Prediction Analysis of Cancer Cells Using ML Classification Algorithms

M. Parvathi[1], Ch. Manaswini[2]

[1]Professor, [2]Student, Department of Electronics and Communication Engineering, BVRIT Hyderabad College of Engineering for Women, Telangana, India

Abstract

Background: Breast Cancer is one of the most occurring cancers among all the diseases in medical science. About 1 in 8 women are suffering from breast cancer in their lifetime. In 2020, an estimated 276,480 new cases of invasive breast cancer are expected to be diagnosed in women along with 48,530 new cases of non-invasive breast cancer. Apart from women, men are also suffering from breast cancer, but the rate of occurrence is low in men.

Aim: To classify the breast cancer cells from original mammographic images using processing steps of Gaussian smoothing, Threshold segmentation, Feature Extraction.

Methodology: Breast cancer datasets are collected and preprocessed using attributes-like Menopause, Node-Capes, INV Nodes, Irradiate and Class. Three Machine Learning classifiers such as Bagging, Naive Bayes, and Naive Bayes Multinomial are applied for the classification analysis.

Results: Bagging classifier gives efficiency in the range of 77-86% when we consider Menopause, Node-Capes and Irradiate attributes and Naive Bayes classifier gives efficiency in the range of 71-78% for INV-Nodes and Class attributes.

Conclusion: It is observed that the bagging classifier gives best efficiency when we consider Menopause, Node-Capes and Irradiate attributes and Naive Bayes is best suits for INV-Nodes and Class attributes.

Keywords: *Weka explorer, dataset, Threshold segmentation, ML Classifiers.*

Introduction

The basic cancer cells are either from forming solid tumors or flood of blood from the abnormal cells. In general cell division is required for body growth and repair. In the cell division, the parent cell divides into sub cells called daughter cells. These daughter cells play a key role further to be used in generating new tissue or to replace the dead cells due to aging or by damage. A healthy cell stops itself by dividing into daughter cell when there is no longer a need for more daughter cells. In contrast, a cancer cell continues to produce copies.

Corresponding Author:
Dr. M. Parvathi
Professor, Department of ECE, BVRIT Hyderabad College of Engineering for Women, Telangana-500049
e-mail: pbmuddapu@gmail.com

Breast Cancer is one of the most exquisite and internecine diseases among all of the diseases in medical science. It is one of the crucial reasons of death among females all over the world. About 1 in 8 women (about 12%) are suffering from invasive breast cancer over the course of their lifetime. In 2020, an estimated 276,480 new cases of invasive breast cancer are expected to be diagnosed in women along with 48,530 new cases of non-invasive breast cancer.

In the year 1994, a research on WEKA Explorer has been done by Geoffrey Holmes, Andrew Donkin and Ian H and a paper named WEKA[1], a Machine Learning Workbench has been published. In this paper there is a detailed explanation given about what is WEKA Explorer.

They have tried enabling the machine learning schemes to be applied directly to the data in the database

in much the same way as systems that perform knowledge discovery in databases. This makes the users to use data from WEKA instantly into their projects comfortably.

In the year 1994, experts from the Department of Information and Computer Science Aalto University School of Science, Espoo, Finland have researched on the Classification with Learning Naive Bayes.In the year 1998, a publishing on Naïve Bayes Multinomial text has been done by Andrew McCallum and Kamal Nigam[2].This publishing concludes that to clarify the confusion by describing the differences and details of these two models, and by empirically comparing their classification performance on five text corpora. From the research they have given a conclusion of what is Naive Bayes algorithm and what is its implementation. Their experimental results on two abstract image datasets demonstrate the advantage of the multiple kernel learning frameworks for image affect detection in terms of feature selection, classification performance, and interpretation.

In the year 2003, there has been a research done on Image Processing techniques preceded by Artificial Neural Network by Zhi-Hua Zhou, & Yuan Jiang[3]. From this publication they have given a conclusion about the comprehensibility being very important for any machine learning technique to be used in computer-aided medical diagnosis.

