**Classification of Breast Lesions on Mammogram Images using Monarch Butterfly Optimization and Support Vector Machine**

***Erkan AKKUR1,\*[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-xxxx-xxxx-xxxx), Fuat TÜRK2[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-xxxx-xxxx-xxxx), Osman EROĞUL3[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-xxxx-xxxx-xxxx)***

*1Faculty of Engineering, Department of Biomedical Engineering, TOBB University of Economics and Technology, Ankara, Turkey*

*2Faculty of Engineering, Computer Engineering, Çankırı Karatekin University, Çankırı, Turkey*

|  |
| --- |
| **Abstract**  Currently, breast cancer affects many women worldwide. In recent years, many Computer-aided diagnosis (CAD) model have been developed for early diagnosis of breast cancer. An efficient CAD model is suggested to identify mammogram images as benign versus malignant in this study. The suggested CAD model constitutes four stages which are image acgusition, segmentation, feature extraction, feature selection and classification process. Gray level run matrix (GLRM) approach is used for feature extraction, while monarch butterfly optimization (MBO) for feature selection process. Support vector machine (SVM) algorithm is preferred for classification process. The suggested model has been tested on a private mammographic dataset. The suggested model (GLRM+MBO+SVM) shows an 0.944 of accuracy for breast lesion classification. Compared with similar studies, our proposed model showed good classification results for the breast lesion classification process. |
| Keywords: Breast cancer, Gray level run matrix, Monarch Butterly optimization, Support vector machine |

1. **Introduction**

Currently, breast cancer (BC) is one of leading cause of death among women worldwide. Early diagnosis of breast cancer is key functionality to reduce the mortality rates. Screening mammography is recommended imaging tool for early diagnosis [[1](#r1)]. Generally, radiologists use mammogram images to classify breast lesions as malignant versus benign. This classification process can be challenging task for radiologists as they interpret many mammogram images daily. In recent years, computer-aided (CAD) systems have been used as a second opinion by the radiologists to make the final diagnosis decision and also possible means to lessen the possibilities human errors. Typical a CAD system includes four main parts; (1) segmenting of breast lesion (2) feature extraction (3) feature selection (4) classification. The performance of a CAD system mainly relies on three issues: extraction of features from an image, selection of optimal features from the extracted features and classification [[2](#r2)]. Therefore, this study focuses on feature extraction, feature selection and classification. This study proposes Gray Level Run Matrix(GLRM) for feature selection, Monarch butterfly optimization (MBO) for feature selection and Support Vector Machine (SVM) for classification. This combination methods (GLRM/WOA/SVM) are the novelty of this study. The main purpose of this study is to investigate whether the proposed MBO feature selection improves the performance of the SVM algorithm.

1. **Literature Survey**

Bajsci et al. [[3](#r3)] used gray level run matrices for feature extraction. The study used decision tree and random forest for classification. The experiments conducted on The Mammographic Image Analysis (MIAS) dataset and they achieved 100% of accuracy.

Punitha et al. [[4](#r4)] utilized a neural network to classify malign and benign breast lesions. The study uses GLCM and GLRM features. Their study achieved 98% of accuracy for Digital Database for Screening Mammography (DDSM).

Mohanty et al. [[5](#f5)] suggested a hybrid model for breast cancer classification. They used 2-D wavelet transform and gray level co-occurance matrix (GLCM) for feature extraction. Forest optimization algorithm was used for feature selection. Support vector machine, k-nearest neighbor and decision tree algorithms were used for classification. The experiments were conducted on MIAS and DDSM datasets. Decision tree-based model achieved the best performance with 100% of accuracy compared to other models.

Jona and Navegeni [[6](#r6)] suggested Genetical swarm optimization (GSO)-SVM based model for breast cancer classification. Gray Level Co-occurance Matrix (GLCM) approach is utilized for feature extraction. The experiments are conducted on MiniMIAS database. The suggested model achieved 94% of accuracy.

Candra [[7](#r7)] et al. suggested an CAD system for breast cancer prediction. They used GLRM and GLCM approach feature extraction. The experiments are conducted on MiniMIAS database. The best performance shower 93.97% of accuracy using polynomial kernel.

Dheeba et al. [[8](#r8)] proposed an CAD model for breast microcalcification classification. They used the combination particle swarm optimization (PSO)-feed forward neural network (FFNN). Laws texture approach is used for feature extraction. The experiments are conducted on MiniMIAS database and clinical database. The suggested model showed 0.97 of accuracy for MIAS and 0.913 of accuracy for clinical databese.

