**A Linear Classifier for Determining the**

**Presence of Malignancy in Breast Tumors**

**Final Report**

**Clemens Vallejo**

**Zayd Hammoudeh**

Table of Contents

[Dataset Overview 3](#_Toc402667269)

[List of References 4](#_Toc402667270)

[Appendix A – Genetic Algorithm Sample Output 5](#_Toc402667271)

[Appendix B – UML Class Diagram for the Breast Cancer Classifier Genetic Algorithm 6](#_Toc402667272)

# Dataset Overview

The source dataset is entitled the “Wisconsin Breast Cancer Database” and was created by a team from the University of Wisconsin Hospitals. The dataset [ 2 ][[1]](#footnote-1) consists of 699 patient tumors[[2]](#footnote-2). Each tumor has 9 distinct features that were individually assigned a value on a 1 to 10 scale; the features are listed below with brief descriptions of the biological differences between benign and malignant tumors [ 3 ].

1. **Clump Thickness:** Benign cells tend to be clumped in monolayers while malignant cells are usually grouped in multilayers.
2. **Uniformity of Cell Size:** Benign cells are more uniform in size while the size of malignant cells can vary significantly.
3. **Uniformity of Cell Shape:** Benign cells have smooth and round edges with a surrounding fibrous capsule that is very well-circumscribed. In contrast, the edges of the malignant cells are usually very distinct and lack this uniformity.
4. **Marginal Adhesion:** A malignant mass is mobile and not attached to surrounding tissue. Benign masses tend to stick more tightly together.
5. **Single Epithelial Cell Size:** Epithelial cells line the cavities and surfaces of structures throughout the body. In benign tumors, surrounding epithelial cells form a single layer with normal cell size; malignant tumors usually have significantly enlarged epithelial cells.
6. **Bare Nuclei:** A bare nucleus is devoid of surrounding cytoplasm (i.e. the rest of the cell). They are more typically seen in benign tumors.
7. **Bland Chromatin:** In benign tumors, the nucleus generally has a uniform texture. In contrast, cancerous cells tend to have coarser nuclei.
8. **Normal Nucleoli:** Nucleoli are small structures in the cell nucleus. In benign cells, the nucleolus is very small and barely visible (if at all). Malignant cells have more prominent nucleoli, and in some cases, they are more numerous as well.
9. **Mitoses:** It is the process in which the cell replicates and divides. Mitosis is rapid and uncontrolled in malignant cells.

Those tumors which are malignant are marked with a “4” while benign tumors are marked with a “2”.

The dataset is published as a text file in comma separated variable (CSV) format.

# List of References

|  |  |
| --- | --- |
| [ ] | Wolberg, W.H. & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis apply to breast Cytology. In Proceedings of the National Academy of Sciences, 87, 9193-9196. |
| [ ] | Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. |
| [ ] | Salama et. al. “Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers.” International Journal of Computer and Information Technology. September 2012. |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Appendix A – Genetic Algorithm Sample Output

Below is a sample output of the genetic algorithm.

After 1 run, the percent correct on the training set is: 97.00

After 2 run, the percent correct on the training set is: 97.00

After 3 run, the percent correct on the training set is: 97.00

After 4 run, the percent correct on the training set is: 97.00

After 5 run, the percent correct on the training set is: 97.00

On the training set, the score for the best solution is: 194

The maximum possible score is: 200

The percent correct is: 97.00

The linear function weights are:

Clump Thickness Weight: -73149407

Cell Size Uniformity Weight: 806500191

Cell Shape Uniformity Weight: 383976807

Marginal Adhesion Weight: 7331932

Single Epithelial Cell Size Weight: -714736227

Bare Nucleoli Weight: 476423725

Bland Chromatin Weight: -100514872

Normal Nucleoli Weight: 48835613

Mitoses Weight: -52238695

Offset Weight: -2094917374

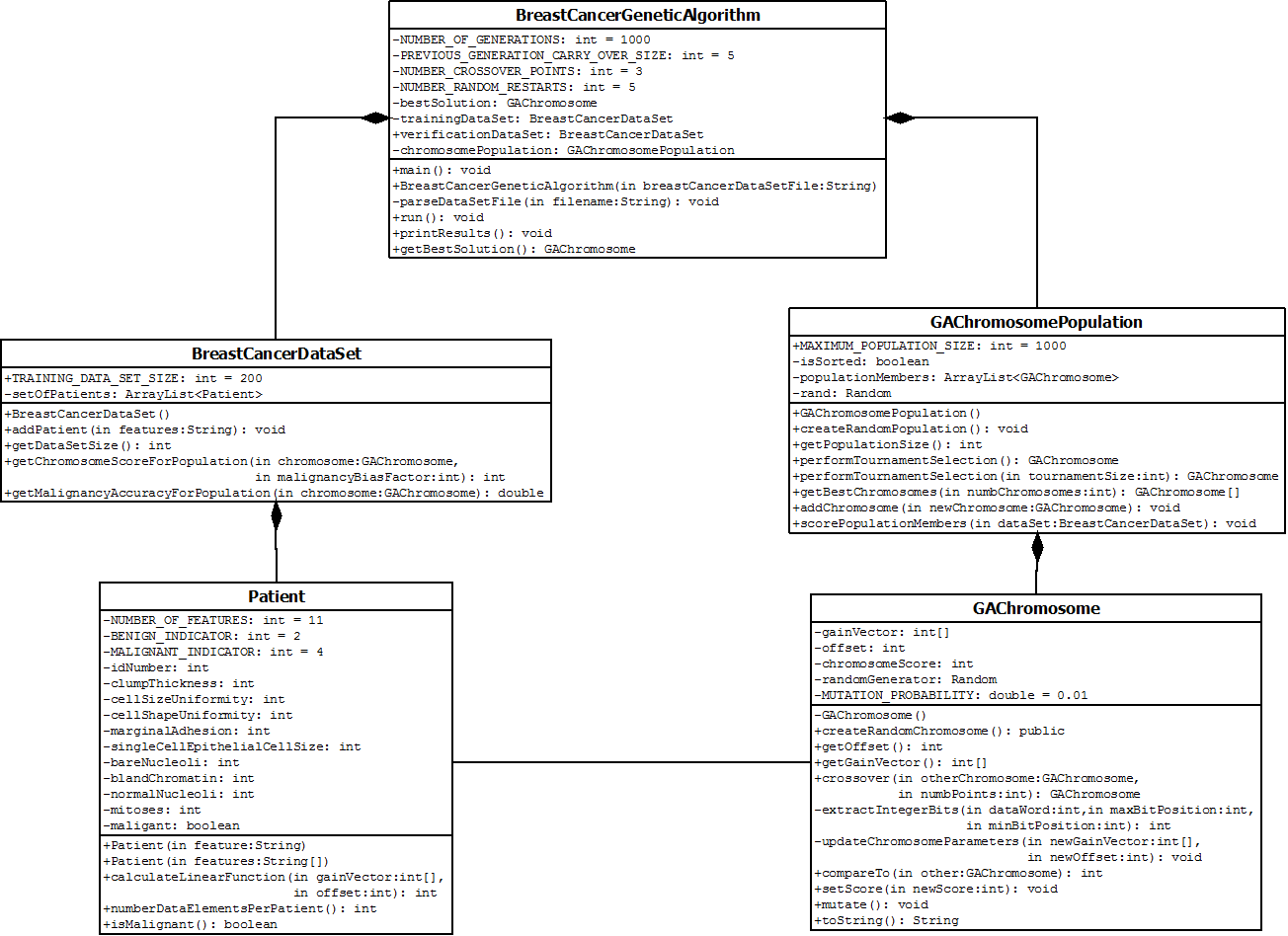
On the verification set, the score for the best solution is: 456

The maximum possible score is: 483

The percent correct is: 94.41

The percentage of malignant tumors correctly categorized is: 92.66%.

# Appendix B – UML Class Diagram for the Breast Cancer Classifier Genetic Algorithm

****

1. The dataset is available in the University of California, Irvine’s Machine Learning Repository. A link to the dataset is: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)>. [↑](#footnote-ref-1)
2. 16 of the 699 tumors have incomplete data so only 683 instances are used in our analysis. [↑](#footnote-ref-2)