Using Support Vector Machine Detection of Breast Cancer in Early stage

[1]Lakshman K, [2]Siddharth B. Dabhade, [3]Sachin Deshmukh, [4]Ranjan Maheshwari, [5]Karan Dabhade
[1]NIELIT Aurangabad, Dr. B. A. M. U. Campus, Aurangabad (MS), India.
[2]MGM's, Dr. G. Y. Pathrikar College of Computer Science and IT, Aurangabad (MS), India.
[3]UDCSIT,Dr. Babasaheb Ambedkar Marathwada University, Aurangabad(MS), India.
[4]RTU, Kota (RJ), India. [1]lakshman.korra@gmail.com, [2]dabhade.siddharth@gmail.com

***Abstract*— The Breast Cancer is disease which tremendously increased in women's nowadays. Mammography is technique of low-powered X-ray diagnosis approach for detection and diagnosis of cancer diseases at early stage. The proposed system shows the solution of two problems. First shows to detect tumors as suspicious regions with a weak contrast to their background and second shows way to extract features which categorize tumors. Hence this classification can be done with SVM, a great method of statistical learning has made significant achievement in various field. Discovered in the early 90's, which led to an interest in machine learning? Here the different types of tumor like Benign, Malignant, or Normal image are classified using the SVM classifier. This techniques shows how easily we can detect region of tumor is present in mammogram images with more than 80% of accuracy rates for linear classification using SVM. The 10-fold cross validation to get an accurate outcome is been used by proposed system. The Wisconsin breast cancer diagnosis data set is referred from UCI machine learning repository. The considering accuracy, sensitivity, specificity, false discovery rate, false omission rate and Matthews's correlation coefficient is appraised in the proposed system. This Provides good result for both training and testing phase. The techniques also shows accuracy of 98.57% and 97.14% by use of Support Vector Machine and K-Nearest Neighbors**

Keyword: ***Breast Cancer, Mammography, preprocessing, Segmentation, SVM classifier,*** **K-nearest etc**

I. INTRODUCTION

More than 6 million of deaths occur due to breast cancer Probability of Deaths with Breast cancer is higher to more than 20 million in 2050. In females Breast cancer is one of the common types happening cancer diseases and there are various techniques like identification of breast cancer is Mammography and Magnetic Resonance Imaging (MRI) based technique. As Breast tumors and masses usually appear in the form of dense regions in mammograms. A typical benign mass has a round, smooth and well circumscribed boundary; on the other hand, a malignant tumor usually has a speculated, rough, and blurry boundary. In biopsy testing [2], the biopsy is occupied from the tissues of the breast the accuracy level is high but its painful and surgery is needed. Hence most of the patients avoid such test for breast cancer detection. Mammogram [3] is the now a days most used technique for the detection of breast cancer which provides the 2D projection images of the breast. There are two kind of Mammogram technique The First technique is mammogram consists with X-ray to check the abnormality in breast this type of tumor is categories into three parts as Normal, benign and malignant.

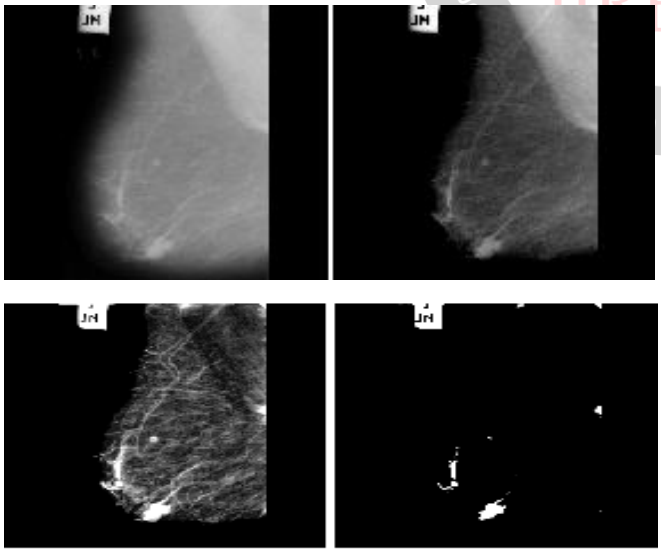
The normal type is without any cancer cells. The tumor is displayed but not of cancerous cells are benign type and a tumor with cancerous Cell is called as malignant type. Several techniques have been already come for analyzing mammogram images

