Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 8, pp. 2213–2225
DOI : 10.47974/JIOS-1784

Design of optimization-based machine learning model with AI for effective breast cancer detection system

Anusha Ampavathi *
Singapore Institute of Technology
1 Punggol Coast Road
Singapore 828608

and

Department of Artificial Intelligence
Vidya Jyothi Institute of Technology
Hyderabad 500075
Telangana
India

B. Sivaneasan †
Specialist Adult Educator Engineering
Singapore Institute of Technology
1 Punggol Coast Road 828608
Singapore

Siva Shankar §
Department of Computer Science and Engineering
KG Reddy College of Engineering and Technology
Hyderabad 501504
Telangana
India

* *E-mail:* Anuampavathi@gmail.com (Corresponding Author)
† *E-mail:* sivaneasan@singaporetech.edu.sg
§ *E-mail:* drsivashankars@gmail.com

Tulika Chakrabarti ‡
Department of Chemistry
Sir Padampat Singhania University
Udaipur 313601
Rajasthan
India

Prasun Chakrabarti @
Department of Computer Science and Engineering
Sir Padampat Singhania University
Udaipur 313601
Rajasthan
India

Martin Margala #
School of Computing and Informatics
University of Louisiana at Lafayette
LA- 70503 (337) 482-6768
U.S.A.

Abstract

Accurate and prompt diagnosis is crucial for early detection and recovery from breast cancer. Many women die from breast cancer because current methods fail to detect it early. To address this challenge, we designed an Artificial Intelligence-based Random Forest (AI-RF) model optimized with Sail Fish Optimization (SFO). Using MRI breast cancer datasets from Kaggle, the system preprocesses data with median filtering and extracts features using the Grey Level Coherence Matrix (GLCM) method. The SFO fitness is updated in the RF model to improve prediction accuracy of breast cancer stages (benign and malignant). The developed model aims to enhance detection accuracy and reduce false rates. Experimental results show the model achieves 99.13% accuracy, 99.33% sensitivity, and 98.85% precision, outperforming existing techniques.

Subject Classification: *92B05, 92C20, 92C55.*

Keywords: *Breast cancer diagnosis system, Machine learning, Sail fish optimization, Random forest, Grey level coherence matrix.*

1. Introduction

Cells in the human body, visible only under a microscope, divide to replace damaged and aged ones [1]. Uncontrolled cell growth can create

‡ *E-mail:* **tulika.chakrabarti@spsu.ac.in**
@ *E-mail:* **drprasun.cse@gmail.com**
# *E-mail:* **Martin.margala@louisiana.edu**

malignant cells, leading to tumours [2]. Breast cancer, a prevalent malignant tumour in women, is the second-leading cause of mortality worldwide. Symptoms include breast lumps, skin wrinkles, shape changes, and red plaques [3]. Early identification and classification of breast cancer help medical professionals provide appropriate therapies and monitor relapses. Early detection can reduce mortality risks and improve survival rates. Mammograms and doctor examinations are vital for diagnosing breast cancer, even without visible symptoms. Medical imaging, including mammography, MRI, ultrasound, and CT, is crucial for early cancer identification [4]. Machine learning (ML) enhances diagnosis accuracy, reducing human error and expediting the process. This study aims to develop an AI-based random forest model with sailfish optimization for precise breast cancer examination and diagnosis, improving diagnosis accuracy and classification performance [5]. Additionally, the main contribution of the developed model is

- Enhance the detection accuracy and reduce the false rate in the breast cancer diagnosis system.
- Improve the accuracy by classifying breast cancer stages using sailfish fitness.
- The gained model performance is validated with other prevailing techniques to prove the efficiency.

This study develops an AI-RF model using Sail Fish Optimization (SFO) for breast cancer detection. MRI datasets from Kaggle are preprocessed with median filtering to enhance quality, and GLCM extracts key features. The Otsu threshold method segments affected regions, and SFO optimizes the RF model for classifying benign and malignant stages. Performance is benchmarked against existing techniques across accuracy, precision, recall, false positive rate, F-measure, and error rate. Sections include a literature survey (section 2), methodology (section 3), detailed results and discussion (section 4), and conclusion (section 5).

