International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

Mammogram Image Classification Using Machine Learning

M. Vanitha[#1], Dr. Sheeja V Francis[*2]
[#]Student, Department of Electronics and Communication Engineering, Jerusalem College of Engineering, Chennai, Tamil Nadu, India
[*]Head of Department, Department of Electronics and Communication Engineering, Jerusalem College of Engineering, Chennai , Tamil Nadu, India

ARTICLEINFO

Article History:
Accepted: 10 April 2023
Published: 29 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
747-750

ABSTRACT

All over the world, breast cancer is the second leading cause of death in women above 40 years of age. To design an efficient classification system for breast cancer diagnosis, one has to use efficient algorithms for feature selection to reduce the feature space of mammogram classification. The current work investigates the use of the hybrid genetic ensemble method for feature selection and classification of masses. A genetic algorithm (GA) is used to select a subset of features and to evaluate the fitness of the selected features, Adaptive boosting (AdaBoost). The selected features are used to classify masses into benign or malignant using AdaBoost classifiers. The results obtained with the proposed method are better when compared with extant research work.

Keywords : Genetic Algorithm, feature selection, benign, malignant, Adaboost classifier.

I. INTRODUCTION

Breast cancer poses a significant threat to women health and is considered the second leading cause of death in women breast cancer is the result of abnormal behavior of normal breast cells which is why breast cells tend to grow uncontrollably and form a tumor that appears as a lump in the breast. Among the various diagnostic methods such as x-ray and ultrasound of the breast digital mammography is the most reliable and cost-effective method to detect the symptoms of early breast cancer and can provide a lot of information about these abnormalities such as masses microcalcifications architectural distortions and bilateral asymmetry. It is advisable that all women who are 35 years and older get screened regularly to prevent this disease digital mammography has many advantages the patient has to spend less time screening the radiologist can quickly forward the images to another physician and they can be easily processed data mining is a machine learning uses mathematical and statistical models to learn from data machine learning plays an important role in biomedical applications where measurement accuracy is a critical factor, for example, machine learning algorithms can help diagnose early breast cancer machine learning tools can identify most predictive features from complex and noisy datasets so false negative and false positive decisions can be significantly reduced leading to better classification accuracy.

II. EXISTING SYSTEM

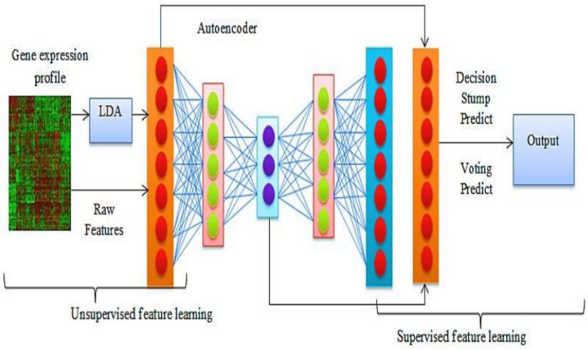


Fig. 1 Diagram of an existing system

III. PROPOSED SYSTEM

Through the proposed system, breast cancer is detected automatically through advanced processing techniques using a machine learning approach. Different parameters of the image are examined such as color conversion, resizing, and filtering. Segmentation algorithms are used to execute K-means algorithms. This information is useful in identifying how many lesions are located throughout the body. After GLCM is used to extract features, the mammogram image is classified using the AdaBoost algorithm. The message box also indicates whether the finding is benign, malignant, or normal. The results of this study help us gain a deeper understanding of machine learning algorithms that are useful for predicting early symptoms. AdaBoost is likely the most accurate classifier, as it has an accuracy rate of 98.24%.

