Classification of Breast Cancer Histopathology Images using Machine Learning Algorithms

Kavya K, Savita K Shetty

Abstract: ***Machine Learning (ML), provides system the capacity to learn instinctively and allows systems to improve themselves with past experience and without being programmed specifically. In the field of Medical Science, ML plays important role. ML is being used to develop new practices in medical science which deals with huge patient data. Breast Cancer is a chronic disease commonly diagnosed in women. According to the survey by WHO, rank of breast cancer is at number one as compared to other cancers in female. BC has two kinds of tumour: Benign Tumour (BT), and Malignant Tumour (MT). BTs are treated as non-cancerous cells. MTs are treated as cancerous cells. The unidentified MTs in time stretch to other organs. Treatment procedure for BT and MT is different. So, it is salient to determine precisely whether a tumour is BT or MT.***
In this proposed model, Histopathology Images are used as dataset. These Histopathology images are pre-processed using Gaussian Blur and K-means Segmentation. The pre-processed data fed into feature extraction model. ML algorithms such as Support Vector Machine (SVM), Random Forest (RF) and Convolution Neural Network (CNN) are applied to extracted features. Performance of these algorithms is analysed using accuracy, precision, recall and F1-score. CNN gives the highest accuracy with 87%.

Keywords: Machine Learning (ML), Breast Cancer (BC), histopathology.

I. INTRODUCTION

According to the survey done by International Agency of Research on Cancer (IARC) Breast Cancer is the deadliest cancer in World for females with incidence rate of 24.7% and mortality rate of 13.4% (rate per 1 lakh person per year) in the year 2018 [1]. It is crucial to detect the type of Breast cancer Tumour in early stages to reduce the mortality rate. Diagnosis of breast is done by three types: (i) Mammogram – It is a kind of x-ray used during early stages of screening (ii) Ultrasound – Breast is scanned to detect solid mass and fluid filled cyst using sound waves (iii) Magnetic Resonance Imaging (MRI) – Which has multiple images of breast at different position to identify the cancer.

Understanding the Histopathology images and to diagnosis the Breast cancer requires well educated Pathologist. Determining the tumour accurately is still a challenging task for pathologists. Early detection of tumour helps in reducing the rate of mortality. The emerged features and technologies in medical science in associated with machine learning help the doctor to predict the cancer in early stage.

ML has major role in medical science in innovating new medical practices. ML has predictive and classification models which deals with huge patient's data and helps in predicting the disease. In Breast Cancer Research, the ML technologies can be used to predict the BC as BT or MT with high accuracy.

In this paper, Histopathology Images are used as dataset. These Histopathology images are pre-processed using Gaussian Blur and K-means Segmentation. Features are extracted from results obtained from image pre-processing model using pixel mean value. Then Traditional ML algorithms such as SVM, RF and CNN are applied to the obtained features. Results obtained from each of the algorithm are compared to identify the best algorithm with highest accuracy. Same is depicted using data visualization.

II. LITERATURE SURVEY

Machine Learning Techniques are the recent trend in the field of Medical Field, especially in cancer prediction. Scientists are coming up with Machine Learning Techniques which helps to create new medical procedure to identify Cancer Tumour Prediction. Many researchers have been done in prediction of Breast Cancer tumour prediction.

This section explains some of the research done to Predict Breast Cancer using ML techniques. And their results have been mentioned and compared here.

Sara Laghmati [2] compared different Classification algorithms applied on Mammographic mass dataset obtained from Breast Imaging – Reporting Data System (BI-RADS) to predict the severity of Breast Cancer. The result obtained tells that the Artificial Neural Networks gives highest accuracy with 84%. This paper explains the ML algorithms applied to BI-RADS dataset. But data should be pre-processed. And Features are to be extracted from the pre-processed data.

Dana Bazazch [3] compares three of the most used ML techniques in terms of Accuracy, Precision, Recall and Area of ROC on Wisconsin Breast Cancer Dataset. Support Vector Machine gives the highest accuracy with 97.4% for Benign Tumors and 96.3% for malignant tumors. The data set used is very small, autoencoders can be used to increase the dataset. For image data, data augmentation can be used to increase the data set size.

Ayush Sharma [4] compares the ML algorithms to predict the breast cancer tumor using Wisconsin Prognostic dataset. Logistic Regression gives the highest accuracy with 96.89%. In this paper, min max normalization is used for pre-processing. But should have used PCA to reduce the feature space since it contains more features.

