EARLY STAGE DETECTION AND CLASSIFICATION OF BREAST CANCER

C SaiDeepReddy, Yeturi Ram Mohan, ChandanaS, Kavya S, Chaitra P C
Department of CSE,
DayanandaSagarAcademy of Technology and Management
yeturirammohan@gmail.com

Abstract. One of the major diseases that affect young to old aged women in recent times is breast cancer. It almost ranks as the first cause for death in women across the world. The survival rate of people suffering with it ranges somewhere between 40% and 60% depending on the development terms of particular countries. Hence, it becomes quite important to be able to diagnose such a disease at a stage as early as possible, so the patient could look out on the available options for treatment. Therefore, in this project, we propose such a breast cancer detection system which predicts the nature of the cancer, either benign or malignant by processing the mammographic image of the patient. The model basically uses a range of digital image processing techniques and also algorithms of ML in the process to output the prediction. It is trained using the MIAS breast cancer dataset. The input image is first resized, gray-scaled, and a gaussian filter is applied on it to remove background noises. It is then segmented and fed to the neural network, which gives the output prediction as an integer value (each value corresponding to a predicted class). The project also has a second stage where the severity of the cancer is also detected by taking input of other detailed attributes of the mammogram.

Keywords: MIAS, resized, gray-scaled, gaussian filter, segmented, benign, malignant, neural network, predicted class, mammogram.

1 Introduction

As the chartspicturize, one driving reason for death in ladies over the entire world is breastcancer.It is a frequently occurring cancer in women, which actually affects almost 2.1 million women every year. It is surveyed to usually affect women morein the better developed regions, although rates areincreasing in all regions worldwide. Although prevention is not a feasible option, improvement of outcomes and survival of breast cancer is an option that we may consider for betterment of the situation. In order for that improvement, an early stage detection of the cancer becomes quite critical. By detecting it at very early stages,it is conceivable to take into consideration increasingly successful treatment to be utilized which may lessen the danger of death from breast cancer growth. There have been a lot of researchers and scientists work-

ICASISET 2020, May 16-17, Chennai, India

DOI 10.4108/eai.16-5-2020.2304093

ing in this field to come up with techniques to allow for early stage detection of breast cancer.

Since this involves a lot of prediction that has to be accurately done, it is most feasible to use machine learning in this particular field. There are a number of ML techniques and algorithms which are used in such cancer detecting models, also giving a good rate of accuracy since it becomes most important to be very sure of the outcome, keeping in mind medical data. Some of them give a great rate of accuracy but may not be able to handle particular inputs, and some vice-versa. Hence, it becomes important to look at every aspect when choosing a particular technique or algorithm in building a model. Some very important ones are robustness, accuracy, time complexity and so on. By making an analysis of such algorithms keeping all parameters in mind, this paper presents a breast cancer detection model using the MIAS dataset to train the model.

The MIAS is a research group in the United Kingdom, which interests itself in studying mammogram images from which they generated a digital database out of the research done. The dataset consists of 322 breast images from 161 patients. This particular dataset is used to train the model to output a proper prediction. The images are first pre-processed using a number of digital image processing techniques after which they are passed on to the neural network. The first stage the image undergoes is resizing, done to ensure all images are of same resolution. The resized image is then grayscaled and a gaussian filter is applied to remove all the background noise that may be present in the image. The smooth filtered image is then passed on for segmentation done by binary conversion. The further steps involve feature extraction and classification. The classification is done using a K-NN classifier, outputting the nature of the cancer. Furthermore, the detailed attributes of the cancer are also used in the project to predict the degree of severity of the cancer using a convolutional neural network.

2 Literature Review

[1] represents the breast cancer detection system that detects and classifies abnormalities or cancerous tissue region in mammograms. This system uses several image processing techniques like pre-processing by using median filter, cropping, segmentation by using Otsu’s thresholding, feature extraction by GLCM and also uses machine learning algorithms like K-Nearest Neighbours formammogramclassification into normal, benign and malignant.Image Database, Image pre-processing, Image segmentation, feature extraction and classification of an Image are the key five stages in this system. From the first stage i.e., Image Database they have used mini – MIAS database which consists of three types of breast images: normal, benign and malignant from 161 patients. The second stage is Image pre-processing where the background noise which is irrelevant and unwanted is being removed by using median filter.The third stage in this system in Image segmentation where the image is partitioned into several meaningful constituent components. This is done by using a technique called

