

## Evaluation Of Machine Learning Classifiers In Breast Cancer Diagnosis

S. Leena Nesamani<sup>1</sup>, S. Nirmala Sugirtha Rajini<sup>2</sup>,

<sup>1</sup>Research Scholar, Department of Computer Applications, Dr.M.G.R.Educational and Research Institute, Maduravoyal, Chennai

<sup>2</sup> Professor, Department of Computer Applications, Dr.M.G.R.Educational and Research Institute, Maduravoyal, Chennai

**Article History:** Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

**Abstract:**Breast Cancer is one among the most deadly diseases that threaten women. It affects women in general but men are also not exceptions to it. Most breast cancers end up fatal except for a few cases. Early diagnosis of the disease helps in successful treatment and cure. Computer Aided Detection (CADe) or Computer Aided Diagnosis (CADx) System helps physicians to take prompt decisions in the field of medical imaging. CAD systems are aimed at identifying the abnormalities at the earliest in the human body which a human professional may fail to detect. Machine learning techniques in the field of medical imaging are increasingly being used in the accurate diagnosis of breast cancer. Machine learning classifiers such as Support Vector Machine and Neural Network are examined in this paper. Breast mammogram images, both normal and pathological images were used in this experiment. The machine learning classifiers were employed to identify the given image as either Benign or Malignant. Performance of both the classifiers was recorded and it was observed that the Neural Network classifier excelled in the diagnosis with 98% accuracy than the Support Vector Machine classifier.

**Keywords:**Breast Cancer, Malignant, Benign, Diagnosis, Prediction, Machine Learning, Classifiers, Accuracy, Support Vector Machine, Neural Network.

### 1. Introduction

The term carcinoma refers to cancer that may be found either in the skin or in any tissue cells found in the internal organs of the human body. Cancers are caused by the mutations of genes that regulate the growth of the cell. The mutation process divides and multiplies the cells in an uncontrolled manner. Adenocarcinoma is a specific type of carcinoma that is formed in the breast region. Most of the breast cancers are tumors that begin in the epithelial cells and grow into the breast tissues. Breast cancer can show up a variety of symptoms, the first being a new lump in the breast but not all lumps are cancer. The other symptoms include pain, swelling, nipple discharge, inverted nipple, lump or swelling under the arm, etc.

Breast cancers are common among women and are a very rare in men. They are not common in men due to the less development of breast cells in male. The centre for disease control and prevention has given the statistics that one out of every hundred breast cancers identified is identified in man. Most of them do not check for signs of lumps as a result male breast cancers are diagnosed in a much later stage.

Stage zero indicates that there are no signs of cancer or abnormal cells in the breast. Stage one indicates that the breast tumor is less than 2cm and has not spread outside the breast. Stage two indicates that the tumor is 2-5cm in size and has spread to the lymph node near the breast. Stage three indicates that the tumor in size 2-5cm is found in the axillary lymph node or in the lymph node near the breastbone. Stage four is usually known as the last stage where the cancer has spread to other distant organs of the body. Survey shows that 600,000 women were dead in 2018 due to breast cancer. The survival rate of cancer is only 50% in India and in other developing countries. Early diagnosis of breast cancer can increase the survival rate.

CAD systems are aimed at identifying the abnormalities at the earliest in the human body which a human professional may fail to detect. Mammograph is an imaging technology used to identify tiny lumps, identify the distortions in the architecture and to predict a mass type as either benign or malignant. Machine learning techniques when applied to CAD systems produce tremendous results in disease diagnosis. The most popular use off of ML is in the classification of the given objects into different classes. The objects in the CAD systems cannot be represented with simple equations which demand the learning from examples paradigm for pattern recognition problems at which ML techniques are proven to be the best.

## 2. Literature Review

E. A. Bayrak et al[1] has used ANN (Artificial Neural Network) and SVM (Support Vector Machine) to predict the presence of cancer. This machine learning concepts are employed on the WBC (Wisconsin Breast Cancer) dataset. The performances of the classifiers measure on various metrics such as precision, recall, ROC and exactness. Final output shows that SVM technique produces better result than other classifier. The models are implemented in the WEKA software.

S. Leena et al[2] developed a new cancer detection model using a combination of Artificial Neural Network and Multi Level Support Vector Machine classifiers for the finding whether a given mammogram images contain benign or malignant type of cancer. The authors concentrated on working on a reduced feature set. Twelve important features were selected for the study using the GLCM feature extraction technique. The performance of the new model proved to be higher than the other existing models which employed a larger feature set.

