A Comparison between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index

Alaa F. Sheta
Computers and Systems Department
Electronics Research Institute Giza, Egypt

Sara Elsir M. Ahmed
Computer Science Department
Sudan University of Science and Technology
Khartoum, Sudan

Hossam Faris
Business Information Tech. Dept.
The University of Jordan
Amman, Jordan

***Abstract*—Obtaining accurate prediction of stock index significantly helps decision maker to take correct actions to develop a better economy. The inability to predict fluctuation of the stock market might cause serious profit loss. The challenge is that we always deal with dynamic market which is influenced by many factors. They include political, financial and reserve occasions. Thus, stable, robust and adaptive approaches which can provide models have the capability to accurately predict stock index are urgently needed. In this paper, we explore the use of Artificial Neural Networks (ANNs) and Support Vector Machines (SVM) to build prediction models for the S&P 500 stock index. We will also show how traditional models such as multiple linear regression (MLR) behave in this case. The developed models will be evaluated and compared based on a number of evaluation criteria.**

***Index Terms*—Stock Market Prediction; S&P 500; Regression; Artificial Neural Networks; Support Vector Machines.**

I. Introduction

Understanding the nature of the relationships between financial markets and the country economy is one of the major components for any financial decision making system [1]–[3]. In the past few decades, stock market prediction became one of the major fields of research due to its wide domain of financial applications. Stock market research field was developed to be dynamic, non-linear, complicated, non-parametric, and chaotic in nature [4]. Much research focuses on improving the quality of index prediction using many traditional and innovative techniques. It was found that significant profit can be achieved even with slight improvement in the prediction since the volume of trading in stock markets is always huge. Thus, financial time series forecasting was explored heavenly in the past. They have shown many characteristics which made them hard to forecast due to the need for traditional statistical method to solve the parameter estimation problems. According to the research developed in this field, we can classify the techniques used to solve the stock market prediction problems to two folds:

- **Econometric Models**: These are statistical based approaches such as linear regression, Auto-regression and Auto-regression Moving Average (ARMA) [5], [6]. There are number of assumptions need to be considered while using these models such as linearity and stationary of the the financial time-series data. Such non-realistic assumptions can degrade the quality of prediction results [7], [8].
- **Soft Computing based Models**: Soft computing is a term that covers artificial intelligence which mimic biological processes. These techniques includes Artificial Neural Networks (ANN) [9], [10], Fuzzy logic (FL) [11], Support Vector Machines (SVM) [12], particle swarm optimization (PSO) [13] and many others.

ANNs known to be one of the successfully developed methods which was widely used in solving many prediction problem in diversity of applications [14]–[18]. ANNs was used to solve variety of problems in financial time series forecasting. For example, prediction of stock price movement was explored in [19]. Authors provided two models for the daily Istanbul Stock Exchange (ISE) National 100 Index using ANN and SVM. Another type of ANN, the radial basis function (RBF) neural network was used to forecast the stock index of the Shanghai Stock Exchange [20]. In [21], ANNs were trained with stock data from NASDAQ, DJIA and STI index. The reported results indicated that augmented ANN models with trading volumes can improve forecasting performance in both medium-and long-term horizons. A comparison between SVM and Backpropagation (BP) ANN in forecasting six major Asian stock markets was reported in [22]. Other soft computing techniques such as Fuzzy Logic (FL) have been used to solve many stock market forecasting problems [23], [24].

Evolutionary computation was also explored to solve the prediction problem for the S&P 500 stock index. Genetic Algorithms (GAs) was used to simultaneously optimize all of a Radial Basis Function (RBF) network parameters such that an efficient time-series is designed and used for business forecasting applications [25]. In [26], author provided a new prediction model for the S&P 500 using Multigene Symbolic Regression Genetic Programming (GP). Multigene GP shows more robust results especially in the validation/testing case than ANN.

In this paper, we present a comparison between traditional regression model, the ANN model and the SVM model for predicting the S&P 500 stock index. This paper is structured as follows. Section II gives a brief idea about the S&P 500 Stock Index in the USA. In Section III, we provide an introduction to linear regression models. A short introduction to ANN and SVM is provided in Section IV and Section V, respectively. The adopted evaluation methods are presented in Section VI. In Section VII, we describe the characteristics of the data set used in this study. We also provide the experimental setup and results produced in this research.

II. S&P 500 Stock Index

The S&P 500, or the Standard & Poor's 500, is an American stock market index. The S&P 500 presented its first stock index in the year 1923. The S&P 500 index with its current form became active on March 4, 1957. The index can be estimated in real time. It is mainly used to measure the stock prices levels. It is computed according to the market capitalization of 500 large companies. These companies are having stock in the The New York Stock Exchange (NYSE) or NASDAQ. The S&P 500 index is computed by S&P Dow Jones Indices. In the past, there were a growing interest on measuring, analyzing and predicting the behavior of the S&P 500 stock index [27]–[29]. John Bogle, Vanguard's founder and former CEO, who started the first S&P index fund in 1975 stated that:

> The rise in the S&P 500 is a virtual twin to the rise in the total U.S. stock market, so of course investors, and especially index fund investors, who received their fair share of those returns, feel wealthier,"

In order to compute the price of the S&P 500 Index, we have to compute the sum of market capitalization of all the 500 stocks and divide it by a factor, which is defined as the Divisor (D). The formula to calculate the S&P 500 Index value is given as:

$$Index\ Level = \frac{\sum P_i \times S_i}{D}$$

P is the price of each stock in the index and S is the number of shares publicly available for each stock.

