

# User Manual

## Breast Cancer Diagnosis Web User Interface

---

**Author: Reda MERZOUKI**

## Contents

1	Summary.....	2
2	Description.....	3
2.1	Attributes description.....	3
2.2	Attributes grading scale.....	3
2.3	Attributes grading signification.....	4
3	Application's benefits.....	5
3.1	The fine needle aspiration technique.....	5
3.2	Machine Learning in the breast cancer diagnosis chain.....	5
4	Materials used for application's development.....	6
5	Tool Utilization.....	7
5.1	Launching.....	7
5.2	Diagnosis.....	7
5.2.1	Malignant case.....	7
5.2.2	Benign case.....	9
6	Acknowledgements.....	10
7	References.....	11

# 1 Summary

This application is intended for pathologists who grade the 9 following features of “Fine Needle Aspiration” in accordance with Wisconsin Dataset's grading scale<sup>1</sup> so that to determine whether a **breast mass** is benign or malignant:

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses

After rigorous tuning and training cycles of 8 models over 489 train data randomly selected from the available dataset, Random Forest algorithm turned out to be the best model.

**98.1%** of 210 test cases are correctly classified and **100%** of the cancerous tumors are perfectly diagnosed by the algorithm.

In other words, the algorithm provides a classification error of **1.9% (+/-1.85)** which is the percentage of incorrect predictions to the number of predictions made, moreover, the algorithm does not miss any cancerous tumor and then it is **100% sensitive to malignancy** (recall score is 100%).

These performance results were obtained with 210 test data that have never been seen by the algorithm during its training step (143 benign tumors / 67 malignant tumors).

Test data is a sample of the available data that has been randomly selected and removed from the available data, such that it is not used during model selection or configuration.

Lastly, please note that algorithm's tuning and training were performed with cross validation method (StratiFiedKfold).

---

<sup>1</sup> Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193—9196.  
<http://www.pnas.org/content/87/23/9193.full.pdf>

## 2 Description

### 2.1 Attributes description

Attributes	Description
Clump Thickness	Assesses if cells are mono or multi-layered
Uniformity of Cell Size	Evaluate the consistency in size of the cells in the sample
Uniformity of Cell Shape	Evaluate the consistency in shape of the cells in the sample
Marginal Adhesion	Quantifies proportion of cells that stick together
Single Epithelial Cell Size	Measures the enlargement of epithelial cells size
Bare Nuclei	Proportion of nuclei surrounded by cytoplasm versus those that are not
Bland Chromatin	Rates the uniform "texture" of the nucleus in a range from fine to coarse
Normal Nucleoli	Determines whether the nucleoli are small and barely visible or larger, more visible, and more plentiful.
Mitoses	Describes the level of mitotic activity

### 2.2 Attributes grading scale

Attributes are integer values belonging to [1,10] interval.

Attributes	Grading_Scale
Clump Thickness	1 to 10
Uniformity of Cell Size	1 to 10
Uniformity of Cell Shape	1 to 10
Marginal Adhesion	1 to 10
Single Epithelial Cell Size	1 to 10
Bare Nuclei	1 to 10
Bland Chromatin	1 to 10
Normal Nucleoli	1 to 10
Mitoses	1 to 10

