An Integrated Clustering Method for Medical Images

Maria Fayez[1], Soha Safwat [2] and Ehab Hassanein [3]
[1,2]Faculty of Computer Science, October University for Modern Sciences and Arts, Giza, Egypt
[3]Faculty of Computers and Information, Cairo University, Giza, Egypt
[1]Mariafayez91@yahoo.com

ABSTRACT

Medical images became a huge problem due the fast growing size of the medical image repositories, thousands of medical images are produced daily. Medical images repositories need to be well organized using an efficient and fast tool to allow researches or medical experts to extract useful information from it in the right time and as fast as possible. Organizing large medical images repositories helps in many fields as in medical fields that can be useful in diagnosis and knowing the history of a patient and in the researching area as it can be mined easily and be a necessary step before many application as content based image retrieval and medical image classification application. The objective of this paper is to implement a new efficient clustering method for medical images. The system contains three main models, the first is to extract features using gray-level co-occurrence matrix and apply PCA for dimensionality reduction, and then k-means clustering is applied. This second model where the 2D wavelet transforms is applied as a feature extraction and feature selection is used to select most efficient attributes, then k-means clustering is applied. The final and proposed method is to combine the two methods and apply k-means clustering.

KEYWORDS

Medical images; Clustering; K-Means; 2D wavelet transform; Gray-level co-occurrence matrix; PCA

1 INTRODUCTION

Due to the huge numbers of medical images generated and created per day, archiving medical images has become a significant problem. Hospitals generate numerous numbers of medical images per day, with different modalities and types. In the future there will be exponentially increase in the amount of medical images generated per day. Variation in medical images makes archiving, classifying, indexing and retrieving medical images a problem as it varies from one person to another. Manually organizing and using those huge numbers of medical images generated is such an expensive and time consuming method. Medical images archiving, indexing and retrieval in a proper way helps in different fields like diagnosis, medical image retrieval, surgical planning and research [1].

There are different modalities of medical images as X-rays which produce images of internal tissues, bones and organs where the transmitted radiation is detected and processed into a two-dimensional shadow image of the body's bones and organs, bone scan, mammogram which is a specific type of X-ray imaging that uses a low dose X-ray system, nuclear medicine where radiation originating from radioactive substances brought into the body is detected outside the body and processed into an image, MRI-scan where hydrogen atoms in the body vibrate by radiating the body with radio waves, this radiation emitted by these atoms is processed into an image and the ultrasound where sound waves going through the body are reflected at the separating surface of two different types of tissue inside the body, these reflections are detected outside the body and processed into an image [2].

Medical images have a very complex nature which leads to difficulty in extracting features. Some main challenges in medical images nature is the noise which is strong, Low resolution is very common in medical images;

most of the medical images is gray level images and not colored ones and variation in the medical images. All these medical images characteristics lead to more difficulty while extracting features [1].

Databases of medical Databases of medical images are important in future diagnosis and research needs. Archiving and organizing these huge amounts of medical images requires automatic and efficient tools which can be used to analyze the content of medical images and archive them in a proper way where they can be searched and used easily in different fields. Medical information systems are important to deliver the needed information in the suitable time and with good efficiency to help in different medical fields. Archiving and organizing huge number of medical images in automatic and fast way helps medical experts in diagnosis task and keeping those images in a patient's personal files which will be used to help medical experts in the patient's medical history [3].

Feature extraction is the process where the content of an image can be described by its color, texture or shape. There are different methods for extracting features from images. Using the suitable feature extraction method to extract the needed and effective visual content from an image is a difficult subject. Each extracted feature specifies some of the object's experimental property or characteristics. Features are mainly classified to two different types which are the general features where features as shape, color and texture can be extracted on any level of abstraction whether pixel, local or global level. The second type is the domain specific level where features are reliant for applications like fingerprint and human faces [4].

Discrete wavelet transform is a feature extraction method used to extract texture features from discrete sample or image. Discrete wavelet transform analyzes signal at multiple resolution and frequency bands to decompose this signal to approximation and detailed information. Scaling and wavelet functions are used to obtain high and low frequency components for each row in the data. Discrete wavelet transform's first level of decomposition decomposes the given image to 4 sub-bands which are information details in the three directions (horizontal, vertical and diagonal) and the approximation of the source image which will be used in the second level of decomposition [5].

