Feature Extraction and Classification For Breast Cancer Identification

Lokesh S
Meghashyam S
Harshitha C

Mrs.Sangeetha R
Batch No. - 14

May 21, 2019

Problem Statement / Definition

PROBLEM STATEMENT : **Feature Extraction and Classification of Micro-Calcification Cluster**

One of the most important stage in building a CAD scheme is the feature extraction and classification part.The features can be calculated from the ROI characteristics such as the size, shape, density and roundness.

The purpose of feature extraction is to reduce the data set by measuring some properties, or features that distinguish one input pattern from the other. The extracted features provide characteristics of the input type to the classifier considering the relevant description of the image.

The feature space is very large and complex due to the wide diversity the normal tissues and the variety of abnormalities.Using excessive features may degrade the performance of the algorithm and increase the complexity of the classifier. Some redundant features should be removed to improve the performance of the classifier.

Motivation of the Work

The radiological definition of cluster micro-calcification is the presence of more than three micro calcification in 1 cm2 area.A given cluster of micro calcification be associated with malignant and benign case.

Distinguishing between malignant and benign cluster is a difficult and time consuming task for radiologists. Only 20 percent to 30 percent biopsy cases recommended by the radiologists turn out to be malignant

Its of crucial importance to design the classification method n such a way to obtain high level of TPF(True Positive Fraction) while maintaining the FPF (false -positive fraction)at its minimum.It has been shown that computerised detection and classification thuds outperform radiologists detection and classification.

In addition, by using the results from the CAD(computer aided diagnosis) the performance of the radiologists can be increased.

Literature Survey
Summary of different methods

- Shape Features :- The shape features are also called morphological features.These kind of features are based on ROI
- Surround Region Dependence :- Four directional-weighted sum extracted form SRD matrix.
- Grey Scale Features :- Feature extraction is done by GLCM statistical method
- Fractal Dimensions :- Feature extraction by multi scale fractal dimensions
- Wavelet and Curvelet features : using texture analysis based on curvelet transform for the classification of the tissues

Literature Survey
Shape Features (from paper 1)

- 1.Eccentricity-feature: It is a scalar that describes the eccentricity of the ellipse which has the same second moment as the mass region. The eccentricity value is obtained from the ratio of the distance between the foci of this ellipse to its major axis length.
- 2.Solidity-feature: It is a scalar that describes the proportion of the pixels in the mass region to the pixels in the convex hull including the mass. It is computed as (Masss area)/ (Convex hulls area).
- 3. DifferenceArea-Hull-RectangularX: It is absolute value of convex hulls area minus convex RectangularXs area.
- 4. DifferenceArea-Mass-RectangularX: It is absolute val-ue of convex RectangularXs area minus masss area.
- 5. Cross-correlation-left: It is the cross correlation value between the RectangularX region and the Left-side region.
- 6. Cross-correlation-right: It is the cross correlation value between the RectangularX region and the Right-side region.

Literature Survey
Shape Features (from paper 1)

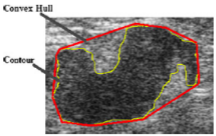


Fig. 4 An example of malignant mass with its Convex Hull- Yellow line is mass's boundaries, Red line is Convex Hull's boundaries

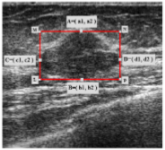


Fig. 5 An example of benign mass with its RectangularX region- Red line is RectangularX boundaries by (M, N, L, P) corner points; (A, B, C, D) are the most external points of the mass contour

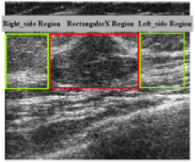


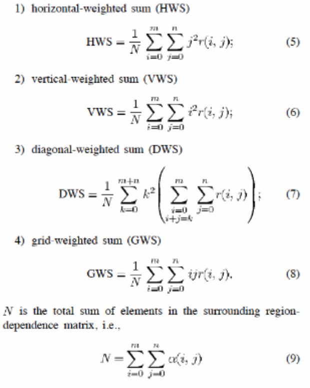
Fig. 6 An example of benign mass with Right_side, RectangularX, Left_side Regions

Literature Survey
Surround Region Dependence (from paper 2)

- The surrounding region-dependence matrix contains the texture information of an ROI
- The texture coarseness or fineness of an image can be interpreted as the distribution of the elements in the matrix
- If a texture is smooth the distribution of elements should be concentrated on or near upper left corner of the matrix.
- If a texture has fine details the distribution of elements should be spread out along the diagonal of the matrix
- The distribution of elements tends to spread to the right and/or lower right corner of the matrix for positive ROI's containing clustered microcalcifications.

