



University of Kashan

Contents lists available at SCJ

Soft Computing Journal

Journal homepage: <https://scj.kashanu.ac.ir/>

# Breast cancer diagnosis and classification improvement based on deep learning and image processing <sup>◇</sup>

Mohsen Eftekharian<sup>1</sup>, PhD student, Ali Nodehi<sup>1,\*</sup>, Assistant Professor<sup>1</sup>Department of Computer Engineering, Gorgan Branch, Islamic Azad University, Gorgan, Iran.

## ARTICLE INFO.

### Article history:

Received May 24, 2022

Accepted December 11, 2022

### Keywords:

Breast cancer

Diagnosis and classification

Quantum inverse MFT algorithm

Social spider algorithm

Convolutional neural network

## ABSTRACT

Nowadays, medical intelligence detection systems have evolved significantly due to advancements in artificial intelligence, however, they face some challenges. Breast cancer diagnosis and classification is one of the medical intelligence systems. There are a variety of screening techniques available to detect breast cancer such as mammography, magnetic resonance imaging, and ultrasound. This research uses the MIAS mammography image dataset and tries to diagnose and classify benign and malignant masses based on image processing and machine learning techniques. Initially, we apply pre-processing for noise reduction and image enhancement using Quantum Inverse MFT, and then image segmentation with the Social Spider Algorithm. The type of mass is then diagnosed by the Convolutional neural network. The results show that the proposed approach has better performance in comparison to others based on some evaluation criteria such as accuracy of 99.57%, sensitivity of 91%, and specificity of 86%.

2322-3707 / © 2022 The Authors. Open access article under the CC BY license.



## 1 Introduction

Breast cancer is a type of cancer that begins in the breast tissue of women with symptoms such as a mass in the breast, breast deformity, skin rash, discharge from the nipple, or partial scaling of the skin. To grow cancer, the gene must regulate growth and cell proliferation. These mutations then become a mass through cell proliferation. Identifying the transporter gene of this cancer can be an important step in predicting breast cancer. A high volume of genetic information is one of the most important problems in representing the large structure and function of biological molecules. Furthermore, one of the most important challenges in bioinformatics is the need to

design and produce methods, algorithms, and tools to convert this large volume of often heterogeneous (low-level) data to higher-level bio-knowledge. The masses are divided into benign and malignant. Visually, benign masses have very smooth and uniform margins, whereas malignant masses have dark and prominent margins [1].

The use of image processing principles and techniques, along with statistical and cognitive pattern recognition in the diagnosis and automated detection of mammograms has reduced human error and increased detection speed. In [2], deep learning techniques are used to diagnose and classify breast tumors. To this end, three different deep learning architectures including GoogLeNet, VGGNet and ResNet are considered and then an analysis is performed between them. The results show that the proposed approach had high accuracy in the diagnosis and classification of tumor areas. In [3], the researchers used the MIAS

<sup>◇</sup> Article type: original research (short paper)

\* Corresponding author.

Email addresses: [m.eftekharian@gorganiau.ac.ir](mailto:m.eftekharian@gorganiau.ac.ir)

(M. Eftekharian), [ali.nodehi@gorganiau.ac.ir](mailto:ali.nodehi@gorganiau.ac.ir)

(A. Nodehi)

