Proceedings of the International Multiconference on Computer Science and Information Technology pp. 11–17
ISBN 978-83-60810-27-9
ISSN 1896-7094

Using data mining for assessing diagnosis of breast cancer

Dr. Medhat Mohamed Ahmed Abdelaal
Statistics and Mathematics Department,
Faculty of Commerce, Ain Shams University.
medhatal@hotmail.com

Muhamed Wael Farouq
Statistics and Mathematics Department,
Faculty of Commerce, Ain Shams University.
m.wael.farouq@gmail.com

Prof. Dr. Hala Abou Sena
Faculty of Medicine, Ain Shams University.

Prof. Dr. Abdel-Badeeh Mohamed Salem
Faculty of Computer Science, Ain Shams University.

***Abstract*—The capability of the classification SVM, Tree Boost and Tree Forest in analyzing the DDSM dataset was investigated for the extraction of the mammographic mass features along with age that discriminates true and false cases. In the present study, SVM technique shows promising results for increasing diagnostic accuracy of classifying the cases witnessed by the largest area under the ROC curve (area under empirical ROC curve =0.79768 and area under binomial ROC curve =0.85323) comparable to empirical ROC and binomial ROC of 0.57575 and 0.58548 for tree forest while least empirical ROC and binomial ROC of 0.53452 and 0.53882 was accounted by tree boost. These results are confirmed by SVM average gain of 1.7323, tree forest average gain of 1.5576 and tree boost average gain of 1.5718.**

***Keywords*- Breast Cancer, Classification Support Vector Machine (SVM), Decision Tree, Receiver Operating Characteristic Curve (ROC), Tree Boost, Tree Forest, Gain.**

I. Introduction

CANCER can develop when cells in a part of the body begin to grow out of control. These extra cells form a mass of tissue, called a growth or tumor. Tumors can be benign or malignant. The scientific discipline whose goal is the classification of objects into a number of categories or classes can be called Pattern Recognition. Objects can be images, signal waveforms or any type of measurement that needs to be classified [11].

The importance of the study is that; the breast cancer refers to life-threatening malignancies that develop in one or both breasts and is the most common form of cancer among women in developed countries. According to American Cancer Society one in eight women will develop breast cancer during their lifetime.

The problem with Breast Cancer Diagnosis is that despite radiographic breast imaging and screening has allowed for more accurate diagnosis of breast cancer, 10% to 30% of malignant cases are not detected for various reasons. There are two errors typical in examining mammograms. They are False Positives (FP) and False Negatives (FN).

The Computer-Aided Diagnosis (CAD) can reduce both the FP and the FN diagnosis rates. CAD is an application of pattern recognition aiming at assisting doctors in making diagnostic decisions. The final diagnosis is made by the doctor. Our aim is to utilize a pattern recognition system in order to assist radiologist with a "Second" opinion by concluding the mammographic mass features that most indicates malignancy.

II. Data Set

This part of the study includes shedding light on the case study used and the collected data description.

An image may be defined as a two-dimensional function, $f(x, y)$, where x and y are spatial (plane) coordinates, and the amplitude of f at any pair of coordinates (x, y) is called the intensity or gray level of the image at that point. When x, y, and the amplitude values of f are all finite, discrete quantities, we call the image a digital image. The field of digital image processing refers to processing digital images by means of a digital computer. Note that a digital image is composed of a finite number of elements, each of which has a particular location and value. These elements are referred to as picture elements, image elements and pixels. Pixel is the term most widely used to denote the elements of a digital image [11].

Computers cannot handle continuous images but only arrays of digital numbers. Thus it is required to represent images as two-dimensional arrays of points. A point on the 2-D grid is called a pixel or pel. Both words are abbreviations of the word picture element. A pixel represents the irradiance at the corresponding grid position. In the simplest case, the pixels are located on a rectangular grid; the position of the pixel is given in the common notation for matrices. The first index, m, denotes the position of the row, the second, n, the position of the column [4].

The principal goal of the image segmentation process is to partition an image into regions of interest that are homogeneous with respect to one or more homogeneity criteria(s) or features. Segmentation is an important tool in medical image processing, and it has been useful in many applications. A segmentation algorithm, in a mammographic context, is an algorithm used to detect something, usually the whole breast or a specific kind of abnormalities, like micro-calcifications or masses [8].

