BREAST CANCER DIAGNOSIS USING IMAGE PROCESSING AND MACHINE LEARNING

Nityasree Upadhyay[1], Prerana Chakraborty[2], Subham Sadhukhan[3]

[1]Department of Computer Science & Engineering, Institute of Engineering & Management, Kolkata
u.nityasree95@gmail.com
[2.]Department of Computer Science & Engineering, Institute of Engineering & Management ,Kolkata
prerana.chaks@gmail.com
[3.] Department of Computer Science & Engineering, Institute of Engineering & Management, Kolkata
subhamsadhukhan95@gmail.com

Abstract: Breast Cancer has become common factor nowadays. In spite of the above fact, there have not been adequate facilities for its analysis within a short time. Therefore, we present here computerized method for cancer detection in its early stage within a very short time. Here we have used *Machine learning* to train a model using a dataset that we have downloaded from 'Kaggle'. After this we analyse a random digital image of a *Fine Needle Aspirate*(FNA) of breast tissue using *Image Processing* to analyse features of nuclei of the cells. We then apply the feature values to our trained model to find whether the tumour developed is *benign* or *malignant*.

Keywords: Machine Learning, Fine Needle Aspirate(FNA), Image Processing, Benign, Malignant

1. Introduction

Breast cancer is the second mainly common disease in women in India and this disease is increasing annually. The lack of awareness initiatives, structured viewing, and affordable treatment facilities continue to result in poor survival. Breast cancer is a disease that occurs when cells in breast tissue change (or mutate) and keep reproducing. These abnormal cells usually cluster together to form a tumour. A tumour is cancerous (or malignant) when these abnormal cells invade other parts of the breast or when they spread (or metastasize) to other areas of the

body through the bloodstream or lymphatic system, a network of vessels and nodes in the body that plays a role in fighting infection.

FNA is performed to collect a sample of cells or fluid from a cyst or solid mass, to allow the cells to be examined under a microscope. Fine needle aspirations may be performed on palpable lumps (lumps which can be felt), or impalpable lumps which have been detected on ultrasound or x-ray. A vacuum or negative pressure is created in the needle and with an in and out motion of the needle, the sample is taken.

\

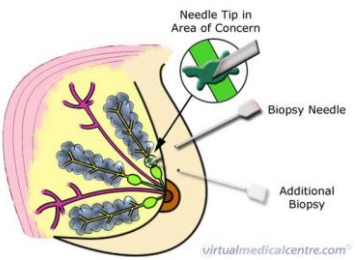

Figure 1: Fine Needle Aspiration (FNA)

1.1 Image Processing

Image Processing is a process in which various operations are performed on an image to get an enhanced image or to derive some information from the image. It is a type of signal processing in which the output may be an image or some useful information related to it.

1.1.1. Types of Image Processing

(i) Analog Image Processing: It is a task conducted on two dimensional analog signals by using various analog means . The most common example is the television image. The television signal is a voltage level which varies in amplitude to represent brightness through the image. By electrically varying the signal, the displayed image appearance is altered.

(ii) Digital Image Processing: The term digital image processing generally refers to processing of a two-dimensional picture by a digital computer. It is a subfield of signals and systems but mainly it focuses in image.

1.2 Machine Learning

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases.

1.2.1 Types of Machine Learning

(i) Supervised learning:

Regression models

Regression models (both linear and non-linear) are used for predicting a real value,If the independent variable is time, then you are forecasting future values, otherwise your model is predicting present but unknown values. Regression technique vary from Linear Regression to SVR and Random Forests Regression

Classification Model

Unlike regression where you predict a continuous number, we use classification to predict a category. There is a wide variety of classification applications from medicine to marketing. Classification models include linear models like Logistic Regression, SVM, and nonlinear ones like K-NN, Kernel SVM and Random Forests.

(ii)Unsupervised learning.

Unsupervised learning is used when we have unlabeled data. Here we cluster our data depending on various features of our dataset.

(iii)Reinforcement learning.

Reinforcement Learning is a branch of Machine Learning, also called Online Learning. It is used to solve interacting problems where the data observed up to time t is considered to decide which action to take at time t + 1.