and INbreast datasets and reduced the noise in the image dataset by using median and Gaussian filters. In this paper, out of 7259 images, 6346 of which are used for training and 913 for testing. They used VGGNet, MobileNet, GoogLeNet, ResNet, and DenseNet to propose a deep ConvNet + SVM hybrid network. The accuracy of this method is 97.8%. In [4], random forest, SVM, decision tree, K-nearest neighbor, and logistic regression methods are investigated with the Wisconsin breast cancer dataset to determine the best machine learning algorithm. From the results of the confusion matrix, accuracy, and precision, it is concluded that the support vector machine performs better than other classifiers with an accuracy of 97.2%. In [5], the authors proposed an approach to reduce the noise of mammography images in the pre-processing phase. They then classify the input data using machine learning techniques such as support vector machine, logistic regression, and K-nearest neighbor. In this approach, 60% of the data is used for training and 40% for testing, and the accuracy is 97.7%. In [6], a CAD system for the automatic detection of tumor types is proposed by using different machine learning algorithms. The results are compared by examining the machine learning algorithms and different group models on two datasets. These results show that the group method outperformed other methods and achieved 98.80% accuracy. In [7], a graph-based semi-supervised learning method is used to propose a new model for analyzing the survival of breast cancer patients. In [8], the Z-Curve mapping method is applied for conversioning DNA alphabetic strings to digital signals. This method uses a Linear Predictive Coding model to analyze resultant data for feature extraction. In addition, this method is beneficiary of a certain singular value decomposition computational approaches to select significant features for reduction of dimension.

In this paper, we propose an approach to diagnose and classify benign and malignant masses based on image processing and machine learning techniques. Hence, the proposed approach has three main parts, including pre-processing, image segmentation, and mass-type diagnosing. In the following section, we describe these parts in detail.

## 2 Proposed method

The proposed approach has three phases: pre-processing, image segmentation, and mass-type diagnosing. In the pre-processing phase, we improve and reduce the noise of mammographic images using a method called Quantum Inverse MFT Filtering. These images are noisy due to the way they are created. They have salt and pepper noises that may be confused with microcalcifications found in mammography images. For this reason, they must be detected and destroyed. Then the image segmentation phase

is implemented using the Social Spider Algorithm (SSA). Finally, the type of mass is diagnosed by the Convolutional Neural Network (CNN).

### 2.1 Pre-processing phase

Every single image in the combination of local threshold and Active Contour is represented by a two-dimensional array of pixels whose values are integers in the range of [0,255]. Local thresholding initializes the images in two steps. Initially, the input noise image is considered as the initial image, which is used to eliminate the image noise. Thus, at the end of the first step, there is a decomposed image. In the second step, the thresholding is performed on the detail coefficients and one of these decomposed parts is randomly selected and sent to a reconstruction operation by Gauss Fading.

In the first step, we define a threshold value for the noise reduction method and then a Quantum Inverse MFT Filtering is applied by three parts in Eqs. (1) to (3). In this noise reduction step, some kinds of noises are determined. Active Contour is applied to specify the variation of these noises to help Quantum Inverse MFT Filtering for more noise reduction. Quantum Inverse MFT Filtering-based local thresholding and Active Contour can produce a much smoother display. A local threshold function and Active Contour with Quantum Inverse MFT Filtering have two main features. First, the function is oscillatory or has a wave appearance such as Eq. (1).

$$\int_{-\infty}^0 |\Psi(t)|^2 dt < \infty \quad (1)$$

Second, the local thresholding is a value which has [0 1] or [0 255] colors. In this case, most of the energy at  $\Psi(t)$  is limited to a finite period of time whose relation is in Eq. (2).

$$\int_{-\infty}^0 \Psi(t) dt = 0 \quad (2)$$

In general, the proposed method reduces the noise using Eq. (3).

$$Method(I) = \left( \sum_{\Omega} \sqrt{1 + \beta^2 |\nabla I|^2} \right) + \frac{\lambda}{2} (I - I_0)^2 \quad (3)$$

The function represented in Eq. (3) is aware of the edges of the image and tries to preserve the important features of the image. In this equation,  $(I - I_0)^2$  guarantees a certain degree of validity between the evaluated image and the original image, in which the evaluated image  $I$  and the  $I_0$  image are noisy. Parameter  $\nabla I$  denotes the period for adjusting the sum of the variations,  $\beta$  and  $\lambda$  are the balancing parameters, and  $\Omega$  is the points in the image. Note that  $\nabla I$ , as a parameter to adjust the sum of the variations, means that one mammography image may

have some noises such as Gaussian, Salt and Pepper or also blur effects. Hence this variation is used to determine the kinds of these noise variations and calculate its sum. Minimizing Eq. (3) reduces the overall image variability while maintaining accuracy [9].

