GENETIC ALGORITHM (GA) OPTIMIZATION USING DIABETES EXPERIMENTAL DATA

Ejiofor C. I[1]&Laud Charles Ochei[2]

Department of Computer Science, University of Port-Harcourt, Port-Harcourt. Nigeria [1]
Robert Gordon University, Aberdeen, United Kingdom,[2]

Abstract

Diabetes comprises of noisy features. This feature hampers classification and prediction for Artificial Intelligence (AI) system. The optimization of diabetes dataset using Genetic Algorithm (GA) exploring its fundamentals identifies the focus of this research. The dataset obtained from Biostat comprising of random samples (fifteen: 15) and parameter variables five: cholestrol, high-density lipoprotein, age, height and weight was used for the optimization. The simulation was Matrix Laboratory (MATLB). The optimized dataset was validated using standard optimization equation resulting in percentage score of Forty-one (41%) percent. This dataset will be using in classifying fuzzy system

Keywords

Genetic Algorithm, Diabetes.

1.INTRODUCTION

Diabetes is a condition where the body fails to utilize the ingested glucose properly. Diabetes is caused by deficiency in insulin production or failure of the body to response to insulin production (Ananya, 2017). Excess glucose overtime within the blood stream can result in eye, kidney and nerve damage. It can also inflame heart disease, stroke and even amputation.

Diabetes has been identified as the fastest growing long term disease that affects millions of people worldwide(MedlinePlus, 2017). The statisticindeed has been alarming across the globe. In 2013 it was estimated that over 382 million people worldwide were suffering from diabetes. In the United Kingdom, 2million persons have been identified as suffering from diabetes with 750, 000 unaware of their current illness. In the United States 25.8 million people or 8.3% of its population have been identified as diabetes sufferer with70 million remaining undiagnosed. In 2010, about 1.9 million new cases of diabetes were also identified in United State with a future prediction of 1 in every 3 Americans having the chance of experiencing diabetes by 2050 (Ananya, 2013).

Diabetes can be categorized into: Type I and Type II. Type I diabetes is also known as insulin-dependent diabetes mellitus and is usually associated with younger adolescence.This form of diabetes is dueto the inability of the body to produce insulin as a result of autoimmune disorder destroying the pancreases and eliminating the chance of insulin production. Approximately 10% of diabetes diagnoses are associated with Type I.Type II diabetes; also known as adult-onset or noninsulin-dependent diabetes. This form of diabetes exists due to failure of the pancreas to produce enough insulin to metabolize glucose which is usually associated with aging.Approximately 90% of all cases of diabetes worldwide are associated with Type II (Medlinplus, 2017).

DOI : 10.5121/ijfls.2017.7302

Diabetes symptoms vary from frequent urination, intense thirst and hunger, weight gain, unusual weight loss, fatigue, male sexual dysfunction, numbness and tingling in hands and feet (MedlinPlus, 2017 and Ananya, 2017).

Treatment and diagnosis of diabetes usually are complex with physician depending on patient's symptoms in collaboration with Age, weight, height, family history and the contributing factor of alcoholism and lack of physical exercise in identifying type I or II diabetes. While the diagnoses have been consistent over time, the complication experienced by novel physician and the non-availability of experienced expert has fostered numerous Artificial Intelligence (AI) models. Although these model has been employed for the prediction of diabetes built to complement the conventional approaches of physician-patient interaction (Mehdi et al., 2012; Meysam and Mahdi, 2016)these models, possibly have suffer for inaccuracies due to noisy training sample which has hampered training cases

This research paper explores genetic algorithm optimization technique for preprocessing diabetes datasetidentifying change variation within the dataset. The variation in change will be explored using standard optimization equation.