III. Regression Analysis

Regression analysis have been used effectively to answer many question in the way we handle system modeling and advance associations between problem variables. It is important to develop such a relationships between variables in many cases such as predicting stock market [13], [14], [30], [31]. It is important to understand how stock index move over time.

A. Single Linear Regression

In order to understand how linear regression works, assume we have n pairs of observations data set $\{x_i, y_j\}_{i=1,..,n}$ as given in Figure 1. Our objective is to develop a simple relationship between the two variables x (i.e. input variable) and y (i.e output variable) so that we can develop a line equation (see Equation 1).

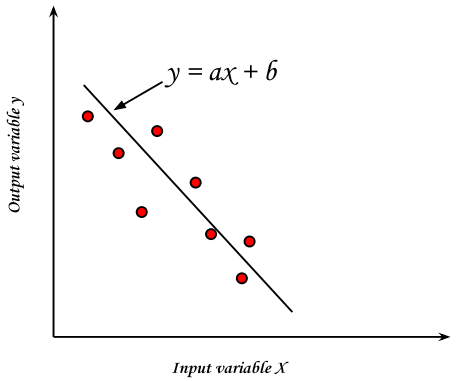

Fig. 1. Simple Linear Model

$$y = a + bx \tag{1}$$

where a is a constant (i.e. bias) and b is the slope of the line. It is more likely that the straight line will not pass by all the points in the graph. Thus, Equation 1 shall be re-written as follows:

$$y = a + bx + \epsilon \tag{2}$$

where ϵ represents the error difference between the values of x_i and y_i at any sample i. Thus, to find the best line that produce the most accurate relationship between x and y. We have to formulate the problem as an optimization problem such that we can search and find the best values of the parameters (i.e. $\hat{a}$ and $\hat{b}$). In this case, we need to solve an error minimization problem. To minimize the sum of the error over the whole data set. We need to minimize the function L given in Equation 3.

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{3}$$

To find the optimal values for $\hat{a}$ and $\hat{b}$ we have to differentiate L with respect to a and b.

$$\begin{aligned} \frac{\partial L}{\partial \hat{a}} &= -2 \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i) = 0 \\ \frac{\partial L}{\partial \hat{b}} &= -2 \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i) x_i = 0 \end{aligned} \tag{4}$$

By simplification of Equations 4, we get to the following two equations:

$$\begin{aligned} n\ \hat{a} + \sum_{i=1}^{n} x_i\ \hat{b} &= \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\ \hat{a} + \sum_{i=1}^{n} x_i^2\ \hat{b} &= \sum_{i=1}^{n} x_i y_i \end{aligned} \tag{5}$$

Equations 5 is called least square (LS) normal equations. The solution of these normal equations produce the least square estimate for $\hat{a}$ and $\hat{b}$.

B. Multiple Linear Regression

The simple linear model Equation 2 can be expanded to a multivariate system of equations as follows:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_j \tag{6}$$

where x_j is the j^{th} independent variable. In this case, we need to use LS estimation to compute the optimal values for the parameters $a_1, \ldots, a_j$. Thus, we have to minimize the optimization function L, which in this case can be presented as:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{a_1} x_1 - \hat{a_2} x_2 - \ldots \hat{a_n} x_n)^2 \tag{7}$$

To get the optimal values of the parameters $\hat{a_1}, \ldots, \hat{a_n}$, we have to compute the differentiation for the functions:

$$\frac{\partial L}{\partial \hat{a_1}} = \frac{\partial L}{\partial \hat{a_2}} = \cdots = \frac{\partial L}{\partial \hat{a_j}} = 0 \tag{8}$$

Solving the set of Equations 8, we can produce the optimal values of the model parameters and solve the multiple regression problem. This solution is more likely to be biased by the available measurements. If you we have large number of observations the computed estimate of the parameters shall be more robust. This technqiue provide poor results when the observations are small in number.

IV. Artificial Neural Networks

ANNs are mathematical models which were inspired from the understanding of some ideas and aspects of the biological neural systems such as the human brain. ANN may be considered as a data processing technique that maps, or relates, some type of input stream of information to an output stream of processing. Variations of ANNs can be used to perform classification, pattern recognition and predictive tasks [15], [19], [20], [22], [30].

Neural network have become very important method for stock market prediction because of their ability to deal with uncertainty and insufficient data sets which change rapidly in very short period of time. In Feedforward (FF) Multilayer Perceptron (MLP), which is one of the most common ANN systems, neurons are organized in layers. Each layer consists of a number of processing elements called neurons; each of which contains a summation function and an activation function. The summation function is given by Equation 9 and an activation function can be a type of sigmoid function as given in Equation 10.

Training examples are used as input the network via the input layer, which is connected to one, or more hidden layers. Information processing takes place in the hidden layer via the connection weights. The hidden layers are connected to an output layer with neurons most likely have linear sigmoid function. A learning algorithms such as the BP one might be used to adjust the ANN weights such that it minimize the error difference between the actual (i.e. desired) output and the ANN output [32]–[34].

$$S = \sum_{i=0}^{n} w_i x_i \tag{9}$$

$$\phi(S) = \frac{1}{1 + e^{-S}} \tag{10}$$

There are number of tuning parameters should be designated before we can use ANN to learn a problem. They include: the number of layers in the hidden layer, the type of sigmoid function for the neurons and the adopted learning algorithm.