Otsu’s thresholding technique, which separates background image from foreground image by setting up a threshold value. The following stage is feature extraction where they used two types of texture features: 1storder texture features (mean, skewness, standard deviation) and 2ndorder GLCM texture features (correlation, homogeneity, entropy, energy, contrast) are computed. The last stage is classification where k-NN is used as classifier for breast image into normal, benign and malignant. This system achieves 92% of true classification accuracy. [2] represents the breast cancer detection system by classifying the mammograms using texture analysis. For textural pattern analysis of a mammogram breast image we use LBP and LGP techniques. The generated patterns are classified using SVM. The whole system undergoes four stages: Mammogram Image Acquisition, Pre-processing, Texture Image and Classification using SVM classifier. MIAS dataset is used as the input to the system which is for Mammogram Image Acquisition step. The pre-processing stage enhances the image in such way to differentiate the gray levels between the desired objects and also reduces the irrelevant noise without disturbing the important texture features. And then LBP is used to obtain texture features where LBP transforms mammographic image into texture image. This texture image is obtained by comparing the centre pixel of a 3X3 matrix with its neighbouringpixels, ifcentre pixel value is greater than its neighbouring value then the neighbour value is made as 0 otherwise 1. These values are taken in a sequential order to form 8-bit decimal code. All the pixels are considered as centre pixel and decimal codes are generated to form new matrix (texture image). LGP is similar but the average of all pixels is computed and replace the centre pixel with this average value. Then the same comparison should be made as in LBP and the new texture image matrix is obtained. These computed values are given to SVM classifier for further classification of an Image. As a result, the accuracy of 91% is obtained with LBP and accuracy of 95% is obtained with LGP.

[3] mainly throws light upon the use of deep learning techniques for mammogram analysis.Mostly it is a study about the various automatic system- aided classification techniques that can be applied to a mammographic image and the diagnosis can be done at an ease in terms of human effort and accuracy.Deep learning involves concepts such as convolutional neural network which are the most used in modern day applications. The paper puts forth the application of CNN in various mammographic image analysis procedures.There are variations of CNNs that are used in a wide variety of applications, namely RCNN (Region based convolutional Neural Network), Fast RCNN, Faster RCNN and YOLO to name a few.These are specialized ways of applying the neural network to perform segmentation on the processed mammogram image.[2] also describes about some of the top-listed datasets that can be used in the study of breast cancer, such as the mini-MIAS,Digital Database for Screening Mammography (DDSM) and the Mammographic Image Database for Automated Analysis (MIDAS) .There are certain basic steps that may be followed with some variation in every breast cancer detection procedure.The first of them being the mammogram preprocessing, enhancement, segmentation, feature enhancement and classification into

benign or malignant. Lastly, it also talks about how the calcification can be detected and classified by using thresholding.

[4] puts forth an understanding of the different machine learning algorithms which are on top of the charts, used in many applications, the advantages and the disadvantages of using a particular algorithm and the performance metrics that were obtained when all those algorithms were experimented on the Wisconsin Diagnosis Breast Cancer Dataset. The paper mainly focuses on three particular algorithms, namely the Random Forest, the K-Nearest Neighbour and the Naive Bayes algorithm. Some of the performance metrics that it considers to compare the algorithm are accuracy percentage of the prediction, the time complexity, the kind of problems that the algorithm can handle (classification and regression) and some other parameters. The dataset that has been used in this paper has 569 instances totally and also lacks any kind of missing values. The target output is either malignant or benign based on all the variables considered as area mean, perimeter mean, texture main and diagnosis, which are the most influential. The training set had 398 observations and the testing set had 171 observations. All the three algorithms were tested on the dataset and it was inferred from the confusion matrix obtained that the K- Nearest Neighbour algorithm gave the most accurate prediction and was the most efficient of all the three.

In [5] IDSS is developed to diagnose breast cancer. Image pre-processing, Segmentation, Feature extraction, and Classification are four main stages designed in IDSS. In [1] we consider MRI images of a patient as an input. Pre-processing- noise and artefacts are removed leaving only the breast region. Therefore, the values of the objects with greater threshold are eliminated. Weiner filter and clahe filter removes noise while removing the background for enhancing image quality. Segmentation - ROIs from various mammogram images are used to segment by gathering similar pixels of image in a connected region or disconnected region based on the intensity of gray level in K-means algorithm using methods like clustering. Hence, K-means algorithm-based regions of mammogram images are partitioned. DWT and GLCSM are methods in feature extraction for ROIs of a mammogram.

