Statistical and Machine Learning Approach in Forex Prediction Based on Empirical Data

Sitti Wetenriajeng Sidehabi
Department of Electrical Engineering
Politeknik ATI Makassar
Makassar, Indonesia
tenri616@gmail.com

Indrabayu[1], Sofyan Tandungan[2]
Department of Informatics Engineering
Universitas Hasanuddin
Makassar, Indonesia
indrabayu@unhas.ac.id, standungan@gmail.com

Abstract—**This study proposed a new insight in comparing common methods used in predicting based on data series i.e statistical method and machine learning. The corresponding techniques are use in predicting Forex (Foreign Exchange) rates. The Statistical method used in this paper is Adaptive Spline Threshold Autoregression (ASTAR), while for machine learning, Support Vector Machine (SVM) and hybrid form of Genetic Algorithm-Neural Network (GA-NN) are chosen. The comparison among the three methods accurate rate is measured in root mean squared error (RMSE). It is found that ASTAR and GA-NN method has advantages depend on the period time intervals.**

Keywords—forex, prediction, ASTAR. GA-NN, SVM, RMSE

I. INTRODUCTION

Forex (Foreign Exchange) is a type of transaction where a party obtains some units in one currency to buy proportion amount in another currency. This exchange is usually conducted in pair currency. The most popular pair and trade worldwide is Euro vs. US Dollar (EUR / USD). In Forex, there are two kinds of analysis, fundamental and technical analysis. Fundamental term refer to the movement of the market in association with news or factors that can affect a country's economy, while technical assessment is mainly observed the supply demand trend through market movements by reading charts and indicators of ongoing market price.

In most cases, Forex rates technical prediction are based on statistical charts and machine learning. It is always interesting to measure up both of this procedures in data series prediction, which none of both scheme is likely better than other for each case [1]. A statistical modelling and forecasting using Auto-Regressive Integrated Moving Average (ARIMA) for Gold Bullion Coin has shown promising result with a MAPE (mean absolute percentage error) within 10% [2]. Artificial Intelligence has been researched as well as statistical and machine learning. With a novel approach for efficient weekly market price forecasting, has come to an outstanding result with 99.62% of accurate rate[3]. Recently, A hybrid methods of Artificial Intelligence also fulfill the 30 minutes time frame prediction [4]. This breakthrough allows a practical application for traders in gaining profit within the time frame with all the price indicators i.e. open, close, high and low are predicted as well. These previous research in price forecasting are conducted thoroughly on single method. This study aim to apply Adaptive Spline Threshold Autoregression (ASTAR), combination of Genetic Algorithm-Neural Network (GA-NN) and Support Vector Machine (SVM) to Forex rates prediction and provide a computational comparison of the performance of these techniques.

A. Adaptive Spline Threshold Autoregression (ASTAR)

ASTAR is a model obtained from modeling nonlinear time series threshold in Multivariate Adaptive Regression Spline (MARS) method where the predictor is the lagged value of time series data [5]. ASTAR has the ability to generate continuous models with underlying limit cycles when the time series data indicate periodic behaviour. Similar to MARS, ASTAR structured by two complementary algorithm. ASTAR has two stepwise algorithm, which help to get basis functions for model and to get the best appropriate model. First step is forward stepwise algorithm, the model obtained has a very complex structure. Second step is backward stepwise algorithm, basis function in the model from the previous step is turn to reach optimum model. ASTAR model example is as follows:

$$Zt = c + \emptyset_1(Z_{t-d1} - t_1)_+ + \emptyset_2(Z_{t-d2} - t_2)_+ \\ +\emptyset_3(Z_{t-d1} - t_1)(Z_{t-d2} - t_2)_+ + \cdots + \varepsilon_t \tag{1}$$

where:
c = constants
$\emptyset$ = coefficient
t_1, t_2 = threshold of each variable $Z_{t\text{-}d1}$, and
$Z_{t\text{-}d2,\, d1,\, d2}$ = lagged predictor variable.

B. Support Vector Machine (SVM)

Support Vector Machine (SVM) is known as a machine learning that uses a pair of input and output data in the form of the desired target. The concept of SVM can be explained simply as the search for the best hyper plane which serves as a separator of two classes in the input space [6].