II. LITERATURE SURVEY

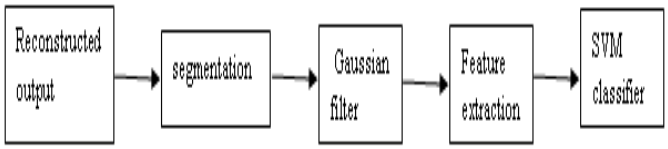
These are few early research made on detection of tumor of Brest cancer. S.Thamarai Selvi et.al introduced an algorithm for tumor detection. The work focuses on two things: detection of suspicious region of tumor with lower contrast and method to extract features from these regions to categorize tumors" Jawad Nagi et.alexplored mammogram segmentation based on automatic technique. This makes use of seeded region growing (SRG) along with morphological pre-processing steps. The segmentation process helps in improving the detection and diagnosis of cancer. Satish Saini and Ritu Vijay designed a system for detecting breast cancer using ANN. Its effect is checked for different number of layers and chooses numbers of layer for optimum result. The prime objective of this research was to increase the diagnostic accuracy of the detection of breast cancer malignancy in Computer

 DOI : 10.35291/2454-9150.2020.0465

Aided Diagnosis (CAD) systems by developing image processing algorithms and to categorize the women into different risk groups. The evaluation of SVM classifier has been considered. Initially, tumors have been detected from mammograms with the aid of morphological processing of breast images. Then classification is done by SVM classifier using the most dominant features namely GLRLM and Difference of Gaussian (DoG) features, which have been extracted from the selected region. The algorithm has achieved an accuracy of 89.11% using SVM classifier. Y.Ireaneus Anna Rejani+, Dr.S.Thamarai Selvi* had proposed system focuses on the solution of two problems. First problem is what is the way to detect tumors as suspicious regions with a very weak Contrast at background and second problem is how to extract features which categorize tumors. The detection method of tumor follows scheme of (a) mammogram enhancement (b) The segmentation of the tumor area. (c) From the segmented tumor area extraction of features. (d) The use of SVM classifier. This can be enhance and define conversion of the image quality to a understandable level. The procedure of Mammogram enhancement includes methods like filtering, top hat operation, DWT. Contrast stretching method is used to increase the contrast of the image. The Mammogram images segmentation plays an important role to advancement in improving the detection and diagnosis of breast cancer also the thresholding an common segmentation method used and the features are extracted from the segmented breast area. Next which regions using the SVM classifier and method was tested on around 75 mammographic images from mini-MIAS database which achieved a 88.75% sensitivity.



(a) Original mammogram. (b) Filtered image. (c) Second level DWT reconstructed mammogram. (d) Tumor segmented output.



As shown in figure 2.1 b. the existing system where the Detection of tumors in mammogram is divided into three main stages. The enhancement procedure is initial step and image enhancement techniques are used to improve an image, To see by modifying the colors or intensities need to increase the signal to noise ratio and to make certain features easier Then intensity adjustment is an image's intensity values to a new range. After the mammogram enhancement segment the tumor area. Then the features are extracted from the segmented mammogram. Now a stage involves the classification using SVM classifier.

III. PROPOSED SYSTEM

The 10-fold cross validation to get an accurate outcome is used in proposed system. The data set of Wisconsin breast cancer diagnosis is taken from UCI machine learning repository. The performance of the proposed system is appraised considering accuracy, sensitivity, specificity, false discovery rate, false omission rate and Matthews's correlation coefficient. Best result of training and testing is achieved by this approach also the techniques achieved the accuracy of 98.57% and 97.14% by Support Vector Machine and K-Nearest Neighbors. The classification model can be able to work with the linear or nonlinear problem. The Logistic Regression, Support Vector Machine (SVM) etc. are utilized for a linear problem. The K-Nearest Neighbors (K-NN), Kernel SVM, Random Forest etc. are used for the nonlinear problem.

