

Accurate Breast Cancer Prediction using Machine Learning Techniques

Vinoth S. M. E., P. Valarmathi



ABSTRACT: Applications of machine learning (ML) have been increasing widely in various fields like recommendation, fault identification and disease prediction. In ML, different algorithms were available and utilized in disease prediction such as heart disease prediction, cancer prediction and other forms disease prediction. In our proposed work, breast cancer prediction using ML has been implemented. Initially breast cancer images have been taken as input, preprocessing steps will be done to remove noisy and irrelevant data from image. Then 2D median filter is a nonlinear operation often used in image processing to reduce "salt and pepper" noise. To increase contrast of image contrast-limited adaptive histogram equalization is used. Segmentation has been implemented and GLCM feature extraction is deployed based on this information classification is implemented. For accurate classification Artificial neural network (ANN) is used to predict whether the patient is affected by breast cancer or not. Compared to other existing method our method predicts results in accurate way.

Keywords: Breast cancer prediction, segmentation, feature extraction, classification and prediction.

I. INTRODUCTION

Among ladies, Breast cancer has gotten one of the most widely identified sickness that prompts passing. By arranging tumors Breast disease can be analyzed. Threatening and considerate tumors are two unique kinds of tumors. To recognize these tumors doctors need a dependable finding method.

In any case, for the most part it is exceptionally hard to separate tumors even by the specialists. For diagnosing tumors subsequently robotization of symptomatic framework is required. To apply AI calculations for recognizing survivability of malignant growths in people numerous specialists have endeavored and it is additionally been demonstrated by the scientists that these calculations work better in identifying disease determination. In identifying malignant growth in human this paper bridges the utilization of AI calculations.

Neural networks:

By the way human sensory systems a neural network is a model that is planned, for example, mind that procedure the data.

To get significance from convoluted or loose information neural systems with their amazing capacity, can be utilized to remove designs and identify patterns that are too intricate to possibly be seen by either people or other PC methods. Numerous neural system models, even organic neural systems expect numerous rearrangements over real natural neural systems.

Such improvements are important to endeavor any numerical examination and to comprehend the planned properties. For diagnostic reason regardless of whether all the properties of the neurons are known, rearrangements are as yet required. Neural systems are versatile factual gadgets. As a component of their presentation this implies they can change (synaptic loads). In ANNs, all the neurons are working simultaneously, which makes ANN to perform errands at a lot quicker rate.

Imaging tests:

Mammogram, Magnetic resonance imaging (MRI) of breast, Ultrasound of breast X-beam of the breast, Tissue biopsy: by a pathologist removal of the tissue of the breast for assessment. Sentinel hub biopsy: Once breast malignant growth is affirmed, patients normally experience sentinel hub biopsy. This assists with recognizing harmful cells in lymph hubs to affirm metastasis of breast malignancy into lymphatic framework. Whenever required, oncologist may likewise arrange extra tests or strategies. In the customary method for diagnosing breast malignant growth a few tests and strategies are done. These tests incorporate Breast test Mammogram Breast ultrasound Biopsy. As an elective we can likewise utilize Machine Learning methods for the grouping of amiable and threatening tumors.

Contribution of the paper:

This process of classifying benign and malignant tumors can be best done by the application of Classification techniques of ML. In this area lot of research is being conducted by the application of various ML and data mining techniques for many different datasets on Breast Cancer.

II. RELATED WORK

Ch. Shravya (2019), presents increasing mortality rate among women breast cancer is the most often identified cancer among women and major reason. As the conclusion of this illness the lesser accessibility of frameworks and physically takes extended periods, there is a need to build up the programmed analysis framework for early location of malignant growth. Information mining procedures contribute a great deal in the advancement of such framework. For the characterization of favorable and dangerous tumor we have utilized order methods from the past information and can anticipate the classification of new info of AI in which the machine is found out.

Manuscript received on February 10, 2020.

Revised Manuscript received on February 20, 2020.

Manuscript published on March 30, 2020.

* Correspondence Author

Vinoth S. M. E*, Department of Computer Science and Engineering, Mookambigai College of Engineering, Tamilnadu, India.