Literature Survey
Surround Region Dependence (from paper 2)

- From the characteristics of the element distribution in the surrounding region- dependence matrix,the following 4 features are extracted



1) horizontal-weighted sum (HWS)

$$\mathrm{HWS} = \frac{1}{N} \sum_{i=0}^{m} \sum_{j=0}^{n} j^2 r(i, j); \qquad (5)$$

2) vertical-weighted sum (VWS)

$$\mathrm{VWS} = \frac{1}{N} \sum_{i=0}^{m} \sum_{j=0}^{n} i^2 r(i, j); \qquad (6)$$

3) diagonal-weighted sum (DWS)

$$\mathrm{DWS} = \frac{1}{N} \sum_{k=0}^{m+n} k^2 \left(\sum_{\substack{i=0 \\ i+j=k}}^{m} \sum_{j=0}^{n} r(i, j) \right); \qquad (7)$$

4) grid-weighted sum (GWS)

$$\mathrm{GWS} = \frac{1}{N} \sum_{i=0}^{m} \sum_{j=0}^{n} ij r(i, j). \qquad (8)$$

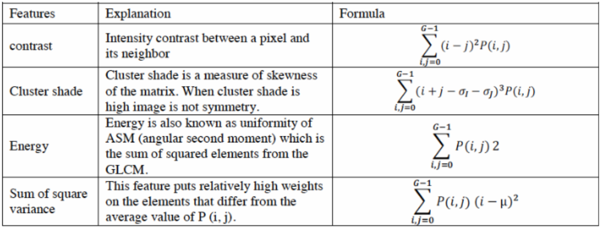
N is the total sum of elements in the surrounding region-dependence matrix, i.e.,

$$N = \sum_{i=0}^{m} \sum_{j=0}^{n} \alpha(i, j) \qquad (9)$$

Literature Survey
Grey Scale Features (from paper3)

- The texture features are extracted using GLCM(grey level co occurrence matrix)
- The GLCM is a two dimensional array which takes into account the specific position of a pixel relative to other pixels
- The GLCM is a tabulation of how often different combination of pixel brightness values occur in an image.
- The texture descriptors derived from GLCM are cluster shade, contrast, energy and sum of square variance.

Literature Survey
Grey Scale Features (from paper3)



Features	Explanation	Formula
contrast	Intensity contrast between a pixel and its neighbor	$\sum_{i,j=0}^{G-1} (i-j)^2 P(i,j)$
Cluster shade	Cluster shade is a measure of skewness of the matrix. When cluster shade is high image is not symmetry.	$\sum_{i,j=0}^{G-1} (i+j-\sigma_I-\sigma_J)^3 P(i,j)$
Energy	Energy is also known as uniformity of ASM (angular second moment) which is the sum of squared elements from the GLCM.	$\sum_{i,j=0}^{G-1} P(i,j)\,2$
Sum of square variance	This feature puts relatively high weights on the elements that differ from the average value of P (i, j).	$\sum_{i,j=0}^{G-1} P(i,j)\,(i-\mu)^2$

- P (i, j) represents the probability that two pixels with a specified separation have grey levels i and j.
- G is the number of grey level used

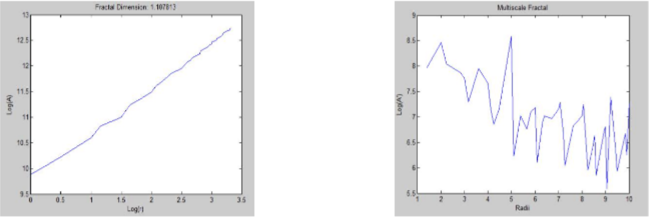
Literature Survey
Fractal Dimensions (from paper 4)

