

# Breast cancer recognition based on performance evaluation of machine learning algorithms

Chiman Haydar Salh<sup>1</sup>, Abbas M. Ali<sup>2</sup>

<sup>1</sup>Department of Information System Engineering, Erbil Technical Engineering College,  
Erbil Polytechnic University, Erbil, Iraq

<sup>2</sup>Department of Software Engineering and Informatics, College of Engineering, Salahaddin University-Erbil, Erbil, Iraq

## Article Info

### Article history:

Received Jan 23, 2022

Revised May 21, 2022

Accepted Jun 8, 2022

### Keywords:

Artificial neural network

Bag of words

Decision tree

K-nearest neighbor

Local binary pattern

Random forests

Support vector machine

## ABSTRACT

Breast cancer is the one common cause of death in both developed worlds and the most death-causing disease diagnosed among women. Early recognition of this condition can help to minimize death rates. The breast problem statement, in brief, is not reliable for accuracy recognition. They have a high degree of classification accuracy as well as diagnostic capabilities. The most common classifications are normal, benign cancer, and malignant cancer. Machine learning (ML) techniques are now widely used in the classification of breast cancer. In this paper, some machine learning technics have been investigated to diagnose breast cancer (BC) on magnetic resonance imaging (MRI) images using multi-step processes. The first step has been to take the MRI image as an input image and have been pre-processing an image, then use feature extraction by using (scale-invariant feature transform (SIFT), histogram of oriented gradient (HOG), local binary patterns (LBP), bag of words (BoW), and edge-oriented histogram (EOH)). Next step we implement the classifying algorithms (KNN, decision tree (DT), naïve Bayes, ANN, SVM, RF, AdaBoost), have been used to detect and classify the normal or breast cancer region for this purpose datasets like ACRIN-Contralateral-Breast-MRI, In and breast cancer MRI dataset) has been collected our breast cancer MRI images from Erbil and Sulaymaniyah hospital the results was 91.9%, the result of ACRIN was 97% and the results Breast Cancer was 92.3%.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Chiman Haydar Salh

Department of Information Communication Technology, Erbil Technology College

Erbil Polytechnic University

Erbil, Iraq

Email: chiman.salh@epu.edu.iq

## 1. INTRODUCTION

Breast cancer is the most common cancer among women around the world. Breast cancer is becoming more common, according to studies conducted in advanced countries. Breast cancer is more common in many Western nations than in poorer regions, according to researchers in Africa and Asia [1]. Although the causes of breast cancer are unknown, early detection is crucial for successful treatment and reduced mortality. Magnetic resonance imaging (MRI) technology has been frequently utilized for the screening and diagnosis of breast illnesses because of its low cost, mobility, noninvasiveness, and real-time imaging [2]. The problem of automated diagnosis of breast cancer till now, is not accurate enough to be used as a reliable way for next the steps of removing the dangerous area (tumor) to be not distributed in the human body [3].

Accurate breast cancer behavior prediction is critical because it supports clinicians in their decision-making process, allowing for more customized therapy for patients and higher recovery possibilities [4]. The underlying characteristics that contribute to early breast cancer recurrence remain a top research concern for doctors and data scientists alike [5]. A range of machine learning algorithms and statistical techniques have improved breast cancer diagnosis and prediction in many research studies on cancer recurrence [4], [5]. These algorithms represented try two situations, the first was machine learning methods k-nearest neighbor (k-NN), and artificial neural network (ANN) [6], and the second is deep learning models like convolutional neural network (CNN), and recurrent neural networks (RNN) [7]. The contribution of our study is most crucial thing is to multi-check the authenticity and accuracy of these impacts because you'll always find something in countless sets of data, if you consider machine learning as a single concept, this is also one of the disadvantages. The goal of this research work is to investigate some machine learning algorithms to predict and diagnose breast cancer, using machine-learning algorithms, and find out which ones are the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision, and sensitivity. The remainder of this work is arranged in the following manner. Section 2 introduces the methodology and findings of earlier breast cancer diagnosis studies. The recommended methodology for our research is described in Section 3. The experiment's results are presented and explained in depth in Section 4. Section 5 concludes the paper.

**2. RELATED WORK**

Many studies have applied machine learning (ML) approaches for breast cancer diagnosis to improve classification accuracy and speed, however, these studies have yet to produce a trustworthy system on which they can rely. Some of the earlier related efforts on breast cancer diagnosis by researchers employing various machine learning methodologies are addressed in this part. Machine learning algorithms are available for breast cancer prediction and diagnosis as shown in Table 1.

Table 1. Comparison with the existing work for breast cancer diagnosis

| Authors                       | Classifiers  | Accuracy (%)          |
|-------------------------------|--|-----------------------|
| Omondiagbe <i>et al.</i> [6]  | Support Vector Machine   | 98.82%                |
| Aishwarja <i>et al.</i> [7]   | SVM, KNN, Naive Bayes, Random forest                           | 98%, 97%, 97%, 96%    |
| Bharat <i>et al.</i> [8]      | naïve Bayes, J48, RBF networks                                 | 97.3%, 96.77%, 93.41% |
| Shravya <i>et al.</i> [9]     | Support Vector Machine   | 92.7%                 |
| Islam <i>et al.</i> [10]      | Support Vector Machine, K-Nearest Neighbors                    | 98.57%, 97.14%        |
| Golagani <i>et al.</i> [11]   | Support Vector Machine   | 99.2%                 |
| MurtiRawat <i>et al.</i> [12] | K-Nearest Neighbors,<br>Logistic Regression Ensemble Learning. | 98.60%<br>97.90%      |

**3. EXISTING METHOD OF BREAST CANCER RECOGNITION**

**3.1. MRI and breast cancer**

Because resonances are the most appealing contrasting alternative for mammography, MRI is an intriguing approach. Furthermore, by identifying the stage of the disease, MRI can assist radiologists and other experts in deciding how to treat breast growth patients as shown in the Figure 1. Exceptionally effective in depicting a large number of breast surgeries or radiation treatments [13].

