Abstract— The aim of our research is to classify digital mammograms into two classes, abnormal microcalcification and normal. Texture is one of the major mammographic characteristics. The statistical textural of Gray Level Cooccurrence Matrix (GLCM) used in characterizing images are contrast, energy and entropy. K-Nearest Neighbor (K-NN) and Fuzzy K-Nearest Neighbor (FK-NN) was proposed for classifying images. The result of K-NN method shows that 77.78% accuracy, 50.00% specificity and 100% sensitivity. The result of FK-NN method shows that 88.89% accuracy, 100% specificity and 80.00% specificity.

Keywords— Fuzzy K-Nearest Neighbor, Gray Level Cooccurrence Matrix, K-Nearest Neighbor, Microcalcification.

I. INTRODUCTION

More than one million breast cancer cases occur annually and more than 400,000 women die each year from this fatal disease as estimated by World health Organization’s International agency for Research on Cancer (IARC) [1]. Though breast cancer leads to death, early detection of breast cancer can increase the survival rate. The current diagnostic method for early detection of breast cancer is mammography. Mammographies are low dose X-ray projections of the breast, and it is the best method for detection cancer at the early stage[2].

Microcalcifications are tiny calcium deposits accumulated in breast tissue and may be an early sign of breast cancer. They appear in mammograms as small bright spots embedded within an homogeneous background [3].

Texture are one of the important features used for many application. Texture features have been widely used in mammograms classification. Extracted texture feature provide information about textural characteristics of the image [4].

In this research, we take texture features from gray level co-occurrence matrix (GLCM) to discriminate between microcalcification and normal tissue in the breast. Three features (contrast, energy and entropy) are extracted from original of region of interest (ROI). And the classifier is used K-Nearest Neighbor and Fuzzy K-Nearest Neighbor.

1. Gray Level Cooccurrence Matrix (GLCM)

In statistical texture analysis, texture feature are computed from the statistical distribution of observed combinations of intensities at specified positions relative to each other in the image. According to the number of intensity points (pixels) in each combination, statistics are classified into first-order, second-order and higher-order statistics.

The Gray Level Cooccurrence Matrix (GLCM) is a way of extracting second order statistical texture feature. A GLCM is a matrix where the number of rows and columns is equal to the number of gray levels, G, in the image. The matrix element P(i, j | Δx,Δy) is the relative frequency with which two pixels, separated by a pixel distance (Δx,Δy), occur within a given neighborhood, one with intensity i and the other with intensity j. One may also say that the matrix element P(i, j | d,θ) contains the second order statistical probability values for changes between gray levels i and j at a particular displacement distance d and at a particular angle (θ).

Figure 1 below illustrates the geometrical relationships of GLCM measurements made for four distances d (d = max(|Δx|, |Δy|)) and angles of θ = 0, π/4, π/2 and 3 π /4 radians under the assumption of angular symmetry.

![Geometry for measurement of gray level coocurrence matrix for 4 distances d and 4 angles θ.](image)

1.3. Given an M xN neighborhood of an input image containing G gray levels from 0 to G − 1, let f(m, n) be the intensity at sample m, line n of the neighborhood. Then

\[ R(i,j|Δx,Δy) = WQ(i,j|Δx,Δy) \]

(1)

where

\[ W = \frac{1}{(M-Δx)(N-Δy)} \]

and
A number of texture features may be extracted from the GLCM. The following features are used:

- **Contrast**:
  \[
  \text{Contrast} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P(i,j) \left| i - A \right|^{n} \quad (2)
  \]

  This measure of contrast or local intensity variation will favour contributions from \( P(i, j) \) away from the diagonal, i.e. \( i \neq j \).

- **Energy**:
  \[
  \text{Energy} = \sqrt{\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P(i,j)^{2}} \quad (3)
  \]

  Energy is the sum of squared elements in the Gray Level Co-occurrence Matrix (GLCM). Energy is also known as uniformity. The range of energy is \([0, 1]\). Energy is 1 for a constant image.

- **Entropy**:
  \[
  \text{Entropy} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P(i,j) \log P(i,j) \quad (4)
  \]

  Inhomogeneous scenes have low first order entropy, while a homogeneous scene has a high entropy.

2. K-Nearest Neighbor

The kNN classifier is the first and foremost extension of the nearest neighbor classifier, which is a versatile multivariate statistical technique. The efficiency of kNN has already been verified against other statistical techniques and some neural networks. It can be defined as, for a given unlabeled sample, the \( k \) “closest” labeled samples in the reference data set are found and \( x \) is assigned to the class that appears most frequently within the \( k \) subset. The kNN only requires an integer \( k \), a set of labeled examples and a measure of “closeness”. The appropriate selection of \( k \) value becomes an apparent problem to solve. If a larger value of \( k \) is chosen, it may suppress the fine structure density of reference set, else if the \( k \) value is chosen too small, it may put more emphasis on chance locations of a very few references. One of the probable solutions is to try on various values of \( k \) and select the \( k \) which gives the best classification accuracy [5].