Gray-level co-occurrence method which is well known as GLCM is used for extracting texture features from mainly gray level images. Gray-level co-occurrence matrix (GLCM) is used by the gray-level co-occurrence method; this matrix contains gray level combination frequencies within pixels pairs in an image. Different directions can be used to construct the gray-level co-occurrence matrix as vertical, horizontal or diagonal which represents the relationship between pixels direction. Finally, some features can be computed from the obtained gray-level co-occurrence matrix such as energy, contrast, entropy and homogeneity [6].

Clustering is a data mining task which is used for unsupervised learning. Clustering is used to discover new groups or categories as it groups data objects to clusters or subjects where similar objects are grouped together in one cluster and different ones belongs to another cluster. Finally, each data object will belong to one and only one group or cluster. There is mainly two types of clustering hierarchal, graph-based and partitioned clustering [7].

K-means clustering algorithm is a prototype-based clustering where the prototype in it is the centroid which is the mean of all data points within the same cluster. K-means clustering algorithm firstly begins using initially K centroids, where K is the number of desired clusters. Then, each data point is assigned to the closest centroid using any distance measure and then each collection of points are assigned to one and only one cluster. Clusters centroids is then updated by calculating the mean of all points within each cluster, after having new or updated centroids the same steps are repeated until no change in centroids occurs [8].

The rest of this paper is organized as follows: section II briefly introduced the medical images clustering steps; section III contains the approaches used in proposed system section. The proposed system is shown in section IV. The experimental results are then discussed in section V. Finally section VI summarizes and concluded the work.

2 MEDICAL IMAGES CLUSTERING

2.1 Medical images

Medical images are fast growing in size due to nowadays technology, There are many and different types of medical imaging system which generates daily different types of medical images with huge number. Some of the medical images types are X-ray which creates a picture to show different body parts in black and whites shadows which differs according amounts of radiation absorbed by the body tissues. CT-scan which makes a cross-sectional picture of the body parts and it is produced from special x-ray equipment. Nuclear Medicine type which show the structure inside the body using a camera that can detect the radioactivity. Ultrasound type can look organs and structure inside the body using high frequency sound waves. MRI which uses radio waves and magnet to look structures and organs inside a body [9]

2.2 Feature extraction

Feature extraction is the process where the visual content of any image can be described using different methods. Feature is a role of one or more than one capacity as each feature represents property or special characteristic of a given image or object. Extracting features from an image is not a trivial process as it is hard to use the most suitable features to extract the needed or correct features. There are two main types of features which are general features as the color, shape or texture features and the domain specific level which is concerned with applications like the fingerprints application and conceptual features. General features that describes the image from its color, shape or features can be divided to pixel level where the features vary at each pixel such as color, local features where features are resulted from subdivision of an image as in image segmentation and edge detection and the global level features where the whole image is used to extract features or only specified area of the image is used to extract features [4].

Color feature extraction methods are very effective in extracting color features in a given image as it is measured for every pixel in an image. In color feature extraction method firstly, the color space need to be defined as RGB or HSV. Secondly features can be extracted using method from the color feature extraction methods as the global color histogram which is a histogram that specifies the whole image in one histogram and the local color histogram which break down the original image to parts and represent each part with histogram. Also, color correlogram is another method for color feature extraction which can describes spatial information and distribution of colors in an image [10].

Shape feature extraction which is concerned by detecting some shape features for the object in an image. There is contor based shape feature extraction method where features are extracted from the shape boundaries or region bas one where the features are extracted from the entire region of the shape in an image [11].

Texture feature extraction which can describe the characteristics of the surface of an image and also visual patterns can be provided. Due the pixel repetition valued information can be extracted and it can describe also the relationship between the image’s surface and the environment. Texture feature extraction methods are commonly used as it is measured for a group of pixels so it gives valued features in describing different images. There is huge and variant number of texture extraction methods. Discrete wavelet transform break down the original image which is called decomposition of an image. Image can be decomposed as much as needed, in each decomposition level 3 sub-bands representing different directions which are the diagonal, horizontal and vertical directions and the approximation image that will be used in the next level of decomposition if needed [9]. Gray-level co-occurrence matrix is a matrix that is constructed from the relationship between a pixel and another. GLCM can be created in different directions some common directions are [0, 90, 180 and 270°]. Finally using any statistical measures as homogeneity, contrast or energy features can be extracted [12].

