Article - Engineering, Technology and Techniques

A Novel Deep Learning-based Whale Optimization Algorithm for Prediction of Breast Cancer

Poonam Rana[1*]
https://orcid.org/0000-0003-1985-4375

Pradeep Kumar Gupta[2]
http://orcid.org/0000-0003-0416-7097

Vineet Sharma[3]
https://orcid.org/0000-0003-4737-4272

[1]ABES Engineering College, Ghaziabad, Uttar Pradesh, India, Scholar AKTU; [2]Jaypee University of Information Technology, Department of CSE, Solan, HP, India; [3]KIET Group of Institutions, Delhi-NCR, Ghaziabad.

Editors-in-Chief: Alexandre Rasi Aoki
Associate Editor: Paulo Vitor Farago

*Correspondence: poonam.rana@abes.ac.in, Tel.: +91-8377841855, (P. R.).

HIGHLIGHTS

- Novel whale optimization algorithm is proposed for prediction of breast cancer.

- Deep learning-based WOA adjusts the CNN structure as per maximum detection accuracy.

- Proposed method achieves 92.4% accuracy in comparison to 90.3%.

- Validity of method is evaluated with magnifying factors like 40x, 100 x, 200x, 400x.

Abstract: Breast cancer is one of the most common cancers among women that cause billions of deaths worldwide. Identification of breast cancer often depends on the examination of digital biomedical photography such as the histopathological images of various health professionals, and clinicians. Analyzing histopathological images is a unique task and always requires special knowledge to conclude investigating these types of images. In this paper, a novel efficient technique has been proposed for the detection and prediction of breast cancer at its early stage. Initially, the dataset of images is used to carry out the pre-processing phase, which helps to transform a human pictorial image into a computer photographic image and adjust the parameters appropriate to the Convolutional neural network (CNN) classifier. Afterward, all the transformed images are assigned to the CNN classifier for the training process. CNN classifies incoming breast cancer clinical images as malignant and benign without prior information about the occurrence of cancer. For parameter optimization of CNN, a deep learning-based whale optimization algorithm (WOA) has been proposed which proficiently and automatically adjusts the CNN network structure by maximizing the detection accuracy. We have also compared the obtained accuracy of the proposed algorithm with a standard CNN and other existing classifiers and it is found that the proposed algorithm supersedes the other existing algorithms.

INTRODUCTION

Recent technological advancements have brought several changes in the lifestyle of a human which in turn have increased the risk for many lifestyle diseases like cancer, diabetes, hypertension, etc. According to statistics as provided by International Institute for Cancer Research (IARC), cancer is increased by 28% from 2006 to 2016, and it is also predicted that 2.7 million new forms of cancer will take place by 2030 [1]. To solve the problem of cancer and to save human life, the early diagnosis of cancer must be performed [2]. Different environmental conditions are considered as the human body introduces a variety of cancers, such as lung cancer, leukemia, cervical cancer, ovarian cancer, vaginal cancer, throat cancer, and breast cancer [3,4]. From a variety of cancers, breast cancer is the 2nd foremost reason for death among females [5].

Breast cancer prediction is essential in its earlier stage to reduce mortality. Diagnosis of breast cancer includes nuclear imaging, computed tomography (CT) scan, ultrasound, magnetic resonance imaging (MRI) scan, microscope, and mammography. However, none of these methods afford a more proficient outcome in the prediction of cancer. Several symptoms could lead to the identification of breast cancer [6]. The breast cancer types are categorized by benign and malignant [7]. The benign can be mentioned as the initial stage of breast cancer. The malignant stage is mentioned as severe cancer in the breast. Breast cancer is characterized by the uncontrolled improvement of abnormal cells on the glands milk-producing situated in the breast [8]. Therefore, the identification of breast cancer is very much significant.

To identify breast cancer, tissue-based detection is mainly done with stained technology. In this approach, tissues should be colored by adding the frequently used hematoxylin and eosin (H&E) element. Through the use of adding components, the images are improved with high resolution to look for foreign elements, types and cell structures. The diagnosis of tumors, the histopathology test is required [9]. The procedure of emerging tools for analyzing histopathological images is hampered by the subsequent tissues. Further, the limitation of feature extraction methods of breast cancer for histopathological images is another challenging task. Some existing feature extraction approaches are known as grey-level correlation matrix (GLCM) and scale-invariant feature transform (SIFT), etc., that depend on monitored information. Moreover, selecting useful features requires previous knowledge of the data, which leads the capability of feature extraction very low and the computational burden too large. In this paper, an effective deep learning-based solution has been proposed to handle these problems which efficiently retrieve information, and extract features automatically from the input data. The proposed technique can solve the difficulties of traditional feature extractions and provide accurate diagnostic results.

MATERIAL AND METHODS

In the past, several Machine Learning (ML) based methods have been established for the detection and classification of breast cancer. Several ML classifiers like Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) [10], Convolution Neural Network (CNN) are used for the detection of breast cancer [11]. Lately, deep learning-based approaches have superseded the conventional machine learning-based approaches [12]. The CNN based deep learning technique has provided better performance for the prediction of different diseases in the image processing task [13,14]. In this section, we have included the recent and related studies of various approaches used for the prediction and detection of breast cancer.

In [6], Tapak and coauthors have developed Linear Discriminant Analysis, Logistic Regression, Adabag, Least Square SVM, SVM, AdaBoost, Random Forest and Naive Bayes for breast cancer prediction analysis. As discussed, the proposed method was used to analyze six different machine learning methods for predicting breast cancer. They have considered the dataset of 550 breast cancer survivors. In [15], Tseng and coauthors have presented four different classifications for predicting breast cancer. The developmental method was evaluated based on serum human epidermal growth factor receptor 2 (SHER2). In this paper, they have considered various classifications such as Bayesian classification algorithms, logistic regression, SVM and random forest. For experiments, they have considered the dataset of 302 patients at least 3 months before the diagnosis of breast cancer. In [16], Arresta and coauthors have developed neurological networks for diagnostic imaging in breast caper histology images (BACH). BACH may be organized for the localization and classification of histopathological classes associated with the entire slide images of the microscope. Microscopic images were considered benign and malignant. The developed method provided excellent

accuracy in microscopic films. In [17], Ouyang and coauthors have used a LASSO (less absolute shrinkage and selection operator) regression for the diagnosis of breast cancer. A nomogram was developed to complete the area under the curve (AUC) of receiver operating characteristic of negative predictive value, positive predictive value, specificity, sensitivity, and accuracy. In this study, they have diagnosed 200 patients with breast cancer. In their results, breast cancer was found to be optimal in terms of the LASSO classifier, and it was compared with other approaches for analyzing the efficiency of the implemented method. In [18] Atrey and coauthors have proposed a dominance-based filtering approach. The proposed machine learning-based approach considers nine features and utilizes the Wisconsin Breast Cancer Dataset (WBCD) to evaluate the performance of the algorithm. They have achieved an accuracy of 98.9% and 99.6% for the top 4 and top 5 dominant features respectively, in their experiments. In [19,20] authors have proposed a novel method for prediction of breast cancer by using whole slide images for sentinel to enhance the accuracy. From the analysis of the state of art, it is essential to provide a system that automatically identifies whether a patient is suffering from breast cancer or not by looking at histopathological images along with a good accuracy rate. To rectify this, a novel classification algorithm has been discussed in the proposed research to obtain better results in comparison to the existing methods.