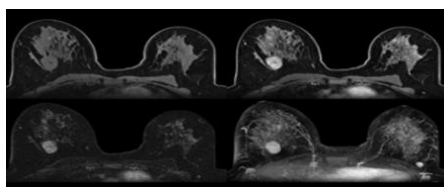


Figure 1. Breast cancer MRI image

**3.2. Feature extraction**

Feature extraction is critical for accurate breast cancer MRI image classification. The feature of the images is compounded in the classification approach based on some specific rules. The fundamental goal of feature extraction is to condense the data while still retaining the majority of the important information. Different types of feature extraction techniques are listed next sub-section.

### 3.2.1. Scale-invariant feature transform (SIFT)

David Lowe created scale-invariant feature transform, a method for extracting distinctive visual features consist from four stages. The first is the identification of the image's representative key point positions. The second step entails calculating picture descriptors, or features, that are positioned at crucial places. The third step is to assign orientations, and the fourth is to fine-tune the location of crucial points. As a result, the approach defines an image as a set of feature vectors that are not affected by image translation, scaling, or rotation. The extraction of SIFT is mainly characterized by two parameters: the peak threshold and the edge threshold [14].

$$F(x/\mu, \alpha) = \frac{\exp\{\frac{x-\mu}{\alpha}\}}{\alpha(1+\exp\{\frac{x-\mu}{\sigma}\})} \quad (1)$$

$$F(x, h) = \frac{1}{nh} \sum_{i=1}^n K\{\frac{x-x_i}{h}\} \quad (2)$$

### 3.2.2. Histogram of oriented gradients (HOG)

One of the most essential parameters in the detection of medical images is texture. It is critical for classification, detection, and segmentation based on intensity and colour in computer image analysis. The histogram of oriented gradient was employed as a feature extraction tool in this study. The computer vision community has effectively used these properties for object detection and localization. The HOG technique is based on the notion that the histogram of local intensity gradients or edge directions in a dense grid may determine the bulk local appearance and shape [15].

### 3.2.3. Edge oriented histogram (EOH)

The computation of edge-oriented histogram, which uses edge pixels around each feature point to generate descriptors, is the first step in the development of these histograms. EOH, on the other hand, is sensitive to fine textures. Short edges, for example, can be recovered from visible images, but the same features cannot be extracted from infrared images' comparable regions. The equation is used to find an edge [16], [17].

$$m_{i,j} = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \quad (3)$$

$$G_x = \text{Sob1}(I_{\text{ROI}}), G_y = \text{Sob1}(I_{\text{ROI}})$$

### 3.2.4. Local binary patterns (LBP)

Local binary patterns is a texture feature extraction method that is commonly utilized in recognition algorithms. The retrieved features can be used to classify breast cancer abnormalities in MRI images [18]. The LBP code for a pixel in a picture is calculated by comparing it to its neighbors. P (number of neighbors) and R (relative distance) are two user-defined parameters for LBP (radius of comparisons) Calculate the sign parameter for the neighborhood pixel using as calculated in (4) [19].

$$LBP_{P,R} = \sum_{p=0}^{P-1} S(g_p - g_c) 2^p, S(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (4)$$

### 3.2.5. Bag of words (BoW)

A k-means clustering technique was utilized to create dictionaries based on the training data in this method. These dictionaries are then used to determine BoW, which is the image's final representation. The Bag of words (BoW) feature coding approach is commonly used for medical and natural picture classification. To create the bag of words model, we must first create a visual vocabulary (a codebook) from the retrieved local descriptors by clustering them. Using K-means, the descriptor sets recovered from the training set are clustered into K-clusters [20], [21].

## 3.3. Machine learning (ML) algorithms

The classification of malignant and benign tumor cells was done using machine learning methods in this study. The parametric examination of seven different machine learning algorithms is included in this research. The following is a brief summary of the methods used in this paper [22].

### 3.3.1. k-nearest neighbor (k-NN)

The non-parametric lazy algorithm is k-nearest neighbor. The nearest neighbors are chosen based on the Euclidean distance between the x and y vectors as calculated in (5). The k-NN outcome varies depending on the value of K. A large value of K will result in class overlap, whereas a smaller value of K will result in faster computations [23].

$$Eucliden\ Distance = \sqrt{\sum_{i=1}^k (X_i - Y_i)^2} \tag{5}$$

**3.3.2. Decision tree (DT)**

A decision tree is a model of decisions and their probable outcomes that looks like a tree. The fundamental algorithm of a DT is called iterative dichotomiser (ID3), which constructs the decision tree using the entropy or information gain of each attribute. The max-depth, min-samples-leaf, and max-leaf-nodes parameters of a decision tree are employed here for tuning [24].

$$E(S) = \sum_{i=1}^c -P_i \log_2 P_i \tag{6}$$

**3.2.3. Random forests (RF)**

Overfitting difficulties in decision trees are solved with random forests. The random forests are collections of trees, with a majority vote deciding the outcome, as shown in Figure 2 [25], [26];

$$MSE = \frac{1}{N} \sum_{i=1}^N -(f_i - y_i)^2 \tag{7}$$

where N is the number of data points, fi is the value returned by the model and (yi) is the actual value for the data point.

**3.2.4. Artificial neural network (ANN)**

An artificial neural network is a biologically driven computational mechanism used for a variety of tasks including pattern recognition, output prediction, clustering, and optimization. The ANN is a popular model of neural networks that offers various advantages, one hidden layer and an output layer as shown in Figure 3 [27].

$$Z = Bias + W_1X_1 + W_2X_2 + \dots + W_nX_n \tag{8}$$

where, Z is the symbol for denotation of the above graphical representation of ANN. W is, are the weights or the beta coefficients, X is, are the independent variables or the inputs, and Bias or intercept=W<sub>0</sub>.

