Comparison of Forecasting Models for Banking Stock: Multiple Linear Regression and Artificial Neural Network

Pichit Boonkrong[1*], Nithipa Arjrith[1] and Junwei Yang[2]

[1]Department of Mathematics,
[2]Department of Information and Communication Technology,
College of Digital Innovation Technology, Rangsit University, Pathum Thani 12000, Thailand
[*]Corresponding author, E-mail: pichit.bk@rsu.ac.th

Abstract

In this paper, two machine learning algorithms including multiple linear regression and artificial neural network are employed as forecasting models for the banking stock of Bangkok Bank Public Company Limited, Thailand. Five predictors including the SET50 Index, Barrick Gold stock price, exchange rate of US dollar, Down Jones Industrial Average, and crude oil price from West Texas Intermediate are taken into account. The historical time series for response and predictors were collected from 238 days from January 1 to December 31, 2020. The sizes of training and test datasets are 162 and 76 (70%: 30%). At the significance level of 0.01, the correlation between the response and each predictor is significantly found. Then, the datasets are recruited into the multiple linear regression and neural network models. Only three predictors including SET50 Index, Barrick Gold Stock, and Down Jones are not removed from the multiple linear regression model whereas all five predictors can be added up into the neural network model. Measuring the model performance by root mean square error found that the neural network is much better than the multiple linear regression model, in which the root mean square error are 0.03874 and 2.53624, respectively. As a result, this paper claims that the signals from SET50 Index, Barrick Gold Stock, and Down Jones are important for making a decision in trading on the stock of Bangkok Bank.

Keywords**:** *Banking stock, Forecasting model, Machine learning, Multiple linear regression, Neural network, Time series*

1. Introduction

Regarding the rapid growth of the world commodity trade, the massive influence of the stock market's performance on the national economy has become the crucial factor behind the macroeconomic overview. The investment in finance attracts lots of attention from the investors to make a profit so that they could take advantage of the financial market using appropriate forecasting models and significant economic factors. Recently, the stock trading in the banking sector spontaneously has been accordingly expanded because of the rapid development of information technology and internet banking.

Forecasting stock price index, various financial time series and forecasting models have been together applied to simulate the future values of an interesting stock price from their historical values. However, many unpredictable factors are affecting the trend of each stock and the return is proportional to the risk, i.e., the investors have to carry a higher risk if they would like to gain more return. Typically, the relationship among different financial time series is early investigated for forecasting model using exploratory factor analysis. Focusing on multi-factor models for stock prediction, input variables are predictors and the output variable is response or target. By the concept in machine learning for a data-driven modeling approach, multiple linear regression (MLR) and artificial neural network (ANN) are popularly voted as effectively applicable algorithms (Gao et al., 2020; Han & Nordin, 2017; Ivanovski et al., 2016; Qi, 1996; Moghaddam et al., 2016; Vijh et al., 2020). Since MLR can appropriately reduce the number of predictors, it is widely employed in many studies on the assessment, prediction, and evaluation of stock price index (Boonkrong & Arjrith, 2018; Jia et al, 2019; Nivetha & Dhaya, 2017). Using the MLR model to forecast the stock price, the four key indicators for the global market including gold price, oil price, USD exchange rate, and Down Jones index have been identified as significant predictors (Jeffres & Atkin, 1996; Meade & Islam, 2015; Sutheebanjard & Premchaiswadi, 2010). In particular with the stock market in Thailand, the SET index is often taken into account because it is a composite index representing the price movement for all common stocks in Thailand (Arjrith & Boonkrong, 2019, Boonkrong et al, 2020). Nevertheless, the relationship

between predictors and target variable is not necessary to be linear, which is the limitation of the MLR model. To model the stock price with a nonlinear relationship, the ANN model is a well-established tool rather than the MLR model. Forecasting financial time series using the ANN model presents good predictive accuracy as ANN can decrease errors between the real output and expected output (Gao et al., 2020; Moghaddam et al., 2016; Qi, 1996; Qiu et al., 2020; Vijh et al., 2020). It is still interesting to investigate the difference, advantages, and disadvantages of both MLR and ANN models for stock prediction. By the aforementioned evidence, this paper focuses on how MLR and ANN models are practically implemented for stock price prediction.

