A Support Vector Regression Model for Forecasting Rainfall

Nasimul Hasan[1], Nayan Chandra Nath[1], Risul Islam Rasel[2]

Department of Computer Science and Engineering, International Islamic University Chittagong, Bangladesh
nasimul_hsn@yahoo.com[1], nayanctg143@gmail.com[1]

***Abstract*—Rainfall prediction is a very important part of weather forecasting. In countries like Bangladesh; which has several seasons a year, rainfall prediction is really a key factor for many sectors. Rainfall data is a time series data and it changes time to time as climate and season changes. Moreover, rainfall depends on several factors as flow of wind, humidity etc., it is very challenging to make a hundred percent perfect prediction. This paper exhibits a robust rainfall prediction technique in view of the recent rainfall data of Bangladesh utilizing Support Vector Regression (SVR), a relapse methodology of Support Vector Machine (SVR). The collected raw data wasn't prepared for using as input of algorithm, thus it had been pre processed manually to suit into the algorithm, then fed to the algorithm. The evaluation results of the study conducted on the data shows that the projected technique performs higher than the conventional frameworks in term of accuracy and process running time. The proposed approach yielded the utmost prediction of 99.92%**

I. INTRODUCTION

Weather (especially rainfall) has a great influence over cultivation, flood and some other sectors, The cultivation of corps relies mostly on rainfall here and it is important to predict how much rainfall may happen in an upcoming season. A perfect rainfall prediction can help us to take decision what to cultivate and what to not. Since many other things such as drought, tourism, transportation, construction etc. also sometimes depends on rainfall; a good prediction can help to take decisions regarding these sectors. From ancient time people are trying to find the pattern of weather and predict weather for their well-being. From the very beginning of science and technology, weather prediction is a very interesting sector of study. Forecasting rainfall is a tough task due to the complication of the physics and different variable which cause rainfall [1]. It is actually a very noisy and deterministically disordered natural event. Some very important factors as monsoon, wind, moisture, heat, rotation of earth etc. plays a very strong role on the rainfall issue.

A lot of intelligent techniques as Artificial Neural have been used to predict rainfall. Among them Artificial Neural Network (ANN) is a commonly used technique to produce a good prediction [2] [3] [4]. Frequent study shows that ANN can work way better than different regression, MA and EMA. ANN frequently displays conflicting and unpredictable execution on boisterous data [5]. Support Vector Machine (SVM) is a supervised classification and regression algorithm. An SVM model represents the samples as point spaces. It is mainly based of decision plane concept. It separates objects with different class by a visible gap as much as possible between the classes. SVM can perform both linear and non-linear classification with super efficiency. Kesheng Lu and Lingzhi Wang [6] showed in their research that Support Vector Machine can perform an efficient rainfall prediction with a low error rate. Their model can produce almost 99% accurate prediction. A. Mellit, A. Massi Pavan & M. Benghanem [7] developed a SVM model which can produce up to 99% accurate prediction for different models. At the point when utilizing SVM, the fundamental issue is defied: how to pick the appropriate kernel and how to set the best kernel function. The best possible parameters setting can enhance the SVM relapse exactness. Diverse kernel capacity and distinctive parameter settings can bring about huge contrasts in execution. However, there are no investigative strategies on the other hand solid heuristics that can direct the client in selecting a fitting part capacity and great parameter values [6].

In this study, different kernel functions of Support Vector Machine (SVM) and an exclusive data preprocessing technique windowing is combined to predict rainfall. Three different models- Total rainfall prediction, maximum rainfall prediction and average rainfall prediction with three different sub models, for 1day ahead, 7 day ahead and 10 day ahead prediction are proposed here for a better rainfall prediction technique.

II. METHODOLOGY

ϵ-insensitive loss function was introduced by Vapnik and then SVM was developed to solve regression problems as well. The process is known as Support Vector Regression (SVR). SVR has an excellent performance regarding the perfection of solving non-linear regression.

Support Vector Regression: Almost all principals of SVM classification are followed by Support Vector Regression (SVR). As the output is real number and there are unlimited possibilities. To handle this problem, a limitation to the tolerance (epsilon) is set to the SVM which would have effectively asked for from the problem. It minimizes $||\omega||^2$ to reduce the model complexity. Here ξ_i and ξ_i^* are slack variables and $i = 1, \ldots, n$ to calculate the difference of training data outside sensitive zone [8] [9].

