VOL. 10, NO. 10, JUNE 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CLASSIFICATION OF SEX CHROMOSOMES USING SVM CLASSIFIER

S. Saranya, P. S. RamaPraba, S. Sathiya Priya and V. Loganathan
Panimalar Institute of Technology, Chennai, Tamil Nadu, India
E-Mail: Saranyas2008@gmail.com

ABSTRACT

Scrutinize of chromosome is a preliminary procedure to detect the most characteristic signs of a disorder that may require for further investigation of medical applications mainly for cancerous. Diagnosis of karryogram is generally very complex, eroding and a time consuming operation. As of now it requires fussy attention to details and calls for meritoriously and trained personnel. Normally chromosomes are essential genomic information carriers which contain 23 pairs. This paper suggests a efficient classifier Support Vector Machine (SVM) for classifying chromosomes of 23 and 24 chromosome which is the sex chromosome in which 23 is X chromosome and 24 is Y chromosome in comparison to the already existing methods such as support vector machine based medial axis and density profiles. The features are extracted based on GLCM (Gray level co-occurrence) feature extraction algorithm which is very effective well known for its high accuracy. X chromosome features and Y chromosome features are extracted based on very effective GLCM algorithm from the segmented image and GLCM features of chromosomes are extracted from the segmented image. As a prerequisite, image segmentation needs to be done by using Fuzzy - C Mean (FCM) procedure to obtain efficient features in coordination with SVM which is used to classify the chromosomes from the available pairs of 23 chromosomes. Using this methodology increased the accuracy of classification results. Simulation results are carried out in MATLAB to support the analysis.

Keywords: chromosome, GLCM, support vector machine (SVM), karryogram, fuzzy C Mean.

1. INTRODUCTION

Every living organism carries the genital information from their parents through a substance called DNA. DNA molecule from cell Nucleus is packed into a thread like structure is known chromosomes. DNA from each Chromosomes are tightly coiled many times sourrounded by proteins support this called histones its structure is shown in Figure-1. Most cells in the human body have pairs of 23 chromosomes, with 46total chromosomes. One from these pairs of chromosome one pair is the gender identification chromosomes (XY: Male and XX: Female), plus some other body characteristics, and the other 22 pairs are autosomal chromosomes (determine the rest of the body's makeup) [2]. Many attempts have been made to improve visual analysis as humans are prone to errors, the need for cytogeneticists and biologist has been made it a tedious and time consuming process to assign chromosome to a class (karyotyping). Figure-1 shows the structure of the chromosome in which telomeres are special structures that can act as a cap and protect the very ends of the area next to each telomere may contain these important genes for growth and development. Each chromosome arm Chromosomes.

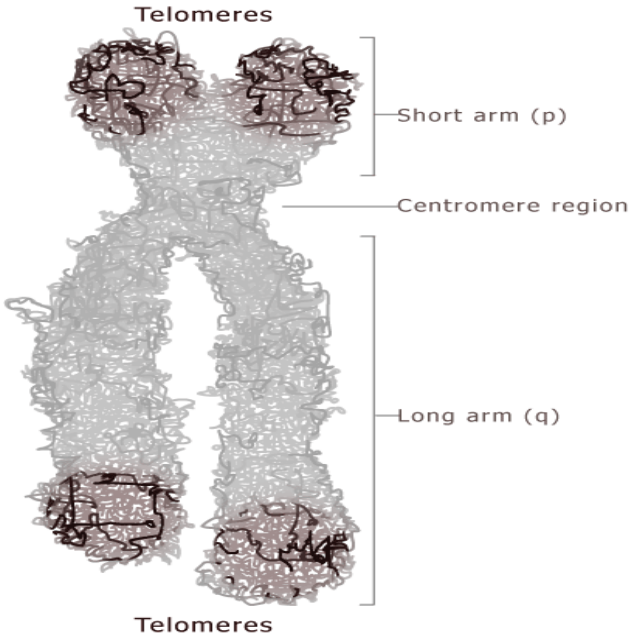

Figure-1. Structure of chromosome.

