IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 67-71

Overview of Breast Cancer Detection using Machine Learning

DikshaRajpal[1], Anil Kumar[2], Sumita Mishra[3]
Department of Electronic and communication Engineering, Amity University Uttar Pradesh,Lucknow, India

Abstract: *Cancer is the second cause of death in the world. Breast Cancer is one of the leading causes of death in women. Breast Cancer is caused due to abnormal growth of cells in the breast tissue. Both women and men can suffer from breast cancer. There are several ways by which breast cancer can be detected the chief being mammograms. However, mammograms have a disadvantage that false results can be detected which can risk a patient's health. It becomes imperative to discover different techniques which are simpler to formulate and are able to operate with various datasets, are cheaper to design and are able to estimate results which have higher accuracy. This paper discusses a basic breast cancer classifier model which uses different machine learning and deep learning algorithms to predict results via images to determine whether the cancer is benign or malignant and future plans for designing better classifier models.*
Keywords:*Breast Cancer, Deep Learning*

I. Introduction

Breast Cancer generally is diagnosed in women however; even men can suffer from breast cancer. Breast cancer begins when there is excessive growth of cells in the breast tissue. These excessive amountsof cells form a tumor in the breast which can often be seen on a X-ray and can be felt as a lump.Breast Cancer is growing at an alarming rate.A decade ago, breast cancer moved ahead of oral cancer, in which India ranks no.1 globally, to become the country's most fatal disease. More than 115,000 new cases are diagnosed each year. The limited therapy centre's that monitor survival say 52% of breast cancer patients in India survive after five years, a 2010 study published in The Lancet found. This pales in correlation with the 89% survival rate in the US and the 82% rate in China.In the past decade the mortality rate due to breast cancer has shown a continuous decline when detected at an early stage. The Cancer Research Institute in UK has predicted that the survival rate of breast cancer is almost 100% if it is detected in the early stages and is as low as 15% if detected in the last stages.

Any breast cancer can be of two types:

- **Benign(non-cancerous)**:These cases are non-cancerous that is non-life threatening. However, sometimes they can turn into cancer.
- **Malignant(cancerous):**- A cancer is detected as malignant when cell growth has been in at a very large amount and has spread a t a very high speed to nearby cells and tissues**.** In this scenario the nucleus of the infected tissue or cell is quite large as compared to that of a non-infected cell, which decreases chances of survival of the patient.

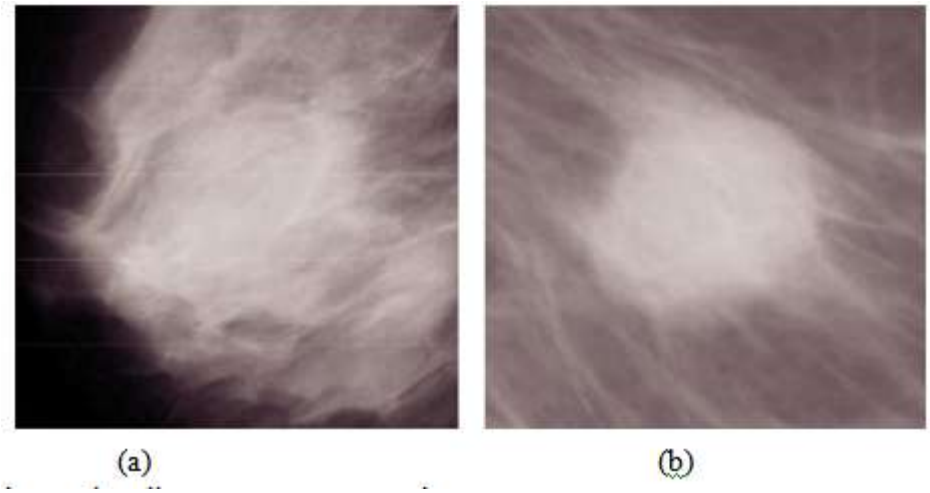

Figure 1: (a, b) show benign and malignant mammogram images

II. Breast Cancer Detection Methods

The current breast cancer detection methods include:

- **Breast exam:-** A breast exam is done by a doctor by checking both breasts and armpit for lymph nodes, making sure there are no lumps or abnormalities present in any place
- **Mammogram:-** X-ray of the breast is called as a mammogram. These are done for testing whether a patient has breast cancer or not. If the x-ray shows any results which are not normal a diagnostic mammogram is generally done for further investigations to gain more insight.
- **Breast ultrasound:-**When we use sound waves to form images of organs inside the body it is referred to as ultrasound**.** It also helps in help in finding out if the breast lump that was detected is just a solid mass or it is a cyst which is filled with fluid.
- **Removing a sample of breast cells for testing (biopsy):-** Biopsy is done with a combination of two things: a needle (special needle) plus any imaging test (eg. X-ray). The process consists of using the needle which is directed or guided by the imaging test. This needle gets a small amount of the infected tissue. The samples collected from biopsy are sent to a laboratory for scrutiny where it is determined whether the cells have cancer or not.
- **Breast magnetic resonance imaging (MRI):-** A breast magnetic resonance imaging machine creates an internal picture of the breast with the help of radio waves and a magnet. Before a breast MRI, the patient is injected with dye. Unlike other types of imaging tests, an MRI doesn't use radiation to create the images.

III. Limitation Of Detection Methods

All these detection methods have certain limits few of them are;

- All breasts are not similar looking in a mammogram, the woman's age plays a role in the breast's density and makes cancer more complicated or easier to see.
- Mammograms cannot detect certain cancers due to the density of the breast tissue and the site of the cancer.
- All cancers may not be visible on ultrasound. Much coagulation seen in a mammography is not visible on ultrasound.
- Most dubious findings on ultrasound that call for biopsy may not be cancers.
- The dye used for the MRI can cause allergies in the patient. It can also lead to an infection at the location of injection of the contrasting dye.
- MRI uses strong magnets which may cause problems for patients with implants for example: pace makers etc.

IV. Breast Cancer Detection Using Deep Learning

The doctors and physicians are highly dependent on the mammograms, MRI and ultrasounds to track the state of the cancer. To decrease this load on the doctors, research was done to form computer systems to which could help in breast cancer diagnosis. This could be achieved by machine learning, artificial intelligence and deep learning.

To detect whether an image is cancerous or not we need to design a general system having:

1. A breast database or dataset
2. Feature selector
3. Classifier
4. Performance or accuracy measurement
5. Output from classifier

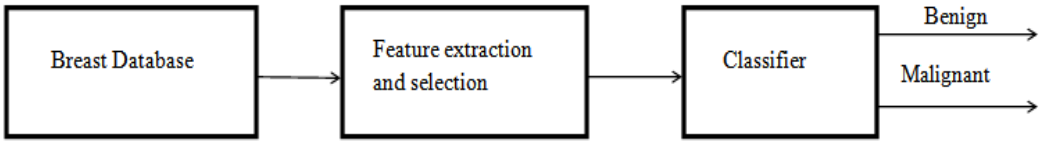

Figure 2:A basic breast classifier model

i. **Database:-**A database basically consists of a set of images on which the investigation or research can be done. Many organizations have introduced image sets which are available for researchers for further work.

ii. **Feature Selector:-**This step basically determines which features are to be extracted from the images in the dataset to determine whether the image is cancerous or not. Feature extraction is important because it helps in extracting useful information from the images which makes it possible to get results.

iii. **Classifier:-**Based on the requirement of the system the breast classifier can be of following types:

- Supervised
- Unsupervised
- Semi-supervised

Below are few classifiers used in breast cancer classification:

- **Stochastic GradientDescent:-**Gradient is basically the measure of variation in the output function corresponding to the slightest variation in the input value. Gradient can also be referred as the derivative of a function. It informs us of the slope or incline of the cost function.