Original image is decomposed into four new sub images where the size of each sub image is one-fourth of the original image. Special distribution attributes based on gray level in a texture image are measured through GLCM. Inputs are taken from the extracted features and produces the output in the form of classification decision. Benign, Malignant, and Normal are the three classes of the output obtained. ANN classifies the mammogram in the proposed IDSS. MIAS is a dataset used for the evaluation ofIDSS with an average accuracy of 96.563% using 10 folds cross validation techniques.

In [6] the correctness of data classification with respect to accuracy, a model and various optimization algorithms are developed where they are feasible for computer aided diagnosis. Wisconsin breast cancer is a dataset with the features of digitized image. Involving data, performing task on the data, choosing the model, Evaluatingthe loss function, Identifying the model's parameter by the learning algorithms and Evaluation the final stage. A DNN is built with sigmoid neuron can classify the task trained where it is implemented on the mammogram datasets.

In [6] three different models are built such as two-layer sigmoid neurons, three-layer sigmoid neurons and four-layer sigmoid neurons. Sigmoid neuron induces a small change in the output with respect to changes in the input. Due to this phenomenon output values ranges from 0 to 1. Optimization algorithms minimize or maximize error function E(x). To minimize the function RMS and SGD optimization algorithms are used. RMS propagation is the gradient descent of weighted average but the only difference is in updating parameters. SGD is linked with a random probability where some samples are selected for each iteration. In [6] 70% data is considered for training and 30% is considered for testing. Training accuracy, testing accuracy, training loss, testing loss for RMS propagation shows various changes but for SGD it approximately shows no change on this model with a sigmoid activation function to minimize the binary cross entropy loss.

[7] represents simple techniques for detection of breast cancer in mammogram. Segmentation, Removal of pectoral muscle and Classification are the three key steps. This system consists of five stages: Median filtering, ROI, removal of pectoral muscle, Feature extraction and SVM based classification. In the first stage i.e., Median filtering is applied to remove all the background noise in the mammogram. The second stage Extraction of breast region it contains two section

1. Generation of binary mask is done through Otsu's thresholding method
2. Connected computing here is to extract the largest component from the binary mask.

The third stage Removal of pectoral muscle a pectoral muscle has highest brightness level when compared to other cancerous tissues so using a canny edge dedication and straight line approximation technique it completely removes all the pectoral muscles in the breast region.

The fourth stage is feature extraction where breast region feature will be extracted from normal and abnormal tissues, feature extraction is performed through GLCM technique which several features will be extracted.

The final stage is SVM based classification the SVM based classifier gives the boundary between positive and negative class features. This system technique is validated on Mini-MIAS database. This can be applied successfully in a CAD system for detection breast cancerous tissues.

[8] represents the detection of breast cancerusing ANFIS - adaptive neuro-fuzzy interference system where the ANFIS is used asclassifier and AR – associative rules technique is basically used as selection of features. The Cuckoo OptimisationAlgorithm(COA) is used to find optimal value of radius. The Wisconsin Breast Cancer Detection (WBCD) which will provide high detection accuracy. Feature selection module, Classification module and Optimisation module are the three key modules. The main goal of this algorithm is to classify whether the cancer is normal, benign or malignant stage.Basically it uses samples of 66% in training phase and 34% for test-

ing proposed methods. The ANFIS and selection features will accurately recognize the tumour type with 99.26% of accuracy.

3 Proposed Methodology

In this project, mainly there are two phases as follows the first phase and the second phase, the first phase is divided into five different modules, they are Image acquisition, Image/data pre-processing, Image segmentation, Feature extraction and Classification.