Minimize:

$$\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (1)$$

Subject to:

978-1-4673-9257-0/15/$31.00

$$\begin{cases} y_i - f(x_{i,\omega}) \leq \epsilon + \xi_i^* \\ f(x_{i,\omega}) - y_i \leq \epsilon + \xi_i^* \\ \xi_i,\, \xi_i^* \geq 1, \ldots, n \end{cases} \tag{2}$$

Kernel Functions:

$$K(X_i, X_j) = \begin{cases} X_i * X_j & Linear \\ (\gamma X_i * X_j + C)^d & Polynomial \\ exp(-\gamma |X_i - X_j|^2) & RBF \\ tanh(\gamma X_i * X_j) + C & Sigmoid \end{cases} \tag{3}$$

Here, $K(X_i, X_j) = \phi(X_i, X_j)$, and γ is the adjustable parameter.

A. Time series data

Rainfall data is time series data. Time series data is a set of values of something where the intervals of the value are listed in such a way that the time period between each attribute is exactly same.

B. Moving Average

A Moving Average (rolling average or running average) is an estimation to analyze data points by making a progression of midpoints of distinctive subsets of the full data set.

$$MA = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{4}$$

Here, $x_1, x_2, \ldots, x_n$ are the average values of the monthly rainfall and n is the total number of observation. Moving average of 7 days was considered for all attributes taken as labels in this study.

C. Windowing operator

Windowing operator is an exclusive operator which can perform better for time series prediction. It converts series sample data into single valued data. The series data must be given as Example Set. The parameter "series representation" defines how the series data is represented by the Example Set.

III. EVALUATION PROCESS

A. RMSE

Root Mean Square Error is a well-known and commonly used evaluation process for regression models.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (yt - \hat{yt})^2}{n}} \tag{5}$$

Here, yt is the original value of a point for a given time period t, n is the total number of fitted points, and $\hat{yt}$ is the fitted forecast value for the time period t.

B. MAE

Mean Average Error (MAE) is an widely used evaluation formula. Mean Absolute Error (MAE) measures how far predicted values are away from observed values.

$$MAE = \frac{SAE}{N} = \frac{\sum_{i=1}^{n} |x_i - \hat{x_i}|}{N} \tag{6}$$

Here, x_i is the actual observations time series, is the estimated or forecasted time series, SAE is the sum of the absolute errors (or deviations), N is the number of non-missing data points.

IV. EXPERIMENT DESIGN

A. Research Data

6 years Rainfall data (2008-2014) of Chittagong, Bangladesh from the Meteorological Department, Bangladesh were collected to perform experiment and evaluation this study. Here we took only the data of March to October of each year as basically in this period rainfall occurs here normally. Three different models for forecasting maximum, total r and average rainfall was built with the same data. 80% of the data were considered as training data and the rest 20% as testing data. Six attributes, Date, total, avg, max, min, MA included both the training and test data. The 'Date' attribute was selected as id for all three models and another attribute ('max' for maximum rainfall prediction, 'avg' for average rainfall prediction and 'total' for total rainfall prediction) was chosen as label. Figure 1 shows the actual rainfall (2008-2014)

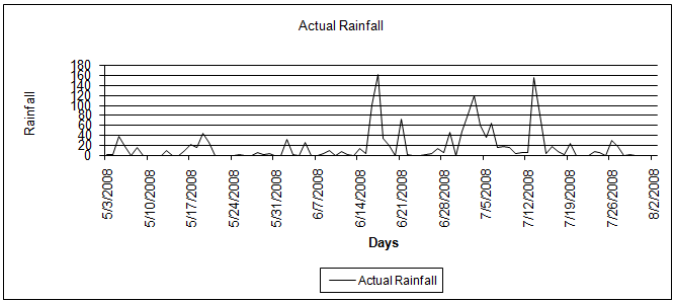

Fig. 1. Actual Rainfall of Chittagong (Total,2008-2014)

B. Model construction and analysis

This step starts with the preprocessing step of data. At the very first Moving Average (MA) technique was applied on the data set. Then windowing operator was used so that it can convert series data to universal dataset and feed the learning process [10]. As learning technique we used Support Vector Regression (SVR). The model structure includes two vital stages, training and testing. Both the stages are continuous process as the main objective of the study was to build a efficient model for rainfall forecasting.