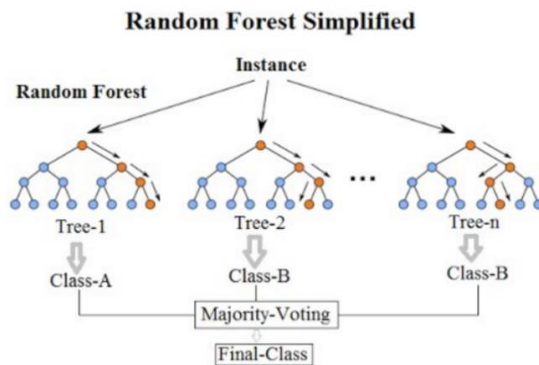


Figure 2. random forests [19]

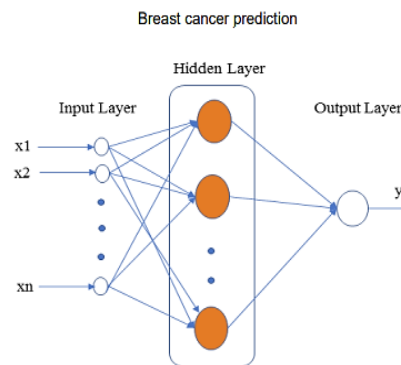


Figure 3. ANN architecture [20]

**3.2.5. Support vector machine (SVM)**

Support vector machines are machine learning algorithms that are based on restricted minimization issues. The dot products of support vectors and the objects must be determined to determine the maximum separation distance between objects. A radial basis kernel, as indicated in (9), is the kernel trick used in this study.

$$K(X_1, X_2) = \exp(-\gamma ||X_1 - X_2||^2) \tag{9}$$

where, K(X<sub>1</sub>,X<sub>2</sub>) is the radial bias equation for points or regions (X<sub>1</sub>,X<sub>2</sub>) and γ is spread of kernel. A low value of γ leads to low decision boundary whereas high values of this parameter give higher decision boundaries [23].

**3.2.6. Naïve Bayes**

The supervised learning algorithm Naïve Bayes classifier is employed for classification. It's based on the Bayes theorem, which involves calculating the likelihood of an event after it's already happened. It is one of the most basic yet powerful machine learning algorithms in use, with applications in a variety of industries. The 'zero-frequency problem' is when an algorithm assigns zero probability to a categorical variable whose category in the test data set was not present in the training dataset [28], [29].

$$P(A/B) = \frac{P(\frac{A}{B})P(A)}{P(B)} \tag{10}$$

where P(A/B) is the posterior probability, P(A) and P(B) are probabilities of the occurrence of events A and B respectively and P(B/A) is the likelihood.

**3.2.7. AdaBoost**

Using regression and classification, this technique is used to predict the existence of breast cancer. It turns weak learners. It obtains the node's weight and adjusts it until proper results are obtained. Despite this, it is vulnerable to feature quality and noise [30].

$$H(x) = sign(\sum_{i=1}^T \alpha_t h_t(x)) \tag{11}$$

**4. PROPOSED SYSTEM**

Analyzing and diagnosing breast cancer diseases based on MRI images needs some common styles to make this process get success some of these stages are, segmentation of MRI images and then extraction features to be analyzed later on. Our proposed work gives the recognizable proof and division of breast tumors from the MRI and image by using four-phase shapes. Starting stride has been taking the test image as an MRI image and after that, image upgrade (image change). Next, Segmentation to black and white. The third step is about feature extraction for the image the methods used in this part it is (SIFT, HOG, LBP, BoW and EOH) Fourth step and finally implement the classifying machine learning algorithms (KNN, decision tree, naïve Bayes, ANN, SVM, RF, AdaBoost) our proposed work result shows the normal and benign and malignant region location as shown in Figure 4.

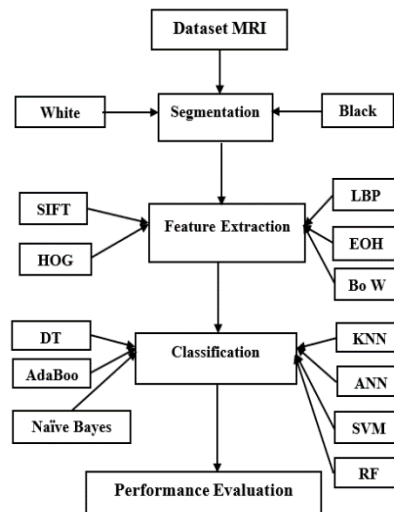


Figure 4. Proposed system diagram

**4.1. Experimental setup**

Three datasets have been used in the proposed study two of them were taken from the internet. The ACRIN first dataset consists of 984 patients but only 969 were included in the primary data analysis due to study criteria. This paper has been containing 1280 images that belong to 128 patients and 10 images for each patient. The breast cancer MRI second dataset includes 400 images from 40 patients, 10 images from each

patient in which 35 of them have cancer and the other 5 patients have only tumor. In this data set, all the patients have cancer. These two data sets were taken between 2014 to 2020, and the ages of the patients are between 25 to 65 years old. The third data set includes 920 images from 92 patients, 10 images for each patient, of which 73 of them have cancer and 19 of them only have lumps. Their ages are between 25 to 70 years old have been collected breast cancer images from Erbil and Sulaymaniyah hospital. The data in all datasets have been resized in (RadiAnt) application, which is a medical program used in all the hospitals. The size of each image is 512 by 512, which is regarded as the best size to show cancer images according to the other rates that have been studied. In addition, two other tools have been used. The first one is (MATLAB 2021b), which is used for the results and the features. The second one is (Weka) which is used for the classifications.

**4.2. Result**

The proposed approach is tested on real MRI images of breast cancer. Volume MRI images of patients who are benign, malignant, or normal make up the three data sets. Five methods for feature extraction and seven algorithms for classifications were tested on the three datasets with preprocessing approaches for resizing images, and the best result was five methods for feature extraction and seven algorithms for classifications. Among them, the best result for three datasets was recorded for method HOG and using ANN for classification: 94% in the Erbil and Sulaymaniyah breast cancer dataset, for second dataset breast cancer MRI method BoW and using ANN for classification: 96% and Naïve Bayes 97% and for the third dataset ACRI-Contralateral-Breast-MRI method BoW and using Adaboost for classification: 99% and method HOG and using ANN for classification: 98%. The Performance of the methods and classifiers are improved and enhanced as shown in Table 2 and Figure 5 (Appendix 1).

