ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

A Survey on Detection of Breast Cancer from Mammogram

M.Chitra Devi[1], B.Sobiya[2]

Associate Professor, Sakthi College of Arts and Science for Women, Tamil Nadu, India[1]

Research Scholar, Department of Computer Science, Sakthi College of Arts and Science for Women, Tamil Nadu, India[2]

ABSTRACT: In this world, most of the females are affected with the breast cancer. In the ranking of cancers, breast cancer is in the second place. Breast cancer is a type of disease which originates from the tissue cells present in or around the breasts. It is commonly arises in the lobules or milk duct. A survey report shows that 15% of breast cancer affected females meets death. Sometimes it extends the lifetime of the affected women by removing the affected breasts. Many techniques supports in the detection of breasts cancer. In our proposed work, the classification using entropy enabled decision tree is presented. The proposed work makes use of the mammogram image for the detection purpose. The performance of the classifier is examined using Accuracy, Specificity, Precision etc. The training phase and the testing phase are utilized. It categorizes the cancer growth as recurrent and non-recurrent accurately. This method will be effective in all the ways.

KEYWORDS: Breast cancer, Decision Tree classifier, Entropy, Classification, Data Mining.

I. INTRODUCTION

Breast cancer detection is taken in process when the affected women or medical intern feels some change or abnormality in the tissues of breasts while diagnosing or by self-analyzing. The factors involved in analyzing the breast cancer are based on age, severity conditions, symptoms and testing reports. In order to present best implementation method, the more research work has been handled. A survey report of National Institute of Cancer Prevention and Research (NICPR) states that, the women are mostly affected by the breast cancer and it is evolving as more common disease now-a-days in women. The 15% of the affected women faces death. The report says that only low number of count of women is surviving by removing or affected breasts or by extending their lifetime with treatment at the earliest stage. In addition, they state that the count is very poor in rural sides when compared to urban regions. The report of Globocan gives information there are 1,44,937 tumors are in existence. The loss in life % is 70,218. The women of 50-64 years and mid of 30 years are affected in most common. The report of Indian Council of Medical Research (IMCR) states that the tumor count will be increase in more than 3lakhs from now in 2020.

Breast cancer is found to be the top most common disease in India and in world level, the second common disease. The report of PBCR and HBCR states that among 16% only 5% of the women are surviving or extending their lifetime. The report of ASCO (America ' s medical report) state that the count of improvement is increased from 75% to 89%. The death count in India is due to the unawareness of the disease. The women are aware of the disease at the third or fourth stage of the breast cancer. The breast cancer has to be finding at the earliest stage. For that, a best implementation method is to be needed. The method has to pave way for the treatment or detection without biopsy. The recurrent is used to refer the tumor extraction at some time and the occurrence of disease stopped at this stage.

The applications with data mining are emerging now-a-days. Data mining technique is inspired by many scientists and researchers. It is much more attracted due to the conversion of large amount of data into information or knowledge. The

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

data extracted from these should be robust, simple and less computational complexity. To acquire meaningful information, many techniques from the fields of computer science, machine learning, mathematics and statistics are helpful. The extraction of meaningful information from the data helps us to discover solutions for many unknown causes. For the up gradation in treatment, the mammography image is analyzed using data mining technique.

Mostly the assessment of breast cancer is analyzed in and around parts of the breasts. Based on various kinds of test reports, the medical intern will precede the treatment. If the test and identification using biopsy is not impossible, the other alternative has to be taken in care. The most common alternative solution is the detection using image analyzing. Some of the image analyzing approaches is diagnostic using mammography, MRI or ultrasound images. Classification also plays an important role in this. Using the classifier, the disease can be identified at the earliest stage. The earliest stage of detection of disease helps to provide treatment accordingly and to extend the lifetime process of the person. Because of this reason, the analysis and detection is much important in the field of medical and bioinformatics. The female breasts include 15-25 milk glands in breasts and it is referred as lobules. Lobules are the part which is the milk producer and holder. Lobules are connected to the nipples via small tubes called duct. The brown color layered part around the nipple is termed as areola. Apart from this, the remaining space of breast is filled with fat and the connecting tissue. Some of the types of breast cancer are Ductal cancer, Lobular cancer, and Sarcoma cancer. The masses in the breast cells are classified into benign and malignant. Malignant is the cancerous cell whereas benign is the non-cancerous cell. Digital mammogram is a imaging technique used for diagnosing breast cancer [13].