## 2.3 Attributes grading signification

Feature	Grade	Signification
Clump Thickness	1	Cells are fully mono-layered
	2	Cells are 90% mono-layered
	3	Cells are 80% mono-layered
	4	Cells are 65% mono-layered
	5	Cells are slightly more mono-layered than multi-layered
	6	Cells are slightly more multi-layered than mono-layered
	7	Cells are 35% mono-layered
	8	Cells are 20% mono-layered
	9	Cells are 10% mono-layered
	10	Cells are multi-layered
Uniformity of Cell Size	1	Cells are completely uniform
	2	Cells are 90% uniform
	3	Cells are 80% uniform
	4	Cells are 65% uniform
	5	Cells are more than 50% uniform
	6	Cells are less than 50% uniform
	7	Cells are 35% uniform
	8	Cells are 20% uniform
	9	Cells are 10% uniform
	10	Cells are inconsistent with their uniformity
Uniformity of Cell Shape	1	Completely uniform
	2	Cells are 90% uniform
	3	Cells are 80% uniform
	4	Cells are 65% uniform
	5	Cells are more than 50% uniform
	6	Cells are less than 50% uniform
	7	Cells are 35% uniform
	8	Cells are 20% uniform
	9	Cells are 10% uniform
	10	Cells are inconsistent with their uniformity
Marginal Adhesion	1	Completely stick together
	2	90% stick together
	3	80% stick together
	4	70% stick together
	5	60% stick together
	6	50% stick together
	7	40% stick together
	8	30% stick together
	9	20% stick together
	10	Cells do not exhibit marginal adhesion
Single Epithelial Cell Size	1	No cells are significantly enlarged
	2	Largest cells appear 20% larger
	3	Largest cells appear 30% larger
	4	Largest cells appear 40% larger
	5	Largest cells appear 50% larger
	6	Largest cells appear 60% larger
	7	Largest cells appear 70% larger
	8	Largest cells appear 80% larger
	9	Largest cells appear 90% larger
	10	Largest cells appear 100% larger
Bare Nuclei	1	Nuclei completely devoid of cytoplasm
	2	20% of nuclei have cytoplasm
	3	30% of nuclei have cytoplasm
	4	40% of nuclei have cytoplasm
	5	50% of nuclei have cytoplasm
	6	60% of nuclei have cytoplasm
	7	70% of nuclei have cytoplasm
	8	80% of nuclei have cytoplasm
	9	90% of nuclei have cytoplasm
	10	All nuclei have cytoplasm
Bland Chromatin	1	Completely fine textured chromatin
	2	Chromatin is 20% coarse
	3	Chromatin is 30% coarse
	4	Chromatin is 40% coarse
	5	Chromatin is 50% coarse
	6	Chromatin is 60% coarse
	7	Chromatin is 70% coarse
	8	Chromatin is 80% coarse
	9	Chromatin is 90% coarse
	10	Chromatin is completely coarse
Normal Nucleoli	1	Nucleoli are completely normal (small, one per cell, barely visible)
	2	20% of nucleoli are abnormal
	3	30% of nucleoli are abnormal
	4	40% of nucleoli are abnormal
	5	50% of nucleoli are abnormal
	6	60% of nucleoli are abnormal
	7	70% of nucleoli are abnormal
	8	80% of nucleoli are abnormal
	9	90% of nucleoli are abnormal
	10	100% of nucleoli are abnormal
Mitoses	1	Mitotic activity is completely normal
	2	20% of mitotic activity appears abnormal
	3	30% mitotic activity appears abnormal
	4	40% mitotic activity appears abnormal
	5	50% mitotic activity appears abnormal
	6	60% mitotic activity appears abnormal
	7	70% mitotic activity appears abnormal
	8	80% mitotic activity appears abnormal
	9	90% mitotic activity appears abnormal
	10	100% mitotic activity appears abnormal

## 3 Application's benefits

### 3.1 The fine needle aspiration technique

The data **Breast Cancer Wisconsin Original Dataset**<sup>2</sup> that were analyzed were derived from fine needle aspiration over breast masses. The Fine Needle Aspiration technique is deemed to be the least invasive of all those practiced to date.

In the context of this medical examination the patient does not undergo surgery. Actually, it is a sample of the tumor fluid by means of a needle, which, if necessary, can be guided using an ultrasound.

This medical examination is:

- Less traumatic for the patient than the surgical biopsy.
- Relatively fast (15 minutes).
- Less expensive than a traditional biopsy.

### 3.2 Machine Learning in the breast cancer diagnosis chain

With Machine Learning, we can automate the tumor diagnosis through the grading of "Fine Needle Aspiration" attributes described in 3.1.

