National College of Ireland

Predicting Asian Stock Market Index Using U.S. Financial Market Indexes and Machine Learning Techniques

MSc Research Project
FinTech

Won Il Kang
Student ID: x20174675

School of Computing
National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: .....Won Il Kang.....

Student ID: .....x20174675.....

Programme: .....MSc FinTech..... **Year:** .....1.....

Module: .....MSc Research Project.....

Supervisor: .....Victor Del Rosal.....

Submission Due Date: .....15.08.2022..... .....

Project Title: .....Predicting Asian Stock Market Index Using U.S. Financial Market Indexes and Machine Learning Techniques .....

Word Count: .....17..... **Page Count**.....6736.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....Won Il Kang.....

Date: .....15.08.2022.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Asian Stock Market Index Using U.S. Financial Market Indexes and Machine Learning Techniques

Won Il Kang
x20174675

Abstract

Among many economic indicators, predictions for stock indices and stock prices are one of the most actively studied topics in economics and computer science. This is because the volatility of stock prices is affected by many variables, making it unpredictable. These variables are microscopically influenced by a company's performance and prospects. Conversely, it is macroscopically influenced by a large number of variables such as politics, diplomacy, security, exchange rates, and monetary policy. However, due to the development of computer hardware technology, computing power has been dramatically improved unlike before. Accordingly, many algorithms used for prediction are improving prediction accuracy. In particular, algorithms called machine learning provide more accurate prediction results that could not be obtained before. This is done by discovering patterns and features among countless data through advanced computing operations. In this study, Asian stock indices were predicted using machine learning algorithms and US financial market indicators. The reason why US financial market indicators were used to predict Asian stock indices is that economic cooperation between countries has been strengthened, and the economic situation of one country can easily spread to other countries. The study expected that the prediction accuracy of Asian stock indices would be improved by using the US financial market indices, which are the most influential in the world. In addition, since the operating times of the US financial market and the Asian financial market do not overlap, this time difference is expected to be easy to construct an actual prediction model in reality. The US financial indicators used in this forecast include the S&P 500, Nasdaq 100, and Dow, which are representative stock indices in the US. In addition, the VIX index, which is a volatility index, and the exchange rate with the country to be predicted were also used for this prediction. Lastly, to improve accuracy, predicted Asian stock indices were limited to the opening price. Those Asian stock indices were Korea's KOSPI, Hong Kong's Hang Seng, and Japan's Nikkei stock index.

1 Introduction

The study of stock price and stock index prediction has been studied in various fields such as economics and computer science for a long time. Numerous models and algorithms have emerged in this regard. Nevertheless, it is virtually impossible to accurately predict the stock index. This is because the variables that affect stock indices are broad and inconsistent. In other words, it is not easy to build an accurate forecasting model by quantifying all variables related to human life, such as politics, diplomacy, security, and resources, as well as economic variables, irregularly affecting the financial market (Vijh, 2020, p599). However, the

development of computer hardware and the emergence of algorithms that maximize computer performance, such as machine learning, has been applied to stock price prediction, improving predictability (Nikou, 2019, p. 172). This is because algorithms such as machine learning use advanced hardware to discover specific patterns that exist in numerous data volumes that are different from the previous ones and use them for prediction. To make an accurate prediction, it is necessary to input as many patterns as possible and a large amount of data is required, this is where machine learning excels as it can handle complex and large volumes of data (Naeini, 2010, p. 132). In this study, various financial markets in the United States and major Asian stock indices from January 2001 to May 2022 were used as data through APIs provided by Yahoo Finance. In addition, various algorithms were used to compare the performance of machine learning algorithms. Deep learning algorithms and traditional machine learning algorithms such as random forests and support vector machines were also used. In addition, a linear regression model, which has shown strong performance up to now, was used for prediction, which became an important indicator for comparing the performance of machine learning algorithms.

This study is to predict Asian stock indices using various financial market indicators in the United States. The US financial indicators used in this study were: S&P 500, Nasdaq 100, and Dow, which are representative stock indices. The VIX index, which is a volatility index, and exchange rates with the target stock index countries were also used. In addition, the Asian stock indices to be predicted are the opening prices of Korea's KOSPI, Japan's Nikkei and Hong Kong's Hang Seng stock index. The reason why the opening prices were selected as the forecast target is that it is the index that can reduce the time difference between the results of the US market and the Asian market the most. This choice in time was expected to improve the accuracy of the predictions. In addition, there are two main reasons for using US financial market indicators in predicting Asian stock prices in this study. First, the US economy is one of the largest in the world, and it affects the economies of other countries to a large extent (Brunschwig, 2011, p. 1). For example, the 2008 subprime mortgage crisis in the United States was not only a problem within the United States but also had a major impact on the global economy. In this regard, Brunschwig (2011) stated that most Asian countries, including China and India, saw their real GDP growth slow down during this period. Since the economic problems of the United States are likely to extend to other global countries, the financial market indicators of the United States can be an important indicator when diagnosing a country's economic situation. The second is the difference in the financial market operating hours. The stock markets trading hours of Korea, Hong Kong, and Japan, which are the subjects of this study, and the US market do not overlap. Therefore, in the real world, the US market data from the previous day can be used to predict the stock indices of the three countries. It can be said that this study has versatility because it is not a one-time analysis of phenomena. This study compared the performance difference between predicting Asian stock prices using US financial market indicators and predicting stock prices using only the history of stock indices itself. Following that, the study will examine which algorithm is most effective to predict stock indices.