V. Support vector machines

Support vector machine is a powerful supervised learning model for prediction and classification. SVM was first introduced by Vladimir Vapnik and his co-workers at AT&T Bell Laboratories [35]. The basic idea of SVM is to map the training data into higher dimensional space using a nonlinear mapping function and then perform linear regression in higher dimensional space in order to separate the data [36]. Data mapping is performed using a predetermined kernel function. Data separation is done by finding the optimal hyperplane (called the Support Vector with the maximum margin from the separated classes. Figure 2 illustrates the idea of the optimal hyperplane in SVM that separates two classes. In the left part of the figure, lines separated data but with small margins while on the right an optimal line separates the data with the maximum margins.

A. Learning Process in SVM

Training SVM can be described as follows; suppose we have a data set $\{x_i, y_j\}_{i=1,..,n}$ where the input vector $x_i \in \Re^d$ and the actual $y_i \in \Re$. The modeling objective of SVM is to find the linear decision function represented in the following equation:

$$f(x) \leq w, \;\; \phi_i(x) > +b \tag{11}$$

where w and b are the weight vector and a constant respectively, which have to be estimated from the data set. ϕ is a nonlinear mapping function. This regression problem can be formulated as to minimize the following regularized risk function:

$$R(C) = \frac{C}{n} \sum_{i=1}^{n} L_\varepsilon(f(x_i), y_i) + \frac{1}{2} \|w\|^2 \tag{12}$$

where $L_\varepsilon(f(x_i), y_i)$ is known as $\varepsilon-$intensive loss function and given by the following equation:

$$L_\varepsilon(f(x), y) = \begin{cases} |f(x) - y| - \varepsilon & |f(x) - y| \geq \varepsilon \\ 0 & otherwise \end{cases} \tag{13}$$

To measure the degree of miss classification to achieve an acceptable degree of error, we use slack variables ξ_i and

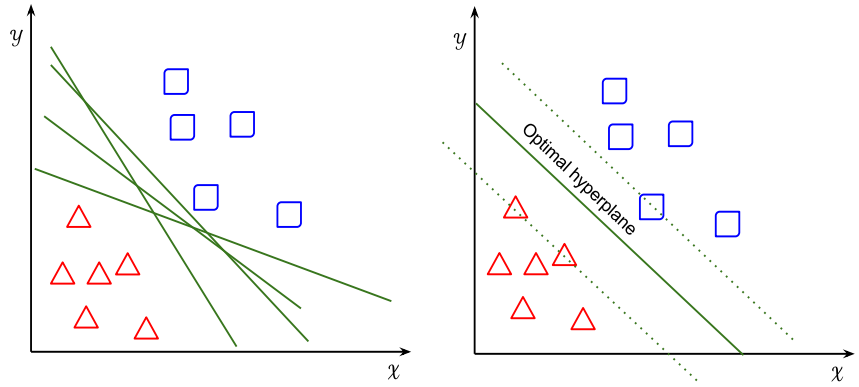

Fig. 2. Optimal hyperplane in Support Vector Machine

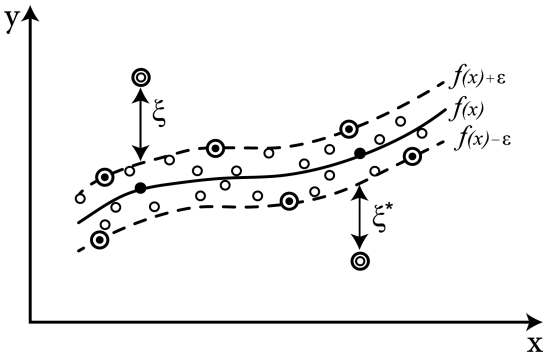

Fig. 3. Optimal hyperplane with slack variables

ξ_i^* as shown in Figure 3. This addition makes the problem presented as a constrained minimum optimization problem (See Equation 14).

$$Min.\ R(w, \xi_i^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{14}$$

Subject to:

$$\begin{cases} y_i - < w, x_i > -b & \leq \varepsilon + \xi_i \\ < w, x_i > +b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \tag{15}$$

where C is a regularized constant greater than zero. Thus it performs a balance between the training error and model flatness. C represents a penalty for prediction error that is greater than ε. ξ_i and ξ_i^* are slack variables that form the distance from actual values to the corresponding boundary values of ε. The objective of SVM is to minimize ξ_i, ξ_i^* and w^2.

The above optimization with constraint can be converted by means of Lagrangian multipliers to a quadratic programming problem. Therefore, the form of the solution can be given by the following equation:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{16}$$

TABLE I
COMMON SVM KERNEL FUNCTIONS

Polynomial Kernel	$K(x_i, x_j) = (x_i.x_j + 1)^d$
Hyperbolic Tangent Kernel	$K(x_i, x_j) = tanh(c_1(x_i.x_j) + c_2)$
Radial Basis Kernel	p: $K(x_i, x_j) = exp(\|x_j - x_i\|/2p^2)$

where α_i and α_i^* are Lagrange multipliers. Equation 16 is subject to the following constraints:

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \tag{17}$$

$$0 \leq \alpha_i \leq C \quad i = 1, \ldots, n$$

$$0 \leq \alpha_i^* \leq C \quad i = 1, \ldots, n$$

$K(.)$ is the kernel function and its values is an inner product of two vectors x_i and x_j in the feature space $\phi(x_i)$ and $\phi(x_j)$ and satisfies the Mercer's condition. Therefore,

$$K(x_i, x_j) = \phi(x_i).\phi(x_j) \tag{18}$$

Some of the most common kernel functions used in the literature are shown in Table I. In general, SVMs have many advantages over classical classification approaches like artificial neural networks, decision trees and others. These advantages include: good performance in high dimensional spaces; and the support vectors rely on a small subset of the training data which gives SVM a great computational advantage.