PanuwatMekha [5] tried applying different Activation functions like tanh, rectifier, Maxout and Max-rectifier to several ML algorithms on Wisconsin Dataset. The result obtained from this is, deep learning method with max-rectifier activation function gives optimized result of 96.99%.

Revised Manuscript Received on August 10, 2020.

Kavya K, Department of Software Engineering, M. S. Ramaiah Institute of Technology, Bengaluru, India.

Savita K. Shetty, Assistant Professor, Department of Computer Science and Engineering M. S. Ramaiah Institute of Technology, Bengaluru, India.

Zhan Xiang [6] used CNN to bring out the required features from histopathology images and to classify the image as BT or MT from SoftMax function. He also applied fine tune technology and data augmentation to avoid over-fitting, and cross validation strategy to enhance the performance of the system. But histopathology images might be noisy and inconsistent. So, it needs to be pre-processed.

Yuqian Li [7] proposed patches screening method to discriminate patches that do not contain any information according to the label. CNN is used to extract smaller and larger patches from histology images. The images are classified as normal, benign and malignant. This method gave 88.89% accuracy.

Most of the researchers used traditional approach for feature extraction. The difficulty with this traditional approach is that it is necessary to choose which features are important in each given image. As the number of classes to classify increases, feature extraction becomes more and more cumbersome.

Here in this proposed model, both traditional and CNN method is used for feature extraction. Convolution Layer is applied for the features extracted from the traditional approach and for the pre-processed data.

III. METHODOLOGY:

In this section, Fig. 1 explains the architecture of the model.

A. Dataset

The dataset used in this project is from kaggle "breast histopathology images". 277,524 histopathology images of size 50 x 50 were retrieved from 162 whole slide images of BC patients. Among these 198,738 images are benign and 78,786 are malignant. Each image name is in the format of: iXYclassT.png. Where, i is the patient ID, X is the x-coordinate from where the image is trimmed, Y is the y-coordinate from where the image is trimmed, and T indicates the type of tumour where 0 is BT and 1 is MT. In this project 70% of dataset is used as training data and remaining 30% of dataset is used as testing data.

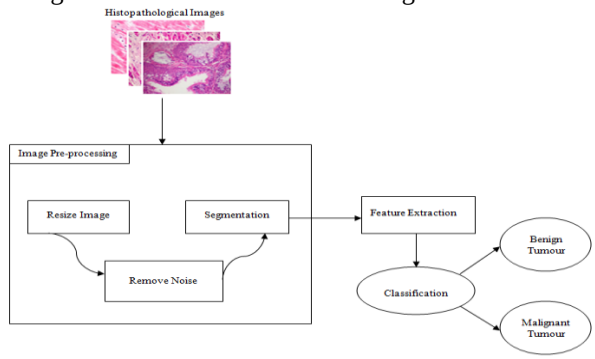

Fig.1. Architecture diagram of the model.

B. Image Pre-processing

Image Pre-processing improves the histopathology image by suppressing unwanted distortions and improving the image features that are necessary for further processing. In this pre-processing stage we undergo 3 steps. Fig.3 shows the image pre-processing.

- **Resize Image**

Original images may vary in size, so, we need to fix a constant size for every image. In this project image bicubic interpolation technique given in equation 1 is used to resize the image.

$$g(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}\, x^i y^j \qquad (1)$$

In this project the fixed size is of each image is (50, 50, 3) where height – 50pixel, width – 50pixel and 3 dimensional. Fig. 2(i) shows the resized image.

- **Remove Noise**

Noise is an undesired detail which deprecates the quality of image. Pixels of the image delineate dissimilar intensity than the accurate value [8]. Noise can deteriorate the features of the image. Noise should be removed at high priority before image processing. According to the paper [9], Gaussian filter outperforms than Weiner filter in medical imaging. In this project we have used Gaussian filter to remove noise. This filter is efficient in computation and degree of smoothening is controllable. The formula of Gaussian filter is given in equation 2. Fig. 2(ii) shows the noise removal.

$$g(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (2)$$

- **Segmentation**

In Image Processing, image segmentation is the procedure of dividing image into multiple regions. Segmentation plays a vital role in medical image processing as it clusters the image and helps in finding the cancerous cell in the given image. Clustering is dividing the pixels into a number of groups. K-means is the mostly used clustering algorithm. K-means clustering is simple and less complexity in computation yet effective in medical image segmentation [10]. Here for Breast cancer dataset we have opted number of clusters as 4. Fig. 2 (iii) shows segmentation of histopathology image using k-means clustering.

