INTERNATIONAL JOURNAL OF ELECTRONICS, MECHANICAL and MECHATRONICS ENGINEERING
Vol.2 Num.1 pp.(72-78)

BREAST CANCER CLASSIFICATION WITH GENETIC PROGRAMMING

Abdurrahim AKGÜNDOGDU

Istanbul University, Engineering Faculty, Electrical and Electronics Eng. Dep. 34320, Avcilar, Istanbul, Turkey

Abstract: *This paper proposes the performance of Genetic Programming (GP) methods on Wisconsin breast cancer data. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset, provided by the University of Wisconsin Hospital, was derived from a group of images using Fine Needle Aspiration (FNA) of the breast. Genetic Programming with different population size was employed to this study. Therefore, GP was trained with 50, 100 and 200 population size and ten-folds cross validation procedure. Results showed %96.6 success rate on 50 population with GP.*

Keywords: *Genetic Programming, Breast Cancer, Classifiers*

1. INTRODUCTION

Breast cancer is the second leading cause of cancer deaths among women in the world (after lung cancer) [1]. This cancer is a disease initially found in the form of tumor in the breast. These tumors are two types: one is benign (non cancerous) and second is malignant (cancerous). These malignant tumors later grow into cancer. However, earlier treatment requires the ability to detect breast cancer in early stages. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones. Thus, finding an accurate and effective diagnosis method is very important. Biopsy is the best way to accurately determine whether the tumor is benign or malignant. Fine needle aspiration (FNA) of breast masses is a cost-effective, non-traumatic, and mostly invasive diagnostic test that obtains information needed to evaluate malignancy (Figure 1).

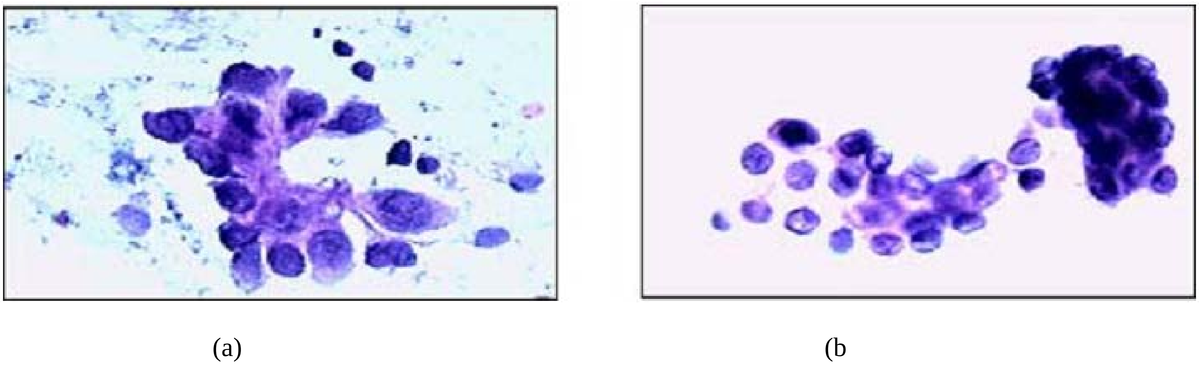


Figure 1: Fine Needle Biopsies of Breast. Malignant (a) and benign (b) breast tumors [2].

The Breast Cancer Diagnosis (BCD) problem has attracted many researchers in computational intelligence, data mining, and statistics fields [3]. Artificial neural networks (ANNs) [4] and support vector machines [5,6] have been recently proposed as a very effective method for pattern recognition, machine learning and data mining. In this paper we examine the performance of Genetic Programming classification on breast cancer data.

2. GENETIC PROGRAMMING

Genetic programming was introduced by Koza in order to automatically generate a program that could solve a given problem [7]. It was similar to the genetic algorithm in many ways, but it was different. The main difference between GP and GA is the representation of the solution. GP creates computer programs as the solution, while GA creates a string of numbers or parameters that impudence the performance of a fixed solution. An individual was represented as a tree composing of functions and terminal representation. Various functions and terminal symbols were advanced for the target application, and classification was one of the aims of genetic programming.