A wide variety of segmentation techniques have been proposed. However, there is no one standard segmentation technique that can produce satisfactory results for all imaging applications. The definition of the goal of segmentation varies according to the goal of the study and the type of image data. Different assumptions about the nature of the analyzed images leads to the use of different algorithms [7].

This section provides the general mammographic image information explaining the mammographic abnormalities then introduces techniques, operations and statistics that form the tools for the development of comprehensive analysis/ diagnosis algorithms.

The digital database for screening mammography (DDSM) is a resource for use by the mammographic image analysis research community. The primary purpose of the database is to facilitate sound research in the development of computer algorithms to aid in screening. Secondary purposes of the database may include the development of algorithms to aid in the diagnosis and the development of teaching or training aids. The database contains approximately 2,500 studies [15].

The most effective method of early detection of the breast cancer is mammograms, certain characteristics in the mammograms determines whether cancer exists or not, breast cancer often presents as a mass with or without presence calcifications.

The location, size, shape, density, and margins of the mass are useful for the radiologist in evaluating the likelihood of cancer. Most benign masses are well circumscribed, compact, and roughly circular or elliptical. Malignant lesions usually have a blurred boundary, an irregular appearance, and sometimes are surrounded by a radiating pattern of linear spicules. Masses are categorized by their shape, density, and margins.

The mass shape is described with a four-point assessment: round, oval, lobular and irregular. The mass margins modify the boundaries. For example the overall shape may be round, but close inspection may reveal scalloping along the border, which may indicate a degree of irregularity or a lobular characteristic. The margins are rated with a 5-point system: circumscribed (well-defined or sharply-defined) margins, microlobulated margins, obscured margins, indistinct margins and spiculated margins as shown in Fig. 1.

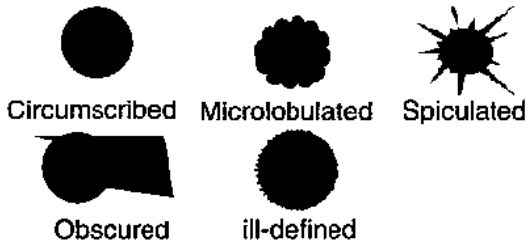

Fig. 1 Mass descriptors for margin

The intensity or the x-ray attenuation of the mass tissue region is described as density. The density here is the relative density, i.e. higher, lower or similar relative to the surrounding tissue. The density is rated on 4-point system:

High density, equal density, low density (lower attenuation, but not fat containing) and fat containing – radiolucent [10].

Numerous statistics can be developed from digital images which aid in describing and analyzing images. This section provides an introduction to a selected group of image statistics. The selected group represents those used in this research; however, they are far from exhaustive. In fact, a significant amount of researches continues developing new statistics for describing and analyzing images. This research is primarily interested in the application of existing statistics. .

(1) Age

(2) Variance

The variance of an image is a measure of the variation of pixel intensities in the image.

(3) Area

The area is defined as the total number of pixels belonging to the object.

(4) X-centroid, Y-centroid

The coordinates of the geometrical center of the object defined with respect to the image origin.

$$x_{centroid} = \sum_{i \in object} i \div A \qquad (1)$$

$$y_{centroid} = \sum_{j \in object} j \div A \qquad (2)$$

where i and j are image pixel coordinates and A is the area of the segment or region of interest (ROI) [9].

(5) Compactness

Compactness is another measure of the object's roundness and is calculated as:

$$compactness = p^2 \div 4\pi A \qquad (3)$$

where P is the object perimeter. Compactness gives the minimal value 1 for circles [12].

(6) Circularity

Area and perimeter are two parameters which describe the size of an object. In order to compare objects which are observed from different distances, it is important to use shape parameters which do not depend on the size of the object on the image plane. The circularity c is defined as:

$$c = p^2 \div A \qquad (4)$$

The circularity is a dimensionless number with a minimum value of $4\pi \approx 12.57$ for circles. The circularity is 16 for a square and $12\sqrt{3} \approx 20.8$ for an equilateral triangle. Generally, it shows large values for elongated objects.

(7) Eccentricity

This is a measure similar to the circularity but with a better defined range. The parameter is extracted from the second-order moments as:

$$\varepsilon = \frac{\left(m_{2,0} - m_{0,2}\right)^2 + 4m_{1,1}^2}{\left(m_{2,0} + m_{0,2}\right)^2} \qquad (5)$$

The eccentricity ranges from 0 to 1, it is zero for circular object and one for line shaped object [3].