SVM was developed by Boser, Guyon, Vapnik, and was first presented in 1992 at the Annual Workshop on Computational Learning Theory. The basic concept of SVM is actually a harmonious combination of computational theories that have existed decades earlier, such as margin hyperplane,

978-1-5090-5548-7/16/$31.00 ©2016 IEEE

kernel and concepts supporting others. However, until 1992, there was no attempt to weave these components.

In contrast to the neural network strategy that seeks hyperplane separation between classes, SVM trying to find the best hyperplane in the input space. The basic principle of SVM is a linear classifier, and further developed in order to work on a non-linear problem by incorporating the concept of the kernel trick on high-dimensional workspace.

The example of linearly separated data is shown in Fig.1. The best hyper plane between two classes can be found by measuring the hyper plane margin and find out the maximum points. Margin is defined as the distance between hyper plane and the closest pattern of each class, which is called support vector. The best hyper plane is defined as the following equation.

$$f(x) = wTx + b \qquad 2)$$

where *x* refers to a training pattern, *w* is referred to as the weight vector and *b* as the bias term.

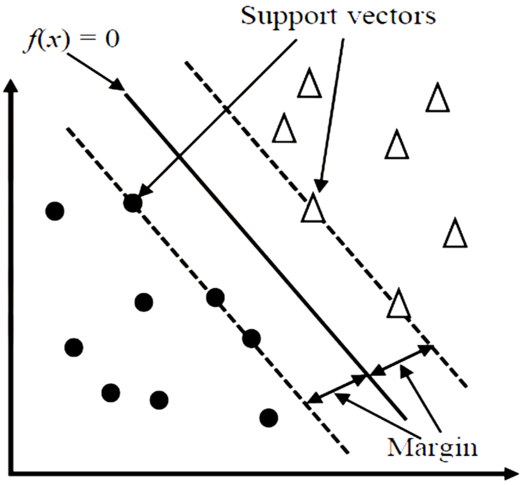


Fig. 1. The example of linearly separated data

The types of SVM kernel that is often used to establish the rules of decision, namely:

1. *A polynomial machine*

$$K(x, x_t) = \ (x, x_t + 1)^d \qquad (3)$$

where *d* is the degree of polynomial kernel

2. *A radial basis function machine*

$$K(x, x_i) = \exp - \left(\frac{\|x - x'\|}{2sig^2}\right) \qquad (4)$$

3. *A two-layer neural network machine*

$$K(x, x_t) = \ \delta[(x, x_t)] = \ \frac{1}{[1 + \{exp(v(x, x_t) - c)\}]} \qquad (5)$$

where v and c are the parameters of sigmoid function.

C. Genetic Algorithm-Neural Network (GANN)

Genetic Algorithm (GA) are algorithms that seeks to apply a comprehension of the natural evolution in problem solving tasks. The approach taken by this algorithm is to merge a various solutions at random within a population and then evaluate them to obtain the best solution [4].

In Genetic Algorithm, procedure for finding the best solution is operated simultaneously on a few solutions known as population. Individuals in a population are specified as chromosomes. First population is randomly generated, then the next population is the result of the chromosomes evolution through generation. In every generation, chromosomes will be evaluated using fitness function. Fitness value determine chromosomes quality in the population.

Artificial Neural Network (ANN) is computer science area that attempt to solve real world problems by proposing a powerful solution. ANN has the capability to learn and generate its own knowledge based no its environment. ANN could be used to model complex relation between inputs and outputs to find patterns in data.

Artificial Neural Network (ANN) is a computation system that its architectural and operation based on the knowledge about biological neuron in human brain. ANN is an artificial representation based on human brain that try to copy the learning process of human brain. ANN models has the ability to analyse, predict and associate. ANN ability can be used to learn and generate rules or operation from a few example or given inputs and make a prediction about possible output or save the characteristic of given inputs.

II. Research Method

A. Input

Historical data from 2007 to September 2012 is prepared for training data. Hourly data of Open, High, Low, and Close is obtained from Meta Trader software then divided according to the input data for the designed application. This data will be used to calculate the prediction diagram used as a reference in determining the value of the actual prediction.

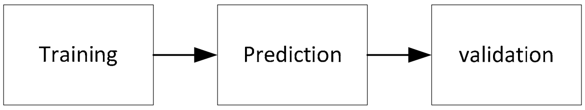


Fig. 2. The proposed scheme of forex prediction

B. System design

The proposed scheme used for ASTAR, SVM and GANN system shown in Fig. 2.