2. Related works

Several studies have proposed advanced models for breast cancer detection and classification using various computational techniques. Yang Wang et al. [6] introduced a Convolution and Deconvolution Neural Network (CDNN) for breast cancer CT image identification and screening. They enhanced images using the Fuzzy C-means Clustering Algorithm

(FCM) and achieved improved segmentation accuracy. Huan-Jung et al. [7] proposed a technique employing Principal Component Analysis (PCA) and Multilayer Perceptron (MLP) to predict breast cancer based on nine variables. Their approach, integrating Support Vector Machines (SVM) in transfer learning, demonstrated effective cancer prevalence analysis. Roseline et al. [8] developed a healthcare IoT-based system using CNNs and ANNs for breast cancer classification, achieving high accuracy with the help of Particle Swarm Optimization (PSO) for feature selection. Shahan et al. [9] introduced an Internet of Medical Things (IoMT) cloud-based approach for the intelligent staging of breast cancer, showing superior accuracy in identifying cancer stages compared to existing methods.

Kiran Jabeen et al. [10] proposed a framework for breast cancer classification from ultrasound images using Deep Learning and feature fusion techniques, achieving an impressive accuracy of 99.1%. Viswanatha Reddy [11] proposed a computer-aided diagnostic (CAD) method utilizing CNN, SVM, and Random Forest classifiers for mammogram analysis, enhancing categorization accuracy. Zijun Sha et al. [12] employed the Grasshopper Optimization (GSO) method for identifying malignant regions in mammography images, achieving high sensitivity and accuracy.

Despite advancements, challenges remain such as high error rates, computational costs, scalability issues, and model complexities. Future research aims to address these to improve overall performance and efficiency in breast cancer detection and classification systems [13].

3. Proposed methodology

Designing an AI-based breast cancer detection system involves developing a random forest model enhanced with Sail Fish Optimization (SFO) for improved accuracy and speed. MRI datasets from Kaggle containing 1480 files are utilized, and divided into Sick (Malignant) and Healthy (Benign) groups for training and validation [13].

Dataset: The dataset is collected from the Kaggle website called Breast Cancer Patients MRIs. It contains a total of 1480 files and the pictures are separated into two groups: Sick (Malignant) and Healthy (Benign). Both categories include 700 MRI scan pictures of healthy and ill patients for training. Each category includes 40 MRI scan pictures of healthy and ill patients for validation. Preprocessing begins with median filtering to remove noise from MRI scans, crucial for enhancing prediction accuracy. This technique effectively reduces high-frequency noise while preserving

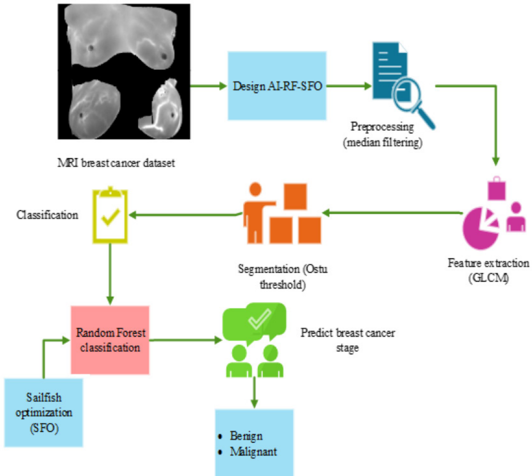

Figure 1

Architecture of the developed model

visual details. Feature extraction employs the grey-level co-occurrence matrix (GLCM), focusing on texture features crucial for accurate breast cancer categorization in MRI images. GLCM extracts seven key texture features: standard deviation, energy, mean, homogeneity, contrast, entropy, and correlation. The classification phase integrates SFO within the RF model to predict breast cancer stages (benign or malignant) more effectively. This novel approach enhances classification results and processing speed, showcasing the model's improved accuracy. Figure 1 illustrates the architecture of this proposed AI-RF-SFO model, designed to advance breast cancer detection systems through innovative optimization techniques.

