A CAD SYSTEM FOR THE DETECTION OF MALIGNANT TUMORS IN DIGITIZED MAMMOGRAM FILMS

Michael A. Yacoub [1], A. S. Mohamed [1], Yasser M. Kadah [1]
[1]Department of Biomedical Engineering, Cairo University, Giza, Egypt
E-mail: Michael.Ghaly@k-space.org

Abstract- **The high incidence of breast cancer in women has increased significantly in the recent years. The most familiar breast tumors types are mass and microcalcification. Mammograms—breast X-ray—are considered the most reliable method in early detection of breast cancer. Computer-aided diagnosis system (CAD) can be very helpful for radiologist in detection and diagnosing abnormalities earlier and faster than traditional screening programs. We proposed in this paper a method to detect malignant tumors which is a three-step process. The first step is ROI extraction of 256 x 256 pixels size. The second step is the feature extraction, where we used a set of 99 features and we found that 83 of these feature are capable of differentiating between normal and cancerous breast tissues. The third step is the classification process. We used the techniques of the minimum distance, the k-Nearest Neighbor (k-NN) and Bayes classifiers to classify between normal and cancerous tissues. We examined the effect of changing the size of ROI extracted from the mammogram on the system by extracting ROI of size 512 x 512. Our computerized scheme was shown to have the potential to detect malignant tumors with a clinically acceptable sensitivity and low false positives.**

Keywords: **-CAD; Mammography; Minimum distance classifier; K-NN classifier and Bayes classifier.**

I. INTRODUCTION

Breast cancer is a leading cause of fatality among all cancers for women. However, the etiologies of breast cancer are unknown and no single dominant cause has emerged. Still, there is no known way of preventing breast cancer but early detection allows treatment before it is spread to other parts of the body. Currently, X-ray mammography is the single most effective, low-cost, and highly sensitive technique for detecting small lesions resulting in at least a 30% reduction in breast cancer deaths. [1]

It may not be feasible to routinely perform a second reading by a radiologist due to financial, technical, and logistical restraints. Therefore, efforts were made to develop a computer-aided detection (CAD) system. [2],[3] CAD can be defined as a diagnosis made to improve radiologists' performance by indicating the sites of potential abnormalities, to reduce the number of missed lesions, and/or by providing quantitative analysis of specific regions in an image to improve diagnosis. CAD systems typically operate as automated "second-opinion" or "double reading" systems. [4]

Computer-aided methods for detecting masses have been investigated using many different techniques. Karssemeijer [5] developed a statistical method for detection of microcalcifications in digital mammograms. The method is based on the use of statistical models and the general framework of Bayesian image analysis. Chan *et al.* [6] investigated a computer-based method for the detection of microcalcification in digital mammograms. The method is based on a difference image technique in which a signal suppressed image is subtracted from a signal enhanced image to remove structured background in the mammogram. Global and local thresholding techniques are then used to extract potential microcalcification signals.

Yu *et* al. [7] presented a computer-aided diagnosis (CAD) system for the automatic detection of clustered microcalcifications in digitized mammograms. The proposed system consists of two main steps. First, potential microcalcification pixels in the mammograms are segmented out by using mixed features consisting of wavelet features and gray level statistical features, and labeled into potential individual microcalcification objects by their spatial connectivity. Second, individual microcalcifications are detected by using a set of 31 features extracted from the potential individual microcalcification objects. The discriminatory power of these features is analyzed using general regression neural networks via sequential forward and sequential backward selection methods. The classifiers used in these two steps are both multilayer feedforward neural networks. The method is applied to a database of 40 mammograms (Nijmegen database) containing 105 clusters of microcalcifications.

Brake *et* al. [8] studied single and multiscale detection of masses in digital mammograms. Scale is an important issue in the automated detection of masses in mammograms, due to the range of possible sizes masses can have. In this work, it was examined if detection of masses can be done at a single scale, or whether it is more appropriate to use the output of the detection method at different scales in a multiscale scheme.

Abou-Chadi *et* al. [9] used a neural network approach for detecting candidate circumscribed lesions in digitized mammograms. The neural network learned using back propagation algorithms. The procedure depends mainly on the major difference between the histogram of the normal tissue and that of the cancerous tissue.

