Available online at
www.mdl.mazedan.com

MCET

www.mazedan.com

www.mazedan.com/mcet

# BREAST CANCER PREDICTION BASED ON THE SUPERVISED MACHINE LEARNING

EI PHYU SIN WIN

MAZEDAN COMPUTER ENGINEERING TRANSACTIONS

e-ISSN: 2583-0414

Article id-MCET0201001

Vol.-2, Issue-1

Received: 17 Jan 2021

Revised: 12 Feb 2021

Accepted: 28 Feb 2021

**Citation:** Win, E. P. S. (2021). Breast Cancer Prediction Based on The Supervised Machine Learning. *Mazedan Computer Engineering Transactions*, 2(1), 1-5.

## Abstract

Today, millions of women around the world suffer from breast cancer on a daily basis. It has been noted that women's beauty lies in their skin, complexion, breasts and buttocks. However, most women are struggling to make ends meet, and their beauty and self-esteem are at stake. Very few people are diagnosed with breast cancer. This is because there is only a lump and no symptoms. If left untreated, it can be life-threatening and can be life-threatening. Therefore, it is important to get an early diagnosis. This finding could lead to a combination of accurate algorithms based on ultrasound imaging to provide breast cancer prognostic results. In this paper, the first ultrasound images were obtained using a combination of adaptive median filter, GMM segmentation, GLM feature extraction, Probabilist Neural Network (PNN), and KNN classifier to identify the stages of breast cancer. It simulates with MATLAB Programming Language.

**Keywords:** BREAST CANCER, ULTRASOUND IMAGE, GMM, PNN AND KNN

## 1. INTRODUCTION

Most women around the world suffer from breast cancer and need special case. Breast cancer can be fatal, but it is not a disease that can be treated in time, it is a disease that can be treated in a timely manner. The onset of breast cancer begins with a tumor on one side of the breast. There may be no initial pain, but after a few months it will become large and hard, and the environment will experiment will experience pain. Taking painkillers can provide temporary relief and then return the pain. In that case, you need to consult a specialist and follow the instructions exactly.

An ultrasound scan of the breast is performed, which is performed by a physician, who then performs medical or surgical procedures. Even after surgery, the incidence of cancer may be high and diagnosis may need to be made. At this point, you may face issues such as high costs and depression. The feelings of a woman who has had her breast amputated and lost her beauty can also lead to a loss of self-confidence.

The procedure provides a way to describe the extent of breast cancer, including tumor size, whether it has distributed to lymph nodes, metastases to distant regions of the body, and biological indicators. The term can be done before or after the surgery of the patient. Physicians use diagnostic experiments to determine the stage of cancer, so this stage may not be completed until all tests have been completed. Understanding the procedure helps doctors determine the best course of action and predict the inpatient's forecast - the likelihood of improvement. There are different stages of characterizations for different types of cancer. This research uses ultrasound pictures to predict the stage of breast cancer. Differential theories are used to classify the stage of prediction such as normal, Begnin and Melanoma. Cancer data sets are also used to obtain accurate data sets. Once you have the images you can test, use the adaptive median filtering method to refine the image. It is then split using the Otsu thresholding method. GLCM is used to feature that segmented image. A probabilistic Neural Network trains and test with database. Finally, the KNN classifier classifies the stage of cancer.

## 2. LITERATURE REVIEW

Over the years, cancer medical images have been researched using a variety of purposes and methods. Their various discoveries and concepts are also summarized below.

The researcher, Bocchi (2008) stated as "Regional growth algorithm, and Microcalcification, which is the traditional sign of breasts cancer, is the proposed identification and classification method, microcalcification. Special attention was paid to the investigation. The mammogram is defined as the lesion of the score model would be used [1], and allows coordinate filtering. On the contrary, the steps to improve microcalcification frame. He found an algorithm linked by a neural classifier. Therefore, another score model will retain existing lesions Used to check

Department of Information Technology Engineering, Technological University (Kyaukse), Myanmar

*Corresponding author email- panthakhin9001@gmail.com

their collection. The next researchers, B.M. Gayathri, T. Santhanam and, C.P. Sumath (2013) [2]. The most important cancer expressions for cancer are often analyzed. The real identity to reduce female mortality and Breast cancer is important. There is no doubt that success is an important step. A new automatic illness Awareness of clustered small calculations in computer systems by Book Joyson. This new method Coding the surface of the breast radiograph based on the concept. The Haralik option is estimated for the SVM classification. Intended with the help of test results obtained Coding techniques have been tried in literature. Haralik's expected time has been significantly reduced. The parameter vector pre-ratio is started Classification at expense of encrypted images and have lots of improvements.