Dr. P. Valarmathi, Professor and Head Department of Computer Science and Engineering, Mookambigai College of Engineering, Tamilnadu, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

This paper is a relative report on the usage of models utilizing Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) is done on the dataset taken from the UCI archive. Regarding the consequences of exactness, accuracy, affectability, explicitness and False Positive Rate the effectiveness of every calculation is estimated and looked at.

B.M.Gayathri et.al (2013), describes breast cancer has become a common factor now-a-days. Notwithstanding the reality, not every single general medical clinic have the offices to analyze breast disease through mammograms. Hanging tight for diagnosing a breast malignant growth for quite a while may build the chance of the disease spreading. Accordingly a modernized breast malignancy conclusion has been created to diminish the time taken to analyze the breast disease and lessen the passing rate. This paper outlines the review on breast malignancy conclusion utilizing different AI calculations and strategies, which are utilized to improve the exactness of foreseeing disease. This review can likewise assist us with knowing about number of papers that are actualized to analyze the breast disease.

Bichen Zheng et.al (2013), describes the development of clinical technologies, different tumor features have been collected for breast cancer diagnosis. Separating all the appropriate component data to help the clinical sickness conclusion is a difficult and tedious undertaking. The target of this examination is to analyze breast malignant growth dependent on the extricated tumor highlights. Highlight extraction and choice are basic to the nature of classifiers established through information mining strategies. To separate helpful data and analyze the tumor, a half and half of K-means and bolster vector machine (K-SVM) calculations is created. The K-implies calculation is used to perceive the shrouded examples of the amiable and dangerous tumors independently. The enrollment of every tumor to these examples is determined and treated as another element in the preparation model. At that point, a help vector machine (SVM) is utilized to acquire the new classifier to separate the approaching tumors.

Kathija, Shajun Nisha (2016), describes the major problems for women that have increased over years. A notable articulation in malignancy society is "Early location implies better odds of endurance". So early identification is essential as to forestall breast malignant growth with progress and lessen profound quality. One of the most dynamic regions of research in directed AI is to read strategies for building great groups of students. The goal of this paper is to discover littlest subset of highlights from Wisconsin Diagnosis Breast Cancer (WDBC) dataset by applying disarray lattice precision and 10-crease cross approval technique that can guarantee profoundly exact troupe order of breast disease as either favorable or dangerous. For arrangement, the breast malignancy information were first grouped by Support Vector Machine (SVM) and Naïve Bayes classifiers, and afterward settle the order procedure.

Babak Bashari Rad and Mervat Adib Bamiah (2018), discusses a few kinds of research have been done on early recognition of breast malignant growth to begin treatment and increment the possibility of endurance. The majority of the examinations focused on mammogram pictures. Nonetheless, mammogram pictures once in a while have a danger of bogus recognition that may jeopardize the patient's wellbeing. It is essential to discover elective

techniques which are simpler to execute and work with various informational indexes, less expensive and more secure, that can deliver a progressively dependable forecast. A cross breed model consolidated is presented here that consist of a few Machine Learning (ML) calculations including Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Decision Tree (DT) for powerful breast malignancy location. The proposed model can be utilized with various information types, for example, picture, blood, and so on.

Wenbin Yue et.al (2018), presents clinical treatment to patients and it elevates the early conclusion of BC can improve the guess and possibility of endurance altogether. Further exact order of amiable tumors can forestall patients experiencing superfluous medicines. In this way, the subject of a lot of research is the right determination of BC and characterization of patients into threatening or kind hearted gatherings. In basic highlights discovery from complex BC datasets, it's one of kind points of interest, AI (ML) is broadly perceived as the strategy of decision in BC design arrangement and estimate demonstrating. Right now, intend to audit ML methods and their applications in BC finding and forecast.

III. PROPOSED METHODOLOGY

In our paper we detect Breast cancer using image processing techniques in thermal Images. The image is acquired and pre processed for further process of Breast Cancer detection. After pre processing the cancerous part is segmented using Adaptive K-means algorithm. Extract some Features from image for Classification of normal and abnormal cancerous tissues. The features are Statistical features. For Classification we use Neural Network Classifier.

Image Acquisition:

Image Acquisition is a process of getting an input image for classify the normal and abnormal tissues.