IV. BLOCK DIAGRAM

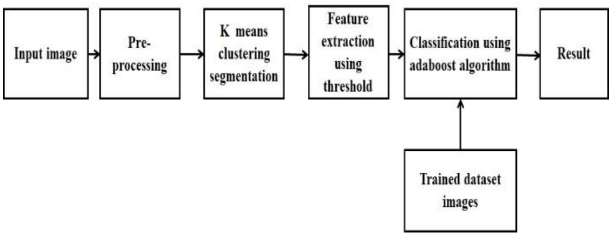

Input image
Pre-processing
K means clustering segmentation
Feature extraction using threshold
Classification using adaboost algorithm
Result
Trained dataset images

Fig. 2 Block diagram of the proposed system

V. ROCESSING STAGES OF THE SYSTEM

A. Data Collection

Mammography Image Analysis Society (MIAS) database was used in this research. Data is in the form of PGM (Portable Gray Map) format. In this research, 74 mammogram images are used for determining the various stages. Out of 74 images, 10 images are classified as normal, 13 images are classified as benign, and of the remaining 50 images, 6 belong to stage I, 19 belong to stage II, 15 belong to stage III, 6 belong to stage IV and 4 belong to stage V. After the acquisition of an image, we perform a median filter on the input image to pre-process it.

B. Pre Processing

Resize the image to make it usable. Adjust the image size so that all images are the same size and used as input. Image filtering is useful in many applications, including smoothing, sharpening, noise reduction, and edge detection. The filter is defined by the kernel, which is a small box applied to each pixel and its neighbors in the image. It is generally used to blur an image or reduce noise. Such reduction is a basic step to enhance the consequences of post-processing (image edge detection). The median filter is generally used to reduce noise in an image, similar to the average filter. However, it is often more effective at preserving useful image detail than the average filter. The median filter sequentially analyzes each pixel in the center and its friends to determine if it is an average advisor in its area. Instead of changing a pixel with the cue of neighboring pixel values, it changes it with the median of those values. The median is calculated by arranging all of the neighboring pixel values in numerical order and then replacing that pixel with a median pixel value. If the neighborhood in question contains an even number of pixels, the average of the two half-pixel values is taken.

C. Image Segmentation

Image segmentation is one of the most widely used methods to accurately classify pixels in an image in a decision orientation application. It divides an image into several distinct regions such that the pixels have high similarity in each region and high contrast between regions. There are different types of clustering: K-means clustering, Fuzzy C-means clustering, mountain clustering method, and subtractive clustering method. One of the most widely used clustering algorithms is K-means clustering. It is computationally simpler and faster than hierarchical clustering. It can also work for a lot of variables. But it produces different cluster results for different numbers of clusters. Therefore, it is necessary to initialize the correct number of clusters. The goal of K-means is simple: group similar data points together and discover underlying patterns. To achieve this goal, K-means searches for a fixed number (k) of clusters in the data set. A cluster refers to a data set of points aggregated due to several similarities. You will define a target number k, which will indicate the number of centroids you need in the data set. The centroid is the virtual or real location that represents the center of the cluster. Each data point is assigned to each cluster by reducing the sum of squares in the cluster. In other words, the K-means algorithm determines the k number of centroids and then allocates each component to the closest cluster, even keeping the cluster very small as much as possible. Mean" in K-means refers to the mean of the data; i.e. find the centroid.

D. Feature Extraction - GLCM

Given an image made up of pixels, each with an intensity (a specific gray level), GLCM is a histogram of how often different combinations of gray levels occur simultaneously in an image or an image section. The texture feature calculation uses the contents of the GLCM to give a measure of the intensity change. The Grayscale co-occurrence matrix method (GLCM) is a way to extract quadratic statistical texture features. The approach that has been used in several applications, ternary and higher order textures, considers relationships between three or more pixels. Theoretically, these are possible but rarely done due to computational time and difficulty of interpretation.

E. Adaboost Classification

Ada Boost classifier is a meta-estimator that event every classifier to the original dataset and then contests additional copies of the classifier to the same dataset, but balancing the weights of misclassified instances so that subsequent classifiers focus more on difficult essential facts. Ada Boost has many advantages due to its ease of use and less parameter adaptation compared to SVM algorithms. Ada Boost can further be used with SVM, despite hypothetically overfitting is not a feature of Ada Boost applications, perhaps because parameters are not agreeable all at once and the learning process is slowed cascading due to stepwise assessment. Ada Boost additionally be used to increment every accuracy of inclined classifiers and times in image/text classification.