In the year 2011, research on another Machine Learning classifier, Bagging. Breast Cancer classification using Bagging was done by authors namely M. A. Pradhan et.al[4]. Through this publication, they have concluded that breast cancer detection is very important in the field of medical science as well as Bioinformatics. The biomedical process like as image processing, the electrical processes like sensing from patient are erroneous because the accuracy of these processes is not stable all the time because of the limited lifetime of instrument.

In the year 2013, there has been a publishing on Digital Image Processing by Tina R. Patil and S.S. Shereker[5].This publication revealed that the digital image processing is far from being a simple transpose of audio signal principles to a two dimensions space.

From the recent publishing of Gaussian Smoothing by Lundin M, Lundin J, Burke HB and ToikkanenS[6], there has been a conclusion about Gaussian Smoothing that denoising method that introduce better smoothness that is much required for lowering SNRs. Summarizing all of these results, wavelet denoising method that introduce relatively little smoothness are generally preferable over Gaussian smoothing for denoising.

Thresholding is a step needed after Gaussian smoothing, in which each pixel in an image has its own threshold, which is estimated by calculating the statistical information of its neighbourhood pixels. Experimental results show that it is apparent to obtain better results by the proposed algorithm than by cannyoperator [7, 8].

Recent publications[9, 10, & 11]reported are on Deep Learning Method for automated breast cancer diagnosis using different classifiers like CNN, KNN, Inception V3, SVM and ANN concludes that better accuracy for detection of Breast cancer and analysis using WEKA explorer.

With the reference to all the papers above it is observed that detection and classification of cancer cells includes many steps to be performed in the present scenario. This paper presents an idea of detection and classification of cancer cells into benign and malignant using few machine learning algorithms by taking few attributes based on which cells are classified. The mammographic images are collected from the specialists; from that the breast cancer cells are identified and further they are classified into Benign and Malignant cells using machine learning classifiers namely Bagging, Naive Bayes and Naive Bayes Multinomial Text. The major attributes considered in our work are Menopause, Node-Capes, INV-Nodes, Irradiated and Class. We have performed the required Image Processing techniques like Gaussian-Smoothing Threshold Segmentation for the preprocessing the mammographic images and then ML algorithms are applied on the extracted features using WEKA Explorer for the classification of cells. Finally the performance of all the three classifiers is analyzed and observed for higher efficiency of a particular classifier pertains to a chosen attribute.

Section 2 discussed about flowchart, the methodology used in our work. Section 3 gives the details on Weka explorer. Section 4 discussed about the attributes which are used for the classification. Section 5 gives the details on machine learning algorithms and classifiers which are used in our work along with conclusions respectively. Section 6 & 7 gives analysis and conclusions respectively.

Flow Chart Description: Before proceeding to classification of cells, the preprocessing steps are required as shown in Fig.1, which are followed like Gaussian smoothing and threshold segmentation for original mammographic images. These are performed in order to do the partition of the image into different segments based on pixel intensity values. After extracting the features of the cancer cells, then machine learning algorithms are applied to the feature extracted data so that the cells are classified as cancerous and non-cancerous.

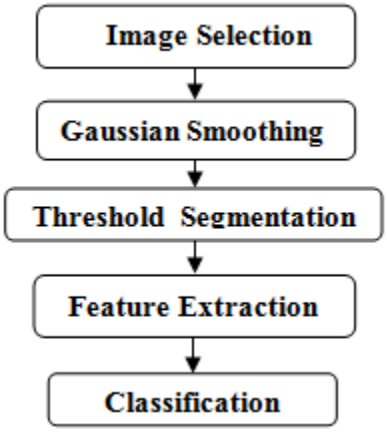
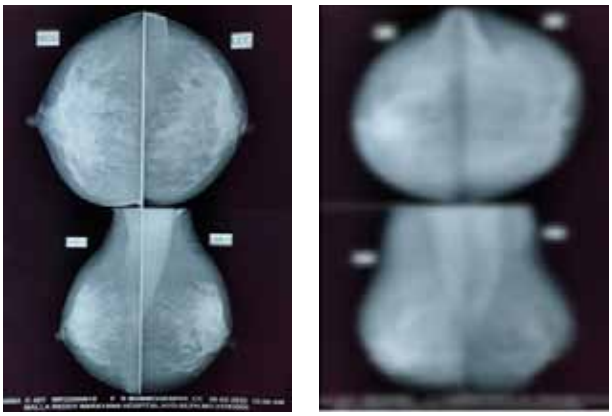