A. Support Vector Machine

It is the generalization of maximal margin classifier which comes with the definition of hyper plane. In an n-dimensional space, the hyper plane is of (n-1) dimension with flat subspace that need not pass through the origin. The hyper plane is not visualized in higher dimension but the notion of an (n-1) dimensional flat subspace still

Applies [10]. If there doesn't exist any linearly separable hyper plane for any dataset, linear classifier can't be formed in that case. If there doesn't exist any linearly separable hyper plane for any dataset, linear classifier can't be formed in that case. Kernel trick have to be applied to maximum-margin hyper planes to develop nonlinear classifier In p-dimensions, a hyper plane is described as follows.

Where β0, β1, β2 … βp are the hypothetical values and Xp are the

Data points in sample space of p dimension.

B. K-Nearest Neighbors

K-Nearest Neighbors (K-NN) algorithm is another supervised machine learning technique used for classification and regression. It does not make any Assumption on the obtained fundamental data, it perform analysis. Initially it gathers data point which is closest, the

 DOI : 10.35291/2454-9150.2020.0465

algorithm than sorts the closest data points in distance from arrival data points. Euclidian distance is suggested method to calculate distance. In KNN, the number of closest data points is usually chosen as an odd number if the number of classes is 2.

IV. THE PROPOSED METHODOLOGY

Wisconsin Breast Cancer (WBC) and name of Brest cancer whose data retrieved from UCI machine learning repository dataset [11]. This dataset comprises of 699 instances, where the cases are labeled as either benign or malignant and 458 (65.50%) of the cases are benign and 241 (34.50%) are malignant the dataset is partitioned into two classes 2 and 4, where 2 denotes the benign class and 4 denotes the malignant class. The benign instances are represented as positive class and the malignant instances are represented as negative class in our study. Finally, the dataset is randomized to guarantee the correct circulation of data. The k-fold cross-validation is used where the given data set is split into k equal size chunks. A single chunk is used for testing and k-1 chunks are used for training.

C. The Performance Measure Indices

of some performance measure indices used to measure performance of machine learning, A confusion matrix for actual and predicted class is formed comprising of TP, FP, TN, and FN to evaluate the parameter. The significance of the terms is given below.

TP = True Positive (Correctly Identified)
TN = True Negative (Incorrectly Identified)
FP = False Positive (Correctly Rejected)
FN = False Negative (Incorrectly Rejected)

The performance of the proposed system is measured by the following formulas:

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity (Sen.)} = \frac{TP}{TP + FN}$$

$$\text{Specificity (S pec)} = \frac{TN}{TN + FP}$$

$$\text{False Discovery Rate (FDR)} == \frac{FP}{TP + FP}$$

$$\text{False Omission Rate (FOR)} = \frac{FN}{TN + FN}$$

Matthews Correlation Coefficient (MCC)

$$\frac{(TP\ X\ TN) - (FP\ X\ FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

In this paper we are using Kernel Support Vector Machine and K-Nearest Neighbors which usually is implemented in a high configuration computer. 10-fold technique is usedi.e. the data set was split into 10 chunks. This technique is utilized to approve the methodical model and 9 folds are utilized for training and the rest one for testing in 10 fold cross validation. We have formed a confusion matrix from the classifier. We have utilized 629 (90%) instances of total data for training both in Support Vector Machine and K-Nearest Neighbors individually. The remaining 70 (10%) instances used for testing both in SVM and K-NN individually.