## 2.2 Image segmentation phase

The Social Spider Algorithm is used for image segmentation. This algorithm formulates the search space of the optimization problem as a very high spider web. Each position on the web presents a practical solution to the optimization problem and all practical answers to the problem have the answer on the web. The thread is used as the vibration transfer medium produced by the spiders. Each spider on the web demonstrates the ability to find the position and fit the solution based on the objective function as well as the ability to find food sources in the position. Spiders can move freely on the web, but cannot leave the web until they have found a practical answer to the optimization problem. When a spider moves to a new position, it produces a vibration that propagates throughout the web. Each vibration holds the information of one spider, and the other spiders keep the information until they receive the vibration.

The image is referred to as the search space or the web in the Social Spider Algorithm. It is necessary to create an initial population of spiders on the search space or image that is web. The spiders move on the web, aiming to accurately identify boundary areas, which have different brightness and light intensity than other parts, based on color means black and white histograms for gray-scaled images or in RGB modes. In fact, spiders walk on the edges, moving to the right for precise alignment by shaking the lines or recognizing edges and light intensity to segment images. The fit function is the segmentation of the whole image area which is the termination condition. Edges are in mammogram images. In segmentation after defining the initial parameters of SSA, spiders should move to find edges due to their brightness value which is determined in the pre-processing phase by local thresholding in [0 1] or [0 255] colors. Fig. 1 shows segmentation with the SSA algorithm.

## 2.3 Classification

Now the output of the noise reduction and mammography imaging phase enters as input to the convolutional neural network. Then, the dot product between the input and the parameters of each neuron is performed and finally, the convolution operation is performed in each layer. After calculating the network output, these results are used to adjust the network training parameters and calculate the error rate. In

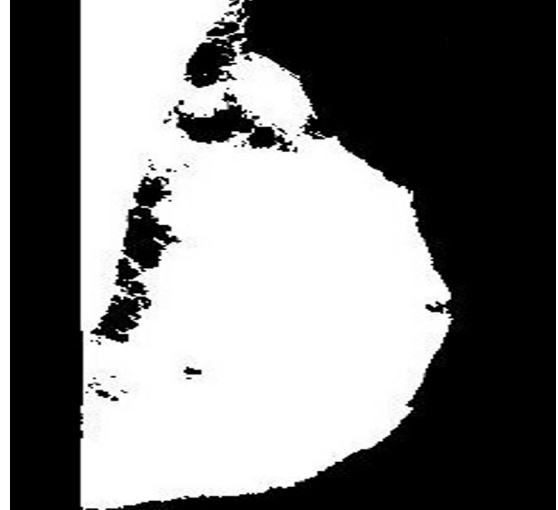


Fig. 1. Segmentation with the SSA algorithm.

the next step, based on the calculated error, the back-propagation phase begins. Now the gradient of each parameter is calculated according to the rule chain and all the related parameters of the neural network change according to their effect. After updating the parameters, the feed-forward step begins. The training ends after a certain number of repetitions. Fig. 2 and Table 1 show the structure of our proposed convolutional neural network which has 16 layers.

## 3 Results

We use MATLAB to simulate our algorithm. The hardware specifications of our device include a 7-core processor with a 3.4 GHz Intel processor, 6 MB of cache memory, and 6 GB of RAM. In this paper, the MIAS dataset is illustrated using statistical properties. In the dataset, there are images with both breast cancer and non-breast cancer features. This dataset is available at <http://peipa.essex.ac.uk/info/mias.html> link. We use evaluation criteria AUC, Mean Square Error (MSE), Accuracy, Sensitivity, and Specificity in the classification part. Table 2 demonstrates the obtained evaluation criteria after the simulation and implementation of the proposed algorithm.