Telemere is present at the end of the Chromosome body which is repeating DNA sequence. To prevent from losing base pair sequence at the end of chromosome telemers are used. They can avoid chromosome fusing each other, whenever chromosomes are separated few telemers are lost because telemers are very short which reaches a critical length there is no replicate. Diseases identified by chromosomes are Turner syndrome (one of the sex chromosome is missing), Down

syndrome more than one copies of each chromosome), Norrie disease (mutation in the gene on the x chromosomes), kinefelter syndrome (chromosomes look like XXY or XXXY), Jacobs syndrome (chromosome looks like XYY), Edwards syndrome (trisomy of 18^{th} chromosome), Patau syndrome (trisomy of 13^{th} chromosomes).

The different Classification of Karyotyping is based on various parameters such as surface area or size, length, shape [3]. Automatic classification of chromosomes images based on a set of pictorial band pattern models of the various chromosome types known as ideograms. These ideograms, illustrate artists' depiction of the chromosomes, which were published as an international system for cytogenetic nomenclature (ISCN) The resulting image is a karyotype image in which all the chromosomes are classified and arranged into a standard display format The identify the chromosomes a multichannel water shed Segmentation [4] and ROI classification technique were used to decompose into a set of homogeneous regions using Bayes classifier which has an accuracy of 82.5%.

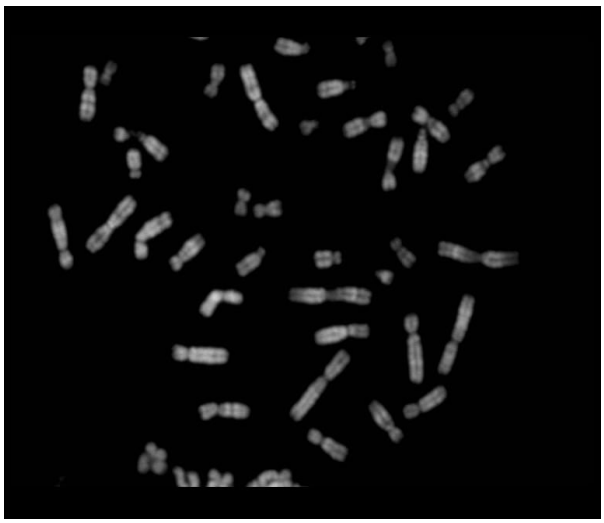

Figure-2. Metaphase chromosome spread image.

A special DNA probes are used to observe the metaphase slides as shown in Figure-2. The fluorescent nucleic acid segment facilitates the notable sites on the chromosomes.

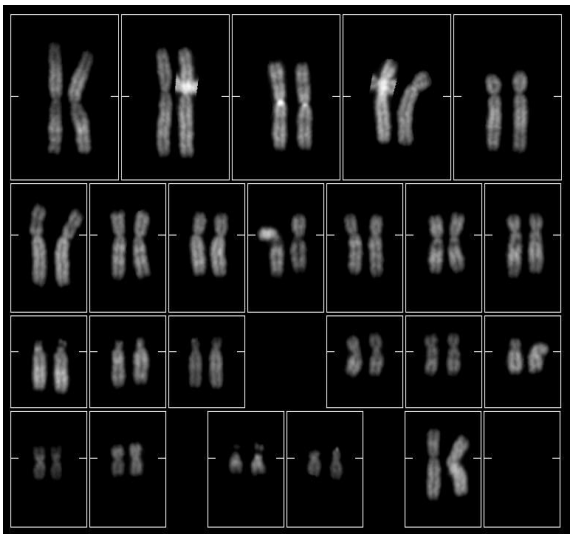

Figure-3. Metaphase chromosome spread image and its karyotype.

Metaphase chromosome images are categorized into 23 pairs (karyotyping). The features are extracted by GLCM and these are classified using SVM classifier.

2. PROPOSED METHOD

The purport of this approach addresses how we can make use of the topological spatial representation to achieve the objectives of epitomizing and relegating chromosome images.