- A typical benign mass has a contour that is round, smooth, and well-defined, whereas a typical malignant tumor has a contour that is spiculated, rough, and ill-defined.
- The significant differences between the boundary shapes of benign masses and malignant tumors may be used to differentiate them by deriving shape factors. Therefore, the fractal dimension may be used to quantify the complexity of an objects boundary.
- The fractal dimension is computed using the ramp of the linear regression of the set of points from a graph log(r) vs logA(r)
  A(r) = total area represented by sum of all points as distance 'r' of dilation area.
- A(r) is obtained from the contour image by the Bouligand-Minkowski method.

Literature Survey
Fractal Dimensions (from paper 4)

- In this approach a multiscale analysis of fractal dimensions which, reveals the variations of the fractal dimensions of an object based on the variations in the scale of the metric space occupied by it.It requires the log(r) x logA‘(r) bi-log graph, where A‘ is the numerical derived point at (r,A(r))
- The Multiscale Fractal Dimension was calculated through the Differentiation by Finite Differences of the vectors of the areas of influence generated by the Bouligand-Minkowski method.
- From the multiscale signals (100*(r/max(r)) vs logA) the inflection points was extracted as a feature vector.

Literature Survey
Fractal Dimensions (from paper 4)



(a) (b)

Fig. 5. Multi fractal dimension method. (a) Fractal Dimension log(*r*) x log*A(r)* graph from the contour of the Fig. 3(d); (b) multiscale fractal dimension graph (100*(t/max(t)) X log(*A'*)).

Literature Survey
Wavelet and Curvlet features (from paper 5)

- Multiresolution allows a preservation of an image according to a certain levels of resolution. It allows as well a zooming in and out on the underlying texture structure. Therefore, the texture extraction is not affected by the size of the pixel neighborhood.
- Multiresolution analysis has been useful in so many applications from image compression to image de-noising and classification
- Some studies using curvelet transform in image processing have been done. The author presented a curvelet approach for the fusion of Magnetic Resonance (MR) and Computed Tomography (CT) images.
- The discrete curvelet transform was proposed by Candes and Donoho [10], from the idea of representing a curve as superposition of functions of various length and width obeying the curvelet scaling law width length2

Literature Survey
Wavelet and Curvlet features (from paper 5)

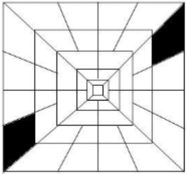


Figure 1. Curvelet tiling in the frequency domain, wedge samples are shaded.

Literature Survey
Hybrid Method (from paper 6)

- In this features from different categories are extracted such a brightness, contrast, shape and texture from micro calcification of the clusters
- Obtaining statistical moments of the gray level histogram of the region. The nth moment of statistics of gray level histogram can be obtained using the equation:

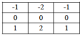
$$\mu_n(v) = \sum_{i=0}^{A-1} (v_i - m)^n \, p(v_i)$$

Literature Survey
Hybrid Method (from paper 6)

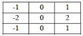
- Where v is a discrete random variable, p(vi) is an estimate of the probability of value vi (i=0,1,2,.A-1) and m is

$$m = \sum_{i=0}^{A-1} v_i \; p(v_i)$$

- Constructing gray level co-occurrence matrix. Co-occurrence matrix describes the texture by finding the occurrence of certain gray levels. Various descriptors can be obtained from the co-occurrence matrix including the maximum probability, entropy, uniformity.
- Applying Sobel edge detection mask. The Sobel edge detection mask are



-1	-2	-1
0	0	0
1	2	1



-1	0	1
-2	0	2
-1	0	1

Literature Survey
Comparison of Classifiers (from papers 7,8)

- Classification, a data mining task which assigns an object to one of several predefined categories based on the attributes of the object. Classification has been studied extensively in statistics, machine learning, neural networks and expert systems over decades. Several Classification methods:
  1. Decision tree algorithms
  2. Bayesian algorithms
  3. Multilayer Perceptron Classifier
  4. Support Vector Machine
  5. KNN Classifier

Literature Survey
Comparison of Classifiers (from papers 7,8)