Early diagnosis is essential to improve the survival rate of breast cancer. Bektaş et al[3] in their study of machine learning algorithms in the diagnosis and classification of breast cancer compared various ML algorithms. Attribute selection methods were used to identify the active genes that are necessary to classify the input as either benign or malignant. The method showed a success rate of 90.72% accuracy making use of 139 active features.

S. Leena et al[4] Ensemble models prove to increase the accuracy of model prediction. Ensemble model uses a collection of classifiers to classify the given data. It then combines the result of the individual classifiers to obtain the final prediction. It is proven that ensemble machines perform better than individual classifiers. The authors have evaluated different ensemble machines in the prediction of breast cancer and tried to identify a new combination of classifiers that will perform better than the existing models.

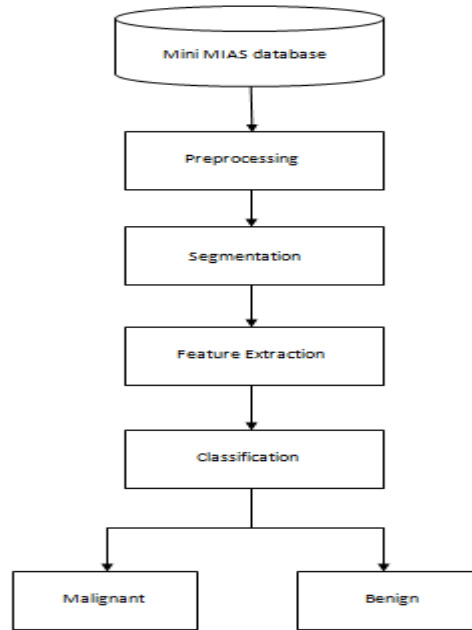
Hiba et al[5] Classification and data mining techniques are effective in identifying breast cancer. In this paper various machine learning techniques have been explored. Methods like SVM, DT(c4.5), NB, KNN have been used on the Wisconsin dataset. The algorithms were evaluated on the basis of their accuracy, sensitivity, precision and specificity. The experiments showed that among all the classifiers that were examined, SVM showed a greater performance of 97.13% with minimal error rates.

Gowri et al[6] Medical data are enormous and cannot be handled easily and effectively without the help of classification techniques. One of the main challenges faced today is to identify an accurate classifier in machine learning. Three methods namely, Decision Tree, SVM, Naïve Bayes were implemented in the study on a real time dataset. It was identified that the Decision Tree algorithm surpassed all other algorithms and proved to be much faster, accurate and more efficient with 91%. The authors have concluded that Decision Tree algorithms are best for handling medical dataset.

Data mining techniques play a vital role in the diagnosis of breast cancer at an early stage. In this paper, Siham et al[7], have proposed propose an approach that enhances the accuracy and thereby improves the performance of three classifiers: Decision Tree, Naïve Bayes, and Sequential Minimal Optimization. We also validate and compare the classifiers on Two benchmark datasets, the Wisconsin Breast Cancer and the Breast Cancer dataset were used to compare the performance of the classifiers. Problems like imbalanced data are addressed in this work by re-sampling the dataset. 10 fold cross validation was used for evaluation. The performances of the classifiers were evaluated in terms of confusion matrix, standard deviation, ROC curve and accuracy. SMO out performed in the WBC dataset and J48 performed well in the Breast Cancer dataset L.G Ahmad[8] analyzed three different ML algorithms which include decision tree DT, ANN, and SVM and identified that SVM showed higher accuracy techniques. This work used the Iranian centre for breast cancer (ICBC) data set

## 3. Proposed System

For the purpose of this study the mini-MIAS dataset[10] was used. The dataset contains 322 single-breast mammogram images which are of size 1024x1024. The primary task here is to classify the breast mammogram images into either benign or malignant class. This work uses two Machine Learning techniques, namely Artificial Neural Network and Support Vector Machine to classify the mammogram images. Finally the performance of both the techniques is compared to identify the best method. Finally assess the performance of the classifiers using accuracy and precision matrices.



**Figure 1 Proposed System Flow Diagram**

Preprocessing is the first step in image processing, where the image is cleaned by removing the unwanted artifacts. The quality of the image is improved using a median filter. The ROI is segmented using the Region growing algorithm. The part which appears to be abnormal is segmented for further processing. The textural features are then selected using GLCM algorithm. The extracted features are fed into two different classifiers and the performances of both the classifiers were studied individually.

Support Vector Machine classifier is commonly used in binary classification problem with linear data. The primary goal of SVM is to identify a hyper-plane. The hyper-plane is represented as the following equation 1.