2.GENETIC ALGORITHM (GA) OVERVIEW

Genetic Algorithm (GA) considered a search and optimization technique based on the adaptation of natural selection process and Evolutionary Algorithms (EA) has found its mark in Artificial intelligence problem domains. GA provides a framework for solving both constrained and unconstrained optimization problems based on a natural selection process which mimics biological evolution (Akbari, 2010). GA is used in arriving at optimal solutions; solutions directing the optimization to the best possible area (Eiben, 1994) through the modification of population of individual solutions.

GA usually creates an optimization process using an initial population. This population encompasses solution seen as candidates with each candidate's solution possess initial possible solutions. In each generation, the fitness of each individual are usually expressed and examined using the objective function created in most cases based on adaptation, user examine, experimental design and trial error (Son et al., 2016). The fittest individual are stochastically selected from the current population with each genome modified based on genetic operator. This modification creates new generation which are repeatedly evaluated. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population (Harik et al., 2006).

GA population size depends largely on the problem domain. This population usually combined numerous possible solutions which are generated randomly forming the search or state space. These solutions are usually attuned toward better solutions (*Taherdangkoo et al., 2012*).
Successive generation solutions are combined with preceding generation to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions are typically more likely to be selected. Novel solutions are produced using parent solution in breeding pool of successive generation by producing children solution (Coffin and Robert, 2008). The optimization process usually terminate based on minimum criteria satisfaction, allocated budget and highest fitness value (Echegoyen et al., 2012).
Solution encoding, objective function identification, identified solution, application of operator and termination are indeed the fundamental components of GA (Coffin and Robert, 2008; Akbari, 2010 Echegoyen et al., 2012).

3.METHODOLOGY:DIABETES DATASET OPTIMIZATION USING MATRIX LABORATORY (MATLAB)

The dataset for simulation was obtained from Biostat. This data served as the experimental data for optimization. The dataset comprises of fifteen samples spread across five decisionvariables: cholestrol, high-density lipoprotein, age, height and weight. Table 3.1 depicts the diabetes simulation data.

Table 3.1: Un-Optimized Diabetes Dataset

Case	Cholesterol	High-Density Lipoprotein	Age	Height	Weight	Weighted Score	Status	Confusion Matrix
Case 1	203	56	46	62	121	488	Type II	TP
Case 2	165	24	29	64	218	500	Type II	FP
Case 3	228	37	58	61	256	640	Type II	TP
Case 4	78	12	67	67	119	343	Type II	TP
Case 5	249	28	64	68	183	592	Type II	TP
Case 6	248	69	34	71	190	612	Type II	TN
Case 7	195	41	30	69	191	526	Type II	TN
Case 8	227	44	37	59	170	537	Type II	TN
Case 9	177	49	45	69	166	506	Type II	TP
Case 10	263	40	55	63	202	623	Type II	TP
Case 11	242	54	60	65	156	577	Type II	TP
Case 12	215	34	38	58	195	540	Type II	TN
Case 13	238	36	27	60	170	531	Type II	TN
Case 14	183	46	40	59	165	493	Type II	TP
Case 15	191	30	36	69	183	509	Type II	TN

