Analysis of Thyroid Disease Using K Means and Fuzzy C Means Algorithm

Kirubha.M[#1], Prinitha.R[#2], P.Preethika[#3], A.Samyuktha[#4]

#1 Assistant Professor, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

#2, #3, #4 Students, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

Abstract

Thyroid disease is now a days most common and the second largest in the field of endocrine. Classification of this disease is primary problem in clinical treatment. Various research studies estimates that about 42 million people in India suffer from thyroid disease. There are a number of possible thyroid diseases and disorders right from simple goiter to thyroiditis and thyroid cancer. This paper is all about classification of thyroid disease into normal and abnormal. Medical imaging system has done lots of research on thyroid segmentation. The effects of the thyroid disease may be uncomfortable but if they are diagnose in a proper way they can be managed and well treated. Sometimes the disease will be a simple goiter and hence it can be cured naturally there is no need of any treatment but sometimes it might lead to cancer which requires removal of the thyroid gland. The defected thyroid gland can be either chemically removed or surgically removed. The diagnosis methods consist of four stages and they are pre-processing of input images where the image is converted into grey scale for better performance, feature selection, feature extraction and feature classification. Feature classification is based on fuzzy c means clustering.

Keywords — *Thyroid Disease, Data Mining, Fuzzy C Means Clustering*

I. INTRODUCTION

Large amount of data set are extracted to identify and analyze the pattern of data using data mining techniques. This technique is used for discovering the knowledge base from the given data set. All the metabolic process in our body is influenced by the hormone produced by the thyroid gland. Human body system is controlled by the hormone produced by the thyroid. This thyroid disease is most wide spreading and has become more common. Its abnormality may be a simple goiter or life-threatening cancer. A simple goiter needs no treatment it is just an enlarged gland. The abnormality is based on the level of thyroid stimulating hormone. There are two cases where the excess amount of thyroid production is hyperthyroidism and less production of hormone is hypothyroidism. A long period of untreated hypothyroidism might lead to myxedema coma. Myxedema coma is very rare but if affected surgery should be taken immediately. It might even lead to weight gain, dry skin and fatigue along with sleepiness.

TSH test is done to check if your thyroid glands are working properly in deep sense checkup is done to check for overactive and underactive conditions that are hyperthyroidism and hypothyroidism respectively. They are usually done on the morning. The normal TSH ranges from 0.4(mIU/L) to 5(mIU/L), when the TSH level is lower than the normal level we call it as underactive called hypothyroidism and when the level is above normal it is overactive called hyperthyroidism. The reasons for the lower and higher level of THS are mainly due to too much of iodine content in your body or graves, disease or due to too much consumption of supplement that naturally contains thyroid hormone.

II. LITERATURE SURVEY

Aswathi A K and Anil Antony used data mining technique to classify the image. Data mining Techniques play a vital role in healthcare organizations such as for decision making, diagnosing disease and giving better treatment to the patients. Thyroid gland plays a major role in maintaining the metabolism of human body. Data mining in health care industry provides a systematic use of the medical data. Thyroid diseases are most common today. Early changes in the thyroid gland will not affect the proper working of the gland. By the early identification of thyroid disorders, better treatment can be provided in the early stage thus can avoid thyroid replacement therapy and thyroid removal up to an extent.

Jameel Ahmed and M. Abdul Rehman Soomrani in have proposed a framework for diagnosing the thyroid disease type. The first phase is data pre-processing in which missing values in the dataset are filled using Medical Data Cleaning (MDC). Second phase is classification. Two SVM classifiers are used here. First one is the multi-SVM used for predicting the thyroid disease type ie, Euthyroid, Hypothyroid, Sub-clinical hypothyroid and Sub-clinical Hyperthyroid.

Prasad have proposed hybrid architecture for identifying the thyroid disease and the disease type. Rough data sets theory (RDS) is used here for finding the missing values in the input. Here, a String Matching System is proposed with Particle Swarm optimization and Artificial Bee Colony Optimization. The system is further enhanced with Rule Based System.