The median filter selects the centre element n of the set numbers (i, j), replaces the central pixel, and then arranges all of the neighbourhoods (a, b) in the order ascending to filter a neighbourhood. The process of median filtering is obtained using eqn. (1).

$$K_{(a,b)} = M(n_{(i,j)}, (i, j) \in N_c) \quad (1)$$

Let, N_c is denoted as the centred neighbourhood on the location (a, b) in an image. The salt and pepper noises are eliminated using the median filter. The median filter is used in this work to eliminate digital noise from the MRI dataset.

- **Correlation**

  The measurement of the linear shape of a grayscale image is correlated and is identified in Eqn (2).

$$C_r = \sum_{a=1}^{m} \sum_{b=1}^{m} \frac{(a-\mu_a)(b-\mu_b)V(a,b)}{\sigma_a \sigma_b} \tag{2}$$

  Let, a and b are denoted as line and column, $V(a,b)$ is considered as congruent matrix probability in the MRI images and m is represented as the quantity of elements in the matrix of maintained length.

- **Standard deviation**

  The standard deviation, which is computed using Eqn (3), is a range of an image's histograms.

$$\sigma = \sqrt{\sum_{a=1}^{m} \sum_{b=1}^{m} V(a,b)(a-\mu_a)^2} \tag{3}$$

- **Energy**

  At multiple coordinates, energy displays the number of pixel pairs in the equivalent matrix's intensity. Equation (4) represents the energy.

$$E_n = \sum_{a=1}^{m} \sum_{b=1}^{m} V(a,b)^2 \tag{4}$$

- **Mean**

  The picture's mean is a measure of its dispersion, which is computed using equation (5)

$$\mu = \sum_{a=1}^{m} \sum_{b=1}^{m} a(V(a,b)) \tag{5}$$

- **Homogeneity**

  A degree of grey matching is displayed in an image when homogeneous. Homogeneity is represented by the equation (6).

$$H_m = \sum_{a=1}^{m} \sum_{b=1}^{m} \frac{V(a,b)^2}{1+(a-b)^2} \tag{6}$$

- **Contrast**

  The amount of diversity or change in grey intensity in an image is called contrast. The equation of (7) gives the contrast.

$$C_n = \sum_{a=1}^{m} \sum_{b=1}^{m} |a-b|^2 V(a,b) \tag{7}$$

- **Entropy**

  Entropy provides details about the image's contents. It provides levels of randomness, intensity, and uncertainty to the image. Entropy is computed using equation (8).

$$E_n = \sum_{a=1}^{m} \sum_{b=1}^{m} - \ln |(V(a,b))| \times V(a,b) \tag{8}$$

Finally, extracted textural features are updated to the segmentation phase for segmenting affected regions of breast cancer. It enhances the classification results.

3.4 *Segmentation using the Ostu threshold method*

The Ostu image segmentation method seeks to identify the ideal threshold for image segmentation. The threshold in Ostu divides the image into foreground and background zones, and the best segmentation threshold is determined by taking the threshold with the most significant amount of variance in the background and foreground regions.

The interclass variance reaches a maximum, φ is the optimum threshold. The input, preprocessed, and segmented results of the developed model are shown in Figure 2.

