Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming

Ahlam A. Sharief
Computer Science Department
Sudan University of Science and Technology
khartoum, Sudan
dreams585@yahoo.com

Alaa Sheta
Software Engineering Department
Zarqa University
Zarqa 13132, Jordan
asheta66@gmail.com

***Abstract*—Diabetes Mellitus is one of the deadly diseases growing at a rapid rate in the developing countries. Diabetes Mellitus is being one of the major contributors to the mortality rate. It is the sixth reason for death worldwide. Early detection of the disease is highly recommended. This paper attempts to enhance the detection of diabetic based on set of attributes collected from the patients to develop a mathematical model using Multigene Symbolic Regression Genetic Programming technique. Genetic Programming (GP) showed significant advantages on evolving nonlinear model which can be used for prediction. The developed GP model is evaluated using Pima Indian data set and showed higher capability and accuracy in detection and diagnosis of Diabetes.**

***Keywords*—Diabetes; Classification; Genetic Programming; Pima Indian data**

I. Introduction

Diabetes is one of the famous diseases that causing death. Based on measured statistics, it is the sixth reason for death worldwide. It was estimated that the world lose about 116 billion per year from medical care costs directly, and cost 580 billion indirectly (death, loss of work because of the deficit). Statistics showed that the high rates of deaths in developing countries are caused by diabetes disease. Early detection of the disease is highly recommended. It is essential to find a way that can help in early predicting this disease. A model with high accuracy, less complex and has efficient performance is urgently needed.

Diabetes Mellitus is simply caused by the failure of the body to produce the right amount of insulin to stabilize the amount of sugar in the body [1]. Most patients suffer this type of body failure are recommended to take insulin injection. This is called diabetes type I. Diabetes type II the patient body rejection to insulin. This type of patient is recommended to undergo certain health meal program as well as performing exercises to lose weight, plus taking oral medication. But heart diseases are likely to strike these patients in the long run [2].

Gestational Diabetes can occur temporarily during Pregnancy which is due to the hormonal changes and usually begins in the fifth or sixth month of pregnancy (between the 24th and 28th weeks). Gestational diabetes usually resolves once the baby is born. However, 25-50% of women with gestational diabetes will eventually develop diabetes later in their life, especially in those who require insulin during pregnancy and those who are overweight after their delivery.

A. General Diabetes Statistics

Due to the wide spread of type II infected diabetes in the USA, a survey was conducted by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in collaboration with the American Diabetes Association [3], the result was 17.9 million have been diagnosed while 5.7 million are unaware that they are infected by the disease. Statistically 23.6 million people in America have been diagnosed type II diabetes positive. Table I show some statistics from the Gestational diabetes in the Middle East and Northern Africa [4]. Some reported statistics of infected people include:

- US women aged 20 years and older form 11.5 million which represent 10.2% of women in USA.
- US men aged 20 years and older form 12 million which represent 11.2% of men in USA.
- people under 20 years form 186,300.
- Adults over 60 years form 12.2 million.
- African Americans aged 20 years and older form 3.7 million (14.7% of all African Americans age 20 years and older).
- Hispanic/Latino Americans form 2.5 million (9.5% of all Hispanic/Latino Americans).
- Caucasian Americans aged 20 years and older form 14.9 million (9.8% of all Caucasian Americans age 20 years and older).

TABLE I. Gestational Diabetes Statistics

Country	Extrapolated Incidence	Population Estimated
Sudan	19.430	39, 148, 162
Iran	33.503	67.503.205
Iraq	12.594	25.374.691
Jordan	2.784	5.611.202
Kuwait	1.120	2.257.549
Lebanon	1.874	3.777.218
Saudi Arabia	12.803	25.795.938
Syria	8.942	18.016.874
UAE	1.252	2.523.915
Yemen	9.938	20.024.867
Egypt	37.778	76.117.421
Libya	2.795	5.631.585

The objective of this work is to explore the advantages of Multigene Symbolic Regression GP to classify the existence or non-existence of diabetic based on data collected from patient

with various nature [5]. The proposed model can predict the class of the patient based on the eight attributes. The model is based on number of measured features of the patients. They include: the number of times pregnant, the results of an oral glucose tolerance test, diastolic blood pressure (mm/Hg), E-Triceps skin fold thickness (mm), 2-h serum insulin (micro U/ml), body mass index, diabetes pedigree function, Age (year).