3.1MATLAB GA SIMULATIONS

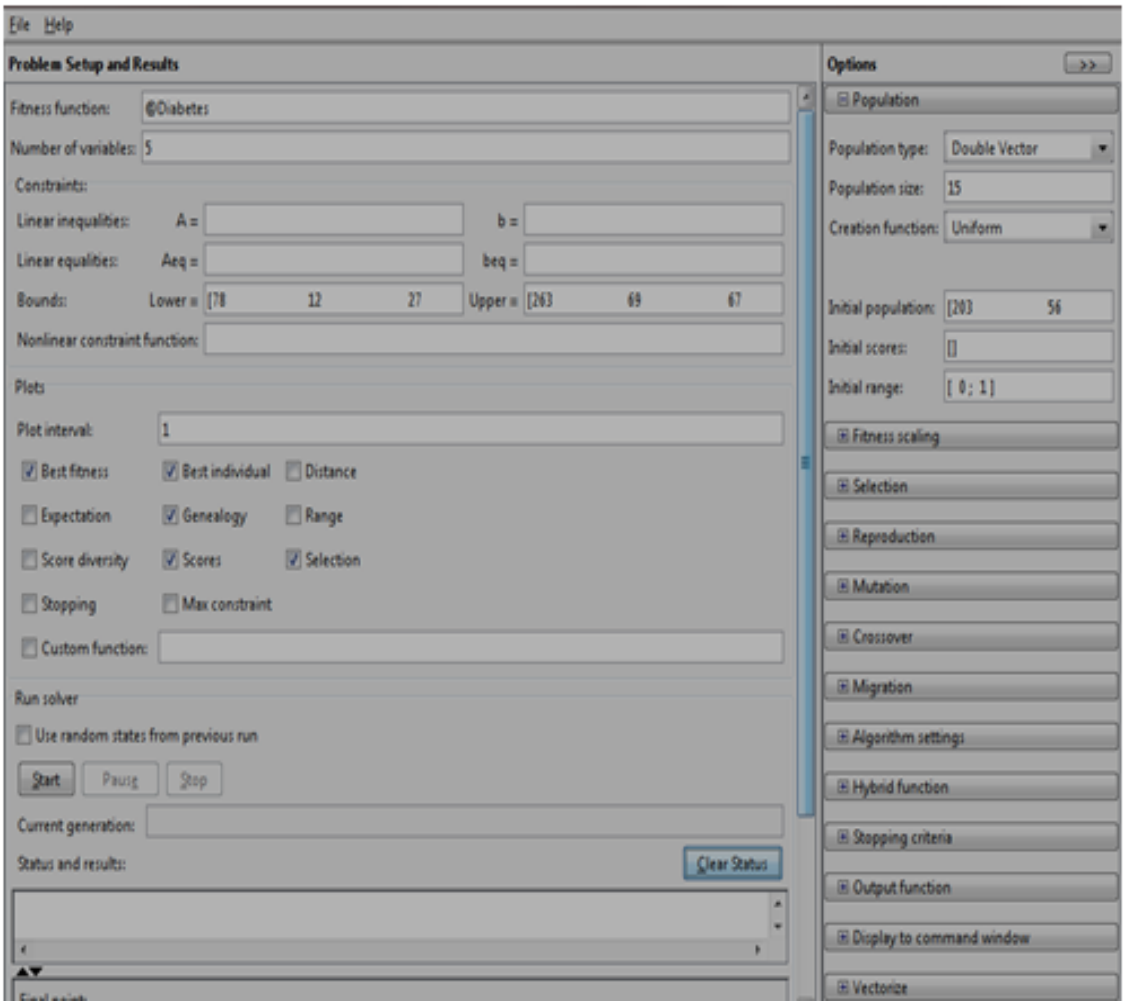

Figure 3.1: GA-Simulation Chart

The chart of Figure3.1 captures the fundamentals of diabetes GA simulation. It shows clearly that five input variable were utilized in collaboration with fifteen samples. The population type for simulation was explored as double while the function was uniformly created. The chart also depicts the weighted score for each column parameter input. The simulations chart also capture best fitness value, best individual value, genealogy, score and selection.

