International Journal for Modern Trends in Science and Technology, 8(08): 203-211, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0808029
Available online at: http://www.ijmtst.com/vol8issue08.html

Diagnosis of Breast Cancer Using K-Means Clustering and C-Means Fuzzy

Zahra Ghaffari[1*], Hamidreza Erfanian[2]

Master of Bioinformatics, University of Science and Culture, Tehran, Iran

To Cite this Article
Zahra Ghaffari and Hamidreza Erfanian. Diagnosis of Breast Cancer Using K-Means Clustering and C-Means Fuzzy. International Journal for Modern Trends in Science and Technology 2022, 8(08), pp. 203-211. https://doi.org/10.46501/IJMTST0808029

Article Info
Received: 30 May 2022; Accepted: 15 June 2022; Published: 21 August 2022.

ABSTRACT

Nowadays, breast cancer is one of the most prevalent cancers in women that its timely diagnosis plays asignificant role in survival and treatment. New methods in image processing and machine learning have led to successful studies to establish breast cancer diagnostic systems by using thermography images. In this research, breast cancer diagnosis was conducted using K-Means clustering and C-Means fuzzy. In this regard, an intelligent methodwas provided for segregating healthy tissue from unhealthy and separating the mass in unhealthy tissue. In the recommended method, clustering was performed using two methods, K-Means and C-Means fuzzy. The results demonstrated that the C-Mean fuzzy method had more accurate results. After the clustering process, the cluster with the highest intensity cluster center was selected as the input of the area growth algorithm and the brightest pixel was selected as the granular point of the area growth method and area suspected to masswas determined according to area growth algorithm. Using the recommended method, an intelligent system was designed to reduce the amount of human error in the diagnosis of cancer and be able to diagnose the breast cancer in patients in the early stages of mass diagnosis.

Keywords: Breast Cancer, K-Means Clustering, C-Means Fuzzy

1. INTRODUCTION

Nowadays, progresses in technology and innovations have led to impressive results in the field of health, and computer diagnostic systems with the help of physicians have been useful in diagnosing and controlling diseases. Cancers are important to be diagnosed in the shortest time possible. The first symptom of breast cancer is the existence of mass, or thickening of the breast, or underarm area (Basch et al., 2017). Other symptoms include clear or bloody secretion from the breast, desquamation at the nipple, inverted nipple, redness and swelling of the breast or underarm, invertedbreast skin in which the skin looks like orange peel, the occurrence of a breast deformity that makes it different from the other breast, a wound in the breast skin that does not heal, a persistent pain in the breast or underarm (Kohler et al., 2017). Inflammatory breast cancer, which is aggressive, is diagnosed late due to the lack of a palpable mass and has critical consequences. Symptoms include fever, swelling, warmth, inverted nipple, itching, redness, and inverted breast skin in the form of orange peel (Ingeman et al., 2015).

Breast cancer is one of the most prevalent cancers in women that its timely diagnosis plays a significant role

in survival and treatment, but even the most common diagnostic techniques such as mammography cannotprovide highly accurate diagnoses. Therefore, better diagnostic techniques need to be considered. Mammography is a special type of CT scan that uses high-resolution X-ray and film methods to be able to diagnose breast tumors well. Low radiation is the strength of this method. Mammography is only used to diagnose breast tumors. Breast cancer mammography images have the ability to help physicians diagnose a disease caused by abnormal cell growth. Developing algorithms and software to analyze these images may also help physicians in their ordinary works. In expert systems, one of the most important applications of data mining techniques is related to disease diagnosis. Automatic diagnostic systems, with the help of medical data, can reduce the time of diagnosis and possible error of experts and make the details more accurate.Studyingthe several automatic diagnostic systems and the combination of support vector machine (SVM) and fuzzy C-Mean and K-Mean, etc. are very effective for diagnosing malignant cancer from benign. Cancer happens when cells begin to grow and multiply uncontrollably (Radh and Rajendiran, 2014).