The paper is structured as follows. In section II, we provide a literature review on the basic research work in the area of diabetic research based on soft computing techniques such as Artificial Neural Networks. In section III, basic process of GP is described. The expansion of Multigene Symbolic Regression GP approach is provided in section IV. The developed results are presented in section VI including the inputs and output of the model, the experimental setup and the developed mathematical GP Model. Finally we introduce the conclusion and future work.

II. Literature Review

The need for an accurate predictor for the diabetes is highly needed. Not only this, but also a predictor that is extremely automated and with less human interference. A diabetic predictor should meet the following specification; efficient modeling, applicability and accuracy and be trusted. It should be compatible with various diagnostic techniques.

Many prediction techniques are used, but the Multi-layer Percepton (MLP) is the most common [6]–[8]. ANN consists of fully connected layers. In the training phase of the prediction, the learning algorithm examines the inputs. While during the testing phase, it examines the outputs and the other unexamined parts during the training phase.

Anthropometrical Body surface scanning data was used to construct a classification model for diabetes type II in [9]. The model applies four data mining approaches. This model is meant to select and point out the appropriate and necessary decision tree for classifying diabetic diseases. It incorporates Artificial Neural Network, Decision Tree, Logistic Regression and Rough sets. In [10] authors used the classification tree for the classification and regression with a binary target. It introduces ten attributes including age, sex, emergency department visits, comorbidity index, dyslipidemia, hypertension, cardiovascular disease, and retinopathy and end-stage renal disease. The cascade learning system which is based on generalized discriminant analysis was introduced by [11]. It has also linked the system with the least square support vector machine in order to perform the classification of diabetes diseases. This uses the classification accuracy, k-fold cross-validation method and confusion matrix.

A method to discover key attributes affecting diabetic diseases was introduced in [12]. The method is called feature selection method. Then it introduced the three classification complementary techniques including Naive Bayes and C4.5. In [13] authors developed and upgraded the Linear Discriminant Analysis (LDA) and integrated it into the automatic diagnosis system. All these models functions primarily in the area of classification. But this method is meant to be accurate and well performed.

The fuzzy approaches have recently become the well-known approaches for improving classification models. Fuzzy Neural Networks (FNNs) and artificial neural networks have been recently integrated hybrid classification model that helps well in diagnosing and classifying the state of the diabetic diseases. This model was presented by [14]. Multi-objective genetic programming approach is proposed by [15] to develop Pareto optimal decision trees in diabetes classification. In [16], GP was used to generate new features by making combinations of the existing diabetes features.

III. Genetic Programming

GP works on a population of individuals, each of which represents a potential solution to a problem. GP was introduced by J. Koza in 1992 at Stanford. A flow chart for GP evolutionary process is shown in Figure 1. In order to solve a problem, it is necessary to specify the following [17]:

- **The terminal set:** A set of input variables or constants.
- **The function set:** A set of domain specific functions used in conjunction with the terminal set to construct potential solutions to a given problem. For symbolic regression this could consist of a set of basic mathematical functions, while Boolean and conditional operators could be included for classification problems.
- **The fitness function:** Fitness is a numeric value assigned to each member of a population to provide a measure of the appropriateness of a solution to the problem in question.
- **The termination criterion:** This is generally a predefined number of generations or an error tolerance on the fitness.

In order to further illustrate the coding procedure and the genetic operators used for GP, a symbolic regression example will be used. Consider the problem of predicting the numeric value of an output variable, y, from two input variables a and b. One possible symbolic representation for y in terms of a and b would be: $y = \frac{a-b}{3}$.

