Nikkei Stock Market Price Index Prediction Using Machine Learning

Lukman A Harahap¹², Rajalida Lipikorn² and Asanobu Kitamoto³

¹ Department of Agricultural Engineering, Faculty of Agriculture, University of Sumatera Utara, Medan, Indonesia
² Machine Intelligence and Multimedia Information Technology Laboratory, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand
³ Research Organization of Information and Systems, Department of Informatics, National Institute of Informatics, Tokyo, Japan

¹lukman.adlin@usu.id, ²rajalida.l@chula.ac.th, ³kitamoto@nii.ac.jp

Abstract. Stock market prediction has always been a difficult process, most of the prediction rely solely on the data of the corresponding stock market. Relationship of gold and oil price with stock market performance has been proven significant in some major world stock index. Prediction of stock market price index using machine learning methods is expected to perform well, with the ability of machine learning method to predict using nonlinear inputs. The methods were commonly able to predict relatively well in predicting the values of Japanese Nikkei 225 and Japanese Nikkei 400 indexes.

1. Introduction
In the machine learning era, prediction of many variables has been done, and claiming relative success, ranging from time series prediction to prediction based on other variables. Methods has also been developed so rapidly, starting from common Artificial Neural Network (ANN) with Multiple Layer Perceptron (MLP) to Deep Neural Network (DNN).

Prediction of stock index price has been done in long time through many methods, namely Fundamental Analysis and Technical Analysis. Fundamental analysis is done based on macroeconomic factors such as state of the economy and industry and microeconomic factors such as internal performance of the company, which in a way will be affecting stock or futures prices[1]. In short-selling strategies, fundamental analysis can be profitable, as can be used to identify overpriced stocks and the predicted the price decline thus making a profit from the stock trade[2]. Technical analysis of a stock market is commonly done by studying the market action on the stock price historically, as it is on common market goods, which is represented graphically. Technical analysis is commonly using techniques such as moving average oscillators, which is data driven[3]. In the case of Singapore Stock Market, the use of technical analysis has been proven to be able to provide significant profit[4].

With the rise of computing age, the use of machine learning in stock market price index prediction also inevitable. Some of recent researches were using machine learning method in a stock market and commodity futures price index, such as a hybrid of teaching learning based optimization and a support
vector machine[5], statistical learning models [6], deep learning based feature engineering [7], and also a modified back propagation neural networks[8].

The objective of this research was to develop a prediction system that can predict a Japanese stock market price index based on exchange rate to American dollars, gold price, American stock market price index and oil price.

2. Data Used
The relationship of stock market price with gold and oil price has been analyzed and proven by many economics experts, such as its statistically significant effect on US Stock Market[9], various effect such as significant effect of world oil price and insignificant effect of world gold price on JCI stock price[10], positive relation of gold price with Indonesia Composite Index [11], while also have various effect on different economic period, such as in the case of Frankfurt Stock Exchange on the year of 2004 to 2016, but overall, the effect of gold price has moderate positive relationship[12], crude oil market also proven to have significant volatility spillover to stock markets of G-7 countries[13]. In Japanese Stock Market case, gold and oil market has useful information on the fluctuations of Japanese macro-financial variables, including stock price and interest rate [14].

Most of the papers does not address directly to the relationship of gold and oil price to Japanese stock market price index, while addressing American stock market price index instead, thus, in a sense, using a currency exchange rate between Japanese Yen and American dollars would also reasonable to use in predicting Japanese stock market price index. The reference to American dollar as a currency also logical since the USD is currently among the strongest currency in the world, if not the strongest[15].

All of data used were in the range of 10 years, which is from 20th July 2009 to July 19th July 2019, except for Nikkei 225 index which is only 3 years, from 25th July 2016 to 19th July 2019. Data used were as follows:

2.1. Japanese Yen to American Dollar exchange rate (DEXJPUS)
The historical data of DEXJPUS were taken from Federal Reserve Bank of St. Louis website[16].

2.2. Philadelphia Gold/Silver (XAU)
The historical data of XAU were taken from investing.com website[17].

2.3. New York Stock Exchange (NYSE)
The historical data of NYSE were taken from investing.com website[18].

2.4. Crude Oil WTI Futures (CLZ9)
The historical data of OIL were taken from investing.com website[19].

2.5. Japanese Nikkei Index N225 (N225)
The historical data of N225 were taken from Nikkei 225 official site[20].

2.6. Japanese Nikkei Index N400 (JPXNK400)
The historical data of JPXNK400 were taken from investing.com website[21].

3. Methodology
3.1. Data set building and data cleaning
Data were put together into a single data set and synchronized based on the date. In total, there are 2589 data points for Nikkei 400 data set and 872 data points for Nikkei 225 data set. Each parameters has missing data, mostly due to market closing period, thus made the compiled data set also has missing data. Commonly, missing data has negative contribution to data distribution, and thus would
propagate into increase of error. In effort to reduce error, the dataset then cleaned by using only complete data set, which resulting in only 2123 data points available for Nikkei 400 data set and 698 data points available for Nikkei 225 data set.