Pre-Processing:

For undertakings with the photos Pre preparing is a normal name at the most diminished level of consideration both data and yield is the data pictures. The purpose of pre taking care of is an improvement of picture data that smoother unfortunate picture data winds or overhaul the some image features huge for the further getting ready. Four classes of picture pre-getting ready procedures as showed by the size of pixel neighborhood that is used for the figuring of new pixel brightness:

Pixel brightness changes

Geometric changes

Pre-preparing techniques that utilization a nearby neighbourhood of the processed pixel,

Picture rebuilding that requires information about the whole picture. If pre processing aims to address some corruption in the picture, the nature of from the earlier data is significant:

1. extremely broad properties of the debasement and Knowledge about the idea of the debasement without a doubt, are expected, 2.

Knowledge about the properties of the picture procurement gadget, the nature of commotion (for the most part its unearthly attributes) is here and there known, 3. Knowledge about articles, which may streamline the pre-handling impressively that are looked for in the picture. During the handling if information about items isn't accessible ahead of time it very well may be evaluated.

Segmentation:

The path toward isolating an electronic picture into various bits (sets of pixels, in any case called super-pixels) is Image division. The depiction of an image into something that is logically significant and more straightforward to inspect is the goal of division is to unravel and also change. To discover things and cut-off points (lines, twists, etc.) in pictures, Picture division is normally utilized. Even more accurately, the route toward allocating an imprint to every pixel in an image is to such a degree, that pixels with a comparable name share certain properties picture division. A lot of shapes removed from the picture or a lot of portions that aggregately spread the whole picture are the result of image segmentation. All of the pixels in an area is practically identical for some trademark or prepared property, for instance, concealing, force, or surface. Contiguous areas are essentially unique regarding the equivalent characteristic(s). A few universally useful calculations and methods have been created for picture division. To be valuable, these strategies should ordinarily be joined with a space's particular information so as to successfully take care of the area's division issues.

Segmentation Using Adaptive K-means Clustering:

At the point when you have unlabeled data K-implies grouping is a sort calculation which is used. The target of this figuring is to find bunches in the data with the amount of social occasions addressed by the variable K. To one of K groups reliant on the features that are given the estimation works iteratively to give out each datum point. In view of highlight likeness information focuses are grouped. The aftereffects of the K-implies bunching calculation are:

1. The centroids of the K clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before taking a gander at the information, bunching permits you to discover and examine the gatherings that have framed naturally. The quantity of gatherings can be resolved by "Picking K" area portrays. In the subsequent gatherings which characterize every centroid of a cluster is an assortment of highlight esteems. Analyzing the centroid include loads can be utilized to subjectively decipher what kind of group each cluster represents.

System architecture:

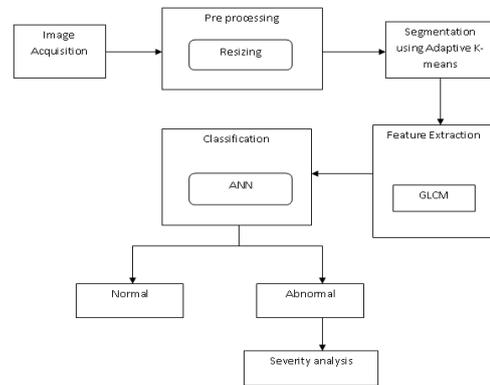


Figure 1: working of proposed system

Feature Extraction:

In machine learning, design acknowledgment and in picture preparing, include extraction begins from an underlying arrangement of estimated information and fabricates determined qualities (highlights) planned to be enlightening and non-excess, encouraging the resulting learning and speculation steps, and at times prompting better human understandings. Highlight extraction is identified with dimensionality decrease.

At the point when the information to a calculation is too huge to be in any way prepared and it is suspected to be excess, at that point it very well may be changed into a decreased arrangement of highlights. Deciding a subset of the underlying highlights is called include choice.

Highlight extraction portrays the significant shape data contained in an example so the assignment of arranging the example is made simple by a conventional strategy. Highlight extraction is done after the preprocessing stage in character acknowledgment framework. The essential errand of example acknowledgment is to take an info design and accurately relegate it as one of the conceivable yield classes. This procedure can be separated into two general stages: Feature determination and Classification. Highlight determination is basic to the entire procedure since the classifier won't have the option to perceive from inadequately chosen highlights. Criteria to pick highlights given by Lippman are: "Highlights ought to contain data required to recognize classes, be harsh toward superfluous changeability in the info, and furthermore be constrained in number, to allow, proficient calculation of discriminant capacities and to confine the measure of preparing information required".

