnings ratio (PE), exchange rate against USD (XCR), crude oil price (OIL), consumer price index (CPI), and gold price (GOLD). Furthermore, this study assumed a double-log functional form; all variables are in log form.

The monthly data used in this study were collected from reliable sources such as International Monetary Fund (IMF), Intercontinental Exchange (ICE), The Bullion Desk, CEIC Data, and central banks and government agencies in the ASEAN-4 nations. Each time-series data was sampled from January 2002 to December 2019, i.e., there were 216 observations.

3.2 Method

The first step was to investigate the data's stationarity using the three forms of the Augmented Dickey-Fuller (ADF) test: intercept and trend, intercept only, and no intercept (Dickey and Fuller, 1981) via Eviews software. At the ADF testing step, we also obtained the maximum integration order (d_max) for each series. Then we chose explanatory variables for the SVM algorithm by conducting a Toda-Yamamoto causality analysis using the augmented VAR (k+d_max) model, i.e., Equation 1 as follows:

$$\ln(SI_{c,t}) = \alpha_c + \sum_{i=1}^{k} \beta_{c,i}^{1} \ln(SI_{c,t-i}) + \sum_{i=k+1}^{d_{max}} \beta_{c,i}^{2} \ln(SI_{c,t-i}) + \sum_{X=1}^{11} \sum_{i=1}^{k} \varphi_{n,c,i}^{1} \ln(X_{n,c,t-i}) + \sum_{X=1}^{11} \sum_{i=k+1}^{d_{max}} \varphi_{n,c,i}^{2} \ln(X_{n,c,t-i}) + u_{c,t} \quad (1)$$

Where $SI_{c,t}$ is the stock market index of each country (c); $X_{c,n,t-i}$ is the explanatory variable (variable n, country c, and time t-i); α_c, $\beta_{c,i}^{1}$, $\beta_{c,i}^{2}$, $\varphi_{n,c,i}^{1}$, and $\varphi_{n,c,i}^{2}$ are estimated parameters; and $u_{c,t}$ is the error term; and k is the optimal time lag on the standard vector autoregressive (VAR) model. We selected the optimal time lag (k) using the Final Prediction Error (*FPE*) criterion (Akaike, 1969). However, in the case that the residual of the standard VAR model encounters the serial correlation, a longer lag length (k) with no serial correlation is investigated by the VAR residual serial correlation LM test.

Equation 1 was estimated by the seemingly unrelated regression (Dagher and Yacoubian, 2012). Then, the testing hypothesis of the causality was examined by the Chi-Square ($\chi 2$) distribution, i.e., modified Wald test. When the null hypothesis of $\varphi_{n,c,i}^{1} = 0$ for all i is rejected, we can conclude that the explanatory variable causes the stock market index.

After the Toda-Yamamoto causality test was conducted to select explanatory variables for the algorithm, the stock market index movement (yt) was set to be 1 or "up" when yt = ln(SIt)-ln(SIt-1) $\geq$ 0 and set to be -1 or "down" when yt = ln(SIt)-ln(SIt-1) $\leq$ 0. The data was then separated into 80 % for the training set and 20 % (43 observations) for the test set. In addition, we designed to use previous months' economic data to predict the stock market movement one month ahead, i.e., all explanatory variables were lagged one month. Then the kernel SVM was conducted by the e1071 package on RStudio (Meyer et al., 2019). This study chose two types of kernel functions: the linear kernel and the sigmoid kernel for a comparison.

4. Results and Discussion

4.1 Unit Root Test

Table 1 illustrates the unit root test and significantly confirms that all candidate variables become stationary after first-differencing, i.e., I(1), except ln(MA_MPI), ln(MA_PE), ln(IN_PE), ln(SG_PE,) ln(TH_MPI), and ln(TH_PE) which are stationary at level, that is, I(0). Thus, in all the four stock markets, the maximum integration order (d_{max}) is set equal to 1 for the Toda-Yamamoto causality test.

