Evolutionary Design of a Carbon Dioxide Emission Prediction Model using Genetic Programming

Abdel Karim Baareh
Information Technology
Computer Science Department
Ajloun College, Al-Balqa Applied University
Ajloun, Jordan

***Abstract*—Weather pollution is considered as one of the most important, dangerous problem that affects our life and the society security from the different sides. The global warming problem affecting the atmosphere is related to the carbon dioxide emission (CO_2) from the different fossil fuels along with temperature. In this paper, this phenomenon is studied to find a solution for preventing and reducing the poison CO_2 gas emerged from affecting the society and reducing the smoke pollution. The developed model consists of four input attributes: the global oil, natural gas, coal, and primary energy consumption and one output the CO_2 gas. The stochastic search algorithm Genetic Programming (GP) was used as an effective and robust tool in building the forecasting model. The model data for both training and testing cases were taken from the years of 1982 to 2000 and 2003 to 2010, respectively. According to the results obtained from the different evaluation criteria, it is nearly obvious that the performance of the GP in carbon gas emission estimation was very good and efficient in solving and dealing with the climate pollution problems.**

Keywords*—Fossil fuels; carbon emission; forecasting; genetic programming

I. Introduction

Weather state and condition is a very important and dangerous issue related to some views health, climate, agriculture, economics, and tourism. Estimating the future events at the proper time is a very important task used to reduce and prevents the risks and the natural disasters. Many researchers were attracted towards this type of problems due to its difficulty and challenges in considering different input variables that should be cautiouslyconsidered, studied and measured to build the accurate forecasting models. The events and processes in the world always change due to the circumstances, so these events should be defined and declared to be processed. Climate pollution related to the carbon emission is a general serious world problem. Many international environmental agencies indicated the increase in CO_2 and greenhouse gas emission worldwide [1]. So protecting the civilization from the gas pollution requires a clear and a strict policy [2]. Different protocols and agreements were held between numerous countries to minimize the greenhouse gas emanation, such as the Kyoto protocol and the United Nations (UN) agreement that confirmed on the continuouspercentage checking and monitoring of the CO_2 emission in the atmosphere to reduce it to the desired levels [3].

Many countries stated and started a new policy to decrease and limit the CO_2 emission. Pollution from CO_2 emission is a serious, critical and real society enemy, for example, the UK Government's declaredclear plans and aims to minimize the CO_2 emissions to 10% from the 1990 base by 2010 and in equivalent to generate 10% of the UK's electricity from renewable sources by 2010. Renewable electricity has become related and equivalent to CO_2 reduction [4]. Different studies were initiated and proposed to find out the relationship between the different energy consumption and CO_2 emission [5]-[9].

In this paper, the stochastic search algorithm Genetic Programming (GP) was used as an effective and powerful tool in building and estimating the forecasted model. GP as a soft computing technique was widely used in different fields to solve some complicated problems such as forecasting in all its type weather, rain, rivers, carbon, etc. [10]-[13]. GP also as a powerful tool was efficiently used in many applications [14], [15] such as economics and sales estimations [16], shift failures [17], estimating prices [18] and stock returns [19]. In this study, the GP technique was applied to deal with important and dangerous phenomena that are the CO_2 gas emitted based on four related inputs the global oil, natural gas (NG), coal, and primary energy (PE) consumption. This paper is organized as follows. Section II describes the collected data. Section III introduces the genetic programming concepts. Section IV presents the different implemented evaluation criteria. Section V describes the genetic programming model. Section VI describes the experimental results. Finally, Section VII presents the conclusion and the future work.

II. Collected Data

The carbon dioxide data set was collected from [20] as shown in Table I. The data set was collected for 31 years from 1980 to 2010. The data were trained for 23 years from 1980 to 2002 and tested for eight years from 2003 to 2010. This work is an extension of the previous work published in [12] using Neural network algorithm.