The overall presentation of this paper is organized as follows. To have an initial insight about the dataset and consider the possible model formation, descriptive statistics and exploratory data visualization are carried out and presented in Section 2. The implementation of MLR and ANN models is conceptually illustrated in Section 3. Subsequently, the empirical results obtained from the use of an available dataset via the formulated models are numerically and visually exhibited in Section 4. Comparing the effectiveness between MLR and ANN models, the Root Mean Square Error is measured for each model and used as the evaluation tool whether one predictive model is better. At the end of this paper, some conclusion is given in Section 5.

2. Data Visualization

The main scope of this paper is to explore the relationship between the response and the predictors. Since banking stocks are a stable business with regular pay dividends and highly flexible financial conditions, they are among the first stocks that investors tend to think of. In Thailand, there are interesting shareholdings of five major banks including Bangkok Bank, Siam Commercial Bank, Kasikorn Bank, Krung Thai Bank, and Bank of Ayudhya. In this study, the banking stock of Bangkok Bank Public Company Limited (BBL) is the response or output because it is the first member of the Stock Exchange of Thailand and one of the leading securities companies in Thailand. Bualuang Securities is a subsidiary of Bangkok Bank, with Bangkok Bank Public Company Limited as the major shareholder to operate a capital market business. Corresponding with the literature reviews, the predictors or the inputs are SET, GOL, USD, DJI, and WTI. Before starting statistical data analysis, data visualization is performed to allow better observation. As presented in Table 1, the description and descriptive statistics of all variables used in this study are summarized.

Table 1 Descriptions and descriptive statistics of variables used in MLR model

Variables	Descriptions	Unit	N	Max	Min	Mean±SD
BBL	The stock of Bangkok Bank	TH Baht	238	163.50	88.00	113.66±18.04
SET	SET50 Index	-	238	1081.22	680.07	880.28±90.07
GOL	Barrick Gold Price	US Dollar	238	30.46	15.67	24.34±3.96
USD	Exchange Rate of US Dollar	TH Baht	238	33.11	29.75	31.26±0.77
DJI	Down Jones Industrial Average	-	238	30393.04	18591.93	26878.45±2509.48
WTI	Crude Oil Price from West Texas Intermediate	US Dollar	238	62.64	10.07	38.31±12.16

Investigating the relationship between BBL to others, the scatterplots and Pearson correlations are employed. As can be seen from Table 2 and the scatterplots in Figure 1, BBL presents its linear relationship to each predictor. At the significance level of 0.01 (2-tailed), there is a large positive correlation between BBL and SET (R= 0.912), a medium negative correlation between BBL and GOL (R= 0.632), a weak negative correlation between BBL and USD (R= -0.493), a weak positive correlation between BBL and DJI (R= 0.495) and a high positive correlation between BBL and WTI (R= 0.716).

Table 2 Descriptions of variables used in MLR model

Pearson Correlation	SET	GOL	USD	DJI	WTI
BBL Stock	0.912**	-0.632**	-0.493**	0.495**	0.716**
Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000
N	238	238	238	238	238

