Filomat 34:1 (2020), 19–34
https://doi.org/10.2298/FIL2001019F

Published by Faculty of Sciences and Mathematics,
University of Niš, Serbia
Available at: http://www.pmf.ni.ac.rs/filomat

Breast Cancer Classification Based on Improved Rough Set Theory Feature Selection

R. M. Farouk[a], Heba I. Mustafa[a], Abd Elmounem Ali[a]

[a]*Mathematics Department, Faculty of Science, Zagazig University, Egypt*

Abstract. Breast cancer is one of the leading causes of death among the women. Mammogram analysis is the most effective method that helps in the early detection of breast cancer. In this paper we have made an attempt to classify the breast tissue based on Statistical features of a mammogram which extracted using simple image processing techniques with rough set theory. The proposed scheme uses texture models to capture the mammographic appearance within the breast. The statistical features extracted are the mean, standard deviation, smoothness, third moment, uniformity and entropy which signify the important texture features of breast tissue. Based on the values of these features of a digital mammogram, we have made an attempt to classify the breast tissue in to three basic categories normal, benign, and malignant given in the data base (mini-MIAS database). This categorization would help a radiologist to detect a normal breast from a cancer affected breast. Rough set theory can be regarded as a new mathematical tool for imperfect data analysis. Rough set based data analysis starts from a data table called a decision table. Each row of a decision table induces a decision rule, which specifies decision (action, results, outcome, etc.). We can know important data rules by using core and reduct which elimination of duplicate rows and elimination of superfluous values of attributes.

1. Introduction

Nowadays breast cancer detection and diagnosis is very important because the treatment is based on radio surgery. Breast cancer is the second leading cause of cancer death among women in the 40-55 ages[13]. Mammography is a key screening tool for breast abnormalities detection, because it allows identification of tumor before being palpable. Interpretations of mammographic images are difficult for the radiologists because of poor quality and noises in the images[1]. The proposed system is a novel scheme with rough set theory for abnormality detection and segmentation in Mammograms images, allows high degree of automation with less computation speed.

The breast is made up of many different types of cells. Breast cancers occur when one type of cell transforms from its normal characteristics and multiplies in an abnormal way. The extra cells form a mass of tissue called a tumor. Tumors are benign or malignant[12]. Many of them are benign and can be successfully removed. Malignant primary breast tumors cause problems by spreading into the normal tissue causing damage to the surrounding areas of the breast. According to the recent research results of

2010 *Mathematics Subject Classification*. Primary 62P10; Secondary 68P10, 68U10
Keywords. Breast cancer classification, Rough Set, Feature selection, Attribute reduction
Received: 29 December 2018; Accepted: 30 August 2019
Communicated by Biljana Popović
Email addresses: rmfarouk1@yahoo.com (R. M. Farouk), dr_heba_ibrahim@yahoo.com (Heba I. Mustafa), abdualimohammed88@gmail.com (Abd Elmounem Ali)

University of Calgary, (Biomedical Engineering Research Group), these have been further classified based on some statistical features. This classification would help a radiologist to determine the breast anatomy (fibro glandular tissue) affected due to Estrogen secretion[2]. Rough set is founded on the assumption that with every object of the universe of discourse some information (data, knowledge) is associated. Objects characterized by the same information are indiscernible (similar) in view of the available information about them. Any set of all indiscernible (similar) objects is called an elementary set, and forms a basic atom of knowledge about the universe. Any union of some elementary sets is referred to as a crisp (precise) set otherwise the set is rough (imprecise, vague). Each rough set has boundary-line cases, i.e. objects which cannot be with certainty classified, by employing the available knowledge, as members of the set or its complement. Obviously rough sets, in contrast to precise sets, cannot be characterized in terms of information about their elements. With any rough set a pair of precise sets, called the lower and the upper approximation of the rough set, is associated. The lower approximation consists of all objects which surely belong to the set and the upper approximation contains all objects which possibly belong to the set. The difference between the upper and the lower approximation constitutes the boundary region of the rough set. Approximations are fundamental concepts of rough set theory. Rough set based data analysis starts from a data table called a decision table, columns of which are labeled by attributes, rows by objects of interest and entries of the table are attribute values. Attributes of the decision table are divided into two disjoint groups called condition and decision attributes, respectively. Each row of a decision table induces a decision rule, which specifies decision (action, results, outcome, etc.) if some conditions are satisfied. If a decision rule uniquely determines decision in terms of conditions the decision rule is certain. Otherwise the decision rule is uncertain. Decision rules are closely connected with approximations[3].

In previous study Mohabey et al. (2000) proposed an interesting strategy for the color image segmentation using the rough set theory [21]. Pal et al. (2002) proposed an integration of the Expectation Maximization (EM) algorithm, rough-set theoretic knowledge Extraction and minimal spanning tree (MST) clustering [22]. Aboul Ella Hassanien (2006) proposed Feature Extraction and Rule Classification Algorithm of Digital Mammography based on Rough Set Theory [23]. Cai et al. (2007) A modified classic fuzzy C-means (FCM) algorithm based on Rough set for image segmentation [24]. Jirava pavel, krupka Jiri (2007) classification model based on rough set and fuzzy sets theory [25]. Jiang et al. (2008) proposed a new method for the image segmentation based on the rough set theory and neural networks [26]. Halder et al. (2012) Proposed a rough set approach for a gray scale image segmentation [27]. Anupama et al. (2013) Rough sets based clustering is analyzed with respect to K-means and Fuzzy C-means algorithms for MRI images of Brain Web database [28]. Reddy et al. (2013) proposed an algorithm which was based on the modified K-means clustering by using rough set theory [17].