III. Statistical Techniques

One of the most useful applications of statistical analysis is the development of a model to explain the relationship between the variables; many types of models have been developed, including classification support vector machines, tree boost and tree forest. This part will focus on the deployed analytical techniques.

SUPPORT VECTOR MACHINES

Methods for analyzing and modeling data can be divided into groups; supervised learning and unsupervised learning. Supervised learning requires input data that has both independent variables and a dependent (target) variable whose value is to be estimated. By various means, the process learns how to model predict the value of some variable, then supervised learning is recommended approach.

Unsupervised learning does not identify a dependent variable (target), but rather treats all of the variables equally. In this case, the goal is not to predict the value of a variable but rather to look for patterns, groupings or other ways to characterize the data that may lead to understanding of the way the data relate. Cluster analysis, correlation, factor analysis (principle components analysis) and statistical measures are examples of unsupervised learning.

One of the most useful applications of statistical analysis is the development of a model to represent and explain the relationship between variables. One of the best state-of-the-art modeling methods including support vector machines (SVM).

In the manner of speaking of SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyper plane is called a feature. The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modeling is to find the optimal hyper plane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other side of the plane. The vectors near the hyper plane are support vectors.

To illustrate classification SVM, let us assume that there is a linear relationship with N observations. The independent variable is x_i and the dependent variable is y_i given that i=1, 2, 3……N. The goal of classification SVM is to produce a linear function which can make the best fit of the dependent variable y_i. The linear function can be expressed as follows:

$$y = f(x) = \langle b \cdot x \rangle + a \tag{6}$$

Where a and b are the classification parameters and is the dot product of b and x.

The dot product of two vectors is defined as:

$$b \cdot x = \sum_{i=1}^{N} b_i x_i = b_1 x_1 + b_2 x_2 + \ldots . b_N x_N$$

The optimum classification function can be found by minimizing the following function:

$$M(b, S) = 0.5 \|b\|^2 + C \sum_{i=1}^{N} \left(S_i^- + S_i^+ \right) \tag{7}$$

The constraint can be as follows:

$$\begin{aligned} & y_i - \langle b, x_i \rangle - a \le \varepsilon + S_i \\ & \langle b, x_i \rangle + a - y_i \le \varepsilon + S_i^* \\ & S_i, S_i^* \ge 0 \end{aligned} \tag{8}$$

Where C is a constant greater than zero, and can be used to determine the trade off the smoothness of f and the amount up to which deviations larger than ε are accepted. and S_i^+ are two slack variables, which can be used to represent the upper and the lower constraints of the classification function. is norm of b. the norm can be as follows:

$$\|b\| = \sqrt{b_1^2 + b_2^2 + \ldots b_N^2}$$

To find the optimum solution of M function, loss function must be determined. Loss function is a function shows the maximum allowed deviation of the predicted values from the observed one. The most recommended loss functions are four. These four functions are: Huber, ε-insensitive, quadratic and Laplace [13]. Fig. 2 shows the difference between the four types of loss functions.

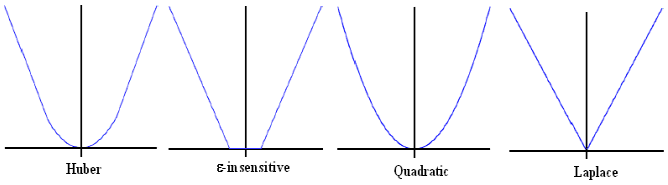

Fig. 2 The four types of loss functions

The first loss function is Huber loss function: it is a robust loss function that has optimal properties when the underlying distribution of the data is unknown. The second one is ε-insensitive loss function; it is an approximation to Huber's loss function but it can reduce sensitivity to the outliers. The third one is quadratic loss function: it corresponds to the predictable least squares error measure. The fourth one is Laplace loss function: it is less sensitive to outliers than the quadratic loss function.

In this paper, the ε-insensitive loss function was selected. For classification SVM model which depends on ε-insensitive loss function: the difference between the estimated values and the observed values of the dependent variable y_i can be calculated. If the difference is less than ε, the classification function is considered to be most popular and accurate [13]. The ε-insensitive loss function can be expressed as follows:

$$|S|_{\varepsilon} = \begin{cases} o & if\ |S| < \varepsilon \\ |S| - \varepsilon & otherwise \end{cases} \quad (9)$$

Most modeling techniques are trying to find the best fit between the observed and predicted values, however, SVM'S ε-insensitive loss function focuses on optimizing a bound around the classification function thus making it stronger against the outliers.