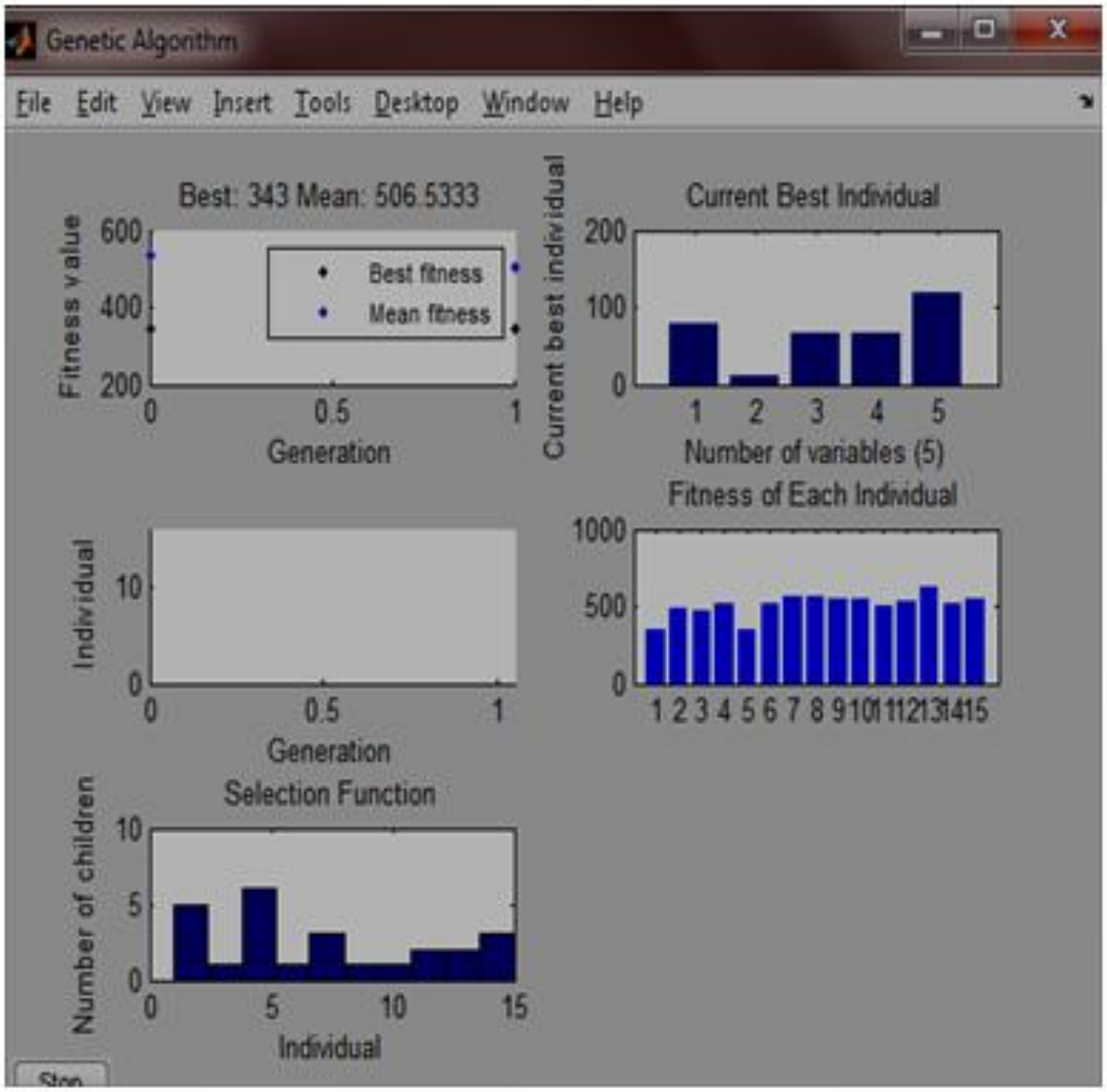

Figure 3.2: Generation 1 (initial) GA-Simulation chart

Figure 3.2 provides the first generation simulation identifying fitness value; with the best fitness value at generation one shown as 343and the over means of 15 samples identified as 506. The individual generation is identified as initial. The selection function picture the probability open to each sample in selecting prospecting children for mating within the next generation, with the fifth individual having the highest section probability of 6. The current best individual provides the best individual score per generation while the fitness score for each individual can be identified successively for each generation. For this generation an average fitness score less than 600 was initially maintained.