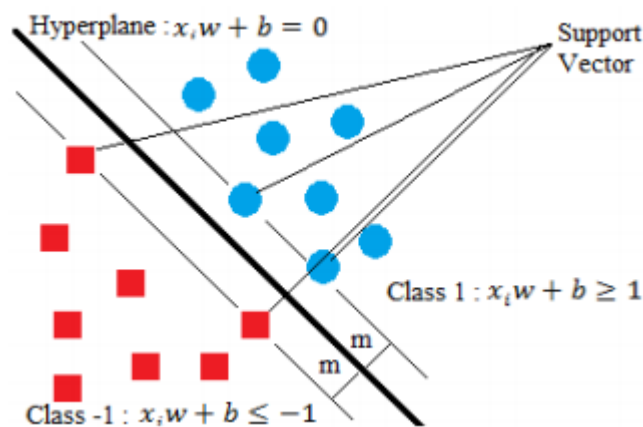
$$f(x)=wx_i + b = 0 \tag{1}$$

Here,  $x_i$  represents the individual data points,  $w$  is the weights assigned to the individual data points and  $b$  is the bias.

Optimal value of margin is got from distance of hyper-plane and the support vector. Classification is represented as

$$\min \frac{1}{2} \|w\|^2 \tag{2}$$

According to T. Nadira, et.al., 2018 The following figure 2 shows the pictorial representation of SVM classifier[16].



**Figure 2 SVM Classifier**

ANN ( Artificial Neural Network) is a mathematical learning model . It is inspired by the functioning of the biological neurons in the brain. It consists of three layers of neurons namely, the input layer which receives the input the model. The next is the hidden layer, where the actual processing takes place. Each neuron in the hidden layer is connected to all other neuron in the next layer forming a network of neurons.

$$f_u(x) = \begin{cases} 1.0, & x \geq x_{th} \\ 0.0, & x < x_{th} \end{cases}; \quad f_b(x) = \begin{cases} 1.0, & x \geq x_{th} \\ -1.0, & x < x_{th} \end{cases}$$

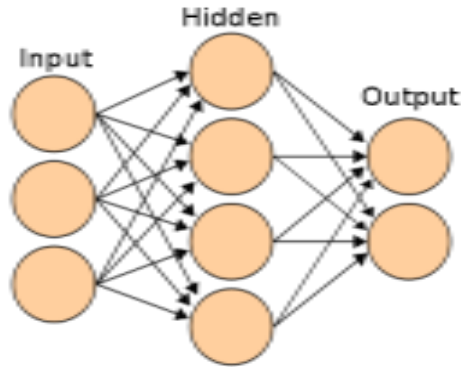


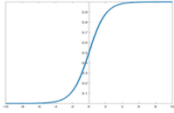
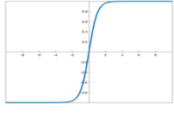
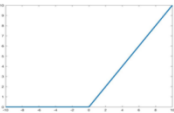
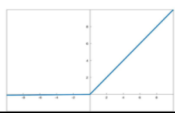
Figure 3 Common Form of NN

The final layer is the output layer. The value of an output neuron will be 1 if it receives a value from the network and 0 otherwise. The activation function decides whether to transfer a value from a neuron to the next level at any point of time. These activation functions could be either linear or non-linear functions. These are used to represent unipolar or bipolar models of neuron. The unipolar and bipolar activation functions are shown in equation (3).

(3)

The following table summarizes the most commonly used activation functions.

Table I. Activation functions

Activation function	Equation	Graph
Sigmoid	$S(x) = \frac{1}{1 + e^{-x}}$	
Tanh	$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	
ReLU	$RELU(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases}$	
Leaky ReLU	$f(x) = \begin{cases} x & \text{if } x > 0 \\ 0.01x & \text{otherwise} \end{cases}$	

#### 4. Result and discussion

Prediction and diagnosis of this disease in earlier stage is significant for better life. Various computing techniques are used to diagnosis the cancer. Machine learning concepts are offering various tools to diagnosis cancer in breast. Here SVM and NN classifiers are used to classify the mammogram images. The classifiers are implemented and tested by using Wisconsin dataset. Finally the performance level of the current work compared

in terms of accuracy and precision value. Metrics like accuracy, precision, recall and f1-score are measured to evaluate the performance of the model.

The misclassifications, made by the model could be measured using true positives (TP), true negatives (TN), false positive (FP) and false negative (FN).