TABLE I. CARBON DIOXIDE DATA SET

Year	Oil Consumption (Mote) X1	NG Consumption (Mote) X2	Coal Consumption (Mote) X3	PE Consumption (Mote) X4	CO2 Emission (Mt) y
1980	2972.2	1296.9	1806.4	6624	19322.4
1981	2863	1309.5	1820.6	6577.5	19073.2
1982	2770.7	1312.5	1846.9	6548.4	18900.7
1983	2748.3	1329	1897.7	6638.2	19072.1
1984	2810.1	1440	1983.2	6960.2	19861
1985	2804.7	1488.3	2056	7137.5	20246.7
1986	2894.1	1503.6	2089.2	7307.5	20688.3
1987	2946.8	1579.6	2169	7555.7	21344.5
1988	3038.8	1654.9	2231.7	7833.5	22052.2
1989	3093	1729.2	2251.2	8001.7	22470.2
1990	3148.6	1769.5	2220.3	8108.7	22613.2
1991	3148.2	1807.5	2196.4	8156	22606.5
1992	3184.8	1817.9	2174.6	8187.6	22656.7
1993	3158	1853.9	2187.6	8257.5	22710.6
1994	3218.7	1865.4	2201.9	8357.6	22980.3
1995	3271.3	1927	2256.2	8577.9	23501.7
1996	3344.9	2020.5	2292.2	8809.5	24089.8
1997	3432.2	2016.8	2301.8	8911.6	24387.1
1998	3455.4	2050.3	2300.2	8986.6	24530.5
1999	3526	2098.4	2316	9151.4	24922.7
2000	3571.6	2176.2	2399.7	9382.4	25576.9
2001	3597.2	2216.6	2412.4	9465.6	25800.8
2002	3632.3	2275.6	2476.7	9651.8	26301.3
2003	3707.4	2353.1	2677.3	9997.8	27508.7
2004	3858.7	2431.8	2858.4	10482	28875.2
2005	3908.5	2511.2	3012.9	10800.9	29826.1
2006	3945.3	2565.6	3164.5	11087.8	30667.6
2007	4007.3	2661.3	3305.6	11398.4	31641.2
2008	3996.5	2731.4	3341.7	11535.8	31915.9
2009	3908.7	2661.4	3305.6	11363.2	31338.8
2010	4028.1	2858.1	3555.8	12002.4	33158.4

III. GENETIC PROGRAMMING CONCEPT

GP is a stochastic search algorithm works on the concept of evolutionary algorithm. This algorithm is drivenby the principles of Darwinian evolution theory and natural selection [21], [22]. GP generates a mathematical model for nonlinear systems in the form of a tree consisting of roots and nodes, where the roots constitute the mathematical operations and the nodes constitute the variables. The formulated tree depth depends on the model functional complexity. An example of GP tree structure is shown in Fig. 1.

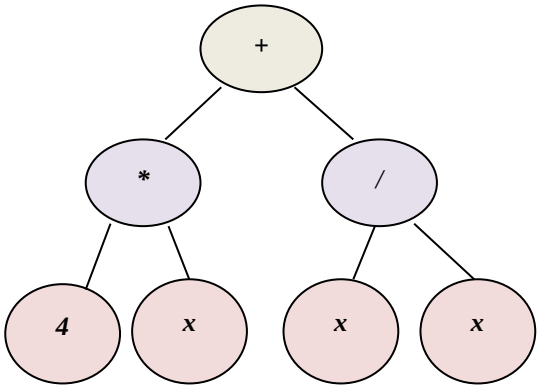

Fig. 1. Example of GP basic tree.

GP was used to encode a computer program in form of a tree structure and evaluate its fitness with respect to the predefined task. In 1991, John Koza suggested LISP programs that deal with various data and structures for a model manipulation due to its flexibility. The GP consists of a population of size n, which is chosen randomly based on the problem. Fig. 2 shows the evolutionary process of GP.

