A Hybrid of Multiple Linear Regression Clustering Model with Support Vector Machine for Colorectal Cancer Tumor Size Prediction

Muhammad Ammar Shafi[1], Mohd Saifullah Rusiman[2], Shuhaida Ismail[3], Muhamad Ghazali Kamardan[4]
Department of Mathematics and Statistics
Universiti Tun Hussein Onn Malaysia, 86400 Pagoh Muar, Johor, Malaysia

***Abstract*—This study proposed the new hybrid model of Multiple Linear Regression Clustering (MLRC) combined with Support Vector Machine (SVM) to predict tumor size of colorectal cancer (CRC). Three models: Multiple Linear Regression (MLR), MLRC and hybrid MLRC with SVM model were compared to get the best model in predicting tumor size of colorectal cancer using two measurement statistical errors. The proposed model of hybrid MLRC with SVM have found two significant clusters whereby, each clusters contained 15 and three significant variables for cluster 1 and 2, respectively. The experiments found that the proposed model tend to be the best model with least value of Mean Square Error (MSE) and Root Mean Square Error (RMSE). This finding has shed light to health practitioner in determining the factors that contribute to colorectal cancer.**

***Keywords*—Colorectal cancer; multiple linear regression; support vector machine; fuzzy c- means; clustering; prediction**

I. Introduction

The colon and rectum is the final portion of the human body digestion tube. The food that humans eat will go through the stomach from mouth to anus. In the stomach, the food is grinded into smaller particle and then enters the small intestine in a careful and controlled manner. The small intestine is where the final stage of food digestion and absorption of the nutrients contained in the food take place. The food that is not digested and absorbed will enter the large intestine or colon and finally to the rectum. In addition, some of the undigested foods accumulated through the years produce bacteria and causes cancer and it is called colorectal cancer. Colorectal cancer is a type of cancer that arise from the inner wall of large intestine [1,2].

However, a cause of colorectal cancer is still unclear. It involves many risk factors including family history, colon polyp and long-standing ulcerative colitis. Symptoms of colorectal cancer are also unclear for detection. Moreover, some of the symptoms of colorectal cancer are too common in the society like anemia, weight loss and many more [3].

Furthermore, information and knowledge about risk of colorectal cancer in Malaysia is still lacking compared to the awareness towards cervical cancer [4, 5, 6, 7, 8]. It might have been one of the reasons behind the increasing number of patient suffering colorectal cancer. It was reported that colorectal cancer causes the third highest number of death among patients after lung cancer and breast cancer by 10.6% [9]. Data in 1995 showed colorectal cancer admission percentage increased from 8.1% to 11.9% [10].

Basically, there are four stages of colorectal cancer. Earlier stages comprise of stage I and II and final stages refer to stage III and IV. According to Malaysian Oncology Society in 2017, stage I refer to the condition where the cancer start to exist in wall of colon or polyp, stage II, III and IV refer to the condition where the cancer has spread through the wall of the colon. There are several methods to reduce the risk of developing colorectal cancer such as treatment, radiotherapy and screening [11, 12, 13]. Nowadays, linear regression and support vector machine model are very popular model among researchers in dealing with various fields [14, 15]. This study plans to find the best model among MLR, MLRC and hybrid model by comparing the value of errors.

II. Materials and Methods

The population of this study is based on secondary data from patients aged between 21 until 90 years old who received treatment at a general hospital around Kuala Lumpur in 2012 with the symptoms or suffering from colon cancer of any four stages. The study includes both male and female patients from various ethnics.

A. Multiple Linear Regression Model (MLR)

Regression analysis proposed by Sir Francis Galton in 19th century who study the relation between the heights of parents and the height of their children. The term regression persists to describe statistical relations between variables. Regression analysis is a statistical method that utilizes the relation between two or more quantitative variables toward predicted variables. MLR model is one of statistical methods which is widely used in many disciplines such as business, the social and behavioral sciences and biological sciences [17].