Stochastic Gradient Descent also referred to as SGD calculates the gradient G for every revised value using "A single training data point" which is decided randomly. The aim is that the new calculated gradient is an approximate stochastic value as compared to the Gradient, which was calculated previously for the entire training data. Therefore, SGD is a much faster way of calculating the gradient and hence, is widely used.

- **K-Means Clustering:-**K means clustering comes under unsupervised learning. In this technique the data provided is divided in to families or groups (also called as clusters). If there is n amount of data it is divided in to k clusters with each data or observation belongs to a cluster with the corresponding nearest mean. This step has to be done iteratively till we have convergence.K-means clustering algorithm helps us in minimizing the square error function.

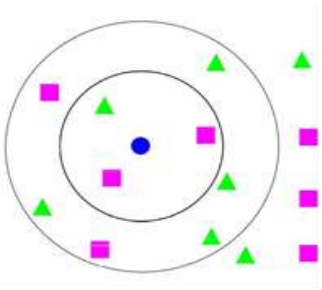

Figure 3:k-Nearest neighbor for breast cancer diagnosis. Blue circle means the test sample;green triangle means malignant BC and Pink Square means benign BC

- **Support Vector Machine (SVM):-**SVM looks for a line or hyper-plane that separates out classes. In real world problems SVM is a critical classifier that determines a hyper plane to segregate data. In case of non–linear problems, SVM has a technique called as kernel trick. These are functions that take low-dimensional input space and convert it to high-dimensional space that is they convert non separable problems into separable ones**.**

iv. **Measuring Performance:-**Measurement of performance is done by the following:

Confusion Matrix:-This is a matrix which is used to define how well the classification model is performing for a data set values for which we already have the results. It is basically a sum-up of estimated results of a classification problem.The confusion matrix shows the results on how a classification prototype is obscured while making estimations.It gives us awareness both about the errors being produced and also about the kind of errors created by the classifier.

Definition of the Terms:

- **True Positive (TP):-** When the known value is positive and the estimation made by the model is also positive.
- **False Negative (FN):-** When the known value is positive and the estimation made by the classifier is negative.

- **True Negative (TN):-** When the known value is negative and the result estimated by the model is also negative.
- **False Positive (FP):-** When the know value is negative and the result estimated by the model is positive.

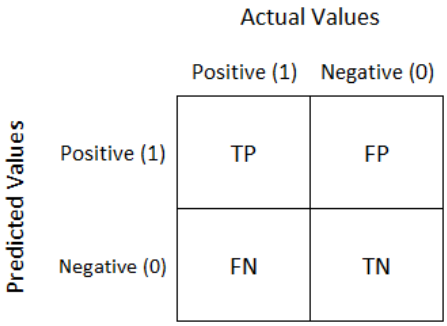

Figure 4: Confusion Matrix

Classification Rate/ Accuracy:-Accuracy or classification rate is given by the formula:

$$\textbf{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Accuracy Disadvantages:** Accuracy supposes equal value of cost functions for all errors (two here). A 99 % value of accuracy may be considered terrible, poor, average, good or excellent subject to the problem.

Recall:-Recall is defined by dividing the entire number of correctly classified positive examples with the entire number of positive examples. A High Recall value symbolises the class is correctly recognized (that is small number of False Negatives).

$$\textbf{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision:-Precision is defined by dividing the entire number of correctly classified true positive examples with the total number true positives and false positives. A High Precision value is an indication of an example detected as positive by the classifier of being indeed positive, which means that we have a very small amount of false positive.

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **High recall value and low precision value:** This signifies that majority of the positive examples are correctly identified by the model which shows that we have low false negative. However; a high number of false positives have been detected.
- **Low recall value and high precision value:** This is an indication that lots of values which were positive were missed by the classifier. This means that a high number of false negative have been identified. However, those that were estimated as positive by the classifier are indeed positive i.e. low False Positive.