**Correlation is significant at the 0.01 level.

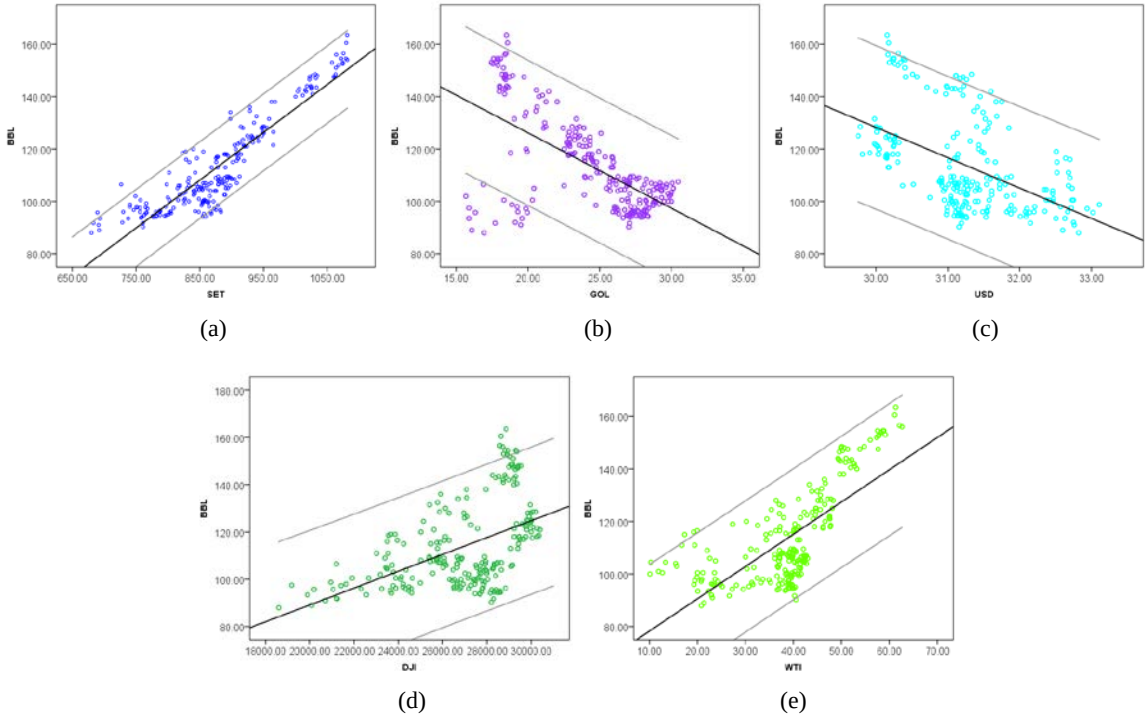


Figure 1 The scatterplot showing the relationship between BBL to each predictor

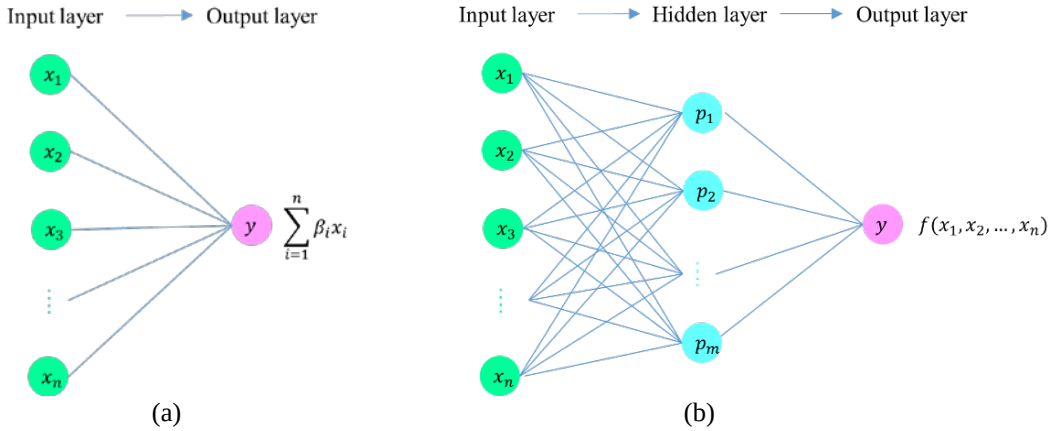


Figure 2 The network diagram of MLR and ANN with one hidden layer

3. Model Implementation

Machine learning algorithms can be applied to almost all data problems. The machine is trained to make a specific decision by historical data. There are two basic models of machine learnings used in this study including MLR and ANN models. As can be seen in Figure 2, the MLR is equivalent to the NN models without a hidden layer. Forming up models, the assumption is started by setting BLB as the response variable whilst GOL, OIL, SET, DJI, and USD as the explanatory variables or predictors. The dataset is classified into two groups including training and testing groups. The proportion between training and testing groups is 162: 76 (70%: 30%). The training data is initially used to train the algorithm, which is used to fit the parameters for the model during the learning process. Then, the validation by testing data is specifically identified for use in tests. Evaluating model performance, the Root Mean Square Error (RMSE) is popularly used as a performance indicator, meaning that the better model shows the smaller RMSE value.

3.1 Multiple Linear Regression Model

The multiple linear regression model (MLR) is one of the most popular statistical models for studying the relationship between independent and dependent variables. The standard form of such a relationship is expressed by the expression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \ . \tag{1}$$

Without considering the intercept β_0, the linear regression model is written as the linear combination where the coefficient β_i is called weight for each predictor x_i.

3.2 Neural Network Model

Based on simple mathematical models of the brain, an artificial neural network is applied to allow complex nonlinear relationships between the response variable and its predictors. Unlike multiple linear regression, the hidden layer is added up between the input layer and output, that is, the inputs to each neuron are combined using a linear combination and the result is then modified by a nonlinear function before being output. As shown in Figure 2(b), the inputs into the hidden neuron *j* are formed as a linear combination such that