In this paper, using machine learning techniques ,we developed models to predict the recurrence of breast cancer by analyzing our data. The next sections of this paper evaluate two classification models (KNN, SVM) to explain the methodology used to conduct the prediction, present experimental results, and the last part of the paper is the conclusion. To estimate validation of the models, accuracy, sensitivity, and specificity were used as criteria, and were compared.

2. Methoodology

2.1 Image Processing

Images of the FNA samples are collected under microscopic view . Then the various features of the cell nuclei are analised which are necessary for prediction of malignancy .The following steps are involved for obtaining the various features.

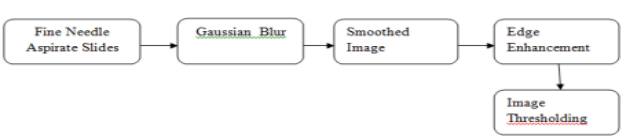

Figure 2Flowchart of processes imvolved to obtain the nuclei

2.1.1 Filtering Image

Filtering is a process for enhancing or modifying an image to emphasize certain features. Filtering involves smoothing, sharpening and edge enhancement.Here Gaussian Blur is used as blurring technique.

Gaussian Blur

Gaussian filtering is used to blur images and remove noise and detail. In one dimension, the Gaussian function is:

$$G(x) = (1/\sqrt[2]{2\pi\sigma^2})e^{\frac{-x^2}{2\sigma^2}}$$

Where σ is the standard deviation of the distribution. The distribution is assumed to have a mean of 0.

Edge Enhancement

Edge enhancement is an image_processing filter that enhances the edge contrast of an image or video in an attempt to improve its sharpness.It identifies sharp edge boundaries in the image and increases the contrast in the area immediately around the edge.Edge enhancement is applied to the blurred image to get the boundaries of the cells and the nuclei.Classical method of edge enhancement includes convolving the image with an operator (a 2-D filter), which is constructed to be sensitive to large gradients in the image .

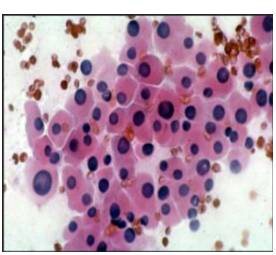

Figure 3Blurred Image

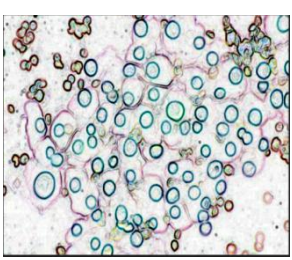

Figure 4:Edge Enhanced Image

2.1.2 Thresholding Image

Image Thresholding is a simple way of segmenting an image into foreground and background.It is a type of segmentation that distinguishes object. The simplest property that pixels in a region can share is intensity. So, a natural way to segment such regions is through thresholding, the separation of light and dark regions.

Thresholding creates binary images by turning all the pixels below the threshold to 0 and all the pixels above threshold to 1.

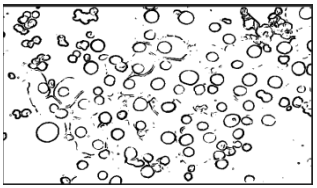

Figure 5 Thresholded Image

Now using MATLAB we find the circles from the image which represent the nuclei. Then we use various function like mean(),Std() , max().For finding the circles we have a function imfindcircles() which returns the coordinate of the centre and the radius length.

2.2 Machine Learning

Today's real-world databases are highly vulnerable to noisy, missing and inconsistent data due to their typically massive size and their likely origin from multiple, miscellaneous sources. Hence data preprocessing is required before applying any machine learning algorithm.

The breast cancer Wisconsin (diagnostic) data-set used here, downloaded from 'Kaggle', contains 31 columns and 569 entries. The data-set requires data-cleaning followed by feature extraction process, data transformation (data normalization, data binning) process. The unnecessary columns have been removed, followed by scaling the entire data-set to bring its mean to 0 and standard deviation to 1. The categorical values of the column 'diagnosis' has been changed to the binary values, that is 'malignant'(cancerous) tumour has been changed to '1' and 'benign'(non-cancerous) tumour has been changed to '0' respectively. Now, the dataset is ready for applying it to machine-learning algorithms, K-Nearest Neighbours and Support Vector Machine respectively.

