International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
529 – 535

A Computational Approach to Predict the Severity of Breast Cancer through Machine Learning Algorithms

K.L.V.G.K. Murthy[#1], Dr. R. J. Rama Sree [*2]
[#1]CSE Department, St. Marys Group of Institutions, Guntur, Research Scholar of Rayalaseema University, Kurnool.
[*2]Professor & Head, Department of Computer Science, Rashtriya Sanskrit VidyaPeeth,Tirupathi,
Research supervisor for Rayalaseema University,Kurnool.

Abstract: Breast cancer is one type of cancer which causes from breast tissue. A lump in the breast, skin dimpling, breast shape changes, fluid from the nipple, or a red scaly patch of skin are some of signs of breast cancer. In the world, cancer is one of the most leading causes of deaths among the women. Among the cancer diseases, breast cancer is especially a concern in women. Mammography is one of the methods for finding tumor in the breast. This method is utilized to detect the cancer which is helpful for the doctor or radiologists. Due to the inexperience's in the field of cancer detection, the abnormality is missed by doctor or Radiologists. Segmentation is very expensive for doctor and radiologists to examine the data in the mammogram. In mammogram the accuracy rate is based on the image segmentation. The recent clustering techniques are presented in this paper for detection of breast cancer. These Classification algorithms have been mostly studied which is applied in a various application areas. To maximize the efficiency of the searching process various clustering techniques are recommended. In this paper, we have presented a survey of Classification techniques.

Keywords: *Mammogram, Classification Techniques, Decision Tree, SVM, Naive bayes, Amelia package.*

******

I. INTRODUCTION

Breast cancer is one of the types of cancer which is caused from the breast tissue. A lump in the breast, dimpling of the skin, a change in a shape, fluid coming from the nipple or a red scaly patch of skin [1] are the some of the symptoms [8] for breast cancer. The distant spread of disease may be causes bone pain, shortness of breath, swollen lymph nodes, and yellow skin [9].

The risk factors which cause the development in breast cancer are being female, lack of physical exercise, obesity, drinking alcohol, hormone replacement therapy during menopause, early age at initial menstruation, and ionizing radiation by having children late or not at all, at age of older and history of family.[1][2] 5-10% of cases are because of genes which are assumed from a person's parents [1] containing BRCA1 and BRCA2 amidst others. Cancers which are developed from ducts is called as ductal carcinomas and which are developed from lobules is called as lobular carcinomas.

Moreover, more than 18 other subtypes are there in the breast cancer. [2] Some other cancers like ductal carcinoma in situ which is developed from pre-invasive lesions.[2] Once diagnosis is done, then further tests are made for determining which treatment has to be done that cancer may respond to [1] and to identify that cancer which is spread beyond the breast.

Breast image processing has various stages. The initial stage is breast image acquisition through mammography. The next stages are pre-processing image, feature extraction, feature selection and classification [5].

One (single reading) or two (double reading) trained professionals are evaluated. These film readers are radiologists and radiotherapists or breast clinicians are non-radiologist physicians which are specializing in disease of breast cancer. In the UK, there is a standard practice for double reading but in the US, it is less common which is significantly improves the sensitivity and specificity of the procedure. Digital mammography or analogue mammography of digitized images may be utilized by Clinical decision support systems. These studies suggest that these do not improve performance only gives a small improvement.

Digital mammography: It is a mammography in which digital receptors and computers are utilized for examining of breast tissue of breast cancer. The electrical signals are manipulating the images which can be read on computer screens to allocate radiologists to view the results more clearly.

For screening, there may be "full field" (FFDM) Digital mammography or "spot view" for breast biopsy. In stereotactic biopsy also there is a utilization of Digital mammography. By utilizing various modalities, like ultrasound or magnetic resonance imaging (MRI), Breast biopsy may be performed. The more marked advancement is expected by radiologists, but the digital mammography effectiveness was found which is comparable to traditional x-ray methods in 2004. Additionally, there may be reduction in radiation and may lead to fewer retests. The Preventive Services Task Force stated insufficient evidence against digital mammography.