$$p = a_j + \sum_{i=1}^{n} \omega_{i,j} x_i \tag{2}$$

where $\omega_{i,j}$ denotes the weight for input x_i. Then, s is modified by an activation function such as sigmoid, hyperbolic tangent, and step functions. The most classic and popular activation function is the sigmoid function defined by

$$s(p) = \frac{1}{1+e^{-p}} \ . \tag{3}$$

Particularly, the number of hidden layers and neurons in each hidden layer must be specified in advance. In this study, there is one hidden layer with three neurons in the neural network model.

3.3 Evaluation of Model Performance

Comparing the performance of both MLR and ANN models, the residual and root mean square error is evaluated. The residual is the differences between observed and predicted values of data. The residual, $y_{u,i} - \hat{y}_{u,i}$, can be positive or negative. Squaring the residuals, averaging the square, and taking the square root to provide the root mean square error. Thus, the root mean square error is defined by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{u,i}(y_{u,i} - \hat{y}_{u,i})^2} \ . \tag{4}$$

RMSE shows how the prediction is against the real value and it is used as the indicator to evaluate the performance of both MLR and ANN models in this study. The better model has the lower value of RMSE.

4. Empirical Results

Dealing with statistical data analysis, both multiple linear regression and neural networks are processed by the software package called IBM SPSS Statistics version 21.0. Comparing the performance of both models, the RMSE for each model is evaluated. The model with the lowest RMSE is considered to be the best model in this study.

4.1 Multiple Linear Regression Model

Regarding the conceptual framework in Section 3.1, three multiple linear regression models are statistically examined at the significant level of 0.05. The intercept β_0 is equal to 0 and each line passes through the origin. From coefficient identification in Table 2, the MLR models are listed as follows:

- $\hat{BBL} = (0.130)SET,\ R^2 = 0.994$
- $\hat{BBL} = (0.165)SET - (1.302)GOL,\ R^2 = 0.998$
- $\hat{BBL} = (0.134)SET - (1.732)GOL + (0.001)DJI,\ R^2 = 0.998$