MLR model uses the least squares estimation technique in order to find the coefficient β_i. Before conducting multiple linear regression analysis, regression model should fulfill the classical assumptions as stated below:

1) Constant variance of residual
2) Residual of normality
3) Multicollinearity checking

The MLR model parameter can be stated as follows:

$$Y_i=\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_k X_{ik} + \varepsilon_i(\beta),\ i=1,\ldots,N \quad (1)$$

where:

Y_i is the value of response variable

B_0, β_1, β_2 and β_j are unknown constant

X_y is value of predictor variable

ε_i is the random error

Regression model is a linear model with multiple linear predictor variables. It is 'multiple' because there are more than one predictor variables in linear parameter. A model which is linear predictor in the parameters is refered as a first-order model.

B. Fuzzy C-Means Methods

Fuzzy C-Means (FCM) is a method of clustering which allows the data to belong into two or more clusters.This model developed by Dunn (1973) and was improved by Bezdek (1981). A large family of fuzzy clustering algorithms is based on minimization of the fuzzy c-means objective function formulated as:

$$J = \sum_{q=1}^{N}\sum_{r=1}^{C} u^2_{qr} d^2_{qr} \quad (2)$$

where z is any real number greater than 1, μ_{qr} is the membership values, d_{qr} represent as the distance according to Euclidean. N is the number of objects and C is the number of clusters. The index q (q=1,..., N) correspond to object number q and the index r (r=1,..,C) correspond to cluster number r. In case of Euclidean distance, the algorithm for minimising J can be summarized by the following steps:

1) Randomly select cluster centers 'c'. Choose the termination tolerance between 0 and 1, then choose fuzziness exponent, $z > 1$.

2) Update distance, d_{qr} for given μ_{qr} by computing the weighted average for each group and the Euclidean distance as:

$$d^2_{qr} = \| x_q - v_r \|^2 \quad , \quad v_r = \frac{\sum_{q=1}^{N} u_{qr}{}^2 X_q}{\sum_{q=1}^{N} u_{qr}{}^z} \quad (3)$$

3) Update membership values as,

$$u_{qr} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{qr}}{d_{qk}}\right)^{\frac{2}{z-1}}} \text{, for } z=1 \quad (4)$$

4) Calculate the objective or criterion J and make iteration in order to minimize the objective function. The iteration repeated for k = 1, 2, …, ∞, then stop the iteration, else repeated step 2.

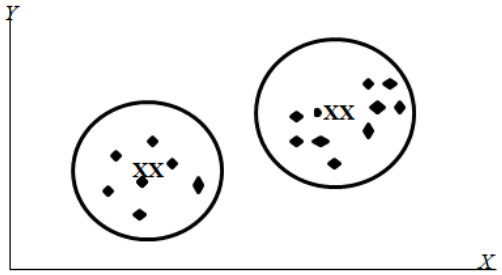

Fig. 1. The FCM Graph with Two Clusters.

The example of FCM cluster is shown in Fig. 1. The clusters have naturally circular shape with the clusters, XX located in the middle for each cluster.

C. Support Vector Machines Model (SVM)

SVM model was proposed by Vapnik in 1963 in order to determine the subtle patterns in complex data sets. An SVM commonly used for classifying objects in prediction or forecasting technique The advantage of SVM is that it has two classifications types which are linear and non-linear [16].

SVM with linear classification is applied in this study. Linear classifier is used to determine to which group an object belongs to. It is done by dividing the groups with a line called hyperplane. The hyperplane linear classification is shown in Fig. 2.

In addition, equation of kernel functions with polynomial kernel as below [18]:

$$K(x, y) = (x^T y + 1)^d \quad (5)$$

D. Hybrid MLR Clustering with SVM Models

Multiple linear regression clustering is proposed in this study which is combination of MLR and FCM method. While, the hybrid is defined as a combination of both MLR clustering model and SVM model [18]. The steps of hybrid MLR clustering with SVM models is shown in Fig. 3.

This study contains three stages in preparing the new hybrid MLRC and SVM model. The first stage is MLR and SVM model will be applied to colorectal cancer data to determine the MSE and RMSE value of the model. In stage two, Pearson correlation was performed to find the highest correlation among the independent variables. Then, the selected variables are put into fuzzy c-means to find the appropriate clustering.

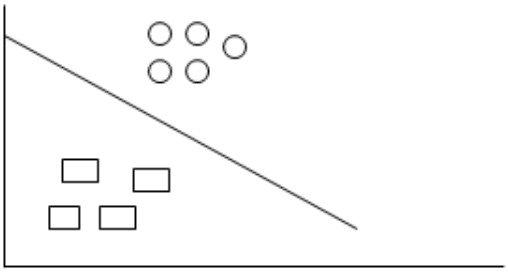

Fig. 2. Linear Classification with a Hyperplane.