2. Problem formulation

We have made an attempt to classify the breast tissue based on features of a mammogram using simple image processing techniques. We have used texture models to capture the mammographic appearance within the breast. Different features can be extracted for a digital mammogram as: Texture Features, Statistical Features and Structural Features. We have used MATLAB for extracting the tumors from input mammogram and for calculating various features. Based on the values of these features of a digital mammogram, breast tissue can be classified. This categorization would help a radiologist to detect a normal breast from a cancer affected breast so as to proceed with further investigation [4].

3. Our technique

The main goals are to perform the following operations:

- Input data
- Detection
- Feature Extraction

- Decision table
- Rough set rules
- Classification
- Segmentation

This forms a basic step in the detection of abnormal breast under computer aided detection system

3.1. Input image

The studying of medical imaging applications plays an important role. In this study, we use the Mini MIAS database that contains 320 images. This dataset consists of 12 class's defined by the Breast Imaging Reporting and Data System (BI-RADS). There are four breast tissue classes (fatty, fibroglandular, heterogeneously dense, and extremely dense) and three health status classes (normal, benign tumor, and malignant tumor) for each breast tissue type[5]. There are 206 normal images, 63 benign and 51 malign (which are considered abnormal)[14].

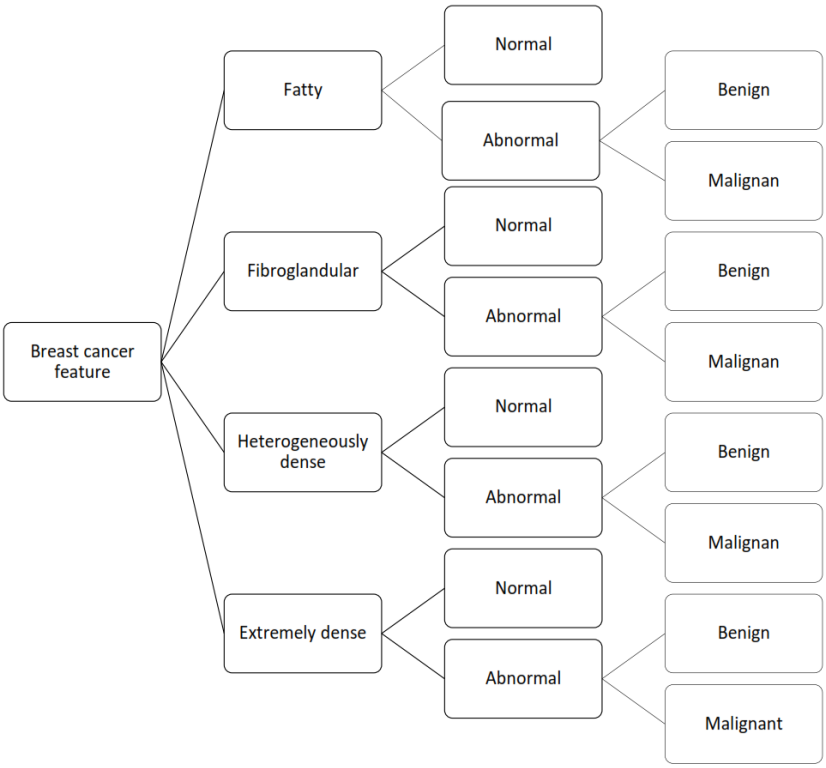

Figure 1: Flowchart designed for classify breast study.

3.2. Image processing

The preprocessing of digital mammograms refers to the enhancement of mammograms intensity and contrast manipulation, noise reduction, filtering, etc.

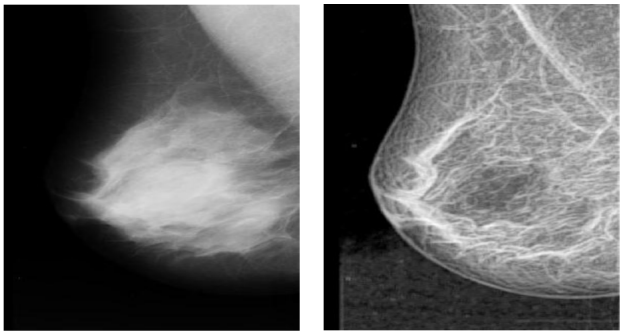

Figure 2: Image processing filter (a) image before filter. (b)image after filter.

3.3. Segmentation

The methods used to separate the region of interest from the background are usually referred as the segmentation process[6]. Segmentation can be carried out using any of the standard techniques like local thresholding, K–means clustering [15].

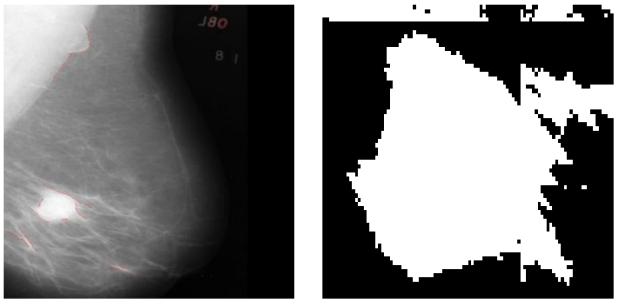

Figure 3: Segmentation image.

3.4. Feature Extraction

In our work, we have extracted Texture features, Statistical features and Structural Features for the segmented tumor from the given input mammogram image[7]. The features that we have extracted are:

Table 1: Contrast group mathematical representations.

no	moment	expression	Description
1	Contrast=inertia=sum of squares variance	$cont = \sum_{i,j} \|i-j\|^2 p(i,j)$	A measure of the image contrast or the amount of local variations present in an image.
2	Inverse Difference Moment (IDM) =Homogeneity	$idm = \sum_{i,j} \frac{p(i,j)}{1+\|i-j\|^2}$	This descriptor has large values in cases where the largest elements in P are along the principal diagonal.
3	Dissimilarity=Directional moment	$dissim = \sum_{i,j} \|i-j\| p(i,j)$	A measure of the image contrast that increase linearly not exponentially.