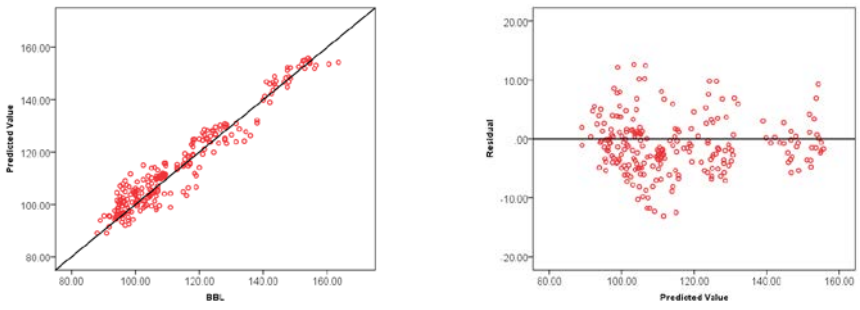
Measuring the goodness of fit in three MLR models, the correlation (R) and coefficient of determination (R^2) are given for each model. After fitting different linear regression models, both R and R^2 are approximately the same, but the standard error of estimate or RMSE of the MRL model with three predictors is the lowest one (4.59602). Therefore, the MRL model $\hat{BBL} = (0.134)SET - (1.732)GOL + (0.001)DJI$ is the best one among the three MRL models. The scatterplot of the predicted response against the actual one (BBL) is shown in Figure 3(a) and the residual plot is also shown in Figure 3(b).

Table 2 Coefficient identification for MLR models.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	SET	0.130	0.001	.997	200.654	.000
2	SET	0.165	0.002	1.269	94.304	.000
	GOL	-1.302	0.063	-.279	-20.730	.000
3	SET	0.134	0.004	1.031	34.335	.000
	GOL	-1.732	0.074	-.371	-23.377	.000
	DJI	0.001	0.000	.330	8.632	.000

a. Dependent Variable: BBL

b. Linear Regression through the Origin



(a) (b)

Figure 3 The predicted values and residuals from the multiple linear regression model

4.2 Artificial Neural Network Model

Using the same predictors as the MLR model, the ANN model has three inputs including SET, GOL, and DJI. There is one hidden layer with three neurons and BBL is still the output. The structural diagram of the ANN model is shown in Figure 4. The sigmoid function is set to be the activation function for hidden and output layers. However, it is still questioned how USD and WTI are removed from the MLR model as they also have a significant correlation with BBL. Thus, both USD and WTI are added to the ANN model again to see whether the precision of the previous ANN model is improved. The structural diagram of the new ANN model is shown in Figure 5. Running multilayer perceptron of neural network in SPSS, estimation of parameters is correspondingly given in Tables 3 and 4 for ANN with three and five inputs, respectively. The activation function for hidden and output layers is assigned to be a sigmoid function.

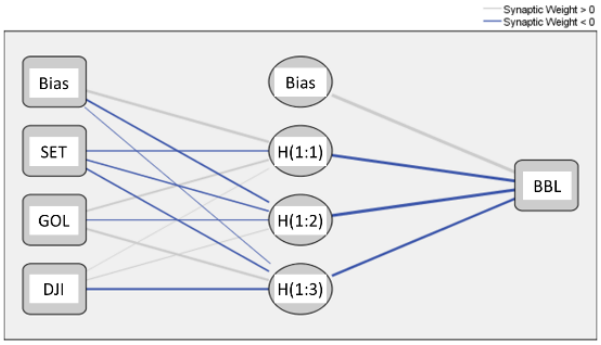


Figure 4 Structure of neural network with three inputs (SET, GOL, DJI) and one output (BBL).