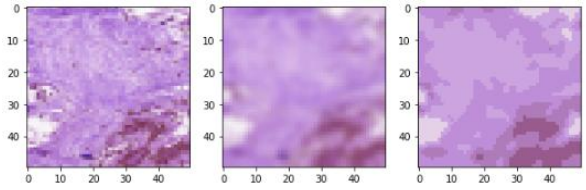

Fig.2. Image Pre-processing – (i) Image Resize, (ii) Noise Removal & (iii) Segmentation

C. Feature Extraction

Feature Extraction results in decreased number of features in pre-processed data from histopathology images by extracting new features from the pre-processed image dataset. Extracting the features improves the accuracy; reduce over fitting, increase the processing speed. In this project mean value technique is used for feature extraction as it reduces the dimensionality. After image pre-processing each image has the size (50*50*3). 50*50 indicates height and width and 3 is the dimension of the image. Dimension stores the RGB colour intensity. In feature extraction mean value of intensity of colours is taken for each pixel thus reducing the image size from (50*50*3) to (50*50) shown in (3).

$$mean = \frac{i(R)+i(G)+i(B)}{3} \qquad (3)$$

Where, i(R), i(G) & i(B) indicates intensity of colors. Fig. 3 shows the feature extracted image.

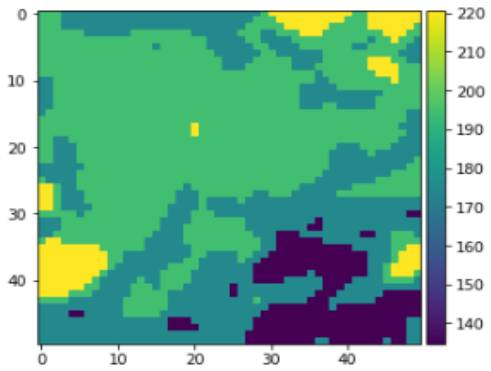

Fig.3. Feature Extraction

D. Classification

The aim of the project is to classify the histopathology image as Breast Cancer Tumor i.e., BT or MT. ML Algorithms such as SVM, RF and CNN can be applied to the feature extracted images of histopathology images. The result will be either Class 0 which indicates the benign tumor and class 1 indicates the malignant tumor.
Finally, the results obtained from various ML algorithms were compared with respect to accuracy given in (4).

$$Accuracy\ (AC) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (4)$$

Where, TP is True Positive
TN is True Negative
FP is False Positive
FN is False Negative.

- **Support Vector Machine (SVM)**

SVM is a supervised ML algorithm for classification and regression problems [10]. It finds the hyper-plan which divides two classes. Fig. 4 shows the result of SVM model.

	precision	recall	f1-score	support
0	0.85	0.83	0.84	1855
1	0.54	0.59	0.56	645
accuracy			0.77	2500
macro avg	0.70	0.71	0.70	2500
weighted avg	0.77	0.77	0.77	2500

Fig. 4. Classification Report of SVM Model

- **Random Forest (RF)**

RF is tree-based Ensemble Learning Algorithms. RF creates group of decision trees (DT) from randomly selected subset of training data. It piles up the poll from all DTs and selects the majority poll. Fig.5 shows the result of RF model.

	precision	recall	f1-score	support
0	0.87	0.93	0.89	4409
1	0.75	0.60	0.66	1591
accuracy			0.84	6000
macro avg	0.81	0.76	0.78	6000
weighted avg	0.83	0.84	0.83	6000

Fig. 5.Classification Report of RF Model

- **Convolution Neural Network (CNN)**

CNN is a Deep Learning Algorithm most commonly used for Image Classification process. A CNN has a Convolution Layer, Rectified Linear Layer (ReLU), Pooling Layer and fully connected Layer. CNN is applied to both feature extracted data and pre-processed data. Rectified linear unit is used as activation function. CNN sequential model is shown in figure 6.

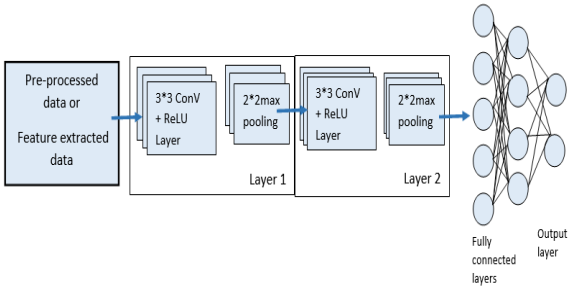

Fig. 6.Convolution Neural Network Model

IV. RESULTS

Image Pre-processing techniques have been implemented to Histopathology image clean the data. This results in uniform size of data (50*50*3). Feature Extraction has been done to pre-processed images resulted in dimensionality reduction to (50*50). ML algorithms such as RF, SVM and CNN are applied to result obtained from Feature Extraction. Percentage of accuracy obtained from each ML algorithm is given in table 1. Convolution Neural network gave the highest accuracy for testing dataset with 87%. Accuracy plot is shown in figure 6 for the same.