Fadi Abu-Amara and Ikhlas Abdel-Qader [3] presented a new system,which is called Computer-Aided Sensing (CAD) for pictures. Mammography is used to identify and classify images. This system uses three modules and it is a function extraction module, the main function of component analysis. It uses autonomous element analysis and a subset of features. The selection module with rough the set model disguised in this system classifier is to display pictures as normal and irregular. Therefore, this classification belongs to a fairly established model.Used to reduce the impact of inconsistent information. Finally, the consequences show that the system has a validity of 84.03% and a percentage of recall 87.28%.

## 3. DESIGN AND METHODOLOGY

This section describes the research methods clearly. The whole process is represented by the flow chart given in Figure-1.

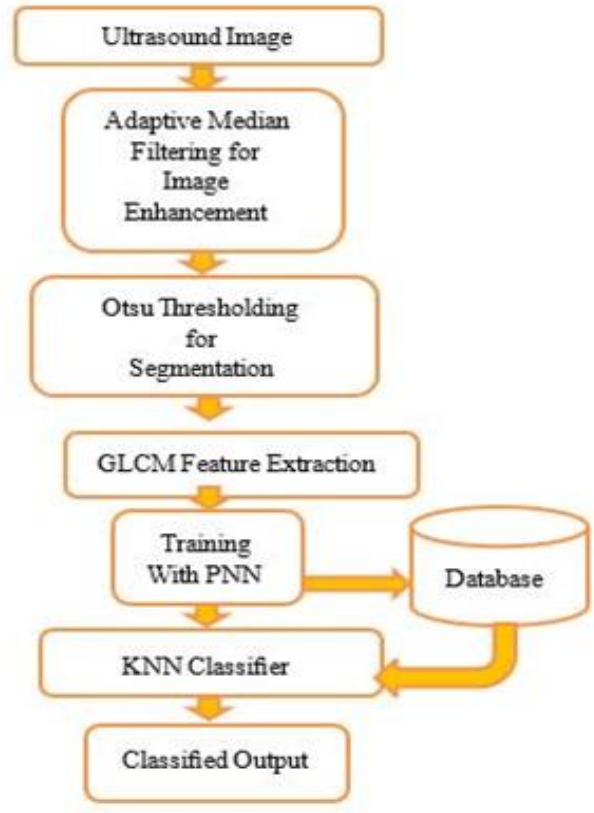


*Figure 1 System Design*

**Data Acquisition**

This journal researcher gets the images they want to test from ultrasound images and search for specific 450 image data sets for Beningn, 200 images for Malign, and 50 images for Normal and research them. The example of data sets is given in figure-2. The stages of breast cancer size are as follow Figure 3(a), the picture also shows how the cancer has spread over time as Figure 3(b), and the nearest view of breast cancer in Figure 3(c).

**Adaptive Median Filtering for Image Enhancement**

To see what tunable median filtering is, first you need to understand what median filtering is and what it does. For various types of digital image processing, the basic tasks are: For each pixel in the electronic picture, it places a neighbor around that point, analyzes the values of all neighboring pixels, according to a specific recursion, and follows it. According to the analysis performed on the neighboring pixels "the neighbor constantly moves and repeats this process for every pixel in the image".

The main intent of these filters is to reduce noise. However, you can use filters to highlight definite features of an image or remove other attributes. Remove words from images using an adaptive average filter. It is the best spatial filter, and it separates the sound of details.