VI. EVALUATION CRITERION

In order to assess the performance of the developed stock market predication models, a number of evaluation criteria will be used to evaluate these models. These criteria are applied to measure how close the real values to the values predicted using the developed models. They include Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and correlation coefficient R. They are given in Equations 19, 20 and 21, respectively.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|(y_i - \hat{y}_i)| \tag{19}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{20}$$

$$R = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2}} \tag{21}$$

where y is actual stock index values, $\hat{y}$ is the estimated values using the proposed technqiues. n is the total number of measurements.

VII. Experimental Results

A. S&P 500 Data Set

In this work, we use 27 potential financial and economic variables that impact the stock movement. The main consideration for selecting the potential variables is whether they have significant influence on the direction of (S&P 500) index in the next week. While some of these features were used in previous studies [30]. The list, the description, and the sources of the potential features are given in Table III show the 27 features of data set.

The categories of these features include: S&P 500 index return in three previous days SPY(t-1), SPY(t-2), SPY(t-3); Financial and economic indicators (Oil, Gold, CTB3M, AAA); The return of the five biggest companies in S&P 500 (XOM, GE, MSFT, PG, JNJ); Exchange rate between USD and three other currencies (USD-Y, USD-GBP, USD-CAD); The return of the four world major indices (HIS, FCHI, FTSE, GDAXI); and S&P 500 trading volume (V).

S&P 500 stock market data set used in our case consists of 27 features and 1192 days of data, which cover five-year period starting 7 December 2009 to 2 September 2014. We sampled the data on a weekly basis such that only 143 samples were used in our experiments. The S&P 500 data were split into 100 samples as training set and data for 43 samples as testing set.

B. Multiple Regression Model

The regression model shall have the following equation system.

$$y = a_0 + \sum_{i=1}^{27} a_i x_i \tag{22}$$

The values of the parameters $a's$ shall be estimated using LS estimation to produce the optimal values of the parameters $\hat{a}'s$. The produced linear regression model can be presented as given in Table II. The actual and Estimated S&P 500 index values based the MLR in both training and testing cases are shown in Figure 4 and Figure 5. The scattered plot of the actual and predicted responses is shown in Figure 6.

