Grenze International Journal of Engineering and Technology, June Issue

Comparative Study of Algorithms used for Predicting Invasive Ductal Carcinoma (IDC)

Suhana R Ghasi[1], Dr. Andhe Dharani[2] and Chandrani Chakravorthy[3]
[1-3]RV College of Engineering, Bengaluru, India
Email: suhanarg.mca19@rvce.edu.in, andhedharani@rvce.edu.in, c.chandrani@gmail.com

***Abstract*—Breast cancer is a disease which we hear about a lot nowadays. It is one of the most widespread diseases. It affects both men and women all around the world. Research says that every year around 2000+ new cases of breast cancer for men and about 2,30,000 new cases for women. Diagnosis of this disease is crucial and its demand the faster treatment for the recover. Correct and early diagnosis of these disease can save the patient. Breast cancer detection is done with the help of mammograms, which are basically X-rays of the breasts. Detection is not easy due to different kinds of uncertainties identified in these mammograms. Machine Learning (ML) techniques can help for the early detection of breast cancer. This paper consists of these techniques which can help doctors for early detection and diagnosis of breast cancer which could greatly enhance the survival rate of patients. In this paper, Support Vector Machine (SVM) and Convolution Neural Network (CNN) model applied for the data set which contains both benign and malignant breast cancer images and generated confusion matrix for validating the result. We also performed a comparative analysis of Logistic regression, SVM, Naive Bays, Neural Networks, Decision tree, Adaboost and Random Forest.**

***Index Terms*— Breast cancer, Support Vector Machine (SVM), Convolution Neural Network (CNN), X-rays, mammograms, histopathological images, Invasive Ductal Carcinoma (IDC).**

I. INTRODUCTION

Breast cancer signifies as a malignant tumour developed from uncontrollable growth of breast cells. It occurs in women and rarely in men. Symptoms include inverted nipple, lumps and nipple discharge, Change in size, shape or appearance of breast. In 2020, it was estimated 2.3 million women diagnosed with Invasive ductal carcinoma (IDC) and 685000 deaths due to Invasive Ductal Carcinoma globally. And in last five years, 7.8 million women were diagnosed with IDC, this ratio reduce to 0.5-1% in men. After diagnosis the survival rate of IDC from last 5 years is 90% in developed countries, 66% in India and 40% in South Africa. Early detection and treatment have proven successful [1]. Now a days there is an increase in rate of Invasive carcinoma incidences in our country, particularly in younger age groups. Detection of IDC is too late and it is directly proportional to the low survival rates in India.

Scientists applied several techniques and methods and many researches are being performed related to the detection of breast cancer The computer science has proven highly beneficial in identifying tiny specs of cancer with accuracy rate more than 95%, much higher than the 65% accuracy rate that radiologists typically achieve. Artificial Intelligence, ML and Neural network are used to detect breast cancer by improving image processing and reducing noise and other aspects which are main reason for decrease in prediction accuracy rate. Deep

Grenze ID: 01.GIJET.8.2.521_1

Learning which is subfield of ML and dependent on AI and ANN is generally used in image processing, CV, histopathological diagnosis and it improve the accuracy rate in breast cancer prediction. In addition to this digital pathology is used to get digitalized high-resolution images of breast carcinoma. Machine learning methods are considered to be highly beneficial for segregating the data into testing and training datasets and analyses the input images based on the characters or the labels assigned to them. ML is classified into supervised, unsupervised and reinforcement learning. The actual way to categorize and detect the breast cancer is by classification and data mining techniques. ML algorithms like SVM algorithm is used to perform non-linear classification and it creates a hyperplane that separates data into two classes with the maximum margin and hyperplane is determined by a subset of the points of the two classes. SVM are well suited and it increases accuracy and effectiveness up to 97% and helps to get better decision Output obtained is fast, effective and reduces the complexity [2].