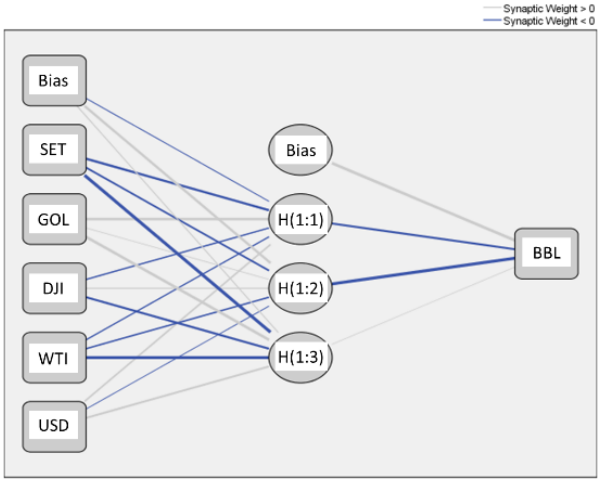


Figure 5 Structure of neural network with five inputs (SET, GOL, DJI, WTI, USD) and one output (BBL).

The dataset is randomly assigned into training (70%) and testing subsets (30%) by the hold-out method, in which the numbers of training and testing data are 162 and 78. The training dataset is used to evaluate the weights and formulate the ANN model whereas the testing dataset is used to validate the ANN model. Corresponding with the ANN model with three predictors, the scatterplot of the predicted response against the actual one (BBL) is shown in Figure 6(a) and the residual plot is also shown in Figure 6(b). For the ANN model with five predictors, the scatterplot of the predicted values and the residual plot are shown in Figures 7(a) and 7(b).

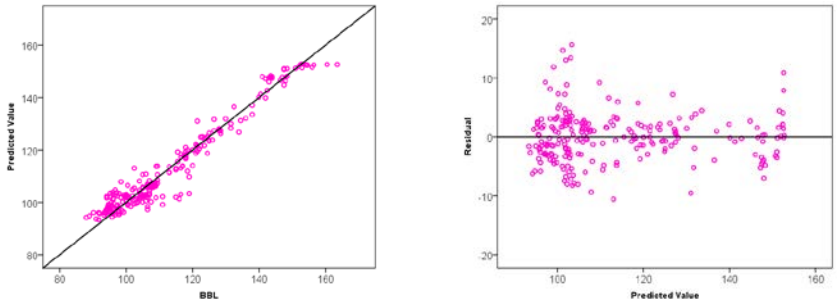
Table 3 Parameter estimates for ANN model with three inputs.

Predictor		Hidden Layer 1			Output Layer
		H(1:1)	H(1:2)	H(1:3)	BBL
Input Layer	(Bias)	-1.000	1.107	-0.091	
	SET	-0.674	-0.904	-0.575	
	GOL	0.662	0.539	0.136	
	DJI	-0.980	0.479	-0.129	
Hidden Layer 1	(Bias)				2.589
	H(1:1)				-1.546
	H(1:2)				-2.221
	H(1:3)				-2.610

Table 4 Parameter estimates for ANN model with five inputs.

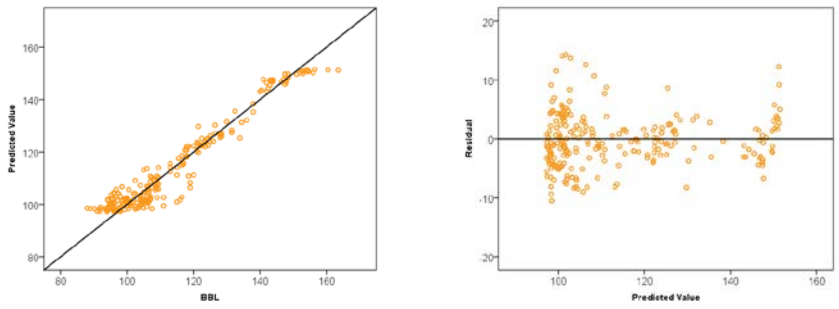
Predictor		Hidden Layer 1			Output Layer
		H(1:1)	H(1:2)	H(1:3)	BBL
Input Layer	(Bias)	-0.262	2.325	0.335	
	SET	-2.159	-1.042	-3.322	
	GOL	2.390	0.123	2.510	
	DJI	-0.718	0.477	-1.788	
	WTI	-0.579	-0.754	-2.922	
	USD	0.983	-0.157	1.374	
Hidden Layer 1	(Bias)				2.841
	H(1:1)				-1.410
	H(1:2)				-3.299
	H(1:3)				0.053