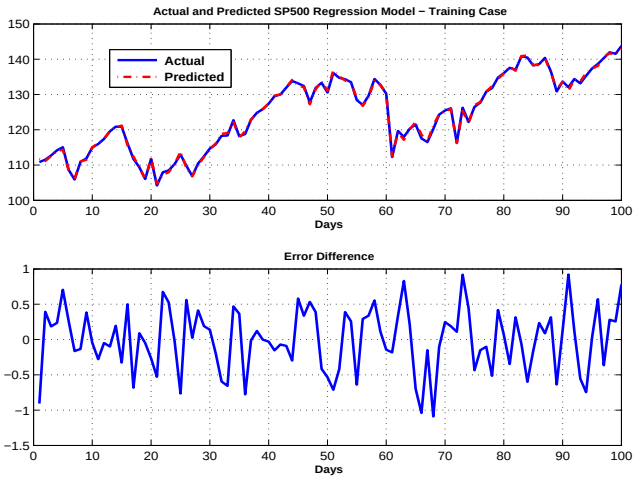

Fig. 4. Regression: Actual and Estimated S&P 500 Index values in Training Case

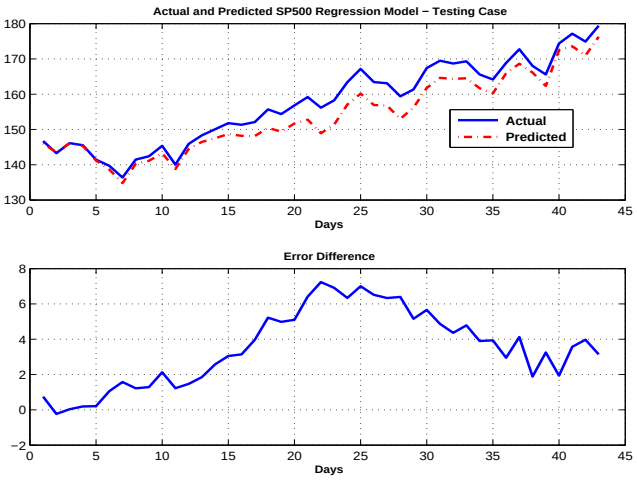

Fig. 5. Regression: Actual and Estimated S&P 500 Index values in Testing Case

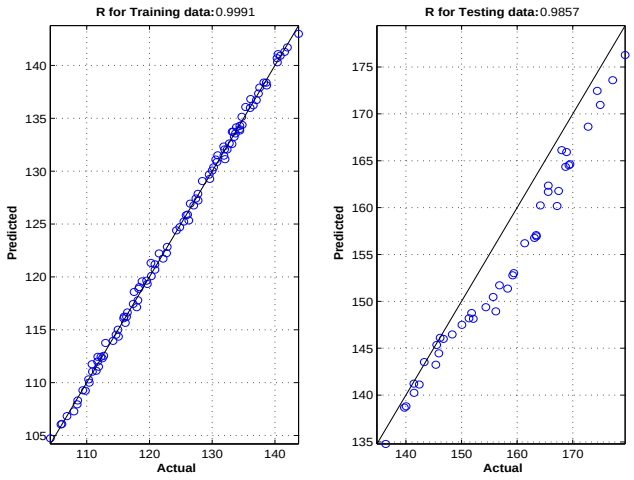

Fig. 6. Regression Scattered Plot

TABLE II
A REGRESSION MODEL WITH INPUTS: $x_1, \ldots, x_{27}$

$$\begin{aligned}\hat{y} &= -0.0234 * x_1 + 0.13 * x_2 + 0.021 * x_3 + 0.021 * x_4 - 0.021 * x_5 \\ &- 10.303 * x_6 + 6.0031 * x_7 + 0.7738 * x_8 + 0.2779 * x_9 - 0.43916 * x_{10} \\ &- 0.27754 * x_{11} + 0.12733 * x_{12} - 0.058638 * x_{13} + 13.646 * x_{14} + 9.5224 * x_{15} \\ &- 0.0003 * x_{16} + 0.24856 * x_{17} - 0.0016 * x_{18} + 0 * x_{19} - 2.334 \times 10^{-9} * x_{20} \\ &+ 0.16257 * x_{21} + 0.63767 * x_{22} - 0.14301 * x_{23} + 0.08 * x_{24} + 0.074 * x_{25} \\ &- 0.0002 * x_{26} + 0.026301 * x_{27} + 6.9312 \end{aligned} \quad (23)$$

TABLE III
THE 27 POTENTIAL INFLUENTIAL FEATURES OF THE S&P 500 INDEX [30]

Variable	Feature	Description
x_1	SPY(t-1)	The return of the S&P 500 index in day $t-1$ Source data: finance.yahoo.com
x_2	SPY(t-2)	The return of the S&P 500 index in day $t-2$ Source data: finance.yahoo.com
x_3	SPY(t-3)	The return of the S&P 500 index in day $t-3$ Source data: finance.yahoo.com
x_4	Oil	Relative change in the price of the crude oil Source data: finance.yahoo.com
x_5	Gold	Relative change in the gold price Source data: www.usagold.com
x_6	CTB3M	Change in the market yield on US Treasury securities at 3-month constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
x_7	AAA	Change in the Moody's yield on seasoned corporate bonds - all industries, Aaa Source data: H.15 Release - Federal Reserve Board of Governors
x_8	XOM	Exxon Mobil stock return in day t-1 Source data: finance.yahoo.com
x_9	GE	General Electric stock return in day t-1 Source data: finance.yahoo.com
x_{10}	MSFT	Micro Soft stock return in day t-1 Source data: finance.yahoo.com
x_{11}	PG	Procter and Gamble stock return in day t-1 Source data: finance.yahoo.com
x_{12}	JNJ	Johnson and Johnson stock return in day t-1 Source data: finance.yahoo.com
x_{13}	USD-Y	Relative change in the exchange rate between US dollar and Japanese yen Source data: OANDA.com
x_{14}	USD-GBP	Relative change in the exchange rate between US dollar and British pound Source data: OANDA.com
x_{15}	USD-CAD	Relative change in the exchange rate between US dollar and Canadian dollar Source data: OANDA.com
x_{16}	HIS	Hang Seng index return in day t-1 Source data: finance.yahoo.com
x_{17}	FCHI	CAC 40 index return in day t-1 Source data: finance.yahoo.com
x_{18}	FTSE	FTSE 100 index return in day t-1 Source data: finance.yahoo.com
x_{19}	GDAXI	DAX index return in day t-1 Source data: finance.yahoo.com
x_{20}	V	Relative change in the trading volume of S&P 500 index Source data: finance.yahoo.com
x_{21}	CTB6M	Change in the market yield on US Treasury securities at 6-month constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
x_{22}	CTB1Y	Change in the market yield on US Treasury securities at 1-year constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
x_{23}	CTB5Y	Change in the market yield on US Treasury securities at 5-year constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
x_{24}	CTB10Y	Change in the market yield on US Treasury securities at 10-year constant maturity, quoted on investment basis Source data: H.15 Release - Federal Reserve Board of Governors
x_{25}	BBB	Change in the Moody's yield on seasoned corporate bonds - all industries, Baa Source data: H.15 Release - Federal Reserve Board of Governors
x_{26}	DJI	Dow Jones Industrial Average index return in day t-1 Source data: finance.yahoo.com
x_{27}	IXIC	NASDAQ composite index return in day t-1 Source data: finance.yahoo.com

C. Developed ANN Model

The proposed architecture of the MLP Network consists of three layers with single hidden layer. Thus input layer of our neural network model has 27 input nodes while the output layer consists of only one node that gives the predicted next week value. Empirically, we found that 20 neurons in the hidden layer achieved the best performance. The BP algorithm is used to train the MLP and update its weight. Table IV shows the settings used for MLP. Figure 7 and Figure 8 show the actual and predicted stock prices for training and testing cases of the developed ANN. The scattered plot for the developed ANN model is shown in Figure 9.

TABLE IV
THE SETTING OF MLP

Maximum number of epochs	500
Number of Hidden layer	1
Number of neurons in hidden layer	20
Learning rate	0.5
Momentum	0.2

D. Developed SVM Model

SVM with an RBF kernel is used to develop the S&P 500 index model. The RBF kernel has many advantages such as the ability to map non-linearly the training data and the ease of implementation [37]–[39]. The values of the parameters C and σ have high influence on the accuracy of the SVM model. Therefore, we used grid search to obtain these values. It was found that the best performance can be obtained with $C = 100$ and $\sigma = 0.01$. Figure 10 and Figure 11 show the actual and predicted stock prices for training and testing cases of the developed SVM model. The scattered plot for the developed SVM model is shown in Figure 12.