Artificial neural network is a collection of connected units called artificial neurons. It uses multiple layers of input information and process to generate output. It understands the relationship and effects between input variables and the outcome. There are so many supervised learning algorithms used in ANN. One of the most common is backward propagation of errors. The artificial neurons strengthen themselves by back propagation technique. The back propagation algorithm is generally based on the error-correction learning rule. The back-propagation process works on the basis of forward and a backward pass among the various layers of the network. Pathologic reports help to confirm that whether the cancer is benign and malignant. A three-layer include input layer of 23 variables, hidden layer of 1000 nodes and an output layer of single node in feed-forward network with MATLAB is used to build Artificial neural network. The risk of malignancy is obtained by a number between 0 and 1 provided by output node [3].

II. RELATED WORK

R. Chtihrakkannan and R. Karthikeyan used bilateral filter and Gaussian blurring on input images to inspect human breast cancer. Later pictures are segmented and GLCM (Grey-level Co-occurrence Matrix) is used for feature extraction, and by Multi-layer Perception classifier containing ANN, SVM, KNN etc. got an average accuracy rate as 89.77% and high accuracy rate as 96% resulting in early diagnosis of breast cancer and henceforth decrease the mortality rate by breast cancer in human [4]. Shashidhar R and Arunakumari B N they have mainly focus on key challenges in detection of breast cancer. They used SVM, k-NN and decision tree classifier on Wisconsin Breast Cancer datasets to categorize the tumours into malignant or benign. The Outcome illustrate that SVM give the high accuracy 98% with low error rate [5]. Vishnukumar K. Patel and Prof. A. C. Suthar presented image enhancement and spatial domain techniques to enhance the digital images of breast cancer tumours. The author further used Gabor Filter Algorithm to get visual cortex of mammals to approximate the localize excise cancerous tissue frequency domain. The work on image segmentation in MATLAB platform using Gabor filter algorithm gave clear picture about normality or abnormality in tumour images [6]. M. Tahmooresi and A. Afshar combined AI and Machine Leaning (ML) methods for prediction and increase accurate decision making. Machine Learning algorithms are introduced with many features from healthcare dataset containing images, x-rays, blood samples etc. Samples of training data are gathered by ANN to train the machine. The hidden data is detected by trained machine and it is used to help searching patterns in IDC patients [7]. Naresh Kumar J, Sean Benhur. J used different techniques of ML and data mining to detect breast cancer. The models are SVM, Adaboost tree and ANN. Further, SGD Classifier model is used for pre-processing and observing the contribution to the loss reduction. The result analysis clearly stated that the performance of the classification is dependent on on the ml algorithm [8]. Guanqing Li and Jiawen Zhang discussed main problems in detecting breast cancer by digitalized images of the Fine Needle Aspiration. The author used deep learning model, seven classification models. Feature selection is used to increase the accuracy rate. Random Forest gives the highest accuracy rate of 96% among seven ML models [9].Nikita Rane and Jean Sunny has given several classification and cluster methods used to predict breast cancer, and website will be build using python framework. Features of dataset are formed by obtaining digitized images from the biopsy reports. The author further discussed that the SVM provide non-probabilistic classifier. The new examples provided by SVM mapped in space and predict to which category they belong [10]. Sivapriya J and Aravind Kumar V used UCI ML repository datasets for study containing 65.5.% malignant and 34.5% benign. Matplotlib and seaborn are used for data visualization. Authors analysed large amount of information to make decision and represented the data using maps, plots, patters, graph etc. and concluded that classification algorithm gives 99% of accuracy and predict the cancer accurately [11]. Hiba Masood have been presented different feature selection techniques like wrapper methods, filter methods and embedded methods. The author further used NodeMcu an open-source