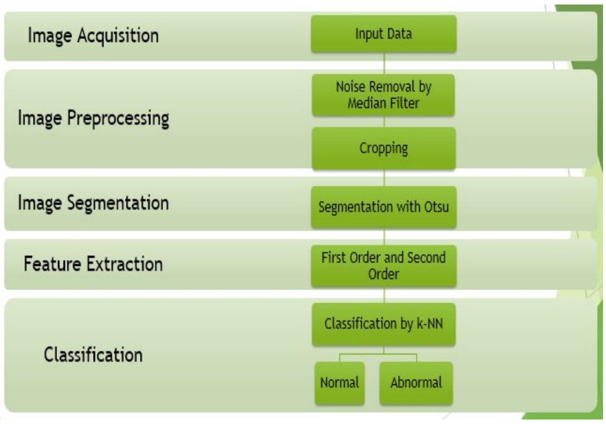


Fig.1. Proposed System Overview

3.1 Image acquisition:

Mini – MIAS database is used as input for this breast cancer detection system. MIAS is nothing but Mammographic Image Analysis Society. It consists of 322 breast images from 161 patients and they are of three types of cancerous tissue: normal, benign and malignant. Of these breast images 208 are normal, 63 are benign & 51 are malignant.

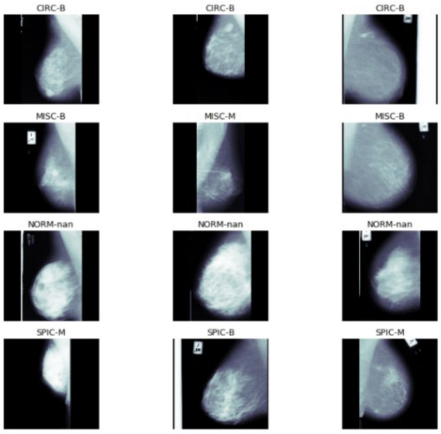


Fig.2. MIAS Dataset

3.2 Image pre-processing:

There are three intermediate stages in pre-processing of an image. They are:

Resizing the Image:.

This is the first stage of pre-processing the image where the images are scaled to a size i.e., suitable for next stage processing. Resizing or distorting an image from one-pixel grid to other is called Image Interpolation. Two categories of Interpolation algorithms adaptive and non-adaptive techniques are used for Image Resizing. In order to increase or decrease total no. of pixels Image Resizing is implemented.

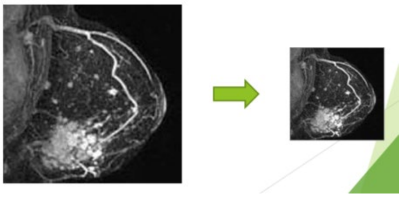


Fig.3. Resizing the Image

Gray Scale Conversion:.

Converting a resized image into a gray scale image i.e.; an image composed exclusively shades of grey. Continuous tone images with unlimited number of shades of gray to an image that a computer can manipulate typically around 16 to 256 levels of intensity is a technique called gray scaling.

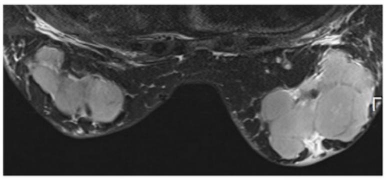


Fig.4. Gray Scale Converted Image

Gaussian Filtering:.

Due to brightness characteristics of a pixel which is a background noise image pre-processing is necessary. It removes noise by Gaussian filter and results in a blur image or smoothed image.Then the image is cropped for further processing.

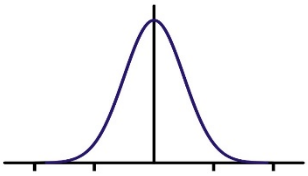


Fig.5. Gaussian Filter

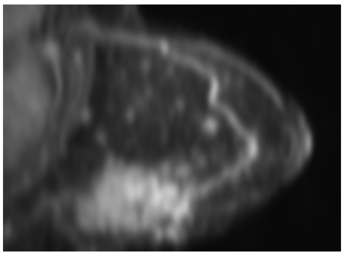


Fig.6. Filtered Image

3.3 Image segmentation:

Image segmentation is a process of partitioning of an image into several constituent components. Here , this stage is divided into two substages- Binary Conversion and Image splitting. Binary conversion is used to segment an image. The algorithm returns a single intensity threshold that separate pixels into two classes mainly foreground and background

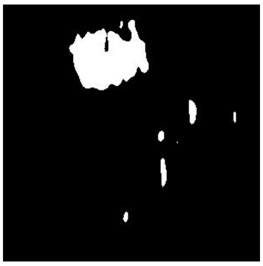


Fig.7. Segmented Image

Splitting up of an Image:.

Here the extracted image is spilt into 16 sub images in-order to rule out the muscle part of the image. After splitting the image, watershed algorithm is used to get the actual white pixel count from the sub images.

