International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Breast Cancer Classification Model to Reduce the Recall Rate of a Breast Cancer Screening

Akshay Laxmish 1, Anoop K N 1, Amit B V 1, Shiva shankar N M 1, Mr. Saravana M K 2

[1]Student, Department of Computer Science Engineering, Jyothy Institute of Technology, Bangalore, India.
[2]Associate Professor, Department of Computer Science Engineering, Jyothy Institute of Technology, Bangalore, India.

ABSTRACT: This project briefs the tests performed to improve the recall rate for breast cancer detection. Different methods were applied and optimizations performed on these methods followed by a comparing the results. Some methods had to be ignored because of very low accuracy. Our final proposed methodology associates techniques such as GLCM feature extraction and wavelet transform to find both mass lesions and micro calcifications with the same accuracy as any of the methods used alone. Also, the feature reduction techniques applied helps in training the model more efficiently, while reducing the training time and keeping the same accuracy.

1. INTRODUCTION

Cancers are a bulk family of diseases that includes abnormal cell growth with the capability to circulate to other parts of the body. Breast cancer is a class of cancer that develops from the breast tissue. Compared to other diseases or other cancers, breast cancer takes a proportionately greater share of finances and concentration. Because of its relatively high frequency and long-term survival rates, exploration is partial towards breast cancer.

Breast cancer is the most constantly found solid cancer and second leading reason of cancer death among U.S. women [3]. Based on collective random trials, screening mammography has been shown to lower breast cancer-related deaths[4]. However, despite this population-based benefit, routine mammography is connected with a extreme threat of false positive testing and may guide to over diagnosis of clinically insignificant lesions [4]. According to the federally-funded Breast Cancer Surveillance Consortium, the overall keenness of digital mammography in the U.S. screening population is 84%, and the overall specificity is 91%. [5] What the Unites States has accomplished today (a decrease in mortality, even with an increase in numbers of women diagnosed with breast cancer) has taken many decades of untiring, persistent efforts. And they stated that, when their death rate was not so high at all. So, in India, the death rate is 70,000 and increasing. Since more patients (in India) turn up in next stages, they do not survive long irrespective of the best treatment they may get, and hence the death rate is fairly high.

2. LITERATURE SURVEY

2.1 REVIEW PAPER ON CLASSIFICATION OF MAMMOGRAPHY [8]

Concepts Introduced:

The efficiency of dimension reduction and normal distribution transformation in enhancing the accuracy of classification has been assessed. The difference in performance of the SVM classifier and the naïve Bayesian classifier was not statistically significant after the transformation. It can exclude a dimension that is good for discriminating positive cases from negative cases and this unsupervised dimension reduction algorithm improved the classification accuracy.

2.2 A SURVEY OF COMPUTER-AIDED DETECTION OF BREAST CANCER WITH MAMMOGRAPHY (2016) [10]

Concepts Introduced:

This paper referes a technique of using mass to help classify benign and malignant mass and microcalcifications. It addresses the concept of Content-based Image Retrieval (CBIR) – picking similar images from the database. This is to assist the radiologist for more better understanding. It may be not applicable to the challenge but it has a nice additional feature to have on board.

2.3 A SURVEY OF COMPUTER-AIDED DETECTION OF BREAST CANCER WITH MAMMOGRAPHY (2016) [10]

Concepts Introduced:

This paper emphasis on the relevance of choosing the best features for enhancement of classification accuracy and therefore investigates different selection methods for mass classification. All the methods investigated are F-score, support vector machine (SVM)-based recursive feature elimination (SVM-RFE) and the accuracy obtained by using a SVM classifier using the features extracted are presented. A fuzzy c-means (FCM) with spatial information constraint is supported that can be integrated into the proposed segmentation method in order to reduce labour cost.