An evolutionary algorithm can be summarized in the following processes [8].
1) GP creates an initial population that consists of a number of individual solutions at random.
2) Randomly select individuals from the population, and compare them with respect to their fitness. The fitness determines the problem the algorithm is expected to solve.
3) Modify an individual with a relatively high fitness using a genetic operator:

4) If the termination criterion is not reached, go to 2.
5) Stop. The best individual represents the best criterion met.

The described procedure [9] is shown in the flowchart of Figure 2.

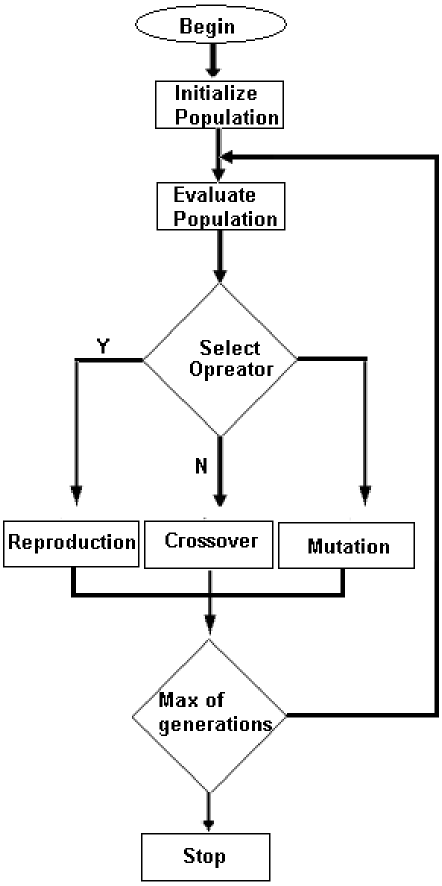


Figure 2: The Structure of Genetic Programming run.

Genetic programming uses tree-like individuals that can represent mathematical expressions, making valuable the application of GP in symbolic regression problems. Tree representation of the GP expression is shown in Fig. 3.

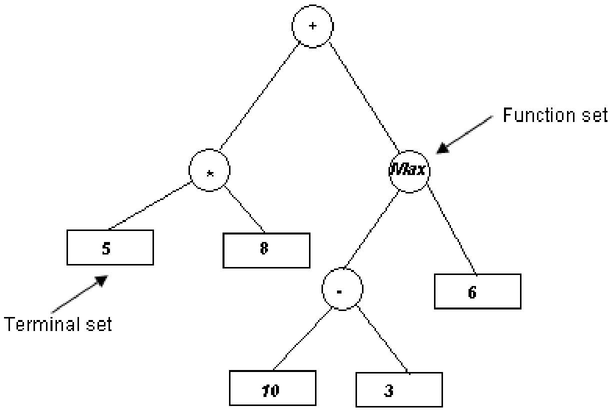


Figure 3: Tree representation of the expression (Max(10-3),6) + (5*8)

3. WISCONSIN DIAGNOSTICS BREAST CANCER (WDBC):

This database is created by William H.Wolberg at University of Wisconsin [10]. This database contains 569 observations among which 357 are benign cases and 212 are malignant cases. We note that in this database that for each observation, there are 30 featured variables. These features are computed from digital images of Fine Needle Aspirates (FNA) of breast masses. These features describe the characteristics of the cell nuclei in the image. Figure 4 shows ribbon 3-D plot of data.