Figure 2 demonstrates how this expression may be represented as a tree structure. With this tree representation, the genetic operators of crossover and mutation must be posed in a fashion that allows the syntax of resulting expressions to be preserved. Figure 3 shows a valid crossover operation where the two parent expressions are given in Equations 1 and 2. The two offspring are given in Equation 3 and 4. Parent 1 (y^1) and Parent 2 (y^2) are presented in Equations 1 and 2. The developed offspring 1 (y^3) and offspring 2 (y^4) are presented in Equation 3 and 4.

$$y^1 = \frac{a-b}{3} \tag{1}$$

$$y^2 = (c-b) \times (a+c) \tag{2}$$

$$y^3 = \frac{a-b}{a+c} \tag{3}$$

$$y^4 = (c-b) \times 3 \tag{4}$$

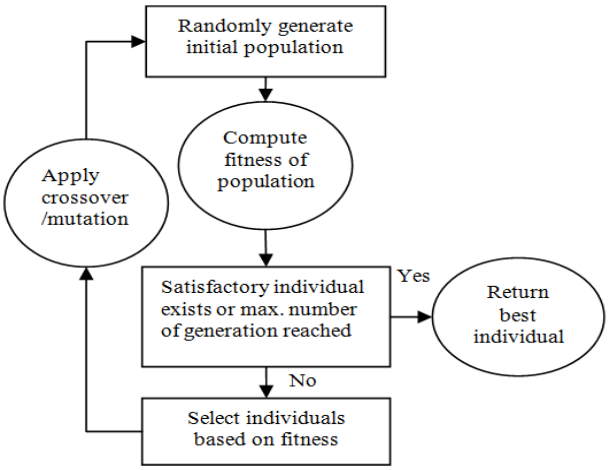

Fig. 1. Flow chart of the GP algorithm [18]

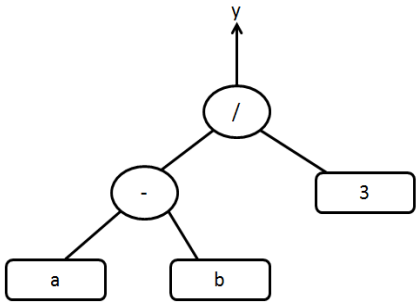

Fig. 2. Representation of a numeric expression using tree structure

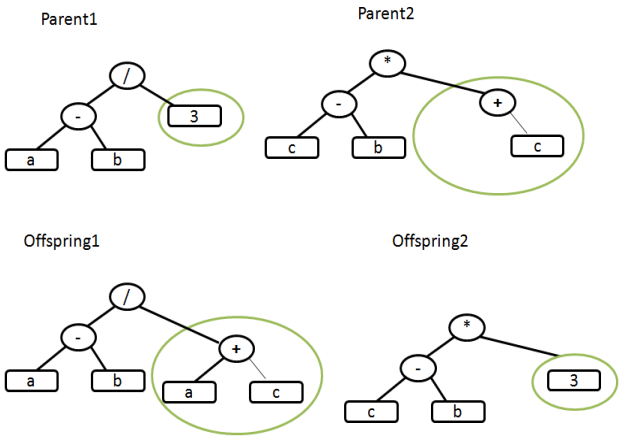

Fig. 3. A typical crossover operation

IV. Multigene Symbolic Regression GP

Typically, symbolic regression is performed by using GP to evolve a population of trees, each of which encodes a mathematical equation that predicts $n \times 1$ vector of outputs y using a corresponding $n \times m$ matrix of inputs X where N is the number of observations of the response variable and M is the number of input (predictor) variables [17]. In contrast, in Multigene symbolic regression each symbolic model (and each member of the GP population) is a weighted linear combination of the outputs from a number of GP trees, where each tree may be considered to be a gene [19]. For example, the Multigene model shown in Figure 4 predicts an output variable using input variables x_1, x_2, x_3. This model structure contains non-linear terms (e.g. the hyperbolic tangent) but is linear in the parameters with respect to the coefficients $\alpha_0, \alpha_1, \alpha_2$.

In practice, the user specifies the maximum number of genes G_{max} and the maximum tree depth D_{max} therefore an exert can control the model complexity. In particular, we have found that enforcing stringent tree depth restrictions (i.e. maximum depths of 4 or 5 nodes) often allows the evolution of relatively compact models that are linear combinations of each model, the linear coefficients are estimated from the training data using ordinary least squares techniques.