firmware that provide flexibility to build IOT based application by helping to operate much faster. Author analysed AdaBoost a ML meta-algorithm and PNN. Probabilistic Neural Network (PNN) replaces sigmoid activation fuction and handles multiclass problem [12]. Vikas Chaurasia and Saurabh Pal mainly focussed on data mining technique and Knowledge discovery in databases (KDD) used to detect IDC. In this paper author has investigated three data mining techniques: RepTree, RBF Network and Simple Logistic classifier. Among them Simple Logistic Classification gives highest accuracy of prediction by 74,7% and total time taken for building the model is approximately o.62 seconds [13]. Ashutosh Kumar Dubey and Umesh Gupta compared K-means and FCM Algorithms on IDC Data. k-means is simple and easy and give highest varience and more accurate value around 97% when compared to FCM. FCM is time consuming and iterative process, and gives an accuracy of 92% with high computation time [14]. Zuherman Rustam and Sri Hartini proposed new approach for breast cancer diagnosis to get high accuracy with less computational time. The author uses modified form of KC-means and FCM. The dataset used contain 201 malignant and 85 benign samples.the accuracy rate of FCM clustering is 85.26% and KC-means is 89.74% [15]. Sweta Bhise and Simran Bepari pr compared the classification results of CNN, SVM, Naïve Bayes, Random Forest and Logistic Regression. The BreakHist dataset from Kaggle repository is used. The dataset consist of four directories. The Classification results predicted that CNN and SVM are the best classifiers [16]. Ram MurtiRawat and Shivam Panchal investigated ANN and SVM for their efficiency and accuracy based on feature selection in breast cancer diagnosis. The maximum accuracy was achieved by 10-fold cross validation and future selection techniques. The Wisconsin Diagnostic breast cancer dataset is used by authors for breast cancer prediction. SVM predicted more accurate value of 97.14% as compared to ANN with accuracy rate of 96.71% [17]. Animesh Hazra and Subrata Kumar Mandal analysed Naïve Bayes and Support vector machine for predicting IDC. Feature selection is achieved by pearson Correlation Coefficient. If Five features are considered The maximum accuracy of 97.3978% was given by Naive Bayes classifier. If we consider 19 features then SVM provide more accurate and good result [18].

III. METHODOLOGY

The proposed technique offers a Support Vector Machine and a Convolutional Neural Network to detect breast cancer order. The presentation of the model is approved against the pictures taken under genuine conditions. SVM is a used to create hyperplane that can segregate p-dimensional spaces into classes.

$$B_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta p X_p \quad \text{(i)}$$

where β_0, β_1, β_2 … β_p are the hypothetical values and X_p are the data points in sample space of p dimension.
The dataset contains 5547 pictures, that are pre-processed with the assistance of the histogram. the correction, which makes up for the permanence between the pictures taken under genuine conditions. These pictures are resized to a standard size picture utilizing the middle square trimming strategy. The CNN-based order model is then prepared and tried [19].

Steps fo prediction of IDC

Step 1: Select the histopathological Images and resize all breast cancer images to standard size.
Step 2: Convert the original image to RGB images.
Step 3: Each pixel value in the K-Means Clustering image segmentation will contain only infected region of breast and predicting disease from histopathological image.

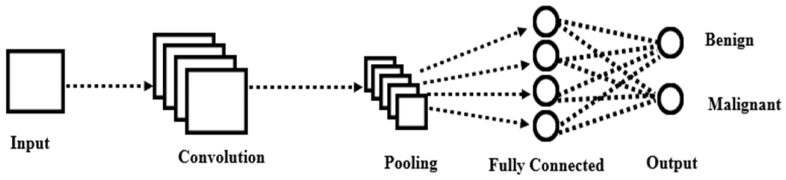

Figure 1: Classification using CNN architecture

Step 4: Threshold segmentation is used for feature extraction of each pixel images.
Step 5: Split dataset into preparing, test and approval set utilizing keras train_test_split order.
Step 6: Pre-process class marks and define our model engineering.
Step 7: The NumPy exhibit gets passed into the Convolution2D layer.
Step 8: Pooling technique considered MaxPooling2D that utilizations (2,2) windows across the

component map just keeping the greatest pixel esteem.