False - Negative Results

False-negative results occur when mammograms appear normal though breast cancer is present. Overall, screening mammograms miss up to 20 percent of breast cancers that are present during screening. The reason for false-negative results is high breast density. Breast contains glandular tissue and connective tissue known as fibroglandular tissue together with fatty tissue. Fatty tissue appears dark on a mammogram, while dense tissue and tumors appear as white areas. Because fibroglandular tissue and tumors may have either more or less density, therefore is quite difficult to detect in women with denser breasts.

False - Positive Results

False – Positive results refers to mammograms for showing abnormalities in tissues, benign tissues. These mammograms should be followed with further test such as ultrasound and/or biopsy to determine whether cancer is present [15].

The research paper is sectioned as follows. The section II presents the review related to the proposed work, section III is to explain the methodologies of proposed work, section IV detailed the results of proposed work and section V concludes the entire paper.

II. RELATED WORK

In the developmental world, identification of breast cancer at the earliest stage using data mining technique is more inspired by all the researchers and medical interns. The review of some existing techniques are discussed in this section.

The author Skevofilakas et al. published a paper in order to increase the provision and visualization of collected data samples using a decision support system. This method is applicable to a particular disease, breast cancer. The author, W.J. Clancey, E.H. Shortliffe, eds., published a paper based on the implementation of the rule-based artificial intelligence to identify the disease at the earliest stage using computer algorithms. This work studied by many researchers for the presence of knowledge, belief networks and other different areas. This method makes diversity in the field of medicine and bio-informatics. Another author Menolascina et al. presented a paper to show the comparison reports of J48, Naive Bayesian Tree, Ant Miner and Gene Expression Programming. These methods are applied to the data samples of identification of CGH based breast cancer analyzing. Another researcher Orlando Anunciacao et. al presented a paper based on the decision tree classifier to identify the disease due to the in taken of alcohol, tobacco

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

related stuffs. They presented the complexity of multiple disease due to these habits instead of one. The researcher Yang et al., proved in his paper that the analyzing of breast cancer identification based on thermographs in gray scale gives better results. He identified the parameters of data samples using Beier-Neely field morphing along with decision trees.

Researchers, A.Soltani Sarvestani proposed a paper based on the comparison of results. They have published this paper by finding a solution to the challenge of high dimension data samples. They have also analyzed the correlated nature of the data, principal component method for this challenge.
The results have found to be satisfied when compared to existing methods. Researcher D. Nauck et. al published a paper using the intelligence system such as hybrid of neural networks with fuzzy technology. By experimental and theoretical tests proved that the hybrid method can make the difference in control, decision-making and data analysis systems in high level.

The author W.Y. Cheng et. al presented a paper based on finding a robust predictive model to identify the breast cancer disease. For the identification purpose used the ample training dataset which is the samples of admitted person ' s gene expression and clinical features. The implementation method is also tested with the newly generated dataset. In the analysis of several tumors, they found signatures called attractor metagene.
The author Sarvestani et al. [18], proposed a paper based on the implementation of neural network structures like SOM, RBF, GRNN, PNN. The methods are tested with the data samples from WBCD and NHBCD. They have proposed this paper to improve the accuracy in detection of breast cancer.

The researcher A. KELE published a paper based on the data extracted using fuzzy rules for the identification of breast cancer. The rules for this extraction are implemented using Neuro Fuzzy Classification tool referred as NEFCLASS. The base rule includes nine rules based on BI-RADS, mass shape and mass margin attributes.
The results concluded that positive value is 75% and negative value is 93%. The author R.R.Janghel et.al proposed paper based on the various models of ANN. They implemented this method for detection, identification and treatment ailment for the affected persons.
The ANN models which proposed are BPN Algorithm, Radial Basis Function Networks (RBFN), Leaning Vector Quantization (LVQ) and Competitive Learning Network (CLN). The results are in the descending order of LVQ, CLN, MLP BPN and RBFN.

III. PROPOSED SYSTEM

The proposed system of detecting the breast cancer at the earliest stage is using the classification method. The classification is defined as the methodology of identifying the model or classifier. The classifier helps in the comparison of data classes. So that, the process can be easily identify the class of entities or objects of unknown label.
The classifier processes as the training phase and testing phase. In the training phase, the relevant data is extracted and it is studied based on the classifier rule. In the testing phase, it can be used for identification purpose. The training data phase is added to improve the accuracy rate. The classification can be based on various approaches like simple classification rules, decision trees, mathematical formulas, or artificial neural networks.