Hence, Multigene GP combines the power of classical linear regression with the ability to capture non-linear behavior without the need to pre-specify the structure of the non-linear model. In [20] it was shown that Multigene symbolic regression can be more accurate and computationally efficient than the standard GP approach for symbolic regression.

Here, the first parent individual contains the genes (G_1 G_2 G_3) and the second contains the genes (G_4 G_5 G_6 G_7) where G_{max} equals to 5. Two randomly selected crossover points are created for each individual. The genes enclosed by the crossover points are denoted by [ ].

$$(G_1 \ [\ G_2\ ]\ G_3)\ (G_4\ [\ G_5\ G_6\ G_7\ ])$$

The genes enclosed by the crossover points are then exchanged resulting in two new individuals as follows:

$$(G_1\ G_5\ G_6\ G_7\ G_3)\ (G_4\ G_2)$$

Two point high level crossover allows the acquisition of new genes for both individuals but also allows genes to be removed. If an exchange of genes results in an individual containing more genes than G_{max} then genes are randomly selected and deleted until the individual contains G_{max} genes.

The user can set the relative probabilities of each of these recombination processes. These processes are grouped into categories called events. The user can then specify the probability of crossover events, direct reproduction events and mutation events. These must sum to one. The user can also specify the probabilities of event subtypes, e.g. the probability of a two point high level crossover taking place once a crossover event has been selected, or the probability of a sub tree mutation once a mutation event has been selected.

An example of Multigene model is shown in Figure 4. The presented model can be introduced mathematically as given in Equation 5. GPTIPS Matlab Toolbox provides default values for each of these probabilities so the user does not need to explicitly set them [21].

$$\alpha_{\mathbf{0}} + \alpha_{\mathbf{1}}(0.41x_1 + tanh(x_2x_3)) + \alpha_{\mathbf{2}}(0.45x_3 + \sqrt{x_2}) \quad (5)$$

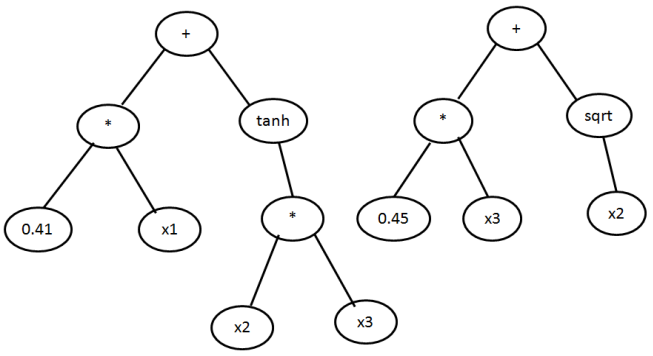

Fig. 4. Example of a Multigene Symbolic Regression Model

V. Performance Criterion

Number of evaluation criterion was computed to evaluate the performance of the developed models. The Route Mean Square (RMS) was used as the fitness function for genetic programming. RMS can be described by Equation 6.

$$RMS = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2} \quad (6)$$

Other performance criterion was used to evaluate the goodness of the developed GP model. The set of criterion used are given as follows:

- Sensitivity (Sens):

$$Sens = \frac{TP}{TP + FN} \quad (7)$$

- Specificity (Spec):

$$Spec = \frac{TN}{FP + TN} \quad (8)$$

- Positive Predicted Value (PPV):

$$PPV = \frac{TP}{TP + FP} \quad (9)$$

- Negative Predicted Value (NPV):

$$NPV = \frac{TN}{FN + TN} \quad (10)$$

- Accuracy (Acc):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

Given that:

- True Positive (TP): Sick people correctly diagnosed as sick.
- False Positive (FP): Healthy people incorrectly identified as sick.
- True Negative (TN): Healthy people correctly identified as healthy.
- False Negative (FN): Sick people incorrectly identified as healthy.