Fig. 1. Preprocessing steps before classification

(a) Image Selection: In this step we have collected a few original mammographic images from specialists, as shown in Fig. 2. Further Gaussian smoothing and threshold segmentation steps are performed as part of image preprocessing techniques.



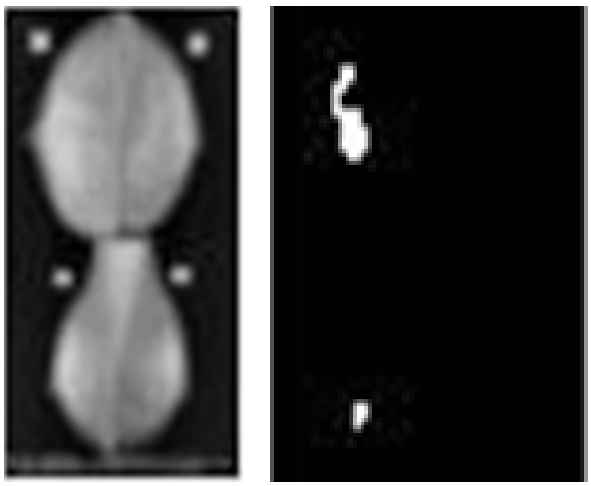
(a) **(b)**

Fig. 2. Results of Gaussian smoothing step, Original mammographic Image (a), Gaussian smoothing image with variance 0.5 (b)

(b) Gaussian Smoothing: In image processing, Gaussian smoothing is the result of blurring an image by a Gaussian function. The result of blurring technique is a smooth blur that resembles the image seen through a translucent screen.

We have performed Gaussian smoothing to the selected mammographic images with two different variance values. After Gaussian smoothing step applied, the resulted images are as shown in Fig. 2a & 2b respectively.

(c) Threshold Segmentation: After performing Gaussian smoothing to the original mammographic images, the smoothened images are segmented using threshold segmentation technique. In this stage, we have assumed threshold values arbitrarily and observed a better threshold value for which the cells are able to categorize. The images after threshold segmentation step applied for the Gaussian outputs are as shown in Fig. 3.



(a) **(b)**

Fig. 3. Threshold segmentation for image (a) using threshold value of 0.75

(d) Feature Extraction: Feature extraction step is carried out on 200 breast mammograms and the results are tabulated in database, which consists of minimum to maximum range of feature values.

(e) Classification: In this stage, we have implemented different Machine Learning classifiers like Bagging, Naive Bayes and Naive Bayes Multinomial Text classifiers to observe number of correctly classified and incorrectly classified cells based on the chosen attributes. Further to classify cancer cells into benign and malignant.

Weka Explorer: Weka contains a collection of visualization tools and algorithms for data analysis

and predictive modeling, together with graphical user interfaces for easy access to these functions. Dataset is collected from the UCI Repository database which consists of mammographic image data in numerical form as shown in Fig.4.

Feature Extracted Dataset

Cell attribute	Menopause	Nodecapes	Inv nodes	Irradiat	Class
1000025	5	1	1	1	2
1002945	5	4	4	5	7
1015425	3	3	1	1	2
1016277	6	8	8	1	3
1017023	4	1	1	3	2
1017122	8	1	1	1	2
1018099	1	2	1	2	1
1018651	2	1	1	1	2
1033078	2	2	1	1	1
1033078	2	5	3	3	3
1035283	4	8	7	5	10
1036172	2	5	3	3	3
1041801	5	7	5	4	3
1043999	1	2	1	1	1
1044572	8	7	5	6	4
1047630	7	4	1	1	1
1048672	4	1	1	2	1
1049815	1	7	3	2	10

Fig. 4. Dataset from UCI Repository database

Attributes Used For Classification:

(a) Menopause: Hormonal imbalance such as continuous exposure to estrogen leads to high risk of breast cancers. In this scenario, women with natural menopause are more likely to prone to develop cancers as twice as high because of hormonal disturbances. Hence, based on Menopause attribute breast cancer cells are detected and classified into benign and malignant.