 ISSN: 2454-4248
529 – 535

It is for the Hubble Space Telescope which is a NASA spin-off is developed. In 2007, 8% of utilization by American screening centers. Systems by Fujifilm Corporation are the utilized around the globe.[citation needed] The expensive of US$300,000 to $500,000 are for GE's units of digital imaging in the US.

3D Mammography: Three-dimensional mammography, also known as digital breast tom synthesis (DBS), tom synthesis, and 3D breast imaging, is a mammogram technology that produces a 3D image by utilizing X-rays. Moreover, usual mammography provides results in numerous positive tests. The effectiveness of Cost is unclear as of 2016. Other concerns are that it maximizes radiation exposure by more than two times. The main aim is to propose a method for identifying the stage of breast cancer malignancy based on cancer size on mammogram image.

This work is described as follows. In Section 2, literature survey is described. We present the proposed method includes the process of segmentation and classification in Section 3. In Section 4, the results are presented. Finally, in Section 5 conclusion is presented.

II.R TOOL

R tool is a free environment software which is utilized for statistical computing and graphics. R tool assembles and runs on Windows and Mac OS [12] and UNIX platforms. The utilization of R tool is for statistical analysis and graphic techniques [10] which is public domain software. A core set of packages, with greater than 7,801 more additional packages (as of January 2016[update]) are included in the R tool installation which is available at the Comprehensive R Archive Network (CRAN), Bio Conductor, Git Hub, and some other repositories.[14] An extensive class of statistical that includes linear and nonlinear modeling, time-series analysis, classical statistical tests, classification, clustering and various graphical functions are given by the R tool.[13]R utilizes collections of packages for performing variant functions [11].

A CRAN project view gives many packages to variant users corresponding to their taste. For approaches of data mining this R package consists of dissimilar functions. This paper differentiates various classification algorithms on datasets by utilizing R which will be useful for researchers working on medical data and biological data as well. Figure 1 shows the utilization of R Studio for this IDE.