VI. Developed Results

A. Model Inputs and Output

Pima Indian is a homogeneous group that inhabits the area around American, but they are popular for being the most infected group with type II diabetes. Pima Indians diabetes data can even be retrieved from UCI Machine Learning Repository's web site [5]. So, they are subject of intense studies in type II diabetes. The data consist of eight input variables and one output (0,1). The GP mathematical model has the inputs and output presented in Table II. We used 500 samples as a training set and 100 samples as a testing set. The data set was normalized according to Equation 12.

$$x^{new} = \frac{x^{old} - x_{min}}{x_{max} - x_{min}} \quad (12)$$

x_{max} and x_{min} are the maximum and minimum values of the array x, respectively. x^{new} is the newly computed value based on the value of x^{old}.

TABLE II. Inputs and Output for Diabetic Prediction Model

Inputs	The number of times pregnant	x_1
	The results of an oral glucose tolerance test	x_2
	Diastolic blood pressure (mm/Hg)	x_3
	E-Triceps skin fold thickness (mm)	x_4
	2-h serum insulin (micro U/ml)	x_5
	Body mass index	x_6
	Diabetes pedigree function	x_7
	Age (year)	x_8
Output	Predicted class	y

B. Experimental Setup

In this research, we adopted a GPTIPS toolbox [21] to develop our results. In GPTIPS, the initial population is constructed by creating individuals that contain randomly generated GP trees with between 1 and G_{max} genes. During the run, genes are acquired and deleted using a tree crossover operator called two point high level crossover. This allows the exchange of genes between individuals and it is used in addition to the standard GP recombination operators.

Some parameters have to be defined by the user at the beginning of the evolutionary process. They include: population size, probability of crossover, mutation probability and the type of the selection mechanism. User has also to setup the maximum number of genes G_{max} where a model is allowed to have. The maximum tree depth D_{max} allows us to change the complexity of the evolved models. Restricting the tree depth helps evolving simple model but it may also reduce the performance of the evolved model.

A prior knowledge on the problem domain helps in designing a function set which could speed up the evolutionary

process for model development. The adopted function set to develop the GP model is given as:

$$F = \{+, -, \times\}$$

TABLE III. GP TUNING PARAMETERS

Population size	100
Number of generation	100
Selection mechanism	Tournament
Max. tree depth	7
Probability of Crossover	0.85
Probability of Mutation	0.1
Max. No. of genes allowed in an individual	6

Crossover was performed with the two-point high-level crossover operator. Once the two parent individuals have been selected, two gene crossover points are selected within each parent. Then the genes enclosed by the crossover points are swapped between parents to form two new offspring.

C. Developed Mathematical GP Model

The data set described earlier was loaded then the Multigene GP was applied using GPTIPS Tool. The parameters of the algorithm were tuned as listed in Table III. In Figure 5, we show the convergence of GP over 100 generations. It can be clearly seen that the final model is a simple and compact mathematical model which is easy to evaluate. The performance measurements for the model was computed and summarized in Table V. The best generated diabetic prediction Multigene GP model is given in Table IV.

TABLE V. PERFORMANCE OF THE GP MODEL WITH $x_1, \ldots, x_8$ AS INPUTS

Criteria	Training	Testing
Sensitivity	0.90881	0.95946
Specificity	0.56593	0.69231
Accuracy	0.78533	0.89873
Positive Predicted Value	0.7803	0.85714
Negative Predicted Value	0.784	0.890

We also explored the idea of considering a subset of the features used to develop the GP model. Thus, we considered the features x_3, x_6 and x_8 to develop the output class y of diabetic type. Running GP with a population size 30 and 100 generations with the same tuning parameters such as tree depth, maximum number of genes, probability of crossover and probability of mutation we produced the results in this case. The performance measurements for the developed GP model was computed and summarized in Table VI. In Figure 6, we show the convergence of GP in the case with less number of features. The developed GP model is presented in Table VII.

TABLE VI. PERFORMANCE OF THE GP MODEL WITH x_3, x_6 AND x_8 AS INPUTS

Criteria	Training	Testing
Sensitivity	0.84591	0.81081
Specificity	0.47253	0.65385
Accuracy	0.73699	0.86957
Positive Predicted Value	0.63704	0.54839
Negative Predicted Value	0.71	0.77

VII. CONCLUSIONS AND FUTURE WORK

In this paper, a GP mathematical model was developed to provide a solution to the diabetic problem. The developed model was able to classify patient type. The developed classification accuracy obtained based on Multigene GP is high with respect to sensitivity, specificity, accuracy, positive predicted and negative predicted values. These evaluation criterions proved that Multigene GP is beneficial for diabetic patient classification. The knowledge gained is comprehensible and can enhance the decision making process by the physician. We plan to expand this research to detect the most significant attributes which indicate diabetic.