Training and testing stages are mentioned below.

C. Training stage

The training stage starts with the retrieving of data from the repository. The data were processed before entering this stage. After this, the role of the attributes in the dataset was assigned. Then one of the attributes (Date) was selected as the id and another as the label. The selection of label was different for three different models. The next step was to set the windowing operator. The parameters of windowing operator were set and then windowing was performed. A special and efficient validation 'Sliding Window Validation' was performed for validation. Afterwards, the main process was fed and the parameters of the Support Vector Regression. Then the model was set to run.

D. Testing stage

In this phase the testing data were retrieved from the repository and the role of the attributes of the data was set and then the id and label were selected as done in the training stage. Like the training stage, the parameters of windowing operator were set and run and then the main process was fed. The performance was compared finally.

V. EXPERIMENT RESULT

A. Windowing operator analysis

At this point we used windowing operator to the time series data to convert it into generic data. Table I shows the windowing operators parameter settings used in this study.

TABLE I
WINDOWING OPERATOR ANALYSIS

Model		Horizon	Window size	Step Size
Total	1 day	1	2	1
	7 day	7	2	1
	10 day	10	2	1
Maximum	1 day	1	3	1
	7 day	7	2	1
	10 day	10	2	1
Average	1 day	1	2	1
	7 day	7	2	1
	10 day	10	2	1

B. Support Vector kernel Analysis

The parameters selecting of Support Vector Regression is a very crucial and important part. The perfection of prediction depends on best setting of parameters. So that, the parameters were chosen carefully and a log was kept to find the best parameter set. Table II shows the parameter settings for Support vector used in this study. Here, C is the constant of complexity and G is the value of kernel gamma.

C. Sliding window validation

As the validation process for this study sliding window validation was used. It is a special validation for series data. It takes a certain window for training and another window for testing. The window then moves to another window and determine the average of all estimation. The parameter "cumulative training" shows if every single previous sample ought to be utilized for preparing (rather than just the present window) [11]. Table III shows the parameter properties of Sliding Window validation and Table IV shows the result of the study.

TABLE II
SUPPORT VECTOR KERNEL ANALYSIS

Model	Days	Kernel Type	Degree	C	G	ϵ	$\epsilon+$	$\epsilon-$
Total	1 day	Anova	1	2000	1	1	1	1
	7 day		1	1000	1	1	1	1
	10 day		1	1000	1	1	1	1
Maximum	1 day	Anova	1	400	1	1	1	1
	7 day		1	700	1	1	1	1
	10 day		1	1000	1	1	1	1
Average	1 day	Anova	1	1000	1	1	1	1
	7 day		1	1000	1	1	1	1
	10 day		1	1000	1	1	1	1

TABLE III
SLIDING WINDOW VALIDATION

Model	Training Window Width	Training Window step	Test Window width	Horizon	Cumulative training
Total	2	1	2	1	No
	2	1	2	7	No
	2	1	2	10	No
Maximum	2	1	2	1	No
	2	1	2	7	No
	2	1	2	10	No
Average	2	1	2	1	No
	2	1	2	7	No
	2	1	2	10	No

D. Error calculation

The error calculated from the actual rainfall value and the predicted rainfall value from the Support Vector Regression model proposed in this study.

E. Graphical Representation of the study

Two different models- Maximum rainfall prediction and Total rainfall prediction are proposed here with 1 day ahead, 7 day ahead and 10 day ahead models for each. Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, Figure 9 and Figure 10 shows the difference between the actual value of rainfall (mm) and the predicted value of rainfall (mm).