In this study, the stock price was predicted using the deep learning method, which has been the most active in recent research. The deep learning methods such as artificial neural networks (ANN) recurrent neural networks (RNN) and long short-term memory (LSTM) were used. In particular, both RNN and LSTM are known to show good performance on time series data (Ma, 2020, p. 4; Nikou, 2019, p. 172). However, as a result, ANN showed better results. In addition, the prediction performance of these deep learning methods was compared with the existing well-known machine learning methods such as support vector machine and random forest. Also, the performance difference was compared. The data used in the forecast were extracted from Yahoo Finance, and the latest data were used for the data period from January 2001 to May 2022. Among the total data, the most recent 100 data were used for accuracy verification and the rest were used for model training. As a result, the artificial neural network and linear regression model showed the best performance, and among them, the artificial neural network showed a slightly good overall performance.

There are many advantages to predicting stock indices. Investors can minimize investment losses or maximize profits through stock price prediction indicators, and economic authorities can predict financial market shocks in advance and promote financial market stability by adjusting public funds. Therefore, accurate prediction of a country's stock index can be used in a wide range of ways. This study showed the possibility of making such stock index predictions more accurately than previous techniques through machine learning and US financial market indicators. As for the structure of the report, Section 2 mentioned the progress of related studies and differences from our study. Section 3 provides an overview of various machine learning techniques and linear regression used in this study, and section 4 mentions the detailed design of each algorithm. Section 5 presents the results obtained through the study, and Section 6 contains the evaluation and interpretation thereof. The last section concludes this report with a comprehensive evaluation of the study and further research directions in the future.

2 Related Work

2.1 Linear Regression

In the study of Abidatul (2014), linear regression was used to predict the Jakarta stock index in real-time using the Android application. In the linear regression modelling method, the stock index of the next business day was predicted using the previous three days rather than the entire previous historical data. However, it was not possible to objectively prove the performance of the proposed model because he mentioned only the performance of the model designed by himself without comparison with other algorithms.

Another study was to predict the direction of stock prices with the public sentiment. Yahya (2015) uses a dataset of company-related tweets extracted from Twitter data and classified them into positive and negative tweets. As a regression model, he predicted the direction of stock price by combining the positive rate of tweets related to a specific company and the stock price values of the previous 5 days in a linear regression formula. In addition, he compared the performance with Neural Network algorithms, Decision Tree, naive Bayes, Random Forest and

Support Vector Machine. As the result, the random forest model presented the highest performance with an accuracy of 67%. However, it showed a low performance of up to 67% despite forecasting the direction of the stock price, not the value of the stock index itself.

For the next study, Emioma (2020) predicted the adjusted closing price on that day using the opening price, high price, low price, and closing price of the day. As a data set, the history of Bank of America stocks for 7 years was extracted using Yahoo Finance API. As a result of the study, an error of about 1.4% was obtained. However, the adjusted closing price with the data of that day is a model that cannot be designed in the real world. This is because historical data is deterministic data, but it can change in real-time in the real world. It is similar to Seethalakshmi (2018) because the adjusted closing price of S&P 500 is the target value. Moreover, in this study, there was no comparison with other algorithmic models, so it was not possible to determine how well the linear regression model performed.

For the last study, Siew (2012) did not predict the stock price but a ranking among stocks. The ranking values were predicted after composing the stock prices in a linear regression equation. Siew (2012)'s attempt is meaningful in that it has a different direction from previous attempts. However, it is difficult to say that the stock price ranking is meaningful information for investors or financial people.

2.2 Support Vector Machines

Sheta (2015) conducted a study to predict the S&P 500 index using SVM. For the prediction, multiple linear regression models and artificial neural networks were also compared to prove the objectivity of the SVM model performance. As a result, the SVM model was the best performance. However, in the artificial neural network, the number of nodes in a single layer was remarkably small, less than 30. It can be said that this did not take full advantage of the algorithms. On the other hand, Vaishnavi (2019) conducted a study to predict Coca-Cola stocks using SVM. The dataset is one year from January 2017 to January 2018 obtained from Quandle. The SVM model was compared with the linear regression model, and the SVM performed better. However, Vaishnavi (2019)'s study has a short dataset of one year. Moreover, it is questionable whether stocks of a single company can be effectively applied to other stocks.

Henrique (2018) conducted a study to predict stock indices in real-time by minute using SVM. This study is unique in that it predicted real-time stock prices in minutes. The duration of the test set was mostly less than one year because of minute-by-minute data. As a result of the study, the predictive results were good for blue-chip stocks. However, blue stocks are normally easy to predict because the flow tends to be stable.

Leung (2014) conducted a study to predict the direction of the S&P 500 stock index using Structure Support Vector Machines (SSVM). SSVM is an algorithm specialized for classification by building categories. Leung (2014) predicted the rise or fall of the stock index. The accuracy was at least 78%. However, Leung (2014) did not prove the objective performance of the model by showing only SSVM as a result.

2.3 Neural Networks

The first study is the study of Naeini (2010), which predicts the stock price of the next day. As a dataset, 1,094 stocks registered in the Iranian stock market from 2000 to 2005 were included. The input data were the high and low prices of the previous day and the average stock price of the last n days. As a result of the study, deep learning showed better performance than the linear regression model. However, since the network layer configuration consists of one layer, the performance of the neural network could not be sufficiently demonstrated. This is because neural networks require a certain number of layers and nodes to store and process sufficient information. In our study, we plan to maximize the performance of the neural network by configuring 3 layers and each layer from 64 to 1024 nodes.

Another study is that of Nikou (2019). In this study, stock price prediction was performed using various types of neural network algorithms and other machine learning algorithms. As data, the MSCI UK index for 3 years from January 2015 was used. The artificial neural network (ANN) consists of two hidden layers and the number of each node is from 5 to 15. In addition, as for the configuration of deep learning long short-term memory (LSTM), the model is composed of 50 to 400 nodes in 50 units. As a result of the study, LSTM showed the best performance, but a disappointment is that the number of layers and the number of nodes in the neural network is considerably smaller than that of LSTM. Therefore, it may not be able to show the maximum performance for ANN. However, in our study, ANN was also designed as a model with three layers, consisting of up to 1024 nodes.