2.1 Segmentation

For classification of the chromosomes segmentation method has been applied. First and foremost step for this is extraction of chromosome features. Prior to that pre processing is carried out to eliminate correcting divergence in zoom for contrast and illumination by means of image normalization. The Figure-4 shows first and second chromosomes. For further processing these images are being used.

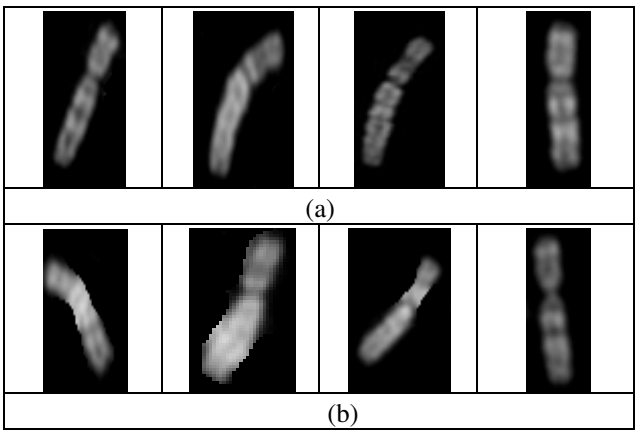

Figure-4. Input images (a) First chromosomes (b) Second chromosomes.

Figure-5 shows the block diagram of the intended chromosome classification, The input image of size (517 x 645 pixels) are fed in to the median filter for preprocessing and the output image of preprocessing and enhanced image is given to the segmentation using Fuzzy c mean algorithm, these segmented output are fed to (GLCM) Grey Level Co-occurrence Matrices are used for feature extraction and then for classification SVM classifier has been used

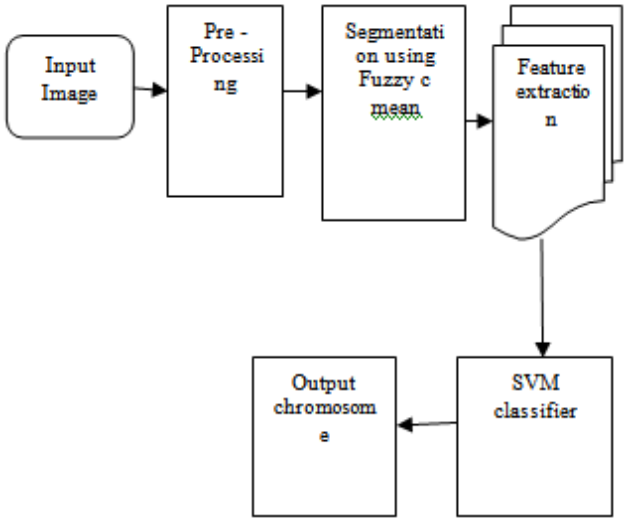

Figure-5. Block diagram for proposed chromosome classification.

2.2 Feature extraction

The important characteristic of an image is a Texture which is distinguished the regions of interest in an image. Grey Level Co-occurrence Matrices (GLCM) is one of the best list and earliest methods in statistical texture analysis. A GLCM is a matrix the number of rows and columns in the pixel is equal to the number of gray levels, G, in the image. The GLCM's are of large dimensionality so that it is sensitive to the size of texture samples that are being evaluated. Steps for generating a symmetrical normalized GLCM are given below:

1. Create an image framework matrix
2. Spatial relation has been decided with reference pixel and neighbour.
3. Count the occurrences of the pixel and fill in the image matrix.
4. To make it symmetrical add the matrix to its transpose.
5. Make the Matrix probabilities into Normalization.

GLCM operation explained with the example shown in Figure-5 the image matrix consists of four different gray levels. Here one pixel offset is used (a reference pixel and its immediate neighbour). If the window is large enough, using a larger offset is possible.

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

Figure-6. The image grey levels (GLs).