In a schematic way, the diagnosis process is performed in two stages and usually lasting one to two weeks<sup>3</sup>:

- 1- Pathologist or laboratory technician grades the attributes of the puncture.
- 2- Pathologist makes the diagnosis thanks to the grading of the attributes.

We propose to automate the second step using Machine Learning ; in our use case we have 9 attributes for each data.

Thus, the predictive tool that we built could greatly contribute to the promotion of FNA, and, in particular, the benefit of the predictive algorithm is fourfold:

- 1- Better medical comfort for the patients.
- 2- Faster and safer diagnosis.
- 3- Technique that can be generalized to any preoperative examination.
- 4- An economic benefit: in fact, surgical biopsies require more resources and are therefore more expensive than FNA.

---

<sup>2</sup> Dr. William H. Wolberg (physician) University of Wisconsin Hospitals, Madison, Wisconsin, USA (1992-07-15). Breast Cancer Wisconsin (Original) Data Set.

[https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Original\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original))

<sup>3</sup> <http://www.imagerie114.fr/specialites/imagerie-de-la-femme/les-prelevements-du-sein-cytoponction-microbiopsie-ou-macrobiopsie/>

## 4 Materials used for application's development

- A computer (Processor : 2GHz, RAM : 8 Go, OS 64 bits).
- Python version 3.5.4 (with Anaconda distribution).
- Scikit-Learn version 0.19.0.
- Pandas version 0.21.1.
- Numpy version 1.13.3.
- Matplotlib version 2.1.1.
- Jupyter Notebook version 5.2.2
- R, version 3.3.1 (2016-06-21).
- Node.js, version 9.4.0 (for the GUI).
- Breast Cancer Wisconsin Original Dataset.

## 5 Tool Utilization

### 5.1 Launching

Open a browser.

The application can be launched with the URL: <https://www.rai-light.com>

Or

<https://bcd/www.rai-light.com>

### 5.2 Diagnosis

#### 5.2.1 Malignant case

##### 5.2.1.1 Probability

- 1- Please enter the 9 graded attributes of breast Fine Needle Aspiration so that to determine whether the breast mass is malignant or benign.
- 2- Click the button "LAUNCH DIAGNOSIS".

Data to Diagnose

Please enter the 9 graded attributes of breast Fine Needle Aspiration.

CLUMP THICKNESS: 1	UNIFORMITY OF CELL SIZE: 8
UNIFORMITY OF CELL SHAPE: 8	MARGINAL ADHESION: 2
SINGLE EPITHELIAL CELL SIZE: 4	BARE NUCLEI: 4
BLAND CHROMATIN: 5	NORMAL NUCLEOLI: 8
MITOSES: 10	

LAUNCH DIAGNOSIS

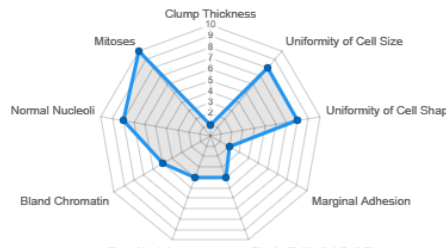
Algorithm's Diagnosis

Tumor Diagnosis: **Malignant**

Probability: **93.25%**

Algorithm provides the diagnosis under the following performances assessed over 210 test data :

- Classification\_error = 1.90 % (+/- 1.85)
- 95% Confidence Interval of the Classification\_error is : [ 0.0006 , 0.0375 ]
- Accuracy\_score = 98.10 %
- Recall\_score = 100.00 %
- Precision\_score = 94.37 %



- 3- You get the **Diagnosis** (in this case Malignant) on the right side of the screen along with the probability and performances of the algorithm. You also get a radar chart that recap all the attributes values you entered.