In order to develop cancer, the gene that regulates cell growth and multiply must be transformed. These mutations will then turn into a mass through cell multiply. By identifying the gene that transmits this cancer, asignificant step can be taken in predicting breast cancer. One of the important applications of data mining technique is related to the field of medicine and diagnosis of diseases. Data mining is a technique for processing and analyzing plentiful data that from its results can obtain a series of useful information. The data mining technique uncovers hidden patterns in the enormous data, and the results can reveal hidden patterns in the eyes of physicians, which are important in prevention and treatment. One of the data mining techniques is information clustering (Andreeva et al., 2004). In this regard, the main objective of this project is early diagnosis of breast cancer patients with some data mining algorithms such as K-Mean and fuzzy C-Mean.

Recommended Method

The objective of this study is providing an automated method for distinguishing healthy tissue from unhealthy that helps the physician make better and more accurate decisions and minimizes unnecessary biopsies. Abundant studies have been conducted to diagnose breast cancer, but still methods improvement is essential. The strength point of this study is the use of a combined method in diagnosing the joint area to the mass with the simultaneous use of two standard images of a breast, which has led to better and more accurate results. The idea originates from the radiologists' definition that a mass is a pervasive lesion that can be seen from at least two different perspectives. During the study and implementation of the methods, both standard images of a breast have been used, and finally the selection of features is based on a combination of the results obtained from both images, which is clearly itssupremacy regarding the use of a single image has been demonstrated.

The development of algorithms and software for analyzing these images may also help physicians in their ordinary works. In expert systems, one of the important applications of data mining techniques is related to disease diagnosis. Automatic diagnostic systems, with the help of medical data, can reduce the time of diagnosis and possible error of experts and make the details more accurate in the diagnosis process. Studying several automatic diagnostic systems and the combination of support vector machine, fuzzy C-Mean, K-Mean, etc. is very effective for distinguishing malignant from benign cancer tumors. Cancer happens when cells begin to grow and multiply uncontrollably (Siu, 2016). One of the important applications of data mining technique is related to the field of medicine and diagnosis of diseases. Data mining is the process of discovering new patterns from a large data set. This technology can utilize large volumes of data through various data mining techniques such as classification, prediction, clustering, and outlier analysis. One of the data mining techniques is information clustering (Kalyankaret al., 2013). Popular algorithms in this field are K-Mean and fuzzy C-Mean algorithms. According to the above mentioned matters, the main objective of this project is early diagnosis of breast cancer patients with some data mining algorithms such as K-Mean and fuzzy C-Mean.

The steps of this case can be divided into three general categories including pre-processing, mass joint initial selection, and feature extraction. The input of this system is two standard images related to one breast and the final output is the classification of the image into two

categories of healthy and unhealthy. The input part of the system consists of two images, CC and MLO, which are processed on both images at the same time. The steps of pre-processing, extraction of thesuspected areato mass, and feature extraction each involve more detailed steps (Thawkar et al., 2017).

1. Data Set

Nowadays, to standardize research, organizations collect labeled data set by researchers and make it available to researchers online. There are various datasets that are ranked based on the validity, number and accuracy of their data. In this research, the mini MIAS data set was used, which is one of the strongest data between digital mammography datasets and has 330 samples, of which 209 are healthy samples, 54 are malignant samples and 67 are benign samples. Mammographic image have a size of 1024 * 1024 and edge of 200 microns pixels in pgm format and in gray color, in the range of [0,255]. The data set keeps the collected information in a separate file called GT5 data for the researcher to have enough information access, which is made available to the public online along with the data set. GT5 data contains information such as healthy or unhealthy, exact location of the mass, benign or malignant, and breast density (Digital Database for Screening Mammography. Available at: http:// marathon.csee.usf.edu /Mammography/ Database.htm).

2. Input Image

The input image is two standard mammographic images, which both images are analyzed simultaneously and side by side,because one of the factors that radiologists look for in mammographic images is the observation of a mass lesion from both perspectives. Figures (1) A and B show an example of two main images with the mentionedperspectives. Figures (1) C and D show the demarcation of the mass in the images used, which are also present next to the main images of the dataset and the demarcations have been done by radiologists.