Classification using Entropy Enabled Decision Tree (E^2DT)
This is the initial process of the proposed system. In this the algorithm steps are presented as a flow chart in the below figure 1. The classification of data is processed using these steps.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

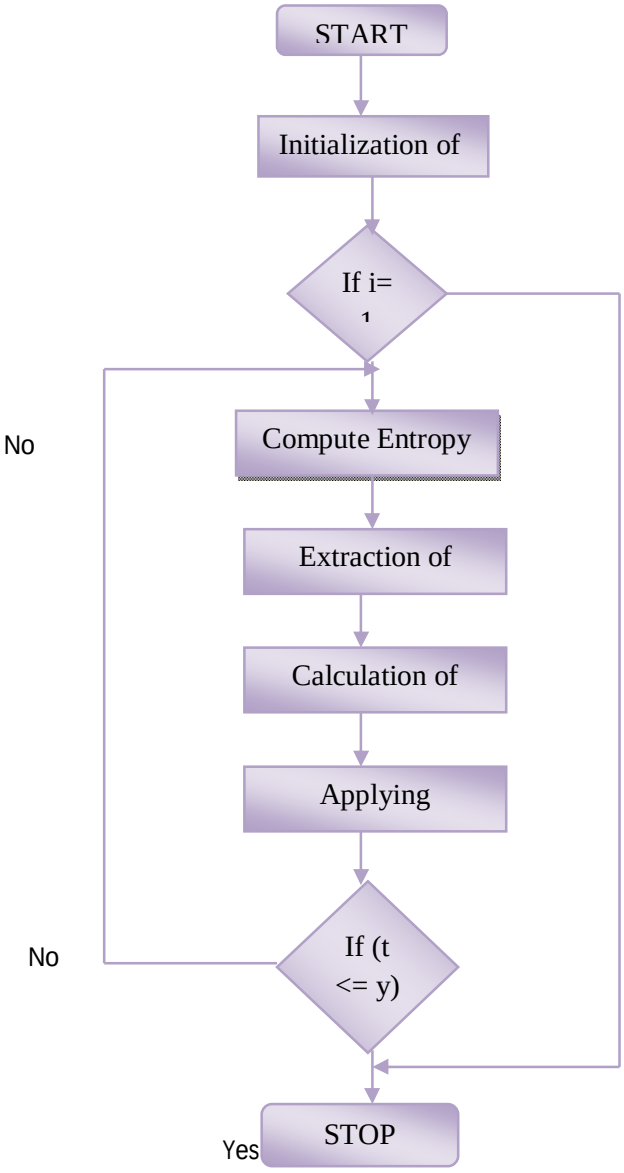

Figure 1 – Steps of Entropy Enabled Decision Tree Classifier Algorithm

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

Steps in algorithm

1) **Feature extraction –** The decision tree is constructed using the relevant features extracted from the dataset. It is a mandatory process in the classification techniques. The relevant features reduce the reworking and false results at the end. For this purpose, the entropy is enabled in the proposed work. The entropy is determined by computing for each attributes in the data samples. The attribute which provides high entropy value is extracted as the relevant feature. The entropy is computed using the below formula,

$$E(A_i) = W * E(A_i) \quad \text{---------- (1)}$$

$$W = 2 * \left(1 - \frac{1}{1+\exp(-E(Ai))}\right) \quad \text{---------- (2)}$$

$$E^2(A_i) = -\sum_{i=1}^{u(Ai)} Pi \log Pi \quad \text{---------- (3)}$$

In the above formula, E is termed to define entropy, $E(A_i)$ is used to describe the entropy of attribute A_i. Similarly $u(A_i)$ is to determine unique values and P_i is the probability of the feature and $E^2(A_i)$ is the entropy of feature attribute. The weight is denoted with the letter W. The computation of entropy is applied to all the features of chunk data. It helps to improve the accuracy and also positive results.

2) **Categorization Rule –** The categorization rule is taken over to make decision of the accurate value to categorize the data into two for the purpose of identifying a node. Based on the results of the previous step, the categorization rule is applied to the extracted relevant features. Based on the parameter, Information Gain, the categorization rule is applied. IG is used based on the unique feature value $u(A_i)$.

IG (A_p, Aq) is used to represent the information gain. The result of the information gain is based on the entropy value $E(A_i)$. The conditional entropy is represented as CE (A_p, Aq). The computation of information gain is as follows.

$$IG(A_p, Aq) = E(A_i) - CE(A_p, Aq) \quad \text{------- (4)}$$

$$CE(A_p, Aq) = \sum_{i=0}^{u(Ai)} Pf\, E(Ap, Aq) \quad \text{----- (5)}$$

$$E(A_p, Aq) = W * E(A_p, Aq) \quad \text{---------- (6)}$$

$$W = 2 * \left(1 - \frac{1}{1+\exp(-E(Ap,Aq))}\right) \quad \text{--------- (7)}$$

$$\text{Finally, } E(A_p, Aq) = \sum_{i=0}^{u(Ai)} P(Ap = p, Aq = q).\log(Ap = p, Aq = q) \quad \text{--- (8)}$$

The decision tree is constructed using the node of the tree. If the node of the tree is decided, the data which are assigned to the branches are sent to the process of feature extraction or classification. And then the output of the extracted feature is sent to the categorization rule process. These two steps are continued until the last node of the tree is reached. If it is identified, it is forwarded to the next step.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

3) **Final Decision -** The final decision is the third step of the algorithm process. The final decision is taken when the entire data sample is categorized and it is noted at the leaf node. It is the step of identification of the classification process.