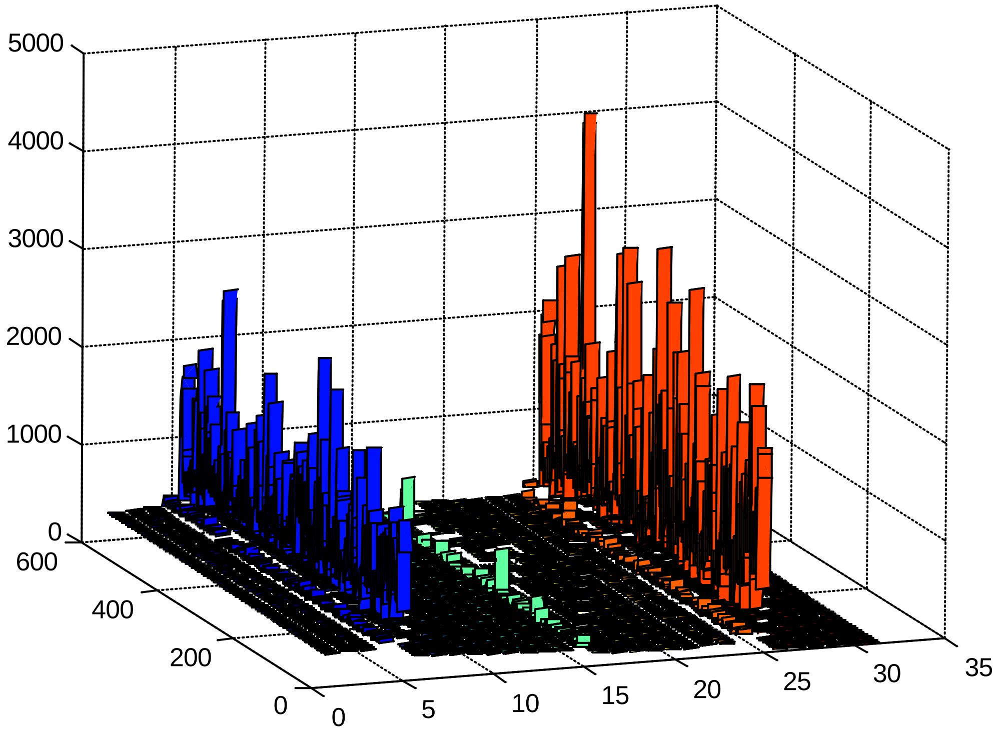


Figure 4: Ribbon 3-D plot of data (obtained from Matlab).

The author of this database considered 10 real-valued features for each cell nucleus:

1. radius (mean of distances from center to points on perimeter);
2. texture (standard deviation of gray-scale values);
3. perimeter;
4. area;
5. smoothness (local variation in radius lengths);
6. compactness ($\frac{(\text{perimeter})^2}{(\text{area-1 })}$);
7. concavity (severity of concave portions of the contour);
8. concave points (number of concave portions of the contour);
9. symmetry;

10. fractal dimension (coastline approximation -1).

They computed the mean, standard error, and worst mean (the mean of the three largest values) of each feature. This process resulted in 30 feature variables for each image.

4. SIMULATION RESULTS

The Genetic Programming was examined benign and malignant classification with WDBC dataset (Figure 5). Different population size was used and tested. Table 1 shows 30 input features that include 569 instances of which 357 are of benign and 212 are of malignant class.

Table 1: Input parameters.

Numbers	Input Features	Numbers	Input Features
1	Mean_Radius	16	Standart_Error_Compactness
2	Mean_Texture	17	Standart_Error_Concavity
3	Mean_Perimeter	18	Standart_Error_Concave_Points
4	Mean_Area	19	Standart_Error_Symmetry
5	Mean_Smoothness	20	Standart_Error_Fractal_Dimension
6	Mean_Compactness	21	Worst_Radius
7	Mean_Concavity	22	Worst_Texture
8	Mean_Concave_Points	23	Worst_Perimeter
9	Mean_Symmetry	24	Worst_Area
10	Fractal_Dimension	25	Worst_Smoothness
11	Standart_Error_Radius	26	Worst_Compactness
12	Standart_Error_Texture	27	Worst_Concavity
13	Standart_Error_Perimeter	28	Worst_Concave_Points
14	Standart_Error_Area	29	Worst_Symmetry
15	Standart_Error_Smoothness	30	Worst_Fractal_Dimension