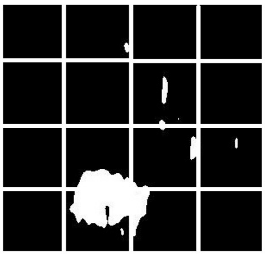


Fig.8. Splitting the Image into 16 sub-images

3.4 Feature extraction:

Feature extraction is a method of capturing visual content of images for indexing and retrieval. We consider two types of texture features for the process of extraction they are 1st order and 2nd order texture features. 1st order texture features are calculated from individual pixel and need not to consider the neighbouring relationship. 2nd order texture features compute the statistical features from the pixel of the neighbouring relationship. Mean, standard deviation, skewness are the examples of 1st order texture features. Energy, entropy, homogeneity, correlation, contrast are the examples of 2nd order texture features.

3.5 Classification:

We use k-NN classifier for the purpose of classifying the image. K-NN classifier works by finding the distances between the query and all the examples in the data, selecting the specified number of examples closest to the query, then votes for the most frequent label. We choose the value of k, by a method known as the elbow method, wherein the process will be repeated for a number of k values, and finally, the one with the least error percentage will be chosen as the optimal k value. This will process and returns an integer value of 0 (Benign) or 1 (Malignant).

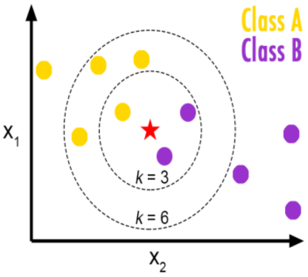
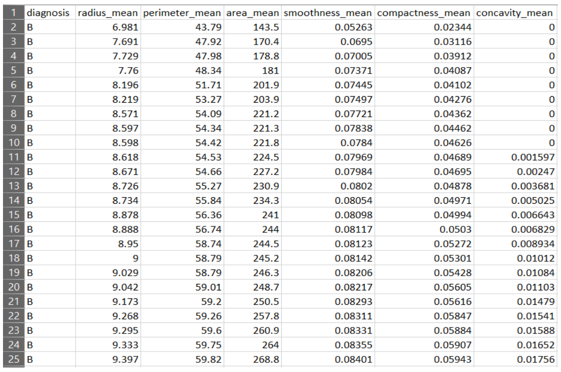


Fig.9. KNN Classifier

Second phase:

After the classification of an image either into benign or malignant, we proceed into further phase. In this phase if an image is labelled as malignant, then we take some other parameters of an image into consideration. Few such parameters are mean, standard deviation, entropy, skewness. By using neural network, we again classify an image into 3 stages they are 1st stage, 2nd stage and 3rd stage, then we suggest the victim to undergo neo-adjuvant chemotherapy.

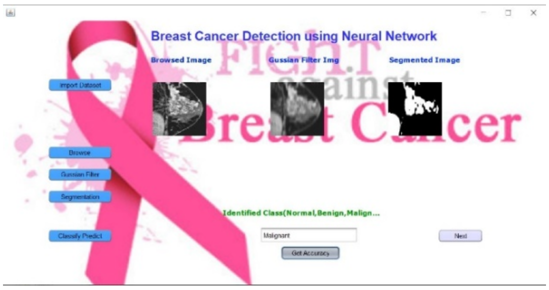


1	diagnosis	radius_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
2	B	6.981	43.79	143.5	0.05263	0.02344	0
3	B	7.691	47.92	170.4	0.0695	0.03116	0
4	B	7.729	47.98	178.8	0.07005	0.03912	0
5	B	7.76	48.34	181	0.07371	0.04087	0
6	B	8.196	51.71	201.9	0.07445	0.04102	0
7	B	8.219	53.27	203.9	0.07497	0.04276	0
8	B	8.571	54.09	221.2	0.07721	0.04362	0
9	B	8.597	54.34	221.3	0.07838	0.04462	0
10	B	8.598	54.42	221.8	0.0784	0.04626	0
11	B	8.618	54.53	224.5	0.07969	0.04689	0.001597
12	B	8.671	54.66	227.2	0.07984	0.04695	0.00247
13	B	8.726	55.27	230.9	0.0802	0.04878	0.003681
14	B	8.734	55.84	234.3	0.08054	0.04971	0.005025
15	B	8.878	56.36	241	0.08098	0.04994	0.006643
16	B	8.888	56.74	244	0.08117	0.0503	0.006829
17	B	8.95	58.74	244.5	0.08123	0.05272	0.008934
18	B	9	58.79	245.2	0.08142	0.05301	0.01012
19	B	9.029	58.79	246.3	0.08206	0.05428	0.01084
20	B	9.042	59.01	248.7	0.08217	0.05605	0.01103
21	B	9.173	59.2	250.5	0.08293	0.05616	0.01479
22	B	9.268	59.26	257.8	0.08311	0.05847	0.01541
23	B	9.295	59.6	260.9	0.08331	0.05884	0.01588
24	B	9.333	59.75	264	0.08355	0.05907	0.01652
25	B	9.397	59.82	268.8	0.08401	0.05943	0.01756