Nakayama *et* al. [10] used a filter bank for the detection of nodular and linear patterns. The filter bank is designed so that the subimages generated the elements of a Hessian matrix at each resolution level. By calculating the small and large eigenvalues, a new filter bank has the following three properties. (1) Nodular patterns of various sizes can be enhanced. (2) Both nodular and linear patterns of various sizes can be enhanced. (3) The original image can be reconstructed with these patterns removed. The filter bank is applied to enhance microcalcifications in mammograms.

In this paper, we present the development of a CAD system for the automatic detection of malignant masses in the breast. The proposed system consists of three major steps: The first step is ROI extraction of 256 x 256 pixels size. The second step is the feature extraction, where we used a set of 99 features and we found that only 83 of these feature are capable of differentiating between normal and cancerous breast tissues. The third step is the classification process; we

used the techniques of the minimum distance, the k-Nearest Neighbor (k-NN) and Bayes classifiers to classify between normal and cancerous tissues. Fig. (1) shows a schematic diagram for the system.

II. METHODOLOGY

A. Mammogram database.

The mammogram images used in this paper are provided by the University of South Florida, the digital database for screening mammography (DDSM) [11]. The dataset consists of digitized mammogram images, composed of both oblique and cranio-caudal views from 16 patients. Each mammogram shows one or more clusters of microcalcifications marked by expert radiologists. The position of individual masses is marked. The location of each cluster of microcalcifications is given in the format of a contour surrounding the mass. The images are digitized from films using the Lumysis scanner with 12 bits depth.

B. Extraction of ROI.

Using the contour supplied by the DDSM for each mammogram, we extracted the ROI of size 256 x 256 pixels with mass centered in the window. We have 51 cancerous and 59 normal ROI.

C. Feature extraction.

A typical mammogram contains a vast amount of heterogeneous information that depicts different tissues, vessels, ducts, chest skin, breast edge, the film, and the X-ray machine characteristics. In order to build a robust diagnostic system towards correctly classifying normal and abnormal regions of mammograms, we have to present all the available information that exists in mammograms to the diagnostic system so that it can easily discriminate between the normal and the abnormal tissue. However, the use of all the heterogeneous information, results to high dimensioned feature vectors that degrade the diagnostic accuracy of the utilized systems significantly as well as increase their computational complexity. Therefore, reliable feature vectors should be considered that reduce the amount of irrelevant information thus producing robust Mammographic descriptors of compact size. In our approach, we examined a set of 99 features were applied to the ROI using a window of size 32 pixels with 32 pixels shift, i.e. no overlap. Table (1) shows the set of features already used.

1- *Mean:* It represents the average gray level in the window [12].

$$\bar{x} = \frac{\sum x}{n} \qquad (1)$$

2- *Standard deviation:* It measures how spread out the values in a data set with respect to the mean [12].

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} \qquad (2)$$

3- *Variance:* A measure of the dispersion of a set of data points around their mean value [12].

$$\text{Var} = s^2 \qquad (3)$$

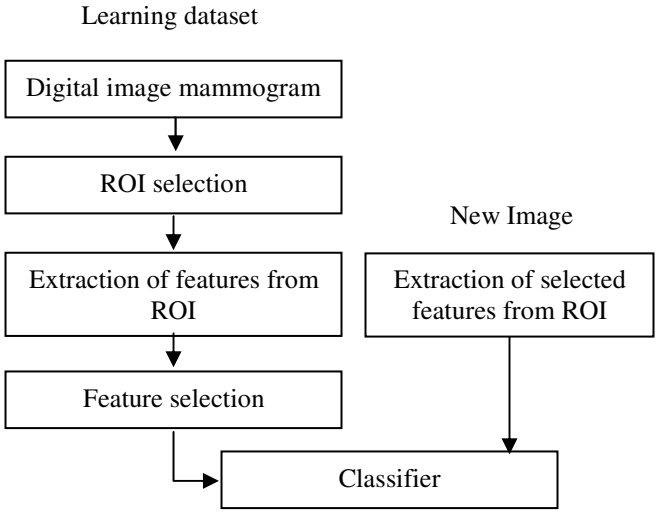

Fig.(1) A schematic diagram for the CAD system.