1) Training : Training data consist of four records such as Open, High, Low, and Close. In ASTAR, each records are trained in order to get best ASTAR model. For SVM case, training input data treated with Kernel calculation of Radial Basis Function (RBF). In GANN, training data input is the records of Open, High, Low, and Close price.

2) Prediction : The next stage is to find the predicted value based on ASTAR model, SVM, and GANN. Forex historical data in October 2012 is used in this process. Prediction process is conducted to predict forex value for one day, one week and one month. The output from this process will be validated with actual data.

3) Validation : To validate the result from prediction stage, Predicted value will be compared with actual data in October 2012 from Meta Trader. Validation is important part to evaluate the performance of the prediction. Root Mean Square Error (RMSE) is used to determine accuracy of the prediction performance. The smaller the value of RMSE correspond to better accuracy. RMSE is defined by:

$$RMSE \ = \sqrt{\frac{1}{N}\sum_{i=h}^{N}(yi - \hat{y}i)^2} \qquad (6)$$

where N is the number of data, yi represent actual value, and $\hat{y}i$ is the predicted value.

C. Output

The Output is the best RMSE from ASTAR model, SVM, and GANN. From a series of simulation, it is found that the best RMSE value is from validation process of one day, one week, and one month time frame.

III. Results And Analysis

A. Adaptive Spline Threshold Autoregression (ASTAR) System

ASTAR is a technique to find the best model from a group of data, thus fully utilizing the data ASTAR past and present to make accurate short-term prediction. Prediction in ASTAR system divided into three periods that is prediction in October 1st 2012, October 1st-5th 2012, and all data in October 2012. Figure 3. shown a comparison between actual and prediction result in charts in October 1st 2012. Table 1 shown RMSE value with ASTAR system in October 1st 2012.

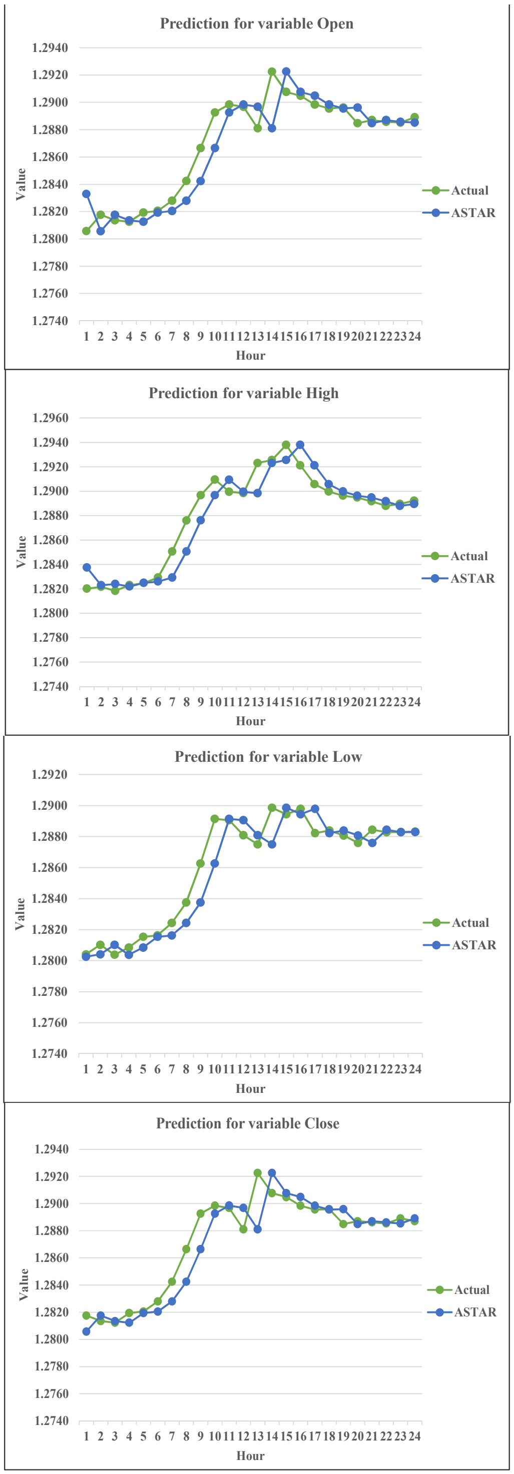


Fig. 3. Comparison of Actual and Predicted value in October 1st 2012 for hourly data in ASTAR system