To find the solution of equation 7, Lagrange optimization must be used, the solution to this optimization problem can be expressed as follows: Minimizes the target function as follows:

$$\begin{aligned} &\varepsilon \sum_{i=1}^{N} \left(\alpha_i + \alpha_i^*\right) \\ &+0.5 \sum_{i=1,j=1}^{N} \left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)\left(x_i \cdot x_j\right) \\ &+\sum_{i=1}^{N} y_i\left(\alpha_i - \alpha_i^*\right) \end{aligned} \quad (10)$$

The constraints can be as follows:

$$\sum_{i=1}^{N} \left(\alpha_i - \alpha_i^*\right) = 0, 0 \le \alpha_i,\ \alpha_i^* \le C \quad (11)$$

SVM models are built around a kernel function that transforms the input data into an n-dimensional space where a hyper plane can be constructed to partition the data.

There are four kernel functions, linear, polynomial, radial basis function (RBF) and sigmoid (S-shaped). There is no way in advance to know which kernel function will be best for an application.

The RBF kernel non-linearity maps samples into a higher dimensional space, so it can handle nonlinear relationships between target categories and predictor attributes; a linear basis function cannot do this. Furthermore, the linear kernel is a special case of the RBF. A sigmoid kernel behaves the same as a RBF kernel for certain parameters. The RBF function has fewer parameters to tune than polynomial kernel, and the RBF kernel has less numerical difficulties.

This is the case of linear classification, but to illustrate the classification case of non-linearity, the data must be firstly linearized by mapping it into a higher dimensional space, called "feature space" by using kernel functions, that linear classification functions can be applied. The most recommended kernel function is the RBF Kernels [6]. The RBF kernel is defined as:

Given that is a kernel parameter.

Insert kernel function in the previous model (10), this model can be adjusted as follows:

The same constraints in (8) can be used, minimize:

$$\begin{aligned} &\varepsilon \sum_{i=1}^{N} \left(\alpha_i + \alpha_i^{\dot{c}}\right) \\ &+0.5 \sum_{i=1,j=1}^{N} \left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right) K\left(x_i \cdot x_j\right) \\ &+\sum_{i=1}^{N} y_i\left(\alpha_i - \alpha_i^*\right) \end{aligned} \quad (12)$$

DECISION TREES

A decision tree is a logical model represented as a binary (two-way split) tree that shows how the value of a target variable can be predicted by using the values of a set of predictor variables. There are many techniques for decision trees. In this paper the authors selected two techniques which are Boosting Trees and Tree Forest.

A decision tree can be used to predict the values of the target variable based on values of the predictor variables.

Each node represents a set of records (rows) from the original dataset. Nodes that do not have child nodes are called "terminal" or "leaf" nodes. The topmost node called the "root" node. Unlike a real tree, decision trees are drawn with their root at the top. The root node represents all of the rows in the dataset.

A decision tree is constructed by a binary split that divides the rows in a node into two groups (child nodes). The same procedure is then used to split the child groups. This process is called "recursive partitioning".

DECISION BOOSTING TREES

"Boosting" is a technique for improving the accuracy of a predictive function by applying the function repeatedly in a series and combining the output of each function with weighting so that the total error of the prediction is minimized. In many cases, the predictive accuracy of such a series greatly exceeds the accuracy of the base function used alone.

The TreeBoost algorithm is optimized for improving the accuracy of models built on decision trees. Research has shown that models built using TreeBoost are among the most accurate of any known modeling technique.

The TreeBoost algorithm is functionally similar to Decision Tree Forests because it creates a tree ensemble, and it uses randomization during the tree creations. However, a random forest builds the trees in parallel and they "vote" on the prediction; whereas TreeBoost creates a series of trees, and the prediction receives incremental improvement by each tree in the series.

Mathematically, a TreeBoost model can be described as:

$$\mathrm{PT} = F_0 + \mathrm{B}_1 T_1(X) + \mathrm{B}_2 T_2(X) + . + B_M T_M(X) \quad (13)$$

Where *PT* is the predicted target, F_0 is the starting value for the series (the median target value for a regression model), X is a vector of "pseudo-residual" values remaining at this point in

the series, $T_1(X)$, $T_2(X)$ are trees fitted to the pseudo-residuals and B_1, B_2, etc. are coefficients of the tree node predicted values that are computed by the TreeBoost algorithm.