TABLE IV
RESULT FOR 1 DAY AHEAD PREDICTION

Model	Date	Actual Rainfall	Predicted Rainfall 1 day
Total	4-Jul-12	0.1	0.19
	5-Jul-12	2.5	1.76
	6-Jul-12	0.0	0.74
	7-Jul-12	1.6	1.17
Maximum	4-Jul-12	0.1	0.85
	5-Jul-12	2	1.26
	6-Jul-12	0	1.04
	7-Jul-12	1	1.11
Average	4-Jul-12	0.0125	0.76
	5-Jul-12	0.3125	0.76
	6-Jul-12	0	0.68
	7-Jul-12	0.2	0.81

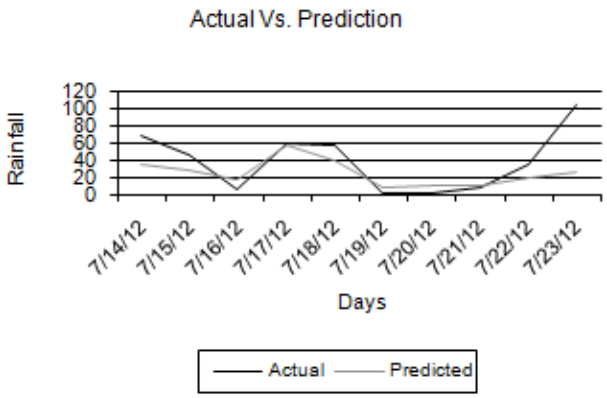

Fig. 2. Actual Vs. Prediction (Total, 1 day Ahead)

TABLE V
RESULT FOR 7 DAY AHEAD PREDICTION

Model	Date	Actual Rainfall	Predicted Rainfall 1 day
Total	10-Jul-12	6.8	1.81
	11-Jul-12	4.6	6.97
	12-Jul-12	78	107.79
	13-Jul-12	115	30.54
Maximum	10-Jul-12	6	1.47
	11-Jul-12	4.6	1.90
	12-Jul-12	53.8	5.48
	13-Jul-12	106	1.52
Average	10-Jul-12	0.85	1.17
	11-Jul-12	0.575	0.06
	12-Jul-12	9.75	14.49
	13-Jul-12	14.375	5.72

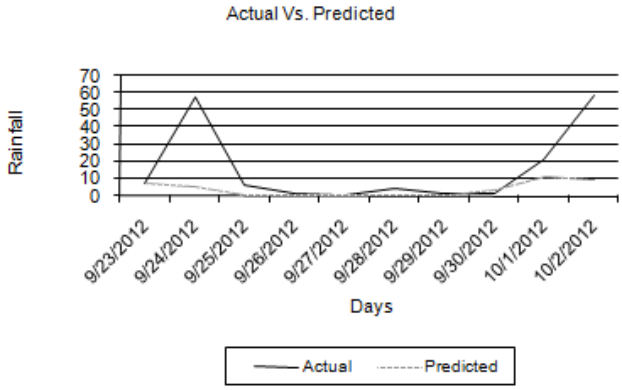

Fig. 3. Actual Vs. Prediction (Total, 7 day Ahead)

TABLE VI
RESULT FOR 10 DAY AHEAD PREDICTION

Model	Date	Actual Rainfall	Predicted Rainfall 1 day
Total	13-Jul-12	115	28.65
	14-Jul-12	66.2	7.83
	15-Jul-12	44.1	1.41
	16-Jul-12	5	2.41
Maximum	13-Jul-12	106	14.10
	14-Jul-12	40	0.38
	15-Jul-12	16	2.63
	16-Jul-12	4	2.91
Average	13-Jul-12	14.375	10.74
	14-Jul-12	8.275	2.93
	15-Jul-12	5.5125	0.51
	16-Jul-12	0.625	1.88

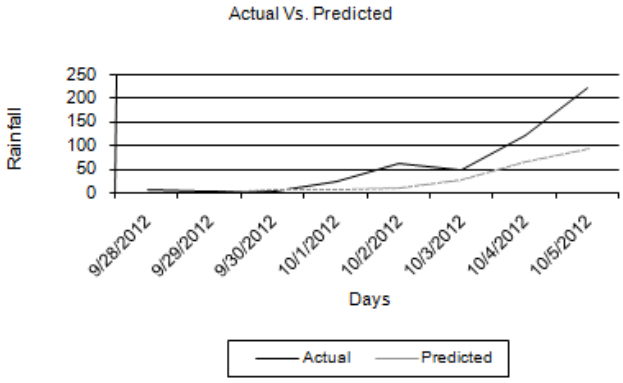

Fig. 4. Actual Vs. Prediction (Total, 10 day Ahead)

