Ghassan Ahmad Ismaeel

College of Pharmacy, University of Mosul, , Mosul, Iraq
ORCID: https://orcid.org/0000-0003-2366-6924

doi:10.15199/48.2023.01.02

Machine Learning to Diagnose Breast Cancer

Abstract. *As the number of breast cancer diseases is increasing rapidly every year, new technologies are utilized to predict and diagnose this disease for better women's lives worldwide. The development of Machine Learning can be utilized to contribute in this sense and help in the early diagnosis of breast cancer. This paper aims to predict and diagnose breast cancer using Machine Learning techniques such as support vector Machine (SVM) and Decision -tree and Nearest neighbour (KNN). The results show the out performance of SVM over the other methods. These methods can be very helpful to predict the breast cancer disease ahead of time.*

Streszczenie. *Ponieważ liczba zachorowań na raka piersi gwałtownie rośnie z roku na rok, nowe technologie są wykorzystywane do przewidywania i diagnozowania tej choroby w celu poprawy życia kobiet na całym świecie. Rozwój uczenia maszynowego może być wykorzystany do wniesienia wkładu w tym sensie i pomocy we wczesnej diagnozie raka piersi. Niniejszy artykuł ma na celu przewidywanie i diagnozowanie raka piersi przy użyciu technik uczenia maszynowego, takich jak maszyna wektora nośnego (SVM) oraz drzewo decyzyjne i najbliższy sąsiad (KNN). Wyniki pokazują wydajność SVM w porównaniu z innymi metodami. Metody te mogą być bardzo pomocne w przewidywaniu zgonów na raka piersi z wyprzedzeniem. (**Uczenie maszynowe do diagnozowania raka piersi**)*

Keywords: Breast cancer disease, machine learning, classification, SVM, KNN.
Słowa kluczowe: Rak piersi, uczenie maszynowe, klasyfikacja, SVM, KNN..

Introduction

Breast cancer is a disease that which the breast cell tissues split abnormally to create a new mass or tissue that is not supposed to be there. Breast cancer is very common in women more than in men and it has a high risk of death [1]. Breast cell production can be divided into benign (not cancer) and malignant (cancer) [2]. The tissue of benign in a normal slow way not to invade the neighbour cells and does not spread randomly. Breast cancer (Malignant) can be detected via some sign of an extra lump or mass in the breast area during the screening exams. Then different kinds of biopsies can be performed [3]. The sooner to discover cancer the better. That is why detecting the abnormal lamp via machine learning is very crucial.

Usually, there is a mammography or magnetic resonance imaging (MRI) test to help for detecting any cancer in the early stage in women's breasts. However, the symptom may not be so clear for any cancer at an early age of that tumour. The images are examined by radiologists to find out if there is any abnormal tumour.

Many machine learning techniques are enabled to help radiologists to detect breast cancer rapidly, and reliably.

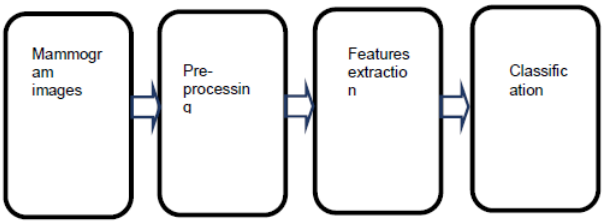

Fig. 1 state-of-art of image analysis.

Patient image preparation

A- Data collection

We used the digital database of the Mammographic Image Analysis Society (MIAS) [4] to implement our method to detect breast cancer. There are 322 images each of (digitized at 50-micron pixel edge that has been reduced to a 200-micron pixel edge) clipped/padded so that every image is 1024x1024 pixels resolution. These images are all indexed as the number and whether or not the breast has cancer or not, which provides a solid platform to test our methods. Furthermore, the images provide the locations of cancer or any abnormal tissues by (x,y) coordinates, which are measured in pixels.

B-image pre-processing

Mammograms are medical images that are difficult to interpret, thus a pre-processing phase is needed in order to improve the image quality and make the segmentation results more accurate. The first step involves the removal of unwanted parts in the background of the mammogram. The main objective of this process is to improve the quality of the image, to make it ready for further processing. Removing the irrelevant parts of the image is done by increasing the contrast of the mammogram using the threshold value.