Table 1: Unit root test

1. Variables for each country							
Malaysia		**Indonesia**		**Singapore**		**Thailand**	
Variable	**I(d)**	**Variable**	**I(d)**	**Variable**	**I(d)**	**Variable**	**I(d)**
ln(MA_SI)	I(1)*	ln(IN_SI)	I(1)*	ln(SG_SI)	I(1)*	ln(TH_SI)	I(1)*
ln(MA_BM)	I(1)*	ln(IN_BM)	I(1)*	ln(SG_BM)	I(1)*	ln(TH_BM)	I(1)*
ln(MA_CPI)	I(1)*	ln(IN_CPI)	I(1)*	ln(SG_CPI)	I(1)*	ln(TH_CPI)	I(1)*
ln(MA_EX)	I(1)**	ln(IN_EX)	I(1)*	ln(SG_EX)	I(1)*	ln(TH_EX)	I(1)*
ln(MA_IM)	I(1)*	ln(IN_IM)	I(1)*	ln(SG_IM)	I(1)*	ln(TH_IM)	I(1)*
ln(MA_MS)	I(1)*	ln(IN_MS)	I(1)*	ln(SG_MS)	I(1)*	ln(TH_MS)	I(1)*
ln(MA_MPI)	I(0)**	ln(IN_MPI)	I(1)*	ln(SG_MPI)	I(1)*	ln(TH_MPI)	*I(0)***
ln(MA_NT)	I(1)*	ln(IN_NT)	I(1)*	ln(SG_NT)	I(1)*	ln(TH_NT)	I(1)*
ln(MA_PE)	I(0)*	ln(IN_PE)	I(0)*	ln(SG_PE)	*I(0)****	ln(TH_PE)	I(0)*
ln(MA_XCR)	I(1)*	ln(IN_XCR)	I(1)*	ln(SG_XCR)	I(1)*	ln(TH_XCR)	I(1)*
2. Variables for all the countries							
Variable	**I(d)**						
Ln(GOLD)	I(1)*						
Ln(OIL)	I(1)*						

Note: Asterisks *, **, and *** indicate significance at 1%, 5%, and 10% level, respectively. All series are tested with intercept and trend form except ln(SG_PE) and ln(TH_MPI), which are tested with intercept.
Source: Authors' calculations.

Table 2: Probability values of Toda-Yamamoto causality test

Variable	**ln(MA_SI)**	**ln(IN_SI)**	**ln(SG_SI)**	**ln(TH_SI)**
ln(BROM)	0.0515***	0.1907	0.0887***	0.0631***
ln(CPI)	0.1834	0.5551	0.0699***	0.2042
ln(EX)	0.6378	0.9189	0.3834	0.3657
ln(IM)	0.6899	0.2666	0.1762	0.2062
ln(MPI)	0.0336**	0.8357	0.3190	0.1746
ln(MS)	0.1969	0.4522	0.1682	0.7442
ln(NUMT)	0.9176	0.0121**	0.0982***	0.0131**
ln(PE)	0.9858	0.9179	0.2227	0.5962
ln(XCR)	0.1078	0.0722***	0.1925	0.1661
ln(GOLD)	0.4436	0.9496	0.0764***	0.4334
ln(OIL)	0.8052	0.4156	0.2304	0.0282**
d_{max}	1	1	1	1
Lag *k* suggsted by FPE	1	2	2	6
Lag *k* suggsted by LM test	2	2	2	6
$k + d_{max}$	3	3	3	7

Note: Asterisks ** and *** indicate significance at 5% and 10% level, respectively.
Source: Authors' calculations.

4.2 Toda-Yamamoto causality test

Table 2 indicates that, in the case of Malaysia, the variable MPI can reject the null hypothesis at the 5% significance level, while BM significantly causes the Malaysian stock index at the confidence level of 90 %. Thus, these two variables are selected for the SVM algorithm to predict the Malaysian stock index movement. In addition, the variables: NT and XCR are selected in the prediction model as they significantly direct the Indonesian stock market at the confidence level of 95 % and 90%, respectively. Also, the causality test suggests that the variables: BM, CPI, NT, and GOLD can explain the Singaporean stock index movement at the confidence level of 90 % and should be included in the SVM model. In the instance of Thailand, the variables: NT and OIL can reject the null hypothesis at the 5% significance level, while BM can explain the Thai stock market at the confidence level of 90%. Thus, these three variables are selected for predicting the Thai stock market movement.