3. Fuzzy k-Nearest Neighbor

The fuzzy k-NN classification method was designed by James M. Keller et.al. 1985 and this method can often improve performance in biological and medical data classification problems.

This method assigns memberships of samples to various classes rather than a particular class as in k-NN method. The following relationship provides class memberships to the sample as a function of the sample’s distance from its k-NN training samples:

\[
u_{k}(x) = \frac{\sum_{j=1}^{k} \left( \left| x - x(j) \right|^{-m(1-2/n)} \right)}{\sum_{j=1}^{k} \left( \left| x - x(j) \right|^{-m(1-2/n)} \right)}
\]

where \( m \) is a fuzzy strength parameter between 1 and 2, that determines how the distance is weighted. The variable \( k \) is the number of nearest neighbors. And \( u_{k}(x) \) is the membership of the test sample \( x \) to class \( i \). \( \left| x - x(j) \right| \) is the distance between the test sample \( x \) and its nearest training samples \( x(j) \). To calculate the distance, various techniques can be used like k-NN. In this study, we used Euclidean distance. \( u_{k}(x(j)) \) is the membership value of the j-th neighbor to the i-th class which can be defined in several ways. The “crispest” way is to give them complete membership in their own class and nonmembership in all other classes. In other words; it assigns 1 if the sample belongs to class otherwise 0. Alternatively, a more fuzzy method can be used to assign the training samples memberships based on the distance from the class mean. After calculating the membership for the test sample, it is assigned to the class with highest membership. In the present study, we used the “crispest” way [6].

II. METHODS

In this study, there are two steps, namely feature extraction and classification. The first step is feature extraction. This step aims to take on the features GLCM of mammograms. They are contrast, energy and entropy at four directions (\( 0°, 45°, 90° \) and 135°). Then, we reduced them by calculating mean of four direction for each feature. The second step is to classify mammogram into two classes, microcalcification case and normal case. k-NN and Fk-NN were proposed as classify methods. We tried various values \( k = 3, 5 \) and \( k = 7 \).

To evaluate the performance of classifiers, sensitivity, specificity and accuracy rate were calculated. The sensitivity is the fraction of microcalcification cases that the classifier predicted as microcalcification, is calculated as MC/NC. The specificity is the fraction of normal cases that the classifier predicted as normal, is calculated as N/Nn. MC is the number of microcalcification cases that are correctly classified as microcalcification and N is the number of normal cases that are correctly classified as normal. Nc and Nn are the number of positive cases and negative cases. The classification accuracy is calculated as \( AC = (MC+N)/(NC+Nn) \). Stages to be carried out in this study can be seen in the block diagram of the system in Figure 2.
Fig 2. Block Diagram of the system

III. RESULT

The resulting images from segmentation stage are shown in Fig. 3.

The first features are calculated from mean of contrast at angles 0, 45, 90, and 135.

![Fig 3 Sample ROI](image)

![Fig 4 Mean Contrast](image)

Figure 4 shows mean contrast for calcification cases and normal cases. The mean contrast of the microcalcification case varies in the range 5.166335 – 15.15134, with an average of 9.495517. The mean contrast of the normal case varies in the range, with an average of 9.581253.

Figure 5 shows mean energy for calcification cases and normal cases. The mean energy of the microcalcification case varies in the range 0.03129 – 0.084664, with an average of 0.052758. The mean energy of the normal case varies in the range 0.02435 – 0.06534, with an average of 0.043673.

IV. CONCLUSIONS

Finally, the performance of classifiers can be evaluated in terms of sensitivity, specificity, and accuracy.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-NN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K=3</td>
<td>0.50</td>
<td>1.00</td>
<td>77.78</td>
</tr>
<tr>
<td>K=5</td>
<td>0.00</td>
<td>0.80</td>
<td>44.44</td>
</tr>
<tr>
<td>K=7</td>
<td>0.00</td>
<td>0.80</td>
<td>33.33</td>
</tr>
<tr>
<td>FK-NN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K=3</td>
<td>1.00</td>
<td>0.80</td>
<td>88.89</td>
</tr>
<tr>
<td>K=5</td>
<td>0.75</td>
<td>0.80</td>
<td>77.78</td>
</tr>
<tr>
<td>K=7</td>
<td>0.75</td>
<td>0.80</td>
<td>77.78</td>
</tr>
</tbody>
</table>

From Table 1 it can be seen that while k=3, both classifier K-NN and FK-NN achieve the best performance. The FK-NN (88.89%) performed better than K-NN (77.78%).
contrast, energy and entropy. Experiments show that for the value \( k = 3 \), the classifier performance is optimum. Using the proposed \( F_k \)-NN classifier, 88.89% accuracy is achieved which is very much encouraging compare to the \( k \)-NN accuracy 77.78%.

ACKNOWLEDGMENT

This work is funded by BOPTN Sains and Technology faculty, Airlangga University. We sincerely and greatly thank anonymous reviewers for thorough evaluation of our work and valuable suggestions.

REFERENCES