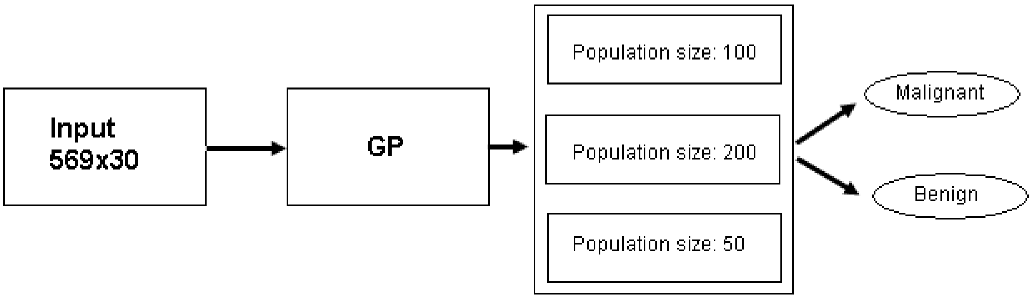


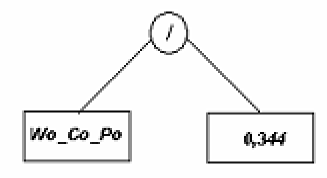
Figure 5: Structure of formation.

4.1. The Improved GP Tree Model

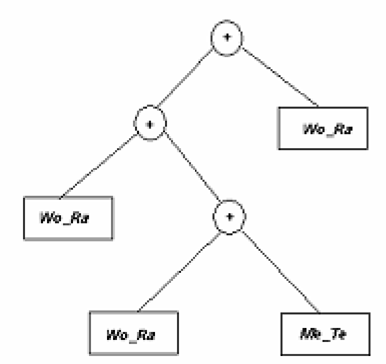
Our model has two classes (malignant and bening). All classes have five elite genetic programs. Table 2 and Table 3 establish elite program's sizes and errors. Figure 6 and figure 7 show models of five elite programs versus malignant and benign class. All simulations are used under Weka program.

Table 2: Elite programs for malignant class.

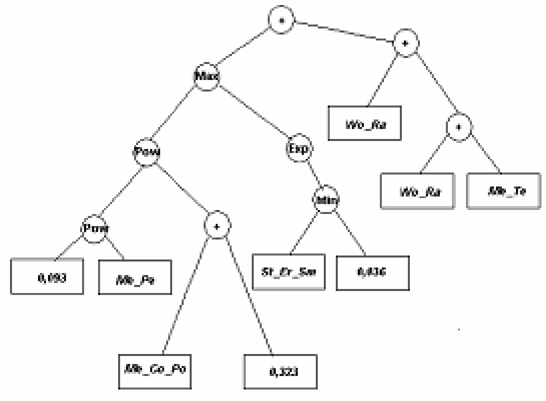
Malignant Class Elite Program No	Size	Training Fitness	Validation Fitness	Error
1	3	0,914	0,892	0,096
2	7	0,906	0,907	0,079
3	18	0,834	0,595	0,128
4	11	0,852	0,846	0,100
5	15	0,614	0,779	0,523



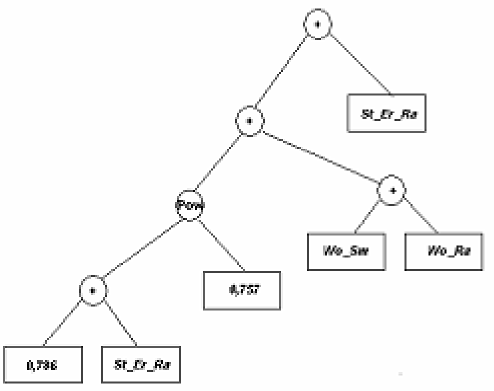
(a)