Proposed architecture for breast cancer detection

Breast cancer detection at the initial stage is an essential task for reducing any casualty of affected humans. In the proposed methodology, we have introduced and assessed a deep learning-based architecture for automated breast cancer detection which is based on the combination of image classification and machine learning and methods. Convolution Neural Network with Whale Optimization Algorithm (WOA) is introduced for breast cancer detection. The detection of breast cancer can be processed with two major stages known as the pre-processing stage, and the classifier stage. The proposed architecture is illustrated in Figure 1.
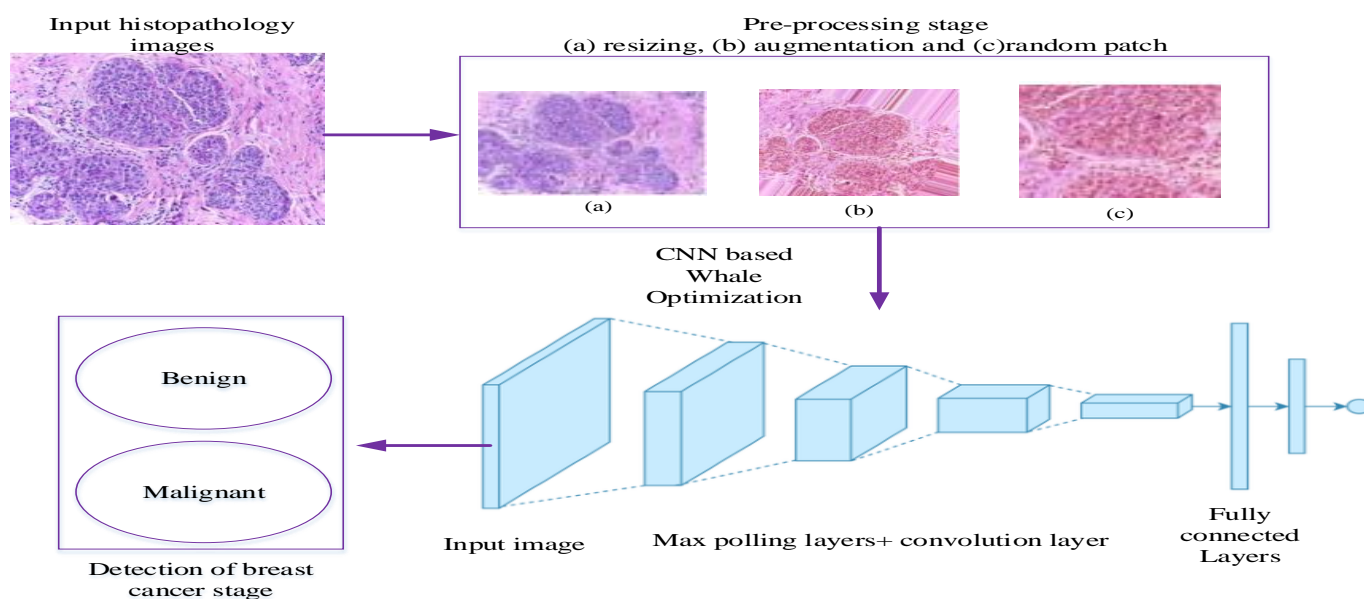

Figure 1. Overall proposed Architecture for breast cancer detection

The CNN based WOA has two different phases for the detection of the benign and malignant stages of breast cancer. The detailed process of the proposed methodology is discussed in the following steps.

The proposed model can used to detect other cancers, but in the current condition we have trained only Breast Cancer images. So this model will not predict the other cancers. But the same configuration of CNN can used to detect other cancers, but prior to this we need to perform training process using the respective cancer images.

Step 1: Dataset Collection and its overview

There are different types of datasets available for stained images of histopathological images for breast cancer such as Wisconsin original data set, MITOS-ATYPIA-14, and BreakHis. In the proposed design, histopathological images are utilized which are collected from the BreakHis. This dataset contains 7909 histopathological images collected from 82 different patients having breast cancer. All these histopathological

images are of pixel resolution with 700 × 460 sizes containing RGB channels. These images are collected with an objective lens and the whole dataset images have four different sub-datasets with the magnification factor 40×, 100×, 200×, and 400×. These sub-datasets are classified as benign and malignant tumors. Out of these patients, 24 patients have benign class and the remaining 58 patients have the malignant class of breast cancer. Benign classes contain Adenosis (A), Fibroadenoma (F), Tubular Adenoma (TA) and Phyllodes Tumor (PT). Malignant tumors include Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), and Papillary Carcinoma (PC). In the implementation process, the benign with subsets is considered as class one and malignant with subsets are considered as class two.

Wisconsin dataset contains attribute type dataset in contrast BreakHis dataset has image type dataset. Compared to Wisconsin dataset, BreakHis dataset is more suitable approach because it contains more different magnifying factor with histopathological images.

Instead of using histopathological images, the same work can also be performed using CT-Scans also. The histopathological images are microscopic images; therefore, they contain high level of information. Due to the usage of CNN, we can use any images, for the disease prediction. But whenever we change the image or disease type, it requires to reperform the training process. Then the proposed model is reliable for any type of images.

Step 2: Pre-processing stage

The main aim is to convert human visual images to computer understanding images with low pixel size. Usually, BreaKHis database contains the high resolution of digital pathological images. The more data the CNN learns, the more features it can generate. Therefore, it is essential to expand the dataset by various pre-processing methods for categorizing breast cancer type. For image-wise classification, here three different approaches used such as presented below.

Image resizing

The main purpose of the image resizing approach is to obtain a lower data volume, which accelerates processing time. The original sizes of the input histopathological images are too large to fit in memory. Therefore, the input images are cropped from the center or random positions in the image by resizing model. The resize scale varies randomly and is generated from 0.1 to 0.9, which generates diverse image sizes. The basic image representation and its resized outcome are given as follows:

Original image size (460, 700, 3)
Resized image size (276, 400, 3) with 0.6 scale

The proposed system also considers the images of different sizes and all the input images are initially resized to the size of 276, 400, 3.

Data augmentation

It is carried out to artificially expand the datasets which can lessen overfitting on the image data and enhance the performance of the algorithm. Overfitting difficulties occur when errors or random noise occur without defining the underlying relationship. With the help of the data augmentation technique, the input images are expanded so that the proposed classifier model can learn large irrelevant patterns during the training stage itself. Therefore, overfitting is avoided and possesses a higher performance. Various data augmentation methods are applied to the input images which are given as follows: sequential rotation by 40 degrees, width shift with factor 0.2, height shift with factor 0.2, shear with factor 0.2, zooming with the range of 0.2, horizontal and vertical flipping. These methods ensure that the dataset is expanded and helps to avoid the overfitting problem during the training phase.

Random patches

These are cropped to generate a patch database for testing and training phases by employing random patches, the obtained image format size is (50, 50, 3). The obtained pre-processed images are fed into the next stage of the proposed classifier algorithm.