Usually, mammogram images require some sort of noise-cleaning to be able to be interpreted with good accuracy and to obtain the best results. Firstly, the noise in the back needs to be removed. This step will lead to the next step to be ready for processing, the step will be via increasing the contrast.

Statistical Features Analysis

A statistical feature is one of the well-known and simplest methods, so far used, to measure the image texture behaviours. Normally, image analysis systems used two methods for extracting the image texture features; i.e., utilizing either 1st order histogram features, the 2nd order of the co-occurrence matrix features, or utilizing higher-order invariant moments features [5].

Proposed Methods

In this paper, supervised machine learning techniques using a minimum distance algorithm are used to locate the areas of interest (AOIs). First, the input image is processed then overlapping windows of size 5 by 5 pixels will scan the whole image looking for the abnormal areas. Different windows are compared with the minimum distance algorithm.

Our method contains three steps: 1) Pre-processing the input image, 2) finding the AOIs, 3) feature engineering extraction using machine learning to locate and identify the mass lesions from mammogram images.

1) pre-processing the input image

Medical images such as Mammograms are usually hard to understand and interpret. Therefore, in order to understand and interpret the images, some image preprocessing techniques such as noise removal can be implemented. This will definitely will increase the quality of the image as well.

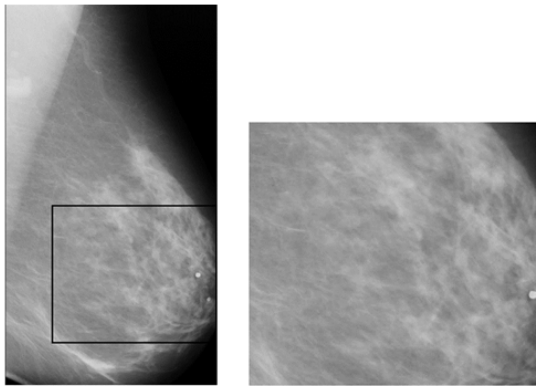

Fig. 2 Normal breast image.

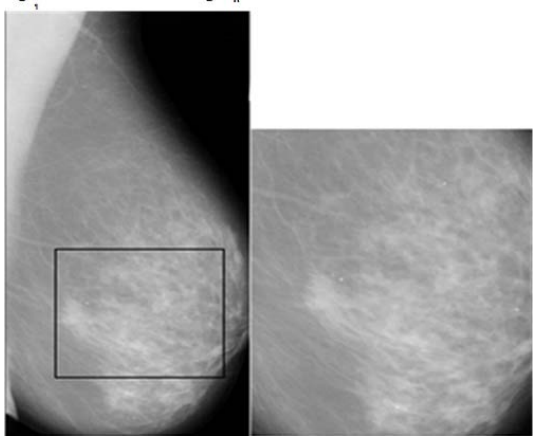

Fig. 3 mdb008 Dense-glandular, ill-defined masses, Malignant image coordinates of the center of abnormality is 318 x359 with an Approximate radius of 27 pixels.

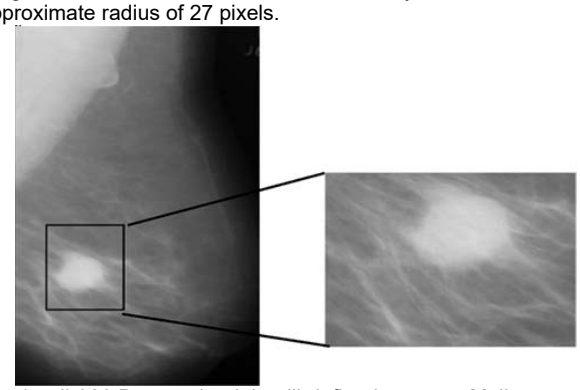

Fig. 4 mdb028 Dense-glandular, ill-defined masses, Malignant image coordinates of the center of abnormality is 318 x359 with an Approximate radius of 27 pixels.

2) pre-processing the input image

Medical images such as Mammograms are usually hard to understand and interpret. Therefore, in order to understand and interpret the images, some image preprocessing techniques such as noise removal can be implemented. This will definitely will increase the quality of the image as well.