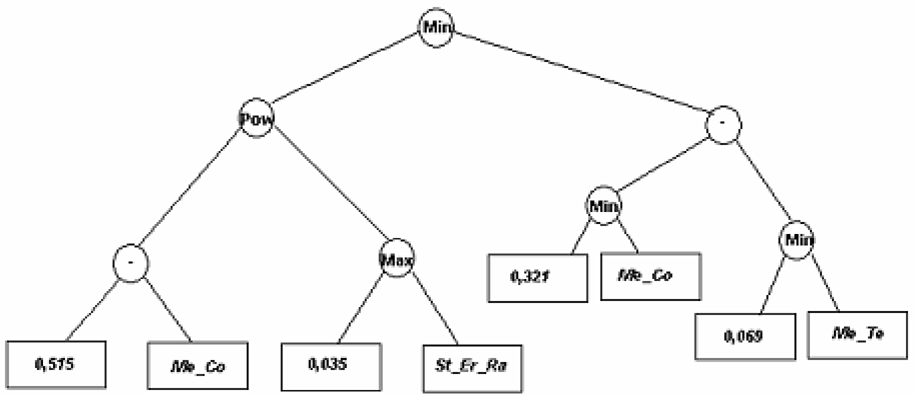
(b)



(c)



(d)



(e)

Figure 6: Tree models of five elite malignant class.

Table 3: Elite programs for benign class.

Benign Class Elite Program No	Size	Training Fitness	Validation Fitness	Error
1	3	0,915	0,919	0,082
2	5	0,776	0,775	0,126
3	15	0,898	0,881	0,172
4	15	0,828	0,779	0,075
5	15	0,703	0,796	0,748