In a study by Tiwari (2017), the opening price of the stock index on the next day is predicted using the Indian Nifty 50 stock index as a data set. As a result of the study, the neural network presented the best performance with an error of 1.81%. However, the difference from our study is that only the history of the stock index, which is the target of prediction, is considered. Our study built a model with better performance by including major US financial market indicators.

2.4 Random Forest

Vijha (2020) predicted the stock index of the next business day using artificial neural networks and random forests. As the target data, Goldman Sachs, Nike, Pyza, JPMorgan, and JNJ were used for learning data from 2009 to 2019. In this study, even though the hidden layer of the artificial neural network has a single configuration, the artificial neural network showed excellent performance. However, it lacks in that the performance of the algorithm cannot be maximized because the input of the data to be predicted is simply limited to the history and the design of the artificial neural network is simple.

Another study predicts the closing price of Nasdaq ETFs. Sun (2020) built a predictive model using multiple linear regression, random forest and long short-term memory (LSTM). As a result of the study, it is interesting to note that multiple linear regression showed the best performance compared to other machine learning algorithms. However, in machine learning algorithms, securing a lot of data indicates predictive performance. However, in this study, the data period is very short, from 2018 to 2019.

Finally, Sadorsky (2021)'s study predicts the direction of stock prices of five ETFs following the Green Energy Index Fund (QCLN). The data period is from January 2009 to September 2020. In this study, the direction of the stock index was predicted using a classification model. As a result of the study, it was possible to build a model with an accuracy of at least 60% or more. However, the fact that this study only targeted ETFs with a specific theme limited its versatility of this study.

2.5 Other Methods

For the first study, Ma (2020) constructed a model to predict the stock price of DELL using the ARIMA model. However, in this study, LSTM showed the better performance than ARIMA. In particular, ARIMA has the disadvantage of being able to predict only with a single input. For another study of ARIMA, Izzah (2014) designed a stock price prediction model using ARIMA with data of Nokia and Zenith Bank in 2010. The predicted result was that Nokia performed better. It is predictable because blue-chip stocks usually have low volatility.

Next is Suharsono (2017)'s study, which predicted stock indices of Thailand, Singapore, Malaysia and the Philippines using Vector Error Correction Models (VECM) and Vector Autoregressive (VAR). In the study of Suharsono (2017), only two algorithms were compared, and VAR showed better performance. However, there was a lack of comparison with other algorithms for objective performance comparison of VAR and VECM.

The last algorithm is reinforcement learning. Reinforcement learning is a type of machine learning that builds an optimized model through compensation when an appropriate value is induced. Lee (2001) predicted Korea's stock index through the TD algorithm. However, it is only stated about the applicability of the algorithm without a performance. Lastly, Yang (2020) conducted a study on stock trading targeting the Dow Jones 30 stock index. As a result of trading forecasts, the average annual return was 13%.

3 Research Methodology

Our study predicts the opening price of stock indices in three Asian countries using major US financial market indicators. Major US indicators include S&P 500, Nasdaq 100, Dow, and VIX. In addition, each country's exchange rate with the US dollar is included. For the prediction, Korea's KOSPI, Japan's Nikkei and Hong Kong's Hang Seng were targeted. The data required for this was extracted using Yahoo Finance API. The data period is from January 2001 to May 2022. Each data consists of the opening price, closing price, low price, high price, adjusted closing price, and volume. However, the exchange rate and VIX are missing the volume. The language used for the analysis was Python, and the program used was PyCharm. The Python interpreter version was 3.9.

In this study, for data pre-processing, the US S&P 500, Nasdaq 100, Dow, exchange rate, and VIX were merged. Based on this data, stock index data from three Asian countries were merged with the merged US market data. Lastly, the opening prices of the predicted indices for the next business day were synthesized together for learning and testing. Based on this data set, the data

were separated into data that includes US data and that does not, respectively. This is to see the prediction performance when US market data are included. Also, in the case of data missing, data for the entire day was deleted. This is the way mainly used in major financial data analysis. Finally, each data was normalized to prevent excessive reflection of specific data when input to future algorithms.

In this way, two types of data sets were obtained: a data set with US financial data and a data set without US financial data. The obtained data set was again divided into a data set for learning and verification. The dataset required for verification was the last 100 most recent data. This was used as a test set to judge the performance of the algorithm, and the rest was used as learning data for each algorithm.

A total of six algorithms are used for this prediction. Those are linear regression, three types of deep learning algorithms, and general machine learning algorithms: random forest and support vector machine. The deep learning algorithms consist of artificial neural network (ANN), long short-term memory (LSTM), and recurrent neural network (RNN). A total of six algorithms were used to predict the opening price of the stock index on the next business day.

Finally, when judging the performance of each model, we plan to evaluate it using the error of each prediction value. Generally, there are three main methods for determining the error: Mean Square Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Here, the MSE value was not used in this study, because each error is squared and averaged. The values tend to be excessively larger than the unit of the actual value, making comparison difficult. Therefore, we used the RMSE, which is the square root value of the MSE value. Another error is the MAE value. This is the average of the sum of the absolute values of all errors. It is useful to check the overall error. The biggest difference between the MAE value and the RMSE value is that RMSE has an error squaring process, so it tends to show a large error. That makes an outlier a large error value. Therefore, it is possible to confirm the suitability of model performance in data showing irregular flow. On the other hand, MAE is a value averaged by absolute values of errors. Therefore, it can be checked whether the model performs well overall. We will evaluate the prediction model in these two ways.

$$\text{MSE} = \frac{1}{\text{N}}\sum_{i}^{N}(pred_i - target_i)^2 \qquad \text{RMSE} = \sqrt{\frac{1}{\text{N}}\sum_{i}^{N}(pred_i - target_i)^2} \qquad \text{MAE} = \frac{1}{\text{N}}\sum_{i}^{N}|pred_i - target_i|$$

4 Design Specification

There are a total of six algorithms used for stock index prediction. These are statistical techniques such as linear regression and deep learning algorithms such as artificial neural network (ANN), recurrent neural network (RNN), long short-term memory (LSTM), and other general machine learning algorithms such as support vector machine and random forest. Referring to the design of each algorithm in order, linear regression analysis is a traditional statistical method of predicting future values using given variables. The variables required for prediction are called explanatory variables or independent variables. Whereas, the variables to

be predicted are called dependent variables or response variables. In linear regression, the dependent variable is expressed as a linear combination of the independent variables, and the coefficient of the independent variable takes the value with the smallest error among the training data.