Fig.10. Attributes and Attribute Values that are considered for the Second phase

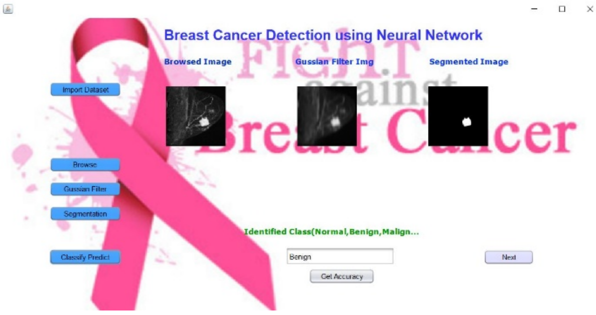
Output:

GUI shows the output of the framework as Fig.11, Fig.12, Fig.13 and Fig.14



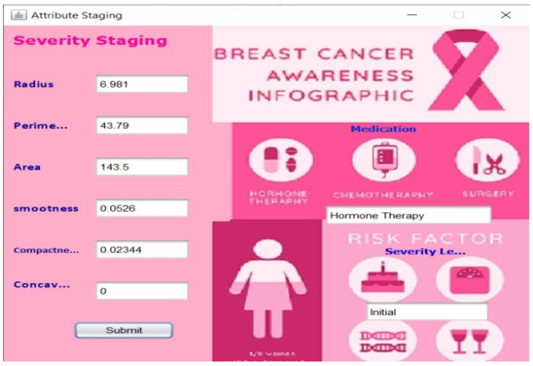
Identified Class(Normal,Benign,Malign...: Malignant

Fig.11. GUI detection system for breast cancer (malignant image)



Identified Class(Normal,Benign,Malign...: Benign

Fig.12. GUI detection system for breast cancer (benign image)



Severity Staging

Radius: 6.981

Perime...: 43.79

Area: 143.5

smootness: 0.0526

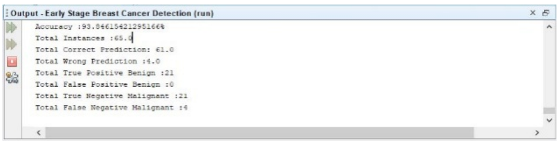
Compactne...: 0.02344

Concav...: 0

Medication: Hormone Therapy

Severity Le...: Initial

Fig.13. GUI detection system for breast cancer Severity Level and Medication Level.



Output - Early Stage Breast Cancer Detection (run)

```
Accuracy :93.84615421295166%
Total Instances :65.0
Total Correct Prediction: 61.0
Total Wrong Prediction :4.0
Total True Positive Benign :21
Total False Positive Benign :0
Total True Negative Malignant :21
Total False Negative Malignant :4
```

Fig.14. Confusion Matrix.

4 Results and Discussion

Different methodologies and processes are implemented on the mini – MIAS database. MIAS database consists of 322 images of breast with 208 normal, 63 benign and 51 malignant images. Among these 322 images we considered 80% of the mini-MIAS dataset for training and 20% for testing i.e; in particular 65 images are taken as a test dataset and remaining 257 images are taken as a train dataset. In the first stage of image pre-processing the imported mammogram image is resized and gray-scaled converted. The pre-processed image is filtered using Gaussian low pass linear digital filter removes the background noise of the image. The obtained image is segmented by dividing the filtered image into several meaningful components. The segmented image is split into 16 sub images to get rid of the muscle tissue in the segmented breast image. The segmentation process extracts the ROI in a pre-processed breast image.