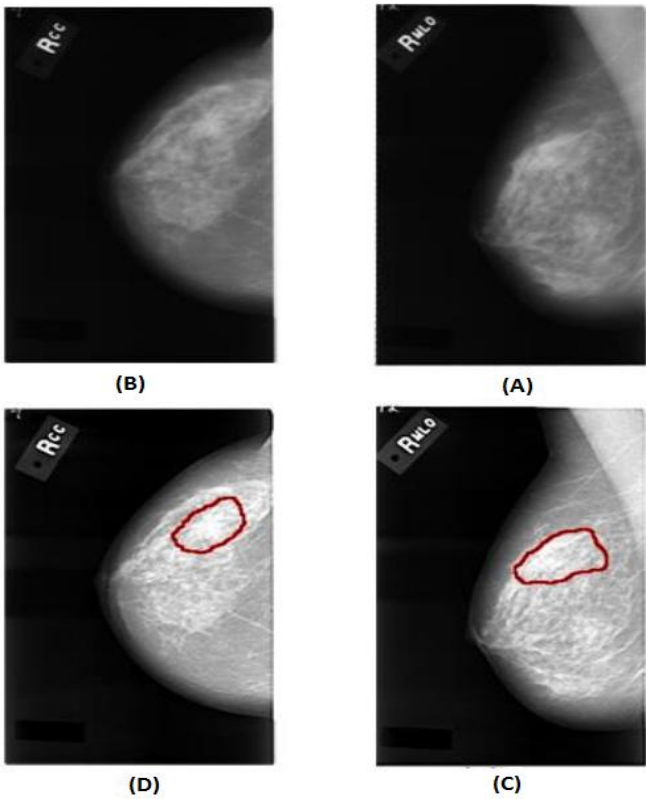
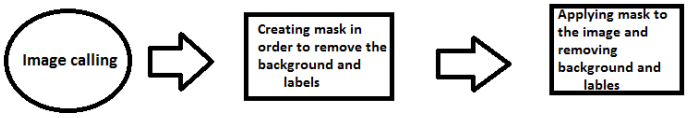


Figure (1): Images used from two different perspectives,(A) and (B) are the main images,(C) and (D) are thedemarcated images of data

3. Pre-processing

Due to the large size of mammogram images and the high volume of calculations required to find the location of unhealthy tissues, performing diagnostic for the entire mammogram surface is time consuming. Therefore, reducing the areas under study, if done without losing valuable information, will help greatlyto the accuracy, speed and volume of calculations. Areas other than the breast in mammographic images are not completely black and do not have intensity level of 0, which makes it difficult to select the area suspected of a mass because in the clustering step, the background is also divided into several separate clusters. In order to remove the background area, a mask is created and after applying the mask to the main image, the background will have an intensity level of 0 and in the clustering step, the whole background will be separated as a cluster. These images also have labels that are removed at this step (Maitra et al., 2012). The flowchart in Figure (2) shows the steps in order.



Pre-processing steps

4. The Co-occurrence Matrix

Haralik et al., to study the structure of different tissuesrecommendedfeatures based on the adjacency matrix, which is one of the most successful methods for studying the properties of different tissues. In this method, first co-occurrence matrixes are calculated for different distances and directions and then a number of features are calculated for each matrix. The following recommended features have been obtained for the C_{ij} matrix (Bhateja et al., 2018).

(A). Energy

1. $$Energy = \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}^2$$

(B). Entropy

2. $$Entropy = \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}^2 \log(C_{ij}^2)$$

(C). Contrast

3. $$Contrast = -\sum_{i=1}^{n}\sum_{j=1}^{m} |i-j| C_{ij}$$

(D). Reverse difference momentum

4. $$\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{C_{ij}}{|i-j|^k} i \neq j\,, k = 1,2$$

(E). Correlation

5. $$Correlation = \frac{\sum_{ij}(i-\mu)(j-\mu)c_{ij}}{\sqrt{var\,(i)var\,(j)}}$$

Findings

Progresses in technology and the construction of various imaging devices in recent decades have made medical images a very important part of the disease diagnosis process in medical centers. Physicians' need for these images to accurately diagnose the disease and the increasing number of these images has also increased the need for automated algorithms to increase accuracy and facilitate the diagnosis process. Meanwhile, the automatic clustering algorithm of medical images has an important role in diagnosing suspicious areas in the images and increasing the speed and accuracy of diagnosis. In regard with this research, breast cancer diagnosis model using K-Mean clustering and C-Mean fuzzy has been presented (Padmavathy et al., 2021).