(b) Node-Capes: One of the main reasons for the spread of breast cancer is the cancer cells get into the blood or lymph system. Once the cancer cells are mixed in the blood, they will spread to other parts of the body. If cancer cells have spread to lymph nodes, there is a higher chance that the cells could have travelled through the lymph system and spread (metastasized) to other parts of the body.

It is more likely to find the cancer in other organs in the body when more lymph nodes are affected with breast cancer cells. Because of this, finding cancer in one or more lymph nodes often affects treatment plans. Usually, we will need surgery to remove one or more lymph nodes to know whether the cancer has spread. Even now, it is not true to say that all women with cancer cells in their lymph nodes may develop metastases. Because, some women with no cancer cells in their lymph nodes may develop metastases later. Hence based on Node-Capes attribute cancer cells are detected and classified as benign and malignant.

(c) INV Nodes: INV Nodes are the number (range 0 - 39) of axillary lymph nodes that contain metastatic breast cancer visible on histological examination. Hence INV Nodes are taken as an attribute to detect and classify cancer cells as benign and malignant.

(d) Irradiate: Accelerated Partial Breast Radiation (APBI) uses high-powered x-rays to kill breast cancer cells. In general, external beam breast is given as a course will take 3 to 6 weeks for the overall treatment to complete. Current research suggests that APBI produces low local recurrence

rates that are comparable to the recurrence rates of whole breast irradiation. Hence Irradiate is to be considered as an important attribute in detecting and classifying cancer cells.

(e) Class: The term class label is usually used in the context of supervised machine learning. Class label is a discrete attribute used as dependent variable. The class label always takes on a finite (as opposed to infinite) number of different values. Hence Class is considered as an attribute in our work.

Machine Learning Algorithm Classifiers:

(a) Bagging: ML algorithms are mainly used in statistical classification and regression. Bagging is an ML ensemble meta-algorithm used to improve the stability and accuracy of ML model. The main advantage is, it reduces variance and further helps in avoid over fitting.

(b) Naive Bayes: Naive Bayes is a simple technique for constructing classifier models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. In case of naive Bayes classifiers, a particular feature will be selected so that its value is independent of the value of any other feature among the class variables.

(c) Naive Bayes Multinomial Text: In the case of a multinomial Naive Bayes classifier, it uses a multinomial distribution for each of the features. It is a specific instance of a Naive Bayes classifier.

Analysis:

(a) Menopause Attribute: We have considered attributes related to Menopause, Node-Capes, INV Nodes, and Irradiate for correctly classified cells, incorrectly classified cells in order to observe the efficiency in each of the algorithm applied. The corresponding results are observed as shown in Table 1, 2 and 3 using the algorithms Bagging, Naïve Bayes and Naïve Bayes Multinomial Text respectively.

Table 1. Attributes and efficiency variation using Bagging algorithm

Attribute/Algorithm	BAGGING				
Catagory of cells	INV Nodes Attribute	Menopause Attribute	Node-Capes attribute	Irradiate Attribute	Class Attribute
CORRECTLY CLASSIFIED CELLS	222	236	240	223	198
INCORRECTLY CLASSIFIED CELLS	64	50	38	63	88
EFFICIENCY(%)	77.6	82.5	86.33	77.97	69.23

Table 2. Attributes and efficiency variation using Naïve Bayes algorithm

Attribute/Algorithm	NAIVE BAYES				
Catagory of cells	INV Nodes Attribute	Menopause Attribute	Node-Capes attribute	Irradiate Attribute	Class Attribute
CORRECTLY CLASSIFIED CELLS	224	233	242	217	205
INCORRECTLY CLASSIFIED CELLS	62	53	36	69	81
EFFICIENCY(%)	78.3	81.46	87.05	75.87	71.67