Table 2. Classification results for feature set diagnosis

| No. of datasets                   | No. of Method | KNN    | ANN    | SVM    | Decision Tree | Naïve Bayes | Random Forest | AdaBoost |
|-----------------------------------|---------------|--------|--------|--------|---------------|-------------|---------------|----------|
| Erbil and Sulaymaniyah Breast MRI | HOG           | 94.13% | 94.13% | 88.7%  | 87.93%        | 53%         | 91.9%         | 88.8%    |
|                                   | EOH           | 85.2%  | 89.2%  | 88.4%  | 86.52%        | 84.78%      | 88.3%         | 86.96%   |
| Breast Cancer MRI                 | SIFT          | 88.9%  | 85.87% | 88.4%  | 87.83%        | 20%         | 88.04%        | 88.04%   |
|                                   | LBP           | 88.18% | 88%    | 88%    | 87.18%        | 72.6%       | 88.59%        | 88%      |
|                                   | Bo W          | 89.1%  | 90%    | 89.9%  | 92%           | 91%         | 89.1%         | 92.3%    |
|                                   | HOG           | 92.3%  | 89%    | 87.3%  | 86.75%        | 56%         | 88.75%        | 87.5%    |
|                                   | EOH           | 85.5%  | 87%    | 87.5%  | 87%           | 78%         | 87.25%        | 87.5%    |
| ACRI-Contralateral-Breast-MRI     | SIFT          | 89.25% | 82.75% | 87.25% | 87%           | 15%         | 86.75%        | 87.25%   |
|                                   | LBP           | 85.25% | 85.75% | 87.5%  | 87%           | 71%         | 88.25         | 86%      |
|                                   | Bo W          | 95%    | 96%    | 95%    | 97%           | 97%         | 95%           | 99%      |
|                                   | HOG           | 97.89% | 98.20% | 93.67% | 93.59%        | 67.7%       | 95.3%         | 93.67%   |
|                                   | EOH           | 90.78% | 93.36% | 93.75% | 93.75%        | 93.75%      | 93.59%        | 93.75%   |
| Breast-MRI                        | SIFT          | 89.38% | 93.67% | 93.75% | 93.59%        | 81.41%      | 93.75%        | 93.75%   |
|                                   | LBP           | 91.32% | 93.81% | 93.75% | 93.28%        | 87.96%      | 93.82%        | 93.75%   |
|                                   | Bo W          | 96%    | 97%    | 96%    | 98%           | 98%         | 96%           | 99%      |

In Figure 5 (see in Appendix) explained overall comparison of three datasets (a), (b), (c) evaluation between five methods and by using seven algorithms in machine learning the feature extraction accuracy by the percentage of five different sample groups of MRI breast cancer images. White gray indicates the accuracy of HOG, gray indicates the accuracy of the EOH, clear gray indicates the accuracy of the SIFT, black indicates the accuracy of the LBP, and blue gray indicates the accuracy of the BoW proposed solution (color figure) and they have been used seven algorithms for classification as shown in the Figure 5.

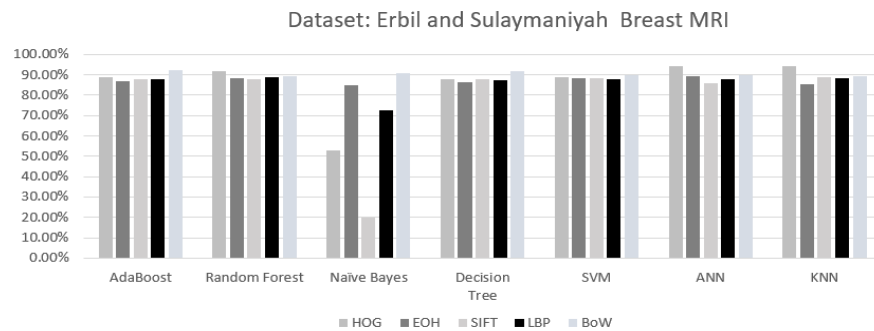
In Tables 3, 4 and 5 in (see in Appendix) the sets are divided into two parts (train and test) parts. There are three methods for dividing the date set. Here, one of the methods is used to show the results, as in the following. The results are better or (clear) in the table.

**5. CONCLUSION**

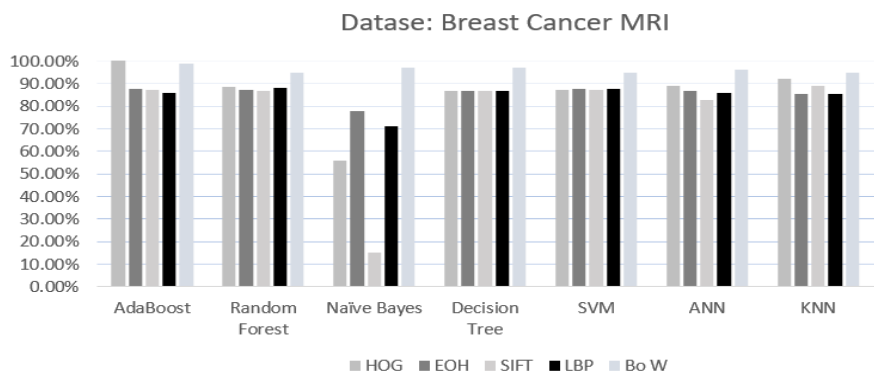
This research work reveals that feature selection and feature extraction by using four methods it is (Sift, HOG, LBP, BoW and EOH) can help improve the diagnosis of benign and malignant tumors using machine learning techniques and in this study, have been employed seven main algorithms: SVM, NB, k-NN, RF, AdaBoost, ANN and decision tree on the three dataset Breast Cancer (original) datasets. Has been tried to compare the efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. Using the bag of words model for breast cancer images proved to provide the best results in classification tasks with SIFT descriptors HOG methods. After an accurate comparison between our models, we found that ANN, SVM and AdaBoost achieved a higher

efficiency of 96%, Precision of 97% and outperforms all others. These models have demonstrated their efficiency in Breast Cancer prediction and diagnosis and achieved the best performance in terms of accuracy and precision. It should be noted that all the results obtained based on the three datasets, it can be considered as a limitation of our work, it is, therefore, necessary to reflect for future works to apply these same algorithms and methods to other databases to confirm the results obtained via this database, as well as, in our future works, has been plan to apply our and other deep learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy.