1. Image Pre-processing

First, the image is called, (Figure 3) shows two standard mammographic images that have been selected as input.

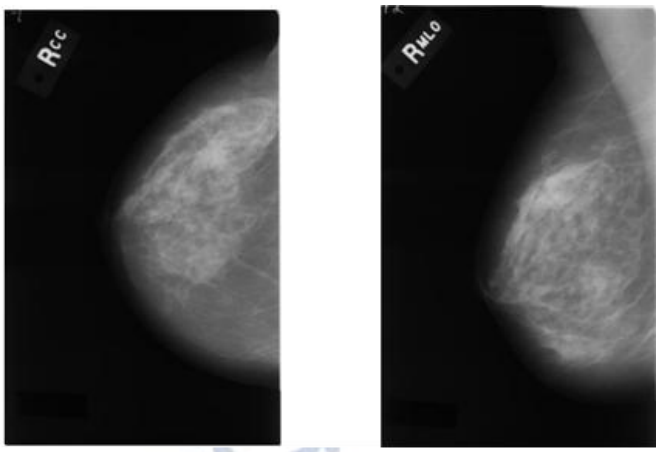


Figure (3): Figure of two standard mammographic images

As mentioned, histogram improvement was used for pre-processing operations. In this case, the brightness of the image pixels was uniformly moderated throughout the image.

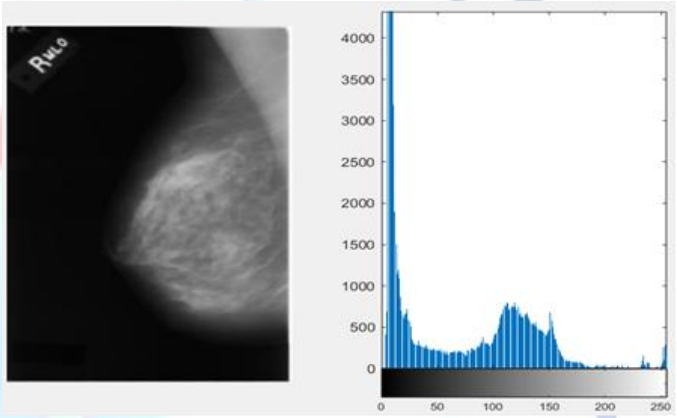


Figure (4): Improving image contrast and image histogram chart

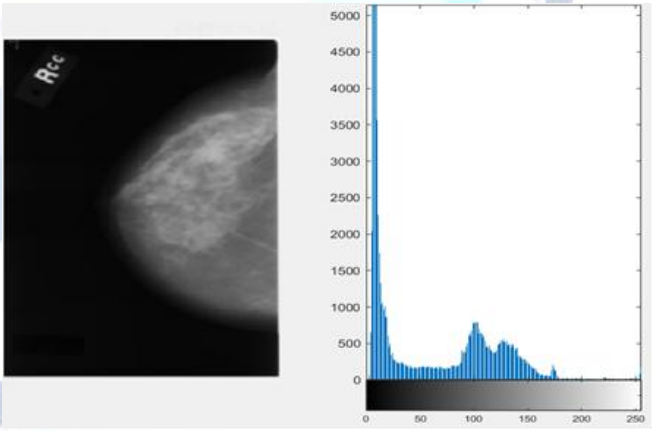


Figure (5): Improving image contrast and image histogram chart

In addition, the Gaussian function was used in the noise removal of pre-processing operation. In this case, to improve the image resolution, a mask with dimension 9 is drawn on the whole image to pixels stretched.