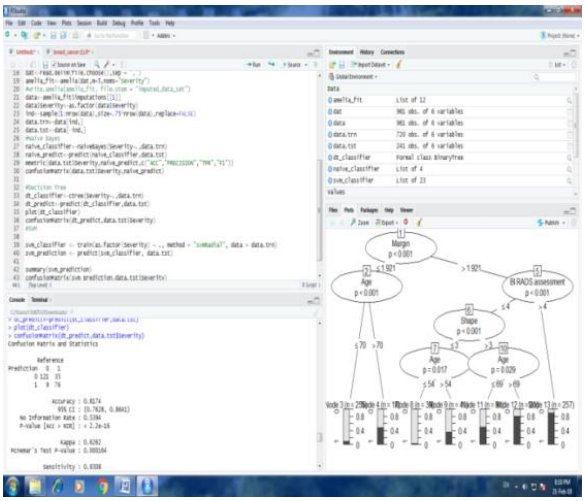

FIG 1: R TOOL STUDIO

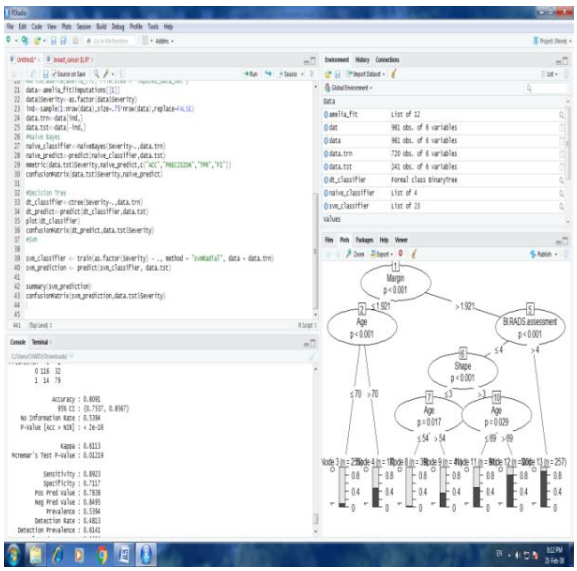

FIG 2: R TOOL STUDIO

III.LITERATURE REVIEW

In the literature review, a brief illustration of the existing researches on data mining techniques is presented with respect to breast cancer. The description of the algorithms is utilized along with their scope and limitation which is presented in the review. The performance of IBK (K-nearest neighbor classifier), BF Tree (Best First Tree) and SMO (Sequential Mi++nimal Optimization) classification techniques are investigated on the data of breast cancer by Chaurasia et al. [3]. The experiment has conducted in Weka data mining tool by the authors. On the basis of time, correctly classified instances and accuracy for assessing the superiority of each algorithm is evaluated. By this estimation, the performance of SMO algorithms has been better in terms of accuracy and low error rate than the other two algorithms. The most significant features for magnifying the prediction accuracy is also identified.

 ISSN: 2454-4248
529 – 535

The approaches of data mining such as naïve bayes and j48 decision trees are assessed by Williams et al. [1] for predicting the possibilities of breast cancer in Nigerian patients. The most efficient and effective model is determined by this analysis. The collection of dataset is made in LASUTH of cancer registry, Ikeja in Lagos, Nigeria which consists of instances 69 with attributes of 17 with the class label. 11 non-modifiable factors and five modifiable factors are carried in the dataset. The experiment is performed through Weka. Finally, by this evaluation, authors declared that the prediction of breast cancer possibilities in j48 decision tree is better with the accuracy values (94.2%), and error rates.

Shajahaan et al. [4] have elucidated the decision trees for prediction of breast cancer and performance of conventional learning algorithms which are supervised like CART, ID3, C4.5 and Naïve Bayes also analyzed. The experiment is made through Weka tool. The authors have five meaningful highlighted attributes which are contemplated for the prediction. The authors have concluded that the random tree serves as the best classification algorithm for breast cancer with higher accuracy in prediction.

Shrivastava et al. [5] have utilized classification techniques for dividing the benign or malignant instances of breast cancer dataset. A decision tree classifier model is created by the authors for the classification. The experimentation is done through Weka tool and for simplifying the prediction task. The authors have declared that the analyses of most of the breast cancer have conducted only through neural network and approaches of decision tree. Hence, decision tree model is taken by utilization of if-then rules for magnifying the functioning of j48 decision trees.

Senturk et al. [6] elucidated the prediction models like Discriminant Analysis (DA), Artificial Neural Networks (ANN), Logistic Regression (LR), Support Vector Machines (SVM), Naïve Bayes (NB), Decision Trees (DT) and K-nearest neighbor (KNN) for the early diagnosis of breast cancer through Rapid Miner Tool. Finally, it can be concluded that the algorithm of SVM is better in the prediction when compared with other six algorithms of the existence and non-existence with 96% of the disease.

Venkatesan et al. [8] analyzed the data of breast cancer by utilizing four classification algorithms such as j48, Alternating Decision Tree (AD Tree), Classification and Regression Trees (CART), and Best First Tree (BF Tree). By the Weka tool, the experiment is performed. The classifier has applied for two test beds –cross validation which uses 10 folds with 9 folds used for training each classifier and 1 fold is used for testing and percentage split uses 2/3 of the dataset for training and 1/3 of the dataset for testing. It can be stated by the authors that the decision trees have a standard construct and easy to understand from which the rules can be extracted. Hence, by the analysis of j48 classifier having the accuracy which is highest with 99%.

Majali et al. [9] have expressed a diagnostic system by which utilizing classification and association approach in data mining. The Frequent Pattern (FP) is utilized in association rule mining for dividing the patterns which are frequently found with benign and malignant instances. The utilization of decision tree algorithm is to predict the possibility of cancer by the age. The authors implemented Fp-growth algorithm for creating the frequent item set without candidate generation which improves the algorithm performance. The authors stated that their algorithm is capable to achieve 94% of prediction accuracy.