The first tree is fitted to the data. The residuals from the first tree are then fed into the second tree which attempts to reduce the error. This process is repeated through a chain of successive trees. The final predicted value is formed by adding the weighted contribution of each tree.

Usually, the individual trees are fairly small, but the full TreeBoost additive series may consist of hundreds of these small trees. TreeBoost models often have a degree of accuracy that cannot be obtained using a large, single-tree model. TreeBoost models are often equal to or superior to any other predictive functions including neural networks. TreeBoost models can handle hundreds or thousands of potential predictor variables. Irrelevant predictor variables are identified automatically and do not affect the predictive model. TreeBoost uses the Huber M-regression loss function which makes it highly resistant to outliers and misclassified cases. TreeBoost procedures are invariant under all (strictly) monotone transformations of the predictor variables. So transformations such as (a*x+b), log(x) or exp(x) do not affect the model. Hence, there is no need for input transformations. The sophisticated and accurate method of surrogate splitters is used for handling missing predictor values. The stochastic element in the TreeBoost algorithm makes it highly resistant to over fitting. Cross-validation and random-row-sampling methods can be used to evaluate the generalization of a TreeBoost model and guard against over fitting. TreeBoost can be applied to regression models and k-class classification problems. TreeBoost can handle both continuous and categorical predictor and target variables.

DECISION TREE FOREST

A Decision Tree Forest is an ensemble of decision trees whose predictions are combined to make the overall prediction for the forest. A decision tree forest is similar to a TreeBoost model in the sense that a large number of trees are grown. However, TreeBoost generates a series of trees with the output of one tree going into the next tree in the series. In contrast, a decision tree forest grows a number of independent trees in parallel, and they do not interact until after all of them have been built.

Both TreeBoost and decision tree forests produce high accuracy models. Decision tree forests use the out of bag data rows for validation of the model. This provides an independent test without requiring a separate data set or holding back rows from the tree construction. The sophisticated and accurate method of surrogate splitters is used for handling missing predictor values. The stochastic element in the decision tree forest algorithm makes it highly resistant to over fitting. Decision tree forests can be applied to regression and classification models.

The primary disadvantage of decision tree forests is that the model is complex and cannot be visualized like a single tree. It is more of a “black box” like a neural network. Because of this, it is advisable to create both a single-tree and a decision tree forest model. The single-tree model can be studied to get an intuitive understanding of how the predictor variables relate, and the decision tree forest model can be used to score the data and generate highly accurate predictions.

IV. Analysis and Results

The measures in Table I were used for comparison purposes. These measures can be used as an indicator of the mean difference between the measured and estimated values.

Table I.
The Results of Applying Classification SVM and Decision Trees

Measures	Classification SVM		Forest Tree		Boosting Tree	
	Training	Validation	Training	Validation	Training	Validation
CV	0.796	0.922	0.852	0.958	0.922	0.988
NMSE	0.479	0.644	0.550	0.695	0.643	0.738
MSE	0.118	0.158	0.135	0.170	0.158	0.181
			145.12		155.11	
MAE	0.331	0.346	5	169.396	3	181.055
MAPE	32.870	37.414	37.675	39.542	47.142	53.658
Var%	36%	30%	42%	36%	69%	58%

Root mean squared error (RMSE) is a quadratic scoring rule which measures the average magnitude of the error. RMSE is the difference between predictions and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means that RMSE is most useful when large errors are particularly undesirable.

The normalized root mean squared error (NMSE) is the RMSE divided by the range of observed values; the value is often expressed as a percentage, where lower values indicate less residual variance.

Mean absolute error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. The MAE is the average over the verification sample of the absolute values of the differences between predictions and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

Proportion of variance (Var%) explained by model variables; this is the best single measure of how well the predicted values match the actual values. If the predicted values exactly match the actual values, then the model would explain 100% of the variance.

According to the previously discussed measures of C.V, NMSE, MSE, MAE and MAPE, the classification SVM was the optimum technique declared by the lowest values of 0.796, 0.479, 0.118, 0.331 and 32.870 respectively for the training set

among techniques applied and validated with values of 0.922, 0.644, 0.158, 0.346 and 37.414 respectively.