1. **Materials and Methods**

The suggested CAD approach includes four main parts namely, data acquisition of mammogram images dataset, feature extraction, feature selection and classification. The proposed of architecture is shown in [Figure 1](#f1).

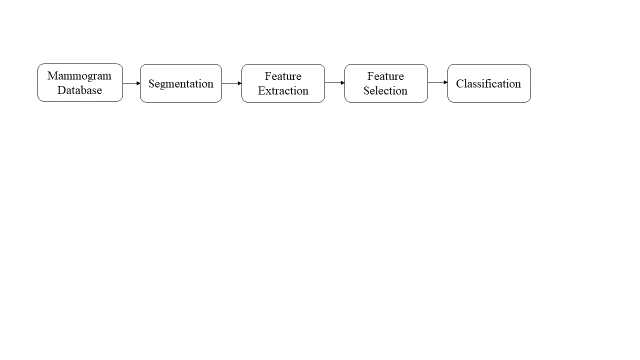
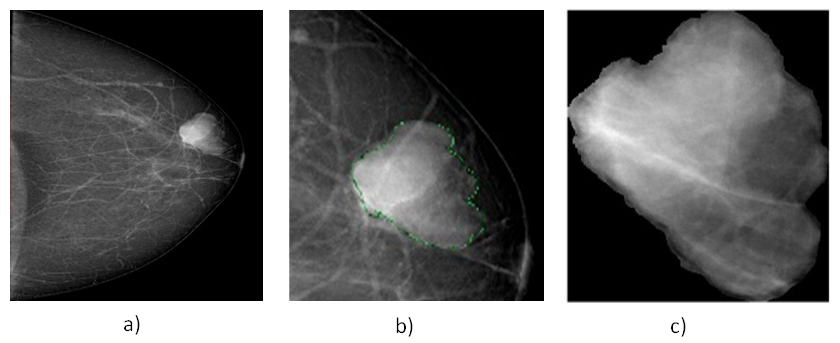


Figure : The framework of the suggested model

**3.1 Mammogram Database**

The proposed methodology was evaluated on a mammography dataset provided by the Department of Radiology at Ankara Education and Research Hospital. This dataset was approved by the ethics committee of Ankara Training and Research Hospital. Due to the retrospective nature of this study, informed consent was waived. All patients who underwent digital mammography were retrieved from the Picture Archiving and Communication System (PACS) between April 2015 and April 2020. All patients underwent mammography using IMS Giotto (Bologna-Italy). The datasets consist of 195 mammogram images (116 images (59%) for malign, 79 images (41%) for benign). The mammogram images were segmented to identify breast lesions. The stages of the segmentation process are demonstrated in Figure 2. First, the mammogram images were retrieved ([Figure 2-a](#f2)). Then, the borders of the breast lesions were determined by the green contour with the radiologists. ([Figure 2-b](#f2)). Finally, the breast lesions were extracted from the image, using the gray level thresholding and morphological operations ([Figure 2-c](#f2)).

****

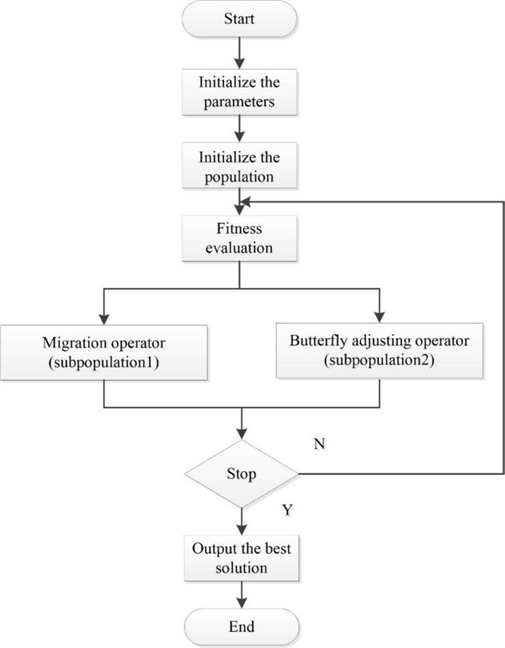
**Figure 2:** a) Original mammogram image, b) Marking of ROI with green contours

**3.2. Feature Extraction**

Mammogram images contain a lot of confidential information based on pixels that the human eye cannot see. Feature extraction is the process of extracting measurable pixel-based information from images. This study, GLRM approach is used for feature extraction. GLRM approach provides about gray level runs (number of pixels with the same intensity in a given direction), describing the texture of an image. These features characterize the distribution of short and long runs of an image in a specific direction. *Short run emphasis (SRE), long run emphasis (LRE), gray-level nonuniformity (GLN), run length nonuniformity (RLN), run percentage (RP), low gray-level run emphasis (LGRE), high gray-level run emphasis (HGRE), short run low gray-level emphasis (SRLGE) short run high gray-level emphasis (SRHGE), long run low gray-level emphasis (LRLGE) and long run high gray level emphasis (LRHGE);* 11 GLRM features were extracted [[3](#r3)].