4) **Process of Labeling -** The process of labeling is marked using the class of maximum number of data. The data which comes under the same class in the chunk, that class is preferred as the class for the labeling node.

IV. DISCUSSION AND FUTURE WORK

In this paper, the breast cancer which affects women in large count is discussed. Many things of breast cancer are presented in this paper. The identification and detection is done using the digital mammography image. The implementation method helps to detect the breast cancer at the earliest stage. It helps to extend the lifetime of the affected women. The classification method implemented in this paper is entropy enabled decision tree classifier. It helps in the classification of cancer as recurrent and non-recurrent accurately. The parameters which are computed for the proposed system are listed below. The parameters computed are used in the purpose of 2 class confusion matrix.

Accuracy – The accuracy is the parameter which is used to compute the overall true values of the proposed system. It is computed based on the ratio of sum of the extracted feature using classification with the total number of feature attributes. The equation to compute accuracy is $Accuracy = TP + TN / TP + TN + FP + FN$.

Precision – Precision is computed based on the true conditions which are identified perfectly. The formula to obtain precision is $Precision = TP / TP + FP$.

False Positive Rate – The false positive (FP) is the condition when the negative values of data are classified as positive by mistake. It is computed using the formula, $FP = FP / FP + TN$.

True Positive Rate – The true positive (TP) is the condition when the positive values are marked as negative incorrectly. It is computed using the equation, $TP = TP / TP + FN$.

Higher precision (HP) – In some analysis of data classification, the higher precision is required to compute. It makes the accuracy in the classification of data samples and works out very well in the implementation. The higher precision is calculated using the below formula, HP = $2 * Precision * TP / Precision + TP$.

Table 1 – Comparison of Chunk and Accuracy

	Size of Chunk			
E^2DT	1	2	3	4
1	0.91492	0.9352	0.92799	0.93854
2	0.79984	0.79987	0.79984	0.79967
3	0.99278	0.99396	0.99394	0.99273

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 5, May 2019

The accuracy rate makes the medical intern to diagnosis and treats as per the report. The decision tree implementation model is experimented with changing the chunk size from 1 to 4. The model gives accuracy gives the high rate to all kinds of data samples. The results obtained for the proposed system is presented in the table 1. In future, the implementation with another classifier or hybrid classifier is planned to check the precision and speed.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Database mining: A performance perspective," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 5, no. 6, pp. 914–925, 1993.
[2] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *Knowledge and data Engineering, IEEE Transactions on*, vol. 8, no. 6, pp. 866–883, 1996.
[3] E. Coyne and S. Borbasi, "Living the experience of breast cancer treatment: The younger womens perspective," *Australian Journal of Advanced Nursing*, vol. 26, no. 4, pp. 6–13, 2009.
[4] W. J. Clancey and E. H. Shortliffe, "Introduction: Medical artificial intelligence programs," *Readings in medical artificial intelligence: the first decade. Reading, Mass.: Addison-Wesley*, pp. 1–17, 1984.
[5] O. Anunciaçao, B.C. Gomes, S.Vinga, J.Gaspar, A.L. Oliveira, and J. Rueff, "A data mining approach for the detection of high-risk breast cancer groups," in *Advances in Bioinformatics*. Springer, 2010, pp. 43–51.
[6] WHO Summary report on HPV & cervical cancer statistics in India (18/03/2008)
[7] Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 . Lyon, France: International Agency for Research on Cancer; 2013.
[8] Bray F, Ren JS, Masuyer E, et al. Estimates of global cancer prevalence for 27 sites in the adult population in 2008.; 2013; Int J Cancer.; 132(5):1133-45.
[9] Vijay Mahadeo Mane and D.V. Jadhav, " Holoentropy enabled-decision tree for automatic classification of diabetic retinopathy using retinal fundus images", Biomedical Engineering, August 2016.
[10] Marios Skevofilakas, Konstantina Nikita, Panagiotis Templaleksis, K. Birbas, I. Kaklamanos, G. Bonatsos, "A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines", pp. 2429 - 2432, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, China, Sept. 2005.