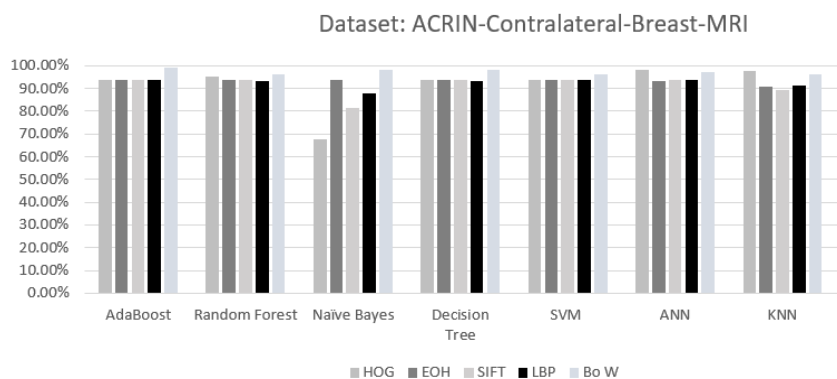
APPENDIX



(a)



(b)



(c)

Figure 5. Overall comparison of three datasets (a) analyses the result of first dataset, (b) analyses the result of second dataset, and (c) analyses the result of the third dataset

Table 3. Accuracy percentage for breast cancer diagnostic (Erbil and Sulaymaniyah dataset) MRI images

| Method | Classifier    | Test Accuracy | Train Accuracy | Sensitivity | f- measure | Precision | Recall |
|--------|---------------|---------------|----------------|-------------|------------|-----------|--------|
| HOG    | KNN           | 92.84%        | 95.65%         | 95.6        | 0.958      | 0.960     | 0.957  |
|        | ANN           | 95.44%        | 95.86%         | 95.9        | 0.959      | 0.960     | 0.959  |
|        | Decision Tree | 88.06%        | 91.08%         | 91.1        | 0.903      | 0.901     | 0.911  |
|        | Naïve Bayes   | 64%           | 72%            | 77.2        | 0.769      | 0.884     | 0.772  |
|        | Random Forest | 90.88%        | 93.69%         | 93.7        | 0.926      | 0.937     | 0.937  |
|        | AdaBoost      | 87.63%        | 91.95%         | 92.0        | 0.911      | 0.912     | 0.920  |
| EOH    | SVM           | 88.06%        | 91.30%         | 91.3        | 0.890      | 0.921     | 0.913  |
|        | KNN           | 86.05%        | 86.5%          | 86.5        | 0.865      | 0.865     | 0.865  |
|        | ANN           | 87.60%        | 88.23%         | 88.2        | 0.854      | 0.852     | 0.882  |
|        | Decision Tree | 87.36%        | 87.60%         | 87.2        | 0.833      | 0.824     | 0.876  |
|        | Naïve Bayes   | 83%           | 84.53%         | 84.5        | 0.829      | 0.817     | 0.845  |
|        | Random Forest | 87.36%        | 88.04%         | 88.00       | 0.835      | 0.824     | 0.880  |
| Sift   | AdaBoost      | 88%           | 88.6%          | 80.7        | 0.840      | 0.900     | 0.807  |
|        | SVM           | 88.0%         | 88.0%          | 88.00       | 0.936      | 0.880     | 0.880  |
|        | KNN           | 88.69%        | 90.86%         | 90.9        | 0.905      | 0.902     | 0.909  |
|        | ANN           | 83.2%         | 87.8%          | 87.8        | 0.882      | 0.868     | 0.878  |
|        | Decision Tree | 87.82%        | 87.82%         | 87.8        | 0.823      | 0.816     | 0.878  |
|        | Naïve Bayes   | 16.7%         | 21%            | 92.5        | 0.898      | 0.873     | 0.925  |
| LBP    | Random Forest | 88.04%        | 88.04%         | 88.7        | 0.818      | 0.880     | 0.887  |
|        | AdaBoost      | 86.95%        | 88.04%         | 88.00       | 0.819      | 0.880     | 0.880  |
|        | SVM           | 87.03%        | 88.60%         | 88.00       | 0.825      | 0.836     | 0.880  |
|        | KNN           | 86.3%         | 93.4%          | 96.0        | 0.963      | 0.956     | 0.960  |
|        | ANN           | 84.7%         | 92.6%          | 96.5        | 0.953      | 0.951     | 0.965  |
|        | Decision Tree | 87.3%         | 90.2%          | 95.8        | 0.949      | 0.933     | 0.958  |
| BoW    | Naïve Bayes   | 65%           | 83%            | 92.4        | 0.908      | 0.893     | 0.924  |
|        | Random Forest | 88.5%         | 93%            | 98.5        | 0.961      | 0.939     | 0.985  |
|        | AdaBoost      | 87.3%         | 88.9%          | 98.8        | 0.940      | 0.897     | 0.988  |
|        | SVM           | 87.6%         | 88.2%          | 1.00        | 0.938      | 0.883     | 1.000  |
|        | KNN           | 87%           | 100%           | 98.5        | 1          | 0.995     | 0.985  |
|        | ANN           | 88.5          | 90%            | 1           | 0.941      | 0.889     | 1      |
|        | Decision Tree | 92.5          | 100%           | 99.5        | 1          | 0.949     | 0.995  |
|        | Naïve Bayes   | 91.5          | 99%            | 93.5        | 0.999      | 0.970     | 0.935  |
|        | Random Forest | 86.6%         | 100%           | 99.8        | 1          | 0.996     | 0.998  |
|        | AdaBoost      | 91%           | 100%           | 99.8        | 1          | 0.999     | 0.998  |
|        | SVM           | 88.5          | 90%            | 1           | 0.941      | 0.881     | 1      |