Step 3: Applied Deep learning-based CNN process for classification of breast cancer

Deep learning-based CNN is introduced for the detection of breast cancer from the pre-processed histopathology images in step 2. Deep learning improves the success rate of detection because of the CNN

train on hierarchical representation. It is capable to extract features because it doesn't need any previous information. Due to its advantages, an automated diagnosis tool has been developed to permit non-specialists to plan their consequent structure of CNN without any prior knowledge of this field. By the inspiration of the brain of mammals (deep structure), the deep learning mechanism has been introduced in machine learning algorithms. The deep learning structure comprises of the various numbers of hidden layers and allows the abstraction of features by possessing various layers. This deep learning network consists of multiple hidden layers known as the convolution layer, subsampling layer, and the fully connected layer which are stated briefly in the following sub-sections. Here both the convolution and subsampling layers in the hidden layer helps to extract the features from low to high levels of features of the input data.

Layer 1: Convolution layer

Initially, a size of S×D input image is convolved by a filter or kernel of b×b size. In the input matrix, each block is convolved independently with the filter and produces a new pixel as the output of this layer. The convolution result of the input image and the filter utilized in this term tends to generate an output image of o features. In general, the output image features attained by the kernel of the convolving layer is represented as the input to the size of feature map j×j. The CNN structure comprises of the large number of convolution layer where its input and output of the subsequently convolutional layer be its feature vector. In each convolution layer, there is a o filter which is convolved with the input. And its depth of the feature map generated (o×) is equivalent to the count of feature map applied in the convolution procedure. At a definite location, each of the feature maps is defined as a specific feature of the input image. Thus, the output for the z-th layer in the convolution is represented as $D_i^{(z)}$ that contains a feature map and is mathematically represented as follows:

$$D_j^{(z)} = B_j^{(z)} + \sum_{i=1}^{b_j^{(z-1)}} L_{j,i}^{(z-1)} \times D_i^{(z)}, \tag{1}$$

Where $B_j^{(z)}$ represents a matrix of bias and $L_{j,i}^{(z-1)}$ is filter size or convolution kernel of b×b that links the feature map of $i-th$ in (z-1) layer with the same layer in $j-th$ feature map. Output $D_j^{(z)}$ the layer contains the feature maps. The convolution layer in first $D_j^{(z-1)}$ in equation (1) is the input space i.e. $D_j^{(0)} = Y_j$.

The kernel presented in the network structure obtains the feature map. After the convolution layer, the activation function is applied for the non-linear transformation of the output convolution layer.

$$X_j^{(z)} = X(D_j^{(z)}), \tag{2}$$

Here $X_j^{(z)}$ represents the activation function output and $D_j^{(z)}$ is the input function which it receives. ReLUs (rectified linear units). is the popular activation function than other traditional functions like sigmoid, tanh. Hence, ReLUs is utilized in this work and is mathematically defined as follows:

$$X_j^{(z)} = max(0, X_j^{(z)}, \tag{3}$$

Equation 3 reduces the interaction and non-linear effects to be adapted by the deep learning model. ReLUs possess the replacement activity where it replaces the output value to zero if it obtains the negative value and at the same time it precedes the same value if it is positive. Compared to the other functions, the activation function performs faster training due to its error deviation. Which becomes slow in the saturation region; therefore, the weight updates almost vanish and effect on vanishing gradient problems.

Layer 2: Sub-sampling layer

In this layer, the chief aspire is to lessen the dimensionality spatially of the feature map extracted from the preceding convolution layer. For performing this, a mask of size a×a is selected; this sub-sampling operation is attained between the feature maps and mask. There are various sum sampling methods such as sum pooling, average pooling, and maximum pooling. Among them, max pooling is an efficient process where each block's maximum value is considered as the output pixel of the image. This layer helps to tolerate the performance of rotation and translation of the convolution layer among input images.

Layer 3: Full connection

Based on the extracted feature from the preceding convolution layer, the classification of breast images is done in this layer. Here, each layer is connected with the previous layer of every neuron is achieved based on the traditional feed-forward neural network (one or more hidden layers). The softmax activation function is used in this output layer.

$$x_j^{(z)} = f\left(z_j^{(z)}\right),$$ (4)

Where:

$$z_j^{(z)} = \sum_{j=1}^{l_j^{(z-1)}} w_{j,i}^{(z)} x_j^{(z-1)},$$ (5)

Here $w_{j,i}^{(z)}$ represents the weights to form the representation of breast cancer class which is tuned by the fully connected layer and the transfer function is denoted as f symbolizes non-linearity. Finally, the classes are labeled from the output signals.

The parameter optimization has the main role in the proposed study for enhancing the performance of deep learning-based CNN architecture. Hence, for optimizing the CNN parameters, a novel WOA has been proposed which is based on the working principle of the whale and its food arranging technique. Also, for achieving the best performance measure of CNN, these training parameters are optimized that lead the lightweight CNN based on the characteristics of WOA and show the structure designed for histopathological breast classification. The Parameter optimization using WOA has been discussed in the subsequent sections.
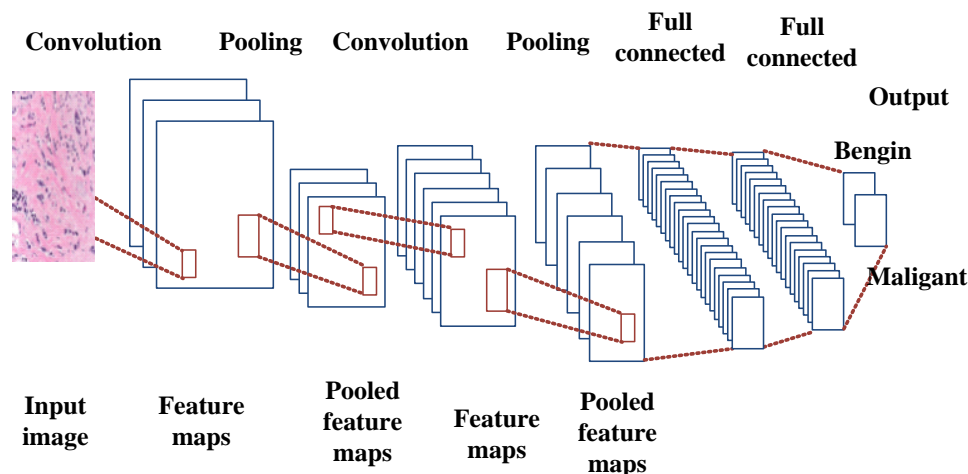

Figure 2. Proposed deep learning-based CNN architecture

Parameters to optimize

Parameters in CNN like the size of the kernel, padding, count of feature maps and the type of pooling are planned to optimize by the proposed optimization algorithm. Here, the parameter named stride is not optimized by WOA to ensure a large search space and make the problem solvable by using CNN.

Proposed deep learning based woa

The proposed deep learning-based WOA helps to enhance the speed of training procedure using optimally selecting the parameters pixel resolution. The working principle of the whale optimization technique is that the humpback whales hunt the prey using three operators namely; searching the prey, encircling the prey and forming a bubble net for the hunting process. The overall process of the whale optimization algorithm is clearly stated in Figure 3.

Mathematical modeling

The arithmetical representation of encircling the prey, spiral bubble-net feeding activities and searching for prey are demonstrated in this segment.