Table 2: Measures related to orderliness mathematical representations.

no	moment	expression	Description
1	Angular second moment(ASM)	$Asm = \sum_{i,j}^{n} p(i,j)^2$	A measure of the homogeneity of an image. Hence it is a suitable measure for detection of disorders in textures. For homogeneous textures value of angular second moment turns out to be small compared to non-homogeneous ones.
2	Energy=Uniformity	$Ene = \sqrt{Asm}$	Energy returns the sum of squared elements in the Grey Level Co-occurrence Matrix (GLCM). Energy is also known as uniformity. The range of energy is [0 1].
3	min probability	$Minimum\{Xi\|i = 1,2,3,....,N\}$	This is simply the largest entry in the matrix.
4	max probability	$Maximum\{Xi\|i = 1,2,3,....,N\}$	This is simply the smallest entry in the matrix.
5	Entropy	$Ent = -\sum_{i=1}^{n} p(i)\log_2 p(i)$	Entropy is a measure of information content. It measures the randomness of intensity distribution.

Table 3: Descriptive statistics group mathematical representations.

no	moment	expression	Description
1	Mean	$\mu = \sum_{i,j}^{n} p(i,j)$	The mean, m of the pixel values in the defined window, estimates the value in the image in which central clustering occurs.
2	Standard Deviation	$\sigma = \sqrt{\sum_{i=1}^{n}(p(i)-\mu)^2}$	The Standard Deviation, σ is the estimate of the mean square deviation of grey pixel value p (i, j) from its mean value Standard deviation describes the dispersion within a local region.
3	Variance	$var = \sqrt{\sigma}$	Variance is the square root of standard deviation.
4	Area descriptor	$ADes = \frac{\sigma}{\mu}$	Area descriptor is Outside the division between standard deviation and mean.
5	Skewness	$skew = \sum_{i=1}^{n}(\frac{p(i)-\mu}{\sigma})^3$	Characterizes the degree of asymmetry of a pixel distribution in the specified window around its mean. Skewness is a pure number that characterizes only the shape of the distribution.
6	Kurtosis	$kurt = \sum_{i=1}^{n}(\frac{p(i)-\mu}{\sigma})^4$	Measures the Peakness or flatness of a distribution relative to a normal distribution.
7	Fourth moment	$M4 = \sum_{i=1}^{n}(p(i)-\mu)^4$	Measures the Peakness of a distribution relative to a normal distribution exponentially (4).
8	Third moment	$M3 = \sum_{i=1}^{n}(p(i)-\mu)^3$	Measures the Peakness of a distribution relative to a normal distribution exponentially (3).
9	Mean energy	$\mu_{energy} = \frac{1}{n}\sum_{i=1}^{n} p(i)^2$	Measures the men energy intensity in the histogram.
10	Energy variance	$var_{energy} = \frac{1}{n-1}\sum_{i=1}^{n}(p(i)^2 - \mu_{energy})^2$	Measures the Energy variance intensity in a region.
11	Correlation	$corr = \sum_{i,j}\frac{(1-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i\sigma_j}$	Correlation returns a measure of how correlated a pixel is to its neighbor over the whole image. The range of correlation is [-1 1].Correlation is 1 or -1 for a perfectly positively or negatively correlated image. Correlation is Nan (Not a Number) for a constant image.
12	Smoothness	$smo = 1 - \frac{1}{1+\sigma^2}$	A measure of grey level contrast that can be used to establish descriptors of relative smoothness.
13	Root Mean square(RMS)	$rms = \sqrt{\frac{\sum_{i=1}^{n}\|p(i)\|^2}{n}}$	The RMS (Root Mean Square) computes the RMS value of each row or column of the input, along vectors of a specified dimension of the input, or of the entire input
14	Similarity	$sim = \sum_{i=1}\frac{p(i,j)}{1+\|i-j\|}$	A first degree measure that increases with less contrast.

3.5. Decision table

A decision table specifies what decisions (actions) should be undertaken when some conditions are satisfied. Most decision problems can be formulated employing decision table formalism; therefore, this tool is particularly useful in decision making. A decision table has two subsets of attributes, called condition C and decision attributes D respectively. Decision table will be denoted $T = (U, A, C, D)$.A decision table specifies what decisions (actions) should be undertaken when some conditions are satisfied. Most decision problems can be formulated employing decision table formalism; therefore, this tool is particularly useful in decision making. A decision table has two subsets of attributes, called condition C and decision attributes D respectively. Decision table will be denoted $T = (U, A, C, D)$.

Table 4: General form of decision table.

U	C				D		
	s_1	s_2	...	s_m	d_1	...	d_k
x_1	f_{11}	f_{12}	...	f_{1m}	g_{11}	...	g_{1k}
x_2	f_{21}	f_{22}	...	f_{2m}	g_{21}	...	g_{2k}
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_n	f_{n1}	f_{n2}	...	f_{nm}	g_{n1}	...	g_{nk}

3.6. Rough set

The idea of rough set was publicized by Professor Zdzislaw Pawlak, a Polish Mathematician, in 1982, for the development of automatic rule generation systems[8]. Using Rough set theory distorted regions of tumor can be detected. Rough set theory can be characterized as a new mathematical tool for imperfect data analysis. The theory has found applications in many domains, such as decision support, engineering, medicine and others[3] .

3.6.1. Fundamental concepts in rough sets

Classical rough set theory often requires a large amount of labeled data, in which concepts information system, indiscernibility, approximation and attribute reduction are key issues.