Table 3. Attributes and efficiency variation using Naïve Bayes MT algorithm

Attribute/Algorithm	NAIVE BAYES MT				
Catagory of cells	INV Nodes Attribute	Menopause Attribute	Node-Capes attribute	Irradiate Attribute	Class Attribute
CORRECTLY CLASSIFIED CELLS	213	150	222	218	201
INCORRECTLY CLASSIFIED CELLS	73	136	56	68	85
EFFICIENCY(%)	74.47	52.44	79.85	76.22	70.27

Efficiency Analysis: After making comparisons among the three classifiers with different attributes we have observed the overall best efficiency using a particular classifier with a particular attribute. Efficiency analysisis shown in Table 4.

Table 4. Efficiency analysis among the chosen attributes

Efficiency Analysis		
ATTRIBUTE	ALGORITHM	EFFICIENCY(%)
MENOPAUSE	BAGGING	82.5
NODE-CAPES	BAGGING	86.33
INV-NODES	NAIVE BAYES	78.3
IRRADIAT	BAGGING	77.9
CLASS	NAIVE BAYES	71.6

Conclusions

In this paper, breast cancer cell classification is done using few machine learning algorithms. Initially few mammographic images were considered and applied few image processing steps for segregation of cancer and non-cancerous cells. The classification further has been used in datasets using attributes like menopause, INV Nodes, Node caps and class. Using efficiency parameter, we have concluded that bagging classifier is best when we use Menopause, Node-Capes, and Irradiate attributes. Similarly Naive Bayes Classifier gives the best efficiency when we consider INV-nodes and Class attributes.

Financial support and sponsorship: Nil

Conflict of Interest: Nil

Ethical Clearance: Taken from institutional ethical committee.

References

1. G. Holmes; A. Donkin and I.H. Witten (1994). "WEKA: A machine learning workbench", in Proceedings: Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.
2. "A Comparison of Event Models for Naive Bayes Text Classification", Andrew McCallum and Kamal Nigam. AAAI-98 Workshop on "Learning for Text Categorization".
3. Zhou ZH, Jiang Y (2003) "Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble", IEEE Transactions in Information Technology Biomedical, Vol:7, Pp: 37-42.
4. M.A.Pradhan, Abdul Rahman, Pushkar Acharya, Ravindra Gawade, Ashish Pateria, "Design of Classifier for Detection of Diabetes using Genetic Programming", in Proceedings: International Conference on Computer Science and Information Technology (ICCSIT'2011), Pattaya, Dec. 2011, Pp: 125-130.
5. Tina R. Patil, Mrs. S.S. Sherika, Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification", in Proceedings: International Journal of Computer Science and Application, Vol.6, No.2, Apr 2013, Pp: 256-261.
6. Lundin M, Lundin J, Burke HB, Tolkien S, Piikani L, "Artificial neural networks applied to survival prediction in breast cancer", Oncology, Vol: 57, Pp: 281-286.
7. Shiping Zhu, Xi Xia, Qingrong Zhang, Kamel Belloulata, "An Image Segmentation Algorithm in Image Processing Based on Threshold Segmentation", in proceedings: Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, 2008, Pp:627-632.
8. Wu HS, Barba J, Gil J. Iterative thresholding for segmentation of cells from noisy images. Journal of Microscopy. 2000 Mar;197(Pt 3), DOI:10.1046/j.13652818.2000.00653.x, Pp:296-304.
9. Kalyani Wadkar, Prashant Pathak, Nikhil Wagh, "Breast Cancer Detection Using Ann Network And Performance Analysis With SVM", International Journal of Computer Engineering & Technology (IJCET), Volume 10, Issue 3, May-June, 2019, pp. 75-86.
10. Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms", Hindawi, Journal of Healthcare Engineering, Article ID 4253641, Pp: 1-11.
11. TanayaPadhi, Praveen Kumar, "Breast Cancer Analysis Using WEKA", in proceedings: 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence-2019), DOI: 978-1-5386-5933-5/19/, Pp: 229-233.