Classification

So as to characterize a lot of information into various classes or classifications, the connection between the information and the classes into which they are ordered must be surely known.

Significant parts of exact grouping

- Learning procedures
- Feature sets

Highlights are qualities of the information components dependent on which the components are allotted to different classes. The quantity of classes, model pixels for each class can be recognized utilizing earlier information.

Artificial Neural Network:

Artificial Neural Networks (ANN) is at present a 'hot' look into region in medication and it is accepted that they will get broad application to biomedical frameworks in the following hardly any years. Right now, the exploration is generally on demonstrating portions of the human body and perceiving sicknesses from different outputs (for example cardiograms, CAT examines, ultrasonic sweeps, and so on.).

Neural networks are perfect in perceiving ailments utilizing filters since there is no compelling reason to give a particular calculation on the most proficient method to distinguish the infection. Neural systems learn by model so the subtleties of how to perceive the illness are not required.

What is required is a lot of models that are illustrative of the considerable number of varieties of the illness. The amount of models isn't as significant as the 'amount'. The models should be chosen cautiously if the framework is to perform dependably and productively.

Artificial Neural Networks (ANNs) are supporting devices for picture handling, regardless of whether right now they are never again considered as the default best answer for any order or relapse issue. ANNs moderate their job as non-parametric classifiers, non-straight relapse administrators, or (UN) regulated element extractors. The part audits the utilizations of ANN system in all the means of the picture preparing chain, beginning from information pre handling and decrease, picture division, up to protest acknowledgment and scene understanding.

IV. RESULT AND DISCUSSION

The session discusses result obtained for every step is explained briefly and effective identification of breast cancer has been identified. Initially image has been chosen for processing and further steps will be processed.

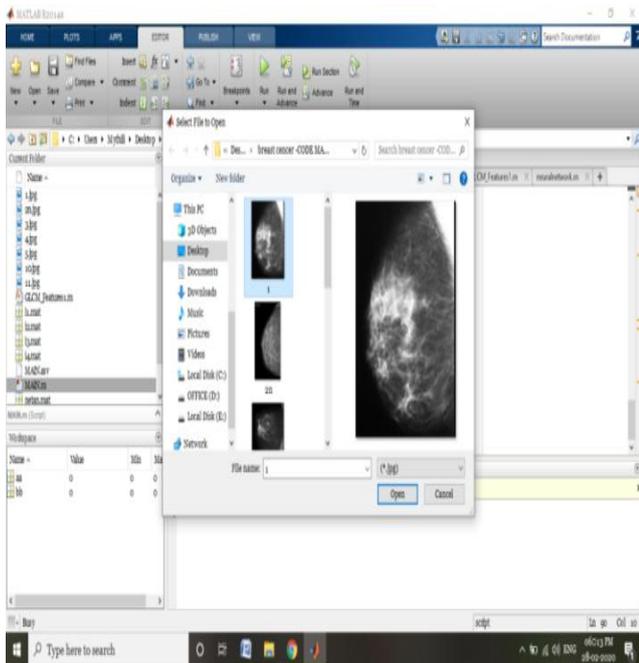


Figure 2: select image to be processed

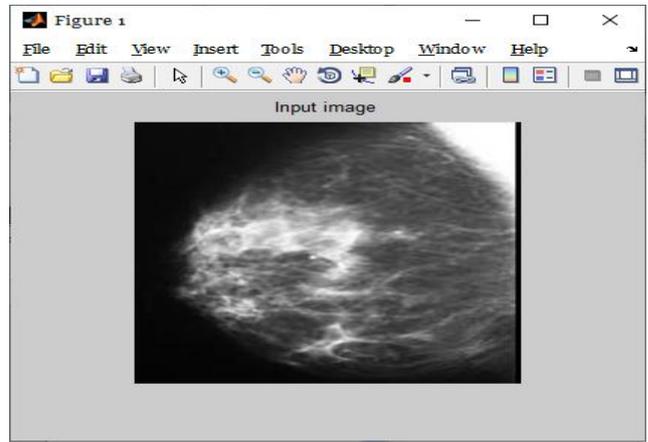


Figure 3: Input image

A median filter is more effective than convolution when the goal is to simultaneously reduce noise and preserve edges. Output obtained after application of median filter.