1. Information system
   The information system T can be expressed as a quadruple $T = (U, A, C, D)$, where U is the set of entities (universe). A is the attribute set. If set A can be divided into conditional attribute set C and decisional attribute set D, i.e., $C \bigcup D = A$ and $C \bigcap D = \phi$.
2. Indiscernibility
   If $P \subseteq R$ and $P \neq \phi$, then $\bigcap P$ (intersection of all equivalence relations belonging to P) is also an equivalence relation, and will be denoted by $IND(P)$, and will be called an indiscernibility relation over P . Moreover
   $$[X]IND(P) = \bigcap_{(R \in P)} [x]R$$
   So $U/IND(P)$ (i.e. the family of all equivalence classes of the equivalence relation $IND(P)$).
3. Approximation of set
   Assume that $X, Y \in U$ and R is the equivalent relationship defined on the universe U . The lower-approximation set of the set X on R is defined as

$$\underline{\mathfrak{R}(x)} = \bigcup \{Y \in U/R : Y \subseteq X\} \tag{1}$$

   Where $\underline{\mathfrak{R}(x)}$ is the maximum set of entities that for sure belong to the set X, also referred to as positive region, denoted by $POS(X)$. Similarly, the upper-approximation set of the set X on R is defined as

$$\overline{\mathfrak{R}(x)} = \bigcup \{Y \in U/R : Y \bigcap R \neq \phi\} \tag{2}$$

Where ϕ is an empty set and $\overline{\mathfrak{R}(x)}$ is the minimum set of entities which possibly belong to set X . Based on the above definitions, the boundary set can be defined as

$$BND(X) = \overline{\mathfrak{R}(x)} - \underline{\mathfrak{R}(x)} \tag{3}$$

If $BND(X)$ is an empty set, the set X reduces to a crisp set on R . If on the other hand, $BND(X)$ is non-empty; X is a rough set on R .

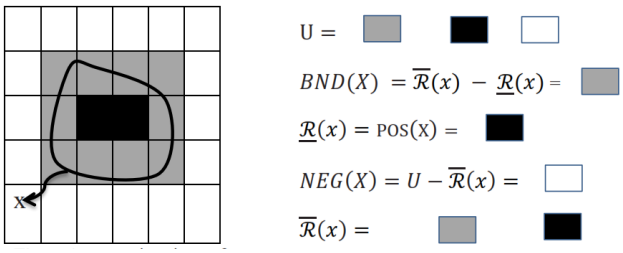

Figure 4: Approximations of set.

4. Attribute reduction
   $Q \subseteq P$ is a reduct of P if Q is independent and $IND(Q) = IND(P)$ .Obviously P may have many reducts[9]. The set of all indispensable relations in P will be called the core of P , and will be denoted $CORE(P)$.

$$CORE(P) = \bigcap RED(P) \tag{4}$$

   Where $RED(P)$ is the family of all reducts of P. Let R and Q be equivalence relations over U, R-positive region of Q , denoted $POS_R(Q)$ we understand the set

$$POS_R(Q) = \bigcup_{(x \in U/Q)} \underline{\mathfrak{R}(x)} \tag{5}$$

   R- Positive region of Q is the set of all object of universe U which can be classified to class of U/Q.

3.6.2. Accuracy measure

Let $K = (U, \mathfrak{R})$ be the knowledge base and $R, Q \subset \mathfrak{R}$. We say Q depend in a degree k from knowledge R if and only if

$$K = \gamma_{R(Q)} = \frac{cardPOS_R(Q)}{cardU} \tag{6}$$

Where card denotes cardinality of the set statistics[18]. We have three phases

1. if $k = 1$, Q totally depends from R , and then all elements of the universe can be classified.
2. If $0 < k < 1$, we say Q roughly depend from R , and then only elements that belong to positive region can be classified.
3. if $k = 0$, we say Q is totally independent from R, and then none of the elements of the universe can be classified.

3.7. Classification

Classification aim at simplifying the uncertain decision table and generating more significant decision rules to classify unseen objects[10]. K-nearest neighbor classification (KNN) is a classification model in which you can alter both the distance metric and the number of nearest neighbors. Because a Classification KNN classifier stores training data, you can use the model to compute resubstitution predictions. Classification KNN performance is directly related to the parameter k , there is no obvious information on the selection of k except that it should be positive and not a multiple of the total number of classes[16]. Alternatively, use the model to classify new observations using the predict method[11].

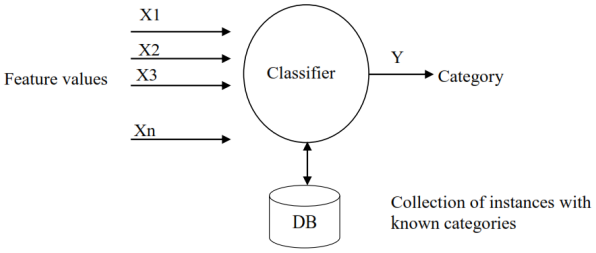

Figure 5: Classifier model.

The presented algorithm was implemented as a toolbox in MATLAB environment. This instrument is further applied

Algorithm 1 classification and segmentation data

Input: breast cancer images from our data base
Output: get new classification of breast cancer and segmentation
procedure GET IMAGE FROM DATA SET, THEN CALCULATE FEATURE EXTRACTION
  for each item `Group no.1 feature extractions` **do**
    `Create matrix` T `called decision table`
    From decision table apply reduct to remove repeat data and overflow values from *Eq*.4
    `Compute lower and upper approximations` *Eq*.1, 2
    `Calculate value of k from` *Eq*.6
    if $0 < k \leq 1$ **then**
      `elements that belong to positive region can be classified`
    else
      `none of the elements of the universe can be classified`
      `calculate new feature extraction`
    end if
    `Apply KNN classifier`
    if `image normal` **then**
      `show it`
    else
      `compute canny segmentation method for abnormal images`
    end if
    `Show result data`
    `compute Accuracy from` *Eq*.7
  end for
  return *Accuracy*
end procedure