Phase 1: Initialization

The initialization phase of the proposed algorithm is generated by developing the initial solution randomly. For instance; after the histopathological image of breast cancer is pre-processed, its pixel size generated by the parameters of CNN is optimally selected with the help of the proposed optimization algorithm. Here, the parameters of CNN like the number of kernels, padding, type of pooling, number of feature maps and number of whales or said to be whale population are randomly initialized. Therefore, random value in the search space is represented as follows:

$$E(u)=(e_1,e_2, \ldots e_h), \tag{6}$$

Here E defines the original population of the whale at h represents the number of interconnection layers to optimize.

Phase 2: Fitness calculation

For automatic breast cancer detection, the fitness function is generated to achieve the best classification measure by maximizing its accuracy and it is evaluated based on the below expression:

$$ff_{E(u)}=maxi \ (Accuracy), \tag{7}$$

Phase 3: Update the position of current solution-Encircling the prey

In this phase, the hunting process by whales is started while noticing the position of prey and then it will encircle the prey. Then, the best solution (whale) is found out which is considered as the finest whale. Towards that best whale, the other whales will move on after updating its position. The update procedure of the whales is depicted by the underneath equations:

$$\vec{V}=\left|\vec{H}\cdot\vec{E}^{best}(u)-\vec{E}(u)\right|, \tag{8}$$

$$\vec{E}(u+1)=\vec{E}^{best}(u)-\vec{C}\cdot\vec{V}, \tag{9}$$

Where u represents the current iteration, $\vec{E}^{best}$ defines the best solution, $\vec{E}$ refers to the current position, $\vec{C}$ and $\vec{H}$ denotes a coefficient vector, $|C *H|$ denotes the absolute point. In addition to this, the coefficient vectors are mathematically represented are follows: $\vec{C}=2\vec{c}\cdot\vec{o}-\vec{c}$ and $\vec{H}=2\cdot\vec{o}$. Where $\vec{c}$ is a sequence of repetitions linearly from 2 to 0, $\vec{o}\in(0,1)$ for both the exploration and exploitation phases.
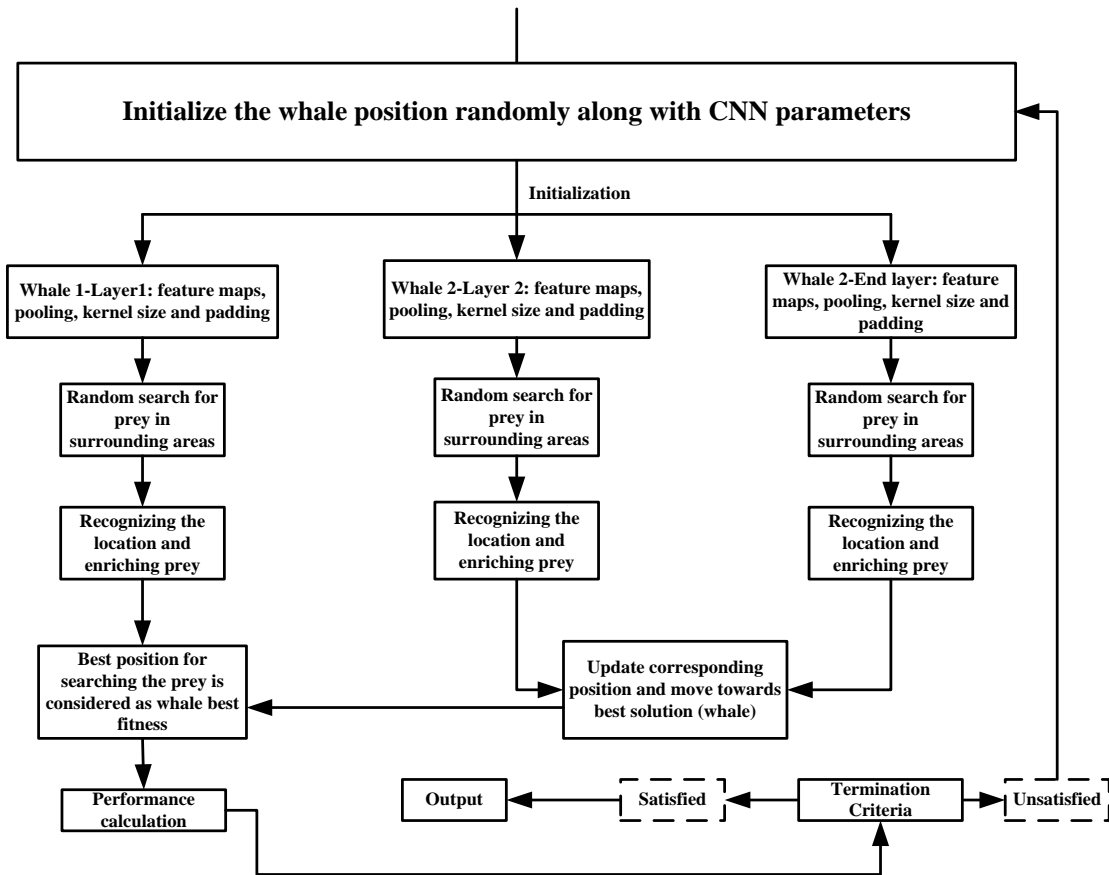
Figure 3. The layout of the proposed optimization algorithm

Exploitation phase:

This phase is also said as Bubble-net attacking technique. There are two mechanisms:

i. Shrinking encircling mechanism: is mathematically given by the following equation: $\vec{C}=2\vec{c}\cdot\vec{o}-\vec{c}$, as said earlier, its $\vec{c}$ value is decreased to reach this performance. Here, $\vec{c}$ is utilized for the reduction of different range of $\vec{C}$. Otherwise, it is stated that in the interval ranges from [-c,c], $\vec{C}$ is an accidental point where c is reduced from 2 to 0. The finding agent's new location can different wherever for $\vec{C} \in [-1,1]$.

ii. Spiral updating position: is calculated between the prey and the position of whale which is derived as follows:

$$\vec{E}(u+1)=\vec{V}_{Dist}\cdot\exp^{m_t s}\cdot\cos(2\textstyle\prod s)+\vec{E}^{best}(u), \tag{10}$$

Where $V_{Dist}=\left|\vec{E}^{best}(u)-\vec{E}(u)\right|$. It is meant to be the distance among the y-th whale and the prey which is denoted as the best solution achieved so far, s is supposed to take value from [-1, 1], m is represented as the shape of the logarithmic spiral. While performing optimization, the location of the whale having a probability of 50 percentage by selecting any of the shrinking or spiral encircling model and its mathematical equation is to be followed:

$$\vec{E}(u+1) = \begin{cases} \vec{E}^{best}(u) - \vec{C} \cdot \vec{V}, & \text{if } P < 0.5 \\ \overrightarrow{V_{Dist}} \cdot \exp^{m_t s} \cdot \cos(2 \prod s) + \vec{E}^{best}(u), & \text{if } P \geq 0.5 \end{cases}, \tag{11}$$

Where $P \in [0,1]$, thus, the humpback whales randomly finding the prey to form a bubble net.