Table 4. Accuracy percentage for breast cancer diagnostic dataset (breast cancer MRI) images

| Method | Classifier    | Test Accuracy | Train Accuracy | Sensitivity | f- measure | Precision | Recall |
|--------|---------------|---------------|----------------|-------------|------------|-----------|--------|
| HOG    | KNN           | 93%           | 94%            | 94.00       | 0.942      | 0.945     | 0.940  |
|        | ANN           | 89%           | 90%            | 90.00       | 0.903      | 0.907     | 0.900  |
|        | Decision Tree | 87.5%         | 88%            | 88.00       | 0.832      | 0.895     | 0.880  |
|        | Naïve Bayes   | 58%           | 66%            | 66.00       | 0.714      | 0.805     | 0.660  |
|        | Random Forest | 89%           | 93%            | 93.00       | 0.923      | 0.927     | 0.930  |
|        | AdaBoost      | 87.5%         | 88.5%          | 88.5        | 0.877      | 0.873     | 0.885  |
| EOH    | SVM           | 87.5%         | 89.5%          | 89.5        | 0.862      | 0.906     | 0.895  |
|        | KNN           | 89.94%        | 92.03%         | 92.00       | 0.922      | 0.923     | 0.920  |
|        | ANN           | 88.44%        | 91.04%         | 91.00       | 0.890      | 0.909     | 0.910  |
|        | Decision Tree | 86.56%        | 90.45%         | 90.5        | 0.904      | 0.903     | 0.905  |
|        | Naïve Bayes   | 74.12%        | 85%            | 85.4        | 0.777      | 0.872     | 0.854  |
|        | Random Forest | 90.04%        | 91.95%         | 92.00       | 0.913      | 0.913     | 0.920  |
| Sift   | AdaBoost      | 87.06%        | 87.93%         | 87.9        | 0.875      | 0.871     | 0.879  |
|        | SVM           | 87.43%        | 87.53%         | 89.6        | 0.934      | 0.878     | 0.876  |
|        | KNN           | 91.5%         | 96.5%          | 96.5        | 0.980      | 0.965     | 0.965  |
|        | ANN           | 83.5%         | 85.5%          | 85.5        | 0.854      | 0.853     | 0.855  |
|        | Decision Tree | 87.5%         | 87.5%          | 87.5        | 0.833      | 0.835     | 0.875  |
|        | Naïve Bayes   | 11%           | 13%            | 81.1        | 0.756      | 0.708     | 0.811  |
| LBP    | Random Forest | 87%           | 88.5%          | 88.5        | 0.840      | 0.898     | 0.885  |
|        | AdaBoost      | 85.5%         | 86%            | 86.00       | 0.823      | 0.816     | 0.860  |
|        | SVM           | 86%           | 88.5%          | 88.5        | 0.868      | 0.866     | 0.885  |
|        | KNN           | 82%           | 91%            | 94.3        | 0.949      | 0.954     | 0.943  |
|        | ANN           | 8.5%          | 91%            | 98.3        | 0.951      | 0.920     | 0.983  |
|        | Decision Tree | 84.5%         | 87.5%          | 1.00        | 0.933      | 0.874     | 1.000  |
| BoW    | Naïve Bayes   | 65%           | 88%            | 93.8        | 0.931      | 0.927     | 0.938  |
|        | Random Forest | 86%           | 92%            | 98.9        | 0.956      | 0.926     | 0.989  |
|        | AdaBoost      | 87%           | 90%            | 97.7        | 0.945      | 0.915     | 0.977  |
|        | SVM           | 87%           | 88%            | 1.000       | 0.936      | 0.880     | 1.000  |
|        | KNN           | 82%           | 91%            | 94.3        | 0.949      | 0.954     | 0.943  |
|        | ANN           | 8.5%          | 91%            | 98.3        | 0.951      | 0.920     | 0.983  |
|        | Decision Tree | 84.5%         | 87.5%          | 1.00        | 0.933      | 0.874     | 1.000  |
|        | Naïve Bayes   | 65%           | 88%            | 93.8        | 0.931      | 0.927     | 0.938  |
|        | Random Forest | 86%           | 92%            | 98.9        | 0.956      | 0.926     | 0.989  |
|        | AdaBoost      | 87%           | 90%            | 97.7        | 0.945      | 0.915     | 0.977  |
|        | SVM           | 87%           | 88%            | 1.000       | 0.936      | 0.880     | 1.000  |



Table 5. Accuracy percentage for breast cancer diagnostic (ACRIN contralateral-breast-MRI) images