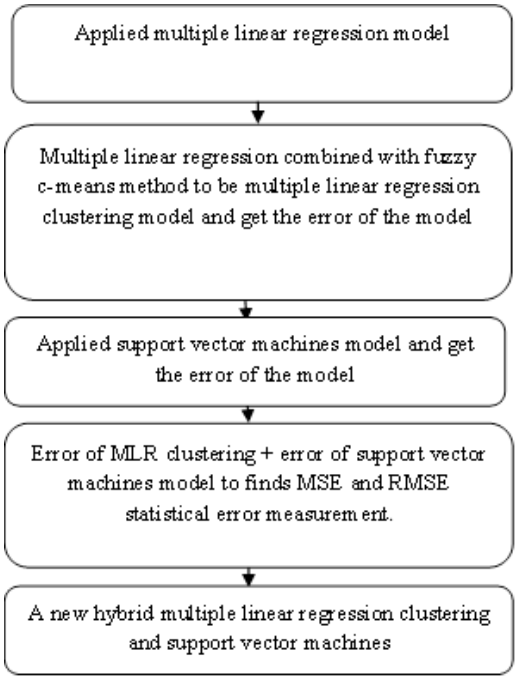

Fig. 3. Steps of Hybrid Model.

Finally, in stage three, the new dataset is obtained based on combination of MLRC and SVM to predict the tumor size of colorectal cancer. The equation for a new dataset as shown is Eq (6).

$$Y = L_t + N_t \tag{6}$$

Where, Y is a new dataset, L_t is the error of linear part (MLR) and N_t is the error of nonlinear part (SVM) of the hybrid model.

The performance for each cluster is evaluated using statistical performances which are MSE and RMSE values. The final equation to find MSE value is as follow:

$$MSE_{final} = \frac{n_1(MSE_1) + N_2(MSE_2)}{n_1 + n_2} \tag{7}$$

Where, n_1 *and* n_2 are the number of data for cluster 1 and cluster 2 respectively. MSE_1, MSE_2 is the value of mean square error for cluster 1 and cluster 2, respectively.

III. Results

This study used secondary data of tumor size for colorectal cancer and contains 180 patients as respondents. The dependent variable is tumor size while 25 factors and symptoms variables are chosen as independent variables. The average age of patients facing colorectal cancer and symptoms is 61 years old. While, the average tumor size in diameter (mm) is 53.45.

The comparison among MLR, MLRC and hybrid of MLRC with SVM model are compared using cross validation statistical techniques which are MSE and RMSE. The model with the lowest of MSE and RMSE value is chosen as the best model to predict the tumor size of colorectal cancer.

A. Multiple Linear Regression Model

The data in this study were tested using three assumptions of MLR which is residual of constant variance, residual of normality and multi-collinearity and all the assumptions were satisfied and fulfilled.

This study applied MLR model in the data of tumor size colorectal cancer with 25 independent variables. Table I shows the parameter of the model.

All the significant variables chosen are important to predict tumor size of colorectal cancer. The estimated multiple linear regression model for colorectal cancer is as follows:

$\hat{Y}$=76.056 + 0.421 age + 3.459 icd10 + 0.961 TNM staging– 16.738 family history + 5.035 Crohn's disease + 5.557 history of cancer – 6.517 gastric + 12.865 ovarian – 4.350 intestinal obstruction –7.943 anemia –3.994 abdominal (8)

