INTERNATIONAL JOURNAL OF
ARTIFICIAL INTELLIGENCE AND EMERGING TECHNOLOGY
VOLUME 5, ISSUE 1, 2022, 67 – 76.

AUTOMATIC CLASSIFICATION OF BREAST CANCER

HAMMAM ABDELAAL

Department of Information Technology, Faculty of Computers and information, Luxor University, Egypt

Abstract:
Breast cancer ranks first among the most common types of cancer, globally, regionally. Artificial intelligence plays an important role in medical sector, especially in improving healthcare for patients, in which the early detection and diagnosis of disease increasing the probability of recovery. This paper with the help of machine learning technique proposes to present a non-invasive method for diagnosing and classify breast diseases based on mammograms and ultrasound images, to extract the statistical features of them (smoothness, perimeter, area, concavity, compactness, symmetry, size, diameter, concave and radius), to identify the breast tissue as malignant tumor, or a benign tumor and predicting in the future at the long term to prevent it. Learning algorithms are used mainly: support vector machine (SVM), multilayer perceptron (MLP), naïve Bayes (NB) and Decision tree (DT) algorithms to build model capable of classifying the breast tissue into malignant or a benign, based on several features reached up to 30 features. The Results showed that SVM achieved higher accuracy which is reached up to 95.89%, followed by MLP classifier with 93.61%, and the NB accuracy which is reached up to 90.62%.

Keywords: Breast Cancer, Supervised Learning, SVM, MLP NB, DT

1. INTRODUCTION

The field of artificial intelligence has advanced to solve several problems in various areas of life, especially in medical sector. Where we can use artificial intelligence through supervised learning algorithms such as support vector machine (SVM), multilayer perceptron (MLP), naïve Bayes (NB) and Decision tree (DT) to make a model that searches for signs of breast cancer through mammograms and predict the breast cancer, AI-based algorithms represent a promising way to improve imaging accuracy digital mammography. AI system is able to classify images into normal, and abnormal (benign tumors or malignant) after learning and testing them as shown in figure 1, so it is able to teach itself and gain experience like humans, then the second stage is to follow up on the condition and the development of its recovery, then the third stage is to add or implement the long-term prediction, the prediction feature at the level of the generation resulting from the individual, from measuring the long-term incidence of the disease in parents in addition to their family history, and then measuring the percentage of giving birth to a generation that has the later gets sick

Building a model that is able to distinguish and classify breast cancer into benign or malignant tumor, based on statistical features extracted from mammograms or ultrasound images [1], these images are as an input for the system to extract some of important the features, based on these features the model can classify and differentiate image categories, and follow up on the recovery and treatment of the patient. In addition to predicting it’s occurring in the future (long-term) based on some other checkup such as pathological history or the genetic factor. In our experiments, we used popular tool WEKA, it is an important tool for applying machine learning algorithms.

INTERNATIONAL JOURNAL OF
ARTIFICIAL INTELLIGENCE AND EMERGING TECHNOLOGY
VOLUME 5, ISSUE 1, 2022, 67 – 76.