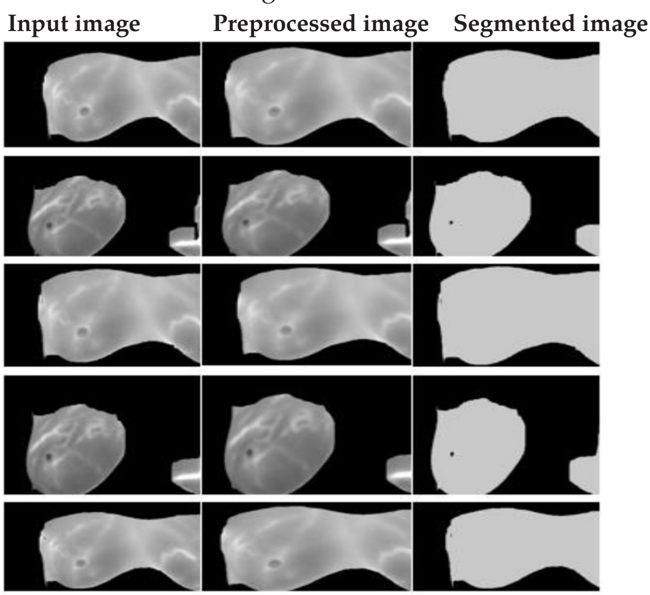

Figure 2

Input, pre-processed, and segmented results of the developed model

The developed model segments the affected region of the breast cancer using their optimum threshold. It segments the affected area from background and foreground images.

Classification using RNN-SFO

Classification involves assigning class labels based on feature-extracted and segmented data, distinguishing between benign and malignant categories. This model optimizes Random Forest (RF) using Sailfish Optimization (SFO) to enhance prediction accuracy. RF employs multiple decision trees on random subsets of data, aggregating their votes for final classification.

Sailfish Optimization is a meta-heuristic algorithm inspired by predatory behaviour in fish schools, aiming to optimize search space exploration. It uses the positions of sailfish (solution candidates) and sardines (supporting variables) to guide search and classification tasks efficiently.

RF models are known for their robustness against noise and high performance in generalization tasks. The Gini index guides RF's decision tree creation, ensuring effective feature splitting and model robustness. Each tree operates independently on bootstrapped samples, aggregating results to enhance classification accuracy. In summary, the combination of RF with SFO enhances processing speed and classification accuracy, making it a potent tool for breast cancer detection and classification tasks. The classified results of the developed model are shown in Figure 3.

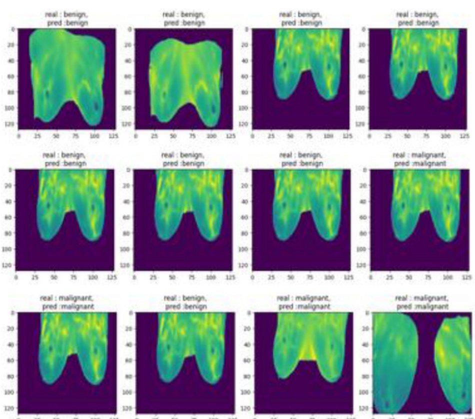

Figure 3

Classified results

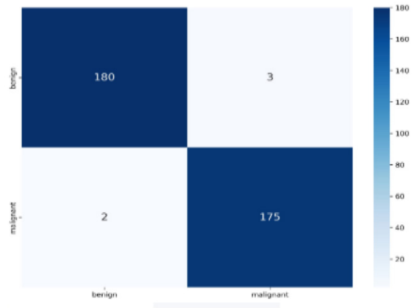

Figure 4

Confusion matrix

4. Results and discussion

Experimental simulations are implemented on the Python platform using MRI breast cancer databases to assess system effectiveness. The process begins with preparing the input dataset via median filtering, enhancing data quality by reducing noise. GLCM extracts various image features to expedite the transformation of raw data into meaningful information. The Otsu thresholding method is then employed to isolate malignant regions from the background.

Subsequently, an enhanced Random Forest (RF) approach integrating Sailfish Optimization (SFO) is utilized for classification. The dataset is split with 30% for testing and 70% for training the model. The developed model's performance is evaluated using a confusion matrix, as depicted in Figure 4.

4.1 *Performance Analysis*

The AI-RF-SFO model's performance is benchmarked against other models across several metrics to validate its effectiveness:

- Accuracy: The model achieved 99.13% accuracy, outperforming PCA-MLP (86.97%), PSO (96.5%), IoMT (97.81%), and GSO (87%) for 100 epochs.
- Sensitivity: With a sensitivity of 99.33%, the model surpassed PCA-MLP (87%), PSO (97%), IoMT (98.86%), and GSO (89%) for 100 epochs.
- Specificity: The model attained 98.73% specificity, surpassing PCA-MLP (84.3%), PSO (95.7%), IoMT (96.5%), and GSO (88%) for 100 epochs.