TABLE I. Parameter of the Model

Independent variables	Beta (β)	Sig. Value
(Constant)	76.056	*0.000
x_1 (Gender)	1.977	0.286
x_2 (Age at Diagnosis (years))	-0.421	*0.000
x_3 (Ethnic Group)	2.231	0.114
x_4 (ICD 10 Site)	3.459	*0.027
x_5 (TNM Staging)	0.961	*0.040
x_6 (Family_History)	-16.738	*0.000
x_7 (Diabetes Mellitus)	0.832	0.679
x_8 (Crohn's Disease)	5.035	*0.012
x_9 (Ulcerative colitis)	1.508	0.432
x_{10} (Polyp)	-1.577	0.396
x_{11} (History of cancer)	5.557	*0.007
x_{12} (Endometrial)	-0.869	0.648
x_{13} (Gastric)	-6.517	*0.001
x_{14} (Small bowel)	-3.33	0.087
x_{15} (Hepatobiliary)	-1.047	0.589
x_{16} (Urinary tract)	2.597	0.176
x_{17} (Ovarian)	12.865	*0.000
x_{18} (Other cancer)	3.248	0.084
x_{19} (Intestinal Obstruction)	-4.35	*0.022
x_{20} (Colorectal)	1.227	0.511
x_{21} (weight_loss)	3.063	0.1
x_{22} (Diarrhoe)	0.75	0.695
x_{23} (Anemia)	-7.943	*0.003
x_{24} (blood_stool)	-2.158	0.233
x_{25} (Abdominal)	-3.994	*0.038

*Significant at 0.05

TABLE II. ANOVA FOR MULTIPLE LINEAR REGRESSION

Sources	Sum of Squares	df	Mean Square	F-Value	P-Value
Regression	21396.590	25	855.864	6.606	0.000
Residual	19951.960	154	129.558 (MSE)		11.3826 (RMSE)
Total	41348.550	179			

In addition, Analysis of Variance (ANOVA) analysis was performed. The result shows that the MSE term is 129.558 and RMSE is 11.3826. The result of ANOVA is shown in Table II.

B. Multiple Linear Regression with Fuzzy C-Means Method

The three assumptions applied for MLR in cluster 1 and cluster 2 models were fulfilled and satisfied. Furthermore, the independent variables x17 (ovarian), x8 (crohn's disease), x5 (TNM staging), x11 (history of cancer) and x1 (gender) were chosen since it has the highest correlation value. The data will be divided into two cluster, cluster 1 and cluster 2. The correlation values among x17, x8, x5, x11 and x1 as shown below in Table III:

Based on Table III, y vs x_{17} show the best cluster results. The final MSE value is calculated based on equation 6. The amount of respondents in the data taken with 180 patients. The results are in Table IV.

1) Cluster 1 (Y vs X17): Cluster 1 (based on X_{17}) for the clustering model between MLR and FCM used 85 data as respondent in the analysis. The measurement of error MSE and RMSE is shown in Table V. The model for cluster 1 is as follow:

$\hat{Y}_1$ = 64.278 - 0.573 age + 4.321 ethnic + 5.836 icd10 + 2.024 TNM Staging - 9.670 family history + 3.130 Crohn's disease + 3.127 ulcerative colitis + 4.924 history of cancer - 6.784 small bowel + 3.414 urinary tract + 13.736 ovarian + 4.193 other cancer – 6.583 intestinal obstruction + 4.677 weight loss + 5.555 diarrhea - 4.315 blood stool. (9)

TABLE III. CORRELATION VALUES

correlation	value	significant
Y vs x_{17} (1)	0.246	0.001
Y vs x_8 (2)	0.145	0.052
Y vs x_5 (3)	0.141	0.058
Y vs x_{11} (4)	0.117	0.119
Y vs x_1 (5)	0.089	0.237

TABLE IV. THE FINAL OF MSE AND RMSE VALUE OF THE MODEL

Correlation	Final MSE
Y vs x_{17}	116.985
Y vs x_8	123.560
Y vs x_5	116.985
Y vs x_{11}	116.985
Y vs x_1	123.560

TABLE V. MSE AND RMSE VALUE OF MODEL CLUSTER 1

Methods	Value
MSE	32.763
RMSE	5.724

TABLE VI. MSE AND RMSE VALUE OF MODEL CLUSTER 2

Methods	Value
MSE	192.341
RMSE	13.868

2) Cluster 2 (Y vs X_{17}): Cluster 2 (based on X_{17}) for the clustering model between MLR and FCM used 95 data as respondent in the analysis. The measurement of error MSE and RMSE is shown in Table VI. The model for cluster 2 is as follow:

$\hat{Y}_2$ = 78.073–22.907 family history – 13.157 gastric + 12.454 ovarian (10)

C. Hybrid MLR Clustering with SVM Model

A measurement statistical error of MSE and RMSE and coefficient of parameter model was applied in a new model hybrid of MLRC and SVM. The error measurement of MSE and RMSE could be evaluated by sum error of MLRC and SVM model to determine the value of MSE and RMSE. The smallest value of error will be the best model to predict tumor size of colorectal cancer. The final MSE and RMSE of the model show in Table VII.