2.4 A NEW FEATURE EXTRACTION FRAMEWORK BASED ON WAVELETS FOR BREAST CANCER DIAGNOSIS [20]

Concepts Introduced:

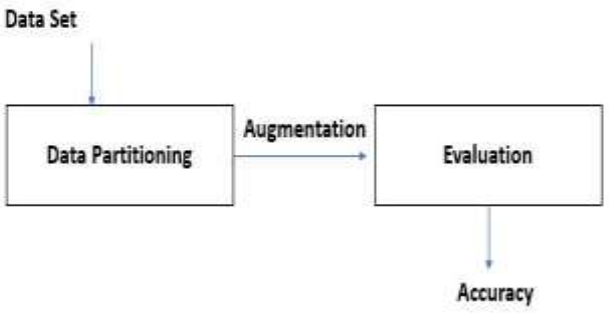
A two class classification study to recognize normal and cancerous breast tissues is conducted. The rotational and scale invariant features are extracted by HOG, DSIFT and LCP descriptors followed by a classification the utilizes SVM, k-NN, Decision trees via 10 fold cross validation. The exact procedure was conducted for a three class classification study and the accuracy was not suitably satisfied. An addition to this study was introduced by applying NLM filter to all the images beforehand.

2.5 COMPUTER-AIDED DETECTION AND CLASSIFICATION OF MICRO-CALCIFICATIONS IN MAMMOGRAMS: A SURVEY [9]

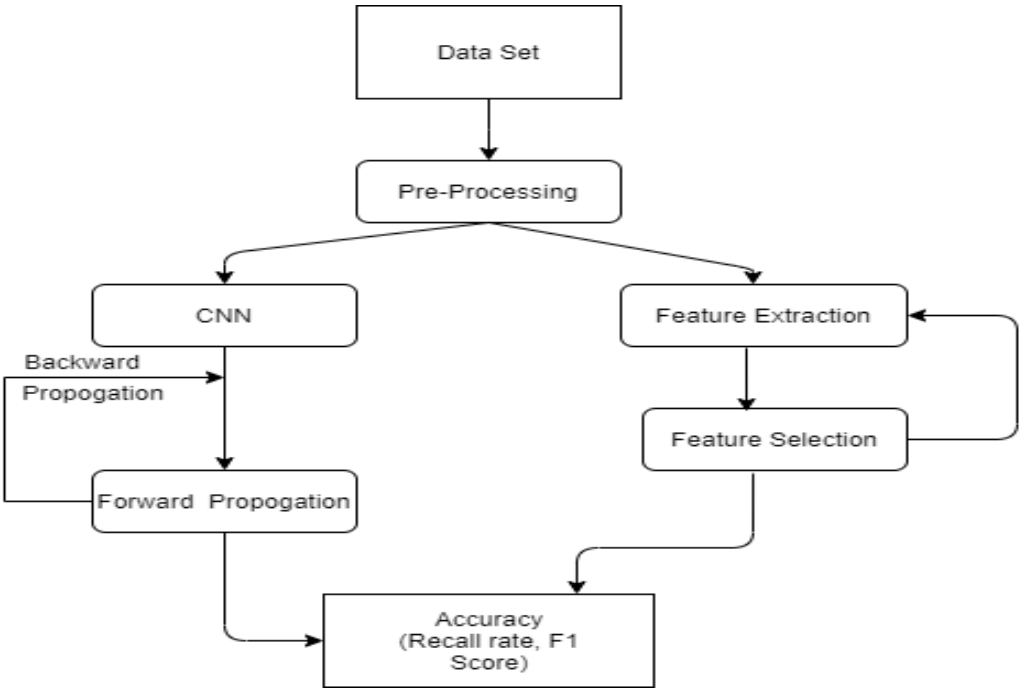
Concepts Introduced:

A complete study of different methods for enhancement of micro calcification such as conventional enhancement techniques, contrast stretching, histogram equalization method, convolution mask enhancement, region or feature based enhancement with the evaluation of the algorithms has been presented. Comparing various segmentation methods such as statistical method, region based approach, mathematical morphology, multiscale analysis and fuzzy approached along with their advantages and disadvantages have been presented. A complete study of micro classification discovery on its features or statistical texture feature ,detected by template matching, Gray level run length method (GLRLM), Gray level difference method (GLDM), Wavelet based method etc,. A complete amalgamation of results from the implementation of algorithms designed for different features from various sources has been categorized.

