

1. Title: Wisconsin Diagnostic Breast Cancer (WDBC)

2. Source Information

a) Creators:

Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792  
wolberg@eagle.surgery.wisc.edu

W. Nick Street, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706  
street@cs.wisc.edu 608-262-6619

Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706  
olvi@cs.wisc.edu

b) Donor: Nick Street

c) Date: November 1995

3. Past Usage:

first usage:

W.N. Street, W.H. Wolberg and O.L. Mangasarian  
Nuclear feature extraction for breast tumor diagnosis.  
IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

OR literature:

O.L. Mangasarian, W.N. Street and W.H. Wolberg.  
Breast cancer diagnosis and prognosis via linear programming.  
Operations Research, 43(4), pages 570-577, July-August 1995.

Medical literature:

W.H. Wolberg, W.N. Street, and O.L. Mangasarian.  
Machine learning techniques to diagnose breast cancer from fine-needle aspirates.  
Cancer Letters 77 (1994) 163-171.

W.H. Wolberg, W.N. Street, and O.L. Mangasarian.  
Image analysis and machine learning applied to breast cancer  
diagnosis and prognosis.  
Analytical and Quantitative Cytology and Histology, Vol. 17  
No. 2, pages 77-87, April 1995.

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian.  
Computerized breast cancer diagnosis and prognosis from fine  
needle aspirates.  
Archives of Surgery 1995;130:511-516.

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian.  
Computer-derived nuclear features distinguish malignant from  
benign breast cytology.  
Human Pathology, 26:792--796, 1995.

See also:

<http://www.cs.wisc.edu/~olvi/uwmp/mpml.html>  
<http://www.cs.wisc.edu/~olvi/uwmp/cancer.html>

Results:

- predicting field 2, diagnosis: B = benign, M = malignant
- sets are linearly separable using all 30 input features
- best predictive accuracy obtained using one separating plane  
in the 3-D space of Worst Area, Worst Smoothness and  
Mean Texture. Estimated accuracy 97.5% using repeated  
10-fold crossvalidations. Classifier has correctly  
diagnosed 176 consecutive new patients as of November  
1995.

#### 4. Relevant information

Features are computed from a digitized image of a fine needle  
aspirate (FNA) of a breast mass. They describe  
characteristics of the cell nuclei present in the image.  
A few of the images can be found at  
<http://www.cs.wisc.edu/~street/images/>

Separating plane described above was obtained using  
Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree  
Construction Via Linear Programming." Proceedings of the 4th  
Midwest Artificial Intelligence and Cognitive Science Society,  
pp. 97-101, 1992], a classification method which uses linear

programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in:

[K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:

```
ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/
```

5. Number of instances: 569

6. Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)

7. Attribute information

1) ID number

2) Diagnosis (M = malignant, B = benign)  
3-32)

Ten real-valued features are computed for each cell nucleus:

- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness ( $\text{perimeter}^2 / \text{area} - 1.0$ )
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry
- j) fractal dimension ("coastline approximation" - 1)

Several of the papers listed above contain detailed descriptions of how these features are computed.

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image,

resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

8. Missing attribute values: none

9. Class distribution: 357 benign, 212 malignant