2.3 Data mining

Data mining is a non trivial process for identifying useful, understandable and valid patterns of data. Data mining helps in analysis and prediction of data, it is also called knowledge discovery of data as knowledge can be gained by analyzing and mining data sets. There are many kinds of information that can be collected as the scientific data, medical data, satellite sensing data, digital media data and e-mail messages data [13].

There are many steps for making the data ready for data mining, the data need to be cleaned, integrated, selected and transformed after these steps the data will be ready for mining and knowledge can be gained. Data that can be mined are the flat files, relational databases, data warehouses, transactional databases and multimedia databases. Data mining has two types of tasks which are the descriptive data mining that describe general properties of the data and the predictive data mining which makes prediction based on available data. [13]

Data mining has many functionalities and variety of knowledge as data characterization which summarize the general features of objects, data discrimination which compare general features of two different classes, association analysis whish discover the association rules, classification which organizes the data in classes, prediction which used to predict class for an unknown data and clustering which is like classification but the class label is not known [13].

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Besides the term data clustering as synonyms like cluster analysis, automatic classification, numerical taxonomy, petrology and typological analysis [14].clustering algorithms can be hierarchical or partitioned. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitioned algorithms determine all clusters at time.

3 PRELIMINARIES

3.1 K-means clustering

K-means clustering algorithm is a partition clustering algorithm. K-means clustering firstly begins with set of centroids where its number varies according to the number of cluster needed to divide the data points. After determining the number of clusters and its initial centroids each data point is assigned to the closest centroid using any of the distance measures. After assigning each data point to a cluster or the closest centroid the centroids will be updated by calculating the mean of the data points in each cluster. Again the distance between the centroida and the data points is calculated to reassign the data points to the closest new centroids and again the new centroids will be calculated until there are no updates in the centroida then the algorithm is done [8].

```
Algorithm 1 : Basic K-means
clustering
```

```
1. Select K points as intial
   centroids
2. Repeat
3. From K clusters by assigning
   each point to its closest
   centroid.
4. Re-compute the centroid of each
   cluster.
5. Until centroids don't change
```

3.2 Gray-level co-occurrence matrix (GLCM)

Gray Level Co-occurrence matrix (GLCM) is calculated through the gray-co-matrix function which is used to calculate the number of occurrences of a pixel with a specific gray level intensity with another pixel having another specific intensity in a specific spatial relationship. GLCM can vary by obtaining it using different distances and directions (commonly used one distance and four directions which are 0°, 90°, 180°, 270°). Finally Haralick‘s statistical measures as homogeneity, contrast, entropy, correlation, energy and variance can be used for calculating texture features from the GLCM [15].

$$GLCM_{homogeneity} = \sum_{i}^{G}\sum_{j}^{G}\frac{M_{i,j}}{1+(i-j)^2} \quad (1)$$

$$GLCM_{energy} = \sum_{i}^{G}\sum_{j}^{G}\left(M_{i,j}\right)^2 \quad (2)$$

$$GLCM_{Contrast} = \sum_{n=0}^{G-1} n^2 \sum_{i}^{G}\sum_{j}^{G}\left(M_{i,j}\right) \quad (3)$$

$$GLCM_{correlation} = \sum_{i}^{G-1}\sum_{j}^{G-1}\frac{(i,j)M(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y} \quad (4)$$

3.3 Discrete wavelet transforms (DWT)

2D discrete wavelet transform is a powerful method that is used in analyzing of images. Discrete wavelet transform can show and represent localized details of an image in space and frequency domain. Mallat's tree algorithm is used to implement the discrete Fourier transform in an efficient way which used iterative linear filtering and down-sampling using the original image to obtain 3sub-bands with high frequency each of which represents a direction (vertical V1, horizontal H1 and diagonal D1) and one another sub-band of low frequency which is the approximation A1. The approximation sub-band created is used to get the second decomposition level so another three sub-bands are extracted (H2, V2 and D2) and another approximation sub-band (A2). The image can be decomposed as much as needed [16].

3.4 Principle component analysis (PCA)

Principle component analysis (PCA) is a technique for extracting features where a linear transformation is used to convert a correlated set of observations to set of variables which are not correlated.

PCA can be also used to decrease a feature vector dimension. Given a number of Eigen vectors more than the number of columns in a feature vector produce feature vector with reduced dimensions. Decreasing the dimensionality of a given set of feature vectors leads to reducing in complexity in the clustering step and gives better results.