3) **locating Area of interest**

To find the Areas of interest AoIs in the breast image tissue, the correlation between pixels inside the scanning window of 5 by 5 is calculated to find the adjoint nearest neighborhood pixels. This can be done by calculating the intensity measure and comparing it with the intensity threshold. If the intensity of the AoI in a particular window is greater than the minimum intensity threshold (for the normal tissue) then an abnormal mass can be identified.

Localization the AoI can be done in 6 steps:

Step 1: input the raw image to the system.

Step 2: define the window size to be scanned.

Step 3: calculate the intensity measure.

Step 4: Those who have high intensity, can be considered as potential abnormal areas.

Step 5: isolate the abnormal tissue.

Fig. 3 and 4 show two mammogram images. Fig. 3 for normal breast tissue and Fig. 4 for abnormal tissue also called (Malignant).

This process is very effective, but the process can be enhanced using some machine learning techniques such as support vector machine (SVM), decision tree, and linear prediction.

4) **Using machine learning techniques**

The machine learning techniques that are used in our paper are:

- A support vector machine (SVM) is a machine classifier that divides datasets into different classes to find the maximum marginal hyperplane via the nearest data [6].
- Random forest or called decision tree is a method of a classifier that uses the tree shape classification technique [7].
- K-nearest neighborhood (KNN) is a classifier technique that is based on the new point labeled to the nearest neighbor [8].

SVM

SVM is one of the supervised learning methods that study data classifications [6]. SVM is the algorithm that is utilized to determine the planes that split the data into different classes. If the data can be split linearly, it is called linear SVM. Figure 5 shows the SVM algorithm planes and data. The description of the plane can be stated in this equation.

$$g(x) = W^T x+b, \quad (1)$$

where w is the vector and b is the offset, x is the data.

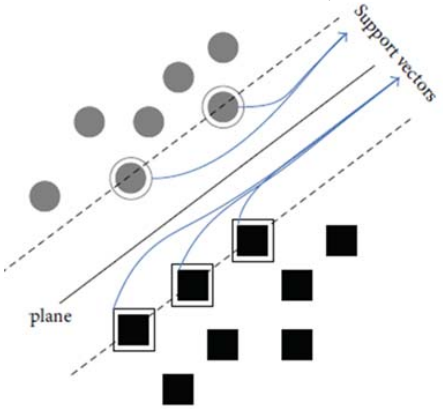

Fig. 5. support vector separated by a linear plane.

SVM can also be used for nonlinear classification problems via something called Kernel function to map the data to higher dimension subspaces to that different hyperplanes can separate between these new sets of data as

$$g(x) = W^T h(x)+b, \quad (2)$$

where h(x) is the mapping process between the non-linear input and the kernel space.

Decision tree

Decision tree is a tool that utilizes a tree shape model of different decisions with different possibilities that are based on outcomes and costs. Decision tree is also known as conditional control branches. Selecting a branch of a tree is based on the statement of different inputs. Furthermore, Decision tree has a decision rules that are associated with target outcomes.

K-nearest Neighbors

KNN is another non-parametric method, that assumes that similar elements live close to each others rather than the non-similar ones.

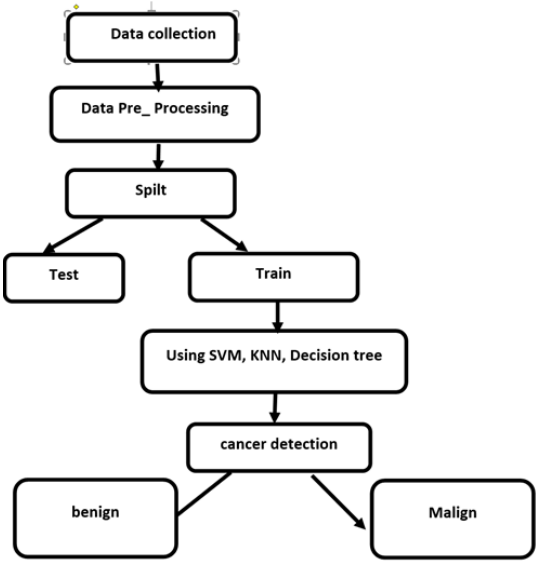

Fig. 6 the cancer detection block diagram.