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Neural network algorithms solve various problems with networks that mimic human neural networks. This network effectively finds mathematical relationships or hidden patterns between input data through learning. The structure of such a neural network mainly consists of an input layer, one or more hidden layers, and an output layer. The input layer is responsible for inputting data, and the output layer is responsible for outputting results. Here, the hidden layers are responsible for information storage and processing, and the hidden layer is composed of one or more layers with multiple nodes. Each node contains specific information, and this information is activated through various mathematical functions. This is called the activation function. The design of these layers is the key setting that determines the performance of the algorithm. In other words, the algorithm operates through the composition of the layer, the number of nodes in each layer, and the composition of the activation function.

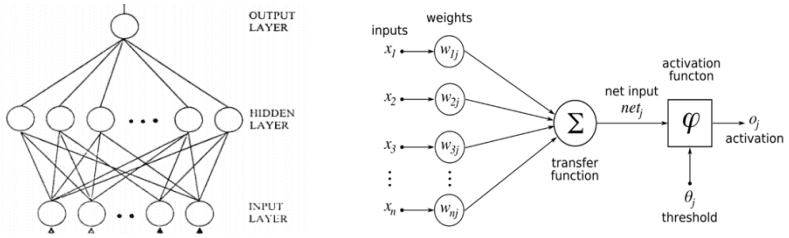

Figure 1: Neural Network Configuration

Among the neural network algorithms, LSTM and RNN are algorithms designed to have strengths in time series data, including the characteristics of the artificial neural networks. These are alternatives created to solve the problem that the previous learning disappears as the artificial neural network continues to learn. This phenomenon is called the Vanishing Gradient Problem. The cause is that, when an error occurs in the learned data, the history data learned in the past disappears as the values of the nodes are appropriately corrected. For reference, the process of modifying node values through errors is called backpropagation. To solve this problem, RNN creates a circular structure, and LSTM preserves long-term learning results by adding memory cells to the neural network. The cyclic structure of RNN gives weight to past learning and reflects it in current learning.

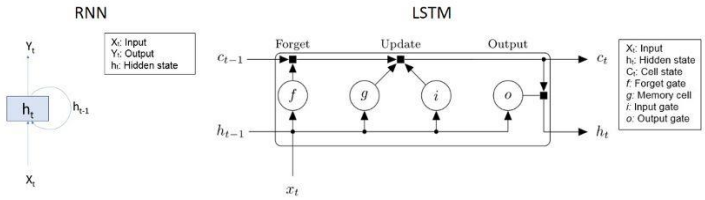

Figure 2: Design of RNN and LSTM

Next, support vector machines are one of the machine learning techniques used for classification and regression. In classification, a line that can separate values is obtained, and future values are classified according to the line. At this time, the line becomes a line with the maximum margin between records. Conversely, in the regression method, a regression line that can include all data to the maximum is calculated based on the distance between the given data, and then the future value is predicted using the line.

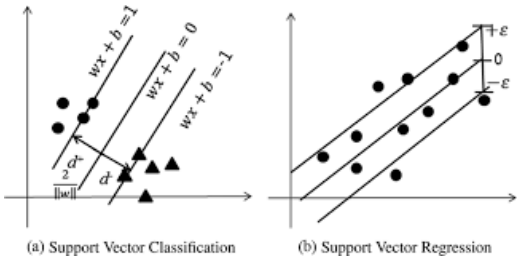

Figure 3: Concept diagram of the operation of the support vector machine

Finally, Random Forest is a type of machine learning that randomly extracts a data set and creates multiple decision trees. The results of these trees are used to predict future values, which are used in regression and classification. When used in regression, the average value of the result values of each decision tree becomes the predicted value. On the other hand, in classification, a majority of values are selected as the predicted values like voting.

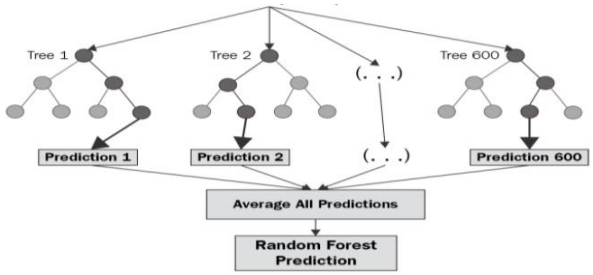

Figure 4: Random Forest Operation Concept

5 Implementation

The analysis environment in our study is as follows. The program was PyCharm and the language was Python. The Python interpreter version is 3.9. All data values were normalized to prevent the excessive influence of specific variables and to enable the comparison of each predicted result. In the case of linear regression analysis, the stock index is the dependent variable as the prediction target. The historical values of the predicted stock index and US financial indicators were expressed as a linear combination. However, by separating data with and without US financial market indicators, it is possible to determine whether there is any performance improvement in the case of data containing US financial market indicators.

In the case of the design of a deep learning algorithm, the algorithm performance varies according to the number of hidden layers, nodes, and activation functions. Therefore, in this study, we tried to derive the design with the best performance by testing as many cases as

possible. First, the batch value was set to 64 as the common design for ANN, LSTM, and RNN, and the loss function was unified as Mean Square Error (MSE). Here, the batch value indicates the number of data to be processed at one time. On the other hand, the loss function is a formula that calculates the error of the predicted value for each batch during training and allows the value of each deep learning network node to be modified through the error. In addition, an early stop function was added to all algorithms to prevent overfitting. Early Stop is a function that terminates learning when there is no performance improvement more than n times during deep learning iterative learning. Here, the maximum iterative learning of all algorithms was set to 1,000, and the learning of the algorithm was designed to stop when there was no performance improvement for 5 consecutive times.