- **Decision tree algorithms:** Frequently used decision tree algorithms are ID3,C4.5,and CART.CART (Classification and Regression Trees)provides better accuracy in classifying the breast cancer data sets than ID3,C4.5 algorithms.
  C4.5 Classifier : C4.5 is an algorithm used to generate a decision tree and these trees can be used for classification. C4.5 uses the concept of information gain to make a tree of classifactory decisions with respect to a previously chosen target classification. The output of the system is available as a symbolic rule base. The cases are scrutinized for patterns that allow the classes to be discriminated. These patterns are then expressed as models, in the form of decision trees, which can be used to classify new cases.Since for the real world databases the decision trees become huge and difficult to understand and interpret

Literature Survey
Comparison of Classifiers (from papers 7,8)

- **Nave Bayes Classifier:** A nave bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions It is based on two simplifying common assumptions: firstly, it assumes that the predictive attributes are conditionally independent given the class and secondly, the values of numeric attributes are normally distributed within each class. Bayes theorem is given as :
  p(C=c—X=x)= p(C=c)p(X=x—C=c)/p(X=x)
- **Multilayer Perceptron Classifier:** A multilayer perceptron is a forward artificial neural network model that maps sets of input data onto a setoff appropriate output. It has distinctive characteristics: 1. Model of neuron in the network includes a nonlinear activation function. 2. The network consists one or more layers of hidden neurons that are not part of input and output of the network. Error back-propagation algorithm is used in this type of classifier.

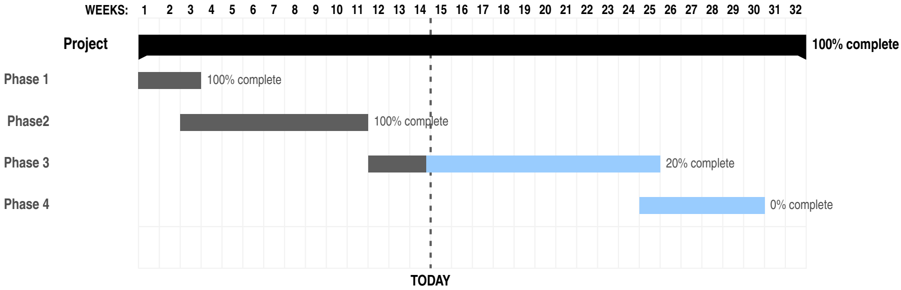
Literature Survey
Comparison of Classifiers (from papers 7,8)

- **Support Vector Machine :** SVM is a technique based on the statistical learning theory.It seperates two classes by determining the linear classifier that maximises the margin . According to the paper on comparitive study of classification by Haowen You and Gearge Rumbe, techniques accuracy of this classifier remained constant at 62.74 percen and progressively better predictions above 90 percent were determined. Initially in this method SVM was trained by varying gamma values and range was selected and classifier was evaluated including functions like polynomial, sigmoid and radial bias function.
- **KNN Classifier:** K Nearest Neighbour is a classifier which uses K value to calculate the collection of nearest data. Mostly the K value is selected as a odd one,but depending upon the various methods and techiques it can differ. The main principle of KNN theory is to count the Euclidian distance between the dots equal to K value, which is nearest to training value.

Requirements

- Hardware :-
- i7 7th gen processor,8gb ram,nvidia gtx 970 (VRAM)
- Software :-
- Python,opencv,numpy,scipy,pandas

Time line of completion of project from Aug 2018-April 2019(Gantt Charts).



- Phase 1 : Problem Statement Definition and Discussion of Project Abstract
- Phase 2 : Literature Survey and finalization of Methodology
- Phase 3 : Implementation
- Phase 4 : Testing and Further Development

References

1 Classification of Benign and Malignant Breast Masses Based on Shape and Texture Features in Sonography Images - Fahimeh Sadat Zakeri and Hamid Behnam and Nasrin Ahmadinejad [J Med Syst (2012) 36:16211627 DOI 10.1007/s10916-010-9624-7]

2 Statistical Textural Features for Detection of Microcalcifications in Digitized Mammograms - Jong Kook Kim and Hyun Wook Park* [IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 18, NO. 3, MARCH 1999]