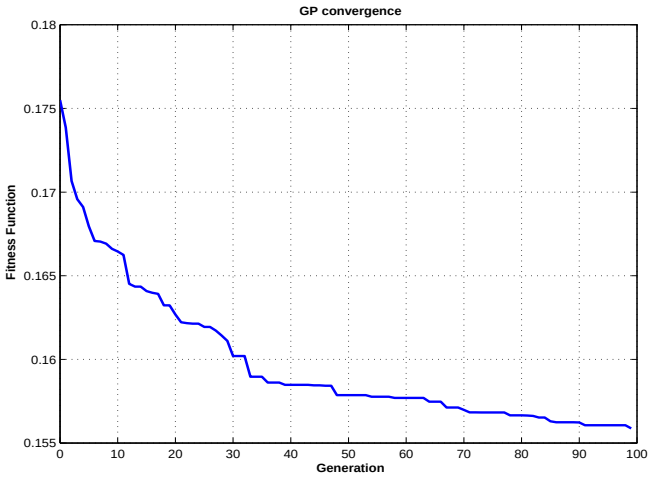

Fig. 5. Convergence of the GP evolutionary process

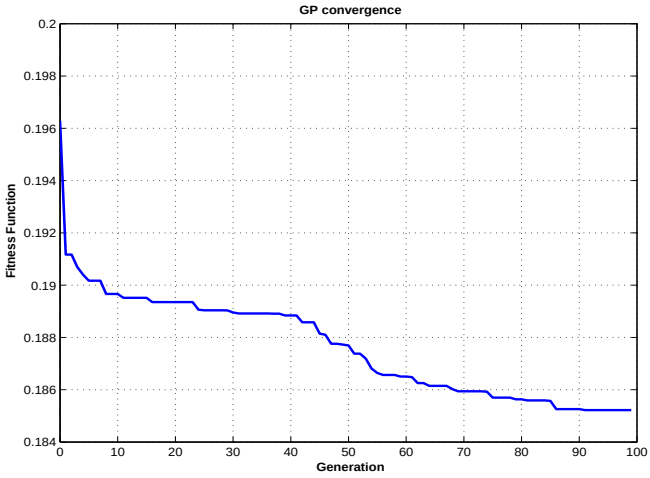

Fig. 6. Convergence of the GP evolutionary process

REFERENCES

[1] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for type-2 diabetic patients," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8102–8108, 2010.

[2] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "A new neural network approach for short-term glucose prediction using continuous glucose monitoring time-series and meal information," *Proceedings of the IEEE International Conference on Engineering Med. Biol. Soc.*, pp. 5653–6, 2011.

[3] G. E. C. Estrada, L. del Re, and E. Renard, "Nonlinear gain in online prediction of blood glucose profile in type 1 diabetic patients," in *CDC*, pp. 1668–1673, 2010.

[4] USA, "US census bureau, population estimates, statistics by country for gestational diabetes," 2004.

TABLE IV. A GP MODEL WITH INPUTS: $x_1, \ldots, x_8$

$$\begin{aligned} y &= 0.3636 * x_1 - 0.658 * x_2 - 0.4626 * x_4 - 0.4626 * x_5 - 0.3636 * x_3 * x_6 + 1.349 * x_2^2 + 1.331 * x_6^2 \\ &+ 0.4626 * x_2 * (x_4 + 2 * x_5) * (-x_6^2 + x_3 + x_4) + 0.658 * x_2 * x_5 * x_7 \\ &- 0.3636 * x_2 * (x_3 + x_4) * (x_1 - x_5) - x_2^2 * x_7 * (3.031 * x_2 - 3.031) \\ &+ 12.38 * x_7 * (x_4 + x_5) * (x_1 - x_5) * (x_3 - x_4) - 0.06898 \end{aligned}$$