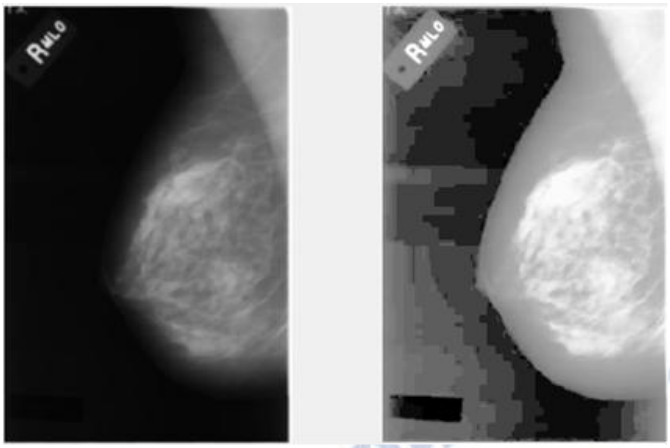


Figure (6): Noise removal with Gaussian function

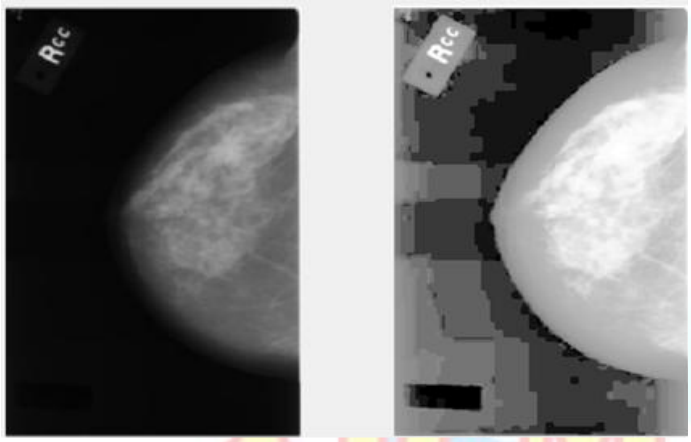


Figure (7): Noise removal with Gaussian function

In the next step, the feature was extracted after the input images. In this step, in order to extract the feature, co-occurrence matrix methods and wavelet method were used. In the wavelet feature extraction method, PCA method was used to reduce the feature dimension.

```
stats =

  struct with fields:

      Contrast: [0.2878 0.1763]
   Correlation: [0.9531 0.9719]
        Energy: [0.3686 0.3685]
   Homogeneity: [0.9663 0.9628]
```

Figure (8): Results of feature extraction with glcm

```
stats =

  struct with fields:

      Contrast: [0.4145 0.1623]
   Correlation: [0.9174 0.9690]
        Energy: [0.4439 0.4465]
   Homogeneity: [0.9666 0.9689]
```

Figure (9): Results of feature extraction with glcm

Figure (10): Binary images to create a mask

As shown in Figure (10), the output of the binary image has labels or particles scattered in places other than the breast tissue, at which point filtration can remove excess particles and labels. By doing these steps, the desired mask is created. Figure (10) demonstrates the removal of labels from the binary image and the creation of the mask.

The mask created in the previous step is applied to the main image and the black dots in the binary image also turn black in the main image (intensity level 0) and the information of theremaining points does not change. The result of applying the mask to the image is demonstrated in Figure (10).

At this step and in order to extract the suspected area to the mass, a combined method has been used, which has ultimately led to the desired result. The pre-processed image is first received as input to this step. Then, K-Mean clustering and C-Mean fuzzy methods are used to cluster the image into 5 clusters. The number of clusters has been selected experimentally after applying to 600 images. Among the mentioned clustering methods, the C-Mean fuzzy method is used as the most suitable method selected. After clustering, the cluster that its cluster center has the highest intensity level is used. In the selected cluster, after removing the pectoral muscle and removing the scattered and fine particles, the brightest pixel is used as the granular point in the area growth method, and after applying the algorithm, the joint area to the mass is extracted. If the joint area to the mass is present in both images, the desired tissue is considered as healthy tissue; otherwise we enter the feature extraction step.

2. Clustering

Clustering methods have been used to separate the suspected area to the mass. Hierarchical clustering methods are used for small data and cannot be used for

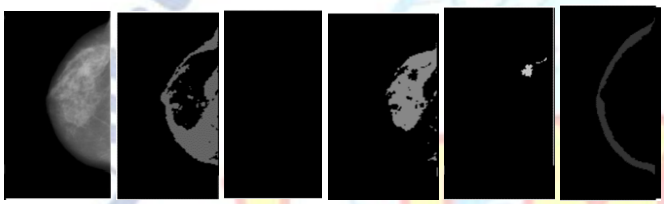
mammographic images that contain large amounts of information.