Table-I: Analysis of ML Algorithms

Dataset	RF	SVM	CNN (applied to pre-processed data)	CNN (applied to feature-extracted data)
Training Dataset	99	96	92	77
Testing Dataset	84	77	87	76

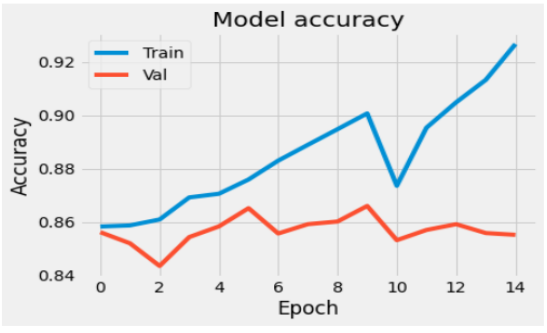

Fig. 7.Accuracy of CNN when applied to pre-processed data

V. CONCLUSION

Breast cancer is at number one as compared to other cancers in female. BC has two kinds of tumor: Benign Tumor (BT), and Malignant Tumor (MT).

MTs are treated as cancerous cells. The unidentified MTs in time stretch to other organs. So, it is salient to determine precisely whether a tumor is BT or MT. ML algorithms such as RF, SVM and CNN are applied to pre-processed histopathology images to classify the image as cancerous or non-cancerous. Results gained by each algorithm are as follows; RF – 84%, SVM – 77% and CNN (applied to pre-processed data) – 87% and CNN (applied to feature extracted data) – 76%. Convolution Neural Network applied to pre-processed data gave the highest accuracy of 87%.
In future the same model can be implemented using GPU and parallel processing which enhance the performance of the model.

REFERENCES

1. International Agency of Research on Cancer, Available: https://www.iarc.fr/cards_page/iarc-research/
2. Sara Laghmati, Machine Learning based System for Prediction of Breast Cancer Severity, 2018.
3. Dana Bazazch, Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis, 2019.
4. Ayush Sharma, Sudhanshu Kulshrestha, Sibi Daniel, Machine Learning Approaches for Breast Cancer Diagnosis and Prognosis, 2017.
5. PanuwatMekha, NutnichaTeeyasuksaet, Deep Learning Algorithms for Predicting Breast Cancer Based on Tumor Cells, 2019.
6. Zhan Xiang, Breast Cancer Diagnosis from Histopathological Image based on Deep Learning, 2019.
7. Yuqian Li, Classification of Breast Cancer Histology Images Using Multi-Size and Discriminative Patches Based on Deep Learning, 2019.
8. Asoke Nath, Image Denoising algorithms: A comparative study of different filtration approaches used in image restoration, International conference on communication systems and network Technologies, 2013.
9. Sukhjinder Kaur, Noise Types and Various Removal Techniques, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 4, Issue 2, February 2015.
10. H.P. Ng, S.H. Ong, K.W.C. Foong1, P.S. Goh, W.L. Nowinski, Medical Image Segmentation Using K-means Clustering and improved Watershed Algorithm, 2006.
11. K. P. Murphy, Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning. Cambridge, Mass.: MIT Press, 2012.
12. International Agency for Research on Cancer (IARC) and World Health Organization (WHO). GLOBOCAN 2018: Age standardized (World) incidence and mortality rates, breast. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-factsheet.pdf
13. Youness Khourdifi, Mohamed Bahaj: Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, 2018

AUTHORS PROFILE

Kavya K is currently pursuing her M.Tech degree in Software Engineering from M. S. Ramaiah Institute of Technology, Bengaluru. She has completed her B.E. in Computer Science and Engineering from Government Engineering College, Ramanagar. Currently, she is working as an Intern in Ernst and Young (EY), Bengaluru, India.Her research area includes Machine Learning and Deep Learning.

Savita K.Shettyis currently working as a Asst. Professor in M. S. Ramaiah Institute of Technology (MSRIT), Bengaluru and pursuing her Ph.D. in Computer Science and Engineering, Visvesvaraya Technological University, Belagavi, Karnataka. Her research interests include Data Analytics, Data Mining and Machine Learning.