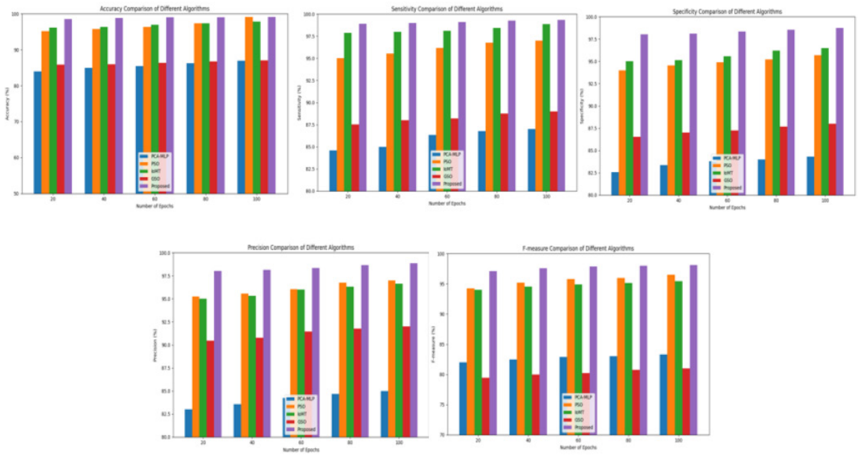

Figure 5

Comparison of a) Accuracy b) Sensitivity c) Specificity d) Precision e) F-measure

- Precision: Achieving 98.85% precision, the model outperformed PCA-MLP (85%), PSO (97%), IoMT (96.65%), and GSO (92%) for 100 epochs.
- F-measure: The model scored 98.13% F-measure, demonstrating superior performance compared to PCA-MLP (83.3%), PSO (96.5%), IoMT (95.43%), and GSO (81%) for 100 epochs.

Figure 5 compares the accuracy, sensitivity, specificity, precision, and F-measure of the designed model with other existing models, showing its superior performance in specificity, precision, and F-measure.

4.2 *Discussion*

Obtaining high-quality data remains a significant challenge in machine learning, yet it is crucial for training accurate forecasting models. The designed model leverages optimization within the RF framework to enhance classification parameters and achieve precise breast cancer prediction, reducing execution time and computational complexity. Its efficiency is validated against existing models, detailed in Table 1 for comprehensive comparison.

Table 1

Overall performance of the developed model

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
FCM [8]	94.3	94	93.21	91.12
PCA&MLP [9]	86.97	86.43	85.34	86
PSO [10]	96.5	95.98	95.23	94.87
DarkNet-53 [11]	99.1	98.86	98.34	98.04
CAD [12]	90.55	90	90.32	89.43
GSO [13]	87	86.54	86.12	86
Proposed (AI-RF-SFO)	99.76	99.54	99.14	99.05

The ROC graph of the developed model is shown in Figure 6.

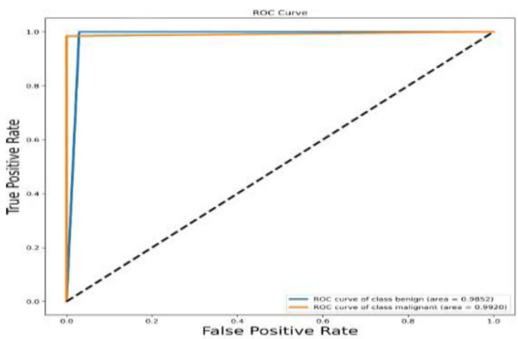

Figure 6

ROC graph

While comparing other models, the designed technique better predicted and classified breast cancer stages. The error and noise rates are also low compared to other models. The developed model takes less execution time to organize the breast cancer stages and improves the performance of the breast cancer diagnosis system.