The detailed design of the ANN consisted of three layers, and the number of nodes per layer varied to 64, 128, 256, 512, and 1024. However, the number of nodes in the previous layer and the next layer was designed so that the next layer was an equal or larger number of nodes to prevent loss of values when transferring values between layers. Also, in the case of ANN, all data are shuffled before learning to prevent a vanishing gradient problem as mentioned in Section 4. Next, the activation function of each layer was changed to 'tanh', 'sigmoid', 'relu', 'softsign', and 'softmax' to find the optimal combination. As a result, the number of nodes and the activation functions showing the best performance are described in Table 1.

Table 1: Detailed design of ANN and its performance

Market*	Layer1	Layer2	Layer3	RMSE	MAE
KOSPI	512 (tanh)	512 (tanh)	512 (tanh)	0.00603	0.00465
HangSeng	64 (tanh)	64 (tanh)	1024 (tanh)	0.01078	0.00798
Nekkei	64 (tanh)	64 (tanh)	1024 (tanh)	0.00710	0.00543

* data-sets that do not have financial features from the US

Next, due to the characteristic of the algorithms, LSTM and RNN were configured as a single layer, and the number of nodes was tested by changing the number of nodes to 64, 128, 256, 512, and 1024. The activation function was also set to 'tanh', 'sigmoid', 'relu', 'softsign', and 'softmax' to find the optimal design. However, for LSTM and RNN, window values must be set due to the characteristics of time series data, and the values were tested by changing the values from 10 to 60 in units of 10. Here, the window size indicates how many recent data values are used to predict the value. As a result, the configuration showing the highest performance is shown in Table 2 and Table 3.

Table 2: Detailed design of LSTM and its performance

Market*	Layer	Act. Func.	Windows	RMSE	MAE
KOSPI	128	relu	20	0.01404	-
	128	sigmoid	20	-	0.01094
HangSeng	1024	relu	20	0.02244	-
HangSeng*	256	relu	50	-	0.01674
Nekkei	128	sigmoid	50	0.01826	-
	512	relu	40	-	0.01462

* data-sets that do not have financial features from the US

Table 3: Detailed design of RNN and its performance

Market*	Layer	Act. Func.	Windows	RMSE	MAE
KOSPI	64	sigmoid	60	0.01396	0.01074
HangSeng	64	sigmoid	30	0.02187	0.01633
Nekkei	1024	softmax	50	0.01814	0.01423

* data-sets that do not have financial features from the US

In the case of a random forest, the number of trees and the depth of the tree must be determined because the decision tree is randomly determined. The value set was tested by changing the number of trees from 200 to 2000 in units of 200. All three countries showed the best performance when the number of trees was 800.

In this study, the support vector machine (SVM) predicted values using the kernel function of rbf. A kernel function is a function that allows two-dimensional values to be extended into a multi-dimensional input space. Therefore, multidimensional values can be processed through SVM. At this time, the gamma and C values must be set for each model. In this study, gamma values were tested while changing to .00001, .00005, .0001, .0002, .0003, .0004, .0005, .0006, .0007, .0008, .0009, .001 values. The model was built by changing the C values to 1, 5, 10, 20, 40, 100, 1000, and 10000. Here, the gamma value determines the distance that one data sample can influence. On the other hand, the C value is a variable that determines how many errors are allowed. The result is shown in Table 4.

Table 4: Detailed design of SVM and its performance

Market*	gamma	C	RMSE	MAE
KOSPI	0.008	1	0.01886	-
	0.00001	10	-	0.01555
HangSeng*	0.0009	40	0.01935	0.01568
Nekkei	0.00001	40	0.01837	0.01460

* data-sets that do not have financial features from the US

6 Evaluation

Before evaluation, all data in this study were normalized to have a value of 0 to 1 to prevent the excessive influence of a specific value. Therefore, the error values were also expressed as small decimal points. To improve the readability required for analysis, in this section, the error values of all results were scaled 100 times. The error values are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). As for the analysis method, we first analysed which algorithms and data performed better for each market, and then mentioned the evaluation of comprehensive research. Tables 6 and 7 show the RMSE and MAE errors by country and algorithm in this stock index prediction.

Table 6: RMSE values by countries and algorithms

RMSE	KOSPI (kp)			Hang Seng (hs)			Nikkei (nk)		
	non_us	us	Imp. (%)	non_us	us	Imp. (%)	non_us	us	Imp. (%)
LR	0.7675	0.6151	**-19.9**	1.1274	1.0960	-2.8	0.9997	0.7003	**-30.0**
LSTM	1.4290	1.4042	-1.7	2.2725	2.2448	-1.2	1.8924	1.8264	-3.5
ANN	0.7605	0.6034	**-20.7**	1.1207	1.0780	-3.8	0.9838	0.7104	**-27.8**
RNN	1.4369	1.3961	-2.8	2.2444	2.1873	-2.5	1.8999	1.8147	-4.5
RF	1.9388	2.4281	25.2	1.1271	1.1251	-0.2	2.4696	2.8101	13.8
SVM	5.3600	1.8868	**-64.8**	1.9352	7.6336	294.5	4.5624	1.8373	**-59.7**