Exploration phase:

This phase also is known as searching the prey. The subsequent equation elaborates on the mathematical form of the exploration phase.

$$\vec{V} = \left| \vec{H} \cdot \vec{E}^{random} - \vec{E} \right|, \tag{12}$$

$$\vec{E}(u+1) = \left| \vec{E}^{random} - \vec{C} \cdot \vec{V} \right|, \tag{13}$$

The random position of the current population is represented as $\vec{E}^{random}$. During the updating process of each solution, the fitness calculation is evaluated for finding the most excellent solution among them.

Based on the obtained best solution, a set of novel solutions is found and the fitness function is calculated for continuing the above solution updating process.

Phase 4: Termination criteria

Atlast, it satisfies the finest parameters of CNN by the hunting behavior of the whale. As a result of finding the optimal solution or best fitness function, the prediction model is qualified. Since the objective function is to improve the accuracy of training data, the prediction model obtained for the best fitness structure is well qualified to predict unknown data.

Testing phase:

After the training process gets over and the proposed model was tested with some set of images in the testing phase. For those images, precision, recall, F1-measure and accuracy for each of the classes have been computed. By calculating all those measures, the overall accuracy is generated. The most interesting part of the proposed algorithm is that the proposed CNN structure extracts the features of an image locally which means that the network will learn specific patterns within the image and can be able to recognize it anywhere in the image. The steps will be repeated until the image is scanned.

In the testing phase the 20% of database images are tested to evaluate the performances of CNN. For example, in this work 1582 images are tested, then the performance is measured based on the number of images predicted correctly.

RESULTS AND DISCUSSIONS

Detection of breast cancer from histopathological images is initiated with the help of the proposed technique by classifying it into two types of classes known as benign and malignant. The performance metrics like accuracy, precision, recall, sensitivity, and specificity are evaluated for the proposed and existing methods that discuss the efficiency of the presented algorithm.

Initially, the pre-processing process is done by using Python, in which for the different images, the resizing, augmentation and random patches are applied. These transformations are done during the run time of the program; hence it doesn't require any additional memory space for the images to store, which is an additional benefit of the work. The implementation detail for the proposed methodology is presented in this section. Some of the sample database images are displayed in Figure 4 and database structure is provided in Table1.
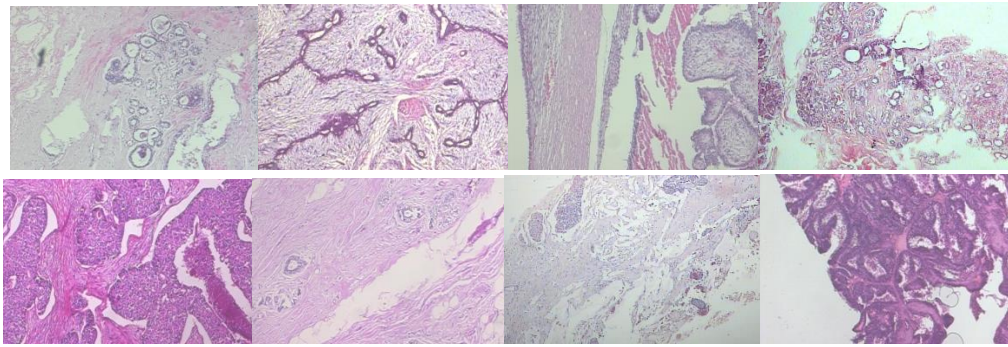
Figure 4. Sample database images

Table 1. Dataset Summary

Magnification Rate	Type of class							
	Sub-class of Benign					Sub-class of Malignant		
	A	F	TA	PT	DC	LC	MC	PC
40×	114	253	109	149	864	156	205	145
100×	113	260	121	150	903	170	222	142
200×	111	264	108	140	896	163	196	135
400×	106	237	115	130	788	137	169	138
Total	444	1014	453	569	3451	626	792	560

Evaluation metrics

The system performance is analyzed by using the most common evaluation metrics like precision, recall, F-measure and accuracy. Various evaluation metrics considered for performance evaluation of the proposed system are as follows:

Precision

Precision provides information about the effectiveness of the proposed system. Precision is defined as the ratio of the number of relevant images retrieved to the number of images retrieved.

$$\text{Precision} = \frac{no(Ret_{Rel})}{no(Ret)}, \tag{14}$$

Where:
$no(Ret_{Rel})$- Number of relevant images retrieved
$no(Ret)$- Number of images retrieved

Recall

Recall provides information about the accuracy of the proposed system. The recall is defined as the ratio of the number of relevant images retrieved to the total number of relevant images in the database.

$$\text{Recall} = \frac{no(Ret_{Rel})}{no(Db_{Rel})}, \tag{15}$$

Where $no(Db_{Rel})$ is the total number of relevant images in the database

F- Measure

The F-measure is an evaluation of a test's accuracy and can be determined as,

$$\text{F-measure} = 2 * \frac{(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}, \tag{16}$$

Performance analysis

The sample outcome obtained for the taken input image after applying the proposed technique with different magnification factor is displayed in Figure 5.
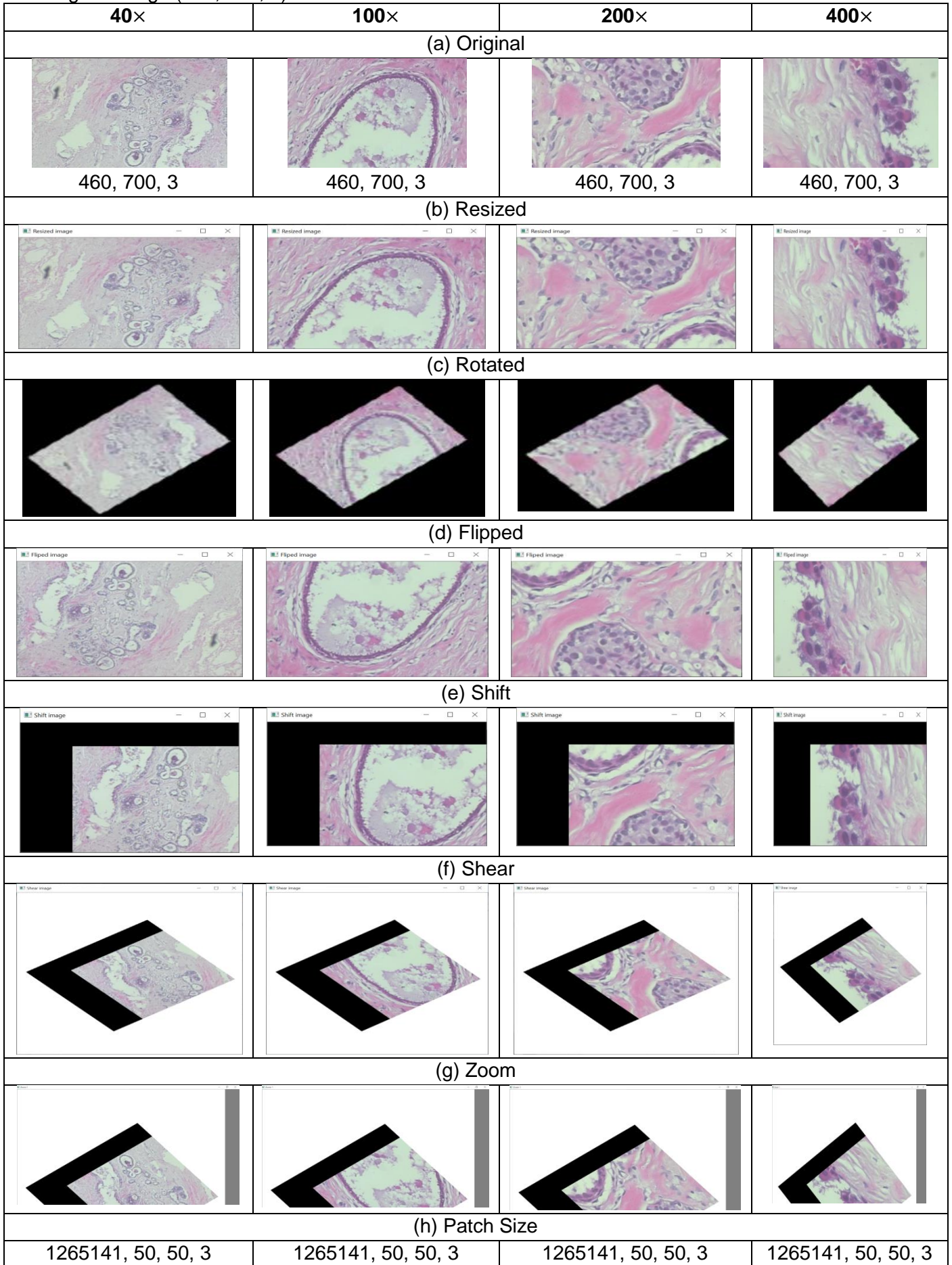
Original image (460, 700, 3)