**3.3. Feature Extraction**

The size of features can affect the performance of classification process due to high computational cost. Therefore, feature selection approaches are used to eliminate redundant and irrelevant features to improve the performance of classification process and reduces the computational cost. In this study, monarch butterfly optimization (MBO) algorithm is used for feature selection. MBO is a swarm-based algorithm which is developed by Wang et al. [[9](#r9)]. This mimics by monarch butterfly migration behavior. MBO algorithm consists of two main equal-sized subpopulation which are subpopulation 1 and subpopulation 2. The migration operator and butterfly adjusting operator generate of two main strategies of this algorithm. The strategies are based on iteration. The global optimal information is separated during iterations and the subpopulations are revised into population again. According to the determining new fitness value, the whole population is divided by two subpopulations When the termination condition is met, this process is ended [[10](#r10)]. The mathematical formulation of MBO algorithm is explained in details in ref. 10.

****

**Figure 3:** The framework of Monarch butterfly optimization [[10](#r10)]

**3.4 Classification**

Classification is the process of assigning unclassified images to their own classes within categories. With the high discriminating ability, machine learning algorithms have been used for breast lesion classification in recent years. In this study, support vector machine algorithm is used to classify breast lesion as malignant versus benign. SVM is a popular machine learning algorithm which uses generally regression and classification process. The primary aim of this algorithm is to build a line or hyperplane which splits the data into classes. Support vectors are defined as data points near the hyperplane, while data points are called margin [[2](#r2)].

**4. Experimental Results and Discussion**

This section of the study presents the results of experiments. Experiments were carried out using the MATLAB 2020a program to validate the predictive model. 10-fold cross validation is used for evaluation of suggested model. SVM algorithm is used to classify breast lesions malignant versus benign. The performance results of SVM are evaluated within the framework of two different scenarios. In the first scenario, the performance of SVM is analyzed without using feature selection method. In the second scenario, the performance of SVM is analyzed after applying the MBO-based method. The parameters for MBO are set as follows: *population size=50*, *max step=1.0, butterfly adjusting rate=5/12, migration period=1.2, migration ratio=5/12 and maxium**generation=50* [[10](#r10)]. Experimental results were evaluated in terms of accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV) and negative predictive value (NPV) performance metrics calculated using the confusion matrix shown in [Figure 4](#f4). The results of both scenarios are summarized in Table 1. The results are evaulated in terms of in terms of accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV) and negative predictive value (NPV). After applying MBO based feature selection, two features are rest which are SRE and RP. Without applying the feature selection method (WFS), the SVM algorithm shows 0.888 of accuracy, 0.899 of sensitivity, 0.87 of specificity, 0.914 of PPV, and 0.849 of NPV. After applying the MBO feature selection method, the SVM algorithm shows 0.944 of accuracy, 0.949 of sensitivity, 0.936 of specificity, 0.957 of PPV, and 0.924 of NPV. After using the MBO feature method, 2 of the 11 GLRM features are selected as distinctive features which are SRE and RP. After the predictive model is created, the efficiency of the model can be checked. For that, the results of without feature selection (WFS) and MBO based method are compared in [Figure 5](#f5). As analyzed [Figure 5](#f5), MBO based method enhances the performance of SVM algorithm in terms of ACC, SEN, SPE, PPV and NPV.

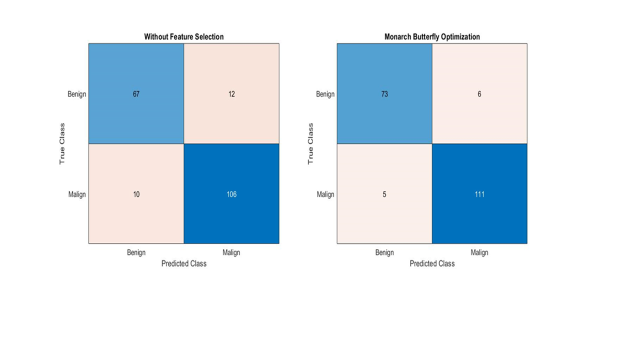


Figure : Confusion matrix of suggested model

Table : Comparison of WFS and MBOA methods

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classification | Method | Selected Features | ACC | SEN | SPE | PPV | NPV |
| SVM | WFS | 11 | 0.888 | 0.899 | 0.87 | 0.914 | 0.849 |
| MBO | 2 | 0.944 | 0.949 | 0.936 | 0.957 | 0.924 |