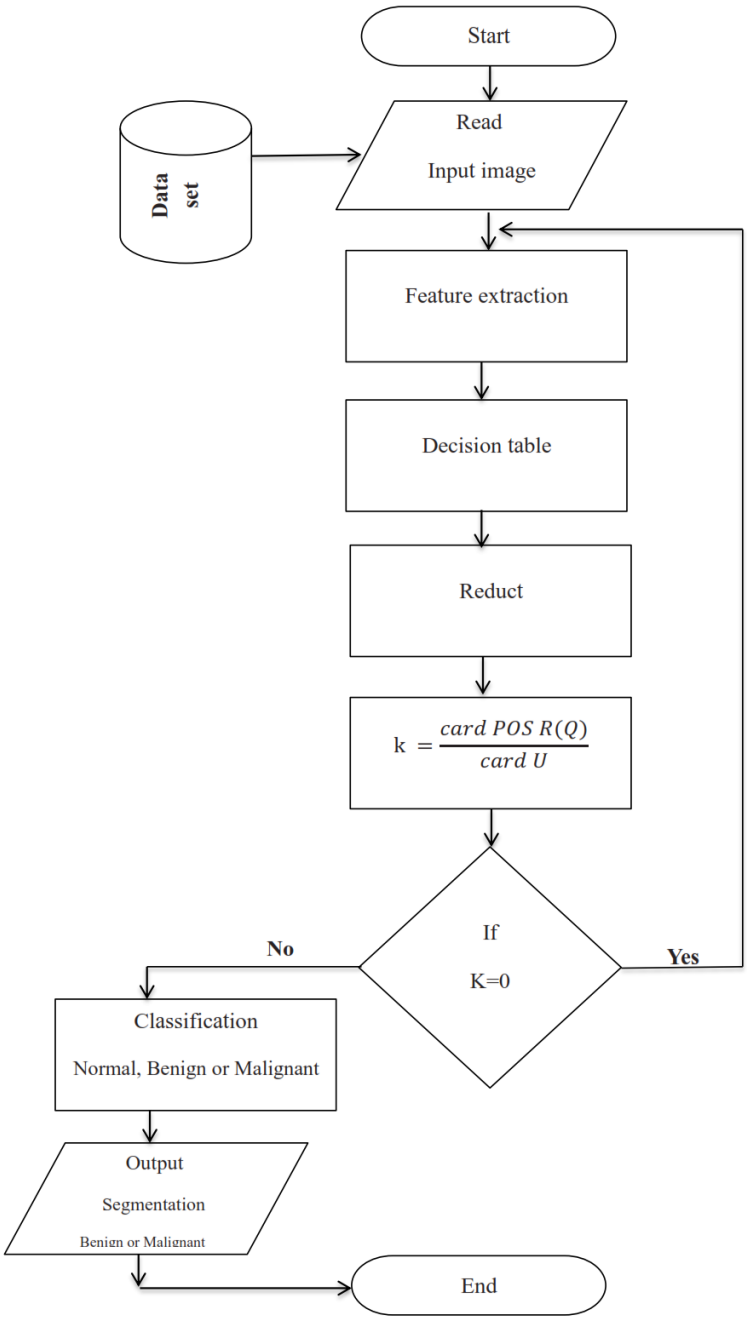

Figure 6: Shows the sequences of steps in which classification occur.

4. Numerical results

We have used MATLAB for extracting the tumors from input mammogram and for calculating various features.

First: Numerical results for measures related to orderliness data set

Table 5: Measures related to orderliness data set.

Image Samples	Uniformity	Energy sqrt(uniformity)	Entropy	max probability	min probability
mam1	0.3401	0.5832	0.40196	0.5814	0.0038147
mam2	0.2487	0.4987	0.47658	0.4963	0.042915
mam3	0.2379	0.4877	0.45951	0.4819	0.0019073
mam4	0.2044	0.4521	0.49022	0.4457	0.062943
mam5	0.1734	0.4164	0.51761	0.412	0.068665
mam6	0.181	0.4254	0.49559	0.4193	0.10395
mam7	0.2684	0.5181	0.44958	0.5158	0.005722
mam8	0.2058	0.4537	0.49878	0.4505	0.00095367
mam9	0.1562	0.3952	0.50232	0.3816	0.060081
mam10	0.3507	0.5922	0.39049	0.5906	0.01049
mam11	0.191	0.4370	0.46683	0.4268	0.044823
mam12	0.1424	0.3774	0.53311	0.3682	0.00095367
mam13	0.3358	0.5795	0.40655	0.5781	0.0047684
mam14	0.2718	0.5213	0.45606	0.5195	0.005722
mam15	0.3619	0.6016	0.36236	0.5985	0.005722
mam16	0.3873	0.6223	0.37094	0.6214	0.014305
mam17	0.4467	0.6684	0.3063	0.6648	0.00095367
mam18	0.382	0.6181	0.32479	0.6097	0.00095367
mam19	0.2073	0.4553	0.50349	0.4528	0.024796
mam20	0.199	0.4461	0.49606	0.4423	0.02861
mam21	0.1434	0.3787	0.52569	0.3653	0.019073
mam22	0.2634	0.5132	0.4379	0.5085	0.002861
mam23	0.145	0.3808	0.51044	0.3654	0.0047684
mam24	0.1458	0.3818	0.50057	0.3652	0.00095367
mam25	0.1701	0.4124	0.54099	0.4096	0.00095367
mam26	0.1908	0.4368	0.52046	0.4342	0.0019073
mam27	0.1388	0.3726	0.53774	0.3652	0.00095367
mam28	0.137	0.3701	0.55805	0.3652	0.0047684
mam29	0.1478	0.3844	0.4855	0.3652	0.00095367
mam30	0.2564	0.5064	0.426977	0.5	0.00095367

Table 6: Measures related to orderliness decision table.