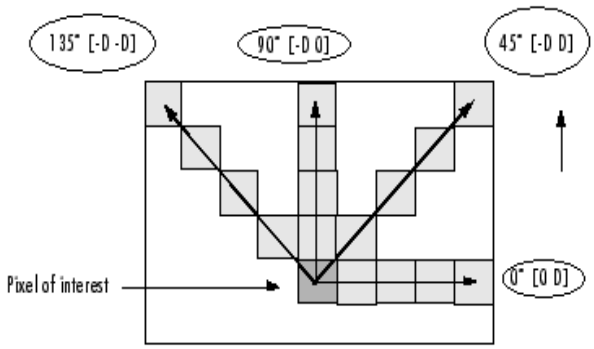

Figure-7. Spatial relation of a pixel.

2	2	1	0
0	2	0	0
0	0	3	1
0	0	0	1

Figure-8. Spatial relationship of GLCM.

2.2.1 Spatial relation of a pixel

The calculation of the image texture symmetrical matrix is needed. To get the symmetrical matrix from the GLCM the merging of different parts of horizontal matrix (2, 2) in the image matrix more likely than (2, 3) Looking at the horizontal GLCM shows that the merging of different parts of image matrix 2, 2 which exist 6 times from the 24 horizontal combinations of pixels in the image 12 direction of east + 12 direction of west. In other hand, 6 is the pixel entry in the horizontal GLCM in the column of third (headed reference pixel value 2) and the row of third (headed neighbor pixel value 2). The definition of the probability of a given matrix resultant is depending on the ratio of the number of times this outcome occurs, by the total number of possible outcomes. The (2, 2) combination repeats 6 times from 24, for a probability of 1/4 or 0.250. The probability of 2, 3 is 1/24 or .042. The equation (1) of the transform shows the GLCM into a close approximation of a probability table. It is only an approximation because an original value probability would require continuous values, and the grey levels are integer values, so they are discrete.

$$P_{i,j} = \frac{V_{i,j}}{\sum_{i,j=0}^{N-1} V_{i,j}} \quad (1)$$

This kind of process is called normalizing the matrix. The Equation (1) indicates the Normalization involves dividing by the sum of values gives the Normalizing of the GLCM Usually, used by extracting secondary features. Features are usually strongly correlated, using more than 4-5 simultaneously is not advisable Need to evaluate several distances d optimal set of features is problem dependent. It is also advisable to preprocess by histogram transform to remove effect of absolute gray level Usually, we want to make the features "rotation" invariant by using the isotropic GLCM Feature are given in the Table-1.

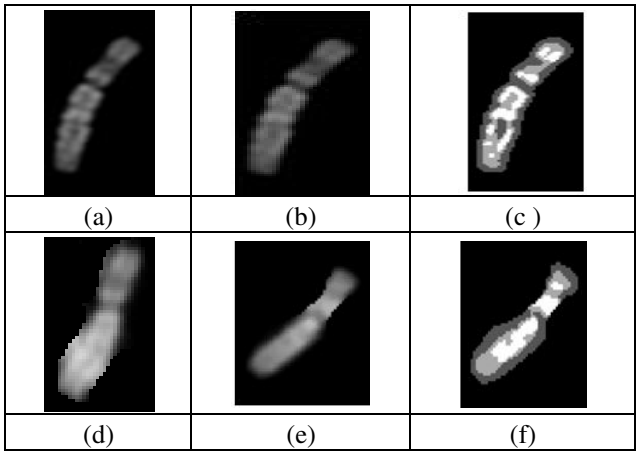

Figure-9. First rows shows the First chromosome (a) original image (b) Preprocessing image (c) Segmented image second row second chromosome (d) original image (e) Preprocessing image (f) Segmented image.

2.3 Classification

SVM classifier has been used to classify the first and second chromosome. Optimum linear separating hyper plane has been used to separate two sets of features. This hyper plane produce maximum and minimum margin between two sets. Outcome of hyper plane is based on border training pattern called support vector. The kernel function determines the efficiency of the SVM. This determines the classification decision function by minimizing the empirical risk, as

$$R = \frac{1}{l}\sum_{i=1}^{L} |f(x_i) - y_i| \quad (2)$$

Table-1. GLCM second order statistical features formula.