To determine the importance of the independent variables of the suggested model; the misclassification rate for the model using the actual data values for all predictors must be calculated. Then for each predictor, it randomly rearranges the values of the predictor and computes the misclassification rate for the model using the rearranged values. The difference between the misclassification rate with the correctly ordered values and the misclassification rate for the rearranged values is used as the measure of importance of the predictor. Table II shows the variables importance for both techniques.

TABLE II.
THE VARIABLES IMPORTANCE OF CLASSIFICATION SVM AND DECISION TREE

Variable	SVM	Forest Tree	Boosting Tree
Age	100	100	100
	22.46		
Variance	2	49.36	36.11
	11.04		
Circularity	4	45.758	34.152
	11.01		
Compactness	8	50.444	31.032
Eccentricity	8.72	35.175	17.044
Centroid X	7.66	40.385	23.212
Centroid Y	3.48	36.636	17.332
Area	1.429	36.503	18.607

The importance score for the most important predictor is scaled to a value of 100. Other predictors will have lower scores. Variance, Circularity, Compactness, Eccentricity, Centroid X, Centroid Y and Area are the less important variables for both SVM and Boosting tree, their importance ranges from 1.429 to 36.11 which is less than 40, then these variables could be ignored in the analysis [14] while Age, Compactness, Circularity and Centroid X possess the highest importance for the Forest Tree with importance of 100, 50.444, 49.36, 45.758 and 40.385 respectively.

The lift and gain is a useful tool for measuring the value of a predictive model. The basic idea of lift and gain is to sort the predicted target values in decreasing order of purity on some target category and then compare the proportion of cases with the category in each bin with the overall proportion. The lift and gain values show how much improvement the model provides in picking out the best of the cases. A gain chart displays cumulative percent of the target value on the vertical axis and cumulative percent of population on the horizontal axis. Cumulative gain is the ratio of the expected outcome using the model to prioritize the prospects divided by the expected outcome of randomization. The straight, diagonal line shows the expected return if no model is used for the population. The curved line shows the expected return using the model. The shaded area between the lines shows the improvement (gain) from the model. The gain of 1.00 means we are not doing any selective targeting.

Figures 3, 4and 5 exploits that by applying boosting tree we get 1.5718 times better outcome than the expected return of no model or randomization comparable to an average gain of 1.5576 for forest trees while the highest average gain is due to applying SVM of 1.7323.

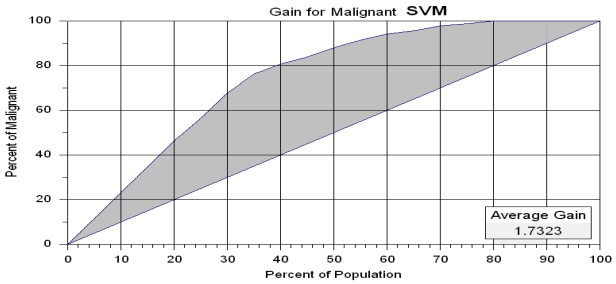

Fig. 3 The gain chart for classification SVM

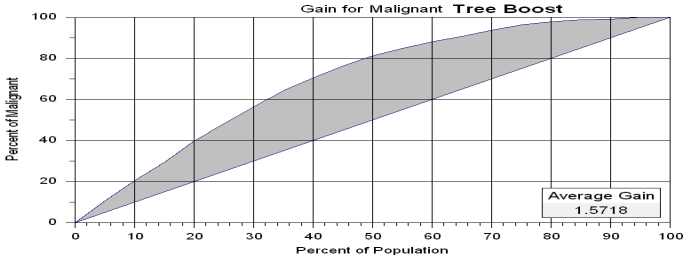

Fig. 4 The gain chart for boosting tree

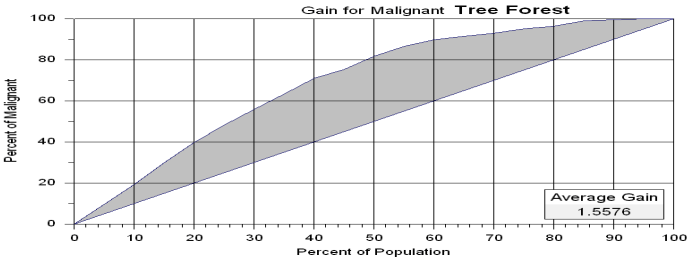

Fig. 5 The gain chart for forest tree

A receiver operating characteristic curve (ROC) curve is a graphical representation of the tradeoff between the false negative and false positive rates for every possible cutoff. Equivalently, the ROC curve is the representation of the tradeoffs between sensitivity and specificity. The area under the curve is a measure of test accuracy. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The slope of the tangent line at a cut point gives the likelihood ratio for that value of the test.