U	C					D
	Uniformity	Energy sqrt(uniformity)	Entropy	max probability	min probability	decision
mam1	0.3401	0.5832	0.40196	0.5814	0.0038147	1
mam2	0.2487	0.4987	0.47658	0.4963	0.042915	1
mam3	0.2379	0.4877	0.45951	0.4819	0.0019073	0
mam4	0.2044	0.4521	0.49022	0.4457	0.062943	0
mam5	0.1734	0.4164	0.51761	0.412	0.068665	1
mam6	0.181	0.4254	0.49559	0.4193	0.10395	0
mam7	0.2684	0.5181	0.44958	0.5158	0.005722	0
mam8	0.2058	0.4537	0.49878	0.4505	0.00095367	0
mam9	0.1562	0.3952	0.50232	0.3816	0.060081	0
mam10	0.3507	0.5922	0.39049	0.5906	0.01049	1
mam11	0.191	0.4370	0.46683	0.4268	0.044823	0
mam12	0.1424	0.3774	0.53311	0.3682	0.00095367	1
mam13	0.3358	0.5795	0.40655	0.5781	0.0047684	1
mam14	0.2718	0.5213	0.45606	0.5195	0.005722	0
mam15	0.3619	0.6016	0.36236	0.5985	0.005722	1
mam16	0.3873	0.6223	0.37094	0.6214	0.014305	0
mam17	0.4467	0.6684	0.3063	0.6648	0.00095367	1
mam18	0.382	0.6181	0.32479	0.6097	0.00095367	0
mam19	0.2073	0.4553	0.50349	0.4528	0.024796	1
mam20	0.199	0.4461	0.49606	0.4423	0.02861	0
mam21	0.1434	0.3787	0.52569	0.3653	0.019073	1
mam22	0.2634	0.5132	0.4379	0.5085	0.002861	0
mam23	0.145	0.3808	0.51044	0.3654	0.0047684	2
mam24	0.1458	0.3818	0.50057	0.3652	0.00095367	0
mam25	0.1701	0.4124	0.54099	0.4096	0.00095367	1
mam26	0.1908	0.4368	0.52046	0.4342	0.0019073	0
mam27	0.1388	0.3726	0.53774	0.3652	0.00095367	0
mam28	0.137	0.3701	0.55805	0.3652	0.0047684	2
mam29	0.1478	0.3844	0.4855	0.3652	0.00095367	0
mam30	0.2564	0.5064	0.426977	0.5	0.00095367	1

In decision table for Measures related to orderliness group 0=normal, 1=Benign and 2=Malignant.

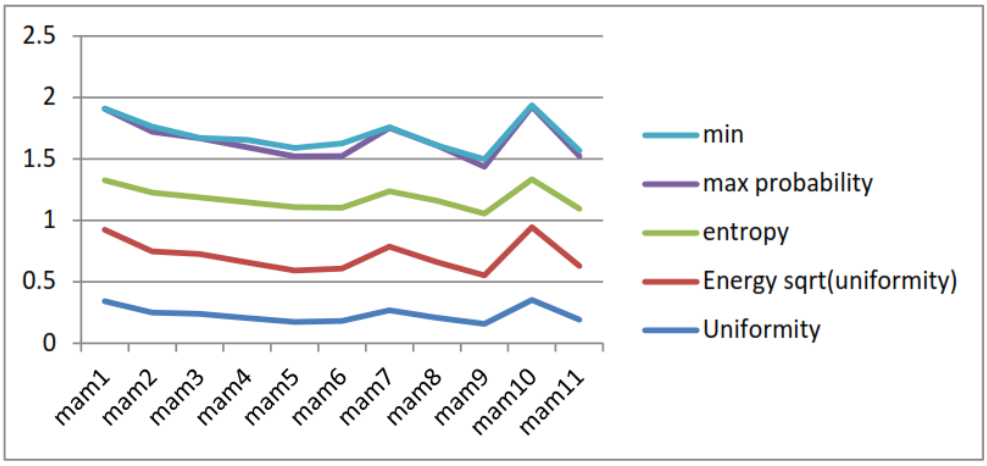

Figure 7: Data representation for Measures related to orderliness group.

Second: Descriptive statistics group data set

Table 7: Some descriptive statistics and their decision table.

U	C										D
	mean	Standard deviation	variance	smoothness	Area Descriptor	skewness	skurtosis	Third moment	Fourth moment	contrast	decision
mam1	36.5348	63.4199	4.02E+03	0.0583	1.7359E+00	1.5346	3.8219	0.0236	0.0146	9.85E+04	1
mam2	49.3319	73.0232	5.33E+03	0.0758	1.4802E+00	1.1902	2.8879	0.0279	0.0194	1.02E+05	1
mam3	52.0066	80.0201	6.40E+03	0.0896	1.5387E+00	1.1485	2.6306	0.0355	0.0255	1.07E+05	0
mam4	59.4984	82.6967	6.84E+03	0.0952	1.3899E+00	0.9093	2.0935	0.031	0.0232	1.10E+05	0
mam5	64.4811	69.7978	4.87E+03	0.0697	1.0825E+00	0.4249	1.5451	0.0087	0.0087	1.24E+05	1
mam6	70.0907	72.7462	5.29E+03	0.0753	1.0379E+00	0.2896	1.4227	0.0067	0.0094	1.07E+05	0
mam7	49.86	74.2953	4.86E+03	0.0695	1.4901E+00	0.913	2.1428	0.0186	0.012	1.10E+05	0
mam8	61.9692	74.2953	5.52E+03	0.0782	1.1989E+00	0.5931	1.6595	0.0147	0.012	9.70E+04	0
mam9	53.5637	73.2029	5.36E+03	0.0761	1.3667E+00	0.9236	2.2758	0.0219	0.0155	1.32E+05	0
mam10	43.1755	69.3156	4.80E+03	0.0688	1.6054E+00	1.2032	2.7173	0.0242	0.0148	5.91E+04	1