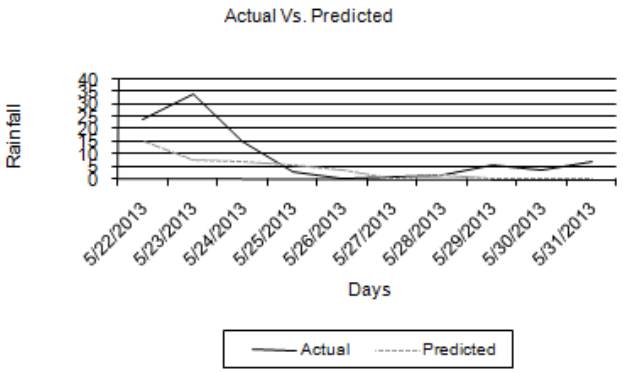

Fig. 5. Actual Vs. Prediction (Max, 1 day Ahead)

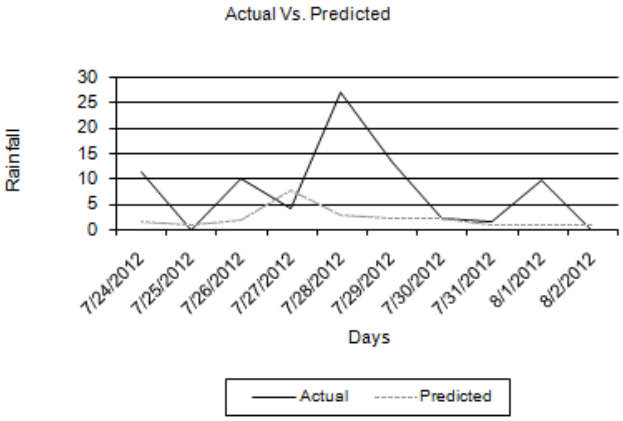

Fig. 6. Actual Vs. Prediction (Max, 7 day Ahead)

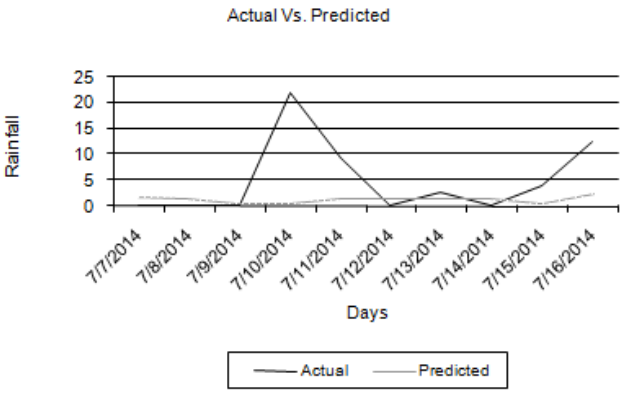

Fig. 7. Actual Vs. Prediction (Max, 10 day ahead)

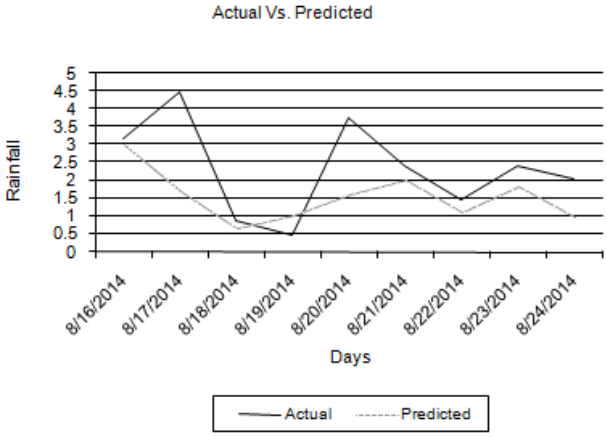

Fig. 8. Actual Vs. Prediction (Average, 1 day ahead)

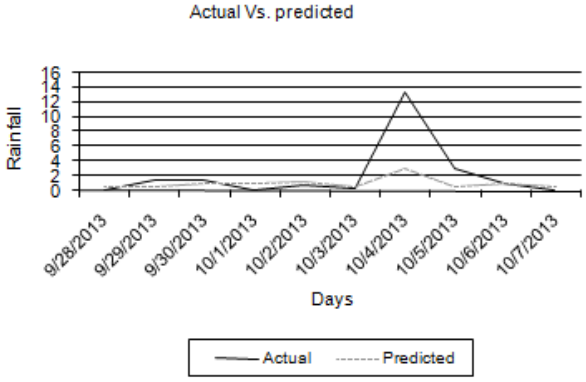

Fig. 9. Actual Vs. Prediction (Average, 7 day ahead)