Step 9: Fit and train information and furthermore assess model on test dataset and ascertain the disarray framework.

Step 10: Confusion Matrix Graph are done using SVM Algorithm and CNN Algorithm.

Step 11: The validation of loss and accuracy result generated using deep learning algorithm.

IV. Result and Discussion

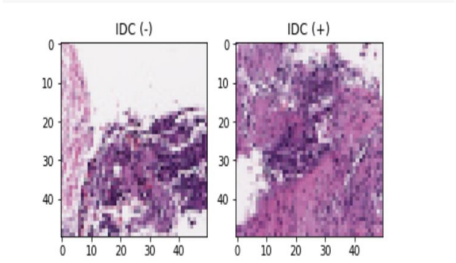

Figure 2: Resize image dataset and RGB color conversion

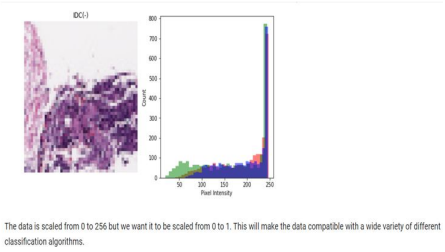

Figure 3: Images with pixel intensity

At the first stage pre-processing of histopathological images be done that commonly involves removal of low frequency, Noise Reduction, sharpening, Clarity, Texture, Masking, portions of images (in Figure 2). It is one of the most important techniques in image processing for enhancing data images prior to computational processing and getting rescale data from the dataset. (a) Resize image dataset (b)RGB color conversion of the original image of the tumour.

The second step involve the tumour images that show pixel intensity that is helpful in predicting the breast cancer detection from the dataset(in Figure 3). It will effectively predict the data from dataset by enhancing the performance of the overall prediction results of benign and malignancy of breast cancer.

The 3rd stage involve the plotting of Learning Curve which is plotted among Training Score versus Cross-validation score(in Figure 4).

In 4th stage the precision and recall is calculated and show the Confusion Matrix providing 89 True Positive, 43 False Negative, 463 False Positive and 515 True Negative values(in Figure 5).

The 5th stage is the comparison of accuracy obtained from various algorithms that depicts the exactness and affectability rate pace of the multitude of strategies chose for the grouping of tumour and disarray grid for approving the proposed framework(in Figure 6).

The 6th stage is the use of CNN Algorithm to process deep learning sequential model by training and testing the sets of histopathology images using deep learning sequential model(in Figure 7).

The 7th stage provides the precision, recall, accuracy, specificity, sensitivity percentage rate of all the methods selected for the classification of benign and malignancy and confusion matrix for validating the proposed system (in Fig. 8).

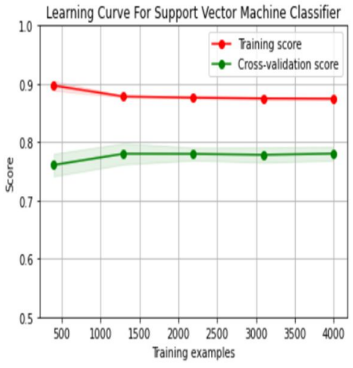

Figure 4: Learning Curve

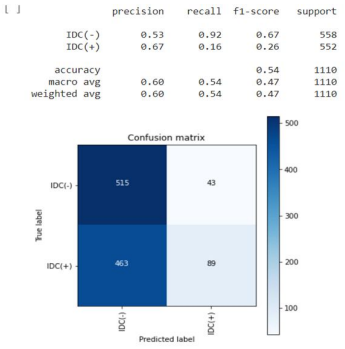

Figure 5: Confusion Matrix

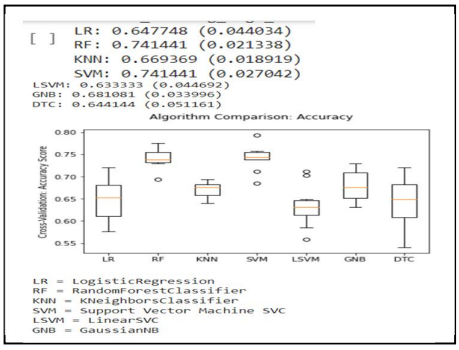

Figure 6: Comparing the accuracy obtained from various learning sequential model and training process of graph