TABLE I. RMSE value for prediction in October 1st 2012 in ASTAR system

	RMSE			
	Open	**High**	**Low**	**Close**
ASTAR	0.001433	0.001209	**0.001104**	0.001318

TABLE II. RMSE value using ASTAR system

	ASTAR RMSE			
	Open	**High**	**Low**	**Close**
Oct 1st 2012	0.001433	0.001209	0.001104	0.001318
Oct 1st-5st 2012	0.001113	**0.000961**	**0.000978**	**0.001084**
Oct 2012	**0.001104**	0.001531	0.001043	0.001099

TABLE III. RMSE value using ASTAR system

	ASTAR RMSE			
	Open	**High**	**Low**	**Close**
October 1st 2012	0.001433	0.001209	0.001104	0.001318
October 1st-5st 2012	0.001113	**0.000961**	**0.000978**	**0.001084**
October 2012	**0.001104**	0.001531	0.001043	0.001099

Table I shown that the lowest RMSE value is variable Low. Table 2 shows RMSE value for prediction in October 1st 2012, October 1st – 5th 2012, and October 2012.

Table II shows the difference of RMSE values for each variable in different prediction time period. Prediction interval from October 1st to 5th 2012 shown a better performance of ASTAR system than forecasting value in one month period of October 2012.

Table III shows the difference of RMSE values for each variable in different prediction time period. Prediction interval from October 1st to 5th 2012 shown a better performance of ASTAR system than forecasting value in one month period of October 2012.

B. Support Vector Machine (SVM) System

The prognosis computing in SVM system comprise of three periods that is prediction in October 1st 2012, October 1st-5th 2012, and whole data in October 2012. Figure 4 shown a comparison between actual and prediction result in charts in October 1st 2012. Table 3 shown RMSE value with SVM system in October 1st 2012.

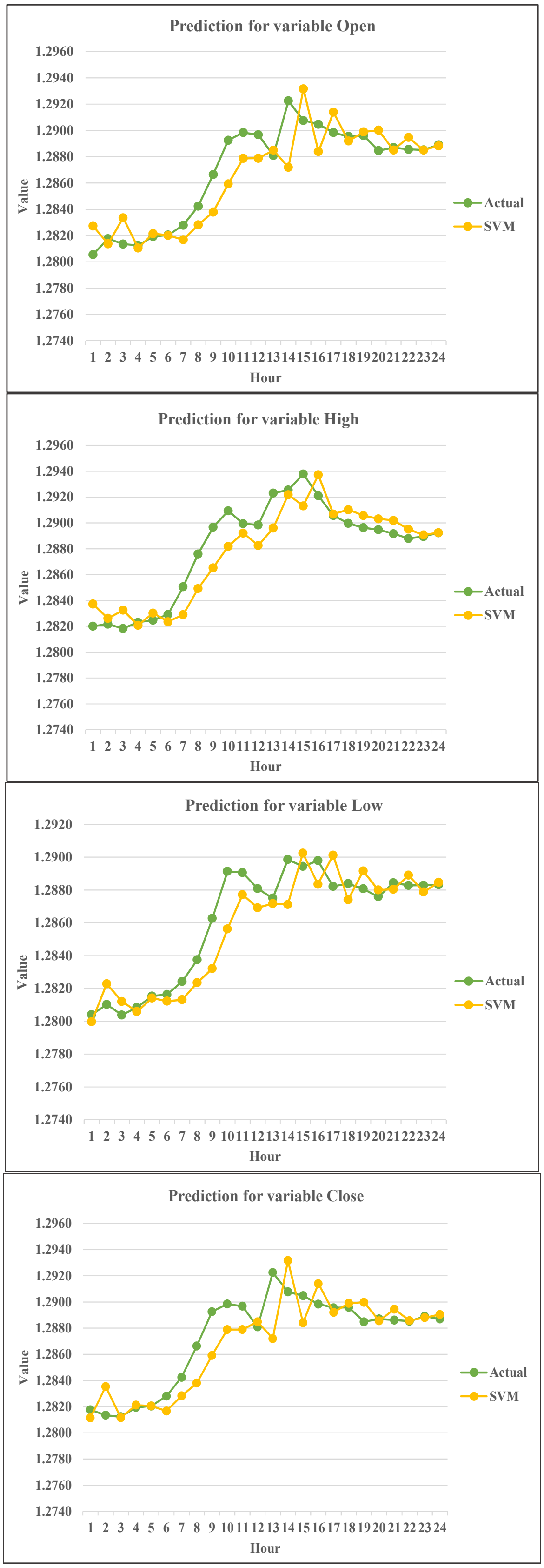


Fig. 4. Comparison of Actual and Predicted value in October 1st 2012 for hourly data in SVM system