Khushboo Chandel, Veenita Kunwar, Sai Sabitha and Tanupriya Choudhury. Saurabh Mukherjee used Data mining

for classification. Data mining is an important research activity in the field of medical sciences since there is a requirement of efficient methodologies for analyzing and detecting diseases. Data mining applications are used for the management of healthcare, health information, patient care system, etc. It also plays a major role in analyzing survivability of a disease. Classification and clustering are the popular data mining techniques used to understand the various parameters of the health data set. In this research work, various classification models are used to classify thyroid disease based on the parameters like TSH, T4U and goiter. Several classification techniques like K-nearest neighbor, support vector machine and Naive Bayes are used. Here the drawback is less concurrency

Qin Yu, Tao Jiang, Aiyun Zhou, Lili Zhang, Cheng Zhang and Pan Xu used ANN for classification of the disease. The objective of this study is to evaluate the diagnostic value of combination of artificial neural networks (ANN) and support vector machine (SVM)-based CAD systems in differentiating malignant from benign thyroid nodes with gray-scale ultrasound images. Two morphological and 65 texture features extracted from regions of interest in 610 2D-ultrasound thyroid node images from 543 patients (207 malignant, 403 benign) were used to develop the ANN and SVM models. Tenfold cross validation evaluated their performance; the best models showed accuracy of 99% for ANN and 100% for SVM Min Zuo,Lan Xiang have proposed the article called improved ensemble classification method of Thyroid Disease Based on random forest.He described machine learning technique is widely used to assist in medical experts in decision making.This paper proposed a new method for thyroid disease classification based on random forest.The
drawbacks is complexity and huge memory.

III. PROPOSED SYSTEM

The proposed system has four stages and they are pre-processing, feature selection, feature extraction and feature classification.

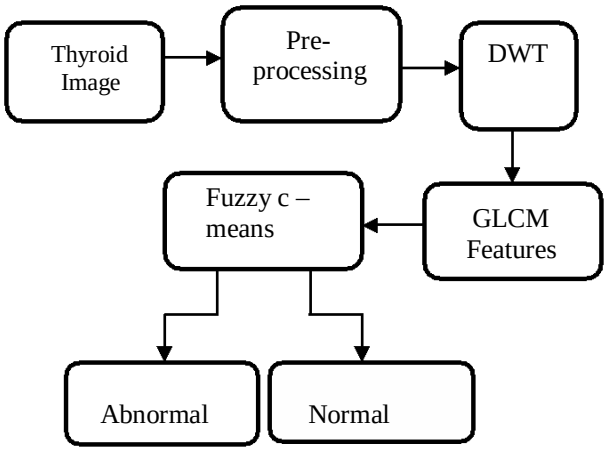

Fig 1: Block Diagram

A. Input Image

An image is a two-dimensional picture, which has a similar appearance to some subject usually a physical object. Images are captured by optical devices such as cameras, mirrors, lenses, telescopes, microscopes and even by natural objects and phenomena, such as the human eye or water surfaces.

Fig 2: Input Image

B. Pre- processing Stage

Input is image. An image is a group of pixels which is generally of two dimensional. Image may be an photography or picture which is captured by an optical devices such as camera.

In this stage the input image is pre-processed. This step is done to adjust the height and width of the given image. The width and height of the image is adjusted by multiplying them with 256*256 since it is two dimensional. Along with this the image is checked if it is colored or not, if colored it is converted to gray scale.

Gray scale is done so that the processing can be done more efficiently. A grayscale Image is digital image, which carries only intensity information. Image processing is a translation between the human visual system and digital imaging devices. pre-processing is done for the improvement of the image data that consist of distortions and to enhances some features that are important for further processing

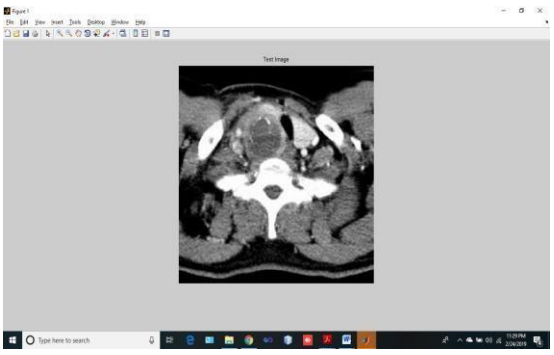

Fig 3: Thyroid Image after Pre-Processing

Images of this type are known as black-and-white, and they are composed exclusively of shades of gray (0-255), varying from black(0) at the weakest intensity to white(255) at the strongest.

C. Feature Selection

Feature selection is based on DWT (Discrete Wavelet Transformation), Here in in this algorithm the input is changed from one domain to another. Certain filters are done to achieve this process. This algorithm extracts only the meaningful information in a time-frequency domain. This algorithm can be implemented in many ways. The oldest implementation of DWT and most known one is the Malaat (pyramidal) algorithm. This transform decomposes the given signal into mutually orthogonal set of wavelets, and this is the main difference from the continuous wavelet transform (CWT), and for the discrete time series sometimes called discrete-time continuous wavelet transforms (DT-CWT). The same image might change based on the time the image is being captured or due to climatic conditions. The image which is given as input might have this kind of problem; to overcome this we use DWT. The image which changes with respect to time and climate is converted from time to frequency domain using filters. The Haar wavelet transform is used here, Haar pairs up input values, storing the difference and passing the sum.