4 PROPOSED SYSTEM

In this section, the proposed method will be discussed. There are mainly three models in the proposed method. In section 4.1, we will describe clustering medical images using gray-level co-occurrence matrix (GLCM) approach mode. In section 4.2, we will describe clustering medical images using discrete wavelet transform (DWT) model. In section 4.3, a combination of the two previous discussed models is used to obtain the final model.

Experiments are conducted on two different medical images datasets. The first one consists of 500 medical images of two different types which are X-rays and CT-scans divided to

5 clusters, 3 clusters are X-ray medical images which are (chest, neck and knee) with 100 images in each class and 2 CT-scans clusters which are (brain and spine) with 100 images in each class. The second one is a data set of 150 X-rays only of 5 cluster which are (hand, skull, chest, backbone and knee

4.1 Clustering medical images using GLCM approach.

Figure 1, shows the first method which is implemented for clustering images using GLCM as the feature extraction method, PCA for dimensionality reduction and k-means clustering algorithm for clustering obtained feature vectors and the final clustering results obtained from applying k-means clustering algorithm will be saved to be used later.

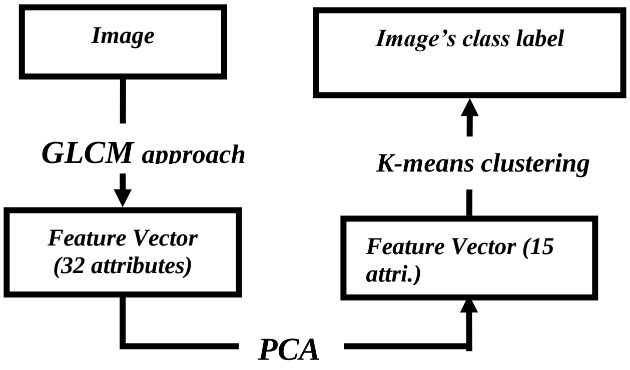

Figure 1: Clustering images using GLCM & k-means

4.1.1 Feature extraction using GLCM

In the proposed method, texture features are extracted using the gray level co-occurrence matrix GLCM. Firstly, GLCM is obtained using 8 directions which are [0°, 45°, 90°,135°,-45°,-90°,-135°and -180°] then we now have eight GLCMs, from each gray level co-occurrence matrix (GLCM) four texture features using Haralick's statistical measures are obtained these measures are contrast, correlation, energy and homogeneity. Finally for each image we will extract feature vector of 32 attributes. Equations from 1-4 shows the four statistical features measures used.

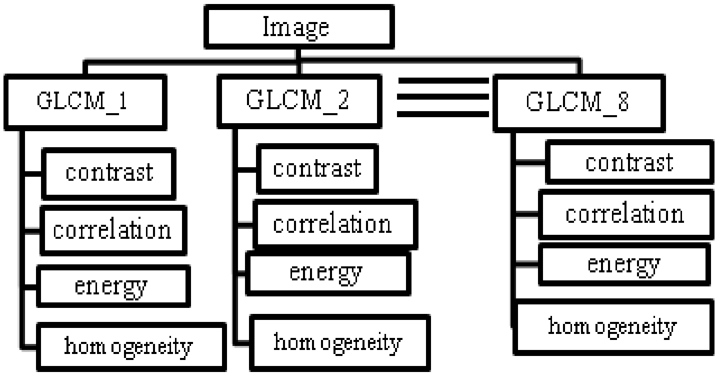

Figure 2: Haralick's statistical measures from GLCM

4.1.2 Using PCA for dimensionality reduction

PCA is used to reduce the dimensionality of the feature vector from length of 32 attributes to 15 attributes. Clustering after applying the PCA on the feature vector database gives better results than before using PCA.

4.1.3 Clustering feature vector using k-means algorithm

Finally, k-means clustering algorithm is used to cluster the extracted feature vectors from the PCA, where the number of clusters is known which is five, centroids of the five clusters are firstly initialized by any random feature vectors to represent each class centroid. City block distance is used to measure the distances between the feature vector and the classes' centeroids. K-means clustering is applied and final clustering results are saved for further use.

```
Algorithm 1 : K-means clustering
Input: S (set of extracted feature vectors),
       K ( 5 )
Output: clusters
Initialize 5 cluster centers (random feature Vectors).
while termination condition is not satisfied do
    Assign instances to the closest cluster center Using city block distance measure.
        $D_{i,j} = \sum_{l=1}^{d} |x_{il} - y_{jl}|$
    Update cluster centers based on the assignment.
end while
```

4.2 Image clustering using DWT and k-means clustering

The following diagram shows the second implemented method. Firstly, 2D wavelet transform approach is used as the feature extraction method. Then feature selection is used to select some features for better results and finally, k-means clustering algorithm is applied to obtain final clustering results and the final clustering results will be saved to be used later.