Sivakami [10] has stated a prediction for the status of disease by retaining a methodology which is hybrid of both Decision Trees (DT) and Support Vector Machines (SVM). The severity of the disease is identified by the strategy of the system which contains two main parts such as information treatment and option extraction, and decision tree- support vector machines. The results which are obtained by the presented model are differentiated with Instance-based Learning (IBL), Naïve Bayes (NB) and Sequential Minimal Optimization (SMO). Hence, a proposed algorithm provides 91% of accuracy which is better than the comparative algorithms.

The j48 decision tree classification algorithm is evaluated by the Sumbaly et al. [11]. This algorithm can predict breast cancer along with the summarization on the breast cancer types, disease symptoms, risk factors and treatment. Hence, the j48 algorithm having the ablility to give 94.5% of accuracy and also suggested that digital mammography and neural network would be the alternative proposals for prediction of breast cancer.

IV. RELATED WORK

Data Mining Techniques For Breast Cancer Analysis: Data mining is a modern field and a powerful technique having various techniques for evaluating the recent real world problems. The raw data is translated into information which is useful by these techniques in various fields of research and discover the patterns for deciding trends of future in medical field. There are various major data mining techniques which have been improved and utilized in projects of data mining for discovery of knowledge from database [1]. Breast Cancer is the one of the leading cause of death in women in developing countries and a second cause in developed countries as per the statistics of national

 ISSN: 2454-4248
529 – 535

cancer institute. The both male and female has a chance for the occurrence of breast cancer. But in female the occurrence is high throughout the world. Breast cancer is an asymptomatic nodule on a mammogram. A modern breast symptom must be taken by their doctors and patients by the probability of an underlying breast cancer at any age.

There are various methods in data mining. Different methods perform various purposes, each method offering its own benefits and limitations. There are the two most common techniques of data mining classification and clustering which are utilized in field of medical science. Additionally, most methods of data mining contains classification category as the applied prediction approaches which assign the patients to a"benign" group that is non-cancerous or a group of "malignant" which is cancerous and produced rules for the similar.

Hence, the diagnostic problems of breast cancer are in the scope of the mostly discussed classification problems. One of the most significant tasks is classification in data mining. The data is mapped in to targets which are predefined. As targets are predefined, it's known as a supervised learning. The major objective of classification is to assemble a classifier which is based on some cases with some attributes for describing the objects or one attribute to regulate the objects category. Then, utilization of classifier is to forecast the attributes of group from the domain depends on the values of some other attributes.

V.MAMMOGRAPHY DATA SET ANALYSIS

The data mining is data analyzing techniques that used to analyze Mammography Data set beforehand stored from various resources to find patterns and trends in Brest Cancer. In additional, it can be applied to enlarge efficiency in detecting severity of Brest cancer. Several studies have discovered various techniques to Detecting the severity of Brest cancer that used too many applications. Such studies can help to speed up the process of detecting the severity of Brest cancer and help the huge data are very difficult and complex. We can take 961 recodes mammography data set. Among them the 720 are in Training Set and 261 are in testing set class data.

WE HAVE 6 ATTRIBUTES RELATED TO MEMOGRAPHY DATA SET

BI-RADS Assess-ent	Age	Shape	Margin	Density	Severity
5	67	3	5	3	1
4	43	1	1	NA	1
5	58	4	5	3	1
4	28	1	1	3	0
5	74	1	5	NA	1
4	65	1	NA	3	0
4	70	NA	NA	3	0
5	42	1	NA	3	0
5	57	1	2	4	0
5	60	NA	5	1	1
5	76	1	4	3	1
3	42	2	1	3	1
4	64	1	NA	3	0

TABLE 1: DATA PREPROCEESING DATA SET ANYASIS

AMELIA (Multiple Imputations of Incomplete Multivariate Data) PACKAGE: Amelia "multiply imputes" data is missing in a single cross-section as a examine, from Data set of Mammography. Our algorithm which is based on bootstrapping is implemented by Amelia II which gives essential the similar answers as the standard IP or EM is approaches and is usually rapid than existing approaches and can handle the numerous variables. Other statistically rigorous imputation software is utilized and it is virtually never crashes. Amelia II also contains functional diagnostics which is fit of multiple imputation models. The program is worked from the R command line or through a graphical user interface which does not require users to know R. The following values are implemented in place of missing values in the Mammography data set.