2.1 Clustering Using the K-Mean Method

In this method, the number of clusters is defined first. After clustering, the image is divided into 5 clusters. As demonstrated in Figures (11) and (12), the background area is clustered as a separate cluster.

(A) (B) (C) (D) (E) (F)

Figure (11): Applying K-Mean clustering algorithm: A) RMLO pre-processed image, B to F) different created clusters



(A) (B) (C) (D) (E) (F)

Figure (12): Applying K-Mean clustering algorithm: A) RCC pre-processed image, B to F) different created clusters

A problem that occurs frequently in this method is the depletion of a cluster during the implementation of the algorithm, which causes the algorithm to remain unfinished and in fact not implemented. And this is exactly one of the disadvantages of this method (if in the iteration of the algorithm the number of data belonging to clusters becomes zero, there is no way to change and improve the continuation of the method). Due to the problem, it was decided to use the C-Mean fuzzy method, the results of which are as follows:

2.2 C-Mean Fuzzy Clustering Method

In this method, the number of clusters is initially determined, which, like the previous method, is considered 5.

(A) (B) (C) (D) (E) (F)

Figure (13): Application of C-Mean fuzzy clustering algorithm: A)RMLO pre-processed image, B to F) different clusters created

(A) (B) (C) (D) (E) (F)

Figure (14): Application of C-Mean fuzzy clustering algorithm: A) RCC pre-processed image, B to F) different clusters created

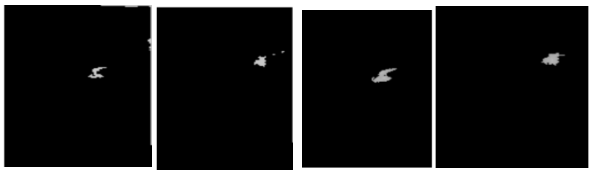
The results obtained by the C-Mean fuzzy method are relatively desirable and the problem created in the K-Mean method has been solved in this method.

2.2.1 Granular Point Selection of Area Growth Method

In the previous step, 5 clusters were created using the C-Mean fuzzy clustering method. In order to select one of the clusters, a cluster with a brighter cluster center (higher intensity level) is selected. In the selected cluster, the brightest pixel is selected as the initial granular point for this algorithm after removing the pectoral muscle and excess edges. At this step, before selecting the brightest pixel, the pectoral muscle can be removed. Because according to experts and as mentioned in scientific and medical references, the onset of cancer is generally from the lobules and mammary ducts, and the possibility of cancer starting from the pectoral is very low.

2.2.2 Area Growth Algorithm

After removing the excess areas and pectoral muscle, the brightest pixel is selected and the area growth algorithm is applied, the results of which are as follows (Figure 15). (A) and (B) are the selected clusters after clustering. After removing the excess edges and the pectoral muscle, the brightest pixel is selected as the granular point of the area. The area growth algorithm application results to Figures (A) and (B) is mentioned in Figures (C) and (D). In order to compare the accuracy of the recommended method in separating the jointarea to the mass, the resulting suspicious area is subtracted from the demarcated image in the dataset. An example of the accuracy of the method is demonstrated in Figure (15).



(A) (B) (C) (D)

Figure (15): Application of area growth algorithm: A) and B) selected clusters in C-Meanfuzzy step, C) and D) application of area growth algorithm and determination of the area suspected to the mass

2.2.3 Feature Extraction

At this step, after observing the suspicious area to the mass in both images, the features are extracted. In this way, the features are calculated separately for each image, and finally, in order to create the feature, a combination of the features ratio of the two images is used, which the accuracy of combined features in separating healthy tissue from unhealthy, compare to the features of single images is very high.

In order to extract the features related to demarcate and the shape of the suspicious area to the mass, respectively, tissue descriptors based on the co-occurrence matrix with c feature have been used. The new feature used in this dissertation is the maximum intensity level of the cluster center in the clustering step that after applying methods on 322 images, which in each image have two different perspectives, the results have been studied.

