International Journals of Advanced Research in Computer Science and Software Engineering
ISSN: 2277-128X (Volume-8, Issue-10)

Research Article | **October 2018**

Analysis of Breast Cancer Cell Based on Hyperchromatic Crowded Group Using Multiple Techniques of Data Mining

Chayanika Sarmah
M.Tech, Department of Information Technology,
Gauhati University, Guwahati, Assam, India
prayashi.pompi@gmail.com

Abstract: In this paper, an effort is made to analyze MGG stain images of breast cell in FNAC which will help in early detection of malignant breast cancer cell. There are many morphological features based on which MGG stain test smear images can be categorized into normal and abnormal classes. Some of them are area, perimeter and presence of hyperchromatic crowded (HCG) group etc. So, in this approach we analyse the breast cell based on a presence in the malignant cell using multiple techniques of data mining (Images are examined). The proposed approach is implemented in WEKA, a java based data mining tool

Keywords: breast cancer, HCG, MGG, FNAC.

I. INTRODUCTION

Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues. Cancer cells can spread to other parts of the body through the blood and lymph systems [1].Cancer is not just one disease but many diseases. There are more than 100 different types of cancer. Most cancers are named for the organ or type of cell in which they start - for example, cancer that begins in the colon is called colon cancer; cancer that begins in melano cytes of the skin is called melanoma.Cancer types can be grouped into broader categories. The main categories of cancer include:

- **Carcinoma** - cancer that begins in the skin or in tissues that line or cover internal organs. There are a number of subtypes of carcinoma, including adeno carcinoma, basal cell carcinoma, squamous cell carcinoma, and transitional cell carcinoma.
- **Sarcoma** - cancer that begins in bone, cartilage, fat, muscle, blood vessels, or other connective or supportive tissue.
- **Leukemia** - cancer that starts in blood-forming tissue such as the bone marrow and causes large numbers of abnormal blood cells to be produced and enter the blood.
- **Lymphoma and myeloma** - cancers that begin in the cells of the immune system.
- **Central nervous system cancers** - cancers that begin in the tissues of the brain and spinal cord.

Origins of Cancer

All cancers begin in cells, the body's basic unit of life. To understand cancer, it's helpful to know what happens when normal cells become cancer cells.

The body is made up of many types of cells. These cells grow and divide in a controlled way to produce more cells as they are needed to keep the body healthy. When cells become old or damaged, they die and are replaced with new cells. However, sometimes this orderly process goes wrong. The genetic material (DNA) of a cell can become damaged or changed, producing mutations that affect normal cell growth and division. When this happens, cells do not die when they should and new cells form when the body does not need them. The extra cells may form a mass of tissue called a tumor. One the methods to identify breast cancer which is used more than the others, is Mammography. But it is frequently seen that different interpretation of radiologists about images is obtained from this way. Another method is Fine needle aspiration cytology [2] (FNAC) and its accuracy is 90%. Therefore, it is better to discover another accurate method.

May-Grunwald-Giemsa(MGG) [3] staining method is used for morphological Counting of blood cells. May-Grünwald test staining combines effect of acidic and alkaline methylene blue. The pH is important factor in staining, so any change will lead to a wrong staining reaction. The limits of most suitable pH are in between 6.5 and 6.8. In this paper an approach is made to develop an automated system for detection and segmentation of abnormal breast cells using the help of MGG stain images.

In our approach we try to classify the breast cells into normal and abnormal classes based on analysis of HCG features.

II. OVERVIEW OF BREAST CANCER

Definition of breast cancer: Cancer that forms in tissues of the breast. The most common type of breast cancer [4] is ductal carcinoma, which begins in the lining of the milk ducts (thin tubes that carry milk from the lobules of the breast to the nipple). Another type of breast cancer is lobular carcinoma, which begins in the lobules (milk glands) of the breast. Invasive breast cancer is breast cancer that has spread from where it began in the breast ducts or lobules to surrounding normal tissue. Breast cancer occurs in both men and women, although male breast cancer is rare.