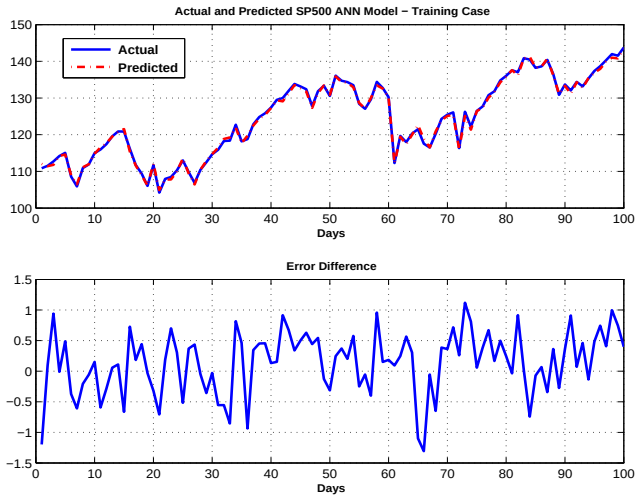

Fig. 7. ANN: Actual and Estimated S&P 500 Index values in Training Case

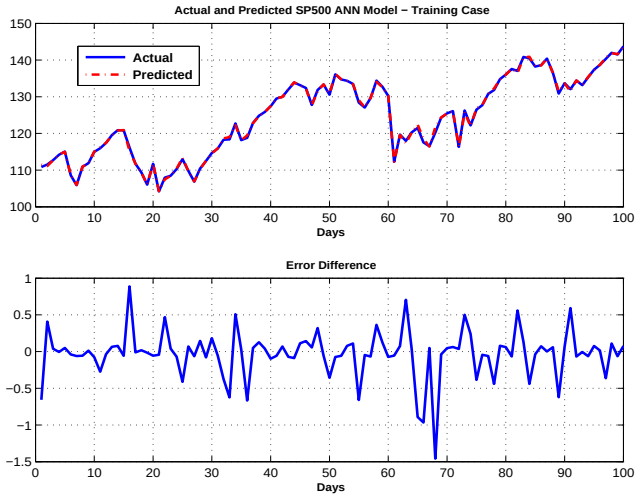

Fig. 10. SVM: Actual and Estimated S&P 500 Index values in Training Case

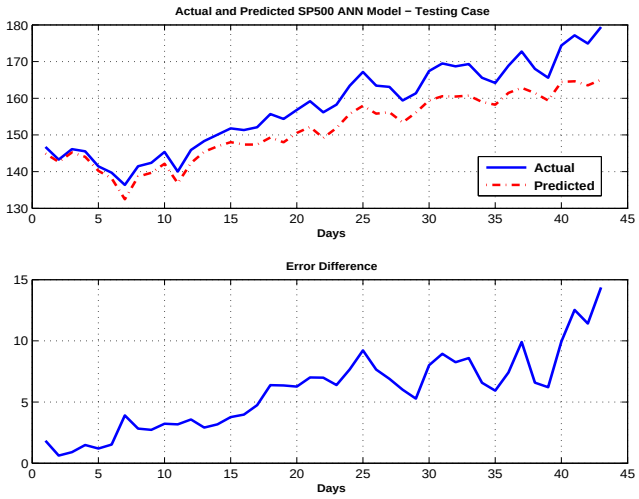

Fig. 8. ANN: Actual and Estimated S&P 500 Index values in Testing Case

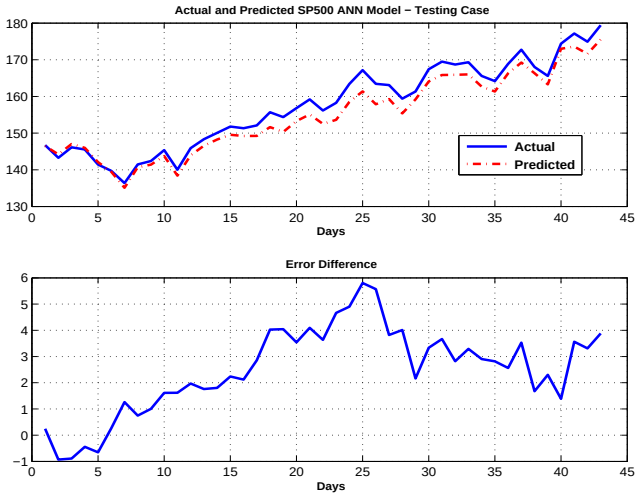

Fig. 11. SVM: Actual and Estimated S&P 500 Index values in Testing Case

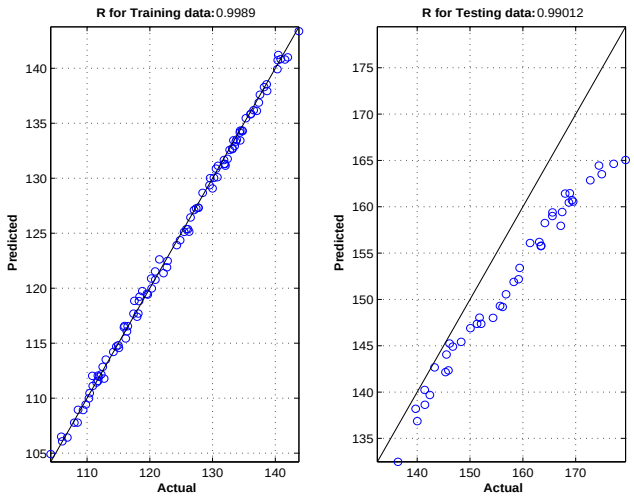

Fig. 9. ANN Scattered Plot

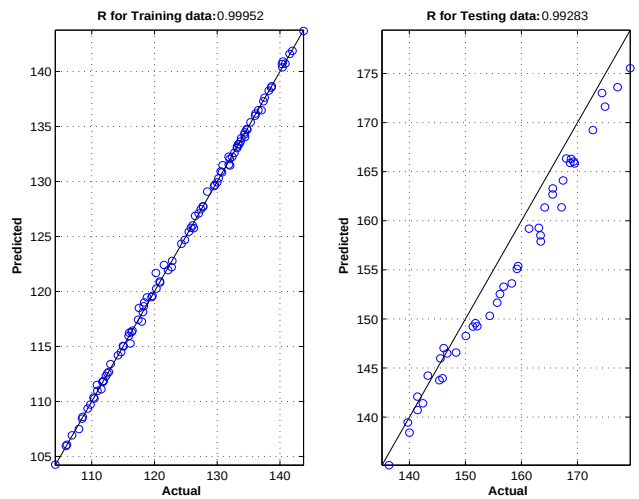

Fig. 12. SVM Scattered Plot

E. Comments on the Results

The calculated evaluation criterion of the regression, MLP and SVM models for training and testing cases are shown in Table V. Based on these results it can be noticed that SVM outperformed the MLP and MLR models in both training and testing cases. SVMs has many advantages such as using various kernels which allows the algorithm to suits many classification problems. SVM are more likely to avoid the problem of falling into local minimum.

VIII. Conclusions and Future Work

In this paper, we explored the use MLP and SVM to develop models for prediction the S&P 500 stock market index. A 27 potential financial and economic variables which impact the stock movement were adopted to build a relationship between the stock index and these variables. The basis for choosing these variables was based on their substantial impact on the course of S&P 500 index. The data set was sampled on a weekly bases. The developed SVM model with RBF kernel model provided good prediction capabilities with respect to the regression and ANN models. The results were validated using number of evaluation criteria. Future research shall focus on exploring other soft computing techniques to solve the stock market prediction problems.

References

[1] S. Hoti, M. McAleer, and L. L. Pauwels, "Multivariate volatility in environmental finance," *Math. Comput. Simul.*, vol. 78, no. 2-3, pp. 189–199, Jul. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.matcom.2008.01.038

[2] Q. Wen, Z. Yang, Y. Song, and P. Jia, "Automatic stock decision support system based on box theory and svm algorithm," *Expert System Application*, vol. 37, no. 2, pp. 1015–1022, Mar. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2009.05.093

[3] M. Kampouridis, A. Alsheddy, and E. Tsang, "On the investigation of hyper-heuristics on a financial forecasting problem," *Annals of Mathematics and Artificial Intelligence*, vol. 68, no. 4, pp. 225–246, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1007/s10472-012-9283-0

[4] T. Z. Tan, C. Quek, and G. S. Ng, "Brain-inspired genetic complementary learning for stock market prediction," in *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2005, 2-4 September 2005, Edinburgh, UK*, 2005, pp. 2653–2660. [Online]. Available: http://dx.doi.org/10.1109/CEC.2005.1555027