4.3 SVM Prediction

The results from the linear kernel and the sigmoid kernel SVM models are illustrated in Table 3. The linear kernel SVM outperforms the sigmoid kernel SVM. The former has an accuracy ranging from 58.14 % to 65.12 %, while the latter performs with an accuracy ranging from 41.86 % to 65.12 %. The findings provide meaningful and useful results as previous studies, e.g., Lahmiri (2011), Kara et al. (2011), Usmani et al. (2016), and Shrimalve and Talekar (2018).

Thus, this finding has challenged the semi-strong form of Efficient Market Hypothesis (EMH) pioneered by Fama (1970), who suggested that current security prices fully reflect all publicly available information, so investors cannot gain excess returns from lagged variables. Therefore, based on the EMH, we can conclude that the stock markets of Singapore (58.14%) and Malaysia (58.14%) are the most efficient among the ASEAN-4 stock markets, followed by Thailand (60.47%) and Indonesia (65.12 %).

Furthermore, based on the average monthly returns simulated by the SVM models, investors could use the linear kernel SVM to earn more returns from stock markets in Thailand (1.06%) and Malaysia (0.89%), whereas the sigmoid kernel SVM would be preferable to boost portfolio returns in the case of the Indonesian (3.35%) and Singapore (0.96%) stock markets.

Table 3: SVM prediction for the ASEAN-4 stock markets

Country	**Linear kernel**				**Sigmoid kernel**			
	Right	Wrong	Accuracy	Monthly return	Right	Wrong	Accuracy	Monthly return
Malaysia	25	18	58.14 %	0.89%	19	24	44.19 %	0.53%
Indonesia	28	15	65.12 %	0.60%	28	15	65.12 %	3.35%
Singapore	25	18	58.14 %	-0.30%	25	18	58.14 %	0.96%
Thailand	26	17	60.47 %	1.06%	18	25	41.86 %	0.18%

Source: Authors' calculations.

5. Conclusion

In summary, this article aims to employ the SVM algorithm and apply the Toda-Yamamoto causality test to select appropriate economic variables to predict the stock market movement of the ASEAN-4: Indonesia, Malaysia, Singapore, and Thailand. Candidates of explanatory variables in this study include manufacturing production index (MPI), the number of tourists (NT), motor sales (MS), export value (EX), import value (IM), broad money (BM), price to earnings ratio (PE), exchange rate against USD (XCR), crude oil price(OIL), consumer price index (CPI), and gold price(GOLD). All the variables are transformed in log form. Time series data are sampled, ranging from January 2002 to December 2019.

Based on the ADF test, most variables are stationary at the first difference, while some variables are stationary at level. Hence, the maximum order integration is set to 1 for the Toda-Yamamoto causality test. According to the causality test, MPI and BM are selected to explain the Malaysian stock market movement; NT and XCR determine the Indonesia stock market; BM, CPI, NT, and GOLD directed the Singaporean stock market movement; and the Thai Stock Exchange index could be explained by BM, NT, and OIL.

The SVM models in this study could predict the movement of the stock markets a month ahead. The linear kernel SVM had accuracy ranging from 58.14 % to 65.12 %, which outperforms the sigmoid kernel SVM. The finding concludes that based on the efficient market theory, the stock markets of Singapore and Malaysia, the most developed nations in ASEAN, are the most efficient, i.e., least accurate, among the ASEAN-4 stock markets. At the same time, investors could strategically employ the linear kernel SVM algorithm to gain excess returns from stock markets in the ASEAN-4 countries, particularly Indonesia (highest accuracy and return) and Thailand.

Reference

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann Inst Stat Math, 21*, 243–247.

Anwar, S., & Ismal, R. (2011). Robustness analysis of artificial neural networks and support vector machine in making prediction. In *Proceedings of the 2011 IEEE Ninth International Symposium on Parallel and Distributed Processing with Applications*. (pp. 256–261). Washington, DC: IEEE Computer Society.

Aras, G., & Yilmaz, M. K. (2008). Price-earnings ratio, dividend yield, and market-to-book ratio to predict return on stock market: Evidence from the emerging markets. *Journal of Global Business and Technology*, *4*(1), 18–30.

Arfaoui, M., & Ben Rejeb, A. (2017). Oil, gold, US dollar and stock market interdependencies: A global analytical insight. *European Journal of Management and Business Economics*, *26*(3), 278–293.