Estimated new cases and deaths from breast cancer in the United States in 2014:

- New cases: 232,670 (female); 2,360 (male)
- Deaths: 40,000 (female); 430 (male)

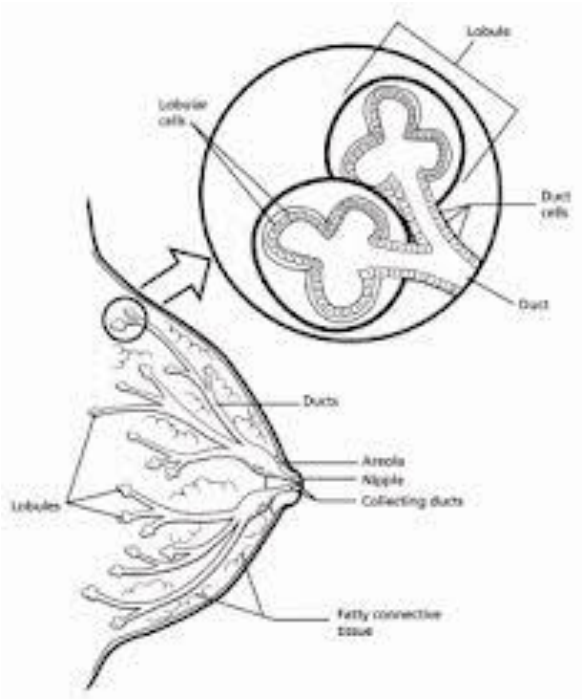

Fig 1

The first symptom of breast cancer that can be noticed is a lump that feels different from the rest of breast tissue [5]. 80% of breast cancer cases are discovered when a woman feels a lump. Earliest breast cancers can detected by a mammogram. Lumps found in the lymph located in armpits can also indicate breast cancer [6].

III. PROPOSED METHOD

The input of the image is taken in .jpg format. The images Sample collected from **Dr B. Barooh Cancer Institute, Guwahati** –

Normal sample = 14

Abnormal sample = 34

Above MGG stain samples are categorized into two main classes Normal and Abnormal.

We analyzed 200 MGG stain samples and tried to draw a conclusion on how a threshold can be generated which will help in distinguishing the Normal Classes and Abnormal Classes. This identified threshold can be fitted in an automatic system to identify the percentage abnormality of a MGG stain test sample. But the quality of the Images matters in this approach. MGG stain images collected from liquid based technology is better for our approach.

3.1 Proposed method for Hyperchromasia identification

For detecting the hyper chromatic Crowded Groups following measures are taken into consideration.

1. Select the image for which HCG to be identified.
2. Crop the particular region of interest (ROI), i.e. the portion which is dark and crowded.
3. Find the number of cells inside the ROI.
4. If number_of_cells>15 then “Hyperchromasia is present” and if <15 then “Hyperchromasia is absent”.

Presence of Hyperchromasia leads to abnormality in breast cells.

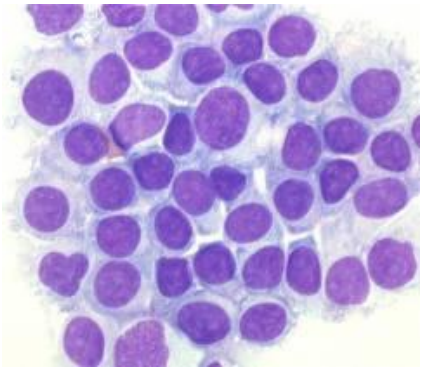

Fig 2 Above figure shows the presence of Hyperchromatic Crowded Group (HCG).

IV. EXPERIMENTAL RESULTS

The breast cancer database comes from the Dr. B.Barooah Cancer Institute and contains real observations of 200 pathological instances. The conditional attributes describe information gained from the digitalized images of the breast mass. Datasets reveals that there are 4 classes of attribute of breast cancer data to prove the Hyperchromatic crowded group. Breast cancer datasets in arff (attribute relation file format).

No of instances: **200**
No of benign cases: **63**
No of malignant cases: **137**
No of attribute: **4**

@relation breast_cancer_2013
@attribute no_of_cell numeric
@attribute image_no numeric
@attribute HCG {Present,Absent}
@attribute class {Benign,Malignant}

