ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

**International Journal of Innovative Research in Computer and Communication Engineering**

*(An ISO 3297: 2007 Certified Organization)*

*Website: www.ijircce.com*

**Vol. 5, Special Issue 3, April 2017**

# Chromosome Abnormality Detection Using K-Means Algorithm

Thulasi Ram. K[1], Dinakar. G[2], Vanitha.L[3]

Student, Dept. of E.C.E., Prathyusha Engineering College, Thiruvallur, Chennai, India[1]

Student, Dept. of E.C.E., Prathyusha Engineering College, Thiruvallur, Chennai, India[2]

Associate Professor, Dept. of E.C.E., Prathyusha Engineering College, Thiruvallur, Chennai, India[3]

**ABSTRACT**: The proposed system classifies the human chromosomes and detects chromosomal disorders automatically without human supervision. Chromosome image is acquired and processed, features are extracted and k-means algorithm is used for classification. Based on the classification, the numerical abnormality in chromosome is diagnosed. The typical number of chromosomes in a human cell is 46. The chromosomes are classified according to their length, width, area, entropy, standard deviation and are compared with the original values of normal chromosomes, such that changes in their dimensions are made as abnormalities in their structure. Identification, classification and presentation of 24 classes into a single picture is defined as Karyotyping. K-mean algorithm is used for classification. Any deviation from the normal karyotype, in terms of chromosome number or structure is known as chromosomal abnormality.

**KEYWORDS:** Karyotype, k-mean, Chromosome abnormality, Syndrome.

## I. INTRODUCTION

Chromosomes are complex structures located in the cell nucleus, basically the "packages" that contain the DNA. They contain thousands of genes, which govern our physical and medical characteristics, such as hair colour, blood type and susceptibility to disease. Under the microscope, chromosomes appear as thin, thread-like structures. They all have a short arm and long arm separated by a primary constriction called the centromere. The short arm is designated as ***'p'*** and the long arm as ***'q'***. Normal human beings have 22 pairs of chromosomes (designated as chromosomes 1-22) and one pair of sex chromosomes; females have two X chromosomes, while males have one X and one Y chromosome. Thus a normal human cell contains 46 chromosomes.

A Chromosome band is defined as a section of a chromosome, which shows relatively darker or lighter stain as compared to the neighboring sections of the same chromosome. All the twenty four pairs of chromosomes have a specific band pattern. Figure 1 shows the Karyotype image of a Female Chromosome.

Human chromosome analysis is an essential task in cytogenetic, especially in prenatal screening and genetic syndrome diagnosis, cancer pathology research. One of the aims of chromosome analysis is the creation of karyotype, which is used to analyze the characteristics of chromosomes and to predict many genetic disorders.

Problems that are faced by the technicians in analyzing the human chromosomes are: (1) Count the number of chromosomes (taking into consideration of overlapping and touching chromosomes), (2) In order to get correct results, this process might be repeated many times with different types of cells. Thus, both the time and effort to accomplish these tasks are relatively long.

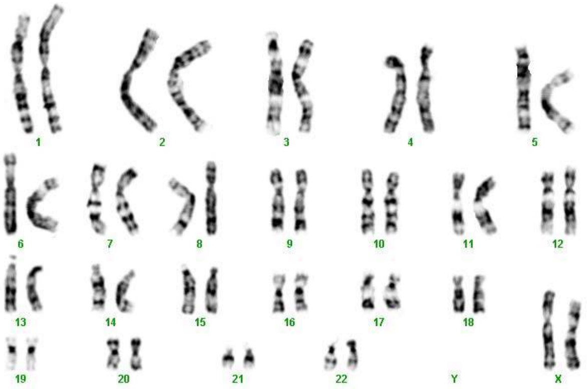

**Figure 1. Karyotype image of a Female Chromosome**

## II. METHODLOGY

Figure 2 shows the block diagram of Pattern Recognition System. The main aim of pattern recognition is the classification of patterns and sub patterns in an image. A pattern recognition system includes:

- Subsystem to define pattern class
- Subsystem to extract selected features
- Subsystem for classification known as classifier.

The classifier used in the proposed system is k-means algorithm.