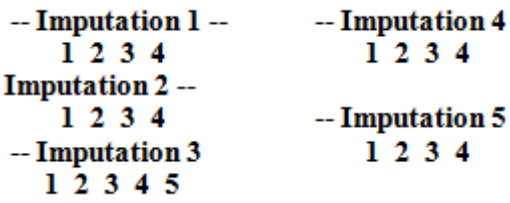

FIG 3: AMELIA IN R PROGRAMMING IT FILLS MISSING VALUES

VI.CLASIFICATION TECHNIQUES

Naive Bayes (NB): The Naive Bayes is a quick method to create statistical predictive models. NB is depends up on the Bayesian theorem. A conditional probability is extracted for the relationships among the attribute values and the class, this classification technique is utilized for analyzing the

 ISSN: 2454-4248
529 – 535

relationship among each attribute and the class for each instance. The probability of each class is enumerated during training by counting how many times it appears in the dataset of training.

This is known as the "prior probability" P(C=c). Moreover, the algorithm also evaluates the probability for the instance x given c with the assumption that the attributes are independent. This probability is the consequence of the possibilities of each single attribute. Then the probabilities can be estimated from the frequencies of the instances in the training set.

Now the following figure 4 shows the Navi Bayes algorithm run in R studio and creates confusion matrix and severity data results.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  95  20
         1  14 112

Accuracy          : 0.8589
95% CI            : (0.8085,0.9003)
No Information Rate : 0.5477
P-Value [Acc > NIR] : <2e-16
               kappa : 0.7166
 McNamara's Test P-Value : 0.3912
       Sensitivity : 0.8716
       Specificity : 0.8485
 Pos Pred Value : 0.8261
 Neg Pred Value : 0.8889
     Prevalence : 0.4523
 Detection Rate : 0.3942
 Detection Prevalence : 0.4772
 Balanced Accuracy : 0.8600
 'Positive' Class : 0
```

FIG 4. THE NAVI BAYES ALGORITHM RUN IN R STUDIO

Decision Tree (C4.5): The Decision tree, each node which is non-terminal presents a test or decision on the data item which is considered. Choice of a definite branch based on the test outcome. For dividing a specific data item, we initialize at root node and follow the assertions down till by reaching a terminal node (or leaf). A decision is created when a terminal node is approached. Decision trees also can be interpreted as a set of rule which is in special form and their hierarchical organizations of rules are characterized.

By using decision tree method data can display decision tree diagram shown in below figure 5.

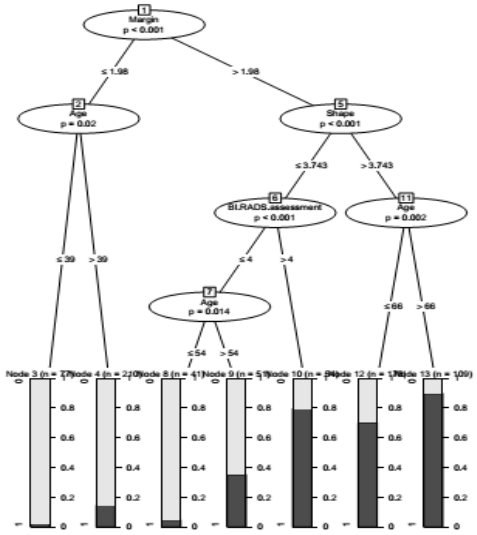

FIG 5. DECISION TREE DIAGRAM

The Confusion Matrix and Statistics for the Decision Tree Algorithm run in R-studio using Weka Tool are as following.