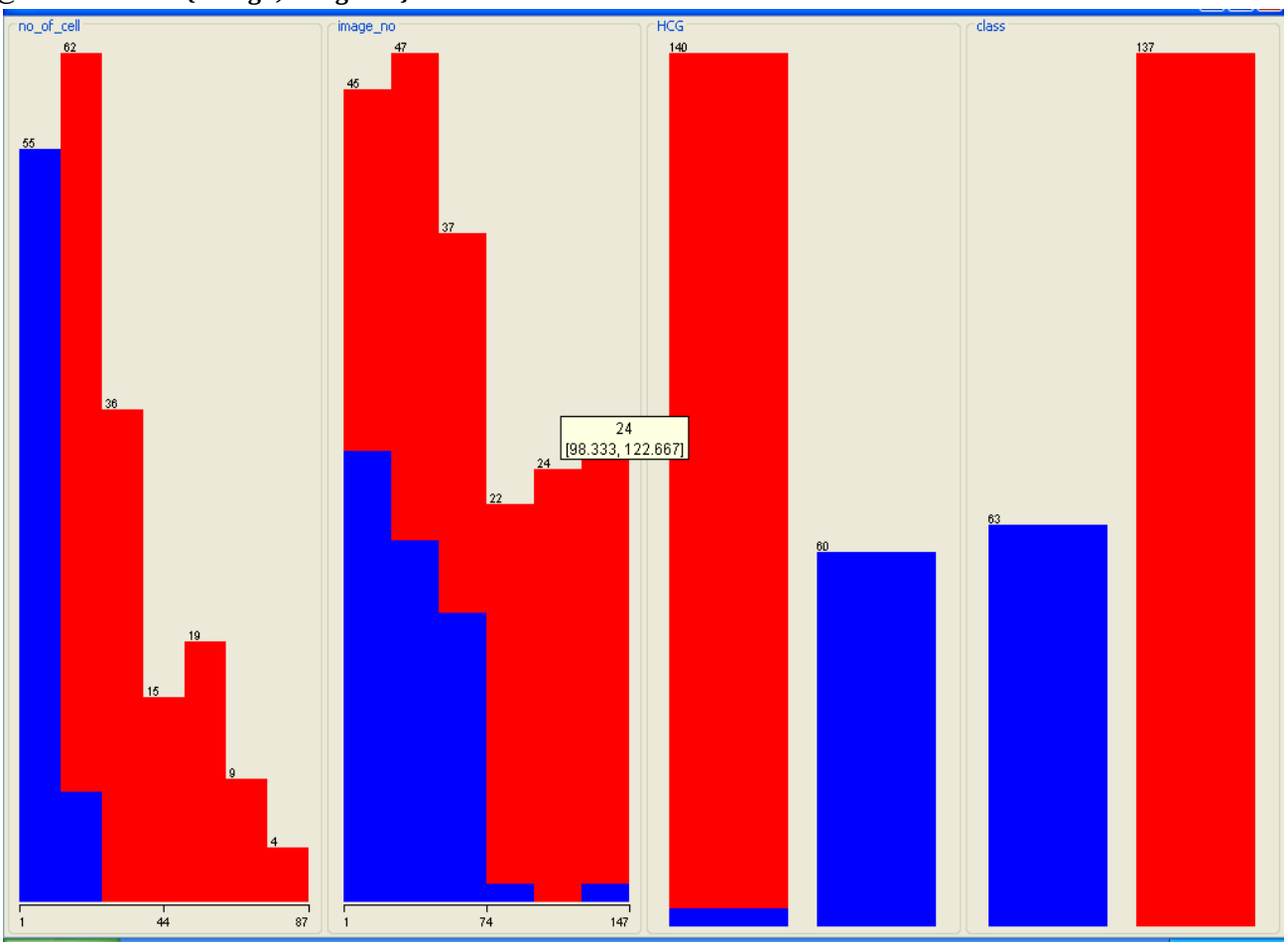

Fig 3 graphical representation of benign and malignant case based on Hyperchromasia

4.1 Classification - 10 fold cross validation on breast-cancer dataset

First we use the data mining tools WEKA to do the training data prediction. In here, we will use 10 fold cross validation on training data to calculate the machine learning rules their performance. The results are as follows:

4.1.1 Results for: Naive Bayes

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: breast_cancer_2013
Instances: 200
Attributes: 4
no_of_cell
image_no
HCG
class
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

	Class	
Attribute	Benign (0.32)	Malignant (0.68)
no_of_cell		
mean	7.0769	34.5255
std. dev.	5.3113	17.7487
weight sum	63	137
precision	2.2632	2.2632
image_no		
Mean	34.9711	75.0426
std. dev.	23.185	42.4206
Weight sum	63	137
Precision	1.0139	1.0139
HCG		
Present	4.0	138.0
Absent	61.0	1.0
[Total]	65.0	139.0