| Method | Classifier    | Test Accuracy | Train Accuracy | Sensitivity | f-measure | Precision | Recall |
|--------|---------------|---------------|----------------|-------------|-----------|-----------|--------|
| HOG    | KNN           | 98.1%         | 98.90%         | 99.8        | 0.989     | 0.990     | 0.998  |
|        | ANN           | 97.5%         | 97.8%          | 97.8        | 0.974     | 0.974     | 0.975  |
|        | Decision Tree | 93.31%        | 95.42%         | 95.00       | 0.921     | 0.949     | 0.950  |
|        | Naïve Bayes   | 72.96%        | 84.68%         | 95.00       | 0.881     | 0.945     | 0.847  |
|        | Random Forest | 94%           | 97%            | 97.00       | 0.966     | 0.971     | 0.970  |
|        | AdaBoost      | 93.53%        | 94.59%         | 94.5        | 0.927     | 0.941     | 0.945  |
| EOH    | SVM           | 93.3          | 95.5%          | 95.3        | 0.941     | 0.951     | 0.953  |
|        | KNN           | 89%           | 94%            | 94.1        | 0.940     | 0.939     | 0.941  |
|        | ANN           | 92.15%        | 95.80%         | 92.8        | 0.902     | 0.878     | 0.928  |
|        | Decision Tree | 93.28%        | 93.42%         | 93.8        | 0.906     | 0.879     | 0.934  |
|        | Naïve Bayes   | 91%           | 93.74%         | 93.7        | 0.907     | 0.903     | 0.937  |
|        | Random Forest | 93.74%        | 94.37%         | 93.7        | 0.934     | 0.938     | 0.944  |
| Sift   | AdaBoost      | 93.5%         | 93.78%         | 100         | 0.906     | 0.938     | 1.000  |
|        | SVM           | 93.75%        | 93.75%         | 93.7        | 0.968     | 0.938     | 0.937  |
|        | KNN           | 88.90%        | 90.87%         | 90.8        | 0.909     | 0.925     | 0.908  |
|        | ANN           | 92.84%        | 94.65%         | 94.8        | 0.947     | 0.946     | 0.948  |
|        | Decision Tree | 93.59%        | 93.59%         | 93.6        | 0.912     | 0.879     | 0.936  |
|        | Naïve Bayes   | 75.62%        | 88.37%         | 91.7        | 0.931     | 0.955     | 0.917  |
| LBP    | Random Forest | 93.7%         | 94%            | 94.1        | 0.915     | 0.944     | 0.941  |
|        | AdaBoost      | 93.75%        | 93.90%         | 93.9        | 0.914     | 0.923     | 0.939  |
|        | SVM           | 93.78%        | 95.59%         | 95.8        | 0.951     | 0.954     | 0.958  |
|        | KNN           | 93%           | 94.2%          | 96.5        | 0.969     | 0.973     | 0.965  |
|        | ANN           | 92.3%         | 95.4%          | 99.9        | 0.976     | 0.961     | 0.999  |
|        | Decision Tree | 93.9%         | 94.5%          | 99.2        | 0.971     | 0.952     | 0.992  |
| BoW    | Naïve Bayes   | 84.3%         | 90%            | 94.0        | 0.947     | 0.954     | 0.940  |
|        | Random Forest | 93.7%         | 95.4%          | 1.00        | 0.976     | 0.954     | 1.000  |
|        | AdaBoost      | 93.4%         | 93.4%          | 99.7        | 0.966     | 0.939     | 0.997  |
|        | SVM           | 93.75%        | 93.75%         | 1.00        | 0.968     | 0.938     | 1.000  |
|        | KNN           | 99%           | 100%           | 1           | 1         | 1         | 1      |
|        | ANN           | 99%           | 100%           | 1           | 1         | 1         | 1      |
| BoW    | Decision Tree | 99%           | 100%           | 1           | 1         | 1         | 1      |
|        | Naïve Bayes   | 99%           | 100%           | 1           | 1         | 1         | 1      |
|        | Random Forest | 99%           | 100%           | 1           | 1         | 1         | 1      |
|        | AdaBoost      | 99%           | 100%           | 1           | 1         | 1         | 1      |
|        | SVM           | 99%           | 100%           | 1           | 1         | 1         | 1      |





## REFERENCES

- [1] M. Coccia, "The increasing risk of mortality in breast cancer: A socioeconomic analysis between countries," *Journal of Social and Administrative Sciences*, vol. 6, no. 4, pp. 218–230, 2019, doi: 10.1453/jsas.v6i4.1972.
- [2] W. X. Liao *et al.*, "Automatic identification of breast ultrasound image based on supervised block-based region segmentation algorithm and features combination migration deep learning model," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 4, pp. 984–993, Apr. 2020, doi: 10.1109/JBHI.2019.2960821.
- [3] V. Kumar, B. K. Mishra, M. Mazzara, D. N. H. Thanh, and A. Verma, "Prediction of malignant and benign breast cancer: a data mining approach in healthcare applications," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 37, 2020, pp. 435–442, doi: 10.1007/978-981-15-0978-0\_43.
- [4] S. J. Lou *et al.*, "Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: A prospective cohort study," *Cancers*, vol. 12, no. 12, pp. 1–15, Dec. 2020, doi: 10.3390/cancers12123817.
- [5] M. O. Roberto Cesar *et al.*, "Method based on data mining techniques for breast cancer recurrence analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12145 LNCS, 2020, pp. 584–596, doi: 10.1007/978-3-030-53956-6\_54.
- [6] D. A. Omondigabe, S. Veeramani, and A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," *IOP Conference Series: Materials Science and Engineering*, vol. 495, no. 1, p. 012033, Jun. 2019, doi: 10.1088/1757-899X/495/1/012033.
- [7] A. I. Aishwarja, N. J. Eva, S. Mushtary, Z. Tasnim, N. I. Khan, and M. N. Islam, "Exploring the machine learning algorithms to find the best features for predicting the breast cancer and its recurrence," in *Intelligent Computing and Optimization*, 2021, pp. 546–558, doi: 10.1007/978-3-030-68154-8\_48.
- [8] A. Bharat, N. Pooja, and R. A. Reddy, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," in *2018 IEEE 3rd International Conference on Circuits, Control, Communication and Computing, I4C 2018*, Oct. 2018, pp. 1–4, doi: 10.1109/CIMCA.2018.8739696.
- [9] C. H. Shrivaya, K. Pravalika, and S. Subhani, "Prediction of breast cancer using supervised machine learning techniques," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 6, pp. 1106–1110, 2019.
- [10] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," in *5th IEEE Region 10 Humanitarian Technology Conference 2017, R10-HTC 2017*, Dec. 2018, vol. 2018-January, pp. 226–229, doi: 10.1109/R10-HTC.2017.8288944.
- [11] P. P. Golagani, T. S. Mahalakshmi, and S. K. Beebi, "Supervised learning breast cancer data set analysis in MATLAB using novel SVM classifier," in *Machine Intelligence and Soft Computing*, 2021, pp. 255–263.
- [12] R. Murtirawat, S. Panchal, V. K. Singh, and Y. Panchal, "Breast cancer detection using k-nearest neighbors, logistic regression and ensemble learning," in *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, Jul. 2020, pp. 534–540, doi: 10.1109/ICESC48915.2020.9155783.