```
Confusion Matrix and Statistics
          Reference
Prediction   0   1
         0 102  26
         1  23  90

      Accuracy : 0.7967
        95% CI : (0.7403, 0.8456)
   No Information Rate : 0.5187
  P-Value [Acc > NIR] : <2e-16
         Kappa : 0.5924
Mcnemar's Test P-Value : 0.7751
       Sensitivity : 0.8160
       Specificity : 0.7759
    Pos Pred Value : 0.7969
    Neg Pred Value : 0.7965
        Prevalence : 0.5187
    Detection Rate : 0.4232
Detection Prevalence : 0.5311
  Balanced Accuracy : 0.7959
   'Positive' Class : 0
```

FIG 6. THE CONFUSION MATRIX AND STATISTICS FOR THE DECISION TREE ALGORITHM

SUPPORT VECTOR MACHINE: In statistics and computer science, SVM is referred as supervised learning method. SVM calculates the data and recognize patterns which are utilized for classification and analysis of regression. A set of data is considered in the standard SVM as input and predicts for each given input. By the two possible classes, the input is attained by creating the SVM as a non-probabilistic linear classifier which is in binary.

 ISSN: 2454-4248
529 – 535

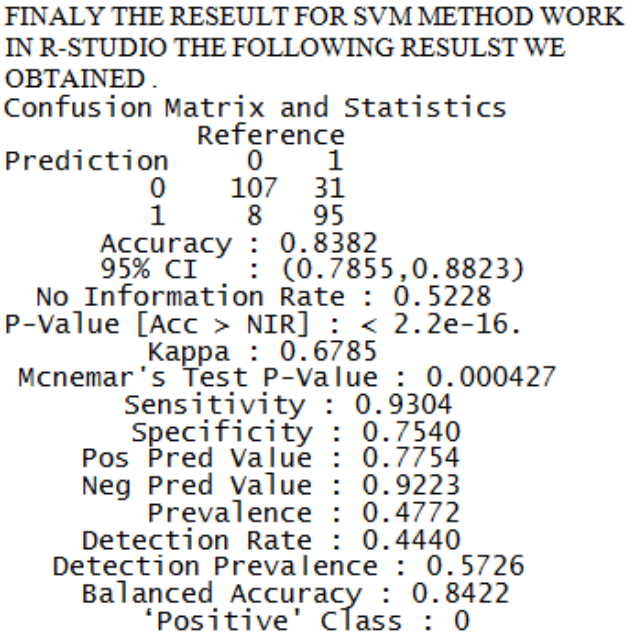

FIG 7. SVM METHOD

The following results are obtained for the Comparative Study purpose.

Sl.No	Classifier algorithm	Accuracy	Sensitivity	Specificity	Detection Rate(+ve Class)
1	Navie Bayes	0.8589	0.8716	0.8485	0.3942
2	SVM	0.8382	0.9304	0.7540	0.4440
3	Decision Tree	0.7967	0.8160	0.7759	0.4232

TABLE 2. COMPARISON TABLE FOR DIFFERENT CLASSIFIER ALGORITHMS

VII.CONCLUSION

Based on the following result table we have conclude that the Navie Bayes is having the highest priority and then , SVM and then Decision Tree algorithms are to best classification for breast cancer based on mammographic Data Set. The application of data mining techniques for predictive analysis is very important in the health field because it gives the power to face diseases earlier and therefore save people's lives through the anticipation of cure. In this work, we utilized three learning algorithm Decision Tree(C4.5), SVM and NB, for predicting patients with **Severity of breast cancer**, and patients who are not suffering from breast cancer.

The paper presented classification Techniques for determining the stage of breast cancer malignancy based on cancer **Sensitivity, Accuracy, Specificity and Detection Rate** on mammogram DATA SET. These methods are tested 961 MEMOGRAPHY DATA SET RECODS FINALY WE GET USING THIS METHODS WE GET 85% OF ACCURATE RESULT to Predict the Severity of breast cancer Simulation results showed that NB classifier proved its performance in predicting with best results in terms of accuracy and minimum detection rate. The diseases which are Anticipating still remains a vital challenge in medical field and pushes us to maximize our efforts in improving more machine learning algorithms to exploit information intelligently and extract the best knowledge from it.

VIII.REFERENCES

[1] American Cancer Society, "Breast Cancer Facts & Fig. s, 2013-2014", American Cancer Society, Inc, 2013.

[2] Acha, B., Rangayyan, R.M., Desautels, J.E.L., "Detection of Micro calcifications in Mammograms", In: Suri, J.S., Rangayyan, R.M. (eds.) Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer. SPIE, Bellingham, 2006.

[3] Highnam R and Brady M, "Mammographic Image Analysis", Kluwer Academic Publishers, British Journal of Radiology, 74(887), 2001.