The next stage performed here is extraction of 1st order and 2nd order features. For example, mean, standard deviation, skewness are some of the 1st order statistical texture features and correlation, homogeneity, entropy, energy, contrast are some of the GLCM texture features that are computed. These extracted features are classified using K-NN classifier. This Classification classifies the breast image into two categories either benign (0) or malignant (1). Hence the model predicts the results at an accuracy of 93.8% along with the confusion matrix. Total Instances, Total Correct Predictions, Total Wrong Predictions, Total True Positive Benign, Total False Positive Benign, Total True Negative Malignant, Total False Negative Malignant are incorporated in the Confusion Matrix.

Finally, after classifying the image whether it is Benign, Malignant or Normal from first phase. In second phase data from MIAS which consists of Attributes and Attribute values as shown in Fig.10. This is considered to recommend the severity level and Medication as shown in Fig.13.

5 Conclusion

As discussed in the above modules, a variety of machine learning algorithms and techniques were assessed for their performance on breast cancer data sets and their results were analysedusing machine learning algorithms in order to predict cancer provides a cutting edge in modern healthcare as it is built so as to give the maximum amount of accuracy, which is utmost important in the healthcare sector.Early stage detection of breast cancers increases the rate of survival of patients immensely, approximately being more than 90%. The detected type of cancer is then diagnosed for the kind of treatment, deciding whether it has to be treated first with neoadjuvant chemotherapy and then sent to surgery or otherwise and the pathological complete remission is monitored for the cancer to be completely healed by the chemotherapy in few cases.

References

1. Than ThanHtay, Su SuMaung, ‘Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image’, ISCIT 2018.
2. NarainPonraj, Poongodi, Merlin Mercy, ‘Texture Analysis of Mammogram for the Detection of Breast Cancer using LBP and LGP: A Comparison’, IEEE 2016.
3. OinamVivek Singh, Dr. PrakashChoudhary,‘A Study on Convolution Neural Network for Breast Cancer Detection’, ICACCP 2019.
4. Shubham Sharma, ArchitAggarwal, TanupriyaChoudhury, ‘Breast Cancer Detection Using Machine Learning Algorithms’,IEEE 2018.
5. HussianAlSalman, NajiahAlmutairi, ‘IDSS: An Intelligent Decision Support System for Breast Cancer Diagnosis’, IEEE 2019.
6. NagadeviDarapureddy, Dr. NagaprekeshKaratapu, Dr. Tirumala Krishna Battula, ‘Implementation of optimization algorithms on Wisconsin Breast cancer dataset using deep neural network’, IEEE-2019.
7. Abdul Qayyum, A. Basit, ‘Automatic based segmentation and cancer dedication via SVM in mammograms’, IEEE 2016.
8. Payamzarbakhsh,Abdoljaliaddeh,HasanDermiral, ‘Early detection of breast cancer using optimized ANFIS and features selection’, IEEE 2017.
9. R. Sangeetha, K. Srikanta Murthy, ‘A Novel Approach for Detection of Breast Cancer at an early stage by Identification of Breast Asymmetry and Microcalcification cancer cells using Digital Image Processing techniques’, IEEE 2017.
10. Ding-Horng Chen and Yi-Chen Chang, Pai-Jun Huang, Chia-Hung Wei, ‘The Correlation Analysis between Breast Density and Cancer Risk Factor in Breast MRI Images’, IEEE 2013.
11. S. T. Ahmed, "A study on multi objective optimal clustering techniques for medical datasets," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp. 174-177, doi: 10.1109/ICCONS.2017.8250704.
12. S. T. Ahmed and K. K. Patil, "An investigative study on motifs extracted features on real time big-data signals," 2016 International Conference on Emerging Technological Trends (ICETT), Kollam, 2016, pp. 1-4, doi: 10.1109/ICETT.2016.7873721.
13. K. D. Singh and S. T. Ahmed, "Systematic Linear Word String Recognition and Evaluation Technique," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 0545-0548, doi: 10.1109/ICCSP48568.2020.9182044
14. Gunashree, M., Ahmed, S. T., Sindhuja, M., Bhumika, P., Anusha, B., &Ishwarya, B. (2020). A New Approach of Multilevel Unsupervised Clustering for Detecting Replication Level in Large Image Set. Procedia Computer Science, 171, 1624-1633. https://doi.org/10.1016/j.procs.2020.04.174