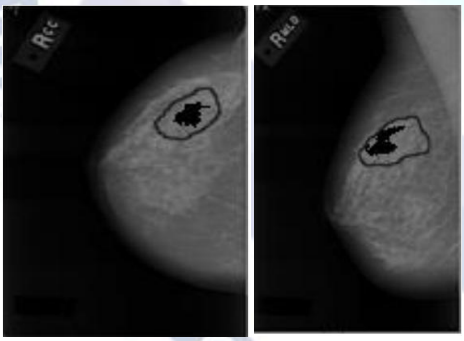
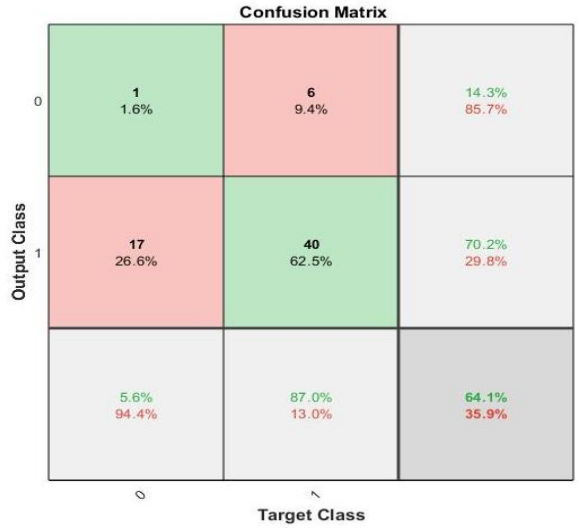


Figure (16): Compares the demarcation conducted with the demarcation in the data. The curves in the image show the data demarcation and the hole inside the curve shows the recommended demarcation of this research.

2.3 Model Evaluation

In order to confirm if machines can be trusted as tools to diagnose disease from images; 322 data (mammographic image) were collected from mini MIAS data center and the accuracy of disease diagnosis by machine was evaluated using MATLAB software and cross-validation algorithm. In this method, 80% of the data is selected as educational data and the remaining 20% as test. Usually this ratio can be changed to 70 to 30. The input data of MATLAB software was in the form of an Excel file and finally the accuracy of the software was calculated. The correct diagnosis accuracy in this machine as shown in the figure below is 64.1%.

Table (1): Confusion matrix



Based on this information, it seems that the absolute reliance on machines in diagnosing disease is a high-risk work and only machines should be used with the help of clustering algorithms and increasing image resolution in timely and more accurate diagnosis of diseases.

Conclusion

The volume of mammography images is very high. Also in these images, a black area can be seen that these areas are not completely black, i.e. the intensity level of all these pixels is not zero, and this makes the separation of the suspected area to the mass difficult. In order to resolve this issue, an attempt was made to remove this area. For this purpose, a suitable mask was created and applied to the main image. The presence of labels also makes it difficult to continue, so these labels were also removed. In order to separate the suspicious area to the mass, a combined method was used. In this way, clustering algorithms were used first. The K-Mean algorithm was used as the first algorithm and the images were divided into 5 clusters. The problem with this method was that the algorithm remained unfinished in the face of depletion cluster, which left the clustering work incomplete, and in fact the

algorithm was not convergent, and this heterogeneity occurs frequently.

The C-Mean fuzzy algorithm solved the problem of the previous method. Considering the advantages of fuzzy C-Mean and considering the similarity of adjacency, this method was also implemented, but in terms of accuracy, it was no different from the K-Mean method. Therefore, the C-Mean fuzzy method was selected as the best method. After clustering, a cluster with a brighter cluster center (higher intensity level) was selected. In the mentioned cluster, the pectoral muscle was removed because, according to experts, the cancer started from the mammary ducts and lobules.

After removing the pectoral muscle, the brightest pixel was selected as the granular point in the area growth method, and after applying the area growth method, the area suspected to the mass was finally selected.