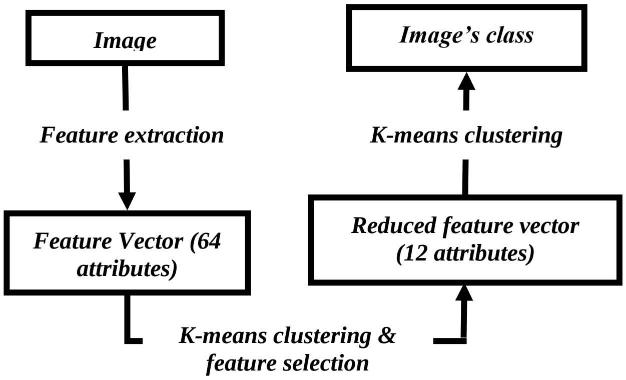
Image
Image's class
Feature extraction
K-means clustering
Feature Vector (64 attributes)
Reduced feature vector (12 attributes)
K-means clustering & feature selection

Figure 3: Clustering images using DWT & k-means

4.2.1 Feature extraction using DWT

For each image in the data set 2D wavelet transform approach is used to extract texture features. Each image in the database is decomposed to 4 levels; from each level 3 sub-bands are extracted for the three directions which are horizontal, vertical, diagonal and one approximation image to be used for the next level of decomposition. Each of the 3 extracted sub-bands from the 2D wavelet transform and the approximation image is used to calculate four feature measures which are maximum, minimum, mean and standard deviation. Now each image is represented by a feature vector of length 64. Equations from 5-8 shows the equations of standard deviation, mean, maximum and minimum respectively.

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} \qquad (5)$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N} \qquad (6)$$

$$Max = \max\,(A) \qquad (7)$$

$$Min = \min\,(A) \qquad (8)$$

Figure 4, shows the first and second level of decomposition using 2D wavelet transform, in the first level of decomposition, the extracted approximation image which will be used in the further decomposition and the 3 sub-bands which are the horizontal, vertical and diagonal respectively are shown.

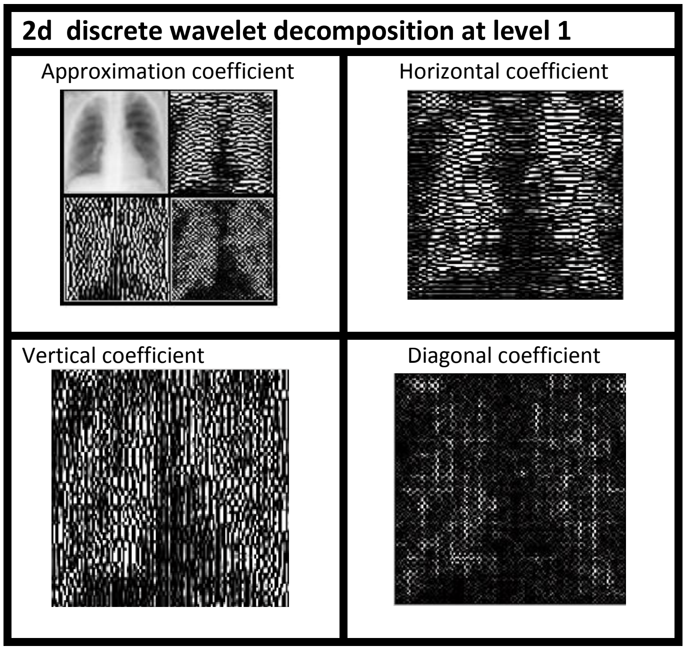
2d discrete wavelet decomposition at level 1
Approximation coefficient
Horizontal coefficient
Vertical coefficient
Diagonal coefficient

Figure 4: 2D discrete wavelet decomposition at level 1

4.2.2 Feature selection

After obtaining the feature vector database extracted using the 2D wavelet transform approach, different combinations from the 64 attributes (the length of the feature vector representing each image) was experiment to select the best attributes that represents each image. Finally, the most 12 attributes which gives good clustering results were selected as the new feature vector and used in K-means clustering algorithm step.