Data set used as input are DEXJPUS, XAU, NYSE and OIL, where the target or label are N225 and N400. Data set were divided into training set and validating or testing set, with two ratios, which are 60:40 and 70:30 for training set and testing set consecutively.

3.2. Prediction using Deep Neural Network (DNN)

DNN used in the prediction has four input nodes, 3 hidden layers with 16 nodes on the first and second layer and 8 nodes in the third layer, third layer connected to a single output node. Training was done in 500 epoch. Using the same architecture, the effect of training data and testing data ratio will be observed. DNN was built using Keras framework[22].

3.3. Prediction using Back Propagation Neural Network (BPNN)

BPNN used in the prediction has four input nodes, a single hidden layer with 16 nodes, connected to a single output node. Training also done in 500 epoch.

3.4. Prediction using Support Vector Machine Regression (SVR)

SVR used in the prediction has four inputs, and one target variable, with hyper parameter configuration as follows: C value was set to 5, epsilon is using default value of 0.1 and kernel used was linear kernel. SVR used in this prediction was built using sklearn package[23].

3.5. Performance measurement

Training phase and testing phase has different set of performance measurement. In training phase, the performance will be measured using Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) for all prediction except SVR. In testing phase, the performance of all prediction was measured using Coefficient of Determination (R²).

4. Results and Discussion

The R² values of prediction result using SVR, DNN and BPNN in can be seen in Table 1 below. Training data ratio is the percentage of data used in the training phase, which is 60%, 70%, 80% and 90%, subsequently.

| Nikkei Index | Prediction Method | Training Data Ratio |
|--------------|-------------------|---------------------|
|              | 60%   | 70%   | 80%   | 90%   |
| N225         | SVR    | 67%   | 64%   | 58%   | 81%   |
|              | DNN    | 68%   | 60%   | 58%   | 79%   |
|              | BPNN   | 66%   | 46%   | 56%   | 82%   |
| N400         | SVR    | 60%   | 37%   | 19%   | 41%   |
|              | DNN    | 72%   | 46%   | 20%   | 38%   |
|              | BPNN   | 63%   | 29%   | 20%   | 66%   |

From the results, it can be seen that each prediction method is performing differently on different Japanese Nikkei Stock Market Index, which are Nikkei 225 index (N225) and Nikkei 400 index (N400). All of the values in Table 1 are derived from Coefficient of Determination (R²) values transformed into percentage form. These values defined how well the corresponding prediction
method would predict the values of Nikkei index given the input of DEXJPUS, XAU, NYSE and CLZ9 of the previous day.

For N225, most of the prediction method performs well in the ratio of 90%, where for N400, most of the prediction method performs well in the ratio of 60%, except for BPNN, which performs relatively well also in 90% ratio. Overall, in terms of N225 prediction, SVR is the best prediction method, and in terms of N400 prediction DNN is the best prediction method. To make it clear, even though the ratio is the same, but the number of dataset in the ratio are different when it is on N225 and N400, which can be seen in Table 2, so practically, 60% of data in N225 means that the network was using 420 data points, where in N400, it used 1279 data points.

| Nikkei Index | Training Data Ratio | Data Points |
|--------------|---------------------|-------------|
| N225         | 60%                 | 420         |
|              | 70%                 | 489         |
|              | 80%                 | 559         |
|              | 90%                 | 629         |
| N400         | 60%                 | 1279        |
|              | 70%                 | 1490        |
|              | 80%                 | 1701        |
|              | 90%                 | 1912        |

The difference of performances of prediction methods with difference ratio could have come from the nature of input and training labels itself, since the methods are all machine learning, which can only memorize how the output dynamically change based on the changes of input data. While large number of data points might have caused overfitting, which in turn, caused further lower accuracy in prediction.

5. Conclusion
All of the prediction methods perform differently on different ration of data points, SVR is the best method for N225 index value prediction, where DNN is the best for N400 index value prediction. SVR, DNN and BPNN performs well in N225 index value prediction with 90% training ratio of data points, where in N400 index value prediction, all three performs relatively well on 60% training ratio of data points, thus it is suggested to use SVR on predicting N225 values with 90% data ratio, and to use DNN on predicting N400 values with 60% data ratio.

6. Acknowledgement
The Authors would like to thank the support of Lembaga Pengelola Dana Pendidikan Republik Indonesia (LPDP-RI) for supporting the submission of this article.