TABLE IV. RMSE value for prediction in October 1st 2012 in SVM

	RMSE			
	Open	High	Low	Close
SVM	0.001827	0.001562	**0.001410**	0.001781

TABLE V. RMSE value for all prediction in ASTAR system

	SVM RMSE			
	Open	High	Low	Close
October 1st 2012	0.001827	0.001562	0.001410	0.001781
October 1st-5st 2012	0.001382	**0.001205**	**0.001170**	0.001369
October 2012	**0.001322**	0.001431	0.001215	**0.001322**

TABLE VI. RMSE value for prediction in October 1st 2012 in GA-NN system

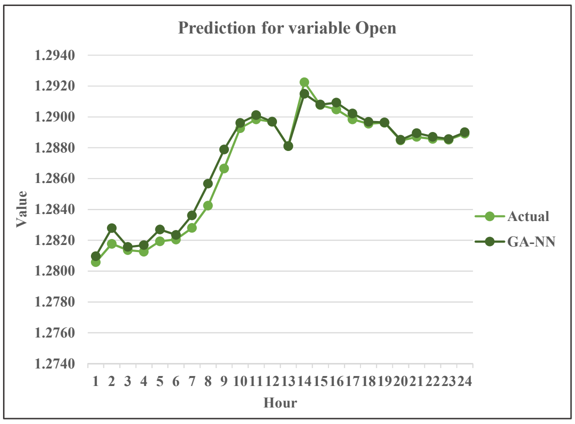
	RMSE			
	Open	High	Low	Close
GA-NN	**0.000559**	0.001747	0.001046	0.001322

From Table IV, it is shown that the lowest RMSE value acquired from the **Low** indicator. Table 4 show RMSE value for prediction in October 1st 2012, October 1st – 5th 2012, and October 2012. Table V shows the difference RMSE values for each variable in different prediction time. Prediction time October 1st until 5th 2012 and October 2012 shown better performance of SVM system when dealing with a lot of data.

C. Genetic Algorithm-Neural Network

Prediction in GA-NN system is divided into three intervals that is prediction in October 1st 2012, October 1st-5th 2012, and all data in October 2012. Figure 5 shown a comparison between actual and prediction result in charts in October 1st 2012. Table 5 shown RMSE value with GA-NN system in October 1st 2012.

Table VI shown that the lowest RMSE value is variable **Open**. Table 6 show RMSE value for prediction in October 1st 2012, October 1st – 5th 2012, and October 2012.