According to the results, our method has good performance in the classification part. Note that the accuracy of 99.57% is obtained in the whole fusion method. Finally, we compare the proposed approach to the state-of-the-art methods in terms of evaluation criteria. To provide the same condition in our comparison, we consider the prior papers that use the MIAS dataset in their approaches and calculate the accuracy as the main criterion for the diagnosis and classification of breast cancer. Table 3 compares the results of the proposed approach with other methods. As the results show, our approach outperforms the others in terms of accuracy

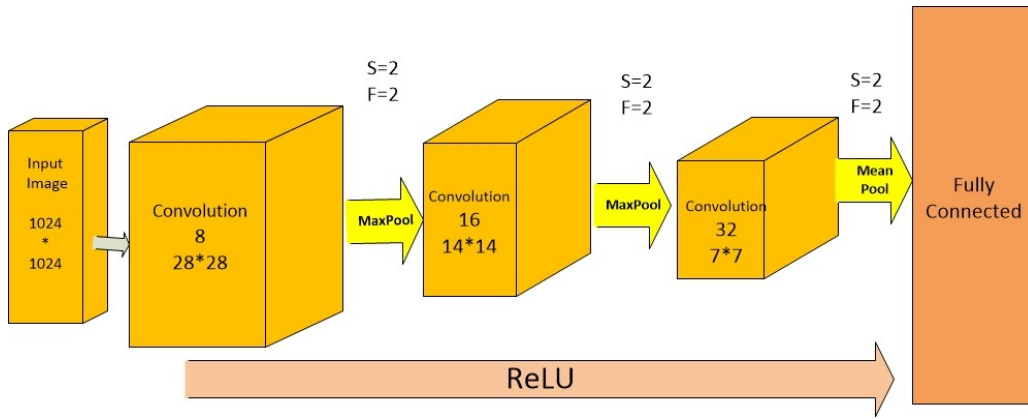


Fig. 2. The proposed CNN structure.

Table 1. The proposed CNN features.

Layer Number	Layer Name	Description
1	Image Input	1204*1204*3
2	Convolution	8 28*28 Convolutions with stride [2 2] and Padding 'same'
3	Batch Normalization	Batch normalization
4	ReLU	ReLU
5	Max Pooling	2*2 Max pooling with stride [2 2] and padding [0 0 0 0]
6	Convolution	16 14*14 Convolution with stride [2 2] and padding 'same'
7	Batch Normalization	Batch normalization
8	ReLU	ReLU
9	Max Pooling	2*2 Max pooling with stride [2 2] and padding [0 0 0 0]
10	Convolution	32 7*7 Convolution with stride [1 1] and padding 'same' and padding 'same'
11	Batch Normalization	Batch normalization
12	ReLU	ReLU
13	Max Pooling	2*2 Mean pooling with stride [2 2] and padding [0 0 0 0]
14	Fully Connected	100 Fully Connected layers
15	Soft Max	Soft Max
16	Classification Out Put	Cross Entropy

Table 2. The obtained evaluation criteria.

Metric	Value
AUC	0.87
MSE	0.01
Accuracy	99.57%
Sensitivity	91.00 %
Specificity	86.00%

#### 4 Conclusion and future work

In this paper, we propose a three-level processing approach. This approach applies Quantum Inverse MFT to enhancement and noise reduction in mammography images. The image segmentation operation is then defined by the operators of the Social Spider Algorithm and CNN for classification. The obtained results show that the proposed method has 99.57%

Table 3. Comparison of proposed approach with the recent methods in terms of accuracy.

Method	Accuracy
M. Togacar, et al., 2020 [10]	98.58%
T. Mahmood, et al., 2021 [3]	97.80%
M. A. Naji, et al., 2021 [4]	97.20%
R. Resmini, et al., 2021 [11]	97.18%
U. Naseem, et al., 2022 [5]	97.70%
S. Safdar, et al., 2022 [6]	98.80%
Proposed approach	99.57 %

accuracy which gains better performance in comparison to the recent methods. In future, we plan to:

- use other standard data sets (for example DDSM),
- use other deep learning structures,

- use the combination of the Social Spider Algorithm with another classifier such as SVM in the segmentation phase to increase the accuracy.