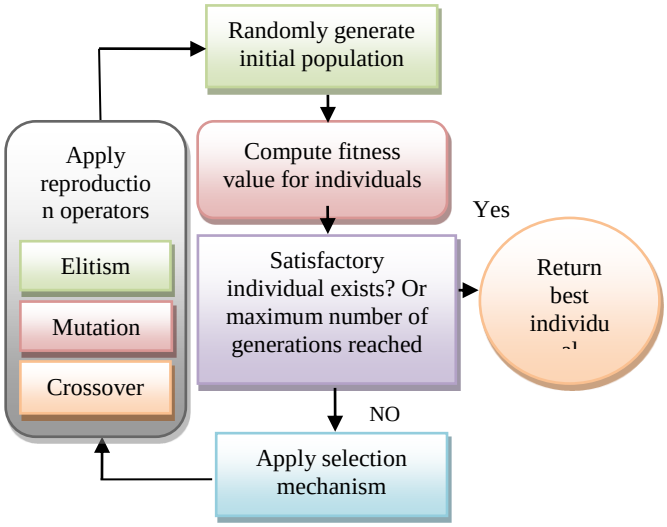

Fig. 2. GP evolutionary process [20].

IV. Evaluation Criteria

In this paper, to solve the modeling problem for the carbon gas (CO_2) estimation, we considered building a model structure that takes into the account the historical measurements of the carbon data during the previous years.

The GP Model was developed using a MATLAB software toolbox called GPTIPS which works as an open source GP Toolbox for MG-GP [23]. GPTIPS defin number of appropriatefunctions for seeking the population of the proper model, such as examining the model behavior, post-run a model simplification function and export the model to some formats, like graphics file, LaTeX expression, symbolic math object or standalone MATLAB file [20]. GP-TIPS can be distinguished by its ability to configure to evolve the multi-gene individuals.

A number of evaluation criteria were used to validate the developed model. These evaluation criteria are the Variance-Accounted-For (VAF), Mean Square Error (MSE), Euclidean distance (ED), Manhattan distance (MD) and Mean magnitude of relative error (MMRE) as shown in equations next.

- Variance-Accounted-For (VAF):

$$\mathrm{VAF} = \left(1 - \frac{\mathrm{var}(y - \hat{y})}{\mathrm{var}(y)}\right) \quad (1)$$

- Mean Square Error (MSE):

$$\mathrm{MSE} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n} \quad (2)$$

- Euclidean distance (ED):

$$\mathrm{ED} = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (3)$$

- Manhattan distance (MD):

$$MD = \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (4)$$

- Mean magnitude of relative error (MMRE)

$$MMRE = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} \quad (5)$$

V. Genetic Programming (GP) Model

The Developed GP model requires the defining and initialization of some important parameters at the beginning of the evolutionary process. These parameters involve the population size, selection mechanism, crossover and mutation probabilities, the maximum number of genes allowed to constitute the multi-gene and many others. The developed GP model tuning parameters are given in Table II.

The complexity of the evolved models will change according to the maximum tree depth. Restricting the tree depth helps to evolve simple model, but it may also reduce the performance of the evolved model. Thus, we need to keep a balance between the depth, the complexity, and required performance.

TABLE II. GP Tuning Parameters

Population size	50
Number of generations	250
Selection mechanism	Tournament
Max. tree depth	10
Probability of crossover	0.85
Probability of mutation	0.001
Max. genes	7
Function set	*, +, -

The GP model can be shown in Fig. 3 where four inputs were applied to the model, the global oil, natural gas, coal, and primary energy consumption to estimate the output CO_2 gas.

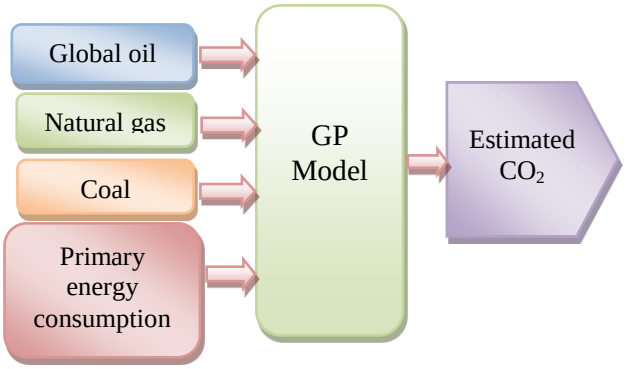

Fig. 3. GP model structure.