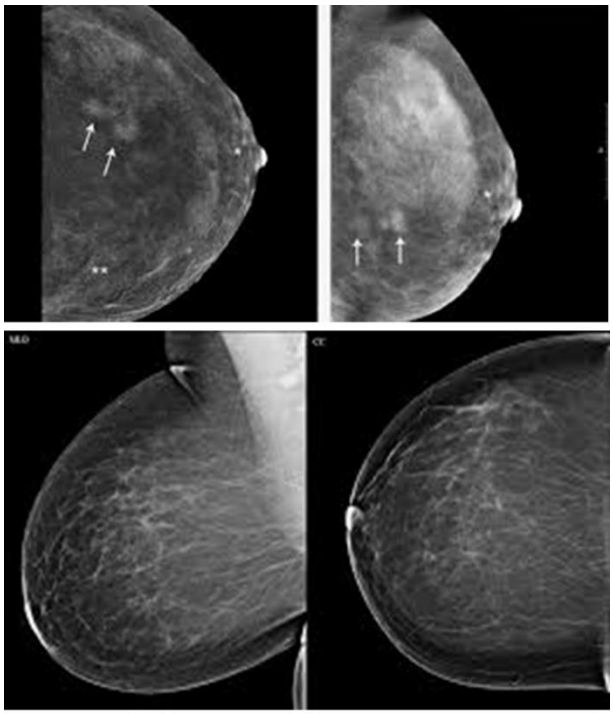

*Figure 2 The example of Datasets [4]*

Level A:
- $A1 = z_{med} - z_{min}$, $A2 = z_{med} - z_{max}$
- If A1 > 0 and A2 < 0, Go to level B
- Else increase window size
- If window size <= $S_{max}$, repeat the level A
- Else the output $z_{xy}$

Level B:
- $B1 = z_{xy} - z_{min}$, $B2 = z_{xy} - z_{max}$
- If B1 > 0 and B2 < 0, output $z_{xy}$
- Else the output $z_{med}$

Level C:

- Let $S_{xy}$ be a neighborhood of the pixel (x,y).
- Let $S_{max}$ be the maximum allowed size of the neighborhood.
- $Z_{min}$ be the minimum the gray value in $S_{xy}$
- $Z_{max}$ be the minimum the gray value in $S_{xy}$
- $Z_{xy}$ be the gray level at the pixel (x, y)
- $Z_{med}$ be the median of the gray values in $S_{xy}$

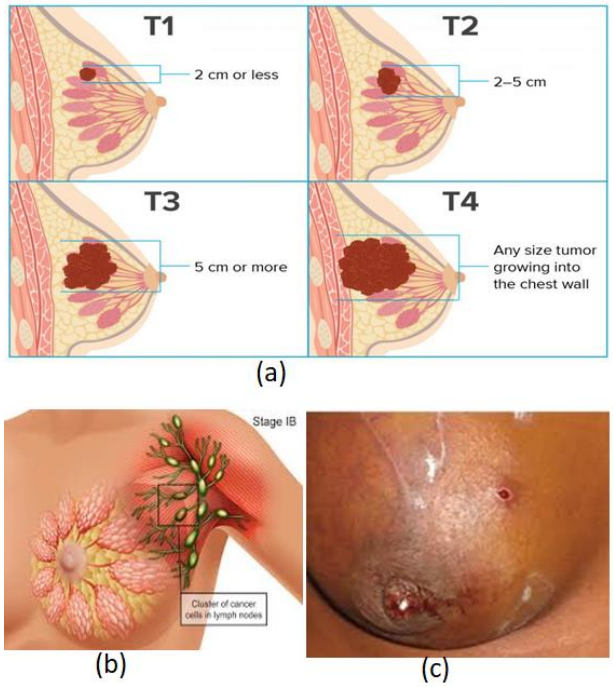


*Figure 3 .(a) The Size of Breast Tumor and (b) The Spreading Stage (c) The nearest view of Breast Tumor [5]*

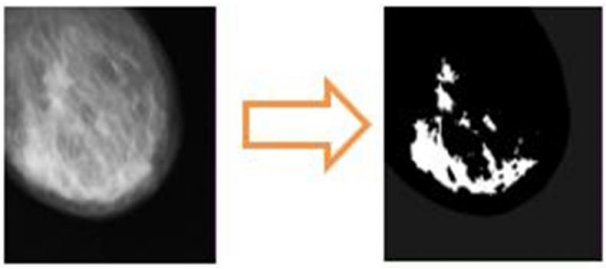

*Figure 4 The Example of Adaptive Median Filtering*

**Segmentation**

Multi-threshold is the process of dividing a grayscale picture into various isolated areas. These methods set at least one edge for a given image and divides it into specific areas of brightness related to a background, and multiple objects can be extended to multiple Otsu methods. Figure 5 below shows an example of a multi-level criterion. The variant between classes can be recreated as follows:

A point (x,y) belongs

- to an object class if T1 < f(x,y) T2
- to another object class if f(x,y) > T2
- to background if f(x,y)

$$g(x,y)\begin{cases} a & if\ f\ (x,y) > T_2 \\ b & if\ T1 \le f\ (x,y)\ \le T_2 \\ c & if\ f\ (x,y) \le T_1 \end{cases}$$