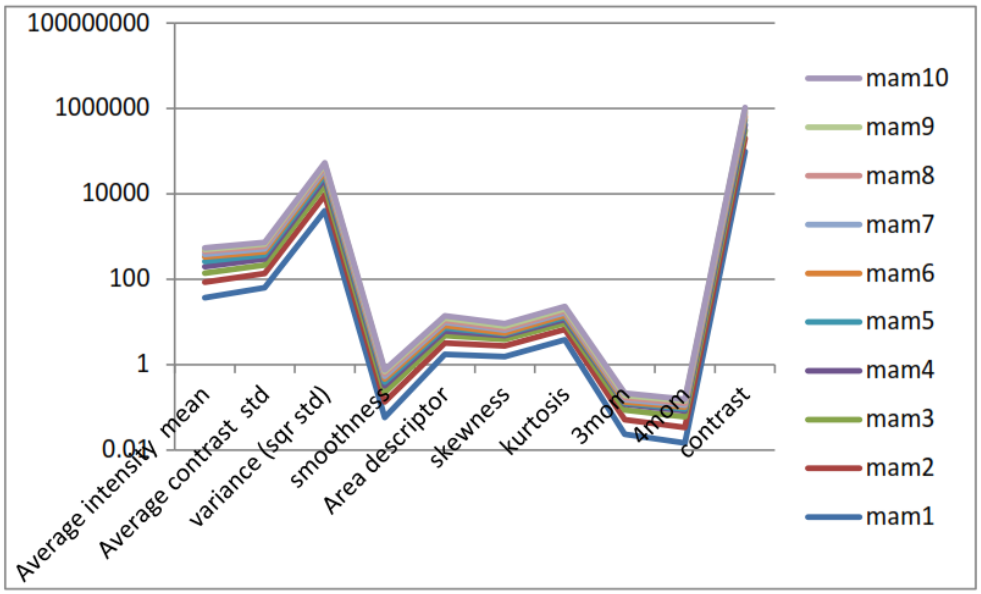

Figure 8: Data representation for descriptive statistics group.

Third: Contrast group data set

Table 8: Contrast group and their decision table.

U	C			D
	contrast	Inverse Difference Moment	Dissimilarity	decision
mam1	9.85	4.9	2.434129	1
mam2	10.19	3.9	2.604502	1
mam3	10.73	4.3	2.549482	0
mam4	10.98	4.1	2.648189	0
mam5	12.36	3	2.940145	1
mam6	10.71	3.7	2.659529	0
mam7	11.05	3.8	2.633394	0
mam8	9.70	4.1	2.473971	0
mam9	13.16	3.3	2.937944	0
mam10	5.91	4.5	2.002846	1
mam11	12.83	3.4	2.891276	0
mam12	10.75	4	2.630655	1
mam13	10.50	4.5	2.529737	1
mam14	8.96	4.1	2.375325	0
mam15	10.64	4.2	2.623059	1
mam16	6.61	4.4	2.09246	0
mam17	9.13	4.3	2.412082	1
mam18	6.92	4	2.194056	0
mam19	13.12	3	3.020435	1
mam20	10.01	3.9	2.527979	0

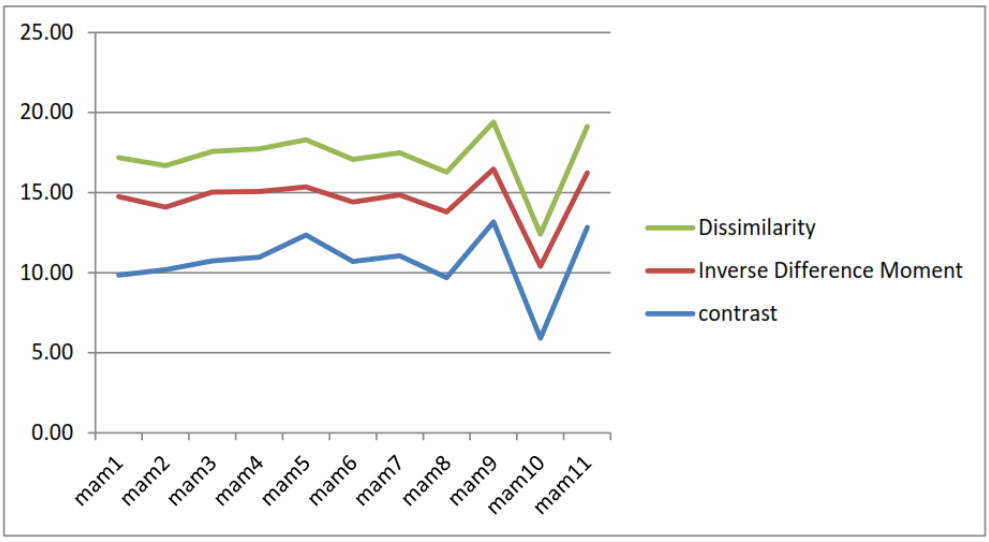

Figure 9: Data representation for contrast group.

1. with the reduction result $Y = [1, 2]$, which means that the feature Energy and uniformity in our system irreducible. The features Entropy, max probability and min probability can be removed without affecting the recognition of system.
2. degree of dependence $k = \frac{2}{5} = 0.4$ this means elements that belong to positive region can be classified.
3. Accuracy of the classification is obtained by using the given equation:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)} \times 100 \tag{7}$$

Where:

True positive=TP: Correctly classified as having breast cancer.

True negative=TN: Correctly classified as not having breast cancer.