Multigene symbolic regression can be defined as a distinctive modification of GP algorithms, where each symbolic model demonstrated by a number of GP trees weighted by a linear combination [24]. In Multigene GP every tree is considered as a "gene" by itself. The predicted output $\hat{y}$ is constituted by adding and combining the weighted outputs that are trees/genes in the Multigene individual with the bias term. Each tree is a function of zero or more of the N input variables $z_1, \ldots, zN$. Mathematically, a Multigene regression model can be written as:

$$\hat{y} = \gamma 0 + \gamma 1 \times \mathrm{Tree1} + \cdots + \gamma \mathrm{M} \times \mathrm{TreeM} \quad (6)$$

Where, $\gamma 0$ represents the bias or offset term while $\gamma_1, \ldots, \gamma_M$ are the gene weights and M is the number of genes (i.e. trees) which constitute the available individual. An example of a multigene model is shown in Fig. 4 and the mathematical model can be shown in (7).

$$y_0 + y_1(\cos(x) * (3 + y)) + y_2 (z/2) \quad (7)$$

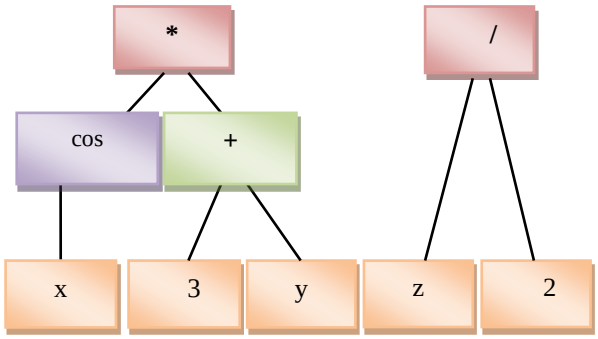

Fig. 4. Example of multi-gene GP model.

VI. EXPERIMENTAL RESULTS

In this paper, the GP model was used to estimate the carbon dioxide gas emission. In our case, four inputs data were used. The inputs are: the oil consumption (X_1), NG consumption (X_2), coal consumption (X_3), PE consumption (X_4) and the output is the CO_2 (y), where the inputs were measured in (Mote) and the output was measured in (Mt). The proposed GP model structure performance was excellent. The estimated CO_2 results for training and testing cases were very close as shown in Tables III and IV. The data were trained for 23 years from 1980 to 2002 and tested for 8 years from 2003 to 2010. Fig. 5 shows the correlation coefficient of the proposed model. In Fig. 6, we show the GP convergence model. In Fig. 7 and Fig. 8, we show the actual and the estimated CO_2 gas emission for training and testing cases.

The mathematical equation promoted for prediction using multi-gene GP can be also shown in (8). The model structure shows a strong linear relationship between the three main attributes Global Oil, Natural Gas and the Coal while the energy consumption was not a significant feature in the modeling process.

$$y = 3.084x_1 + 2.338x_2 + 3.974x_3 - (3.65.10^{-6})x_1x_2$$
$$-(3.65.10^{-6})x_1x_3 - (1.825.10^{-6})x_1{}^2 - 29.47 \qquad (8)$$

In Table V, we calculated the error values through a number of validation criteria for both training and testing cases.

TABLE III. ACTUAL AND ESTIMATED CO_2 - TRAINING CASE

Years	Training Case	
	Actual CO_2 (y)	**Estimated CO_2 ($\hat{y}$)**
1980	19322.4	19322.45
1981	19073.2	19073.02
1982	18900.7	18900.86
1983	19072.1	19072.09
1984	19861	19861.17
1985	20246.7	20246.40
1986	20688.3	20688.30
1987	21344.5	21344.64
1988	22052.2	22052.19
1989	22470.2	22470.25
1990	22613.2	22613.11
1991	22606.5	22606.43
1992	22656.7	22656.84
1993	22710.6	22710.59
1994	22980.3	22980.54
1995	23501.7	23501.70
1996	24089.8	24089.77
1997	24387.1	24386.99
1998	24530.5	24530.54
1999	24922.7	24922.74
2000	25576.9	25576.87
2001	25800.8	25800.64
2002	26301.3	26301.54