40×	100×	200×	400×
(a) Original			
460, 700, 3	460, 700, 3	460, 700, 3	460, 700, 3
(b) Resized			
(c) Rotated			
(d) Flipped			
(e) Shift			
(f) Shear			
(g) Zoom			
(h) Patch Size			
1265141, 50, 50, 3	1265141, 50, 50, 3	1265141, 50, 50, 3	1265141, 50, 50, 3

Figure 5. Detailed structure for obtained Pre-processed images (a) Original, (b) Resized, (c) Rotated, (d) Flipped, (e) Shift, (f) Shear, (g) Zoom, and (h) Patch Size.

The result shows that data pre-processing and augmentation significantly improve the performance of the model and helps to avoid overfitting due to the raw image set. In the used dataset, out of 7909 total number of images, 80% of images have been used for the training phase and the remaining 20% images have been testing phase. From performed experiments, it is found that using a larger training set can bring negligible performance improvements. The detailed count of the images for the training and testing process is shown in Table 2:

Table 2. Types of benign and malignant class with its count

Types of classes	
Benign	**Malignant**
C1 (adenosis): 444	C1 (ductal_carcinoma): 145
C2 (fibroadenoma): 1014	C2 (lobular_carcinoma): 142
C3 (phyllodes tumor): 453	C3 (mucinous_carcinoma): 135
C4 (tubular adenoma): 569	C4 (papillary_carcinoma): 560

For experimentation, each of the classes is taken as C1 to C4 for both benign and malignant types as shown in Table 3.

Table 3. Count of predicted and actual image for different classes using proposed deep learning-based WOA.

			Predicted							
			Benign				**Malignant**			
			C1	C2	C3	C4	C1	C2	C3	C4
Actual	**Benign**	C1	70	8	1	3	6			
		C2	7	154	3	3	16	5	8	4
		C3	3		75		4		7	1
		C4		8		88		4	4	
	Malignant	C1	1	2	1	1	19	1	3	1
		C2		1	8			16	1	2
		C3		2		2	2	1	19	
		C4	2	2	8		12		10	80

Table 4. Achieved outcome for both benign and malignant classes using existing CNN classifier

			Predicted							
			Benign				**Malignant**			
			C1	C2	C3	C4	C1	C2	C3	C4
Actual	**Benign**	C1	67	6	2	4	6	1	1	
		C2	9	143	6	7	15	5	10	5
		C3	4	2	68	1	5	2	6	2
		C4	3	6	1	81	2	5	3	3
	Malignant	C1	1	3	2	2	14	2	3	1
		C2		1	7	2	1	14	1	2
		C3	1	2	2	3	1	2	13	3
		C4	3	4	7	3	11	3	9	74

Table 4, elaborates on the results obtained by the proposed deep learning-based WOA technique. After analyzing both Table 3 and Table 4, it is clear that the value obtained by the proposed methodology is better than the existing default CNN technique. For instance, considering the proposed CNN based WOA method for benign of adenosis class C1: the correctly identified image count is 70, for class C2: 8 images are wrongly

classified, for C3:1 image, and C4: 3 images have been classified. Here, 6 images obtained as C1 for class malignant. Likewise, it is calculated for the existing CNN method and it is found that in the existing technique, the correctly labeled classes count is 67 and the remaining 20 images are wrongly classified. Table 5 elaborates on the different measures like Precision, Recall, and F1-Score used for the evaluation of proposed Methodology which is based on CNN and Whale optimization algorithm for different types of images available in the dataset. Whereas, Table 6 elaborates the different measures like Precision, Recall and F1-Score for the existing Convolution Neural Network Methodology used for different types of images available in the dataset.

Table 5. Comparison of different measures for the proposed methodology

Proposed CNN based WOA			
Image type	Precision	Recall	F1-score
benign_adenosis_100X	0.02	0.09	0.04
benign_adenosis_200X	0.05	0.14	0.07
benign_adenosis_400X	0.02	0.05	0.03
benign_adenosis_40X	0.02	0.05	0.02
benign_fibroadenoma_100X	0.06	0.17	0.09
benign_fibroadenoma_200X	0.05	0.13	0.07
benign_fibroadenoma_400X	0.01	0.02	0.02
benign_fibroadenoma_40X	0.09	0.1	0.1
malignant_ductal_carcinoma_100X	0.15	0.18	0.16
malignant_ductal_carcinoma_200X	0.18	0.15	0.16
malignant_ductal_carcinoma_400X	0.24	0.26	0.25
malignant_ductal_carcinoma_40X	0.13	0.1	0.11

Table 6. Comparison of different measures for existing methodology

Existing CNN			
Image type	Precision	Recall	F1-score
benign_adenosis_100X	0.01	0.07	0.03
benign_adenosis_200X	0.04	0.11	0.06
benign_adenosis_400X	0.01	0.04	0.02
benign_adenosis_40X	0.01	0.04	0.01
benign_fibroadenoma_100X	0.05	0.14	0.07
benign_fibroadenoma_200X	0.03	0.12	0.06
benign_fibroadenoma_400X	0.00	0.01	0.01
benign_fibroadenoma_40X	0.06	0.05	0.0
malignant_ductal_carcinoma_100X	0.12	0.15	0.13
malignant_ductal_carcinoma_200X	0.17	0.13	0.14
malignant_ductal_carcinoma_400X	0.22	0.24	0.23
malignant_ductal_carcinoma_40X	0.11	0.0	0.09