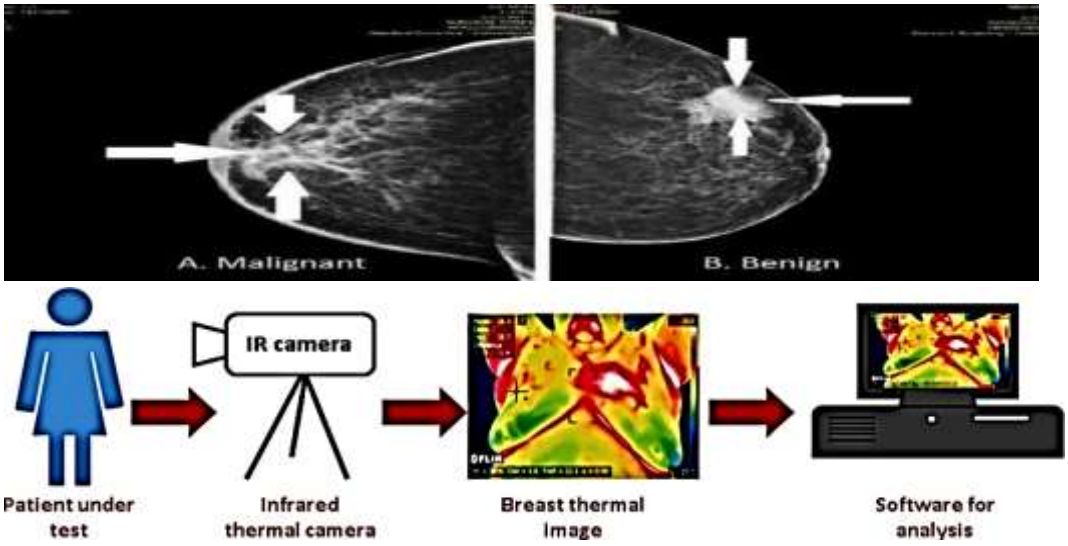

Fig. 1 AI system for breast cancer investigation

2. PROPOSED SYSTEM

Breast cancer prediction is a process based on a deriving a classifier model; this model can describe and distinguish Breast cancer classes. The derived model is based on the analysis of a set of mammograms whose class label are known, to extract the features form these images, and feature selection to select the most important features [2][3], which are more related to target class. Then classifier learning process, it aims to classify the Breast cancer into two categories, these categories are Benign and Malignant, according to features of the images that distinguish of each class of them.

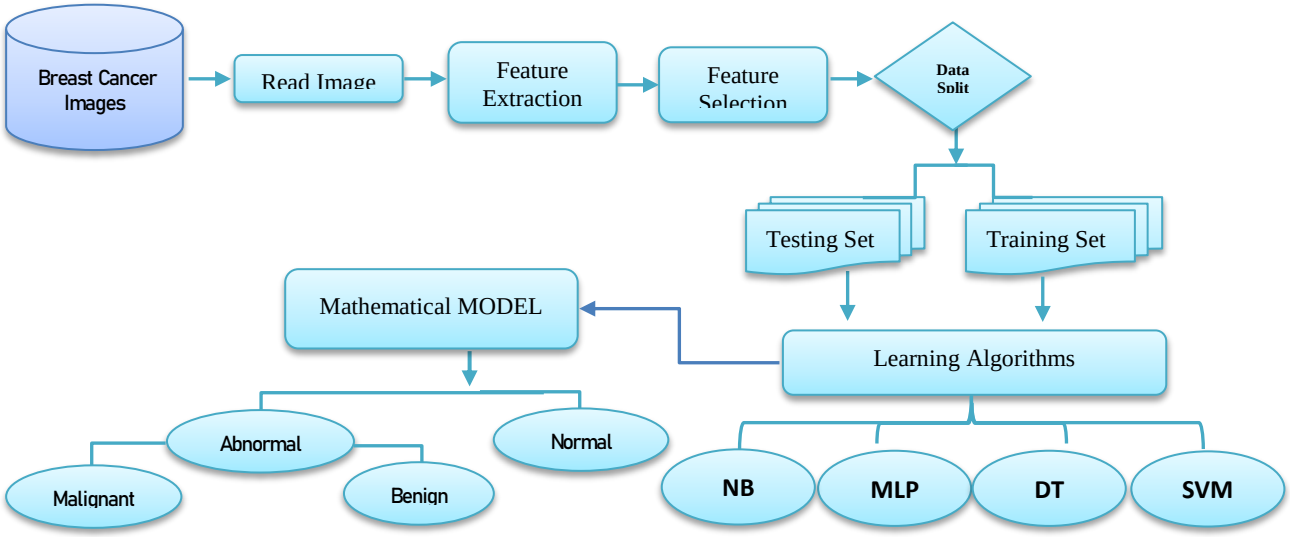

Fig. 2 Outline of the Breast cancer detection system

The proposed system is consisting of different stages as shown in figure 2, each stage contains some of steps, started with the read mammograms as input for the system and ended with the final decision of identify breast tissue as normal, or abnormal (malignant tumor, or a benign tumor) Read of image to extract the features, then these feature are selected, to select the most important features, which are more related to target class, using Information gain technique, it used to measure the dependence between features and labels and calculates the gain between the (i-TH) feature f_i and the class labels, after calculating both of the expected information needed to classify a tuple in D, and the expected information required for each feature [4]

The data set is spited into training and testing using 10 folds cross-validation (CV) and percentage split (PS) methods. Data set contains of 570 samples as shown in figure 3, each sample has 30 features mainly: smoothness mean, perimeter mean, compactness mean, symmetry mean, perimeter worst and concavity worst. In cross validation [5], the dataset is divided into 10 folds.