TABLE IV. ACTUAL AND ESTIMATED CO_2 - TESTING CASE

years	Testing Case	
	Actual CO_2 (y)	**Estimated CO_2 ($\hat{y}$)**
2003	27508.7	27508.24
2004	28875.2	28874.02
2005	29826.1	29824.85
2006	30667.6	30665.52
2007	31641.2	31638.96
2008	31915.9	31913.53
2009	31338.8	31336.95
2010	33158.4	33155.47

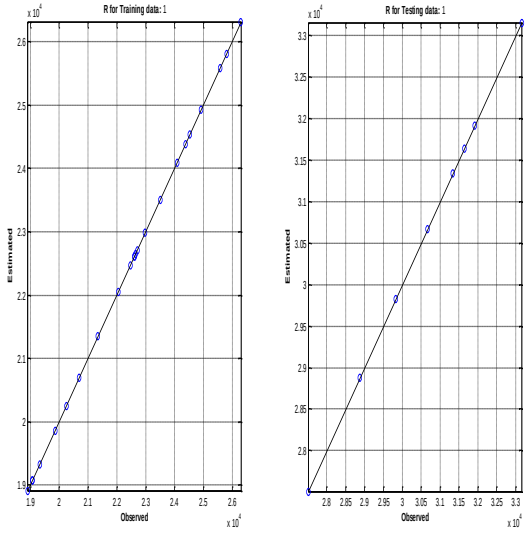

Fig. 5. Correlation coefficient of the proposed GP.

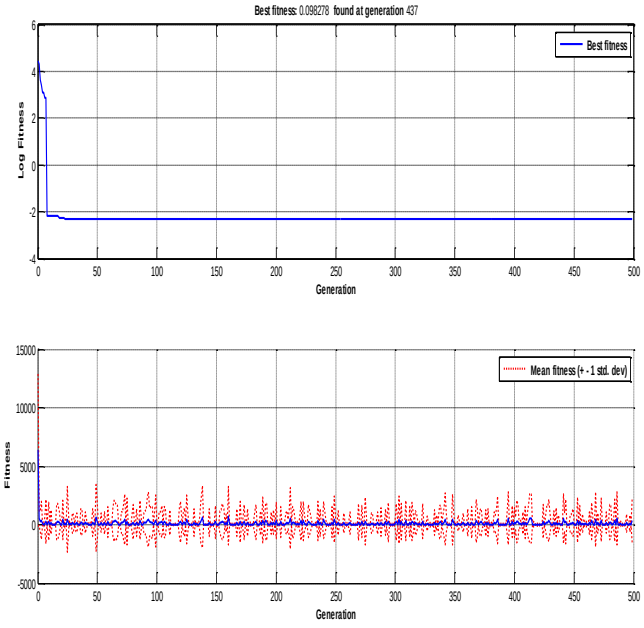

Fig. 6. GP convergence.

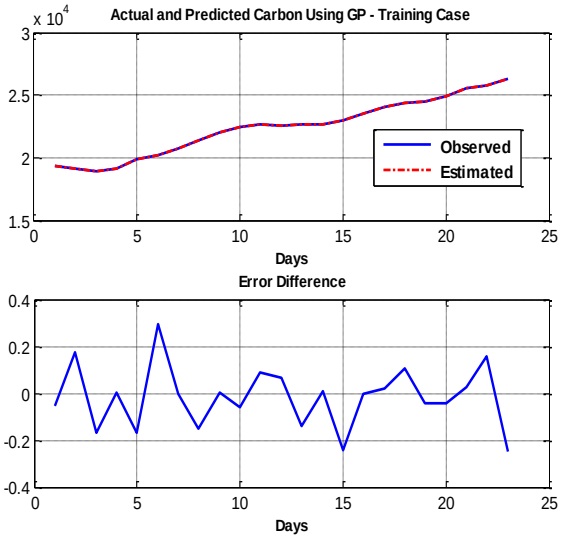

Fig. 7. Actual and predicted CO_2 – training case.