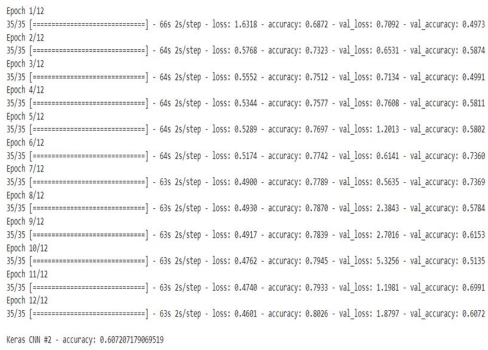

Figure 7: CNN Algorithm using the process of deep Algorithm

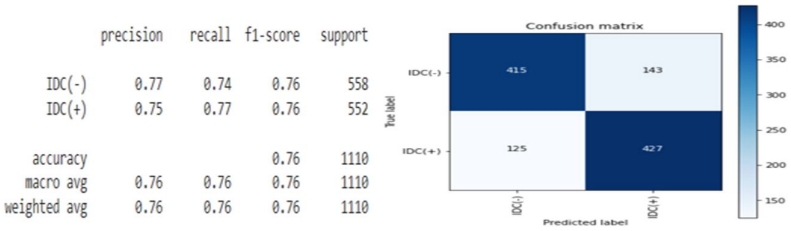

Figure 8 :CNN Algorithm using result prediction of precision, recall, specificity, sensitivity and confusion matrix

The comparison of various algorithm also discussed in the table I.

TABLE I: COMPARISONS OF VARIOUS ALGORITHMS RESULT BASED ON THE VARIOUS

Algorithm	**Primary Problem**	**Predictors**	**Power**	**Raw Implementation**	**Interpretability**	**Regression**	**Normalization**
k-NN[20]	Multiclass or binary	Numeric	Medium	Easy	Good	No	Required
Perceptron[20]	Binary	Numeric	Low	Easy	Good	No	No
Logistic Regression[20]	Binary	Numeric	Low	Easy	Good	No	No
Linear Discriminant analysis [20]	Binary	Numeric	Low	Medium	Medium	No	No
Naive Bays [20]	Multiclass or binary	Categorical	Medium	Medium	Good	No	Required
Decision Tree [20]	Multiclass or binary	Numeric or categorical	High	Difficult	Good	Yes	No
Random Forest [20]	Multiclass or binary	Numeric or categorical	High	Difficult	Good	Yes	No
Adaboost [20]	Multiclass or binary	Numeric or categorical	High	Medium	Medium	Yes	No
SVM [20]	Binary	Numeric or categorical	High	Very Difficult	Medium	No	Yes
Convolution Neural Networks [20]	Multiclass or binary	Numeric or categorical	High	Very Difficult	Weak	Yes	yes

V. CONCLUSION

Breast cancer is one of the significant causes of death in men and women. Early detection plays an important role in saving the life of men/women. At present, only few accurate prognostic and predictive factors are used clinically for managing the patients with breast cancer. Here, by making use of Clustering, high accuracy can be achieved in detection of effected cell shapes with exact marking on detected contours. The proposed system helps to enhance the breast cancer prediction by selecting optimal features. Detection of Invasive ductal carcinoma can be done with the help of modern machine learning and artificial neural network algorithms. In this paper, we focused on influence of machine learning and neural network in breast cancer detection.

REFERENCES

[1] 'WHO|Breastcancer',WHO.http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).