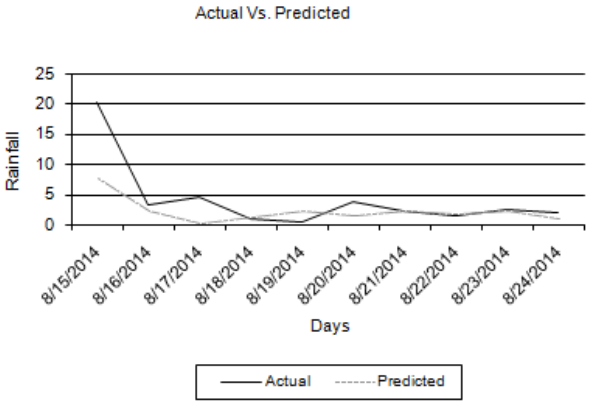

Fig. 10. Actual Vs. Prediction (Average, 10 day ahead)

TABLE VII
ERROR CALCULATION

Model	Days	RMSE	MAE
Total	1 day	18.60	0.18
	7 day	24.95	4.99
	10 day	24.97	86.35
Maximum	1 day	11.39	0.75
	7 day	15.99	4.53
	10 day	14.26	91.90
Average	1 day	2.85	0.45
	7 day	3.30	0.31
	10 day	3.54	3.63

VI. CONCLUSION AND FUTURE WORK

A. Discussion

This paper presents a robust rainfall prediction technique using Support Vector Regression. The result of the experiment done with the data of Chittagong, Bangladesh shows that the proposed model can forecast more accurately in comparison with the regular technique used. The proposed model can predict better than any model before one day. The 7 days ahead model also performs very well than conventional processes.

B. Limitation and Future Work

Only Moving Average and Windowing operator were used as data preprocessing step in this study. The data was taken for only Chittagong area of Bangladesh. In future more data preprocessing steps and different algorithms will be used. Data from other stations of different countries and area will be used to get a universal model.

REFERENCES

[1] L. Xiong and K. M. O'Connor, "An empirical method to improve the prediction limits of the glue methodology in rainfall–runoff modeling," *Journal of Hydrology*, vol. 349, no. 1, pp. 115–124, 2008.

[2] J. Wu, L. Huang, and X. Pan, "A novel bayesian additive regression trees ensemble model based on linear regression and nonlinear regression for torrential rain forecasting," in *Computational Science and Optimization (CSO), 2010 Third International Joint Conference on*, vol. 2. IEEE, 2010, pp. 466–470.

[3] J. Wu and E. Chen, "A novel nonparametric regression ensemble for rainfall forecasting using particle swarm optimization technique coupled with artificial neural network," in *Advances in Neural Networks–ISNN 2009*. Springer, 2009, pp. 49–58.

[4] G.-F. Lin and L.-H. Chen, "Application of an artificial neural network to typhoon rainfall forecasting," *Hydrological Processes*, vol. 19, no. 9, pp. 1825–1837, 2005.

[5] W.-C. Hong, "Rainfall forecasting by technological machine learning models," *Applied Mathematics and Computation*, vol. 200, no. 1, pp. 41–57, 2008.

[6] K. Lu and L. Wang, "A novel nonlinear combination model based on support vector machine for rainfall prediction," in *Computational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on*. IEEE, 2011, pp. 1343–1346.

[7] A. Mellit, A. M. Pavan, and M. Benghanem, "Least squares support vector machine for short-term prediction of meteorological time series," *Theoretical and applied climatology*, vol. 111, no. 1-2, pp. 297–307, 2013.

[8] L. K. Lai, "stock forecasting using support vector machine," *Machine Learning*, 2010.

[9] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[10] R. I. Rasel, N. Sultana, and P. Meesad, "An efficient modelling approach for forecasting financial time series data using support vector regression and windowing operators," *International Journal of Computational Intelligence Studies*, vol. 4, no. 2, pp. 134–150, 2015.

[11] L. K. Lai, "stock forecasting using support vector machine," *Machine Learning*, 2010.