3 Comparitive Study On Feature Extraction Method For Breast Cancer Classification - NITHYA,SANTHI [Journal of Theoretical and Applied Information Technology,30th November 2011,vol 33 No 2]

4 Automated Feature Extraction from Breast Masses using Multiscale Fractal Dimension - Jos Robson de Souza Filho, Carolina Yukari Veludo Watanabe [2017 Workshop of Computer Vision]

5 Curvelet Based Feature Extraction Method for Breast Cancer Diagnosis in Digital Mammogram -Mohamed Meselhy Eltoukhy, Ibrahima Faye, Brahim Belhaouari Samir [ResearchGate]

References

6 A Novel Approach for Detection of Breast Cancer at an early stage using Digital Image Processing techniques - Sangeetha R,Dr. Srikanta Murthy K

7 Competitive Study of Classification Techniques on Breast Cancer FNA Biopsy Data - Daniele Soria ,Jonathan M. Garibaldi,Elia Biganzoli,Ian O. Ellis [A Direct Path to Intelligent Tools]

8 A Comparison of Three Different Methods for Classification of Breast Cancer Data - Haowen You and George Rumbe [ResearchGate]

Methodology

- **Feature Extraction :**
  Many features such as brightness, contrast, size, shape and texture can be extracted from microcalcification clusters by using following methods:

1 Obtaining statistical moments of the gray level histogram of the region. The nth moment of statistics of gray level histogram can be obtained using the equation:

$$\mu_n(v) = \sum_{i=0}^{A-1} (v_i - m)^n \, p(v_i)$$

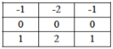
Methodology

- Where v is a discrete random variable, p(vi) is an estimate of the probability of value vi (i=0,1,2,.A-1) and m is

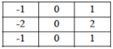
$$m = \sum_{i=0}^{A-1} v_i \ p(v_i)$$

2. Constructing gray level co-occurrence matrix. Cooccurrence matrix describes the texture by finding the occurrence of certain gray levels. Various descriptors can be obtained from the co-occurrence matrix including the maximum probability, entropy, uniformity.