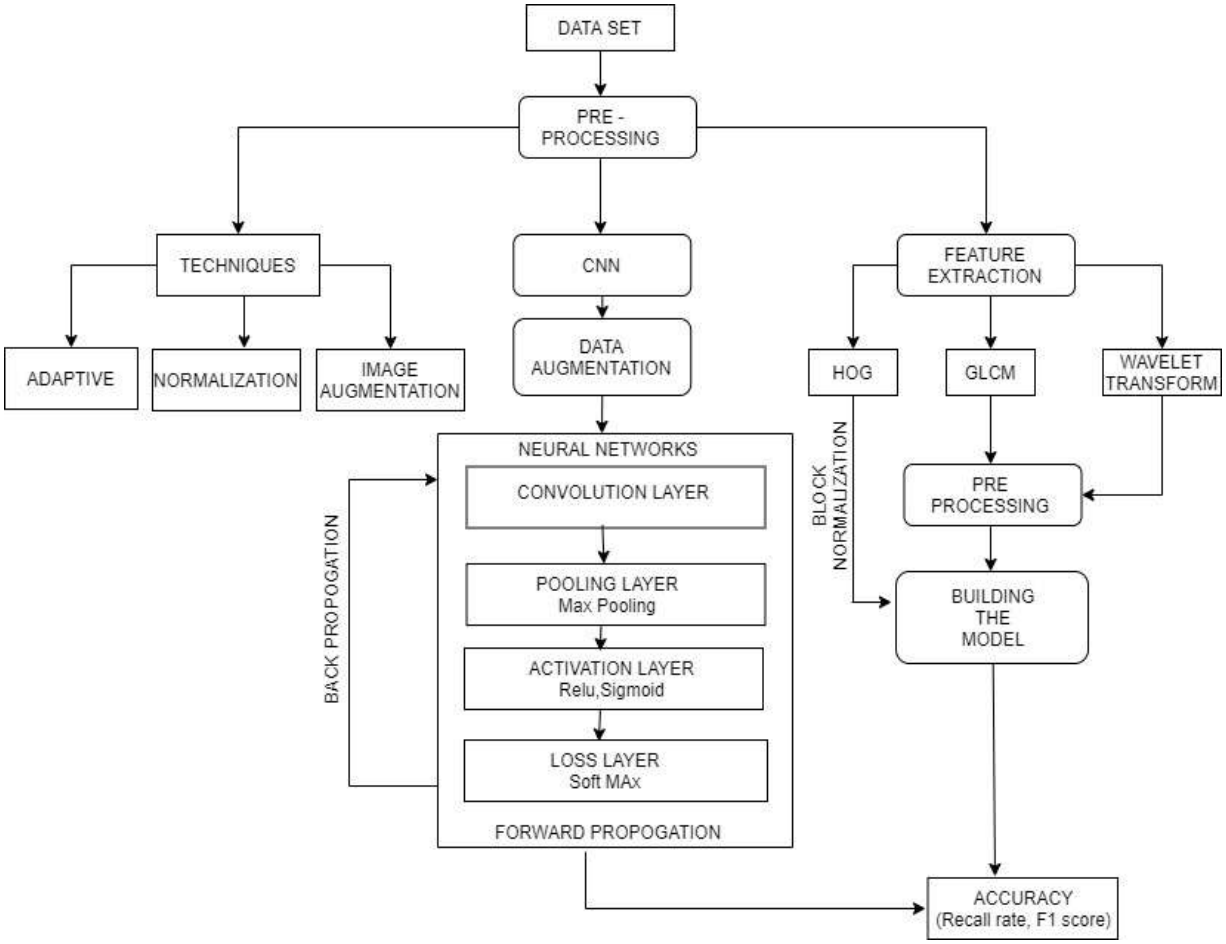
3. SYSTEM ARCHITECTURE



DFD 0


Data Set
Pre-Processing
CNN
Feature Extraction
Backward
Propogation
Feature Selection
Forward Propogation
Accuracy
(Recall rate, F1
Score)