### 5.2.1.2 Malignant case in 2D through Principal Components Analysis (PCA)

After having displayed the Diagnosis, the application displays the data to diagnose in 2 dimensions through a 2D PCA.

Useful details are given along with this 2D representation.

#### 2D Visualization through Principal Components Analysis (PCA)

This representation is the projection's result of the 9 normalized attributes of 210 test data and data to diagnose on the two principal components.

PCA 2D allows us then to reduce data's dimension from 9D to 2D. Our Algorithm applied for diagnosis step is now fitted over 2D reduced training data.

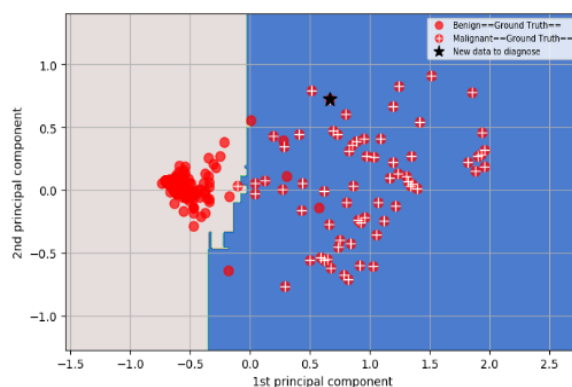
Then the Data Visualization function displays the data to be diagnosed in a 2D graph as well as 210 labeled test data not seen by the Algorithm. This function also plots the decision boundary of the algorithm applied on test data.

Algorithm classifies 2D data with the following performances assessed over 210 test data:

- 2D Classification\_error = 2.86 % (+/- 2.25)
- 95% Confidence Interval of the 2D Classification\_error is: [ 0.0060 , 0.0511 ]
- accuracy\_score = 97.14 %
- recall\_score = 98.51 %
- precision\_score = 92.96 %

The Algorithm's boundary divides the space onto two areas with two different colors:

- Gray area is the one of benign cases.
- Blue area is the one of malignant cases.



**CAVEATS :** Please note that the algorithm provides better performances on original data than on reduced data. In case 2D graphical representation provides different result from the DIAGNOSIS computed on original data, please use the result provided by the DIAGNOSIS.

The data to diagnose is the “black star” ; you can see clearly that the “black star” is located in the “malignant area”.

**Please note that 2D graphical representation could provide different results from the DIAGNOSIS computed on original data. More generally, please use this 2D representation as a secondary information.**

**In other words, as a user of this application, your decision as a healthcare professional should rely only and exclusively on the diagnosis result provided in “Algorithm's Diagnosis” section.**

## 5.2.2 Benign case

### 5.2.2.1 Probability

Same steps than in section 5.2.1.1

Data to Diagnose

Please enter the 9 graded attributes of breast Fine Needle Aspiration.

CLUMP THICKNESS: 2	UNIFORMITY OF CELL SIZE: 2
UNIFORMITY OF CELL SHAPE: 2	MARGINAL ADHESION: 2
SINGLE EPITHELIAL CELL SIZE: 3	BARE NUCLEI: 2
BLAND CHROMATIN: 3	NORMAL NUCLEOLI: 3
MITOSES: 8	

LAUNCH DIAGNOSIS

Algorithm's Diagnosis

Tumor Diagnosis: **Benign**

Probability: **86.89%**

Algorithm provides the diagnosis under the following performances assessed over 210 test data :

- Classification\_error = 1.90 % (+/- 1.85)
- 95% Confidence Interval of the Classification\_error is : [ 0.0006 , 0.0375 ]
- Accuracy\_score = 98.10 %
- Recall\_score = 100.00 %
- Precision\_score = 94.37 %

### 5.2.2.2 Benign case in 2D through Principal Components Analysis (PCA)

See section 5.2.1.2 for more details.

#### 2D Visualization through Principal Components Analysis (PCA)

This representation is the projection's result of the 9 normalized attributes of 210 test data and data to diagnose on the two principal components.