### Conflict of interest

The authors declare that they have no conflict of interest.

### References

- [1] N. Kavva, N. Sriraam, N. Usha, B. Hiremath, A. Suresh, D. Sharath, B. Venkatraman, and M. Menaka, "Breast cancer lesion detection from cranial-caudal view of mammogram images using statistical and texture features extraction," *Int. J. Biomed. Clin. Eng. (IJBCE)*, vol. 9, no. 1, p. 17, doi: 10.4018/IJBCE.2020010102.
- [2] M. Kikuchi, T. Hayashida, R. Watanuki, A. Nakashoji, Y. Kawai, A. Nagayama, T. Seki, M. Takahashi, and Y. Kitagawa, "Abstract p1-02-09: Diagnostic system of breast ultrasound images using convolutional neural network," vol. 80, 02 2020, pp. P1-02, doi: 10.1158/1538-7445.SABCS19-P1-02-09.
- [3] T. Mahmood, J. Li, Y. Pei, and F. Akhtar, "An automated in-depth feature learning algorithm for breast abnormality prognosis and robust characterization from mammography images using deep transfer learning," *Biology*, vol. 10, no. 9, 2021, doi: 10.3390/biology10090859.
- [4] M. A. Naji, S. E. Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis," *Procedia Comput. Sci.*, vol. 191, pp. 487-492, 2021, doi: 10.1016/j.procs.2021.07.062.
- [5] U. Naseem, J. Rashid, L. Ali, J. Kim, Q. Emadul-Haq, M. J. Awan, and M. Imran, "An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers," *IEEE Access*, vol. 10, pp. 78 242-78 252, 2022, doi: 10.1109/ACCESS.2022.3174599.
- [6] S. Safdar, M. Rizwan, T. R. Gadekallu, A. R. Javed, M. K. I. Rahmani, K. Jawad, and S. Bhatia, "Bio-imaging-based machine learning algorithm for breast cancer detection," *Diagnostics*, vol. 12, no. 5, p. 1134, 2022, doi: 10.3390/diagnostics12051134.
- [7] R. Taimourei-Yansary, M. Mirzarezaee, M. Sadeghi, and B. Nadjar Araabi, "Predicting invasive disease-free survival time in breast cancer patients using semi-supervised graph-based machine learning techniques," *Soft Comput. J.*, vol. 10, no. 1, pp. 48-69, 2021, doi: 10.22052/scj.2022.243330.1039 [In Persian].
- [8] A. Khodaei and B. Mozaffary Tazeh-Kand, "An innovative approach in order for discrimination of cancer and non-cancer dna sequences by lpc and svd based algorithms," *Soft Comput. J.*, vol. 3, no. 2, pp. 42-53, 2015, doi: 10.1001.1.23223707.1393.3.2.57.0 [In Persian].
- [9] M. Eftekharian, A. Nodehi, and R. Enayatifar, "Improved noise reduction, segmentation, and classification of cancer masses by quantum inverse-matched filter, social spider algorithm, and improved elm," *Soft Comput. J.*, vol. 10, no. 1, pp. 70-89, 2021, doi: 10.22052/scj.2022.243230.1018 [In Persian].
- [10] M. Togacar, B. Ergen, and Z. Comert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," *Med. Hypotheses*, vol. 135, p. 109503, 2020, doi: 10.1016/j.mehy.2019.109503.
- [11] R. Resmini, L. F. da Silva, A. S. Araújo, P. R. T. Medeiros, D. C. Muchaluat-Saade, and A. Conci, "Combining genetic algorithms and SVM for breast cancer diagnosis using infrared thermography," *Sensors*, vol. 21, no. 14, p. 4802, 2021, doi: 10.3390/S21144802.