The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question. Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test.

The empirical estimate of area under the ROC curve (±SE) for the SVM was (0.79768±0.02762) while the binomial esti-

mate of area under the ROC curve was (0.85323±0.02537). Empirical estimate of area under the ROC curve of the boosting tree was (0.53452±0.03474) while the binomial estimate of area under the ROC curve was (0.53882±0.03900). Empirical estimate of area under the ROC curve of the forest tree was (0.57575±0.03439) while the binomial estimate of area under the ROC curve was (0.58548±0.03843).As discussed above and illustrated in fig. 6, SVM possess the highest area under the curve among the discussed techniques.

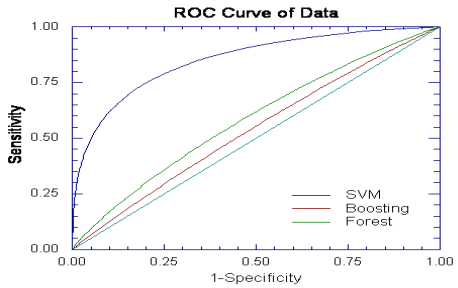

Fig. 6 The ROC for SVM, Forest Trees and Boosting Trees.

IV. Conclusion and Future Work

The results of applying the three classification techniques for extraction of the most important mammographic mass features showed promising and superior results for classification SVM over decision trees witnessed by minimal error measures ans maximum average gain.

Other risk factors could be used to aid the analysis such as genetic risk factors, previous breast radiation, and previous abnormal breast biopsy. Other proposed mammographic mass features includes: center of gravity, sphericity, inertia-shape, mean radius and max radius. Automated detection and classification of other types of mammographic lesions as micro calcifications and distorted architecture. The use of automated detection and segmentation techniques not only in mammographic images but also in MRI.

References

[1] Abdelaal, Medhat Mohamed , Muhamed Wael Farouq ,Hala Abou Sena, Abdel-Badeeh Mohamed Salem . *Using pattern recognition approach for providing second opinion of breast cancer diagnosis*. IEEE International Conference on Informatics and Systems, INFOS 2010.Cairo.Egypt.

[2] Abdelaal, Medhat Mohamed. *Application of regression support vector machine to estimate the fisheries parameters in Lake Nasser*. International Society for Business and Industrial Statistics Conference ISBIS-2010. Portoroz. Slovenia.

[3] Bernd Jahne, 2004, “Practical handbook on image processing for scientific and technical applications”, Florida, CRC press.

[4] Bernd Jahne. “Digital image processing”, Verlag Berlin, Heidelberg, Springer, 2002.

[5] Chang, C. and C. Lin, "A library for support vector machines". 2005, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[6] Cristianini, N. and J. Shawe-Taylor, "An introduction to support vector machines and other kernel-based learning methods". New York, NY: Cambridge University Press, 2000.

[7] Jadwign Rogowska. “Overview and fundamentals of medical image segmentation”, Handbook of Medical Imaging, San Diego, Academic Press, 2000

[8] John Terry. “Computer assisted screening of digital mammogram images”, The Department of Computer Science, University of Southern Mississippi, 2003

[9] Mohamed Sameti. 1998, “Detection of soft tissue abnormalities in mammographic images for early diagnosis of breast cancer”

[10] Monika Shinde, 2003, “Computer aided diagnosis in digital mammography”, Department of Computer Science and Engineering, College of Engineering, University of South Florida.

[11] Rafael C. Gonzalez, Richard E. Woods. “Digital image processing”, New Jersey, Prentice Hall, 2002.

[12] Rangaraj M. Rangayyan. Biomedical image analysis, Florida, CRC Press, 2005.

[13] Smola, A.J. and A. Scholkopf. "A tutorial on support vector regression". NeuroCOLT2 Technical Report NC2-TR- 1998-030.

[14] Thomas, D. R. Zhu, P. Decady, Y. J. *Point estimates and confidence intervals for variable importance in multiple linear regression*. Journal of Educational and Behavioral Statistics, 2007, Vol 32; Num B 1, pages 61-91.

[15] University of South Florida, Digital Mammography Home Page. http:marathon.csee.usf.edu/Mammography/Database.html.2009.