1) Cluster 1 (Y vs X_{17}): Cluster 1 (based on X_{17}) for the clustering model between MLR clustering and SVM used 85 data as respondent in the analysis. The MSE and RMSE values and error of cluster 1 are shown in Table VIII and Fig. 4, respectively.

TABLE VII. MSE AND RMSE VALUE OF MODEL

Methods	Value
MSE	78.661
RMSE	8.869

TABLE VIII. MEASUREMENT ERROR OF CLUSTER 1

Methods	Value
MSE	44.979
RMSE	6.707

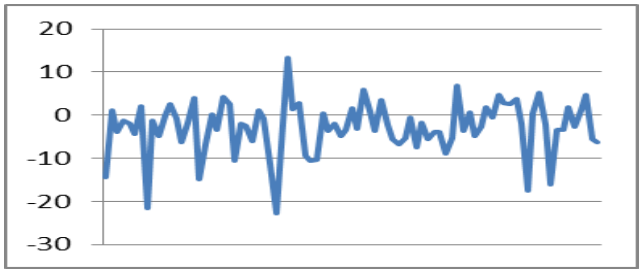

Fig. 4. The Error of Cluster 1.

There are several independents variables of the model which are significant to predict tumor size of colorectal cancer and the MLRC and SVM model in Cluster 1 for the model is as follows:

$\hat{Y}$ = 64.278 - 0.573 age + 4.321 ethnic + 5.836 icd10 + 2.024 TNM Staging - 9.670 family history + 3.130 Crohn's disease + 3.127 ulcerative colitis + 4.924 history of cancer - 6.784 small bowel + 3.414 urinary tract + 13.736 ovarian + 4.193 other cancer – 6.583 intestinal obstruction + 4.677 weight loss + 5.555 diarrhoea - 4.315 blood stool (11)

2) Cluster 2 (Y vs X_{17}): Cluster 2 (based on X_{17}) for the clustering model between MLRC and SVM model used 95 data as respondent in the analysis. The three assumptions were fulfilled and satisfied. The MSE and RMSE values and error of cluster 2 are shown in Table IX and Fig. 5, respectively.

TABLE IX. MEASUREMENT ERROR OF CLUSTER 2

Methods	Value
MSE	107.482
RMSE	10.367

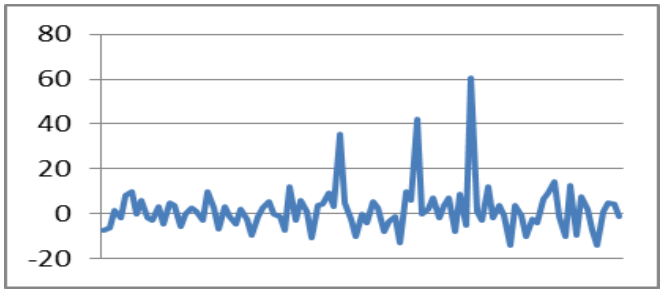

Fig. 5. The Error of Cluster 2.

The MLRC and SVM model for colorectal cancer model in Cluster 2 is then predicted as follows:

$\hat{Y}$=78.073–22.907 family history–13.157 gastric + 12.454 ovarian (12)

IV. CONCLUSION

This study proposed a hybrid of MLRC with SVM model. The MSE and RMSE have been used to measure the effectiveness of the model in predicting tumor size of colorectal cancer based on the factors and symptoms of colorectal cancer.

It was found that the proposed model of MLRC with SVM has yield a good result in predicting the tumor size of colorectal cancer suffered by patients in general hospitals of Kuala Lumpur. The results showed that MSE and RMSE for a new hybrid of MLRC and SVM model are 78.661 and 8.869 respectively.

Based on hybrid MLRC with SVM, there are 16 significant independent variables that are found to be a contributed factors to colorectal cancer which are age, ethnic, icd10, TNM staging, family history, Crohn's disease, ulcerative colitis, history of cancer, small bowel, urinary tract, ovarian, other cancer, intestinal obstruction, weight loss, diarrhea and blood stool. On the other hand, only three independent variables are significant in cluster 2 which are family history, gastric and ovarian. The summary of the error of the models is shown in the Table X.