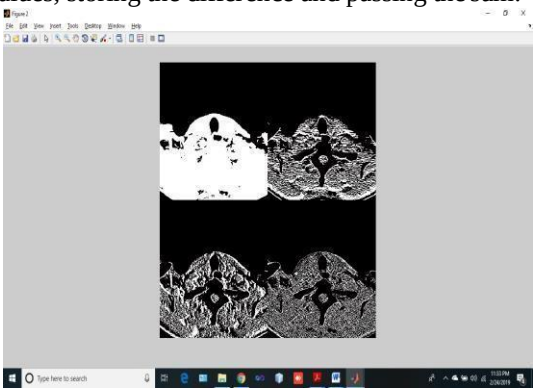

Fig 4: Thyroid Image After Application of DWT

This process is recursive. Pairing up the sums to prove the next scale, this leads to differences and a final sum. A key advantage it captures both frequency and location information.

D. Feature extraction

A statistical method of examining texture which considers the spatial relationship of pixels is known as the gray-level co-occurrence matrix (GLCM), which is also known as the gray-level spatial dependence matrix. The GLCM uses gray co matrix.

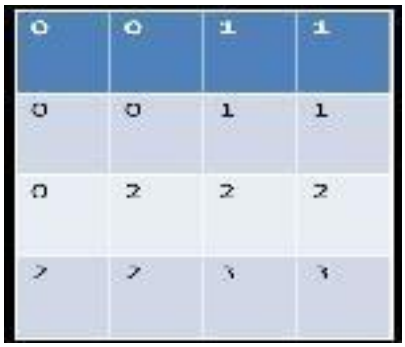

Fig 5: General Form of GLCM

The above image is the general form of GLCM. GLCM functions characterize the texture of an image by calculating how often the pairs of pixel with specific intensity values occur in a specified spatial relationship in an image, and then extracting statistical measures from this matrix. It is also referred as co-occurrence distribution.

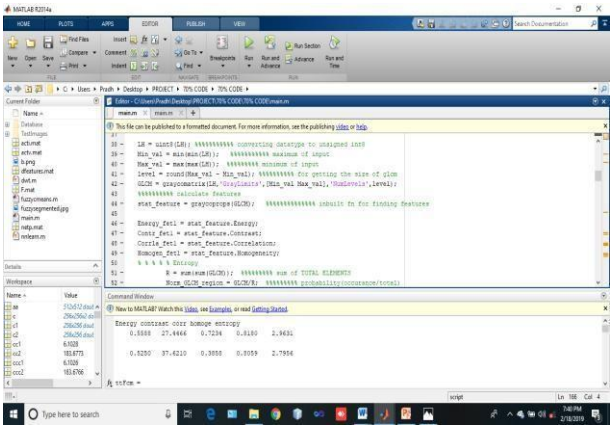

Fig 6: Texture Information Of Thyroid Image Using GLCM

We can get the values of energy, entropy, contrast and homogeneity of the given image.

1. Energy

Energy is also called Uniformity or Angular second moment. It measures the textural uniformity that is pixel pair repetitions in the image. It also detects disorders in textures. Energy reaches a maximum value equal to one.

2. Entropy

Measure the disorder of complexity of an image. The entropy is large when the image is not textually uniform Entropy is strongly but inversely corrected to energy.

3. Contrast

Measure the spatial frequency of an image It is the difference between the highest and the lowest valves of a contiguous set of pixels It is measures the amount of local variations present in the image

4 .Homogeneity

Homogeneity is also called inverse difference moment which measures the image homogeneity as it assumes larger values for smaller gray tones differences in pair elements It is more sensitive to the presence of near diagonal element in the GLCM It has maximum value when all element in the image are same

E. Feature classification

Feature classification is based on two clustering algorithm. The clustering algorithms used are K-Means Clustering and another one is Fuzzy C Means algorithm.