After running each ANN model for 10 iterations, parameter estimates and RMSE are averaged and compared. The performance for each model is shown in Table 2 by its RMSE, in which the better model gives the smaller value of RMSE. The traditional MLR model gives higher RMSE than the ANN model. The best prediction from the MLR model gives RMSE of 2.53624 whilst the ANN model gives only 0.03874. Using the same three predictors as the MLR model, the ANN model shows a slightly different outcome.



(a) (b)

Figure 6 The predicted values and residuals from ANN model with three inputs



(a) (b)

Figure 7 The predicted values and residuals from the ANN model with five inputs

Table 2 Evaluation of model performance by RMSE

RMSE	Predictors	SET	SET & GOL	SET, GOL & DJI	SET, GOL, DJI, WTI & USD
MLR Model	Training (n = 162)	8.82155	5.26343	4.59602	-
	Testing (n = 76)	4.69378	2.77367	2.53624	-
ANN Model	Training (n = 162)	0.06074	0.04548	0.04151	0.04039
	Testing (n = 76)	0.04427	0.04311	0.04055	0.03874

5. Conclusion

Forecasting the banking stock of Bangkok Bank Public Company Limited (BBL), five predictors including SET, GOL, DJI, WTI, and USD are taken into account. A comparison between two forecasting models including MLR and ANN has been made. The statistical and graphical evidence is carried out. Initially, the correlation between BBL and all five predictors is examined by Pearson correlation. It is found that all five predictors are statistically related to the response BBL. Subsequently, the MLR and ANN models are formulated and the datasets are fed into both models. Each data array is split into training and testing subsets. The training subset with 162 data points is used to figure out the weights for each model whilst the testing subset with 76 data points is used for model validation. In each round of the process, the training and testing subsets are randomly assigned. Both machine algorithms have been performed ten times and the numerical results including weights and RMSEs are then averaged. Considering RMSE values, the ANN model is better than MRL for all cases. Three predictors including SET, GOL, and DJI are kept in the MLR model in which its RMSE is 2.53624, but RMSE is 0.04055 in the ANN model. By the fact that BBL stock is included in SET, they always have a linear relationship to each other. To improve the accuracy of the ANN model, all five predictors including SET, GOL, DJI, WTI, and USD are taken into consideration. As a consequence, the updated ANN model has only an RMSE of 0.03874 showing the best numerical result in this study. However, the RMSEs from the ANN model with three and five predictors do not even show a large difference since MLR has accordingly reduced the number of predictors. As a result, the former three predictors (SET, GOL, and DJI) are more important than the latter two predictors (WTI and USD). In other words, it is remarked that the signals from SET, GOL, and DJI should be considered when trading in the banking stock of Bangkok Bank Public Company Limited.

Typically, machine learning is used for making a prediction based on historical data. Predicting how the stock market will perform involves many factors so that it is one of the most difficult things to do. Due to the non-linear nature of the financial stock markets, the basic MLR model shows its limited implementation. The examples of some popular machine learnings used in stock prediction are support vector machine (SVM), random forest (RF), artificial neural network (ANN), recurrent neural network (RNN), and long short-term memory (LSTM). Still, each machine learning has its special characteristics suitable for a particular dataset.

Therefore, the challenging matter for future research in this field is to develop the new predictive model and discover how to minimize RMSE to obtain the best model performance.

6. Acknowledgements

The authors would like to gratefully acknowledge the Stock Exchange of Thailand (SET) for the available datasets used in this study. Furthermore, we would like to express our deep gratitude to the anonymous referees for their valuable suggestions to improve the manuscript.