References
[1] A. Ganti and T. Segal. (2019, 1 November). *Fundamental Analysis*. Available: https://www.investopedia.com/terms/f/fundamentalanalysis.asp
[2] P. M. Dechow, A. P. Hutton, L. Meulbroek, and R. G. Sloan, "Short-sellers, fundamental analysis, and stock returns," *Journal of Financial Economics* vol. 61, no. 1, pp. 77-106, 2001.
[3] R. D. Edwards, J. Magee, and W. H. C. Bassetti, *Technical Analysis of Stock Trends 11th Edition*. Boca Raton: Taylor & Francis Group, 2018.
[4] W.-K. Wong, M. Manzur, and B.-K. Chew, "How rewarding is technical analysis? Evidence from Singapore stock market," Applied Financial Economics, vol. 13, no. 7, pp. 543-551, 2003.

[5] S. P. Das and S. Padhy, "A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting," International Journal of Machine Learning and Cybernetics, vol. 9, no. 1, pp. 97-111, 2018.

[6] C. Liu, J. Wang, D. Xiao, and Q. Liang, "Forecasting S&P 500 Stock Index Using Statistical Learning Models," Open Journal of Statistics, vol. 6, pp. 1067 - 1075, 2016.

[7] W. Long, Z. Lu, and L. Cui, "Deep learning-based feature engineering for stock price movement prediction," Knowledge-Based Systems vol. 164, pp. 163 - 173, 2019.

[8] F. Mithani, S. Machchhar, and F. Jasdanwala, "A Modified BPN Approach For Stock Market Prediction," presented at the IEEE International Conference on Computational Intelligence and Computing Research, Chennai, India, 15-17 December 2016, 2016. Available: https://ieeexplore.ieee.org/document/7919718/citations#citations

[9] M. Rahman and M. Mustafa, "Effects of Crude Oil and Gold Prices on US Stock Market: Evidence for USA from ARDL Bounds Testing," Finance and Market, vol. 3, no. 1, pp. 1 - 9, 2018.

[10] F. H. Syarofi and H. Muhamam, "The Impact of the Domestic Interest Rates, Exchange Rate, World Oil Prices, World Gold Prices, DJIA, Nikkei 225 and HSI on the JCI," in The Second International Conference on Finance, Bali, Indonesia, 2014.

[11] D. A. Kusumawati and N. Asandimitra, "Impact of Global Index, Gold Price and Macro Economic Variable for Indonesia Composite Index," Research Journal of Finance and Accounting, vol. 8, no. 2, pp. 53 - 62, 2017.

[12] M. Al-Ameer, W. Hammad, A. Ismail, and A. Hamdan, "The Relationship of Gold Price with the Stock Market: The Case of Frankfurt Stock Exchange," International Journal of Economics and Policy, vol. 8, no. 5, pp. 357-371, 2018.

[13] R. Khalfaoui, M. Boutahar, and H. Boubaker, "Analyzing volatility spillovers and hedging between oil and stock markets: Evidence from wavelet analysis," Energy Economics vol. 49, pp. 540 – 549, 28 March 2015 2015.

[14] T.-H. Le and Y. Chang, "Dynamics between strategic commodities and financial variables: Evidence from Japan," Resources Policy, vol. 50, pp. 1 - 9 2016.

[15] K. Amadeo. (2019, 24 October). Why the US Dollar Is the Global Currency. Available: https://www.thefinance.com/world-currency-3305931

[16] B. o. G. o. t. F. R. S. (US). (2019, 9 August). Japan / U.S. Foreign Exchange Rate. Available: https://fred.stlouisfed.org/series/DEXJPUS

[17] Investing.com. (2019, 9 August). Philadelphia Gold/Silver Historical Data. Available: https://www.investing.com/indices/philadelphia-gold-silver-historical-data

[18] Investing.com. (2019, 9 August). NYSE Composite Historical Data. Available: https://www.investing.com/indices/nysse-composite-historical-data

[19] Investing.com. (2019, 9 August). Crude Oil WTI Futures Historical Data. Available: https://www.investing.com/commodities/crude-oil-historical-data

[20] N. Inc. (2019, 9 August 2019). Nikkei Stock Average (Nikkei 225). Available: https://indexes.nikkei.co.jp/en/nkave/index?type=download

[21] Investing.com. (2019, 9 August). JPX-Nikkei 400 Historical Data. Available: https://www.investing.com/indices/jpx-nikkei-400-historical-data

[22] keras-team. (23 February). Deep Learning for humans. Available: https://github.com/keras-team/keras

[23] Anonimus. (3 November). sklearn.svm.SVR documentation. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html