Table 7: MAE values by countries and algorithms

MAE	KOSPI (kp)			Hang Seng (hs)			Nikkei (nk)		
	non_us	us	Imp. (%)	non_us	us	Imp. (%)	non_us	us	Imp. (%)
LR	0.6336	0.4568	**-27.9**	0.8195	0.8168	-0.3	0.8440	0.5514	**-34.7**
LSTM	1.1268	1.0949	-2.8	1.6741	1.7230	2.9	1.5415	1.4627	-5.1
ANN	0.6330	0.4651	**-26.5**	0.8169	0.7978	-2.3	0.8190	0.5435	**-33.6**
RNN	1.1231	1.0747	-4.3	1.6619	1.6334	-1.7	1.5423	1.4239	-7.7
RF	1.5320	1.9108	24.7	0.8305	0.8299	-0.1	2.0038	2.3094	15.3
SVM	5.2026	1.5553	**-70.1**	1.5680	6.3176	302.9	4.4034	1.4604	**-66.8**

6.1 KOSPI

In the KOSPI, the Korean stock index, all algorithms except for random forest (RF) performed better when the US financial market index was included in the prediction. In particular, linear regression (LR) and artificial neural network (ANN) showed the best performance. When US financial market indicators were included, there were performance improvements of 19.9% for LR and 20.7% for ANN on the RMSE basis. On the other hand, the performance improvement was 27.9% and 26.5%, respectively, based on MAE. Between the two algorithms, the ANN showed better performance in RMSE and linear regression presented a better performance in MAE. The fact that RMSE showed good performance can be interpreted as the algorithm, ANN, showing good performance for outlier prediction. On the other hand, LR, which performed well in MAE, can be said to be suitable for low variability flows.

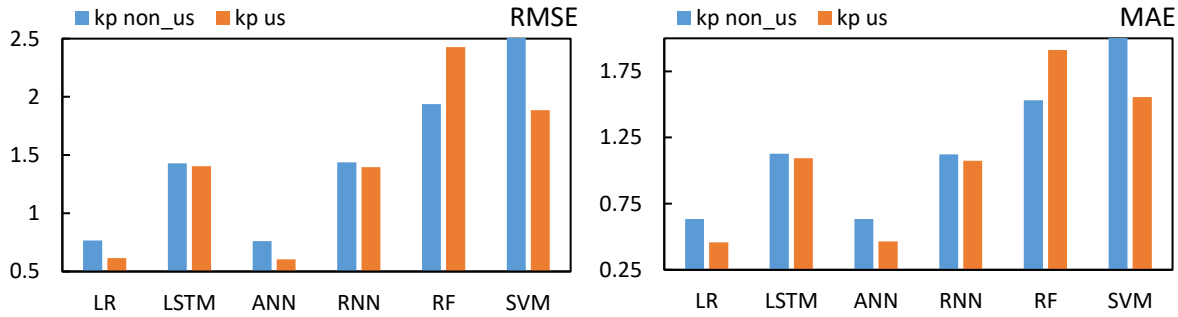

Figure 5: RMSE, MAE of the KOSPI index predictions by algorithms

6.2 Hang Seng

In the Hang Seng market in Hong Kong, all algorithms except support vector machine (SVM) and long short-term memory (LSTM) performed better when US financial market indicators were included in stock index prediction. In particular, linear regression (LR) and artificial neural network (ANN) presented the best performance in the KOSPI. When the US financial

market index was included, there were performance improvements of 2.8% for LR and 3.8% for ANN on the RMSE basis. On the other hand, MAE showed a performance improvement of 0.3% and 2.3%, respectively. In the Hang Seng index, in both RMSE and MAE, ANN performs better than LR, suggesting that neural networks perform well in both high and low volatility market flows. However, compared to the KOSPI and the Nikkei, the Hang Seng index shows the smallest performance improvement when the U.S. stock index is included. It indicates that it is the most independent market of the U.S. economy.

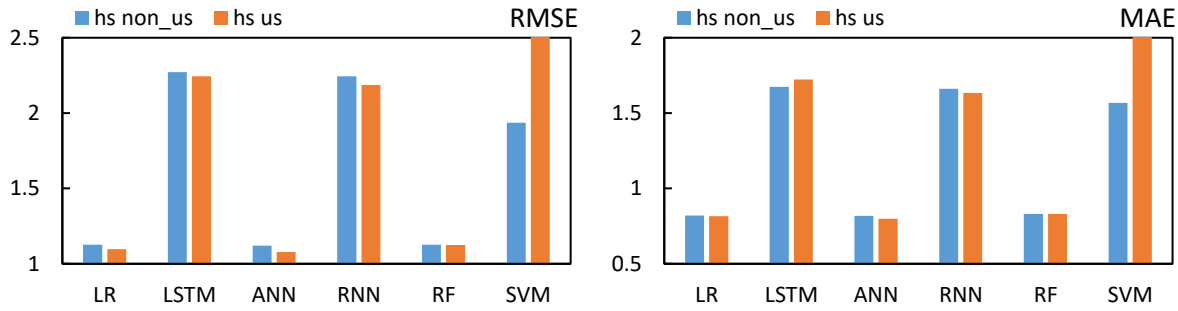

Figure 6: RMSE, MAE of the Hang Seng index predictions by algorithms

6.3 Nikkei

In the Nikkei, the Japanese stock index, all algorithms except for random forest (RF) performed better when the US financial market index was included. Similarly, linear regression (LR) and artificial neural network (ANN) performed the best like the other two stock markets. When the US financial market indicators were included, there were performance improvements of 30.0% for LR and 27.8% for ANN on the RMSE basis, and 34.7% and 33.6% on the MAE, respectively. This is the biggest performance improvement compared to the other two stock markets. Therefore, it can be interpreted that the market is most affected by the US financial market indicators. Here, the ANN performed better in MAE, and LR performed better in RMSE. The fact that the RMSE showed good performance can be interpreted as indicating that LR shows better predictive performance for outliers such as large variability flows. On the other hand, ANN that performed well in MAE can be said to perform well in a field with low volatility.