At this step, it is time to extract the feature. Tissue features based on co-occurrence matrix were used. The difference point in this study is the simultaneous use of two images related to one breast. According to experts, the mass is a pervasive lesion that can be seen from at least two different perspectives, and according to this issue, both existing images were used. It should be noted that according to the above mentioned matter, in the separation step of the joint area, in some healthy images, the joint area was extracted from only one image and in another image no area was extracted and the healthiness of relevant image can be specified at this step and without the need to extract the feature. Among the extracted features, feature C and the brightest center of the cluster perform the separation most accurately. To calculate feature C, it is necessary to calculate the perimeter and area of the joint area. For this purpose, the image obtained from the growth of the area was binary and using the morphological operators, the holes in the image were filled. Sobel edge detector was used to calculate the perimeter. Finally, the accuracy of the proposed features was examined and compared.

Finally, it is noteworthy that due to the accuracy of diagnosis, which was obtained from the cross-validation method on the same 322 images (64.1%),the best way for radiologists to diagnose cancer more quickly and accurately is to use clustering to reduce human error. Due to the fact that in using different algorithms in using machines, the amount of accuracy and sensitivity is usually low, in real data, if we use machines and programs, the amount of error can be high. Therefore, the best use of different data mining algorithms to reduce human error and early diagnosis can be clustering, which according to the clustering results, C-Mean fuzzy clustering helps a lot in this issue. In order to continue this research, the following methods are suggested:

1. Use optimization methods to select the best number of clusters.
2. The pectoral muscle is removed automatically.
3. Use all four images related to a person and existing relationships studied.
4. Check the fractal features.
5. Unhealthy images should be classified into benign and malignant.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Basch, E., Deal, A. M., Dueck, A. C., Scher, H. I., Kris, M. G., Hudis, C., &Schrag, D. (2017). Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. Jama, 318(2), 197-198.

[2] Kohler, R. E., Gopal, S., Miller, A. R., Lee, C. N., Reeve, B. B., Weiner, B. J., & Wheeler, S. B. (2017). A framework for improving early detection of breast cancer in sub-Saharan Africa: A qualitative study of help-seeking behaviors among Malawian women. Patient education and counseling, 100(1), 167-173.

[3] Ingeman, M. L., Christensen, M. B., Bro, F., Knudsen, S. T., &Vedsted, P. (2015). The Danish cancer pathway for patients with serious non-specific symptoms and signs of cancer–a cross-sectional study of patient characteristics and cancer probability. BMC cancer, 15(1), 1-11.

[4] Radha, R., &Rajendiran, P. (2014, February). Using K-means clustering technique to study of breast cancer. In 2014 World Congress on Computing and Communication Technologies (pp. 211-214). IEEE.

[5] Andreeva, P., Dimitrova, M., & Radeva, P. (2004, April). Data mining learning models and algorithms for medical applications. In Proceedings of the 18-th Conference on Systems for Automation of Engineering and Research SAER (pp. 11-18).

[6] Siu, A. L., & US Preventive Services Task Force. (2016). Screening for breast cancer: US Preventive Services Task Force recommendation statement. Annals of internal medicine, 164(4), 279-296.

[7] Kalyankar, M. A., & Alaspurkar, S. J. (2013). Data mining technique to analyse the metrological data. International Journal of Advanced Research in Computer Science and Software Engineering, 3(2).

[8] Thawkar, S., R.J.I.J.o.I.E, (2017). Ingolikar, and Systems, Automatic detection and classification of masses in digital mammograms. 2017. 10(1): p. 65-74.

[9] Digital Database for Screening Mammography. Available at: http:// marathon.csee.usf.edu/Mammography/Database.htm

[10] Maitra, I. K., Nag, S., & Bandyopadhyay, S. K. (2012). Technique for preprocessing of digital mammogram. Computer methods and programs in biomedicine, 107(2), 175-188.

[11] Bhateja, V., Gautam, A., Tiwari, A., Bao, L. N., Satapathy, S. C., Nhu, N. G., & Le, D. N. (2018). Haralick features-based classification of mammograms using SVM. In Information Systems Design and Intelligent Applications (pp. 787-795). Springer, Singapore.

[12] Padmavathy, T. V., Vimalkumar, M. N., Nagarajan, S., Babu, G. C., & Parthasarathy, P. (2021). Performance analysis of pre-cancerous mammographic image enhancement feature using non-subsampled shearlet transform. Multimedia Tools and Applications, 80(18), 26997-27012.