4.2.3 Clustering images using K-means algorithm

Finally, k-means clustering algorithm is used to cluster the selected feature attributes only which are 12 attributes for each image, where the number of clusters is known which is five, centroids of the five clusters are firstly initialized by any random feature vectors to represent each class centroid. City block distance is used to measure the distances between the feature vector and the classes' centeroids. K-means clustering is applied and final clustering results are saved to be used later.

4.3 Clustering images using DWT and GLCM approaches

4.3.1 Combining the two models

By combining the previous two implemented models. For each image in the database, feature vector of length 27 attributes is extracted (15 attribute from the GLCM model & 12 attribute from the DWT model). This new extracted feature vector database will be used by k-means clustering algorithm.

4.3.2 Clustering using k means

Finally, k-means clustering algorithm is used to cluster the new feature vectors database that contains feature vector of length 27 attributes representing each image in the database, where the number of clusters is known which is five, centroids of the five clusters are firstly initialized by any random feature vectors to represent each class centroid. City block distance is used to measure the distances between the feature vector and the classes' centeroids. K-means clustering is applied and final clustering results are saved to be used later.

4.3.3 Selecting the mode

The last step in our proposed method is that we have saved the clustering results obtained from clustering extracted features vectors database using the 2D wavelet transform approach, gray-level co-occurrence matrix GLCM approach and using both feature vectors. Finally we have 3 clusters obtained for each image in the database, Mode is used to select which cluster will be the winner one.

5 EXPERIMENTAL RESULTS

In this section we describe the results obtained from clustering medical images using gray-level co-occurrence matrix (GLCM) approach, 2D discrete wavelet transform approach and the final proposed . The proposed method is implemented using MATLAB and two datasets are used to test the performance of the two proposed method.

5.1 Data Sets

Two datasets were used to test the accuracy of the proposed system:

The first one consists of 500 medical images of two different types which are X-rays and CT-scans divided to 5 clusters, 3 clusters are X-ray medical images which are (chest, neck and knee) with 100 images in each class and 2 CT-scans clusters which are (brain and spine) with 100 images in each class.

The second one is a data set of 150 X-rays only of 5 clusters which are (hand, skull, chest, backbone and knee)

5.2 Evaluation Method

Since, we have data set as ground truth with pre-labeled classes, and then accuracy measure can be used to calculate the performance of each cluster and the overall performance.

$$accuracy = \left(\frac{no.\ of\ correctly\ clustered\ images}{total\ no.\,of\ images}\right) \times 100$$

5.3 Experimental Results

Using data set 1, the system gives overall accuracy for clustering using GLCM approach of 75.8% , for clustering using the 2D wavelet transform approach method overall accuracy of 87.4%, for clustering using both feature extraction methods overall accuracy of 87.6% and finally when applying mode the proposed system give overall accuracy of 88.8%. This shows that the proposed method gives better overall performance than the other two methods. Table 2 shows the performance for each cluster

Table 2: Accuracy Results using data set1

Cluster	Chest X-ray	Neck X-ray	Knee X-ray	Brain CT-scan	Spine CT-scan
GLCM	77%	58%	54%	99%	91%
DWT	96%	80%	66%	96%	99%
Proposed method	100%	82%	66%	97%	99%

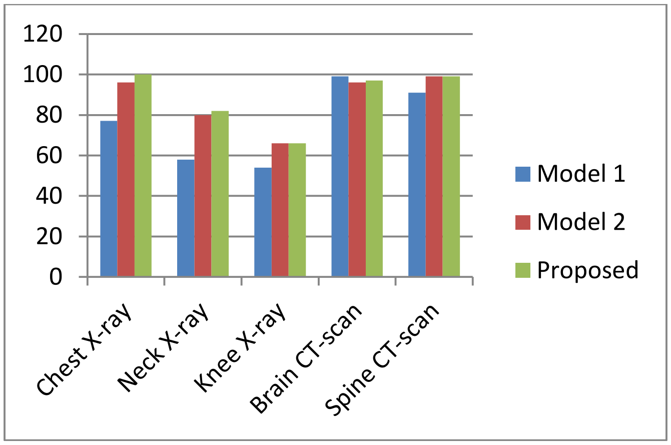

Figure 4: Results using the first dataset

According to table 2 the clusters of CT-scans medical images gives better accuracy results using the proposed method as their texture nature is more better than x-rays which have complex texture nature and much more noise.