We use 9 of those parts for training and reserve one tenth for testing. We repeat this procedure 10 times each time reserving a different tenth for testing and calculate the accuracy for this iteration, and the end we get the overall accuracy, by calculate the mean these 10 measures. While in the percentage split, the dataset is divided randomly into 70% of the data set is used to train the model and 30% is used to test the model as shown in figure 4

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010
20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690
19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740
11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140
20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800
12.45	15.70	82.57	477.1	0.12780	0.17000	0.15780

Fig. 3 Sample of dataset

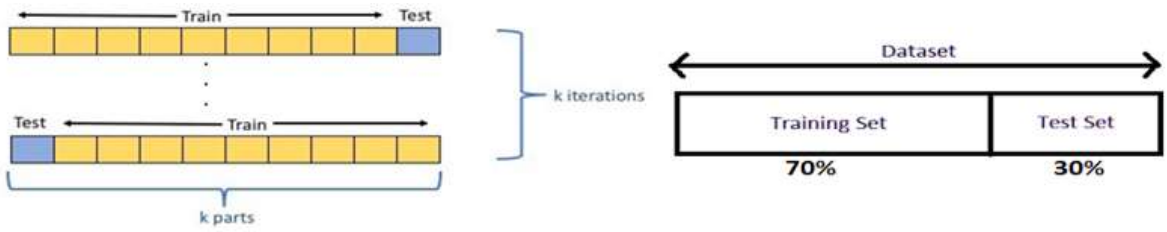

Fig. 4 Cross-validation and percentage split methods

3. LEARNING ALGORITHMS

There are many algorithms that are used to train the classification model that can be used it to predict the classes of new images is unseen samples. In supervised learning, the data are labeled with pre-defined classes such as Multinomial, Logistic Regression, SGD Classifier, SVC, Linear SVC, Nu SVC, MLP, SVM and NB Classifiers. In this study we used the most of these algorithms to identify the best classifier gives high accuracy, to predict the class of breast cancer, based on mammography; it is a special type of x-ray imaging used to create detailed images of the breast.

After the experimental results, we reported the best three classifiers of them, which have a highest accuracy mainly: SVM, MLP, VB and DT that have improved its accuracy using combination method (stacking algorithm). The learning process is two stages, the first Learning (training): Learn a model using the training data, and the second is testing, test the model using unseen test data to assess the model accuracy

4. CLASSIFIER EVALUATION

Classifier evaluation and testing are the final stage, it evaluates and testing of classifier; to validate of experimental results and determine the classifier capability to differentiating and classify image class to take the right final decision. The accuracy of the model is evaluated using different measures; mainly the Recall, Precision, and F-measure according to the following equations [6], based on determine the TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative), as shown in figure 5

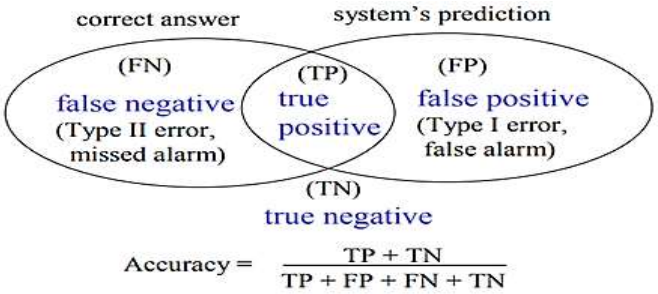

Fig. 5 the rate of correctly predicted two topics

TP and TN are the number of samples which are correctly assigned to the given category
FP and FN are the number of samples which are incorrectly not assigned to the category

$$Acc = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \quad (1) \quad ,$$

$$Precision = \frac{T_P}{T_P+F_P} \quad (2)$$

$$Sensitivity = \frac{T_P}{T_P+F_N} \quad (3) \quad , \quad F-measure = \frac{2(precision * recall)}{precision + recall} \quad (4)$$

5. EXPERIMENTAL RESULTS

Many classifiers have been used in this study, to identify the best classifier gives high accuracy. Table 1 and figure 6 show the overall accuracy using percentage split and cross-validation methods for each classifier (SVM, MLP, NB, and DT) as individual classifier, based on the number of instances that are correctly and incorrectly classified. The Experimental results have been evaluated and tested by WEKA tool.