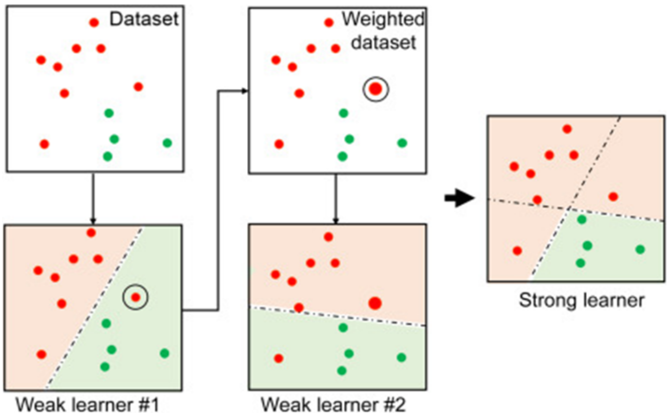
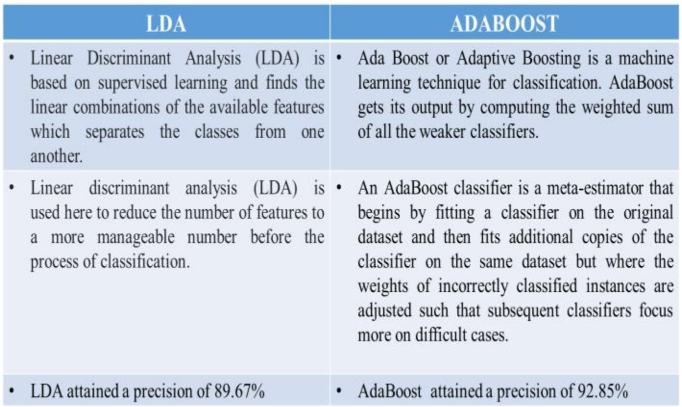


Fig. 3 Adaboost classifier

TABLE I
COMPARISON BETWEEN LDA AND ADABOOST



LDA	ADABOOST
• Linear Discriminant Analysis (LDA) is based on supervised learning and finds the linear combinations of the available features which separates the classes from one another.	• Ada Boost or Adaptive Boosting is a machine learning technique for classification. AdaBoost gets its output by computing the weighted sum of all the weaker classifiers.
• Linear discriminant analysis (LDA) is used here to reduce the number of features to a more manageable number before the process of classification.	• An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.
• LDA attained a precision of 89.67%	• AdaBoost attained a precision of 92.85%

VI. CONCLUSION

Breast cancer is a disease that disturbs entire premenopausal women but also younger youth class due to underlying risk factors analogous to lifestyle changes and elevated exposure to unnatural chemicals that have become a part of our activity. Risk components analogous to age, family history, and environmental circumstance be authorized to be the highest accepted explanation of malignancy. Image processing techniques alike median filtering, k-mean segmentation, GLCM, and AdaBoost classifier are explained and established in the proposed system. The biggest group of the activity that achieved the highest classification accuracy was used as input for training and testing the Ada Boost classifiers. The results of the design display that Ada Boost is first-rate in both RF and simple DT. Ada Boost achieved high classification accuracy and low FPR. The proposed method composes it conceivable to improve the examination of breast cancer.

VII. REFERENCES

[1]. Zhang, Xinfeng, Dianning He, Yue Zheng, Huaibi Huo, Simiao Li, Ruimei Chai, and Ting Liu. "Deep learning based analysis of breast cancer using advanced ensemble classifier and linear discriminant analysis." IEEE Access 8 (2020): 120208-120217.

[2]. Lu, Hao-Chun, El-Wui Loh, and Shih-Chen Huang. "The Classification of Mammogram Using Convolutional Neural Network with Specific Image Preprocessing for Breast Cancer Detection." In 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 9-12. IEEE, 2019.

[3]. Cahoon, Tobias Christian, Melanie A. Sutton, and James C. Bezdek. "Breast cancer detection using image processing techniques." In Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063), vol. 2, pp. 973-976. IEEE, 2018.

[4]. Huang, Min-Wei, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, and Chih-Fong Tsai. "SVM and SVM ensembles in breast cancer prediction." PloS one 12, no. 1 (2017): e0161501.

Cite this article as :

M. Vanitha, Dr. Sheeja V Francis, "Mammogram Image Classification Using Machine Learning", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 2, pp. 747-750, March-April 2023.
Journal URL : https://ijsrset.com/IJSRSET23102115