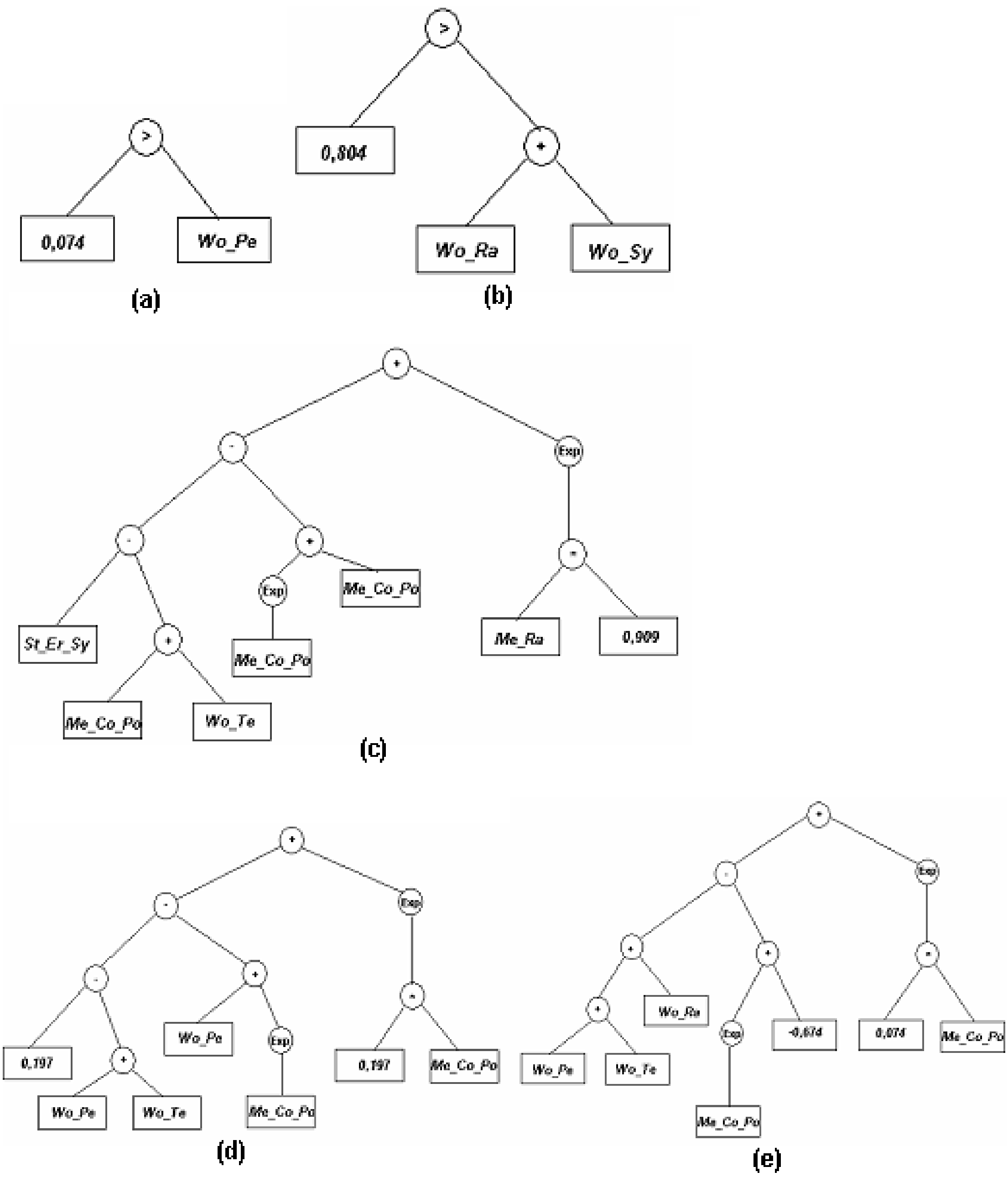


Figure 7: Tree models of five elite benign class.

5. CONCLUSION

In this paper, the performance of Genetic Programming classification was examined with Wisconsin breast cancer dataset (WBCD). 10-fold Cross-validation approach with respectively 100, 200 and 50 population size has been tested with GP system. Accuracy of 96.6% founded at 50 population size with 10-fold Cross-validation. As illustrated in Table 4, confusion matrix is seen with all three population size with success rates.

Table 4: Confusion Matrix

Confusion Matrix				
Population size	**Malignant**	**Benign**		**Successes Rate**
100	**197**	**15**	**Malignant**	**96.1336%**
	7	**350**	**Benign**	
200	**197**	**15**	**Malignant**	**95.2548%**
	12	**345**	**Benign**	
50	**202**	**10**	**Malignant**	**96.6608%**
	9	**348**	**Benign**	

6. REFERENCES

1. http://www.imagins.com/breast
health_cancer.asp.

2. Street, W., Wolberg, W. and Mangasarian, O. , “Machine Learning Techniques to Diagnose Breast Cancer from Image-Processed Nuclear Features of Fine Needle Aspirates”, Cancer Letters Vol. 77 pages:163-171, 1994

3. P. E. H. R. O. Duda and D. G. Stock, editors. *Pattern Classification*, Second Edition. JohnWiley & sons,New york, 2001.

4. Furundzic D., Djordjevic, and Bekic A. J., Neural Networks approach to early breast cancer detection. *Systems Architecture*, 44:617-633, 1998.

5. J. S.-T. N. Cristianini, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

6. C. Hsu and C. Lin, A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, (13):415–425, 2002.

7. J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press, 1992.

8. J.R. Koza, F.H. Bennett III, D. Andre, M.A. Keane, *Genetic Programming III*, Morgan Kaufmann Publ. Inc., 1999.

9. Tsakonas A. “A comparison of classification accuracy of four genetic programming-evolved intelligent structures” *Information Sciences*. 176(6):691-724, 2006

10. Wolberg W. H. and Mangasarian O. L., Multisurface method of pattern separation for medical diagnosis applied to breast cytology, in: *Proceedings of the USA National Academy of Sciences* 87, pp. 91939196, 1990