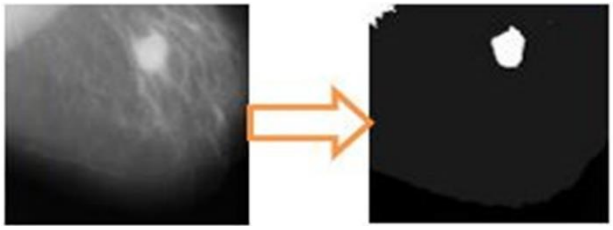

*Figure 5 The Result of Multi Level Thresholding*

**GLCM Feature Extraction**

Surface analysis using the Grayscale Cohesion Matrix (GLCM) calculates how often a specified spatial relationship between a specific value and a pair of pixels in an image is formed, generates a GLCM, and then separating the matrix performs a static measurement and figure

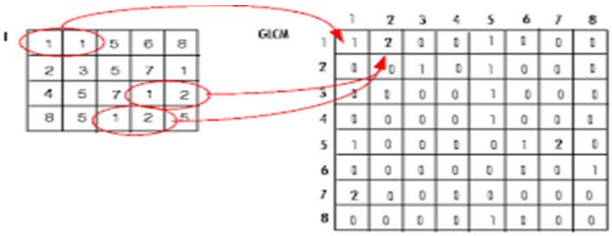


*Figure 6 The Example of GLCM Matrix [5]*

**Probabilistic Neural Network (PNN)**

Donald F. Specht (1990) offered a methodology to formulate the weighted-neighbor technique in the shape of a neural network. He called this a "Probabilistic Neural Network". It is based on the hypothesis of Bayesian categorization and the approximation of probability density function. It is essential to categorize the input vectors into one of the two categories in Bayesian the optimum behavior.

PNN is a response neural network used in categorization and pattern recognition programme. In this algorithm, the parent probability distribution (PDF) for each class is estimated by the Parka window and non-parameter functions. The PNN consists of three layers: the input layer, the radiating basal layer, and the scrambled layer. The radiative basale estimates the vector spacing between the input vector and the row weight vector in the weight matrix. These distances are non-linearly adjusted by the radial basis function, then the competing layers find the shortest distance among them, and the distance finds the training pattern closest to the pattern. The data entry network topology is shown in Figure 7. Symbols and notations are used as used in the Neural Network Design Document [6]. These symbols and notations are also used in the MATLAB Neural Network Toolbox. It will be displayed in Fig-7.

The probabilistic neural network (PNN) has a hierarchy of three nodes. The figure below shows the known PNN architecture of a class with K = 2, but can be extended to multiple K classes. The input (left) class consists of nodes. There is one for every n input attribute of the vector attribute. Each hidden node receives all input attribute vectors X because they are fanout nodes that are isolated

to the hidden level nodes. All (or intermediate) input nodes for each of these features. The hidden nodes are divided into groups. One group of each K class is described in the figure. Each hidden node of the class kth group includes a Gaussian function based on the class attribute vector. (For each simulated feature vector) has Gauss) Every Gauss in the class group fetches the value of the activity to the node with the same output level in that class, so it works under who has the output node and PNN.

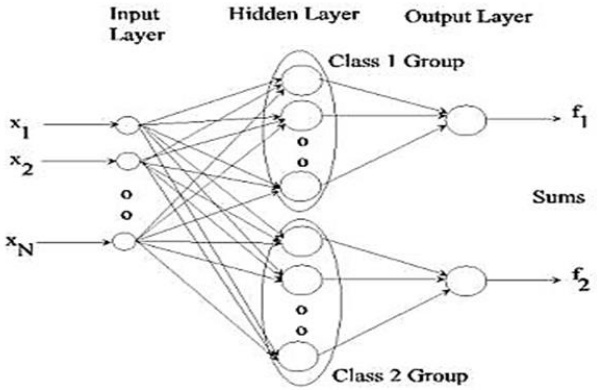


*Figure 7 Three-layer Architecture PNN Network [ 7]*

$g_1(x) = [1/ (2(\pi\sigma^2)^N] \exp \{-\|x-x^{(p)}\|^2/ (2\sigma^2)\}$ [1]

$g_2(y) = [1/ (2(\pi\sigma^2)^N] \exp \{-\|y-y^{(q)}\|^2/ (2\sigma^2)\}$ [2]

$f_1(x) = [1/ (2(\pi\sigma^2)^N] (1/P)\sum_{(p=1,P)} exp \{-||x - x(p)||2/ (2\sigma^2)\}$ [3]

$f_1(y) = [1/ (2(\pi\sigma^2)^N] (1/P)\sum_{(p=1,P)} exp \{-||x - x(p)||2/ (2\sigma^2)\}$ [4]

Let output node for class k= 1 or 2,

P = feature vector $\{x^{(p)}$: p= 1..., P} label as Class 1

Q= feature vector $\{y^{(q)}$: q= 1..., R} label as Class 2

Where $x^{(p)}$ and $y^{(q)}$ are N dimension of vector.$\sigma$ =value, which one -half the average distance.