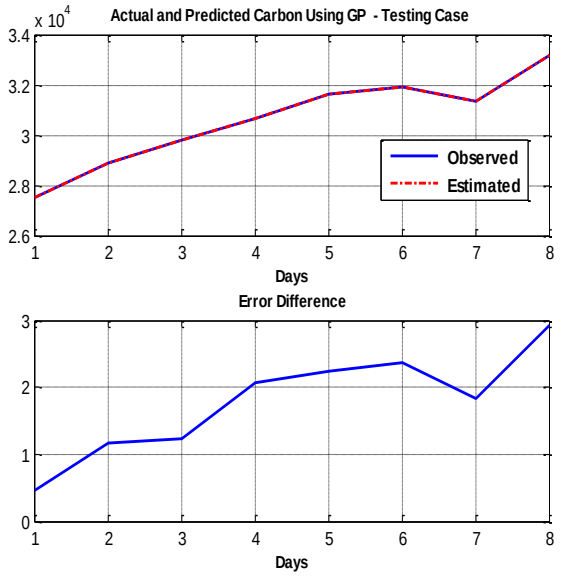

Fig. 8. Actual and predicted CO_2 – testing case.

TABLE V. VALIDATION CRITERIA FOR CO_2

Model	Evaluation Criteria				
	VAF	*MSE*	*ED*	*MD*	*MMRE*
Training	100	0.017	0.629	0.098	4.46e-06
Testing	100	3.744	5.473	1.788	5.72e-05

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we provided an evolutionary model based multigene GP to predict the carbon dioxide emission and we compared the obtained result with actual one to measure the efficiency and strength of GP algorithm in forecasting, for both training and testing cases. From the obtained results, it was shown that the developed model is quite accurate. We can clearly see the solidity and the efficiency of GP in handling and estimating the CO_2 gas. We plan to extend our research to include other paradigms of evolutionary modeling to solve various related environmental problems.

REFERENCES

[1] H. Davoudpourand M. S. Ahadi, "The Potential for Greenhouse Gases Mitigation in Household Sector of Iran: Cases of Price Reform/Efficiency Improvement and Sce- nario for 2000-2010," Energy Policy, Vol. 34, No. 1, pp. 40-49, 2006.

[2] M. R. Lotfalipour, M. A. Falahi and M. Ashena, "Economic Growth, CO_2 Emissions, and Fossil Fuels Consumption in Iran," Energy Policy, Vol. 2010, No. 35, pp. 5115-5120, 2010.

[3] I. A. Samoilov, A. I. Nakhutin, "Esimation and Meddium-Term Forecasting of Anthropogenic Carbon Diox- ide and Methods Emission in Russia with Statistical Methods," Vol. 34, No. 6, pp. 348-353, 2009.

[4] W. David, "Reduction in Carbon Dioxide Emissions: Estimating the Potential Contribution from Wind Power," Renewable Energy Foundation, December 2004.

[5] M. A. Behrang,. E. Assareh, M. R. Assari and A. Ghan- barzadeh, "Using Bees Algorithm and Artificial Neural Network to Forecast World Carbon Dioxide Emission," Energy Sources, Part A: Recovery, Utilization, and Envi- ronmental Effects, Vol. 33, No. 19, pp. 1747-1759, 2011.

[6] H. T. Pao and C. M. Tsai, "Modeling and Forecasting the CO_2 Emissions, Energy Consumption, and Economic Growth in Brazil," Energy, Vol. 36, No. 5, 2011, pp. 2450-2458.

[7] C. Saleh, N. R. Dzakiyullah , J. B. Nugroho, "Carbon dioxide emission prediction using support vector machine", IOP Conf. Series: Materials Science and Engineering 114 (2016).