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	197	98.5 %
Incorrectly Classified Instances	3	1.5 %
Kappa statistic	0.9648	
Mean absolute error	0.0203	
Root mean squared error	0.1207	
Relative absolute error	4.7003 %	
Root relative squared error	25.9829 %	
Total Number of Instances	200	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.952	0	1	0.952	0.976	0.99	Benign
1	0.048	0.979	1	0.989	0.99	Malignant
Weighted Avg.	0.985	0.033	0.985	0.985	0.985	0.99

=== Confusion Matrix ===

```
  a   b   <-- classified as
 60   3 |   a = Benign
  0 137 |   b = Malignant
```

4.1.2 Results for: J48 decision tree (implementation of C4.5)

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: breast_cancer_2013
Instances: 200
Attributes: 4
no_of_cell
image_no
HCG
class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

HCG = Present: Malignant (140.0/3.0)
HCG = Absent: Benign (60.0)

Number of Leaves: 2

Size of the tree: 3

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	197	98.5	%
Incorrectly Classified Instances	3	1.5	%
Kappa statistic	0.9648		
Mean absolute error	0.0295		
Root mean squared error	0.1219		
Relative absolute error	6.8175	%	
Root relative squared error	26.2275	%	
Total Number of Instances	200		

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall F-Measure	ROC	Area	Class
0.952	0	1	0.952	0.976	0.959	Benign
1	0.048	0.979	1	0.989	0.959	Malignant
Weighted Avg.	0.985	0.033	0.985	0.985	0.985	0.959

=== Confusion Matrix ===

```
  a   b   <-- classified as
 60   3 |   a = Benign
  0 137 |   b = Malignant
```

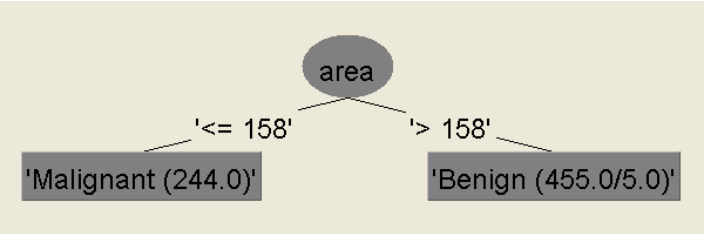

Fig 4 J48 decision tree

4.1.3Results for: JRip (implementation of the RIPPER rule learner)

=== Run information ===

Scheme:weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation: breast_cancer_2013
Instances: 200
Attributes: 4
no_of_cell
image_no
HCG
Class
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

JRIP rules:
==========

(no_of_cell <= 14) => Class=Benign (60.0/0.0)
=> Class=Malignant (140.0/3.0)

Number of Rules: 2

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	197	98.5 %
Incorrectly Classified Instances	3	1.5 %
Kappa statistic	0.9648	
Mean absolute error	0.0295	
Root mean squared error	0.1219	

Relative absolute error	6.8175 %
Root relative squared error	26.2275 %
Total Number of Instances	200

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Area	Class
0.952	0	1	0.952		0.976	0.959	Benign
1	0.048	0.979	1		0.989	0.959	Malignant
Weighted Avg.	0.985	0.033	0.985		0.985	0.985	0.959

=== Confusion Matrix ===

a	b	<-- classified as
60	3	a = Benign
0	137	b = Malignant

4.2 Classification – Compare with test and training data set

Above machine learning tools will use in this section to diagnose cancer dataset. To construct a dataset with 200 instances. The first 121 instances in the dataset are chosen as the training data, and the remaining 79 as the test data.