PCA 2D allows us then to reduce data's dimension from 9D to 2D. Our Algorithm applied for diagnosis step is now fitted over 2D reduced training data.

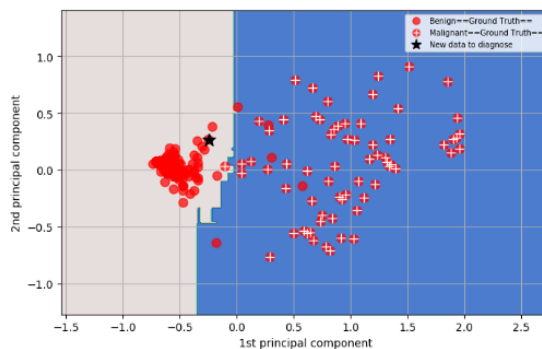
Then the Data Visualization function displays the data to be diagnosed in a 2D graph as well as 210 labeled test data not seen by the Algorithm. This function also plots the decision boundary of the algorithm applied on test data.

Algorithm classifies 2D data with the following performances assessed over 210 test data:

- 2D Classification\_error = 2.86 % (+/- 2.25)
- 95% Confidence Interval of the 2D Classification\_error is: [ 0.0060 , 0.0511 ]
- accuracy\_score = 97.14 %
- recall\_score = 98.51 %
- precision\_score = 92.96 %

The Algorithm's boundary divides the space onto two areas with two different colors:

- Gray area is the one of benign cases.
- Blue area is the one of malignant cases.



CAVEATS : Please note that the algorithm provides better performances on original data than on reduced data. In case 2D graphical representation provides different result from the DIAGNOSIS computed on original data, please use the result provided by the DIAGNOSIS.

The “black star” is clearly located in the “benign” area.

## 6 Acknowledgements

In the early 1990's, Professors William H. Wolberg and Olvi L. Mangasarian at the University of Wisconsin published a near 700-sample dataset of breast cancer masses.

These masses had been biopsied via fine needle aspirates.

Nine cytological characteristics of breast FNAs were valued on a scale of 1 to 10, with 1 being the closest to benign and 10 the most anaplastic.

This data was then published to the University of California Irvine's Machine Learning Repository as public domain.

I am grateful for access to this data, as it provided my algorithm with training and testing data.

The data was also very appropriate for the classification task.

I would also like to acknowledge Brittany Wenger for her contribution in this domain.

In 2012, based on Wisconsin database, Brittany Wenger provided a service<sup>4</sup> built on a neural nets algorithm.

I thank Axel Tessier for his great contribution on the web user interface of my application and also for making a secured server available for this application.

Finally, I would like to thank my family for their continuous support throughout this project.

---

<sup>4</sup> <http://cloud4cancer.appspot.com/>

## 7 References

Dr. William H. Wolberg (physician) University of Wisconsin Hospitals, Madison, Wisconsin, USA (1992-07-15). Breast Cancer Wisconsin (Original) Data Set.

Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193--9196.

Brittany Wenger, <http://cloud4cancer.appspot.com>.

Haibo He, Eduardo A. Garcia (September 2009). Learning from Imbalanced Data, IEEE Transactions on knowledge and data engineering VOL. 21, NO. 9.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer (2002). SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321--357.

Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009). The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition, Springer.

Patrick H. Winston (Fall 2010). Learning: Support Vector Machines, MIT 6.034 Artificial Intelligence.

Patrick H. Winston (Fall 2010). Learning: Boosting, MIT 6.034 Artificial Intelligence.  
<http://ocw.mit.edu/6-034F10>

Pierre-André Cornillon, Arnaud Guyader, François Husson, Nicolas Jégou, Julie Josse, Maela Kloaberg, Eric Matzner-Lober, Laurent Rouvière (2012). Statistiques avec R, Presses Universitaires de Rennes.

Gilbert Saporta (2011) Probabilités Analyse des données et Statistique, Editions Technip.