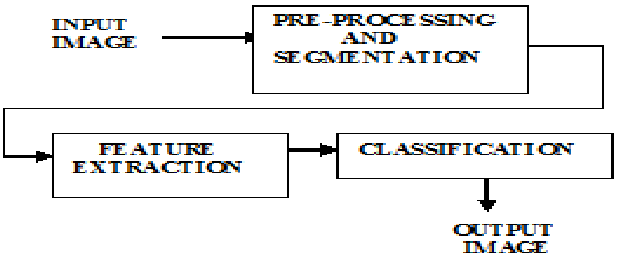

**Figure 2. Block diagram of Pattern Recognition System**

### 2.1 Modules

In the proposed system there are four modules:

Pre-processing
Feature extraction
Multi-Layer Hybrid neural network for classification
Identification of chromosome abnormalities

### 2.2 Pre-Processing

The purpose of pre-processing is to remove the noise from the image. This is required for the reliable extraction of features as feature extraction algorithms give poor results in the presence of a noisy background.

### 2.3Feature Extraction

The three main approaches for feature extraction and classification based on the type of features are as follows:

- Statistical approach
- Syntactic or structural approach

- Spectral approach.
-

In case of statistical approach, pattern is defined by a set of statistically extracted features represented as vector in multidimensional feature space. The statistical features could be based on first-order, second-order, or higher-order statistics of gray level of an image.

In case of syntactic approach, texture is defined by texture primitives, which are spatially organized according to placement rules to generate complete pattern. In syntactic pattern recognition, a formal analogy is drawn between the structural pattern and the syntax of language.

In case of spectral method, textures are defined by spatial frequencies and are evaluated by autocorrelation function of a texture.

As a comparison between the above-mentioned three approaches, spectral frequency-based methods are less efficient, while statistical methods are particularly useful for random patterns; while for complex patterns, structural methods give better results. The present approach is hybridization of syntactic and statistical approaches for texture-based segmentation and classification with neural network as a classifier tool. In this scheme, first- and second-order statistical features of the texture-primitive cell are used for segmentation and classification. In contrast with the syntactic approach, instead of using rules and grammar to represent pattern in terms of sentences, we use analysis by synthesis method.

The features are extracted by determining the medial line of a chromosome by applying the medial axis transformation (MAT). Thinning algorithm is used that iteratively deletes edge points of a region subject to the constraints that deletion of these points does not remove end points, does not break connectedness.

The following are the different features extracted from chromosomes:

(1) Relativelength
(2) Relative Area
(3) Centromericindex(C.I.)

First-order primitives

(4) Mean
(5) Standard deviation
(6) Entropy

Second-order statistical features based on gray-level co-occurrence matrices (GLCMs) computed for primitive texture cell.

(7) Contrast
(8) Correlation
(9) Variance

**1. Relativelength:**

The length of each chromosome is determined b counting the number of pixels in the medial line. The relative length of the i-th chromosome($l_{ri}$) can be obtained by normalizing theme dial axis length using the following equation.

$$l_{ri} = \frac{l_i}{l_t} \qquad \text{--- (1)}$$

where$l_i$ ($i$=1, 2… 24)is the length of $i$-th chromosome

$l_t$ is the total length of all 46 Chromosomes of one cell.

**2. Relative area:**

The relative area of the i-th chromosome($A_{ri}$)can be obtained by counting the pixels of the chromosomebody and by normalizing the areas usingthe following equation.

$$A_{ri} = \frac{A_i}{A_t} \qquad \text{--- (2)}$$

where$A_i$ ($i$=1,2,…,24)is the area of $i$-th chromosome and $A_t$ is the total area of all 46 chromosomes of onecell.

**3. Centromericindex(C.I.):**

$$\text{C.I.} = \frac{\text{short arm length}}{\text{whole length of medial axis}} \qquad \text{--- (3)}$$

**4. Mean:**

--- (4)

where $p_{r,s}$ is pixel at location (r,s).

**5. Standard deviation:**

--- (5)

**6. Entropy:**

It is a measure of randomness

--- (6)

where $p(b) \approx N(b)/n^2$ for $\{0 \le b \le L\text{-}1\}$,

L is the number of different values which pixels can adopt

N(b) = number of pixels of amplitude (b) in the pixel window of size 'n×n'.