[2] Konstantina kourou Themis, P Exarchos Konstantinos, P Exarchos Michalis, V Karamouzis Dimitrios Fotiadis,"Machine learning applications in cancer prognosis and prediction", Computational andStructural Biotechnology Journal, Volume 13, 2015, pages 8-17

[3] Hiba Masood, "Breast Cancer Detection using Machine Learning Algorithm", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 02 | Feb 2021 p-ISSN: 2395-0072 page (s): 738-745

[4] R.Chtihrakkannan, P.Kavitha, T.Mangayarkarasi, R.Karthikeyan, "Breast Cancer Detection using Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019,page (s):3123-3125

[5] Shashidhar R,Arunakumari B N,Naziya Farheen H S,Puneeth S B,Santhosh Kumar R and Roopa M, "Breast Cancer Detection Using Supervised Machine Learning Algorithm" , Asian Journal of Convergence in Technology ISSN NO: 2350-1146 I.F-5.11, page (s):26-29

[6] Vishnukumar K. Patel1 , Prof. Syed Uvaid2 , Prof. A. C. Suthar 3, "Mammogram of Breast Cancer detection Based using Image Enhancement Algorithm", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 8, August 2012) page(s): 144-145

[7] M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath and M. A. Bamiah, "Early Detection of Breast Cancer Using Machine Learning Techniques", Journal of Telecommunication, Electronic and Computer Engineering, e-ISSN: 2289-8131 Vol. 10 No. 3-2, page (s):21-23

[8] Naresh,Kumar J, 2Sean Benhur. J, 3Dr.S.Kanchana, "Classification Analysis of Genetic Variations In Cancer Diagnosis By Multiclass Classifiers", International Journal of Creative Research Thoughts (IJCRT), Volume 9, Issue 2 February 2021 | ISSN: 2320-2882, page (S): 2147-2152

[9] Guanqing Li, Jiawen Zhang, Zhiyuan Lou, "BREAST CANCER DISANOSIS USING MACHINE LEARNING METHOD", University of California San Diego, La Jolla, CA 92093-0238

[10] Nikita Rane, Jean Sunny, Rucha Kanade, Prof. Sulochana Devi, "Breast Cancer Classification and Prediction using Machine Learning", International Journal of Engineering Research & Technology ISSN: 2278-0181 IJERTV9IS020280 Vol. 9 Issue 02, February-2020,page (s):576-579

[11] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S, "Breast Cancer Prediction using Machine Learning", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019, page (s):4879-4880

[12] Hiba Masood, "Breast Cancer Detection using Machine Learning Algorithm", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 08 Issue: 02 | Feb 2021 p-ISSN: 2395-0072 page (s): 738-745

[13] Vikas Chaurasia, Saurabh Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", International Journal of Computer Science and Mobile Computing (IJCSMC), Vol. 3, Issue. 1, January 2014, ISSN 2320–088X, page (s):10 – 22

[14] Ashutosh Kumar Dubey, Umesh Gupta , Sonal Jain, "Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data", " International journal of computer assisted radiology and surgery, vol. 11, no. 11, pp. 2033–2047, 2016. Page (s):18-28

[15] Zuherman Rustam, Sri Hartini, "Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel", 9th Annual Basic Science International Conference 2019 (BaSIC 2019), y, vol. 11, no. 11, pp. 2033–2047, 2019.page (s):1-8

[16] Sweta Bhise, Simran Bepari, Shrutika Gadekar, Deepmala Kale, Aishwarya Singh Gaur, Dr. Shailendra Aswale, "Breast Cancer Detection using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV10IS070064 Vol. 10 Issue 07, July-2021, page (s):99-102

[17] Ram MurtiRawat, Shivam Panchal, Vivek Kumar Singh, Yash Panchal, "Breast Cancer Detection Using K-Nearest Neighbours, Logistic Regression and Ensemble Learning", e International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4, page (S):534-540

[18] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 145 – No.2, July 2016, page (s):39-44

[19] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, 'Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis', Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[20] 'Aisoma| Useful Comparison tables for AI', https://www.aisoma.de/useful-comparison-tables-for-ai-data-science-iot/ (accessed September 17,2019)