- [13] S. Radhakrishna *et al.*, "Role of magnetic resonance imaging in breast cancer management," *South Asian Journal of Cancer*, vol. 07, no. 02, pp. 069–071, Apr. 2018, doi: 10.4103/sajc.sajc\_104\_18.
- [14] S. Gheisari, D. Catchpoole, A. Charlton, Z. Melegh, E. Gradhand, and P. Kennedy, "Computer aided classification of neuroblastoma histological images using scale invariant feature transform with feature encoding," *Diagnostics*, vol. 8, no. 3, p. 56, Aug. 2018, doi: 10.3390/diagnostics8030056.
- [15] A. H. Farhan and M. Y. Kamil, "Texture analysis of mammogram using histogram of oriented gradients method," *IOP Conference Series: Materials Science and Engineering*, vol. 881, no. 1, p. 012149, Jul. 2020, doi: 10.1088/1757-899X/881/1/012149.
- [16] Q. Wu, G. Xu, Y. Cheng, W. Dong, L. Ma, and Z. Li, "Histogram of maximal point-edge orientation for multi-source image matching," *International Journal of Remote Sensing*, vol. 41, no. 14, pp. 5166–5185, Jul. 2020, doi: 10.1080/01431161.2020.1727055.
- [17] S. Naresh and S. Vani, "Breast cancer detection using local binary patterns," *International Journal of Computer Applications*, vol. 123, no. 16, pp. 6–9, Aug. 2015, doi: 10.5120/ijca2015905726.
- [18] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access*, vol. 6, pp. 24680–24693, 2018, doi: 10.1109/ACCESS.2018.2831280.
- [19] P. Gupta and S. Garg, "Breast Cancer prediction using varying parameters of machine learning models," *Procedia Computer Science*, vol. 171, pp. 593–601, 2020, doi: 10.1016/j.procs.2020.04.064.
- [20] K. Atrey, Y. Sharma, N. K. Bodhey, and B. K. Singh, "Breast cancer prediction using dominance-based feature filtering approach: A comparative investigation in machine learning archetype," *Brazilian Archives of Biology and Technology*, vol. 62, pp. 1–15, 2019, doi: 10.1590/1678-4324-2019180486.
- [21] S. H. Yesuf, "Breast cancer detection using machine learning techniques," *International Journal of Advanced Research in Computer Science*, vol. 10, no. 5, pp. 27–33, Oct. 2019, doi: 10.26483/ijarcs.v10i5.6464.
- [22] M. N. Elgedawy, "Prediction of breast cancer using random forest, support vector machines and naïve Bayes," *International Journal of Engineering and Computer Science*, vol. 6, no. 1, pp. 19884–19889, Jan. 2017, doi: 10.18535/ijecs/v6i1.07.
- [23] N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis," *IEEE Access*, vol. 8, pp. 150360–150376, 2020, doi: 10.1109/ACCESS.2020.3016715.
- [24] V. P. C. Magboo and M. S. A. Magboo, "Machine Learning Classifiers on Breast Cancer Recurrences," *Procedia Computer Science*, vol. 192, pp. 2742–2752, 2021, doi.org/10.1007/978-981-15-7205-0\_10
- [25] T. L. Octaviani and Z. Rustam, "Random forest for breast cancer prediction," in *AIP Conference Proceedings*, 2019, vol. 2168, p. 020050, doi: 10.1063/1.5132477.
- [26] L. Sisters, *Matthews Correlation Coefficient: when to use it and when to avoid it*, ed. May. *Towards Data Science*, 2020.
- [27] R. Joseph Manoj, M. D. Anto Praveena, and K. Vijayakumar, "An ACO–ANN based feature selection algorithm for big data," *Cluster Computing*, vol. 22, no. S2, pp. 3953–3960, Mar. 2019, doi: 10.1007/s10586-018-2550-z.
- [28] A. F. M. Agarap, "On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset," in *ACM International Conference Proceeding Series*, 2018, pp. 5–9, doi: 10.1145/3184066.3184080.
- [29] M. O. F. Goni, F. M. S. Hasnain, M. A. I. Siddique, O. Jyoti, and M. H. Rahaman, "Breast cancer detection using deep neural network," in *ICCIT 2020 - 23rd International Conference on Computer and Information Technology, Proceedings*, Dec. 2020, pp. 1–5, doi: 10.1109/ICCIT51783.2020.9392705.
- [30] O. I. Obaid, M. A. Mohammed, M. K. Abd Ghani, S. A. Mostafa, and F. T. Al-Dhief, "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer," *International Journal of Engineering and Technology (UAE)*, vol. 7, no. 4.36 Special Issue 36, pp. 160–166, 2018, doi: http://dx.doi.org/10.14419/ijet.v7i4.36.23737.

## BIOGRAPHIES OF AUTHORS



**Chiman Haydar Salih**     is Asst. Lecture in Erbil Technology College, Erbil Polytechnic University She had received Master Degree in Electrical and Electronics Engineering, Turkey, and now she is a PhD Student in Erbil Technical Engineering College, Erbil Polytechnic University Major field (Computer vision, Machine Learning). She can be contacted at email: [chiman.salh@epu.edu.iq](mailto:chiman.salh@epu.edu.iq).



**Dr. Abbas M. Ali**     is an Asst. Prof. He graduated from Salahaddin University, collage of engineering. Ph.D. in computer vision, UKM, Malaysia. Head of software engineering and informatics dept college of engineering. He can be contacted at email: [Abbas.mohamad@su.edu.krd](mailto:Abbas.mohamad@su.edu.krd).