F-measure:-F-measure is a quantity or value that is representation for both precision and recall, which instead of arithmetic mean uses harmonic mean. This is done because extreme values are punished more by harmonic mean.

The value of F-Measure is always nearer to both the smaller values of precision and recall.

$$\textbf{F} - \textbf{Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

V. Conclusion

This paper gives a general idea about the concept of breast cancer and how this cancer is affecting the population worldwide. It also explains the current breast cancer diagnosis methods and their limitations. We discuss the need for machine learning techniques for breast cancer prediction as it reduces the burden on the doctors and physicians.Machine learning and deep learning techniques have shown a substantial amount of ability to improve estimation and classification accuracy. We discuss various classification models being used for a breast cancer classifier for the datasets available in the market.Although many algorithms have generated high level of accuracy there is still a need for different algorithms for breast cancer diagnostics to achieve better results and generate higher level of accuracies by using different mechanisms.

References

[1]. X. Liu, J. Shi, S. Zhou, and M. Lu, "An iterated Laplacian based semi-supervised dimensionality reduction for classification of breast cancer on ultrasound images," in Proceedings of the 201436th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '14), pp. 4679–4682, USA,August 2014.

[2]. Y. Qiu, Y. Wang, S. Yan et al., "An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology," in Proceedings of theMedical Imaging 2016: Computer-Aided Diagnosis, SPIE. Digital Library, San Diego, California, USA, March 2016.

[3]. M. Taheri, G. Hamer, S. H. Son, and S. Y. Shin, "Enhanced breast cancer classification with automatic thresholding using SVM and Harris corner detection," in Proceedings of the InternationalConference on Research in Adaptive and Convergent Systems (RACS '16), pp. 56–60, ACM, Odense, Denmark, October 2016.

[4]. N. C. Mhala and S. H. Bhandari, "Improved approach towards classification of histopathology images using bag-of-features," in Proceedings of the 2016 International Conference on Signal andInformation Processing (IConSIP '16), IEEE, Vishnupuri, India,October 2016.

[5]. N. Zemmal, N. Azizi, M. Sellami, and N. Dey, "Automated classification of mammographic abnormalities using transduc-tive semi supervised learning algorithm," in Proceedings of theMediterranean Conference on Information & Communication Technologies 2015, A. El Oualkadi, F. Choubani, and A. ElMoussati, Eds., pp. 657–662, Springer International Publishing, Cham, 2016.

[6]. N. Zemmal, N. Azizi, N. Dey, and M. Sellami, "Adaptative S3VM semi supervised learning with features cooperation for breast cancer classification," Journal of Medical Imaging and HealthInformatics, vol. 6, no. 4, pp. 957–967, 2016.

[7]. H. Rezaeilouyeh, A.Mollahosseini, and M. H.Mahoor, "Microscopic medical image classification framework via deep learningand shearlet transform," Journal of Medical Imaging, vol. 3, no. 4, Article ID044501, 2016.

[8]. D. O. Tambasco Bruno, M. Z. Do Nascimento, R. P. Ramos, V.R. Batista, L. A. Neves, and A. S. Martins, "LBP operators on curvelet coefficients as an algorithmto describe texture in breast cancer tissues," Expert Systems with Applications, vol. 55, pp. 29–340, 2016.

[9]. H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis,"Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.

[10]. A. I. Pritom, M. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," in Proceedings of the 19thInternational Conference on Computer and Information Technology(ICCIT '16), pp. 310–314, December 2016.

[11]. L. A. Salazar-Licea, J. C. Pedraza-Ortega, A. Pastrana-Palma, and M. A. Aceves-Fernandez, "Location of mammograms ROI's and reduction of false-positive," Computer Methods and Pro-grams in Biomedicine,vol.143,pp.97–111,2017.