4- *Spreadness:* It measure amount of circularity of region of interest [13].

$$f = \frac{\sum_i \sum_j s(i,j)(i-i_o)^2 + \sum_i \sum_j s(i,j)(j-j_o)^2}{\sum_i \sum_j s(i,j)} \qquad (4)$$

5- *Entropy:* A statistical measure of randomness that can be used to characterize the texture of the image [12].

$$S_E = -\sum_{b=0}^{L-1} P(b) \log_2 \{P(b)\} \qquad (5)$$

6- *Invariant moments:* The set of moments is invariant to translation, rotation, and scale change [14].

7- *Percentile and Cumulative Frequency:* Cumulative relative frequency, or cumulative percentage, gives the percentage of having a measurement less than or equal to the upper boundary of the class interval.

The cumulative frequency graph provides a class of important statistics known as percentiles or percentile scores. The 90th percentile, for example, is the numerical value that exceeds 90% of the values in the data set and is exceeded by only 10% of them. We measured the percentiles at percentages ranging from 10% - 90% [15].

8- *Skewness:* It is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right [12], [15].

$$y = \frac{E(x-\mu)^3}{\sigma^3} \qquad (6)$$

9- *Kurtosis:* It is a measurement of how outlier-prone a distribution is [12].

$$K = \frac{E(x-\mu)^4}{\sigma^4} \qquad (7)$$

10- *Median contrast:* [7]

$$C(i,j) = p(i,j) - \text{median}(y(l,m), l,m \text{ found in window}) \qquad (8)$$

11- *Normalized gray level:* [7]

$$s(i,j) = \frac{p(i,j) - mean(y(l,m): l,m \in Window)}{std(y(l,m): l,m \in Window)} \qquad (9)$$

12- *Local binary partition:* Very simple and useful texture measure. For each pixel P in the image, the eight neighbors are examined to see if their intensity is greater than that of P. The results from the eight neighbors are used to construct an eight-digit binary number $b_1b_2b_3b_4b_5b_6b_7b_8$. Where $b_i = 0$ if the intensity of the i^{th} neighbor is less than or equal to that of P and 1 otherwise. A histogram of these numbers is used to represent the texture of the image [16].

13- *Edge frequency texture features:* A number of edge detectors can be used to yield an edge image from an original image. We can compute an edge dependent texture description function that finds the difference between pixels at specified distances. A total of 15 features are extracted for this [17].

$$E = |f(i,j) - f(i+d,j)| + |f(i,j) - f(i-d,j)| + |f(i,j) - f(i,j+d)| + f(i,j) - f(i,j-d)|$$

Where d varies from 1 to 15. (10)

14- Law's texture: It's a texture energy approach that measures the amount of variation within a fixed size window. A set of nine 5x5 masks is used to compute texture energy, which is then presented by a vector of nine numbers for each pixel of the image being analyzed [16].

15- Second order statistics: This category of parameters describes the gray level spatial inter-relationships and hence, represents efficient measures of the gray level texture homogeneity. These parameters are derived using the gray level co-occurrence matrix. The definition of this matrix is as follows [18].

$$\begin{aligned} Co(i,j) = \frac{1}{N} \text{ cardinality } \{[(k,l),(m,n)] \in \text{ROI}: \\ |k-m| = dx,\ |l-n| = dy, \\ \text{sign}\,[(k-l)\cdot(l-n)] = \text{sign}\,(dx \cdot dy), \\ g(k,l) = i,\ g(m,n) = j\} \end{aligned} \quad (11)$$

Where *Co(i,j)* is the gray level co-occurrence matrix entry at gray levels *i,j* , *g(i,j)* is the gray level of the pixel *(i,j)* in the ROI, *N* is the total numbers of pixels in the ROI, and *(dx,dy)* is a prescribed neighborhood definition. Three parameters are derived from this matrix and are defined as follows

* Contrast: Measures the local variations in the gray-level co-occurrence matrix.

$$\text{CON} = \sum_{i,j \in G} (i-j)^2 \cdot Co(i,j). \quad (12)$$

* Energy: Provides the sum of squared elements in the GLCM, also known as uniformity or the angular second moment.

$$\text{ASM} = \sum_{i,j \in G} [Co(i,j)]^2. \quad (13)$$

* Homogeneity: Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

$$\textbf{HOM.} = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \quad (14)$$