False positive=FP: Classified as having breast cancer but actually they don't have cancer.

False negative=FN: Classified as not having breast cancer but actually they have cancer[19].

Table 9: Classification result accuracy.

no	sample	Knn classifier	TP	TN	FP	FN	accuracy	average
1	5		1	4	0	0	100	86.8
2	50		30	12	6	2	84	
3	100		56	22	12	10	78	
4	100		64	20	10	6	84	
5	200		121	55	9	15	88	

5. Conclusion

In this study, the tumor identification and the investigation are carried out with the help of rough theory for the potential use of mini-breast data for improving the tumor classification. Using rough set theory feature extraction is performed efficiently and this made segmentation easier varying tissue characteristics[20]. Although there is no known way to prevent breast cancer, mortality can be reduced only with early diagnosis. Therefore, the computer aided diagnosis (CAD) systems are very important as they allow radiologists to reconsider mammogram images with increased sensitivity of detection and diagnosis. The health status classification is performed with two consequent stages, where the normal and abnormal mammograms are determined first, and the abnormal defined mammograms are then classified as benign and malignant. These results helpful for radiologists to make more accurate breast cancer diagnoses.

References

[1] J Suckling et al., The Mammographic Image Analysis Society Digital Mammogram Database Exerpta Medica, International Congress Series 1069 (1994) 375–378.
[2] General cancer classification, staging, and grouping systems, (2016), Available online: http://stedmansonline.com/webFiles/Dict-Stedmans28/APP21.pdf.
[3] Slowinski R., Rough Set Approach to Decision Analysis, AI Expert, March (1995) 19–25.
[4] Heine JJ, Scott CG, Sellers TA, Brandt KR, Serie DJ, Wu FF,Morton MJ, chueler BA, Couch FJ, Olson JE, Pankratz VS, Vachon CM A novel automated mammographic density measure and breast cancer risk. J Natl Cancer Inst 104 (2012) 1028--1037.
[5] The Mammographic Image Analysis Society: Mini Mammography Database, Available online: http://peipa.essex.ac.uk/info/mias.html.
[6] Huang QH, Lee SY, Liu LZ, Lu MH, Jin LW, Li AH A robust graph-based segmentation method for breast tumors in ultrasound images. Ultrasonics 52(2) (2012) 266--275 .
[7] Haralick R.M., Statistical and Structural Approaches to Texture Proceeding of the IEEE, 67(5) (1979).
[8] Pawlak Z., Rough sets, Int. J. Computer and Information Science, 11(4) (1982) 341–356.
[9] Kryszkiewicz M., Rybinski H., Finding reducts in composed information systems, Fundamental Informatica 27(2-3)(1996)183–196.
[10] Gardezi S.J.S., Faye I., Bornot J.M.S., Kamel N., Hussain M., Mammogram classification using dynamic time warping, Multimed. Tools Appl. 77 (2017) 1–2 .
[11] General knn classification, Available online https://www.mathworks.com/help/stats/classificationknn.html
[12] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2018, Ca-a Cancer J. Clin. 68(1) (2018) 7–30.
[13] National breast cancer foundation, (2017), Available online: http://www.nationalbreastcancer.org/about-breast-cancer.
[14] Suckling J. et al., Mammographic image analysis society (mias) database v1. (2015) 21.
[15] K. Kamnitsas et al., Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Med. Image Anal. 36 (2017) 61--78.
[16] I. Iskli Esener, S. Ergin, and T. Yüksel, A New Feature Ensemble with a Multistage Classification Scheme for Breast Cancer Diagnosis, Journal of Healthcare Engineering, Article ID 3895164, 15 pages, Volume 2017.
[17] Reddy, V. E.,Reddy, E. S. Image segmentation using rough set based fuzzy K-means algorithm, International Journal of Computers and Applications 74(14) (2013)36--40.
[18] Asit K. Das, Shampa Senguptab, Siddhartha Bhattacharyya A group incremental feature selection for classification using rough set theory based genetic algorithm, Applied Soft Computing 65 (2018) 400--411.
[19] Madhu Kumari, Vijendra Singh, Breast Cancer Prediction system International Conference on Computational Intelligence and Data Science, Procedia Computer Science 132 (2018) 371--376.
[20] Kaya Y, Uyar M., A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease, Applied Soft Computing 1 13(8) (2013) 3429–3438.
[21] Mohabey A., Ray A. K., Rough set theory-based segmentation of color images, In Proceedings of 19th International Conference of the North American Fuzzy Information Processing Society (2000) 338–342.
[22] Pal S. K., Mitra P., Multispectral image segmentation using the rough-set-initialized EM algorithm, IEEE Transactions on Geoscience and Remote Sensing 40(11) (2002) 2495–2501.
[23] Aboul Ella Hassanien and Jafar M. Ali, Rough Set Approach for Classification of Breast Cancer Mammogram Images, Image and Vision Computing (2006), http://www.cba.edu.kw/abo
[24] Weiling Cai, Songcan Chen, Daoqiang Zhang, Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recogn (2007) 825–838.
[25] Jirava Pavel, Krupka Jiri, Cclassification model based on rough set and fuzzy sets, 6th WSEAS Int. Conference on Computational Intelligence, Man-Machine Systems and Cybernetics, Tenerife, Spain, (2007) 14-16.
[26] Jiang et al., He Wei, Nnew method for the image segmentation based on the rough set theory and neural networks, 5th International Conference on Visual Information Engineering (2008).
[27] Halder A, Dasgupta, A Color image segmentation using rough set based K-means algorithm, International Journal of Computers and Applications (2012) 32-38.
[28] Anupama N., Kumar S. S., Reddy E. S., Rough Set based MRI Medical image segmentation using optimized initial centroids, International Journal of Emerging Technologies in Computational and Applied Sciences 6(1) (2013) 90–98.