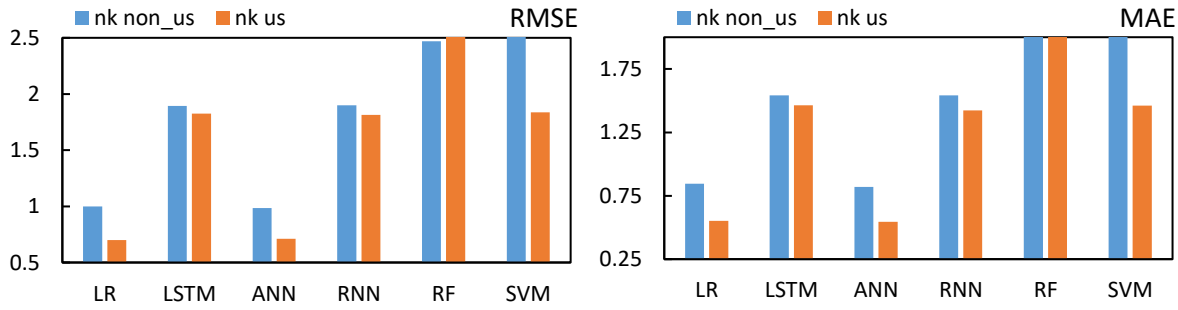

Figure 7: RMSE, MAE of the Nikkei index predictions by algorithms

6.4 Discussion

The KOSPI, Hang Seng, and Nikkei indices, which are the stock index prediction targets of our study, all performed better when the US financial market indicators were included.

However, in the case of Hang Seng, there was a relatively small improvement. It can be inferred that, because Hong Kong has a close economic influence with China, the market has the least link with the US financial market. Generally, LR and ANN show the best performance, so it can be said that both algorithms show good performance in predicting stock indices.

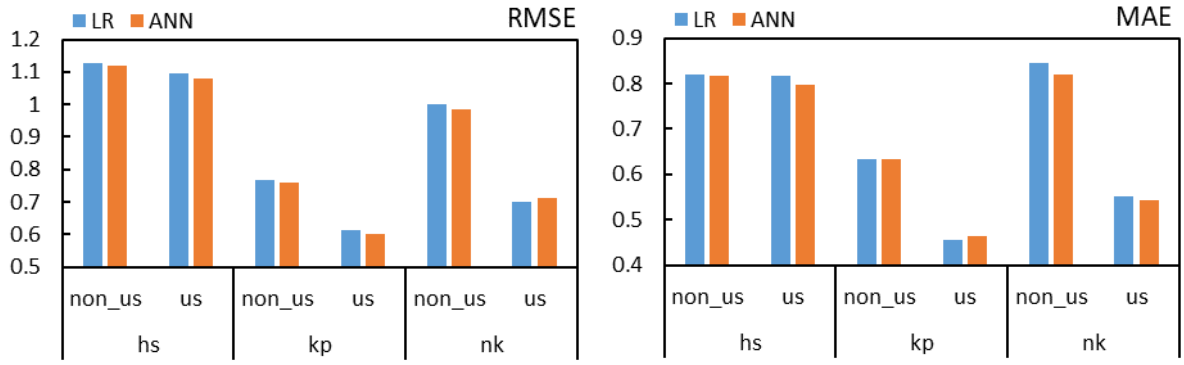

Figure 8: RMSE, MAE values of ANN, LR

In this study, the LSTM and RNN algorithms specialized in time series data did not show good performance. Despite the time series-specific feature, the overall result at Figure 9 below shows that the LSTM and RNN follow the preceding data more than the other two algorithms on the right. The reason for this result can be inferred that the two algorithms have a memory function to store the flow of past data, so overfitting that the previous data has an excessive influence appears. Therefore, it can be seen that LSTM and RNN algorithms are not suitable for predicting stock indices even though they are specialized for time series data.

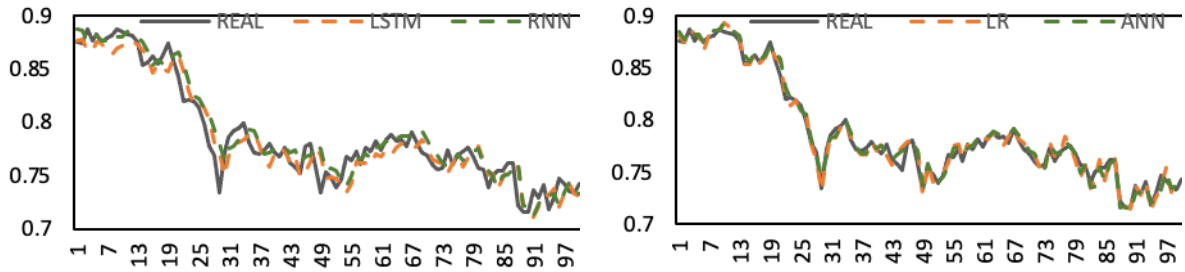

Figure 9: KOSPI Market Forecast Results by Algorithms

7 Conclusion and Future Work

Our study predicts the opening price of stock indices of three Asian countries using US financial market indicators. The target stock markets are Korea's KOSPI, Hong Kong's Hang Seng, and Japan's Nikkei. The US financial market indicators include S&P 500, Nasdaq 100, and Dow, along with the VIX, the volatility index of the US stock market. Also, the exchange rates against the dollar in each country are included in the predicting model. The algorithms used in the prediction are linear regression (LR), artificial neural network (ANN), recurrent neural network (RNN), long short-term memory (LSTM), support vector machine (SVM), and random forest (RF). For performance evaluation, error values of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were used. Because RMSE squares the error, it is highly responsive to data outliers. Therefore, it is possible to evaluate the performance of

outlier prediction using RMSE. On the other hand, because MAE adds the absolute values of errors, it was possible to evaluate the performance of the overall prediction.