[5] A. C. Harvey and P. H. J. Todd, "Forecasting economic time series with structural and box-jenkins models: A case study," *Journal of Business & Economic Statistics*, vol. 1, no. 4, pp. 299–307, 1983. [Online]. Available: http://dx.doi.org/10.2307/1391661

[6] Y. B. Wijaya, S. Kom, and T. A. Napitupulu, "Stock price prediction: Comparison of ARIMA and artificial neural network methods - an indonesia stock's case," in *Proceedings of the 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, ser. ACT'10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 176–179. [Online]. Available: http://dx.doi.org/10.1109/ACT.2010.45

[7] L. Yu, S. Wang, and K. K. Lai, "A neural-network-based nonlinear metamodeling approach to financial time series forecasting," *Appl. Soft Comput.*, vol. 9, no. 2, pp. 563–574, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.asoc.2008.08.001

[8] S. Walczak, "An empirical analysis of data requirements for financial forecasting with neural networks," *J. Manage. Inf. Syst.*, vol. 17, no. 4, pp. 203–222, Mar. 2001. [Online]. Available: http://dl.acm.org/citation.cfm?id=1289668.1289677

[9] M.-D. Cubiles-de-la Vega, R. Pino-Mejías, A. Pascual-Acosta, and J. Muñoz García, "Building neural network forecasting models from time series ARIMA models: A procedure and a comparative analysis," *Intell. Data Anal.*, vol. 6, no. 1, pp. 53–65, Jan. 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=1293993.1293996

[10] R. Majhi, G. Panda, and G. Sahoo, "Efficient prediction of exchange rates with low complexity artificial neural network models," *Expert System Application*, vol. 36, no. 1, pp. 181–189, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2007.09.005

[11] M. R. Hassan, "A combination of hidden markov model and fuzzy model for stock market forecasting," *Neurocomputing*, vol. 72, no. 1618, pp. 3439 – 3446, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231209001805

[12] W. Huang, Y. Nakamori, and S. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & Operations Research*, vol. 32, no. 10, pp. 2513–2522, 2005.

[13] R. Majhi, G. Panda, G. Sahoo, A. Panda, and A. Choubey, "Prediction of S&P 500 and DJIA stock indices using particle swarm optimization technique," in *The IEEE World Congress on Computational Intelligence on Evolutionary Computation (CEC2008)*. IEEE, 2008, pp. 1276–1282.

[14] Y. Zhang and L. Wu, "Stock market prediction of S&P 500 via combination of improved bco approach and bp neural network," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 8849–8854, 2009.

[15] S. O. Olatunji, M. Al-Ahmadi, M. Elshafei, and Y. A. Fallatah, "Saudi arabia stock prices forecasting using artificial neural networks," pp. 81–86, 2011.