In the various obtained comparative results where Figure 6 visualizes the precision parameter, Figure 7 visualizes the recall parameter, and Figure 8 visualizes the F1-score parameter for various types of images available in the dataset such as benign_adenosis_100X**,** benign_adenosis_200X, etc. and represents the comparison between proposed methodology and existing methodology. From the various obtained comparison results, this can be seen that the proposed CNN-WOA method is more reliable than any other

existing technique. When equating Precision, Recall, and F1-score, the proposed method demonstrates 92.42% accuracy in comparison to existing CNN achieves 90.34% accuracy. Therefore, from the detailed analysis, it is clear that the proposed methodology attains better performance with higher accuracy when compared to another classifier algorithm.
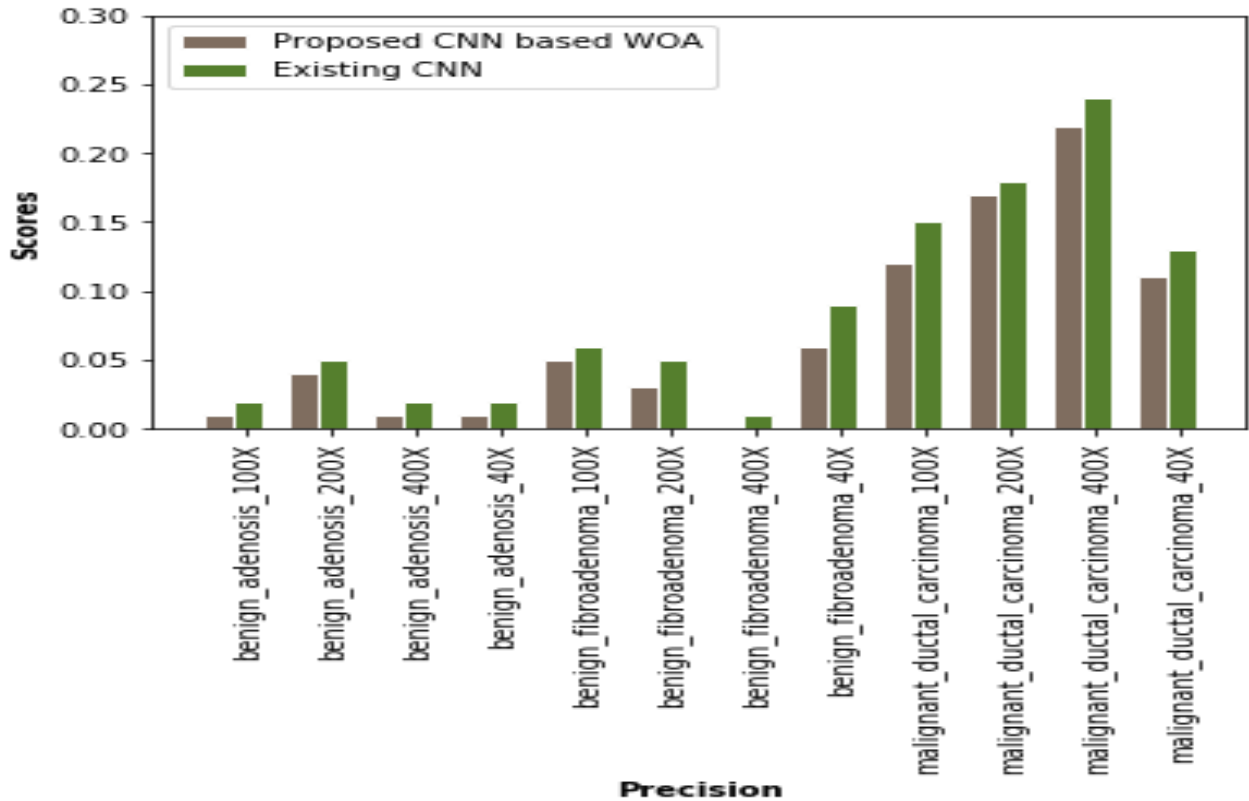
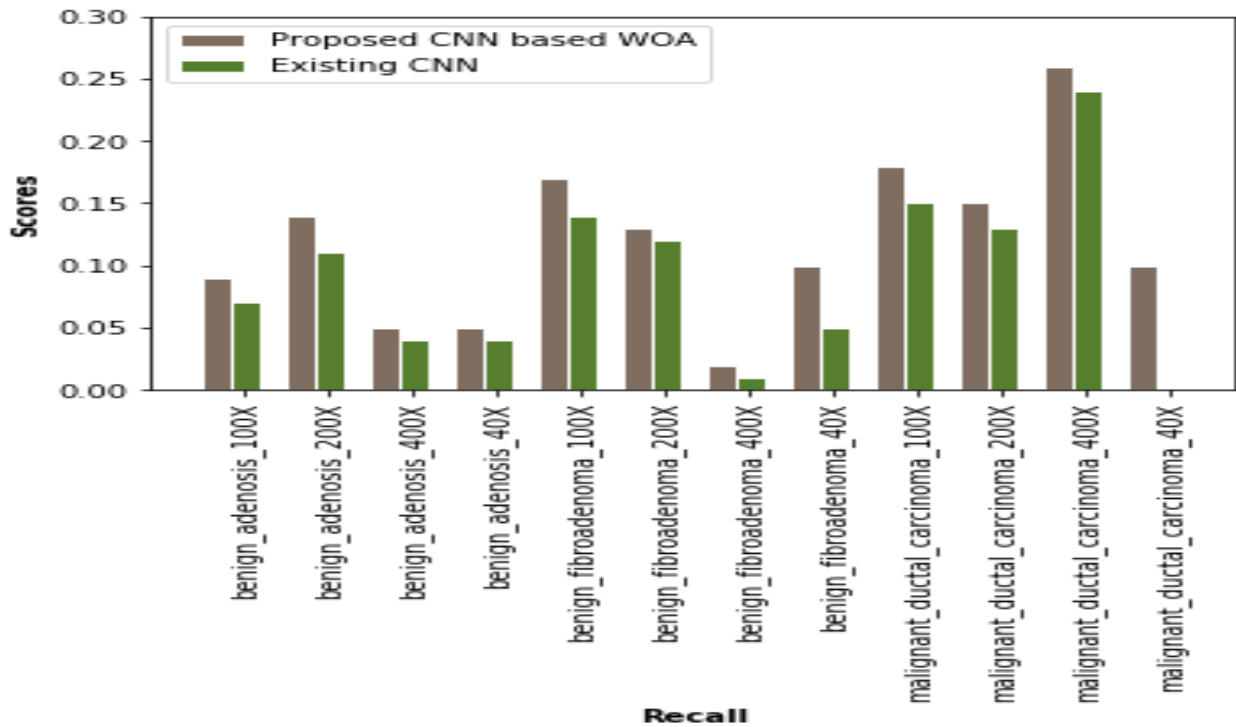