4.2.1 Results for training data: Naive Bayes

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: breast_cancer_2013
Instances: 121
Attributes: 4
no_of_cell
image_no
HCG
class
Test mode:evaluate on training data

=== Classifier model (full training set) ===

Naive Bayes Classifier

	Class	
Attribute	Benign	Malignant
	(0.28)	(0.72)
no_of_cell		
mean	7.9211	34.5457
std. dev.	5.5232	18.3201
weight sum	34	87
precision	2.2632	2.2632
image_no		
mean	22.2328	48.0225
std. dev.	20.7536	26.8276
weight sum	34	87
precision	1.0215	1.0215
HCG		

Present	3.0	88.0
Absent	33.0	1.0
[total]	36.0	89.0

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances	119	98.3471 %
Incorrectly Classified Instances	2	1.6529 %
Kappa statistic	0.9583	
Mean absolute error	0.0254	
Root mean squared error	0.1267	
Relative absolute error	6.2507 %	
Root relative squared error	28.1847 %	
Total Number of Instances	121	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Area	Class
0.941	0	1	0.941		0.97	0.991	Benign
1	0.059	0.978	1		0.989	0.991	Malignant
Weighted Avg.	0.983	0.042	0.984		0.983	0.983	0.991

=== Confusion Matrix ===

```
 a  b   <-- classified as
32  2 |  a = Benign
 0 87 |  b = Malignant
```

4.2.2 Results for test data: Naive Bayes
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: breast_cancer_2013
Instances: 79
Attributes: 4
no_of_cell
image_no
HCG
class
Test mode:user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

Naive Bayes Classifier

Attribute	Class: Benign (0.37)	Class: Malignant (0.63)
no_of_cell		
mean	5.931	34.2925
std. dev.	4.9595	16.5106

weight sum	29	50
precision	2.6875	2.6875

image_no

mean	50.2069	121.9782
std. dev.	15.9091	14.8706
weight sum	29	50
precision	1.4545	1.4545

HCG

Present	2.0	51.0
Absent	29.0	1.0
[total]	31.0	52.0

Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances	78	98.7342 %
Incorrectly Classified Instances	1	1.2658 %
Kappa statistic	0.9726	
Mean absolute error	0.0127	
Root mean squared error	0.1125	
Relative absolute error	2.7207 %	
Root relative squared error	23.341 %	
Total Number of Instances	79	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.966	0	1	0.966		0.982 0.994	Benign
1	0.034	0.98	1		0.99 0.994	Malignant
Weighted Avg.	0.987	0.022	0.988		0.987 0.987	0.994

=== Confusion Matrix ===

a	b	<-- classified as
28	1	a = Benign
0	50	b = Malignant

4.2.3 Results for training data: J48 decision tree (implementation of C4.5)
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: breast_cancer_2013
Instances: 121
Attributes: 4
no_of_cell
image_no
HCG
class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===
J48 pruned tree

HCG = Present: Malignant (89.0/2.0)
HCG = Absent: Benign (32.0)

Number of Leaves : 2

Size of the tree : 3

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	119	98.3471 %
Incorrectly Classified Instances	2	1.6529 %
Kappa statistic	0.9583	
Mean absolute error	0.0325	
Root mean squared error	0.1284	
Relative absolute error	8.0055 %	
Root relative squared error	28.5509 %	
Total Number of Instances	121	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Area	Class
0.941	0	1	0.941		0.97	0.947	Benign
1	0.059	0.978	1		0.989	0.947	Malignant
Weighted Avg.	0.983	0.042	0.984		0.983	0.983	0.947

=== Confusion Matrix ===

a	b	<-- classified as
32	2	a = Benign
0	87	b = Malignant

4.2.4 Results for test data: J48 decision tree (implementation of C4.5)
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: breast_cancer_2013
Instances: 79
Attributes: 4
no_of_cell
image_no
HCG
class
Test mode:user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

Naive Bayes Classifier

Attribute	Class Benign (0.37)	Malignant (0.63)
no_of_cell		
mean	5.931	34.2925
std. dev.	4.9595	16.5106
weight sum	29	50
precision	2.6875	2.6875
image_no		
mean	50.2069	121.9782
std. dev.	15.9091	14.8706
weight sum	29	50
precision	1.4545	1.4545
HCG		
Present	2.0	51.0
Absent	29.0	1.0
[total]	31.0	52.0

Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances	78	98.7342 %
Incorrectly Classified Instances	1	1.2658 %
Kappa statistic	0.9726	
Mean absolute error	0.0127	
Root mean squared error	0.1125	
Relative absolute error	2.7207 %	
Root relative squared error	23.341 %	
Total Number of Instances	79	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.966	0	1	0.966		0.982 0.994	Benign
1	0.034	0.98	1		0.99 0.994	Malignant
Weighted Avg.	0.987	0.022	0.988		0.987 0.987	0.994