In this forecast, we could see how the forecasting performance differs when US financial indicators are included as a predictive model. In particular, for ANN and LR, Korea's KOSPI and Japan's Nikkei both showed a performance improvement of at least 19.9% in both RMSE and MAE. It proves that we can have better forecasting performance for those two markets when US financial market indicators were included. However, in the case of Hang Seng of Hong Kong, although there was a slight performance improvement, the performance was up to 3.8%. It is assumed that the effect was less in the US financial market than in the other two markets. Nevertheless, all three countries showed performance improvement when US financial indicators were included.

Among the algorithms used for the prediction, LR and ANN showed the most outstanding performance than other algorithms. Specifically, in Korea's KOSPI, ANN was superior to linear regression only in RMSE, and in Japan's Nikkei, ANN was superior to linear regression in MAE than LR. On the other hand, in the Hang Seng market, the ANN showed the best performance in both RMSE and MAE. Nevertheless, there was no significant difference in performance between the two algorithms. Nevertheless, in the case of ANN, since learning performance can be improved in a better computing performance environment, ANN has shown the possibility of showing the best performance in predicting stock indices.

However, in this study, it was not possible to confirm specific values that have a major influence on the Asian stock market. It is because the performance evaluation was made on all indicators of the US financial market. In other words, we have not shown how the US stock market index, volatility index, and exchange rate affect Asian stock markets. Also, since this study only targeted the opening price of stock indices, it was not possible to know how different the US financial market indicators performed in the prediction of the closing price, high price, and low price of stock indices. These two cases can be important information to investors or financial authorities, so we leave it as a future task.

References

Sonia Brunschwig, Bruno Carrasco, Tadateru Hayashi and Hiranya Mukhopadhyay (2011). The Global Financial Crisis: Impact on Asia and Emerging Consensus, *Asian Development Bank* (3)

Perry Sadorsky (2021). Random Forests Approach to Predicting Clean Energy Stock Prices, *Journal of Risk and Financial Management*

Zhen Sun and Shangmei Zhao (2020). Machine Learning in Stock Price Forecast, *E3S Web of Conferences (EBLDM)* 214

Mehar Vijha, Deeksha Chandolab, Vinay Anand Tikkiwalb and Arun Kumarc, Stock Closing Price Prediction using Machine Learning Techniques (2020). *International Conference on Computational Intelligence and Data Science (ICCIDS)* 167: 599-606

Agus Suharsono, Auliya Aziza and Wara Pramesti (2017). Comparison of vector autoregressive (VAR) and vector error correction models (VECM) for index of ASEAN stock price, *AIP Conference Proceedings* 1913

Hongyang Yang and Xiao-Yang Liu (2020). Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy, *ACM International Conference on AI in Finance (ICAIF)*

Qihang Ma (2020). Comparison of ARIMA, ANN and LSTM for Stock Price Prediction, *E3S Web of Conferences (ISEESE)* 218.

Bruno Miranda Henrique, Vinicius Amorim Sobreirom and Herbert Kimura (2018). Stock price prediction using support vector regression on daily and up to the minute prices, *The Journal of Finance and Data Science* 4:183-201.

Alaa F. Sheta, Sara Elsir M. Ahmed and Hossam Faris (2015). A Comparison between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index, *(IJARAI) International Journal of Advanced Research in Artificial Intelligence* 4:55-63.

Ramaswamy Seethalakshmi (2018). Analysis of stock market predictor variables using Linear Regression, *International Journal of Pure and Applied Mathematics* 119:369-378.

Carson Kai-Sang Leung, Richard Kyle MacKinnon and Yang Wang (2014). A Machine Learning Approach for Stock Price Prediction, *International Defence Exhibition and Seminar (IDEAS)*, 274-277.

C. C. Emioma and S. O. Edeki (2020). Stock price prediction using machine learning on least-squares linear regression basis, *International Conference on Recent Trends in Applied Research (ICoRTAR)*

Vaishnavi Gururaj, Shriya V R and Dr. Ashwini K (2019). Stock Market Prediction using Linear Regression and Support Vector Machines, *International Journal of Applied Engineering Research* 14:1931-1934.

Mahla Nikou, Gholamreza Mansourfar and Jamshid Bagherzadeh (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithm, *Intelligent Systems in Accounting Finance & Management* 26:164-174.

Abidatul Izzah, Yuita Arum Sari, Ratna Widyastuti and Toga Aldila Cinderatama (2017). Mobile App for Stock Prediction Using Improved Multiple Linear Regression, *International Conference on Sustainable Information Engineering and Technology* 150-154.

Han Lock Siew and Md Jan Nordin (2012). Regression Techniques for the Prediction of Stock Price Trend, *International Conference on Statistics in Science, Business and Engineering (ICSSBE)*

Mahdi Pakdaman Naeini, Hamidreza Taremian and Homa Baradaran Hashemi (2010). Stock Market Value Prediction Using Neural Networks, *International Conference on Computer Information Systems and Industrial Management Applications (CISIM)* 132-136.

Shashank Tiwari, Akshay Bharadwaj and Dr. Sudha Gupta (2017). Stock Price Prediction Using Data Analytics, *International Conference on Advances in Computing, Communication and Control (ICAC3)*

Yahya Eru and Bayu Distiawan Trisedya (2015). Stock Price Prediction using Linear Regression based on Sentiment Analysis, *IEEE International Conference on Advanced Computer Science and Information Systems (ICACSIS)* 147-154.

Jae Won lee (2001). STOCK PRICE PREDICTION USING REINFORCEMENT LEARNING, *International Society for Industrial Ecology (ISIE)* 690-695.

Abidatul Izzah, Aderemi O. Adewumi and Charles K. Ayo (2014). Stock Price Prediction Using the ARIMA Model, *UKSim-AMSS International Conference on Computer Modelling and Simulation* 107-112.

Sidra Mehtab, Jaydip Sen and Abhishek Dutta (2020). Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models, *Communications in Computer and Information Science (CCIS)* 1366: 88-106.