Table (1) List of features used for discriminating normal and abnormal tissues

No.	Feature description	Discriminate	
		256 x 256	512 x 512
1	Mean	yes	yes
2	Standard deviation	yes	yes
3	Variance	yes	yes
4	Spreadness	No	No
5	Entropy	yes	No
6-12	7 Invariant Moments	7-8-9	6-7-8-9
13-21	Percentile	yes	yes
22	Skewness	yes	No
23	Kurtosis	No	yes
24	Median contrast	yes	yes
25	Normalized gray level	No	yes
26	Local Binary Partition	yes	yes
27-42	Edge frequency texture	yes	yes
43-51	Law's texture	yes	yes
52-67	Contrast	58:64	yes
68-83	Energy	yes	yes
84-99	Homogeneity	yes	yes

D. Feature selection.

After the extraction of the previously mentioned features, it is found that not all the features can differentiate between normal and abnormal tissues. We applied a hypothesis test to decide whether the feature can discriminate or not.

E. Classifier.

The classification process is divided into the learning phase and the recognition phase. In the learning phase, known data are given and the feature parameters are calculated by the processing which precedes classification. Separately, the data on a candidate region which has already been decided as a tumor or as normal are given, and the classifier is trained. We used the learning set for this phase which consists of 40 cancerous ROI and 40 normal ROI. In the recognition phase, unknown data are given and the classification is performed using the classifier after learning. Breast cancer image diagnosis assistance is the task in the recognition phase. We used a testing set for this phase which consisted of 11 cancerous ROI and 19 normal ROI.

We used the minimum distance, the Voting K-Nearest Neighbor (K-NN) and Bayes classifiers [18].

1- Minimum distance classifier:

Suppose we are given a set of prototype points p_1,….., p_n, one for each of the n classes w_1,……., w_n. The minimum-distance classifier assigns a pattern x to the class w_i associated with the nearest point w_i according to the minimum Euclidian distance.

Each cluster will be composed from the vectors of its type. The normal cluster will be the 83 normal vectors of the features. The cancerous cluster will be the 83 cancerous vectors of the features. To get the center of gravity of each cluster, we get the average value of each vector. So, each cluster center will be a vector of size 83 x 1, where 83 represent the 83 features. Thus, we have one vector of 83 x 1 represents each cluster.

2- K-Nearest Neighbor (K-NN) classifier:

Suppose we are given a set of prototype points p_1 ,….., p_n, one for each of the n classes w_1 ,……., w_n . The minimum-distance classifier assigns a pattern x to the class w_i associated with the nearest point w_i.

The cluster is formed from some vectors; each one is of size 83 x 1. The number of these vectors equal the number of sample images used in learning the system. So, the normal cluster is composed from 40 vectors of size 83 x 1. The cancerous cluster is composed from 40 vectors of size 83x 1.

3- Bayes Classifier:

The Bayes decision rule classifies an observation (test sample) to the class that has the highest a *posteriori* probability among the two clusters. In this study, the dataset is assumed to have a Gaussian conditional density function and the a *priori* probabilities are assumed to be equal for the two classes that is:

$$f_{\mathbf{X}}(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}_i|^{1/2}} \cdot \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{M}_i)^T\Sigma_i^{-1}(\mathbf{x}-\mathbf{M}_i)\right] \quad (14)$$

And

$$P(w_i) = 1/2, \qquad i= 1, 2 \quad (15)$$

Where

$\mathbf{x}$ 83 x 1 data sample from the random vector X.

$\mathbf{M}_i$ 83 x 1 vectors representing the sample mean of class i

$\mathbf{\Sigma}_i$ 8 x 8 matrix representing the covariance matrix of class i

The Bayes decision rule is: Choose class $j \in \{1, 2\}$.

$$f_{\mathbf{X}}(\mathbf{x}|\omega_j)P(\omega_j) = \max\{f_X(\mathbf{x}|\omega_i)P(\omega_i)|i = 1, 2 \quad (16)$$

Since the covariance matrices are expected to be different for each class. This equation can be evaluated numerically.

F. Multi-size ROI:

We examined the system by applying the same features on different ROI of size 512 x 512. We used a learning set consisting of 20 cancerous and 20 normal ROI; while the testing set consist of 10 cancerous and 10 normal ROI. A set of 93 features were capable of differentiating between the two clusters. Table (1) shows the set of features already used.