The co-occurrence matrix characterizes the spatial interrelationships of the gray tones in an image. The values of the co-occurrence matrix elements present relative frequencies with which two neighbouring pixels are separated by distance d and at angle θ appear on the image. One of them has gray level i and other j, and their joint probability of occurrence is given by Pi,j.

**7. Contrast**:

Contrast is defined as

$$\sum_{n=0}^{N_g} n^2 , \qquad \text{--- (7)}$$

$$\text{where } n = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P_{i,j} \cdot |i-j| \qquad \text{--- (8)}$$

Ng = number of gray levels.

**8. Correlation:**

The correlation feature is a measure of gray-level linear dependency of the image.

Correlation feature is defines as

$$\text{Correlation} = \frac{\sum_i \sum_j (i,j) P_{i,j} - \mu^2}{\sigma^2} \qquad \text{--- (11)}$$

where μ and $\sigma$ are the mean deviation and standard deviation of the co-occurrence matrix respectively.

**9. Variance**:

Variance is the measure that tells us by how much the gray level are varying from the mean.

$$\text{Variance} = \sum_i \sum_j (i,j) P_{i,j} - \mu^2 \qquad \text{--- (13)}$$

## 2.4 Chromosome classifier

Chromosome classification is done using k-means algorithm. The K means algorithm will do the following three steps until convergence

Iterate until *stable* (= no object move group):

1) Determine the centroid coordinate

2) Determine the distance of each object to the centroids
3) Group the object based on minimum distance (find the closest centroid)

### 2.5 Chromosome Abnormality Detection

Chromosome abnormalities or anomalies usually occur when there is an error in cell division. Any deviation from the normal karyotype, in terms of chromosome number or structure, is known as a chromosome abnormality. The chromosomal abnormalities are classified as: Numerical abnormalities and Structural abnormalities.

**Numerical Abnormality:**

When an individual is missing either a chromosome from a pair (monosomy) or has more than two chromosomes of a pair (trisomy), it is called Numerical Abnormality. The set of abnormalities associated with specific symptoms is known as Syndrome. An example of a condition caused by numerical abnormalities is Down syndrome, three copies of chromosome 21, rather than two. Turner Syndrome is an example of monosomy, where the individual - in this case a female - is born with only one sex chromosome(X) instead of normal two X chromosome.

**Structural Abnormalities:**

When the chromosome's structure is altered it is called as Structural Abnormalities. This can take several forms as deletions, duplications, translocations inversions, rings.

This paper, deals only with Numerical abnormality. The output of the neural network is used to determine the Numerical abnormality. If the number of chromosomes in a group is less than or greater than the required number, then numerical abnormality is reported.

## III. IMPLEMENTATION RESULTS

The input to the feature extraction algorithm is the chromosome images. The pattern vectors (features) extracted from the images is given as input to the SOM neural network classifier. Large database are required for the classifier to perform the classification correctly. In this system a sample of 100 chromosome images each consisting of 46 chromosomes is given as input to the SOM. The Classification accuracy and error rate is obtained by using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correctly classified Chromosomes}}{\text{Total number of Chromosomes}} \quad \text{--- (16)}$$

$$\text{Error Rate} = \frac{\text{Number of misclassified Chromosomes}}{\text{Total Number of Chromosomes}} \quad \text{--- (17)}$$

| ALGORITHM | CLASSIFICATION ACCURACY | ERROR RATE |
|---|---|---|
| K-Mean | 84 % | 16 % |

**Chromosome Abnormality Detection**

The output of the Second stage classifier is checked for abnormal number of chromosomes. The abnormal Female chromosomes showing Turner’s syndrome is shown in Figure 5. The output of this stage is shown in Table 4 which shows only one X chromosome instead of two X chromosomes.