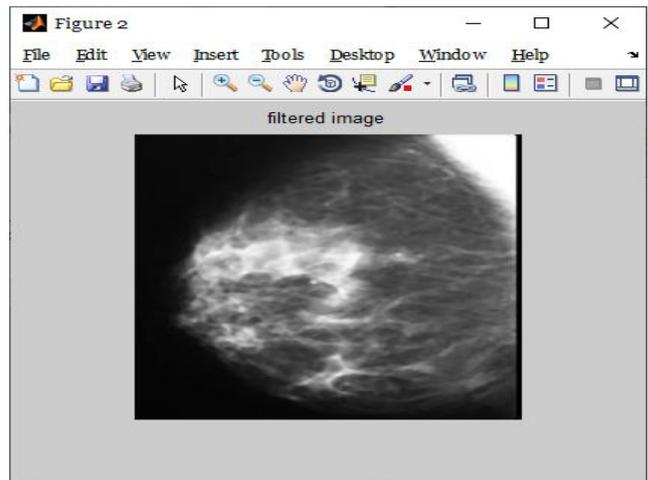


Figure 4: Median filtered image

CLAHE operates on small regions in the image, called tiles, rather than the entire image. adaptive histogram equalization calculates the contrast transform function for each tile individually. Each tile's contrast is enhanced, so that the histogram of the output region approximately matches the histogram specified by the 'Distribution' value.

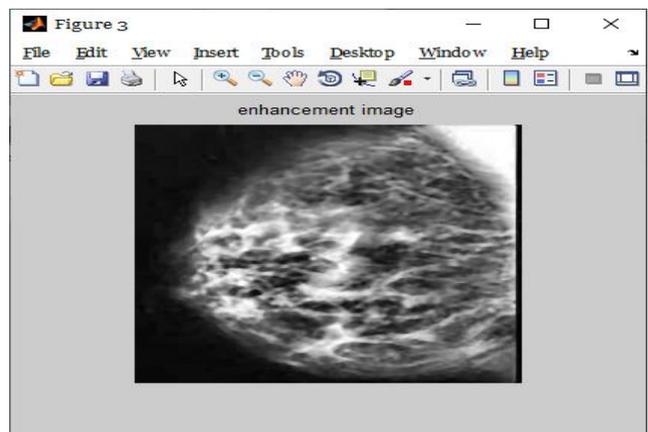


Figure 5: contrast enhanced image

K-means clustering algorithm is an unsupervised algorithm and it is used to segment the interest area from the background.

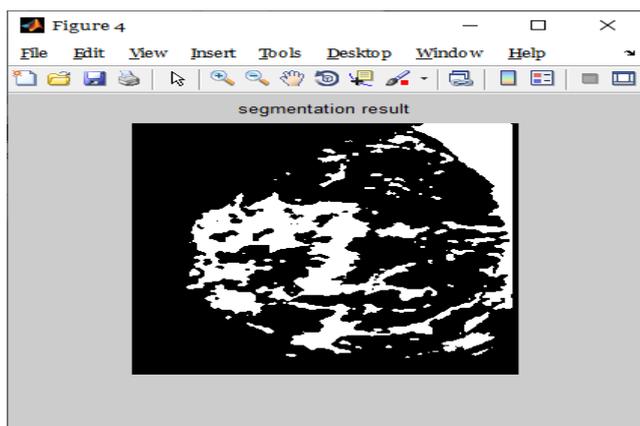


Figure 6: segmented image

Result obtained:

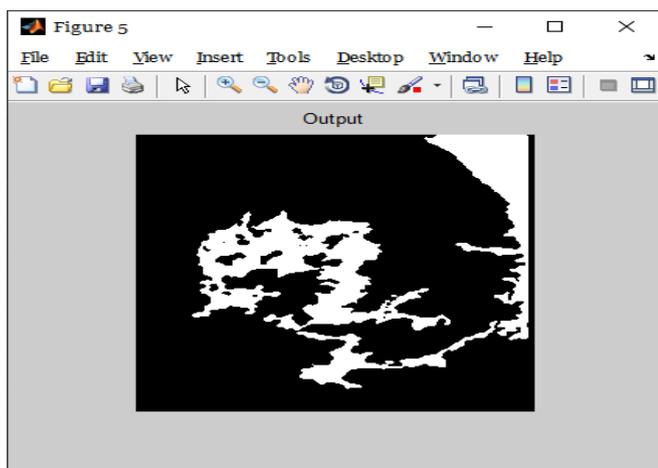


Figure 7: cancer detected image

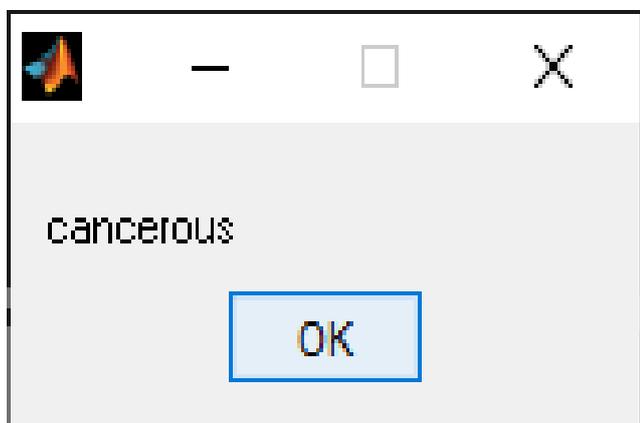


Figure 8: breast cancer affected

V. CONCLUSION:

Detection of breast cancer growth through picture preparing is a productive strategy. Right now of disease is finished by picture preparing strategies. Anyway pre-preparing, division, include extraction and arrangement have different calculations. In our work, 2-D middle channel was utilized for best commotion expulsion and for upgrading difference

of picture versatile histogram balance was actualized. The yield acquired from these two procedure is exposed to include extraction here required piece of the picture has be extricated and experienced grouping. Outstanding amongst other grouping calculations is ANN which achieves most extreme precision in forecast. Consequently our outcome achieves most extreme outcomes contrasted with different techniques.

REFERENCES

1. Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-6, April 2019.
2. B.M.Gayathri ,C.P.Sumathi and T.Santhanam, "breast cancer diagnosis using machine learning algorithms –a survey" International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013.
3. Bichen Zheng, Sang Won Yoon, Sarah S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid 4 of K-means and support vector machine algorithms" Expert Systems with Applications 5 September 2013.
4. Kathija, Shajun Nisha, "Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques" International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 12, December 2016.
5. Babak Bashari Rad and Mervat Adib Bamiah, "Early Detection of Breast Cancer Using Machine Learning Techniques" Journal of Telecommunication, Electronic and Computer Engineering 30 September 2018.
6. Wenbin Yue 1 ID , Zidong Wang 1, Hongwei Chen 2 ID and Annette Payne 1 and Xiaohui Liu, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis" Designs 2018.
7. Gouda I. Salama, Abdelhalim and Magdy Abd-elghany Zeid, "Breast Cancer Diagnosis On Three Different Datasets Using Multi-Classifiers", International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012.
8. Pitchumani Angayarkanni, Nadira Banu Kamal, "Automatic Classification Of Mammogram MRI Using Dendograms", Asian Journal Of Computer Science And Information Technology 2: 4, 78, 81, 2012.
9. Sadhana and Sankareswari, "A Proportional Learning Of Classifiers Using Breast Cancer Datasets", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 3, Issue. 11, pg.223 – 232, November 2014.
10. LeenaVig , "Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset", Open Access Library Journal, Volume 1 | e660,2014.

AUTHORS PROFILE



Dr. P. Valarmathi, is currently a Professor and Head in the department of Computer Science and Engineering at Mookambigai College of Engineering, Keeranur, Pudukkottai District Tamilnadu, India. She obtained her M.E. from Vinayaka Mission University, Ariyanur, Tamilnadu, India and Ph.D. from Anna University, Chennai, Tamilnadu, India. Her research interests include image Processing and data mining. She is a lifetime member of CSI Technical Education.



S. Vinoth, is currently pursuing his Masters of Engineering in Computer Science and Engineering at Mookambigai College of Engineering, Keeranur, Pudukkottai District Tamilnadu, India. He was completed his undergraduate in Thangavelu Engineering College at Chennai Tamilnadu, India in the stream of Computer Science and Engineering.