**K-nearest neighbor (kNN)**

The nearest friend (KNN) is a supervised machine learning algorithm that can be used to solve classification and regression problems, while KNN is a real-life algorithm. People often influence those around us. Developing friends hosts the activities. Our parents also shaped our character. If users grow up with people who like sports, they will like sports more, of course, KNN can also and Easy to use and there is no presumption of advance information, but the estimated time is very long because you can find the distance between each data point.

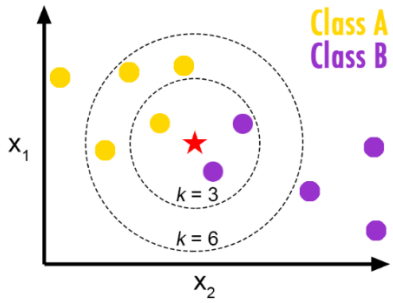


*Figure 8 Data displayed on a graph [8]*

The KNN estimates the distance between data points. For this, we use the easy Euclidean Distance equation as follow;

d(p,q)=d(q,p)=

$$\sqrt{(q1-p1)2+(q2-p2)2+\cdots+(qn-pn)2}$$

$$=\sqrt{\sum_{i=1}^{n}(qi-pi)2}$$

The upper equation is n-dimensional. Here, they feature machine learning. Data points with the shortest length from the experiment position are considered to belong to the same class.

**Experimental Results**

This system consists of six push buttons such as Test Image, Image Enhancement, Segmentation, Load Database, Training and Classifier. Test Image Button is clicked from the Test Folder, which the users want to test.Second is the test image to enhance more and more using the Adaptive Madian Filtering method.

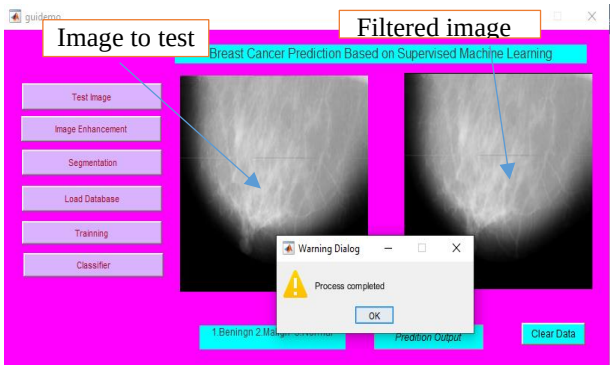


*Figure 9. Breast Cancer Testing Image and Filtered Image*

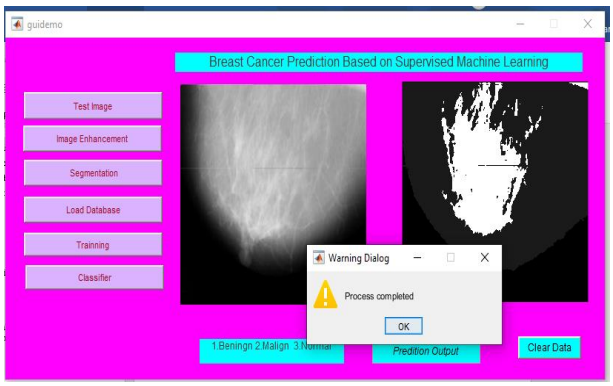


*Figure 10 The Segmented Image*

Third Button, which sigments it using Otsus' Thresholding.After this stage has been done, segmented image is featured using GLCM.The featured image is training with PNN and tests with Database. So, Database Button and Training Button is pressed.Finally, classifier is

clicked to distinguish 1.Benign, 2.Malgn ,and 3.Normal Breast Cancer.Bottom of the GUI shows Prediction Output Button to view cancer situation.Clear Button is used to clear the GUI data and to start next. The step by step results are shown as the following.