5. Conclusion

The proposed AI-RF-SFO model categorizes breast cancer into benign and malignant stages, leveraging RF optimization to enhance performance. For 100 epochs, it achieved 99.13% accuracy, 99.33% sensitivity, 98.73%

specificity, 98.85% precision, and 98.13% F-measure. In a 60-epoch scenario, it reached 99.76% accuracy, 99.54% sensitivity, 99.14% specificity, and 99.05% precision, outperforming other models by 33%, 27%, 17%, and 34% respectively, showcasing its efficiency and reliability. Despite potential computational complexities, future research aims to enhance accuracy through advanced computational intelligence techniques and larger datasets, integrating optimization and deep learning methods for improved breast cancer detection.

References

[1] B. Saylan and S. Cinaroglu, "Opinion mining and machine learning analysis: What emotions Twitter data tell us about telemedicine?" *Journal of Statistics and Management Systems*, vol. 27, no. 3, pp. 605–633 (2024).

[2] A. Pareek, P. Arora, S. Madan, and N. Gupta, "Telecom customer churn prediction model: Analysis of machine learning techniques for churn prediction and factor identification in the telecom sector," *Journal of Information and Optimization Sciences,* vol. 45, no. 2, pp. 613–630 (2024).

[3] K. D. Vidhate, P. Nema, and T. Hasarmani, "Modelling for forecastng energy consumption using SBO optimization and machine learning," *Journal of Information and Optimization Sciences,* vol. 45, no. 2, pp. 605–612 (2024).

[4] R. Selvaraj, S. Vidyanandini, and S. R. Nayak, "Radial radio mean labelling of Grotzsch graph and some graphs," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 27, no. 4, pp. 1377–1387 (2024).

[5] B. K. Padhi, S. Chakravarty, B. Naik, S. R. Nayak, and R. C. Poonia, "Leveraging ensemble learning for enhanced security in credit card transaction fraudulent within smart cities for cybersecurity challenges," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 27, no. 4, pp. 1233–1246 (2024).

[6] D. K. J. B. Saini, S. K. Pawar, S. P. Tondare, A. S. Nigade, J. Morbale, and M. Gangwar, "Improve QoS for multi-body sensor analytics in smart healthcare system using machine learning algorithm," *Journal of Interdisciplinary Mathematics,* vol. 26, no. 3, pp. 393–405 (2023).

[7] A. Seethalakshmy, T. Tamilvizhi, K. N. Sowjanya, and B. Bhoomeshwar, "Deep learning-based computed tomography image classifica-

tion of COVID-19 patients," *Journal of Interdisciplinary Mathematics*, vol. 26, no. 3, pp. 371–381 (2023).

[8] A. A. Heydari and S. H. Z. Al-Thalabi, "Using a binary logistic regression model to diagnose the effect of factors causing cancer: An applied study on a sample of patients at the Oncology Hospital in Baghdad," *Journal of Statistics and Management Systems*, vol. 25, no. 8, pp. 2005–2017 (2022).

[9] A. Yala, P. G. Mikhael, C. Lehman, G. Lin, F. Strand, Y. L. Wan, et al., "Optimizing risk-based breast cancer screening policies with reinforcement learning," *Nature Medicine*, vol. 28, no. 1, pp. 136–143 (2022).

[10] E. Michael, H. Ma, H. Li, and S. Qi, "An optimized framework for breast cancer classification using machine learning," *BioMed Research International*, vol. 2022, no. 1, p. 8482022 (2022).

[11] S. Punitha, F. Al-Turjman, and T. Stephan, "An automated breast cancer diagnosis using feature selection and parameter optimization in ANN," *Computers & Electrical Engineering* (2021).

[12] E. Houssein, M. M. Emam, and A. A. Abdelmgeid, "An optimized deep learning architecture for breast cancer diagnosis based on improved marine predators algorithm," *Neural Computing and Applications,* pp. 18015–18033 (2022).

[13] Y. Suh, J. J. Jung, and B. Cho, "Automated breast cancer detection in digital mammograms of various densities via deep learning," *Journal of Personalized Medicine*, vol. 10, no. 4, p. 211 (2020).

Received February, 2024