3. Applying Sobel Edge Detection Mask.

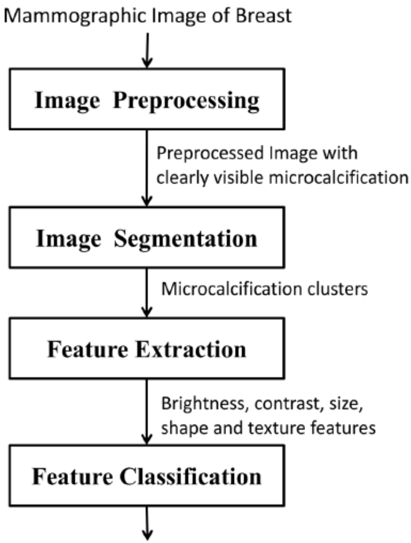


-1	-2	-1
0	0	0
1	2	1



-1	0	1
-2	0	2
-1	0	1

High Level Design



Mammographic Image of Breast

Image Preprocessing

Preprocessed Image with clearly visible microcalcification

Image Segmentation

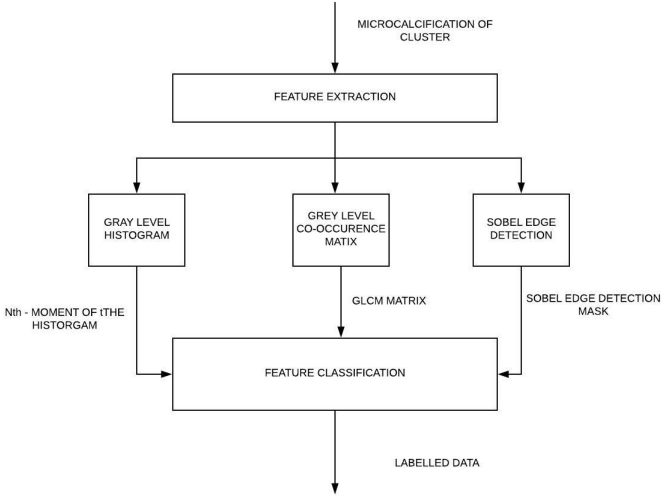
Microcalcification clusters

Feature Extraction

Brightness, contrast, size, shape and texture features

Feature Classification

Detailed Design


MICROCALCIFICATION OF CLUSTER
FEATURE EXTRACTION
GRAY LEVEL HISTOGRAM
GREY LEVEL CO-OCCURENCE MATIX
SOBEL EDGE DETECTION
Nth - MOMENT OF tTHE HISTORGAM
GLCM MATRIX
SOBEL EDGE DETECTION MASK
FEATURE CLASSIFICATION
LABELLED DATA

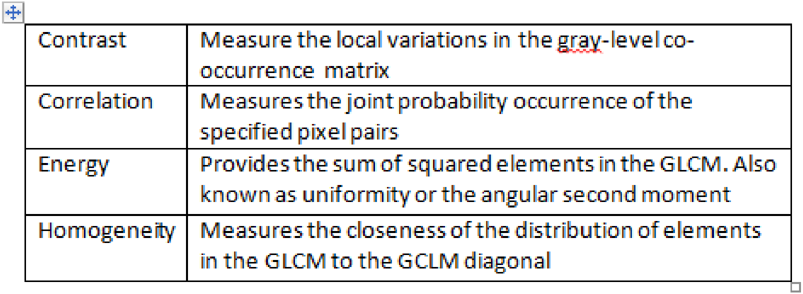
Implementation

- **Feature Extraction - GLCM**
- Feature extraction is very important part of pattern classification.GLCM is a histogram of co-occuring greyscale values at agiven offset over an image. GLCM is obtained by summing up all co-occurences of grey scale values at aspecified offset over an image.
- The specified offsets are:

1. Distance-distance between two pixelsof interest.The distances ranges from 1 to 30.
2. Angle-The directions in which the distances are compared, that is horizontal(0deg), vertical(90deg), and diagonal(45deg).

Implementation

- Four statistical measures such as correlation, energy, contrast, homogeneity are computed based on GLCM



Contrast	Measure the local variations in the gray-level co-occurrence matrix
Correlation	Measures the joint probability occurrence of the specified pixel pairs
Energy	Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment
Homogeneity	Measures the closeness of the distribution of elements in the GLCM to the GCLM diagonal

- For each angle, top 4 distances that achieve the best results are selected.Hence a combination of 12 pairs of distances and angle is choosen.
- For these 12 pairs, the above statistical measures are calculated.

Implementation

- **Gray Level Histogram**

One of the simplest approaches for describing texture is to use the statistical moments of the gray level histogram of the image

The histogram-based features used in this work are nth order statistics that include mean, variance, skewness and kurtosis. Let v be a random variable denoting image gray levels and p(vi) is an estimate of the probability of value vi, i = 0,1,2,3,.A-1, be the corresponding histogram, where A is the number of distinct gray levels.

Mean: The mean gives the average gray level of each region and it is useful only as a rough idea of intensity not really texture.

$$m = \sum_{i=0}^{A-1} v_i \, p(v_i)$$

Implementation

Variance: The variance gives the amount of gray level fluctuations from the mean gray level value.

Skewness: Skewness is a measure of the asymmetry of the gray levels 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.

Kurtosis: Kurtosis is a measure of how outlier-prone a distribution is. It describes the shape of the tail of the histogram.

The nth moment of statistics of gray level histogram can be obtained using the equation:

$$\mu_n(v) = \sum_{i=0}^{A-1} (v_i - m)^n \, p(v_i)$$

Implementation

- **Feature Classification - Random Forest Classifier**

  Random forest classifier creates a set of decision trees from randomly selected subset of training set and gets prediction from each tree and selects the best solution by means of voting.