DFD 1


DATA SET
PRE -
PROCESSING
TECHNIQUES
CNN
FEATURE
EXTRACTION
ADAPTIVE
NORMALIZATION
IMAGE
AUGMENTATION
DATA
AUGMENTATION
HOG
GLCM
WAVELET
TRANSFORM
NEURAL NETWORKS
CONVOLUTION LAYER
POOLING LAYER
Max Pooling
ACTIVATION LAYER
Relu,Sigmoid
LOSS LAYER
Soft MAx
FORWARD PROPOGATION
BACK PROPOGATION
BLOCK
NORMALIZATION
PRE
PROCESSING
BUILDING
THE
MODEL
ACCURACY
(Recall rate, F1 score)

DFD 2

4. IMPLEMENTATION

STUDY 4.1: CNN

Initialisation:

Our dataset consists of images in the DICOM format. So, the first step was to convert the dataset into either .jpg or .png formats. So We have used the'mogrify' command obtainable in the Linux to convert the .dcm images to .png images. png conversions are lossless conversions, thus they are chosen for image processing and analysis.

The command used:

mogrify --format .png *.dcm

Preprocessing:

The images were then resized to 150 pixels x 150 pixels and stored in a separate directory. Then keras is used to read these image into the dataset, X. The corresponding ground truth values were also read into Y.

Building the model:

In the next step, X and Y were to split data into test data and train data. The splitting is a 80:20 split applying the 'test_train_split()' function provided by the 'sklearn' library. They were then transformed to float and normalized between 0 and 1. This was to decrease complexities while performing calculations. X and Y were also transformed to categorical data to define that there is no intrinsic ordering between them. Converting X and Y into categorical also enhances the dimensionality, greatly affecting the training time and testing time of the model. The most important step was the build the CNN architecture with the hepl of the 'sequential()' model provided by the 'keras' library. The actual layers and their parameters like mask size, number of filters, stride, etc. can be manipulated.

1. **Model 1**

Layer	1	2	3	4	5	6
Stage	conv+relu+max	conv+relu+max	dropout	full+relu	dropout	full+softmax
No. of channels	32	32	-	256	-	2
Filter size	5x5	5x5	-	-	-	-
Pooling size	2x2	2x2	-	-	-	-
Dropout factor	-	-	0.5	-	0.5	-

50	10

2. **Model 2**

Layer	1	2	3	4	5	6	7	8
Stage	conv+relu	conv+relu+max	conv+relu+max	conv+relu+max	conv+relu+max	full	full	full+softmax
No. of channels	32	64	128	256	512	256	64	2
Filter size	3x3	32x3	32x3	64x3	64x3	-	-	-
Pooling size	-	2x2	2x2	2x2	2x2	-	-	-
Pooling stride	-	2	2	2	2	-	-	-
Dropout factor	-	0.5	0.5	0.5	0.5	-	-	-

10	10

3. Model 3

Layer	1	2	3	4	5	6	7
Stage	conv+relu+max	conv+relu+max	conv+relu+max	conv+relu+max	dropout	full+relu	full
No. of channels	16	16	16	16	-	128	2
Filter size	7x7	5x5	5x5	5x5	-	-	-
Pooling size	3x3	3x3	3x3	3x3	-	-	-
Pooling stride	2	2	2	2			
Dropout factor	-	-	-	-	0.5	-	-

20	10

The images were given as an input to the architecture for training. The CNN performs back-propagation and sets the weights of the various convolution filters accordingly. The training stage also needs many parameters to be set like:

The batch size: Number of images given at a time to the GPU/CPU for processing.

Number of epochs: Number of times each image is taken for training.

Specification of the validation (test) data.

The model is the tested on the test data for validation.

STUDY4. 2: CNN with Augmentation

Augmentation:

Data augmentation was needed as:

a. Only 500 samples were provided for training purposes.

b. The data was heavily skewed. Only 34 out of the 500 samples were that of the positive class.

Various augmentation approaches are usable in the Keras library such as shear, rotation, zoom, etc. These techniques were used on the positive samples to the dataset size from 500 to 1666. This dataset is used for other computations.

Preprocessing:

The images were first resized to 150 pixels x 150 pixels, then adaptive histogram equalization was applied to each of these images in order to enhance the contrast.

Building the model:

This data was then provisoned to the CNN architecture. The process was tried out on the three different models as described in the previous section. Model 1 gave the best performance of the three.