Precision is a metric that is used to measure the number of correct positive predictions made by the model.

It is measured as the ratio of correctly predicted positive samples divided by the total number of positive samples that were predicted by the mode.

In classification problem with two classes, precision is calculated as:

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}$$

The value of 0.0 implies no precision and a value of 1.0 implies perfect precision. Recall is a metric that calculates the number of correct positive predictions out of all the positive predictions that were made.

In a classification problem with two classes, recall is estimated as the number of true positives values divided by the sum of true positives and false negatives values. Recall is given by the formula:

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}$$

The result lies between 0.0 and 1.0 which represents no recall or perfect recall. The accuracy is a single measure that is used to summarize model performance in one line. F-score is a metric that combines both precision and recall into a single measure. The f-score is calculated as follows:

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

This is also called as the F1-Score. The higher the F1-Score, the better is the model.

The SVM classifier provided the following results. The accuracy achieved by the SVM classifier was 95%.

	precision	recall	f1-score	support
malignant	0.96	0.92	0.94	53
benign	0.96	0.98	0.97	90
avg / total	0.96	0.96	0.96	143

Prediction Accuracy: 0.958042

Figure4 Performance of SVM Classifier

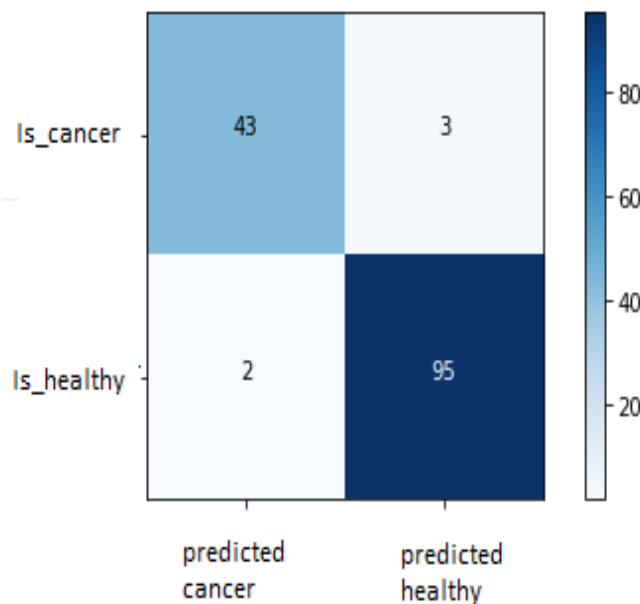


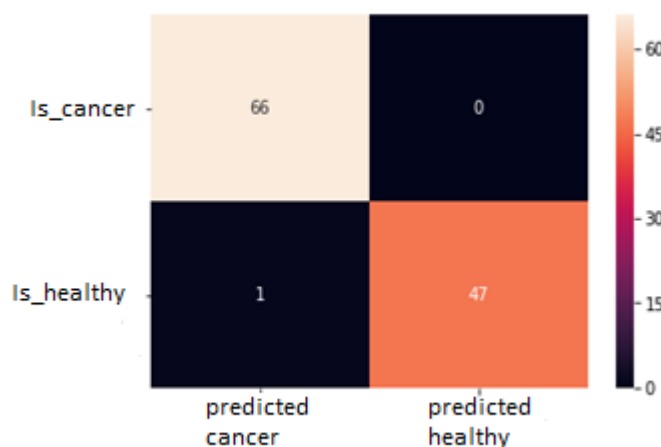
Figure 5 Confusion matrix of SVM classifier

The confusion matrix provides an exact distribution of the number of samples that were incorrectly classified under each category.

	precision	recall	f1-score	support
0.0	1.00	0.98	0.99	48
1.0	0.99	1.00	0.99	66
avg / total	0.99	0.99	0.99	114

**Figure 6 Performance of NN Classifier**

The prediction got better with the NN classifier. It can be seen from the confusion matrix heat map that only one sample has been predicted wrongly as cancer instead of a healthy image. Accuracy achieved was 98%. Compared with SVM technique, NN approach produces better result.



**Figure 7: Confusion matrix of NN classifier**

## 5. Conclusion

Breast Cancer is a very scary disease which is common in females. The survival rate of cancer is low due to the late identification of the disease using traditional methods of diagnosis. Modern CAD system enables the early diagnosis of the disease which tends to increase the survival rate. Machine learning classifiers assists healthcare professionals to diagnosis breast cancer at the earlier stage. This proposed work uses SVM and ANN classifiers to predict the disease. Mammogram images are provided as the input to the current prediction system. Unwanted data are removed by using preprocessing concepts and images are segmented by using cropping or thresholding techniques. Finally the images are classified by using the SVM and NN classifier. The NN classifier produces better result in terms of accuracy and precision value. Optimization techniques could be adopted to get higher results in future.