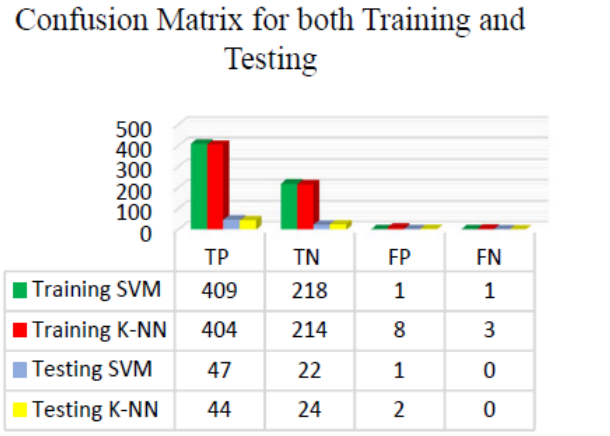


Fig. 1. Confusion matrix both in training and testing phases

For the prediction of breast cancer

Here we have shown the results obtained from both training phase and testing phase, initially of testing phase as shown below in table 1. Fig.1 the comparison between testing phases with SVM and existing models.

The confusion matrix graphical representation for each modality is shown fig. 1 depicts that the correctly identified value is comparatively higher in SVM and K-NN for training and testing. Among the 629 instances 409 and 404 is correctly identified in SVM and K-NN respectively. The performance measures indices are calculated both for training and testing using the above-described equation.

Table.1.1: Comparison between testing methods of proposed system with existing methods.

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)
PSVM[12]	96	97.3684	93.4426
St-SVM [12]	94.86	95.65	93.33
SSVM[12]	96.5714	96.5812	96.55
NSVM[12]	96.5714	96.5812	96.5517
LPSVM[12]	97.1429	98.2456	95.082
Proposed model using SVM	98.57	100	95.65
Proposed model using K-NN	97.14	100	92.31

As we can see the sensitivity of the proposed model using SVM and K-NN model is 100% and the Accuracy & specificity using SVM are 98.57% & 95.65%, and the Accuracy & specificity using K-NN are 97.14% & 92.31%. On comparing with above methods shown in table 1.1 the methods like LPSVM, NSVM are around 96-97 % of Accuracy, sensitivity and specificity on average. Similarly PSVM Gives 96% accuracy.97% of sensitivity

 DOI : 10.35291/2454-9150.2020.0465

and around 93% of specificity and St-SVM and SSVM shows 94% on average of these three mentioned parameters. This shows that the proposed system has more accuracy, sensitivity & specificity.

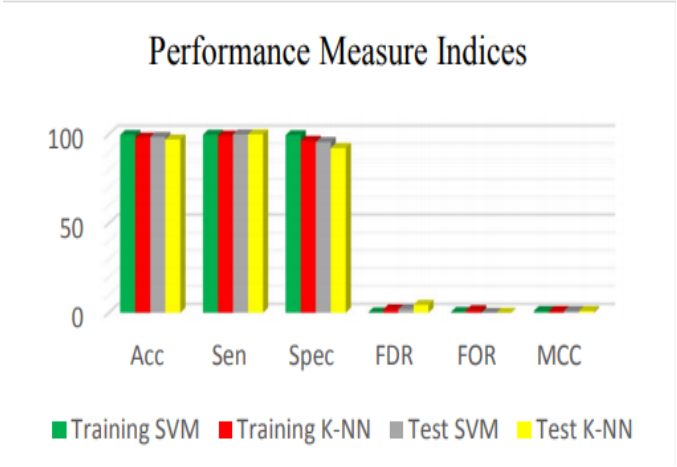


Fig.1.2. Training and testing phase performance measure indices for breast cancer prediction.

TABLE II. PERFORMANCE MEASURE INDICES

Parameters	Training Phase		Testing Phase	
	SVM	K-NN	SVM	K-NN
Accuracy (%)	99.68	98.25	98.57	97.14
Sensitivity (%)	99.76	99.26	100	100
Specificity (%)	99.54	96.40	95.65	92.31
Geometric Mean of Sensitivity and Specificity (%)	99.65	97.83	97.83	96.16
False Discovery Rate (%)	0.24	1.94	2.08	4.35
False Omission Rate (%)	0.46	1.38	0	0
Matthews Correlation Coefficient	0.99	0.96	0.97	0.94