Results

Performance measure

To evaluate the performance of a classifier, there are many ways. To measure the performance of a classifier, confusion matrix, as shown in Table (1), display the correct and incorrect classification results. In the table, TP, TN, FP, and FN represent the true positive, the true negative, the false positive, and the false negative, respectively.

Table 1: the confusion table.

Actual value	Recognized Value	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

To measure the accuracy metric, the accuracy can be calculated via

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \quad (3)$$

The sensitivity is

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \quad (4)$$

The F-score depends on the precision and the recall

$$\text{precision} = \frac{TP}{TP+FP}, \quad (5)$$

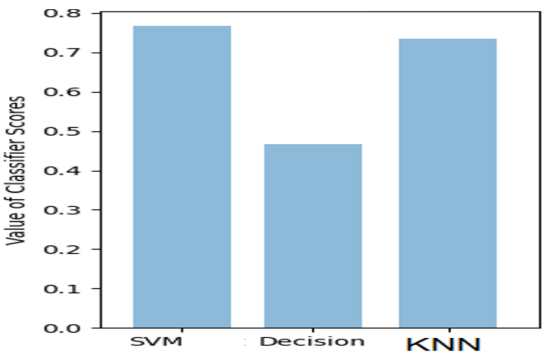

Fig. 7 The classification comparison plot of the SVM, decision tree, and KNN.

Table 2,3, and 4 show the confusion matrices for the three methods under study: SVM, decision tree, and KNN at k=6. Table 5 shows the accuracy, sensitivity and precision for the three methods. The results show good performance for SVM over other methods to predict the diseases. SVM outperform the decision tree and KNN in term of accuracy, sensitivity and precision.

Table 2 SVM result

Actual value	Recognized Value	
	Positive	Negative
Positive	343	9
Negative	13	199

Table 3 Decision tree result

Actual value	Recognized Value	
	Positive	Negative
Positive	334	23
Negative	138	74

Table 4 KNN result (k=6)

Actual value	Recognized Value	
	Positive	Negative
Positive	346	11
Negative	28	184

Table 5 the result parameters with the three methods

	SVM	Decision	KNN
Accuracy %	96	71.7	93.1
Sensitivity %	97.4	93.5	96.9
Precision %	96.3	70.1	92.5

Conclusion

In this paper we study and could use different machine learning algorithms to classify and predict the breast cancer diseases that is very common these days. The methods show good results to predict the diseases. SVM outperform the decision tree and KNN in term of accuracy, sensitivity and precision.

Authors

Ghassan Ahmad Ismaeel is an assistance lecturer at the College of Pharmacy, University of Mosul, Mosul, Iraq. His email is ghassanaldabbagh@uomosul.edu.iq .

REFERENCES

[1] Johnson, R. H., Anders, C. K., Litton, J. K., Ruddy, K. J., & Bleyer, A. (2018). Breast cancer in adolescents and young adults. Pediatric blood & cancer, 65(12), e27397.

[2] Allugunti, Viswanatha Reddy. "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms." International Journal of Engineering in Computer Science 4.1 (2022): 49-56.

[3] Eduati, Federica, et al. "A microfluidics platform for combinatorial drug screening on cancer biopsies." Nature communications 9.1 (2018): 1-13.

[4] SUCKLING J, P. "The mammographic image analysis society digital mammogram database." Digital Mammo (1994): 375-386.

[5] Li, Mengmeng, et al. "Ependymoma and pilocytic astrocytoma: Differentiation using radiomics approach based on machine learning." Journal of Clinical Neuroscience 78 (2020): 175-180.

[6] Osisanwo, F. Y., et al. "Supervised machine learning algorithms: classification and comparison." International Journal of Computer Trends and Technology (IJCTT) 48.3 (2017): 128-138.

[7] Kulkarni, Arun D., and Barrett Lowe. "Random forest algorithm for land cover classification." (2016).

[8] Zhang, Hao, et al. "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition." 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Vol. 2. IEEE, 2006.