[4] Maitra, I.K., Nag S., Bandyopadhyay S.K., "Identification of Abnormal Masses in Digital Mammography Images", International Journal of Computer Graphics, 2(1), 2011.

[5] Bozek, J., Mustra, M., Delac, K., and Grgic, M., "A Survey of Image Processing Algorithms in Digital Mammography", Multimedia Signal Processing and Communications Studies in Computational Intelligence Volume 231, 2009, pp 631-657.

[6] Yasmin, M., Sharif, M., and Mohsin, S., "Survey Paper on Diagnosis of Breast Cancer Using Image Processing Technique", Research Journal of Recent Sciences, Vol. 2(10), 88-98, October 2013.

[7] American Joint Committee on Cancer, "Breast Cancer Staging", AJCC 7th Edition Staging Posters, 2009.

[8] Pradeep, N., Girisha, H., Sreepathi, B., and Karibasappa, K., "Feature Extraction of Mammograms", International Journal of Bioinformatics Research, Volume 4, Issue 1, 2012, pp.-241-244.

[9] Seguin, M., and Sankuru, B., "Survey over image thresholding techniques and quantitative performance evaluation". Journal of Electronic Imaging, 13 (1): 146–165. 2004.

[10] Maitra, I.K., Nag S., Bandyopadhyay S.K., "A Novel Edge Detection Algorithm for Digital Mammogram", International Journal of Information and Communication Technology Research, Vol 2 No.2, February 2012.

[11] Kamdi, S.," Image Segmentation and Region Growing Algorithm", International Journal of Computer Technology and Electronics Engineering (IJCTEE), Vol 2 Issue no.1, October 2011.

[12] Priya, D.S., and Sarojini, B., "Breast Cancer Detection in Mammogram Images Using Region-Growing and Contour Based Segmentation Techniques", International Journal of Computer & Organization Trends, Vol.3 Issue 8, September 2013.

 ISSN: 2454-4248
529 – 535

[13] Martins, Lido., Junior, G.B., Silva, A.C., Paiva, A.C. and Gattass, M.,"Detection of Masses in Digital Mammograms using K-Means and Support Vector Machine", Electronic Letters on Computer Vision and Image Analysis 8(2) : 39-50, 2009.

[14] Suckling, J., "The Mammographic Image Analysis Society Digital Mammogram Database", Excerpt Medical, International Congress 1069, pp 375-378, 1994.

AUTHORS' PROFILE

Karavadi Laksmi Venu Gopala Krishna Murthy, pursuing Ph.D from Rayalaseema University in Computer Science, under the guidance of Dr.R.J.Rama Sree, HoD&Dean, CS Department, Rashtriya Sanskrit Vidya Peeth, Thirupathi, received his M.Tech in computer science and Engineering from JNTU Hyderabad University. He is having 16 years of teaching experience. and he is published various research papers in national conferences/ Journals and international journals/ Conferences .His research areas of interest are Cloud Computing, Data Structures, Parallel processing, Data warehousing and Data Mining.

Dr. R.J.Rama Sree, HOD&Dean, Rashtriya Sanskrit Vidya Peeth, , Department of Computer Science, Tirupathi. She Received her Ph.D from S.V.Women's University, Thirupathi, received M.S in Software System from BITS,Pilani. She has 20 years of teaching experience and published more than 25 National and International Journal Papers. She conducted so many workshops and conferences as a part of Curriculam. Her interested research areas are Data Mining, Natural Language Processing and Internet Of Things. She guided so many research students in their research areas.