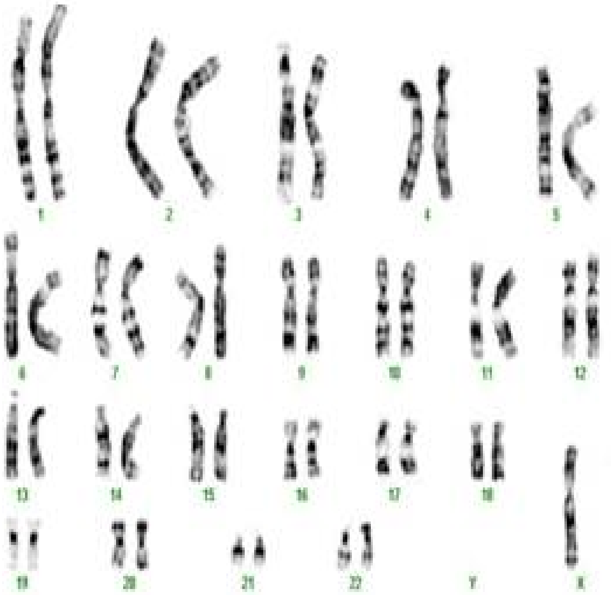

**Figure 5. Turner's syndrome – One X Chromosome missing in Female**

### 3.3. Tools for Experiment

In this study, MATLAB is used as a programming language for feature extraction, chromosome classification and detection of chromosomal abnormalities.

**Table 4. Chromosome Abnormality diagnosis**

| Class Number | Number of Chromosomes | Class Number | Number of Chromosomes |
|---|---|---|---|
| 1 | 2 | 13 | 2 |
| 2 | 2 | 14 | 2 |
| 3 | 2 | 15 | 2 |
| 4 | 2 | 16 | 2 |
| 5 | 2 | 17 | 2 |
| 6 | 2 | 18 | 2 |
| 7 | 2 | 19 | 2 |
| 8 | 2 | 20 | 2 |
| 9 | 2 | 21 | 2 |
| 10 | 2 | 22 | 2 |
| 11 | 2 | X | 1 |
| 12 | 2 | Y | 0 |
| Total Number of Chromosomes : 45 | | | |
| **Diagnosis :** Patient is suffering from Turner's Syndrome | | | |

## IV. CONCLUSION AND FUTURE WORK

The classifier K- Mean's algorithm is used to classify the chromosome data and the system is capable to detect the numerical abnormalities of the chromosomes. This work concentrates with the diagnosis of numerical abnormality of chromosomes. This work can be extended to diagnose different structural abnormality of chromosomes in future.

## REFERENCES

[1] Guisong Liu and Xiaobin Wang, " An Integrated Intrusion Detection System by using Multiple Neural Networks", IEEE,pp22-28,2008

[2] Syed Zahid Hassan and Brijesh Verma, "A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases", Seventh International Conference on Intelligent Systems Design and Applications,pp.503-507,2007

[3] J.Cho, S.Y.Ryu, S.H.Woo,"A Study for the Hierarchical Artificial Neural Network Model for Giemsa stained Human Chromosome Classification", IEEE conference,pp.4588-4591,2004.

[4] Ibrahiem M. M. El Emary," On the application of Artificial Neural Networks in Analyzing and classifying the Human Chromosomes", Journal of Computer science 2 (1): 72-75, 2006,ISSN 1549-3636 © 2006 Science Publications

[5] B.D.Singh, "Genetics", Kalyani Publishers, 2005.

[6] Earl Gose, Richard Johnson-baugh, Steve Jost, "Pattern recognition and image analysis", Prentice Hall of India Private Limited, New Delhi,2000

[7] Laurene Fausett, Fundamentals of Neural Networks, Pearson Education,2007

[8] Simon Haykin, Neural Networks – A comprehensive foundation, Pearson Education,2001

[9] Robert C. Gonzalez, Richard E.Woods, Steven L.Eddins, Digital Image Processing using MATLAB, Pearson Education, 2005

[10] Duane Hanselman, Bruce Littlefield, Mastering MATLAB 7, Pearson Education, 2008.

[11] http://www.genome.gov/

[12] http://www.ncbi.nlm.nih.gov

[13] www.genetics.org

[14] www.ias.ac.in/jgenet

[15] http://www.vivo.colostate.edu/hbooks/genetics/medgen/index.html