Figure 6. Comparison plot for the measure precision

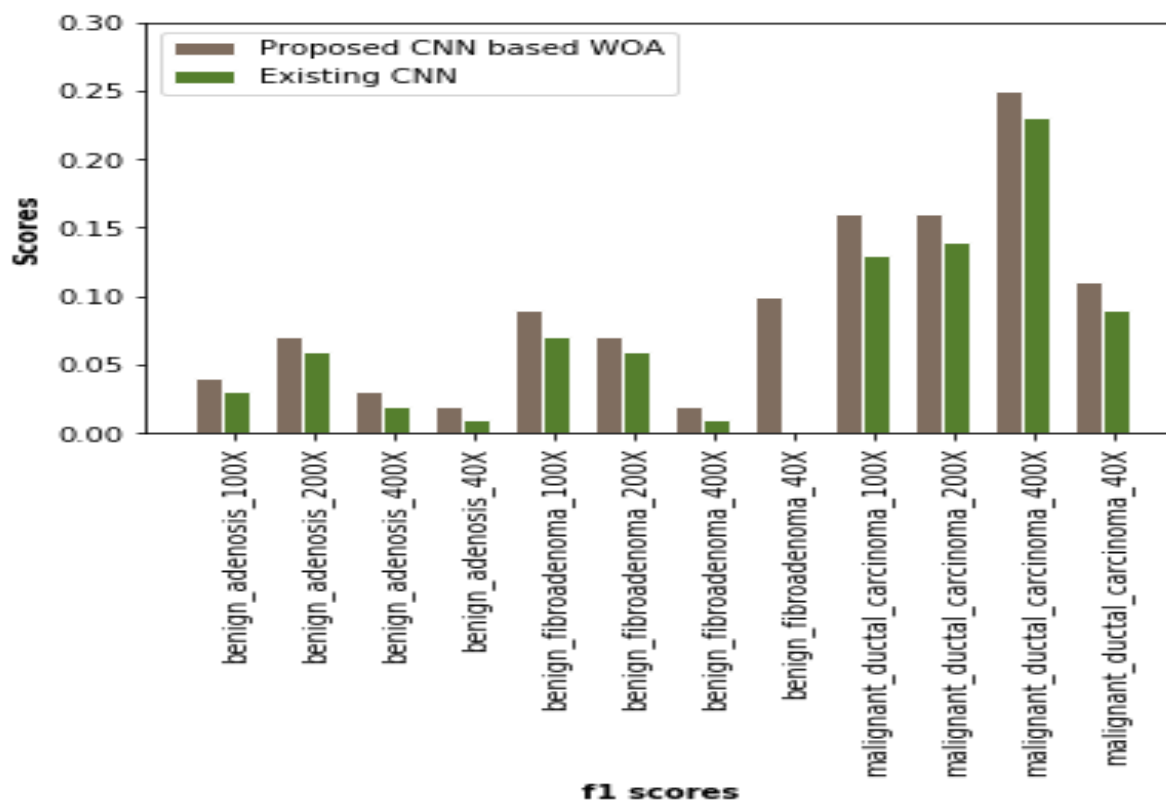**Figure 7.**Obtained recall graph for proposed and existing technique

Figure 8. Comparison plot for F1-score using the proposed and existing technique

CONCLUSION

Breast cancer detection using CNN is intended to speed up the method by assisting specialists for breast cancer diagnosis inefficient manner. The proposed model focuses on the main concern of increasing the accuracy of classifying breast cancer using histopathological images. The proposed system shows peak performance when compared with other existing approaches in terms of accuracy. The proposed model initially collects the histopathological images from the BreaKHis dataset which are fed into the pre-processing stage to avoid overfitting problems by minimizing the image size. Afterward, those pre-processed images are taken as the input to classification stage where the deep learning-based CNN classifier train those images to classify benign and malignant classes and its parameters are optimized using WOA. These procedures have helped the proposed approach correctly and effectively recognize benign and malignant images. To extend this work, a comparison between different machine learning algorithms will be done in the future.

Conflicts of Interest: "The authors declare no conflict of interest."

REFERENCES

1. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019;104(1):21–34.
2. Kanimozhi U, Ganapathy S, Manjula D, Kannan A. An Intelligent Risk Prediction System for Breast Cancer Using Fuzzy Temporal Rules. Natl Acad Sci Lett. 2019;42(3):227–32.
3. Gupta K, Janghel R. Dimensionality reduction-based breast cancer classification using machine learning. In: Advances in Intelligent Systems and Computing. 2019; 798:133-46.
4. Sangaiah I, Vincent Antony Kumar A. Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: application to breast cancer prediction. Cluster Comput. 2019;22(s3):6899–906.
5. Park JV, Park SJ, Yoo JS. Finding characteristics of exceptional breast cancer subpopulations using subgroup mining and statistical test. Expert Syst Appl. 2019;118:553–62.
6. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clin Epidemiol Glob Heal. 2019;7(3):293–9.

7. Grapin M, Coutant C, Riedinger J, Ladoire S, Brunotte F, Cochet A, et al. Combination of breast imaging parameters obtained from 18 F-FDG PET and CT scan can improve the prediction of breast-conserving surgery after neoadjuvant chemotherapy in luminal/HER2-negative breast cancer. Eur J Radiol. 2019; 113::81–88. DOI: 10.1016/j.ejrad.2019.02.005

8. Wasan RK, Morel JC, Iqbal A, Michell MJ, Rahim RR, Peacock C, et al. Can digital breast tomosynthesis accurately predict whether circumscribed masses are benign or malignant in a screening population? Clin Radiol. 2019;74(4):327.e1-327.e5.

9. Herent P, Schmauch B, Jehanno P, Dehaene O, Saillard C, Balleyguier C, et al. Detection and characterization of MRI breast lesions using deep learning. Diagn Interv Imaging. 2019;100(4):219–25.

10. Alqudah A, Alqudah M. Sliding Window Based Support Vector Machine System for Classification of Breast Cancer Using Histopathological Microscopic Images. IETE J Res. 2019 Mar 7:1–9.DOI: 10.1080/03772063.2019.1583610

11. Angara S, Robinson M, Guillén-Rondon P. Convolutional Neural Networks for Breast Cancer Histopathological Image Classification. In: 4th International Conference on Big Data and Information Analytics: Theories, Algorithms and Applications in Data Science, BigDIA 2018. 2019.

12. Xie J, Liu R, Luttrell J, Zhang C. Deep learning based analysis of histopathological images of breast cancer. Front Genet. 2019;10(FEB):1–19.

13. Kassani SH, Kassani PH, Wesolowski MJ, Schneider KA, Deters R. Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks. 2019;

14. Sudharshana PJ, Petitjean C, Spanhol F, Oliveira L, Honeine P, Sudharshana PJ, et al. Multiple Instance Learning for Histopathological Breast Cancer Images To cite this version : HAL Id : hal-01965039 Multiple Instance Learning for Histopathological Breast Cancer Images. 2019;

15. Tseng YJ, Huang CE, Wen CN, Lai PY, Wu MH, Sun YC, et al. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. Int J Med Inform. 2019;128(5):79–86.

16. Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, et al. BACH: Grand challenge on breast cancer histology images. Med Image Anal. 2019;56:122–39.

17. Ouyang F sheng, Guo B liang, Huang X yi, Ouyang L zhu, Zhou C ru, Zhang R, et al. A nomogram for individual prediction of vascular invasion in primary breast cancer. Eur J Radiol. 2019;110(1):30–8.

18. Atrey K, Sharma Y, Bodhey NK, Singh BK. Breast cancer prediction using dominance-based feature filtering approach: A comparative investigation in machine learning archetype. Brazilian Arch Biol Technol. 2019;62:1–15.

19. Zhou LQ, Wu XL, Huang SY, Wu GG, Ye HR, Wei Q, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. Radiology. 2020;

20. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep. 2019;