1. K-Means Clustering

Image segmentation is defined as the classification of an image into different groups. There are different methods used for segmenting the image and one of the most popular methods is k-means clustering algorithm. K-means clustering is a type of unsupervised learning that is used to groups the images. This algorithm is used for unlabeled data. The algorithm classifies a given data set into a certain number of clusters. Initially k is assumed where k is the number of clusters and they are fixed throughout the problem. Define a k center that is we define one center for each cluster. These centers should be placed in a correct way because difference in the location causes different result. The better choice for placing the cluster is placing them far away from each other. The next step is to take each point in the given data set and associate it to the nearest center, First iteration is completed when no more point is pending. After this we need to re-calculate k new centroids for the clusters resulting from the previous step. A loop has been generated. As a result of this loop we may notice that the k centers change their location in every iteration. The iteration continues until no more changes are done.

The objective function is defined as

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left| |x_i^{(j)} c_j| \right|^2$$

where,

$||x_i - v_j||$ is the Euclidean distance between x_i and v_j.

c_i is the number of data points in i^{th} cluster.

c is the number of cluster centers.

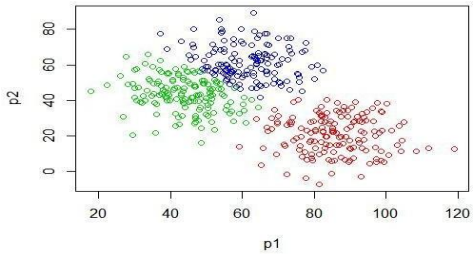

Fig 7: K-means Clustering

Algorithm

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

Step 1 : Randomly select c cluster centers.

Step 2 : Calculate the distance between each data point and cluster centers.

Step 3 : Assign the data point to the cluster center whose distance from the cluster center is minimum of all the other cluster centers..

Step 4 : Recalculate the new cluster center using:

$$V_i = (1/c_i) \sum_{j=1}^{c_i} X_i$$

Where, c_i represents the number of data points.

Step 5 : Recalculate the distance between each data point and new obtained cluster centers.

Step 6 : If no data point was changed then stop, otherwise repeat from step 3

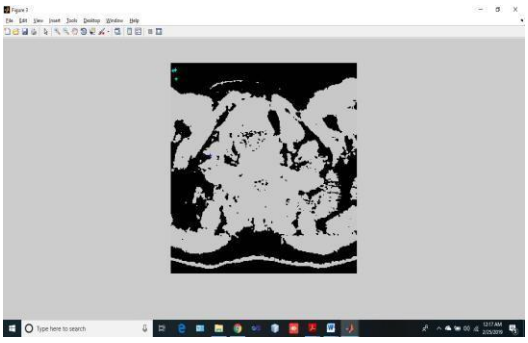

Fig 8: K-Means Clustering Implementation

2. Fuzzy C Means

Fuzzy c means is the simplest and unsupervised learning algorithms that solve the well-known clustering problem. This method is frequently used in pattern recognition. Fuzzy c means clustering is a method of vector quantization which is from signal processing. This method is popular for cluster analysis in data mining. Fuzzy c-means (FCM) is a method which produces two or more clusters by clustering a single data. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Summation of membership of each data point should be equal to one. Implementation of fuzzy c means algorithm is same as k means algorithm and the only advantage is that it works even with overlapped data.

The objective function is:

$$J_m = \sum_{i=0}^{N}\sum_{j=0}^{C} u_{ji}^{m} \left| |x_{i-} c_j| \right|^2$$

Where,
m is any real number greater than1,
N is the number of data,
C is the number of clusters,
u_{ij} is the degree of membership of x_i in the cluster j,
x_i is the ith of d-dimensional measured data,
c_j is the d-dimension center of the cluster

Algorithm:

Step 1 : Randomly selects cluster center

Step 2 : calculate the fuzzy membership 'μ_{ij}' using formula:

$$\mu_{ij} = 1/\sum_{k=1}^{c} (d_{ij}/d_{ik})^{(\frac{2}{m}-1)}$$

Step 3 : Compute the fuzzy centers 'v_j'

Step 4 : Repeat step 2 and 3 until the minimum 'J' value is achieved

where,
k is the iteration step.
β is the termination criterion between [0, 1].
U = $(\mu_{ij})_{n*c}$' is the fuzzy membership matrix. J is the objective function.

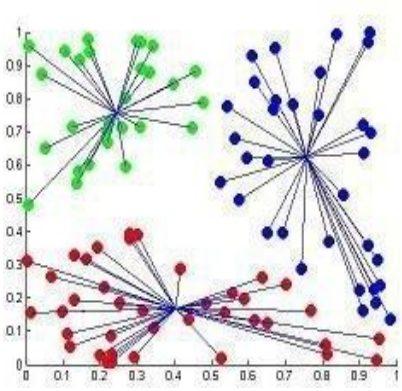

Fig 9: Fuzzy C Means Clustering

The output of this algorithm will be like the above image Fuzzy c means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. This algorithm gives best result for overlapped data set and comparatively better then k-means algorithm.