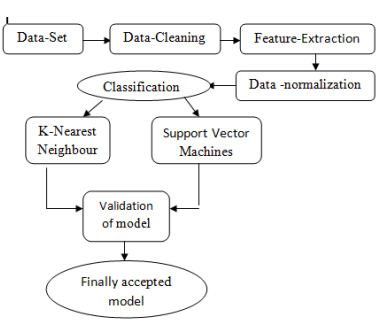

Figure 7 Machine Learning Process

2.2.1 K-Nearest Neighbours

Thisis supervised learning algorithm technique that dealswith‘claassification’problem. Let us assume that there is a dataset where all the data points belong to of the two classes; class A or class B; denoted by yellow and violet dots respectively in the figure below. Classification problem demands the allotting a new data-point denoted by a red star, to any of the two classes, class A or class B.

The algorithm in its initial steps requires choosing a suitable value of ‘K’. The next step of the algorithm states to choose ‘K’ number of nearest neighbours to the red data-point. Normally the nearest neighbours are chosen accoding to ascending order of the square of the distance of its neighbours to the required datapoint. An example illustrating the method for values K=3 and K=6 have been shown in the figure below. After this the number of neighbours for both classes A and B are calculated. The class having more number of neighbours to the assigned data-point is assigned to the data-point. For example, in the given figure for K=3, there are more number of neighbours belonging to class B, so the red data-point is assigned class B.

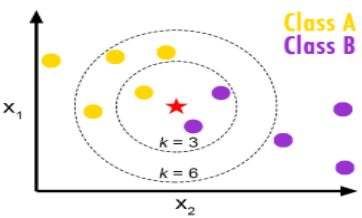

Figure 6 K-Nearest Neighbours

2.2.2 Support Vector Machine (SVM)

Support Vector Machine(SVM) is supervised learning algorithm, with associated learning algrorithms that analyze data and recognize patterns, used for classification and regression analysis. An SVM classifier algorithm builds a model that assigns a new data-point to one of the category or other, making it a non-probabilistic binary linear classifier. The new data-point to be classified is mapped to either of the categories, based on the side of the gap, created by the training data-points as wide enough possible, which side of the gap it falls. The hyperplane is drawn in such a way that maximizes the distance from each of the margins surrounding each of thecategories, known as support vectors.

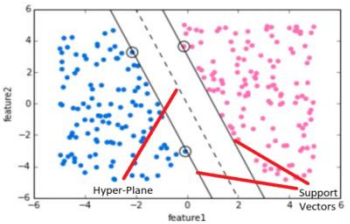

Figure 7: Hyperplanes and Support Vectors

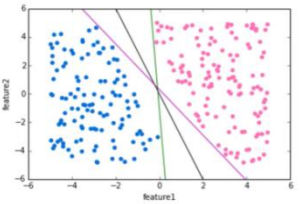

Figure 8: Wrong Approach for separation of data-

The feature values so obtained from image processing is applied to the model that has been prepared by machine learning to predict whether the breast tissue is malignant or benign.

3. Conclusion

Comparing to all other cancers, breast cancer is one of the major causes of death in women. So, the early detection of breast cancer is needed in reducing life losses. This early breast cancer cell detection can be predicted with the help of modern machine learning techniques.The efficiency of this entire model is quite high .Thus in future a software can be developed where the FNA slide images can be uploaded and it will automatically process the image to get the feature values and will automatically apply it to the model to predict the result. Thus it will lead to a very fast diagnosis of breast cancer .

References

1. https://in.mathworks.com/help/images/examples/detect-and-measure-circular-objects-in-an-image.html
2. http://www.drkmm.com/resources/INTRODUCTION_TO_IMAGE_PROCESSING_29aug06.pdf
3. https://in.mathworks.com/help/matlab/learn_matlab/array-indexing.html
4. http://www.wbi.net.au/fine-needle-aspiration-biopsy-fna/
5. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/version/2#
6. .https://www.myvmc.com/investigations/fine-needle-aspiration-biopsy-fna/
7. Shelly Gupta, Dharminder Kumar ,Anand Sharma, DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS, Shelly Gupta et al./ Indian Journal of Computer Science and Engineering (IJCSE)
8. https://en.wikipedia.org/wiki/Support_vector_machine
9. . Prognosis and Diagnosis of Breast Cancer Using Interactive Dashboard Through Big Data Analytics, Bio- Techonology an Indian Journal.
10. .Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Introduction to Statistical Learning with Applications in R, Springer