## References

- E. A. Bayrak, P. Kırıcı & T. Ensari(2019), "Comparison of Machine Learning Methods for Breast Cancer Diagnosis,"*IEEE Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1-3.
- LeenaNesamani.S, NirmalaSugirthaRajini.S, "Evaluation of Ensemble Machines in Breast Cancer Prediction", *Advances in Parallel Computing, Vol. 37 Intelligent Systems and Computer Technology*, pp391-395, ISBN978-1-64368-102-3 (print) | 978-1-64368-103-0 (online),© 2020 The authors and IOS Press
- B. Bektaş & S. Babur(2016), "Machine learning based performance development for diagnosis of breast cancer,"*Medical Technologies National Congress (TIPTEKNO)*, pp. 1-4.

- D. Leena Nesamani.S, Nirmala Sugirtha Rajini.S, “Evaluation of Ensemble Machines in Breast Cancer Prediction”, *Advances in Parallel Computing, Vol. 37 Intelligent Systems and Computer Technology*, pp391-395, ISBN978-1-64368-102-3 (print) | 978-1-64368-103-0 (online),© 2020 The authors and IOS Press.
- E. Hiba Asria ,Hajar Mousannifb ,Hassan Al Moatassime c ,Thomas Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis”, *The 6th International Symposium on Frontiers in Ambient and Mobile Systems, Procedia Computer Science 83 ( 2016 ) 1064 – 1069*
- F. T.Gowri, Dr.S.Geetha, “Breast Cancer Prediction using Supervised Machine Learning Algorithms”, *International Research Journal of Engineering and Technology*, Volume: 07 Issue: 08 | Aug 2020.
- G. Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, Gunter Saak,”Analysis of Breast Cancer Detection Using Different Machine Learning Techniques”, © Springer Nature Singapore Pte Ltd. 2020 Y. Tan et al. (Eds.): DMBD 2020, CCIS 1234, pp. 108–117, 2020.
- H. L.G.Ahmad, A.T.Eshlagh, A.Poorebrahimi, M.Ebrahimi and A.R.Razavi ”Using three machine learning techniques for predicting breast cancer recurrence”, *Journal of Health and medical informatics*, Vol.4, No.2, pp.13, April 2013.
- I. .Umadevi, S. NirmalaSugirthaRajini, A. Punitha and VijiVinod(2020), “Performance Evaluation Of Machine Learning Algorithms In DimensionalityReduction”, *International Journal of Advanced Science and Technology*, Vol. 29, No. 9s, pp. 3845-3853.
- J. The mini-MIAS database of mammograms. <http://peipa.essex.ac.uk/info/mias.html>,2018
- K. B. Bavani, S. NirmalaSugirthaRajini, M.S. Josephine, & V. Prasannakumari (2019), “Heart Disease Prediction System based on Decision Tree Classifier”, *Jour of Adv Research in Dynamical & Control Systems*, Vol. 11, 10-Special Issue.
- L. S.Umadevi, S. NirmalaSugirthaRajini,A. Punitha&VijiVinod, (2020), “ Dimensionality Reduction in Machine Learning Technique using Principal Component Analysis”, *Test Engineering and Management*, January - February 2020 ISSN: 0193 - 4120 Page No. 14546 - 14552 .
- M. Kiruthika , C & NirmalaSugirthaRajini,S(2014), “An Ill-identified Classification to Predict Cardiac Disease Using Data Clustering”, *International Journal of Data Mining Techniques and Applications*, vol. 03,pp. 321-324, ISSN:2278-2419.
- N. Abirami .P & NirmalaSugirthaRajini.S(2021), “Deep Learning Approaches for Disease Prediction”, 7<sup>th</sup> International Conference on Electrical Energy Systems”, Sri SivasubraminaNadar College of Engineering Kalavakkam, Chennai.
- O. Bavani B, NirmalaSugirthaRajini, S(2020), “Prediction of Cronory Syndrome Using Machine Learning Algorithms”, 5th International Conference on Digital Transformation Industry 4.0 and future Business Organised by GL Bajaj Insititute of Management and Research Greater Noida on November 21, 2020
- P. T. Nadira and Z. Rustam(2018), “Classification of cancer data using support vector machines with features selection method based on global artificial bee colony”, *AIP Conference Proceedings*.