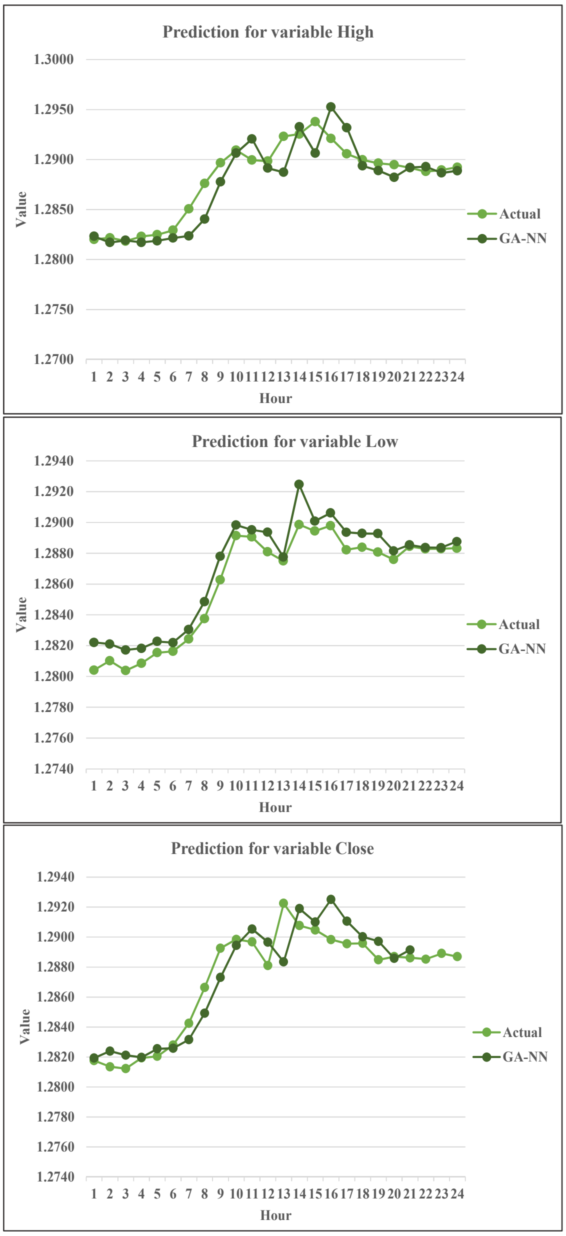


Figure 5. Comparison of Actual and Predicted value in October 1st 2012 for hourly data in GA-NN system

TABLE VII. RMSE value for all prediction in GA-NN system

	GA-NN RMSE			
	Open	High	Low	Close
October 1st 2012	0.000559	0.001747	0.001046	0.0013218
October 1st-5st 2012	0.000307	0.001345	**0.000854**	0.0013220
October 2012	**0.000293**	**0.001040**	0.001232	**0.001295**

TABLE VIII. Comparison of RMSE value for prediction in 1 October 2012

	RMSE			
	Open	High	Low	Close
ASTAR	0.001433	**0.001209**	0.001104	**0.001318**
SVM	0.001827	0.001562	0.001410	0.001781
GA-NN	**0.000559**	0.001747	**0.001046**	0.001322

TABLE IX. Comparison of RMSE value for prediction in 1-5 October 2012

	RMSE			
	Open	High	Low	Close
ASTAR	0.001113	**0.000961**	0.000978	**0.001084**
SVM	0.001382	0.001205	0.001170	0.001369
GA-NN	**0.000307**	0.001345	**0.000854**	0.001144

TABLE X. Comparison of RMSE value for prediction in October 2012

	RMSE			
	Open	High	Low	Close
ASTAR	0.001104	0.001531	**0.001043**	**0.001099**
SVM	0.001322	0.001431	0.001215	0.001322
GA-NN	**0.000293**	**0.001040**	0.001232	0.001295

Table VI shows the difference RMSE values for each variable in different prediction time. Prediction time from October 1st to 5th 2012 shown GA-NN system best performance when dealing with a lot of data.

D. Comparing ASTAR, SVM, and GA-NN System

From Table VIII to Table X, For 1 and 5 days intervals, ASTAR shows better results in term of **High** and **Close** variable because its data show periodic behaviour. On the contrary GA-NN gives an opposite result for the same variables but shows better results in term of **Open** and **Low**. When it comes to longer periods of observations, different results emerge where the **Open** and **High** prediction are better with GA-NN and the rest variables is best from ASTAR forecasting. From this point of view, traders would have wider option in the future trading especially in dealing with volatile currency pairs.

IV. Conclusion

Performance comparison of Statistical and Machine Learning approach has been shown in this paper. From three time periods of observation i.e. 1 day, 5 days, and 30 days, each methods has benefited outcomes depend on the time periods required. The only exception is SVM that always gives average and normalize results compare to ASTAR and GA-NN.

References

[1] Indrabayu, N. Harun, M. S. Pallu, and A. Achmad, “Statistic Approach versus Artificial Intelligence for Rainfall Prediction Based on Data Series,” *ResearchGate*, vol. 5, no. 2, pp. 1962–1969, Apr. 2013.

[2] L. Abdullah, “ARIMA Model for Gold Bullion Coin Selling Prices Forecasting,” *Int. J. Adv. Appl. Sci.*, vol. 1, no. 4, pp. 153–158, Dec. 2012.

[3] Z. H. U. Quan-yin, Y. I. N. Yong-hu, Y. a. N. Yun-yang, and G. U. Tian-feng, “A Novel Efficient Adaptive Sliding Window Model for Week-ahead Price Forecasting,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 3, pp. 2219–2226, Mar. 2014.

[4] A. Sespajayadi, Indrabayu, and I. Nurtanio, “Technical data analysis for movement prediction of Euro to USD using Genetic Algorithm-Neural Network,” in *2015 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2015, pp. 23–26.

[5] P. A. W. Lewis and J. G. Stevens, “Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS),” *J. Am. Stat. Assoc.*, vol. 86, no. 416, pp. 864–877, Dec. 1991.

[6] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.