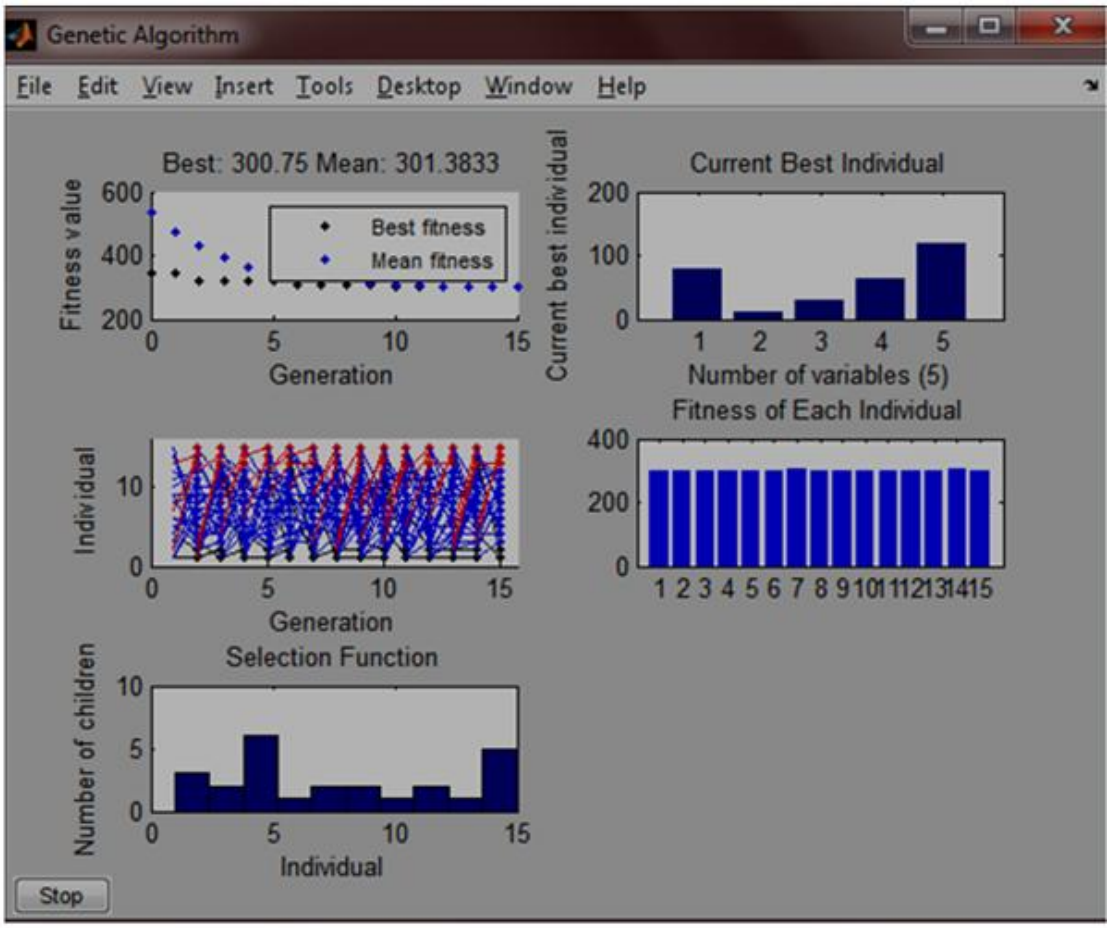

Figure 3.3: Generation 15 (final) GA-Simulation chart

Figure 3.3 provides the final generation simulation identifying fitness value; with the best fitness value at generation fifteen shown as 307 and the over means of 15 samples identified as 301. The individual generation is identified as fifteen. The selection function picture the probability open to each sample in selecting prospecting children for mating within the next generation, with the fifth individual having the highest section probability of 7. The current best individual provides the best individual score per generation while the fitness score for each individual can be identified successively for each generation. For this generation an average fitness score above 300 was maintained.