III. RESULTS & DISCUSSION

A. Feature Extraction and selection.

We applied the previously mentioned 99 features using a windowsize of 32 pixels and a window shift of 32 pixels. i.e no overlap. Features are tested using a hypothesis test to decide whether or not this feature can discriminate between normal and abnormal tissues using a significance level of 0.05. The hypothesis indicated that only 16 features can't discriminate between the two clusters.

B. Classifiers.

Results differed by applying different type of classifiers due to the fact that each classifier has its own method for the formulation of the normal and cancerous clusters upon which it decides whether a test ROI is concidered cancerous or normal.

We measured, quantitatively, the detection performance of the classifiers by computing the sensitivity and specificity on the data. The sensitivity is the conditional probability of detecting a disease while there is in fact a cancerous breast. The specificity is the conditional probability of detecting a normal breast while the breast is indeed normal.

In the terms of the false-negative rate and the false-positive rate:

Sensitivity = 1 – false-negative rate (17)

Specificity = 1 – false-positive rate (18)

False-negative rate: the probability that the classification result indicates a normal breast while the true diagnosis is indeed a breast disease (i.e. positive). This case should be completely avoided since it represents a danger to the patient.

False-positive rate: the probability that the classification result indicates a breast disease while the true diagnosis is indeed a normal breast (i.e. negative). This case can be tolerated, but should be as infrequent as possible.

1- Minimum distance classifier:

Minimum distance classifier is considered one of the simplest types of classifiers. It depends on how much the tested ROI is correlated with each of the two clusters. The system detected 28 images from 40 cancerous images and detected 15 images from 40 normal images. This gave a sensitivity of 70 % and a specificity of 37.5 % for the learning set. When applying the testing set, the system detected 10 images from 11 cancerous images and detected 7 images from 19 normal images. This gave a sensitivity of 91% and a specificity of 37%.

2- K-Nearest Neighbor (K-NN) classifier:

K-Nearest Neighbor (K-NN) classifier is somehow more dependant than the minimum distance classifier.By testing the learning set and using the k value of 1, the system detected 40 images from 40 cancerous images and detected 40 images from 40 normal images. This gives a sensitivity of 100 % and a specificity of 100 %. By using the k value of 3, the system detected 33 images from 40 cancerous images and detected 23 images from 40 normal images. This gives a sensitivity of 83% and a specificity of 58 %.

By testing the testing set and using the k value of 1, the system detected 10 images from 11 cancerous images and detected 7 images from 19 normal images. This gives a sensitivity of 91 % and a specificity of 37 %. By using the k value of 3, the system detected 10 images from 11 cancerous images and detected 4 images from 19 normal images. This gives a sensitivity of 91% and a specificity of 21%.

3-Bayes classifier:

Bayes classifier is considered one of the statistical classifiers. It depends on how much the tested ROI has a distribution which is correlated with each of that of the two clusters. The system detected 40 images from 40 cancerous images and detected 40 images from 40 normal images. This gave a sensitivity of 100 % and a specificity of 100 % for the learning set. When applying the testing set, the system detected 8 images from 11 cancerous images and detected 11 images from 19 normal images. This gave a sensitivity of 73% and a specificity of 58%. Table (2) illustrates the results obtained using the classifiers for each of the ROI sizes.

Evaluating the results obtained, it's found that for ROI of size 256 x 256, the best results obtained for the learning set using both K-NN classifier with K=1 and Bayes classifier (sensitivity: 100%, specificity: 100%) while for the testing set, sensitivity of the system didn't considerably change using all the classification techniques. Evaluating the results for the second group of size 512 x 512, it is found that the best results for the learning set is obtained using K-NN classifier with K=1, while for the testing set best results are obtained using the minimum distance classifier.

Comparing the results obtained from the K-NN classifier with different values of K, it's found that using K-NN with K=3 is better than that with K=1 for the testing set.

Comparing the results obtained when applying on the two different ROI sets, it's found that for the 256 x 256 ROI, the system shows better sensitivity than that of the 512 x 512 ROI. While the system shows better specificity when dealing with 512 x 512 ROI than the other set except for the Bayes classifier (Learning = 30%, Testing = 20%).