[16] T.-S. Chang, "A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14 846–14 851, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2011.05.063

[17] E. Guresen, G. Kayakutlu, and T. U. Daim, "Using artificial neural network models in stock market index prediction," *Expert System Application*, vol. 38, no. 8, pp. 10 389–10 397, Aug. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2011.02.068

[18] P.-C. Chang, D.-D. Wang, and C.-L. Zhou, "A novel model by evolving partially connected neural network for stock price trend forecasting," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 611–620, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2011.07.051

[19] Y. Kara, M. Acar Boyacioglu, and O. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5311–5319, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.10.027

[20] W. Shen, X. Guo, C. Wu, and D. Wu, "Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm," *Know.-Based Syst.*, vol. 24, no. 3, pp. 378–385, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2010.11.001

[21] X. Zhu, H. Wang, L. Xu, and H. Li, "Predicting stock index increments by neural networks: The role of trading volume under different horizons," *Expert System Application*, vol. 34, no. 4, pp. 3043–3054, May 2008. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2007.06.023

[22] W.-H. Chen, J.-Y. Shih, and S. Wu, "Comparison of support-vector machines and back propagation neural networks in forecasting the six major asian stock markets," *International Journal of Electron Finance*, vol. 1, no. 1, pp. 49–67, Jan. 2006. [Online]. Available: http://dx.doi.org/10.1504/IJEF.2006.008837

[23] P.-C. Chang, C.-Y. Fan, and J.-L. Lin, "Trend discovery in financial time series data using a case based fuzzy decision tree," *Expert System Application*, vol. 38, no. 5, pp. 6070–6080, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.11.006

[24] M. R. Hassan, K. Ramamohanarao, J. Kamruzzaman, M. Rahman, and M. Maruf Hossain, "A HMM-based adaptive fuzzy inference system for stock market forecasting," *Neurocomput.*, vol. 104, pp. 10–25, Mar. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2012.09.017

[25] A. F. Sheta and K. De Jong, "Time-series forecasting using ga-tuned radial basis functions," *Inf. Sci. Inf. Comput. Sci.*, vol. 133, no. 3-4, pp. 221–228, Apr. 2001. [Online]. Available: http://dx.doi.org/10.1016/S0020-0255(01)00086-X

[26] A. Sheta, S. E. M. Ahmed, and H. Faris, "Evolving stock market prediction models using multi-gene symbolic regression genetic programming," *Artificial Intelligence and Machine Learning AIML*, vol. 15, pp. 11–20, 6 2015.

TABLE V
EVALUATION CRITERIA FOR THE DEVELOPED MODELS

	Regression		ANN		SVM-RBF	
	Training	Testing	Training	Testing	Training	Testing
Correlation coefficient	0.998	0.995	0.999	0.990	0.9995	0.9928
Mean absolute error	0.373	4.961	0.433	5.869	0.1976	2.6454
Root mean squared error	0.482	5.749	0.529	6.666	0.3263	3.0006
Relative absolute error	4.429%	11.838%	4.683%	58.335%	2.134%	7.993%
Root relative squared error	4.982%	12.313%	5.040%	57.569%	3.109%	8.5579%

[27] D. D. Thomakos, T. Wang, and J. Wu, "Market timing and cap rotation," *Math. Comput. Model.*, vol. 46, no. 1-2, pp. 278–291, Jul. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.mcm.2006.12.036

[28] S. Lahmiri, "Multi-scaling analysis of the S&P500 under different regimes in wavelet domain," *Int. J. Strateg. Decis. Sci.*, vol. 5, no. 2, pp. 43–55, Apr. 2014. [Online]. Available: http://dx.doi.org/10.4018/ijsds.2014040104

[29] S. Lahmiri, M. Boukadoum, and S. Chartier, "Exploring information categories and artificial neural networks numerical algorithms in S&P500 trend prediction: A comparative study," *Int. J. Strateg. Decis. Sci.*, vol. 5, no. 1, pp. 76–94, Jan. 2014. [Online]. Available: http://dx.doi.org/10.4018/IJSDS.2014010105

[30] S. T. A. Niaki and S. Hoseinzade, "Forecasting s&p500 index using artificial neural networks and design of experiments," *Journal of Industrial Engineering International*, vol. 9, no. 1, pp. 1–9, 2013.

[31] A. Sheta, H. Faris, and M. Alkasassbeh, "A genetic programming model for S&P 500 stock market prediction," *International Journal of Control and Automation*, vol. 6, pp. 303–314, 2013.

[32] A. Nigrin, *Neural networks for pattern recognition*, ser. A Bradford book. Cambridge, Mass, London: The MIT Press, 1993. [Online]. Available: http://opac.inria.fr/record=b1126290

[33] J. Leonard and M. A. Kramer, "Improvement of the backpropagation algorithm for training neural networks," *Computer Chemical Engineering*, vol. 14, pp. 337–343, 1990.

[34] A. K. Jain, J. Mao, and K. Mohiuddin, "Artificial neural networks: A tutorial," *IEEE Computer*, vol. 29, pp. 31–44, 1996.

[35] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[36] ——, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 5, pp. 988–999, 1999.

[37] Y. B. Dibike, S. Velickov, D. Solomatine, and M. B. Abbott, "Model induction with support vector machines: introduction and applications," *Journal of Computing in Civil Engineering*, 2001.

[38] R. Noori, M. Abdoli, A. A. Ghasrodashti, and M. Jalili Ghazizade, "Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: A case study of mashhad," *Environmental Progress & Sustainable Energy*, vol. 28, no. 2, pp. 249–258, 2009.

[39] W. Wen-chuan, X. Dong-mei, C. Kwok-wing, and C. Shouyu, "Improved annual rainfall-runoff forecasting using pso–svm model based on eemd," 2013.