TABLE X. SUMMARY ERROR OF MODELS

Model of linear regression	MSE	RMSE
MLR	129.558	11.383
MLR clustering	116.985	10.816
A new hybrid model (MLR clustering and support vector machines model)	78.661	8.869

In future, the new hybrid model of MLRC and SVM model can be applied to various fields especially for vagueness data and complex data. Moreover, the study using hybrid model can be used in order to predict the factors and symptoms of colorectal cancer and hence, can reduce the mortality rate.

ACKNOWLEDGMENT

This research is supported by the Universiti Tun Hussein Onn Malaysia under the TIER 1 grant scheme vot number H232.

REFERENCES

[1] ''Healthline,'' colon cancer, 2011. Retrieved from healthline.com

[2] ''MedicineNet,'' colon cancer, 2011. Retrieved from http''//MedicineNet.com

[3] Obi, J. C, Imianvan, A. A, ''Fuzzy neaural approach for colon cancer prediction,'' Scientia Africana, vol. 11, pp. 65-76, 2012.

[4] Harmy, M.Y, Norwati, D, Norhayati, M.N, Amry, A.R, ''Knowledge and attitude of colorectal cancerscreening among moderate risk patients in west Malaysia,'' Asian Pacific J Cancer prev, vol. 12, pp. 1957-60, 2011.

[5] Hoffman, R. M., Rhyne, R. L., Helitzer, D. L, ''Barriers of colorectal cancer screening: physician and general population perspectives, New mexico, 2006,'' Preventing Chronic Diseases, vol. 289, pp. 2492-2493, 2011.

[6] Lim, G.C.C, ''Overview of cancer in Malaysia,'' Jpn Jclin Oncol, vol. 32, pp. s37-42, 2011.

[7] ''Second report of the national cancer registry,', 2010.. [http:// www.makna.org.my/NCR/].

[8] Vincent, J. Hochhalter, A. K., Broglio, K., Avots-Avotin, A. E., ''Surveys respondents planning to have screening colonoscopy report unique barriers,'' The Permanente Journal, vol. 15, pp. 4-11, 2011.

[9] World Health Organization Data, ''Publications of the World Health Organization,'' The WHO Web Site, 2010.

[10] Ministry of Health, Malaysia, ''Information and Documentation System Unit. Planning & Development Division,'' 2010.

[11] Pereira M. Graca, Ana Paula Figuiredo and Frank D. Fincham, ''Anxiety, Depression, Traumatic Stress and Quality of Life in Colorectal Cancer after Different Treatments: A Study With Portuguese Patients and Their Partners,'' European Journal of Oncology Nursing, pp. 1-6, 2011.

[12] Roslani April Camilla, Taufiq Abdullah and Kulenthran Arumugam, ''Screening for Colorectal Neoplasias with Fecal Occult Blood Tests: False-positive Impact of Non-Dietary Restriction,'' Asian Pacific Journal of Cancer Prevention, vol. 13, pp. 237-241, 2012.

[13] Yusoff Harmy Mohamed, Norwati Daud, Norhayati Mohd Noor and Amry Abdul Rahim, ''Participants and barriers to colorectal cancer screening in Malaysia,'' Asian Pacific J Cancer Prev, vol. 13, pp. 3983-3987, 2012.

[14] Bin, M.A.S, Bin, M.S.R, Yusof, N.S.H.C, ''Determinants Status Of Patient After Receiving Treatment At Intensive Care Unit: A Case Study In Johor Bahru,'' Computer, Communications, and Control Technology (I4CT), 2014 International Conference on, pp. 80 – 82, 2014.

[15] Shafi, M.A., Rusiman, M.S, ''The Use Of Fuzzy Linear Regression Models For Tumor Size In Colorectal Cancer In Hospital Of Malaysia,'' Applied Mathematical Sciences, vol. 9, no. 56, pp. 2749-2759, 2015.

[16] Jaon Bell, ''Machine learning hands on for developers and technical professionals,'' John Wiley & Sons,Inc, 2015.

[17] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li, ''Applied linear statistical models,'' Mc Graw Hill Fifth Edition, 2004.

[18] Ping Fei Pai, Chih-Sheng Lin, '' A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting,'' The International Journal of Management Science, pp. 497-505, 2004.