IV. CONCLUSION

Automated breast cancer detection has been studied for more than 20 years; the CAD mammography systems for microcalcification detection have gone from crude tools in the research laboratory to commercial systems. Other challenges for future research which may increase the capability of the system.

REFERENCES

[1] Heng-Da Cheng, Yui Man Lui, and Rita I. Freimanis, "A Novel Approach to Microcalcification Detection Using Fuzzy Logic Technique" IEEE transactions on medical imaging, vol. 17, no. 3, June 1998.

[2] Winsberg F, Elkin M, Macy J, Bordaz V, Weymouth W. "Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis". Radiology 1967; 89:211–5.

[3] Christiane Marx, Ansgar Malich, Mirjam Facius, Uta Grebenstein, Dieter Sauner, Stefan O.R. Pfleiderer, Werner A. Kaiser "Are unnecessary follow-up procedures induced by computer-aided diagnosis (CAD) in mammography? Comparison of Mammographic diagnosis with and without use of CAD" European Journal of Radiology 51 (2004) 66–72.

[4] Paul Sajda,, Clay Spence and John Pearson "Learning Contextual Relationships in Mammograms Using a Hierarchical Pyramid Neural Network IEEE transactions on medical imaging, vol. 21, no. 3, march 2002.

[5] N. Karssemeijer, "Recognition of clustered microcalcifications using a random field mode, biomedical image processing and biomedical visualization," Proc. SPIE, vol. 1905, pp. 776–786, 1993.

[6] H. P. Chan, K. Doi, C. J. Vyborny, K. L. Lam, and R. A. Schmidt, "Computer-aided detection of microcalcifications in mammograms methodology and preliminary clinical study," Investigative Radiol., vol. 23, pp. 664–671, 1988.

[7] Songyang Yu and Ling Guan "A CAD System for the Automatic Detection of Clustered Microcalcifications in Digitized Mammogram Films" IEEE transactions on medical imaging, vol. 19, no. 2, february 2000.

[8] Guido M. te Brake and Nico Karssemeijer "Single and Multiscale Detection of Masses in Digital Mammograms" IEEE transactions on medical imaging, vol. 18, no. 7, July 1999.

[9] Fatma E.Z. Abou-Chadi, Noha Youssry, Alaa' M. Elsayad, " A Neural Network Approach For Mass Detection In Digitized Mammograms", ACBME.

[10] Ryohei Nakayama and Yoshikazu Uchiyama "Development of New Filter Bank for Detection of Nodular Patterns and Linear Patterns in Medical Images" Systems and Computers in Japan, Vol. 36, No. 13, 2005.

[11]http://marathon.csee.usf.edu/Mammography/Database.html

[12] Digital Image Processing: PIKS Inside, Third Edition. William K. Pratt 2001 John Wiley & Sons, Inc.

[13] Hidefumi Kobatake, Masayuki Murakami, Hideya Takeo, and Sigeru Nawano, "Computerized Detection of Malignant Tumors on Digital Mammograms," IEEE transactions on medical imaging, vol. 18, no. 5, May 1999.

[14] Rafael Gonzalez and Richard woods, Digital images processing. Prentice Hall, 2002.

[15] CHAP T. LE, Introductory Biostatistics. A John Wiley & Sons Publication.

[16] Linda Shapiro, Computer vision, 2000.

[17] J. Strand and T. Taxt, "Local frequency features for texture classification", Pattern Recognition, vol. 27, no. 10, pp. 1397-1406, 1994.

[18] Yasser M. Kadah, Aly A farag, Ahmed M. badawy and Abou-Baker M. Youssef, "Classification algorithm for quantitative tissue characterization of diffuse liver disease from ultrasound," IEEE transactions on medical imaging, vol. 15, no. 4, August 1996.

Table (2) Results obtained using the classifiers.

		Classifier							
		Minimum Distance		K-Nearest Neighbor (K-NN) classifier				Bayes	
				K = 1		K = 3			
		L	T	L	T	L	T	L	T
256 x 256	Sensitivity	70%	91%	100%	91%	83%	91%	100%	73%
	Specificity	37.50%	37%	100%	37%	58%	21%	100%	58%
512 x 512	Sensitivity	90%	90%	100%	30%	95%	40%	70%	100%
	Specificity	60%	70%	100%	80%	70%	90%	30%	20%