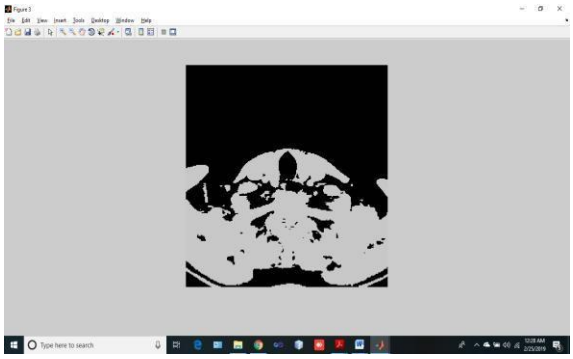

Fig 10: Fuzzy C Means Implementation

IV. CONCLUSION

Thyroid disease is the second largest disease in the endrocrine field. The correct classification of disease is important part of clinical diagnosis. In this study, thyroid disease data set is clustered based on fuzzy algorithm. In this fuzzy k-means algorithm used for classification of thyroid disease .The data is first pre-proccessed, selected, extracted and then classified defect or normal class. This algorithm gives best result for overlapped data also.

REFERENCES

[1] Anil and Antony, "An Intelligent System for Thyroid Disease Classification and Diagnosis" IEEE-2018.

[2] Veenita,Kunwar Sai and Sabitha. "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques" Springer-February 2017.

[3] Aiyun Zhou,Lili Zhang, Cheng Zhang, "Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images" Springer 2017.

[4] Geetha,Santhosh Baboo., "An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayseian Prediction Method", Global Journals Inc. (USA) 2016.

[5] Yuanyuan Zhang, Min Zuo, "Improved Ensemble Classification Method of Thyroid Disease Based on Random Forest.", IEEE - 2016.

[6] Ferreiraa Carvalho, "Kernel fuzzy c-means with automatic variable weighting ", Springer, 2014.

[7] Huang M and Xia Z, "The range of the value for the fuzzifier of the fuzzy c-means algorithm", Indian J. of Science and Technology, vol.7-2012

[8] Fatemeh Saiti1, Afsaneh Alavi Naini "Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM",IEEE, 2009.

[9] Dai Y, Ru B " Feature selection of high-dimensional biomedical data using improved SFLA for disease diagnosis" IEEE, 2015.

[10] M. Sato-Ilic, "Comparitive Analysis Of Fuzzy In Segmentation," Procedia Computer Science 6 (2011) 358–363

[11] K.Geetha , Capt. S. Santhosh Baboo, "An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayseian Prediction Method", Global Journal of Computer Science and Technology: E Network, Web & Security Volume 16 Issue 1 Version 1.0 Year 2016.

[12] Qin Yu, Tao Jiang, Aiyun Zhou, Lili Zhang, Cheng Zhang, Pan Xu, "Computer Aided Diagnosis of malignant or benign thyroid nodes based on ultrasound images", Springer, 2017.

[13] Khushboo Chandel, Veenita Kunwar , Sai Sabitha ,Tanupriya Choudhury "A comparitive study on thyroid disease detection using k-nearest neighbor and naïve bayes classification techniques" Springer, 2017.

[14] Jamil Ahmed Chandio, M. Abdul Rehman Soomrani, "TDTD: Thyroid disease type diagnostics", Intelligent Systems Engineering, 2016 International Conference .

[15] Prasad, T. Sreenivasa Rao, M. Surendra Prasad Babu "Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms", Springer, 2015.

[16] Ali Keles, Aytürk Keles. "ESTDD: Expert system for thyroid diseases diagnosis." Expert Systems with Applications An International Journal, 2008, 34(1): 242–246.

[17] Ozyilmaz L, Yildirim T. "Diagnosis of thyroid disease using artificial neural network methods" International Conference on Neural Information Processing. 2002:2033- 2036 vol.4.

[18] Feyzullah Temurtas." A comparative study on thyroid disease diagnosis using neural networks." Expert Syst. Appl.,2009:944-949

[19] Isa I S, Saad Z, Omar S "Suitable MLP Network Activation Functions for Breast Cancer and Thyroid Disease Detection. Computational Intelligence, Modelling and Simulation", International Conference on. IEEE, 2010:39-44.

[20] Li, M. and Zhou, Z. "Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples".*IEEE*,2007.