Study4. 3: Feature Extraction Using HOG

Preprocessing:

The same preprocessing techniques that were used in the CNN phase were used in all the other phases. The augmented images were first resized to 150x150 and then adaptive histogram equalization was applied.

Feature Extraction:

All feature extraction phases were performed in MATLAB and the corresponding feature matrices were used to train a classifier. 'VL_FEAT', which is an external library for MATLAB was employed to obtain the HOG of an image. 'vl_hog' takes the image, the number of orientations and cell size as the input and produces a HOG matrix.

The HOG feature extraction algorithm was applied on the mammograms with 18 different orientations for a block size of 16x16. The HOG variant utilized for this extraction process was UoCTTI (University of Chicago Toyota Technological Institute) [1] , and tensors with (8x8x58) size were attained for each tissue patch. These tensors were then reshaped into the matrices with a size of (8x464). Six different time-domain features (energy, mean, standard deviation, maximum, skewness, and kurtosis) [2] were extracted from the columns of these matrices so that matrices with a size of (8x6) are attained. These matrices were converted into column vectors. [3] Thus, we extracted 48 attributes for each image.This technique was linked to all images to get the feature matrix.

Building the model:

The feature matrix obtained from the feature extraction phase was read into X. The class labels were read into Y. Y was then converted to categorical. A decision tree was then trained using this X and Y.

Study4.4: Feature Extraction Using GLCM

Preprocessing:

The augmented images were resized to 150x15 and then adaptive histogram equalization was applied.

Feature Extraction:

We extract the texture features using the GLCM from the images in this case. MATLAB provides an in-built function 'graycomatrix' that generated the GLCM of the images. The GLCM was obtained for 4 different orientations: 0, 45, 90 and 135. Then the following are extracted from the GLCM:

a. autocorrelation
b. contrast
c. correlation
d. cluster prominence
e. cluster shade
f. energy
g. entropy
h. homogeneity
i. maximum probability
j. sum of squares / sum of average / sum of variance
k. sum of entropy
l. difference in variance
m. difference in entropy
n. information measure of correlation

Building the model:

The feature matrix attained from the feature extraction phase was read into X. The corresponding class labels were read into Y. Y is then converted to categorical. Before training the decision tree, another optimisation is done in order to enhance the training time, i.e. reducing the dimensionality of X. The correlation between the columns (attributes) was calculated by using the Chi-square method, as the attributes are categorical. The top 5 non-co-related attributes were chosen. Thus, the dimensionality was reduced from 18 to 5. A decision tree was then trained using these attributes and Y.

The runtime complexity of constructing a binary decision tree is mnlog(n), where m is the amount of features and n is the amount of samples[8].The runtime is directly proportional to the number of attribute in each image. Since the number of dimensions was reduced by a factor of (18/5) = 3.6, the runtime was improved by a factor of 3.6.

Study 4.5: Feature Extraction Using Wavelet Transform for Microcalcifications

Preprocessing:

The augmented images were resized into v150x150 and then adaptive histogram equalization was applied.

Feature Extraction:

A combination of GLCM and Wavelet Transform features were extracted[5]. 'wavedec2()' function performs the multi-level wavelet transform on images[6]. It takes the level and the wavelet filter as inputs. The number of levels, 'maxlev' is calculated using the function 'wmaxlev()'[7]. The wavelet transform of each image was obtained at each level, i, where 1<=i<=maxlev. For each of the wavelet transform obtained, five time-domain features were extracted: mean, total sum, standard deviation, skewness and kurtosis. This matrix was then augmented with the 18 GLCM features. The final matrix had 23 attributes per image.

Building the model:

The exact same procedure was used as in the previous,i.e. GLCM study

5. RESULTS

5.1 ACCURACY MEASURES

Dice co-efficient [20]

The Dice index is a statistic applied for equating the resemblance of two samples. The Dice is also understood as F1 score or Dice similarity coefficient (DSC). The index is understood by various other names, particularly the Sørensen index or Dice's coefficient. Other alterations consists of the "similarity coefficient" or "index". It is applied in image segmentation, in particular for equating algorithm output against reference masks in medical applications.