=== Confusion Matrix ===

a	b	<-- classified as
28	1	a = Benign
0	50	b = Malignant

4.2.5. Results for training data: JRip (implementation of the RIPPER rule learner)
=== Run information ===

Scheme:weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation: breast_cancer_2013
Instances: 121
Attributes: 4

no_of_cell
image_no
HCG
class
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

JRIP rules:
==========

(no_of_cell <= 14) => class=Benign (32.0/0.0)
=> class=Malignant (89.0/2.0)

Number of Rules : 2

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	117	96.6942 %
Incorrectly Classified Instances	4	3.3058 %
Kappa statistic	0.9182	
Mean absolute error	0.0487	
Root mean squared error	0.1741	
Relative absolute error	11.9902 %	
Root relative squared error	38.7096 %	
Total Number of Instances	121	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Area	Class
0.941	0.023	0.941	0.941		0.941	0.946	Benign
0.977	0.059	0.977	0.977		0.977	0.946	Malignant
Weighted Avg.	0.967	0.049	0.967		0.967	0.967	0.946

=== Confusion Matrix ===

a	b	<-- classified as
32	2	a = Benign
2	85	b = Malignant

4.2.6 Results for test data: JRip (implementation of the RIPPER rule learner)
=== Run information ===

Scheme:weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation: breast_cancer_2013
Instances: 79
Attributes: 4
no_of_cell
image_no
HCG
class
Test mode:user supplied test set: size unknown (reading incrementally)


```
=== Classifier model (full training set) ===

JRIP rules:
===========

(no_of_cell <= 14) => class=Benign (28.0/0.0)
 => class=Malignant (51.0/1.0)

Number of Rules : 2

Time taken to build model: 0.05 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          78               98.7342 %
Incorrectly Classified Instances        1                1.2658  %
Kappa statistic                         0.9726
Mean absolute error                     0.0248
Root mean squared error                 0.1114
Relative absolute error                 5.3314 %
Root relative squared error             23.111  %
Total Number of Instances               79

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.966    0          1          0.966      0.982      0.983    Benign
                 1        0.034      0.98       1          0.99       0.983    Malignant
Weighted Avg.    0.987    0.022      0.988      0.987      0.987      0.983

=== Confusion Matrix ===

  a  b   <-- classified as
 28  1 |  a = Benign
  0 50 |  b = Malignant
```

4.2.7 Summary

This section presents summary tables for scheme accuracy and running times.

Table 4.2.7.1: Accuracy and running time summary table for 10 fold cross validation

Model	Running time	10 fold cross val.
Naive Bayes	.02 seconds	98.5%
J48 decision tree (C4.5)	.06 seconds	98.5%
JRip (RIPPER rule learner)	.02 seconds	98.5%

Table 4.2.7.2: Accuracy for training and test data between different models

Model	Training Data	Test Data
Naive Bayes	98.34%	98.73%
J48 decision tree (C4.5)	98.34%	98.73%
JRip (RIPPER rule learner)	96.69	98.73%

V. CONCLUSION

Medical images have various limitations such as low quality, presence of noise and human error in interpretation .Digital image processing can help the pathologists to a great extent. So this type of automatic detection of breast cancer can help in early detection and diagnosis which can save patients.

REFERENCES

[1] http://www.cancer.gov/cancertopics/cancerlibrary

[2] Alkhair Abd Almahmoud Idris, Muhammed Sidahmed Hussain, "Comparison of the efficacy of three stains used for the detection of cytological changes in Sudanese females with breast lumps"Sudanese journal of public health. 2009,vol.4, No.2.

[3] www.polysciences.com

[4] American Cancer Society, Nov 2013

[5] Sariego J (2010). "Breast cancer in the young patient". *The American surgeon* 76 (12): 1397–1401.

[6] Florescu A, Amir E, Bouganim N, Clemons M (2011). "Immune therapy for breast cancer in 2010—hype or hope?". *Current Oncology*18 (1): e9–e18. .

[7] N. Lassouaoui, L. Hamami, and N. Nouali, 2007,Morphological Description of Cervical Cell Images for the Pathological Recognition, World Academy of Science, Engineering and Technology,pg 49-51

[8] Hao Yuan Kueh, Eugenio Marco, Mike Springer and Sivaraj Sivaramakrishnan, 2008, Image analysis for biology, MBL Physiology Course.