[8] J. Wang, H. Chang, "Forecasting Carbon Dioxide Emissions in China Using Optimization Grey Model", JOURNAL OF COMPUTERS, VOL. 8, NO. 1, January 2013.

[9] H. Chiroma, S. Abdul-kareem, A. Khan, N. M. Nawi, A. Ya'u Gital, L. Shuib, A. I. Abubakar, M. Z. Rahman, T. Herawan, "Global Warming: Predicting OPEC Carbon Dioxide Emissions from Petroleum Consumption Using Neural Network and Hybrid Cuckoo Search Algorithm", VOL.10, NO. 8, 2015.

[10] A. K. Baareh, A. Sheta, K. AL Khnaifes, "Forecasting River Flow in the USA: A Comparison between Auto Regression and Neural Network Non-Parametric Models," Proceedings of the 6th WSEAS International Conference on Simulation, Modeling and Optimization, Lisbon, 22-24, pp. 7-12, September 2006.

[11] A. K. Baareh, A. Sheta, K. AL Khnaifes, Forecasting River Flow in the USA: A Comparison between Auto-Regression and Neural Network Non-Parametric Models, Journal of Computer Science Vol. 2, No.10, pp. 775-780, 2006.

[12] A. K. Baareh, "Solving the Carbon DioxideEmission Estimation Problem: An Artificial Neural Network Model", Journal of Software Engineering and Applications, Vol. 6 No. 7, PP. 338-342, 2013.

[13] A. Sheta, H. Faris, A. k. Baareh, "Predicting Stock Market Exchange Prices for the Reserve Bankof Australia Using Auto-Regressive Feed forward Neural Network Model", International Review on Computers and Software (I.RE.CO.S.), Vol. 10, N. 7, July 2015.

[14] N. Karunanithi, W. Grenney, D. Whitley and K. Bovee, "Neural Networks for River Flow Prediction," Journal of Computing in Civil Engg, Vol. 8, No. 2, pp. 371- 379,, 1993.

[15] P. R. Bulando and J. Salas, "Forecasting of Short-Term Rainfall Using ARMA Models," Journal of Hydrology, Vol. 144, No. 1-4, pp. 193-211, 1993.

[16] H. Hruschka, "Determining Market Response Functions by Neural Networks Modeling: A Comparison to Econo- metric Techniques," European Journal of Operational Research, Vol. 66, pp. 867-888, 1993.

[17] E. Y. Li, "Artificial Neural Networks and Their Business Applications," Information and Managements, Vol. 27, No. 5, pp. 303-313, 1994.

[18] K. Chakraborty, "Forecasting the Behavior of Multivari- able Time Series Using Neural Networks," Neural Net- works, Vol. 5, pp. 962-970, 1992.

[19] G. Swales and Y. Yoon, "Applying Artificial Neural Net- works to Investment Analysis," Financial Analyst Jour- nal, Vol. 48, No. 5, pp. 78-82, 1992.

[20] H. Kavoosi, M. H. Saidi, M. Kavoosi and M. Bohrng, "Forecast Global Carbon Dioxide Emission by Use of Genetic Algorithm (GA)," IJCSI International Journal of Computer Science Issues, Vol. 9, No. 5, pp. 418- 427, 2012.

[21] J. Koza, "Evolving a computer program to generate random numbers using the genetic programming paradigm," in Proceedings of the Fourth International Conference on GeneticAlgorithms. Morgan Kaufmann, La Jolla,CA, 1991.

[22] J. R. Koza, Genetic Programming: On the Programming ofComputers by Means of Natural Selection. Cambridge, MA,USA: MIT Press, 1992.

[23] D. P. Searson, D. E. Leahy, and M. J.Willis, "GPTIPS: an open source genetic programming toolbox for multigene symbolic regression," in Proceedings of the International Multiconference of Engineers and Computer Scientists, Hong Kong, 17-19March, pp. 77–80.

[24] A. Sheta and H. Faris. Improving production quality of a hot rolling industrial process via genetic programming model. International Journal of Computer Applications in Technology, 49(3/4), 2014. Special Issue on: "Computational Optimisation and Engineering Applications"