TABLE VII. A GP MODEL WITH INPUTS: x_3, x_6 AND x_8

$$\begin{aligned} y &= 0.3748 * x_6^2 * (x_3 * (x_6 + x_8) * (x_3^2 + x_8) + 2.66) * (x_8 - x_6 + 2.559) - 1.435 * (x_3 - x_8) * (x_6 - x_8) \\ &- 0.03142 * x_6 * ((x_6 - x_8) * (2 * x_6 + x_8) + 2 * x_3 * x_6 * x_8 * (x_6 + 5.125)) * (x_8 - x_6 * (x_3 - x_6)) * (2.559 * x_6 - 2.559 * x_8 + 69.97 * x_6 * x_8) \\ &+ 0.02144 \end{aligned}$$

[5] M. W. Aslam and A. K. Nandi, "Detection of diabetics using genetic programming," in *European Signal Processing Conference*, no. 18, (Aalborg, Denmark), August 2010.

[6] L. M. Silva, J. M. de Sá, and L. A. Alexandre, "Data classification with multilayer perceptrons using a generalized error function," *Neural Networks*, vol. 21, no. 9, pp. 1302–1310, 2008.

[7] A. Elkamel, S. Abdul-Wahab, W. Bouhamra, and E. Alper, "Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach," *Advances in Environmental Research*, vol. 5, no. 1, pp. 47 – 59, 2001.

[8] R. S. Selvaraj, K. Elampari, R. Gayathri, and S. J. Jeyakumar, "A neural network model for short term prediction of surface ozone at tropical city," *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5306–5312, 2010.

[9] C.-T. Su, C.-H. Yang, K.-H. Hsu, and W.-K. Chiou, "Data mining for the diagnosis of type ii diabetes from three-dimensional body surface anthropometrical scanning data," *Computers & Mathematics with Applications*, vol. 51, no. 6-7, pp. 1075–1092, 2006.

[10] J. T. Tennis, "Three spheres of classification research: Emergence, encyclopedism, and ecology," in *ASIS SIG/CR Classification Research Workshop*, 2002.

[11] K. Polat, S. Güneş, and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine," *Expert Syst. Appl.*, vol. 34, pp. 482–487, Jan. 2008.

[12] Y. Huang, P. J. McCullagh, N. D. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," *Artificial Intelligence in Medicine*, vol. 41, no. 3, pp. 251–262, 2007.

[13] D. Çalisir and E. Dogantekin, "An automatic diabetes diagnosis system based on lda-wavelet support vector machine classifier," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8311–8315, 2011.

[14] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, no. 12, pp. 82 – 89, 2008.

[15] E. Mugambi and A. Hunter, "Multi-objective genetic programming optimization of decision trees for classifying medical data," in *Knowledge-Based Intelligent Information and Engineering Systems* (V. Palade, R. Howlett, and L. Jain, eds.), vol. 2773 of *Lecture Notes in Computer Science*, pp. 293–299, Springer Berlin Heidelberg, 2003.

[16] M. W. Aslam, Z. Zhu, and A. K. Nandi, "Feature generation using genetic programming with comparative partner selection for diabetes classification," *Expert Systems with Applications*, vol. 40, no. 13, pp. 5402 – 5412, 2013.

[17] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, MA, USA: MIT Press, 1992.

[18] A. F. Sheta, H. Faris, and E. Öznergiz, "Improving production quality of a hot-rolling industrial process via genetic programming model," *Int. J. Comput. Appl. Technol.*, vol. 49, pp. 239–250, June 2014.

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

[20] M. Hinchliffe, H. Hiden, B. McKay, M. Willis, M. Tham, and G. Barton, "Modelling chemical process systems using a multi-gene genetic programming algorithm," in *Late Breaking Papers at the Genetic Programming 1996 Conference Stanford University July 28-31, 1996* (J. R. Koza, ed.), (Stanford University, CA, USA), pp. 56–65, Stanford Bookstore, 28–31 July 1996.

[21] 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 Multi-conference of Engineers and Computer Scientists 2010 (IMECS 2010)*, vol. 1, (Hong Kong), pp. 77–80, 17-19 Mar. 2010.