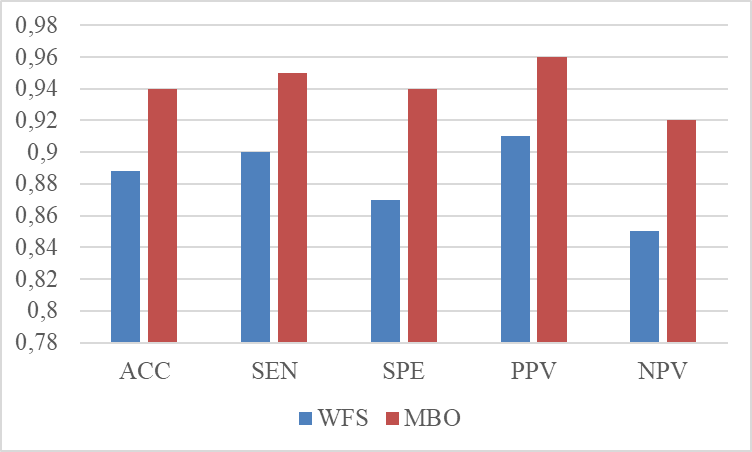


Figure : Comparison of WFS and MBO model

1. **Conclusion**

An efficient CAD scheme for mammographic breast lesion classification has been suggested in this study. GLRM approach is used for feature extraction. Further, the most discriminating features are selected using monarch butterfly optimization (MBOA). SVM algorithm is used for breast lesion classification. The experimental analyzes are conducted on a private mammographic dataset. The proposed scheme (GLRM+MBO+SVM) shows 0.944 of accuracy. Our proposed method yields similar results when compared with similar recent studies, as shown in [Table 2](#t2). As a result, we believe that the suggested method shows a good result in the diagnosis of breast cancer. n the future, other feature selection methods and classification methods can be considered as potential alternatives to the proposed scheme.

**Table 2:** Comparison of suggested with some recent similar studies

|  |  |  |  |
| --- | --- | --- | --- |
| **References** | **Proposed Method** | **Dataset** | **ACC** |
| [[6](#r6)] | GLCM+GSO+SVM | MIAS | 0.94 |
| [[7](#r7)] | GLRM+SVM | MIAS | 0.939 |
| [[8](#r8)] | PSO-FFNN | MIAS  Private Dataset | 0,976  0,913 |
| Proposed method | GLRM+MBO+SVM | Private Dataset | 0.944 |

**References**

1. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249.
2. Hassan, N. M., Hamad, S., & Mahar, K. (2022). Mammogram breast cancer CAD systems for mass detection and classification: a review. *Multimedia Tools and Applications*, 1-
3. Bajcsi, A., Andreica, A., & Chira, C. (2021). Towards feature selection for digital mammogram classification. *Procedia Computer Science*, *192*, 632-641.
4. Punitha, S., Amuthan, A., & Joseph, K. S. (2018). Benign and malignant breast cancer segmentation using optimized region growing technique. *Future Computing and Informatics Journal*, *3*(2), 348-358.
5. Mohanty, F., Rup, S., Dash, B., Majhi, B., & Swamy, M. N. S. (2020). Digital mammogram classification using 2D-BDWT and GLCM features with FOA-based feature selection approach. *Neural Computing and Applications*, *32*(11), 7029-7043.
6. Jona, J., & Nagaveni, N. (2012). A hybrid swarm optimization approach for feature set reduction in digital mammograms. *WSEAS Trans Inf Sci Appl*, *9*(11), 340-349.
7. Candra, D., Novitasari, R., Lubab, A., Sawiji, A., Asyhar, A.H., 2019. Application of feature extraction for breast cancer using one order statistic, GLCM, GLRLM, and GLDM. Advances in Science, Technology and Engineering Systems Journal 4, 115–120.
8. Dheeba, J., & Selvi, S. (2012). A swarm optimized neural network system for classification of microcalcification in mammograms. *Journal of medical systems*, *36*(5).
9. Wang, G. G., Deb, S., & Cui, Z. (2019). Monarch butterfly optimization. *Neural computing and applications*, *31*(7), 1995-2014.
10. Feng, Y., Deb, S., Wang, G. G., & Alavi, A. H. (2021). Monarch butterfly optimization: a comprehensive review. *Expert Systems with Applications*, *168*, 114418.