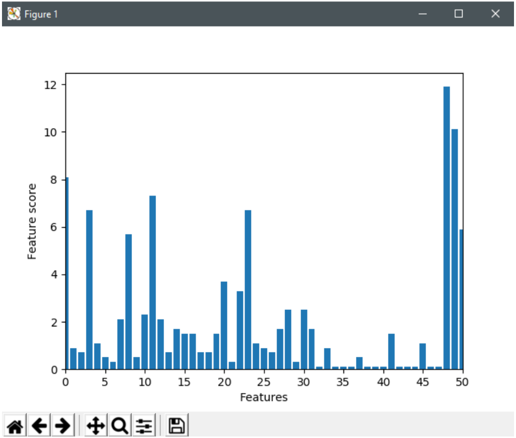
  - No overfitting problem
  - Highly accurate and robust method

  It works in four steps:

  1. Select random samples from a given dataset.
  2. Construct a decision tree for each sample and get a prediction result from each decision tree.
  3. Perform a vote for each predicted result.
  4. Select the prediction result with the most votes as the final prediction.

Results

- The final feature set is constructed which is fed as input to random forest classifier which classifies test data into malignant or benign using selected features.



Results

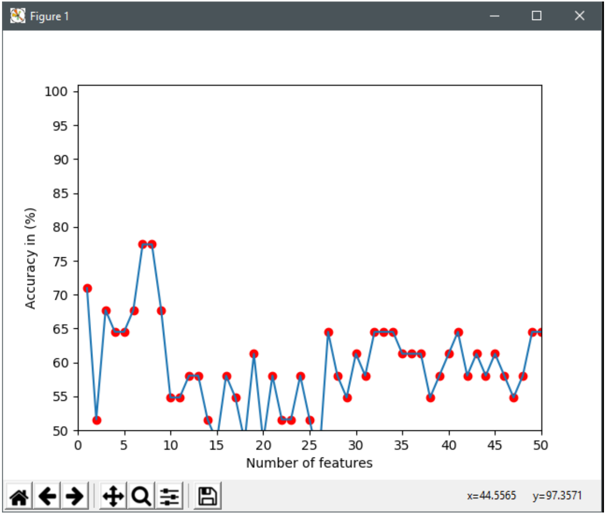
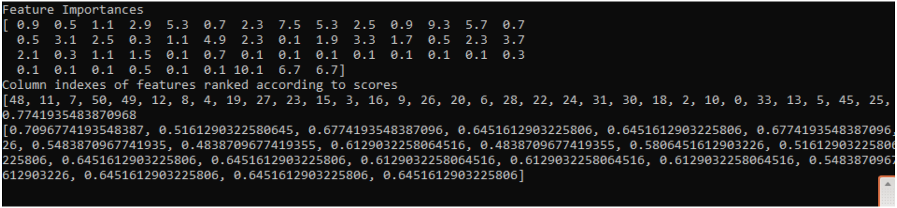


Figure 1
Accuracy in (%)
100
95
90
85
80
75
70
65
60
55
50
0
5
10
15
20
25
30
35
40
45
50
Number of features
x=44.5565 y=97.3571

Results



```
Feature Importances
[ 0.9  0.5  1.1  2.9  5.3  0.7  2.3  7.5  5.3  2.5  0.9  9.3  5.7  0.7
  0.5  3.1  2.5  0.3  1.1  4.9  2.3  0.1  1.9  3.3  1.7  0.5  2.3  3.7
  2.1  0.3  1.1  1.5  0.1  0.7  0.1  0.1  0.1  0.1  0.1  0.1  0.1  0.3
  0.1  0.1  0.1  0.5  0.1  0.1 10.1  6.7  6.7]
Column indexes of features ranked according to scores
[48, 11, 7, 50, 49, 12, 8, 4, 19, 27, 23, 15, 3, 16, 9, 26, 20, 6, 28, 22, 24, 31, 30, 18, 2, 10, 0, 33, 13, 5, 45, 25,
0.7741935483870968
[0.7096774193548387, 0.5161290322580645, 0.6774193548387096, 0.6451612903225806, 0.6451612903225806, 0.6774193548387096,
26, 0.5483870967741935, 0.4838709677419355, 0.6129032258064516, 0.4838709677419355, 0.5806451612903226, 0.51612903225806
225806, 0.6451612903225806, 0.6451612903225806, 0.6129032258064516, 0.6129032258064516, 0.6129032258064516, 0.5483870967
612903226, 0.6451612903225806, 0.6451612903225806, 0.6451612903225806]
```