As shown in above table II. The performance measure indices is shown of both training and testing phase The accuracy of the training phase and testing phase of SVM is 99.68 % and 98.57 whereas for K-NN is 98.25% and 97.14%. Sensitivity of SVM is 99.76% and 100% in training and testing, In K-NN its 99.26 & 100%.The St-SVM shows lower performance in testing phase. The authors in [12] used 4 fold techniques where we utilized 10 fold techniques

V. CONCLUSION

The paper predict Brest cancer which may causes many deaths of ladies in world so early detection of disease will save lots of lives. The proposed system identifies Brest cancer by SVM and K-nearest neighbor. Using Python for SVM classification is most effective in classifying the diagnostic data set into the two classes in view of the seriousness of the cancer. The accuracy of 99.68% in SVM in training phase is obtained. The proposed system has more accuracy, sensitivity & specificity on testing phase on comparison with existing methods As shown in Table.1.1. The obtained classifier by supervised machine learning techniques will be supportive in the medical field of disorders and proper diagnosing. This SVM classifier may used for helping radiologists making accurate and quicker diagnostic decision, which might reduce unnecessary biopsies for detection of tumour area. It might reduce the hazards of radiologist for taking quick decision

REFERENCES

E. Karthikeyan , S. Venkatakrishnan" Breast Cancer Classification using SVM Classifier" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019

[2] Prediction of Breast Cancer Using Support Vector Machine and K-Nearest.Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, and Md. Kamrul Hasan"

[3] EARLY DETECTION OF BREAST CANCER USING SVM CLASSIFIER" Y.Ireaneus Anna Rejani et al International Journal on Computer Science and Engineering Vol.1(3), 2009, 127-130"

[4] Detection and Classification Technique of Breast Cancer Using Multi Kernel SVM Classifier Approach" Pramit Brata Chanda, Subir Kumar Sarkar"

[5] DURGESH K. SRIVASTAVA, 2LEKHA BHAMBHU" DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE" Journal of Theoretical and Applied Information Technology© 2005 - 2009 JATIT.

[6] Theodoros Evgeniou and Massimiliano Pontil MIT, E25-201,Cambridge, MA 02139,USA. "WORKSHOP ON SUPPORT VECTOR MACHINES: THEORY AND APPLICATIONS"

[7] Simon Tong" Support Vector Machine Active Learning with Applications to Text Classification" Journal of Machine Learning Research (2001)45-66.

[8] Chih-Wei Hsu, Chih-Chung Chang, and Chih-JenLin" A Practical Guide to Support Vector Classification"

[9] M. K. Hasan, M. M. Islam and M. M. A. Hashem, "Mathematical model development to detect breast cancer using multigene genetic programming 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, pp. 574-579, 2016

[10] Breast Cancer Wisconsin (Original) Data Set, [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/breast Aug 25 2017

[11] E.C.Fear, P.M.Meaney, and M.A.Stuchly,"Microwaves for breast cancer detection", IEEE potentials, vol.22, pp.12-18, February-March 2003.

[13] A.Rajesh, Dr.E.Mohan "Classification of Mammogram Using Wave Atom Transform and Support Vector Machine Classifier," International Journal of Computer Science and; Information Technologies– Volume 7 ,issue 2, 467-470, Feb 2016.

[14] Satish Saini and Ritu Vijay, "Design and Analysis of Breast Cancer Detection System Using Mammogram Features Extraction", in International Journal of Electronics and Computer Science Engineering, Vol. 3, Number 4, 2015, pp 406-409.

15] J. Blagojce, K. Ivan, T. Katarina, D. Ivika and L. uzana,"Mammographic Image Classification Using Texture Features", 9th conference for Informatics and Information Technology, (2012).

[16] Snehal A. Mane, Dr. K. V. Kulhalli "Gabor Wavelet analysis for mammogram in Breast Cancer Detection", International Journal on Recent and Innovation Trends in Computing and Communication 2015,Volume:2, Issue 4, ISSN:2321-8169.

 DOI : 10.35291/2454-9150.2020.0465