Loss function

Loss function is a scalar value that we try to minimize through out the training of the model. There are several loss functions that can be applied to report the loss incurred.

Recall score

The recall is basically the capacity of the classifier to detect all the positive samplesIn pattern identification and information retrieval binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved.. Therefore, recall is the capability of the classifier to detect all the positive samples and thus can be used as a useful accuracy metric in the classification and prediction process.

Neural Network

The training process victimized some variant of the Delta Rule, which takes the inputs, the output, and the actual output and the processing element which initiates with the calculated difference between the genuine outputs and the wanted outputs. Using this violation, connection weights are accelerated in proportion to the voilation times. The elaborate portion of this algorithm is figuring which input gave away the most to a wrong output and how the input must be defined to correct the violation.

5.2 STUDY 1: CNN

	Training Accuracy	Training Loss	Testing Loss
1	93.25	37.13	26.50
2	93.25	108.8	112.8
3	93.25	25.29	26.00

Results of Study 1(CNN)

5.3 STUDY 2: CNN WITH AUGMENTATION

The following was decided from the output of the CNN models:

- loss – total loss encountered in one epoch in the training set
- acc – accuracy results on the training set

A different metric , dice-coefficient , was used to evaluate the robustness of the model. The coefficient is 63.70 and the validation loss stands at 60.87.

5.4 STUDY 5: WAVELET TRANSFORM FOR MICROCALCIFICATION DETECTION

Accuracy: 88.94. Recall score: 91.67

Model	Dice Coefficient	Accuracy
CNN Arch 1		93
CNN Arch 1		93
CNN Arch 3		93
CNN with Augmentation	63.7	
HOG		93.6
GLCM		95.37. Recall rate: 99.18
Wavelet transform		88.94. Recall rate: 91.67

Comparison of all studies

5.5 STUDY 3-5

The following table briefs the results obtained by manually extracting features for mass lesion detection:

	Accuracy	Recall Score
Hog	93.86	99.25
Glcm	95.37	99.18

Results of studies 2-5

CONCLUSIONS

This project helped in assuring a foundation for us and provided us with the required motivation to venture into various new, creative and interesting techniques in the broad field of image processing, especifically digital mammography. It encouraged us to learn about the techniques in image pre-processing, machine learning, feature extraction, feature selection

and best ways to consolidate them. We also learnt the utility, advantages and limitations of various techniques like convolutional neural networks, decision trees, SVMs, augmentation techniques, etc.

We found that the results obtained from our current implementation are good and can play a dynamic role in assisting radiologists detect breast cancer at a much earlier stage. Our methodology enhances upon the training time resulting in better efficiency, while also maintaining the same accuracy along with reduction in unnecessary recall of patients. This can greatly help reduce anxiety and potential morbidity, in case of patients who are called for retesting, and also to reduce testing costs.

This project made us to realise the fundamental importance of science and technology and how the different principles presented can be modelled for the improvement of the society. This project involved an inter-disciplinary effort taken towards the mutual enhancement of medical and engineering approaches.

REFERENCES

[1] Brown, Anthony, "Cancer bias puts breasts first". The Guardian. London (Oct 2001).

[2] Arnst, Catherin, "A Gender Gap in Cancer". Bloomberg Businessweek. ISSN 00077135 (June,2007).

[3] American Cancer Society: http://www.cancer.org/cancer/breastcancer/detailedguide/breastcancer-key-statistics

[4] https://en.wikipedia.org/wiki/Convolutional_neural_network

[5] Browne, and Saeed Shiry Ghidary. "Convolutional neural networks for image processing: an application in robot vision." Australasian Joint Conference on Artificial Intelligence. Springer Berlin Heidelberg, 2003.

[6] http://cs231n.github.io/convolutional-networks/

[7] https://www.tensorflow.org/versions/r0.11/tutorials/deep_cnn/index.html

[8] Graham, Benjamin (2014-12-18). "Fractional Max-Pooling". arXiv:1412.6071 [cs.CV].

[9] https://github.com/fchollet/keras/wiki

[10] http://www.breastcancerindia.net/statistics/stat_global.html