Table 3.2: Optimized Diabetes Dataset

Case	Cholesterol	High-Density Lipoprotein	AGE	Height	Weight	Weighted Score	Status	Confusion Matrix
Case 1	78	12	67	67	119	343	Type II	TP
Case 2	78	12	67	67	119	343	Type II	FP
Case 3	78	36	27	67	119	327	Type II	TP
Case 4	78	12	40	59	121	310	Type II	TP
Case 5	78	12	36	67	119	312	Type II	TP
Case 6	79	13	27	62	119	300	Type II	TN
Case 7	78	12	27	66	119	302	Type II	TN
Case 8	78	12	67	66	119	342	Type II	TN
Case 9	78	12	45	62	119	316	Type II	TP
Case 10	78	12	37	69	119	315	Type II	TP
Case 11	78	12	27	67	119	303	Type II	TP
Case 12	78	12	29	64	119	302	Type II	TN
Case 13	78	12	34	67	119	310	Type II	TN
Case 14	78	12	36	59	119	303	Type II	TP
Case 15	78	12	29	63	119	300	Type II	TN

Table 3.2 provides the change obtained succeeding optimization. The values cut across cholesterol, high density lipoporotein, Age, height and weight and showed the variation in data changed compared to previous data value appearing on table 3.1.

4.VALIDATION OF OPTIMIZED DIABETES DATASET

Validation provides a definite proof in ascertaining the variation in change and determining percentage optimization. It measures how much the fundamental components have been optimized. Equation 4.1 determines these changes.

$$DoO = \frac{|R_0 - M_0|}{M_0} \quad . \quad . \quad . \quad (4.1)$$

Where

R_0= summation of fitness values of optimized dataset

M_0=summation of fitness values of the non-optimized dataset

Table 4.2, captures the non–optimized and the optimized dataset the dataset are exemplified from case 1-10 and 110.

Table 4.3: Non- Optimized and Optimized fitness Values

SN	M_0 : Σ (Non- Optimized fitness Values)	R_0: Σ (Optimized Fitness Values)
Case 1	488	343
Case 2	500	343
Case 3	640	327
Case 4	343	310
Case 5	592	312
Case 6	612	300
Case 7	526	302
Case 8	537	342
Case 9	506	316
Case 10	623	315
Case 11	577	303
Case 12	540	302
Case 13	531	310
Case 14	493	303
Case 15	509	300
Total	8017	4728

Degree of Optimization (DoO) = [4728 - 8017]/8017
= 0.4102
= 0.4102 * 100
= 41.1%

The degree of optimization shows clearly that 41% variation change has occurred within the dataset that has been optimized. The graph of figure 4.1 graphical depicts this percentage change.This variation in change has improved the given dataset and subsequently eliminated noisy features.