Using data set 2,The system gives overall accuracy for clustering using GLCM approach of 68% , for clustering using the 2D wavelet transform approach method overall accuracy of 78.6%, for clustering using both feature extraction methods overall accuracy of 87.6% and finally when applying mode the proposed system give overall accuracy of 88.6%. This shows that the proposed method gives better overall performance than clustering medical images using DWT and GLCM method separately.

Table 3: Accuracy Results using data set 2

Cluster	Hand X-ray	Skull X-ray	Chest X-ray	Backbone X-ray	Knee X-ray
GLCM	36.6%	57.6%	66.6%	90%	87%
DWT	43.3%	92.3%	69.6%	93.3%	96.7%
Proposed method	76.6%	96.1%	93.9%	93.3%	83.8%

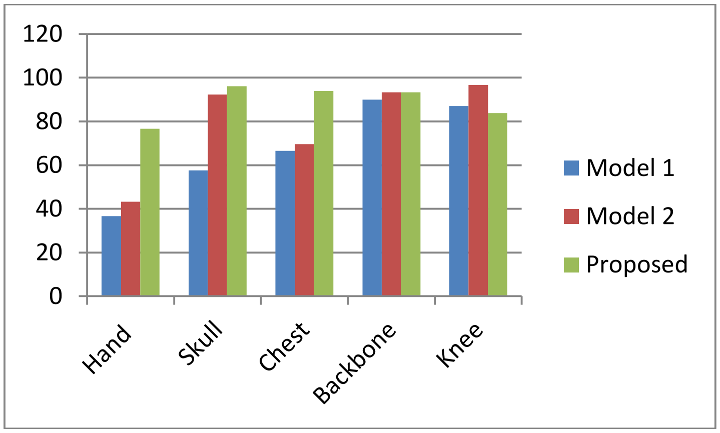

Figure 5: Results using the second dataset

According to the results shown in table 3 and the proposed method gives better performance with all clusters except with the Knee cluster due to the great variation in the knee x-rays positions and the small number of samples used per class.

6 CONCLUSIONS

In this paper a new proposed system for clustering medical images is implemented and tested. Two different datasets were used to test the proposed method. In data set 1 the propose method showed better results with medical images of type CT-scan more than the x-rays.

In data set 2 where all the images were x-rays the proposed system gives more poor results but it shows great accuracy as the number of medical images was not large so there were no much variation within the same class. The proposed system gives better results if the number of images used is large especially if cluster has different positions medical images as the hand and knee x-rays cluster

REFERENCES

[1] S.Malar Selvi and Mrs.C.Kavitha,Content Based Medical Image Retrieval System (CBMIRS) Using Patch Based Representation

[2] Diagnsis images

[3] Chhanda Ray, Krishnendu Sasmal,(2010).A New Approach for Clustering of X-ray Images.

[4] Dr. T. Karthikeyan1, P. Manikandaprabhu,(2014). A Study on Discrete Wavelet Transform based Texture Feature Extraction for Image Mining

[5] Balwinder Singh, Gurbinder Kaur,(2011). Intensity based image segmentation using wavelet analysis and clusteringtechniques.

[6] Asadollah Shahbahrami1, Demid Borodin and Ben Juurlink. Comparison Between Color and Texture Features for Image Retrieval.

[7] Lior Rokach.Data mining and knoweldge discovery handbook.

[8] Chhanda Ray, Krishnendu Sasmal. Cluster Analysis: Basic concepts and algorithms.

[9] Hodder Arnold,(2012). Introduction to medical imaging.

[10] Garima Tripathi,(2014). Review on color and texture feature extraction techniques.

[11] Ying Liu, Dengsheng Zhang, Guojun Lu,Wei-Ying,(2006). A survey of content-based image retrieval with high-level semantics.

[12] Ramamurthy, B. and K.R. Chandran l,(2012). Content Based Medical Image Retrieval with Texture Content Using Gray Level Co-occurrence Matrix and K-Means Clustering Algorithms.

[13] osmar R. zaiane,(1999).A Introduction to data mining.

[14] Han, J. and Kamber,(2001). Data Mining: Concepts and Techniques s.

[15] Ramamurthy, B. and K.R. Chandran, (2012).Content Based Medical Image Retrieval with Texture Content Using Gray Level Co-occurrence Matrix and K-Means Clustering Algorithms.

[16] Nadia Baaziz, Omar Abahmane and Rokia Missaoui ,(n.d).Texture feature extraction in the spatial-frequency domain for content-based image retrieval.