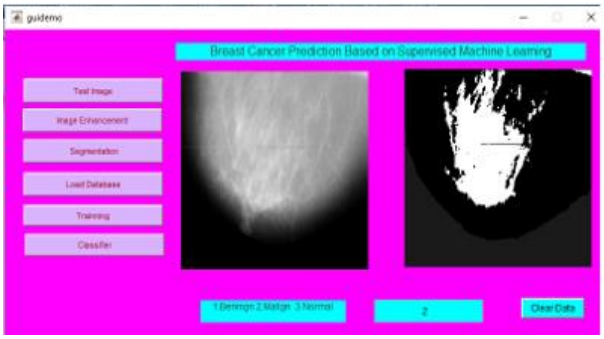


*Figure 11 The Prediction Result*

Accuracy value of Prediction=$\frac{Number\ of\ correct\ Prediction}{Total\ number\ of\ Prediction}$

For Binary classification is Accuracy value of Prediction =$\frac{TP+TN}{TP+TN+FP+FN}$

Where TP = True Positives, TN= True Negatives, FP = False Positives, and FN =False Negatives

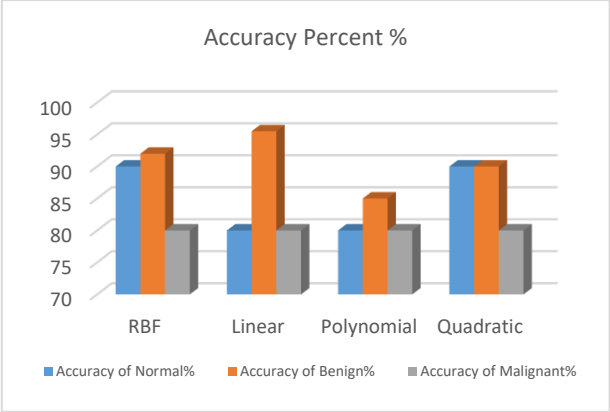


*Figure 12 Kernel Accuracy Percent*

## 4. CONCLUSION

Proper recognition of breast tumors is very important. This study suggests appropriate treatment. Appropriate system design for breast detection Malignant tumors is the standard procedure for breast cancer Diagnosis. The system classifies mammograms as follows: three types: moderate, gentle, cunning, and high rates. More than 95% systems to assist and support physicians or, a professional doctor will quickly make a mammography diagnosis.

It became clear that it was possible. Forecast the proximity of a tumor and at the same time Classify whether the existing is benign tumor or malignant. That is stress in patients with oncology. Otherwise, you will get an error.

## REFERENCES

[1] Bocchi L, Coppini G, Nori J and Valli G (2008), "Detection of single and clustered micro-calcifications in mammograms using fractals models and neural networks", Medical Engineering & Physics, Vol.26, pp.303-312. https://issuu.com/irjet/docs/irjet-v6i6619

[2] B.M.Gayathri. 1, C.P. Sumathi2 and T.Santhanam (2013), BREAST CANCER DIAGNOSIS USING MACHINE LEARNING ALGORITHMS –A SURVEY ,Published at International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013. http://airccse.org/journal/ijdps/papers/4313ijdps09.pdf

[3] Fadi Abu-Amara and Ikhlas Abdel-Qader, "Hybrid Mammogram Classification Using Rough Set and Fuzzy Classifier", International Journal of Biomedical Imaging, Volume 2009, July 2009. https://www.sciencepubco.com/index.php/ijet/article/view/18978

[4] www.Cancer Datasets from Research gates. https://sites.google.com/site/aacruzr/image-datasets

[5] Google Source, stage of breast cancer, https://www.google.com/search?q=stage+of+breast+cancer&sxsrf=ALeKk02dEwhnEz3kajgunLdK0TIgnpwH1Q:1614091561325&source=lnms&tbm=isch&sa=X&ved=2ahUKEwji-ZbRn4DvAhW

[6] Bouyahia S, Mbainaibeye J and Ellouze N (2009), "Wavelet based micro-calcifications detection in digitized mammograms, ICGST-GVIP Journal", Vol.8, No.5, pp.23-31. https://pt.slideshare.net/irjetjournal/irjet-comparison-of-breast-cancer-detection-using-probabilistic-neural-network-and-support-vector-machine

[7] Mohebali, Behshad; Tahmassebi, Amirhessam; Meyer-Baese, Anke; Gandomi, Amir H. (2020). Probabilistic neural networks: a brief overview of theory, implementation, and application. Elsevier. pp. 347–367. doi:10.1016/B978-0-12-816514-0.00014-X.

[8] Piryonesi S. Madeh; El-Diraby Tamer E. (2020-06-01). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transportation Engineering, Part B: Pavements. 146 (2): 04020022. doi:10.1061/JPEODX.0000175.