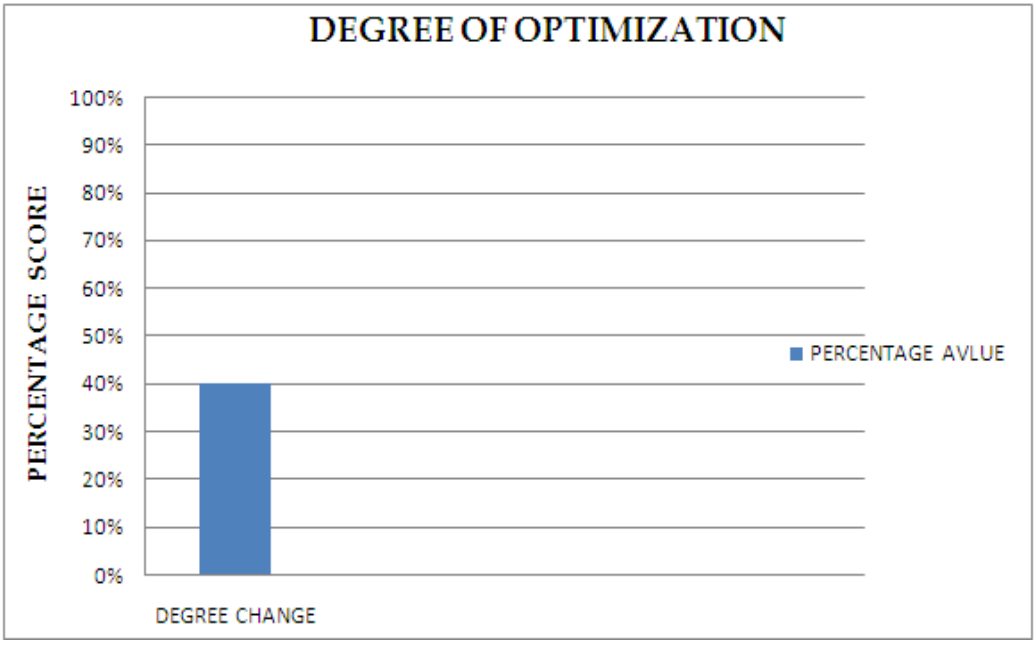

Figure 4.1: Degree of Change for Diabetes data

The graph of figure 4.1 provides a percentage change of 40%. This change is perceived as the stall change value. The point at which change optimization on the same dataset could be stochastic possible as it run into infinite

5. CONCLUSION

This research has established the usefulness of genetic Algorithm as an optimization tool in ascertaining optimal samples for used for prediction. The data obtained from biostat have been optimized using an appropriate objective function and associated fitness values. Matric laboratory interface provided simulation interfaces captured variation change within the dataset. The validation using standard optimization equation captured an optimization change of 41%. This dataset will be utilized probably in provided classification for fuzzy system.

REFERENCES

[1] Akbari Z. (2010). "A multilevel evolutionary algorithm for optimizing numerical functions" IJIEC 2 (2011): 419–430

[2] Ananya (2017), What is Diabetes, retrieved online from https://www.news-medical.net/health/What-is-Diabetes.aspx

[3] Coffin, D.; S., Robert E. (2008). "Linkage Learning in Estimation of Distribution Algorithms". Linkage in Evolutionary Computation. Springer Berlin Heidelberg: 141–156. doi:10.1007/978-3-540-85068-7_7.

[4] Eiben, A. E. et al (1994). Genetic algorithms with multi-parent recombination, PPSN III: Proceedings of the International Conference on Evolutionary Computation. The Third Conference on Parallel Problem Solving from Nature: 78–87. ISBN 3-540-58484-6.

[5] Echegoyen, C.; Mendiburu, A. Santana, R.; Lozano, J. A. (2012). "On the Taxonomy of Optimization Problems under Estimation of Distribution Algorithms". Evolutionary Computation. 21 (3): 471–495. ISSN 1063-6560. doi:10.1162/EVCO_a_00095.

[6] Harik G. R.; Lobo, F. G.; Sastry, K. (2006), Linkage Learning via Probabilistic Modeling in the Extended Compact Genetic Algorithm (ECGA),Scalable Optimization via Probabilistic Modeling

[7] Springer Berlin Heidelberg: 39–61. doi:10.1007/978-3-540-34954-9_3.

[8] MedlinePlus (2017), Diabetes, retrieved online from http:// www.medlineplus.com

[9] Mehdi K., Saeede E. and Jamshid P. (2012), Diagnosing Diabetes Type II Using a Soft Intelligent Binary Classification Model, Review of Bioinformatics and Biometrics (RBB) Volume 1 Issue 1, December 2012 9-23.

[10] Meysam J., and Mahdi M. (2016), Comparison of Predictive Models for the Early Diagnosis of Diabetes, Kournal of Health Information Research, Vol 22(2), Pp.95-100

[11] Son D. D., Kazem A. and Romeo M. (2016), Maximsing Performance of Genetic Algorithm Solver in Matlab, Advance online publication:, Pp. 1-9

[12] Taherdangkoo, M.; Paziresh, M.; Yazdi, M.; Bagheri, M. H. (2012). An efficient algorithm for function optimization: modified stem cells algorithm, Central European Journal of Engineering. 3 (1): 36–50. doi:10.2478/s13531-012-0047-8.