7. References

Arjrith, N. & Boonkrong, P. (2019). Analysis of influencing factors for trading value in telecommunication company, *Proceedings of RSU Research Conference, pp. 434–439, Pathum Thani, Thailand. Retrieved from* https://rsucon.rsu.ac.th/files/proceedings/inter2019/IN19-221.pdf

Boonkrong, P. & Arjrith, N. (2018), Impact of weekdays on the return rate of stock price index: evidence from the stock exchange of Thailand, *Journal of Finance and Accounting*, *6*(1): 35–41.

Boonkrong, P., Arjrith, N. & Sangsawad, S. (2020). Multiple linear regression for technical outlook in telecom stock price, *Proceedings of RSU International Research Conference*, pp. 1178–1185, *Pathum Thani, Thailand. Retrieved from* https://rsucon.rsu.ac.th/files/proceedings/inter2020/IN20-081.pdf

Budhani, N., Jha, C. K. & Budhani, S. K. (2014). Prediction of stock market using artificial neural network, *Proceedings of 2014 International Conference of Soft Computing Techniques for Engineering and Technology (ICSCTET), pp. 1–8, Bhimtal, India. Retrieved from* https://*ieeexplore.ieee.org/abstract/document/7371196*

Gao, P., Zhang, R., & Yang X. (2020). The application of stock index price prediction with neural network. *Mathematical and Computational Applications*, 25(53), 1–16. https://doi.org/10.3390/mca2503 0053

Han, L. S., & Nordin, J. (2017). Integrated multiple linear regression–one rule classification model for the prediction of stock price trend. *Journal of Computer Sciences*, *13*(9), 422-429.

Ivanovski, Z., Ivanovska, N., & Narasanov, Z. (2016), The regression analysis of stock returns at MSE. *Journal of Modern Accounting and Auditing*, *12*(4), 217–224.

Jeffres, L. & Atkin, D. (1996). Predicting use of technologies for communication and consumer needs. *Journal of Broadcasting and Electronic Media*, *40*, 318–330

Jia, L., Wang, F., Sun, H., & Li, H. (2019). An empirical research of the impact of monetary policy on stock market returns–based on the perspective of investor structure. *Chinese Control and Decision Conference*, 4158–4163. doi: 10.1109/CCDC.2019.8833475

Meade, N., & Islam, T. (2015). Forecasting in telecommunications and ICT–A review. *International Journal of Forecasting*, *31*(4), 1105–1126.

Moghaddam, A. H., Moghaddam, M. H. & Esfandyari, M. (2016), Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, *21*(41): 89–93.

Nivetha, R. Y., & Dhaya, C. (2017). Developing a Prediction Model for Stock Analysis. *International Conference on Technical Advancements in Computers and Communications*, 1–3. doi: 10.1109/ICTACC.2017.11

Qi, M. (1996). Handbook of Statistics. *Financial applications of artificial neural networks* (pp. 529–552) Netherlands: Elsevier.

Qiu, J., Wang, B. & Zhou, C. (2020). Forecasting stock prices with long-short term memory neural network based on attention mechanism. *PLoS ONE*, *15*(1), 1–15. doi: 10.1371/journal.pone.0227222

Sharma, A., Bhuriya, D., & Singh, U. (2017). Survey of stock market prediction using machine learning approach. Paper presented at the 2017 International Conference of Electronics, Communication and Aerospace Technology, Coimbatore, India.

Sutheebanjard, P., & Premchaiswadi, W. (2010). Stock exchange of Thailand Index prediction using back propagation neural networks. Paper presented at the the 2nd International Conference on Computer and Network Technology, Bangkok, Thailand.

Vijh, M., Chandola, D., Tikkiwal, V.A., & Kumar A. (2020). Stock closing price prediction using machine learning techniques, *Procedia Computer Science*, *167*, 599–606. doi: 10.1016/j.procs.2020.03.326