Arora, S., Bhattacharjee, D., Nasipuri, M., Malik, L., Kundu, M., & Basu, D.K. (2010). Performance comparison of SVM and ANN for handwritten Devnagari character recognition. *International Journal of Computer Science Issues*, *7*(3), 1-10.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. (pp. 144–152). New York: Association for Computing Machinery.

Chen, M.-H., Kim, W. G., & Kim, H. J. (2005). The impact of macroeconomic and non-macroeconomic forces on hotel stock returns. *International Journal of Hospitality Management*, *24*(2), 243–258.

Chiu, D. Y., & Chen, P. J. (2009). Dynamically exploring internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm. *Expert Systems with Applications, 36*(2), 1240-1248.

Dagher, L., & Yacoubian, T. (2012). The causal relationship between energy consumption and economic growth in Lebanon. *Energy Policy*, *50*, 795–801.

Dickey, D., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, *49*(4), 1057–1072.

Dutta, A. (2018). Implied volatility linkages between the U.S. and emerging equity markets: A note. *Global Finance Journal*, *35*, 138–146.

Emir, S., Dincer, H., & Timor, M. (2012). A stock selection model based on fundamental and technical analysis variables by using artificial neural networks and support vector machines. *Review of Economics and Finance, 2,* 106-122.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

Forson, J. A., & Janrattanagul, J. (2014). Selected macroeconomic variables and stock market movements: Empirical evidence from Thailand. *Contemporary Economics*, *8*(2), 154–174.

Gokmenoglu, K. K., & Fazlollahi, N. (2015). The interactions among gold, oil, and stock market: Evidence from S&P500. *Procedia Economics and Finance*, *25*, 478–488.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438.

Grigoryan, H. (2016). A stock market prediction method based on Support Vector Machines (SVM) and Independent Component Analysis (ICA). *Database Systems Journal*, *7*(1), 12–21.

Hsing, Y. (2011) The stock market and macroeconomic variables in a BRICS country and policy implications. *International Journal of Economics and Financial Issues, 1*(1), 12-18.

Hussainey, K., & Ngoc, L. K. (2009) The impact of macroeconomic indicators on Vietnamese stock prices. *The Journal of Risk Finance, 10*(4), 321-332.

Kara, Y., Boyacioglu, M. A., & Baykan, O. K. (2011) 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*, *38*(5), 5311-5319.

Khan, M. K., Teng, J. Z., Pervaiz, J., & Chaudhary, S. K. (2017). Nexuses between economic factors and stock returns in China. *International Journal of Economics and Finance*, *9*(9), 182–191.

Lahmiri, S. (2011). A comparison of PNN and SVM for stock market trend prediction using economic and technical information. *International Journal of Computer Applications*, *29*(3), 24–30.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch,F., & Chang, C. (2019). Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Retrieved from https://CRAN.R-project.org/package=e1071.

Rasiah, V., & Ratneswary, R. (2010). Macroeconomic activity and the Malaysian stock market: Empirical evidence of dynamic relations. *The International Journal of Business and Finance Research*, *4*(2), 59–69.

Ren, C. (2012). ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging. *Knowledge-Based Systems, 26*, 144-153.

Sadeghzadeh, K. (2018). The effects of microeconomic factors on the stock market: A panel for the stock exchange in Istanbul ARDL analysis. *Theoretical and Applied Economics*, *25*(3), 113–134.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shrimalve, H. H., & Talekar, S.A. (2018). Comparative analysis of stock market prediction system using SVM and ANN. *International Journal of Computer Applications*, *6*(2), 59–64.

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, *66*(1), 225–250.

Usmani, M., Adil, S. H., Raza, K., & Ali, S. S. A. (2016). Stock market prediction using machine learning techniques. *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*. (pp. 322–327). Washington, DC: IEEE Computer Society.

Vapnik, V., & Chervonekis, A., 1964. A note on one class of perceptrons. *Automation and Remote Control, 25*, 112-120.

Wang, Y. (2014). Stock price direction prediction by directly using prices data: An empirical study on the KOSPI and HIS. *International Journal of Business Intelligence and Data Mining*, *9*(2), 145-160.

Chaengkham, S., & Wianwiwat, S., 2021. The impacts of macroeconomic and financial indicators on stock market index: Evidence from Thailand. *International Journal of Trade and Global Markets, 14*(2), 197-205.