Table 2 Overall accuracy using percentage split and cross-validation methods for each classifier

Method Classifier	Correctly Instances	Incorrectly Instances	Percentage split	Cross-validation
SVM	557	12	95.41%	95.89%
MLP	544	25	94.37%	93.61%
NB	527	42	89.19%	90.62%
DT	530	39	89.71%	91.15%

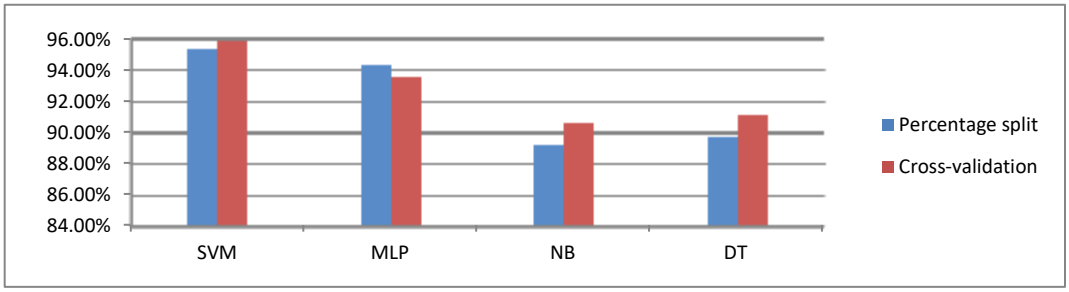

Fig. 6 Overall accuracy using percentage split and cross-validation methods for each classifier

Table 3 and table 4 show the results of the F1-score, Recall, Precision, MCC and ROC for each category (Malignant and Benign), using cross-validation, which generated by the two classifiers mainly: SVM and MLP

Class	TP_Rate	Precision	Recall	F-Measure	MCC	ROC_Area	PRC_Area
Malignant	0.928	0.975	0.928	0.951	0.935	0.953	0.943
Benign	0.977	0.95	0.977	0.963	0.935	0.953	0.949
Weighted_Avg.	0.959	0.959	0.959	0.959	0.935	0.953	0.947

Table 3 F-measure, Recall, Precision and MCC for each category using SVM Classifier using CV method

Table 4 F-measure, Recall, Precision and MCC for each category using MLP Classifier using CV method

Class	TP_Rate	Precision	Recall	F-Measure	MCC	ROC_Area	PRC_Area
Malignant	0.923	0.919	0.923	0.921	0.886	0.97	0.969
Benign	0.944	0.946	0.944	0.945	0.886	0.97	0.972
Weighted_Avg.	0.936	0.936	0.936	0.936	0.886	0.97	0.971

6. CONCLUDING

Artificial intelligence systems are used extensively in medical applications, such as the diagnostic and treatment, in additional to predict the disease before it occurs or discovers in its early stages, using learning algorithms for building a classifier model. In this study, we build classifier model using SVM, MLP, NB and DT and DT. They can describe and distinguish breast cancer type into benign or malignant tumor, based on the analysis of a set of mammograms. The Results showed that SVM achieved higher accuracy which is reached up to 97.89%, followed by MLP classifier with 95.61% % respectively, and improved the classification accuracy for NB and DT classifiers, they reached up to 96.78% and 97.72% respectively, the model is evaluated using two methods mainly: 10 folds cross-validation (CV) and percentage split, but the best is cross-validation, it depends resampling that uses different fold of the data to test and train the classifier on different iterations, and thus gives correct accuracy when testing it using unseen samples.

RREFERENCES

1. M Wei, Y Du, X Wu, J Zhu (2019) Automatic classification of benign and malignant breast tumors in ultrasound image with texture and morphological features. in 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID), pp. 126-130.

2. H. M. Abdelaal, B. R. Elemary, and H. A. Youness, "Classification of hadith according to its content based on supervised learning algorithms," IEEE Access, vol. 7, pp. 152379-152387, 2019.

3. Yousif A. Alhaj, Wiraj Udara, H M. Abdelaal, "Efficient Feature Representation Based on the Effect of Words Frequency for Arabic Documents Classification", Conference Paper: November 2018

4. Jiawei Han, Michelins Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", the Third Edition University of Illinois at Urbana Champaign, 2012 Elsevier Inc.

5. H. M. Abdelaal and H. A. Youness, “Knowledge Discovery in the Hadith according to the reliability and memory of the reporters using Machine learning techniques, Digital Object Identifier 10.1109/ACCESS.2019.2944118 " IEEE Access,. September-2019

6. H. Sawaf, J. Zaplo and H. Ney, "Statistical Classification Methods for Arabic News Articles", in In the Workshop on Arabic Natural Language Processing, Toulouse, France, 2001