Feature	Formula
Mean	$\sum_{i=0}^{L-1} (x_i * p(x_i))$
Variance	$\sum_{i=0}^{L-1} [(x_i - m)^2 * p(x_i)]$
Standard Deviation	$\sqrt{\{\sum_{i=0}^{L-1} [(x_i - m)^2 * p(x_i)]\}}$
Kurtosis	$\sum_{i=0}^{L-1} [(x_i - m)^4 * p(x_i)]$
Skewness	$\sum_{i=0}^{L-1} [(x_i - m)^3 * p(x_i)]$
Entropy	$-\sum_{i=0}^{L-1} (p(x_i) * \log 2 p(x_i))$
Contrast	$VAR = \sum_{i=0}^{G-1} n^2 \{ \sum_{j=0}^{G} (P(i,j)) \}$
Sum of Squares and variance	$VAR = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} (i-\mu)^2 P(i,j)$
Cluster shade	$SHD = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \{i + j - \mu_x - \mu_y\}^3 \times P(i,j)$
Cluster Prominence	$PRM = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \{i + j - \mu_x - \mu_y\}^4 \times P(i,j)$
Difference Entropy	$DEN = -\sum_{i=0}^{G-1} P_{x+y}(i) + \log(P_{x+y}(i))$
Inertia	$INR = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} \{i-j\}^2 P(i,j)$
Sum of Entropy	$SEN = -\sum_{i=0}^{2G-2} P_{x+y}(i) + \log(P_{x+y}(i))$
Difference of Entropy	$DEN = -\sum_{i=0}^{G-1} P_{x+y}(i) + \log(P_{x+y}(i))$
Sum of Average	$AVE = \sum_{i=0}^{2G-2} iP_{x+y}(i)$
Information measure of correlation 1	$IMC1 = \frac{H_{XY} - H_{XY}}{MAX\{HX, HY\}}$
Information measure of correlation 2	$IMC2 = \sqrt{1 - \exp[-2.0(HXYZ - HXY)}$

Where l and f represent the size of examples and the classification decision function, respectively.

In SVM, an optimal separating hyperplane is the primary concern is determines a low-generalization error. A typical SVM classification decision function in the linearly separable problem is

$$f_{\overline{w,b}} = sig(w, x+b) \quad (3)$$

Whose separating hyperplane is determined by giving the largest margin of separation between different classes? This hyperplane bisects the shortest line between the convex hulls of the two classes, thus is required to satisfy the following constrained minimization as:

$$Min = \frac{1}{2} w^T w$$
$$y_i (w, x_i + b) \geq 1 \quad (4)$$

Table-2. Second order values of first and second chromosome.

	IMAGES #	AUTOCORRLATION	CONTRAST	CORR	CORR	CLUSTERPROMINENCe	CLUSTERSHADe	Dissimilarity	Energy	Entropy	Homog	Homog	Maximum probability	Variance	Sum average	SumVariance	Sum entropy	Difference variance	Difference entropy	corr1	corr2	Inverse difference	(INN)
FIRST CHROMOSOME	1	7.79	0.81	0.88	0.88	539.48	59.34	0.45	0.39	2.01	0.83	0.81	0.62	8.12	4.35	20.51	1.61	0.81	0.89	-0.44	0.82	0.95	0.99
	2	5.18	0.71	0.86	0.86	423.06	51.18	0.35	0.54	1.49	0.87	0.86	0.74	5.48	3.51	14.39	1.21	0.71	0.75	-0.42	0.74	0.96	0.99
	3	4.62	0.56	0.87	0.87	393.90	48.74	0.30	0.61	1.26	0.89	0.88	0.78	4.85	3.30	13.25	1.04	0.56	0.65	-0.46	0.73	0.97	0.99
	4	7.51	0.47	0.93	0.93	543.96	62.47	0.30	0.44	1.76	0.87	0.87	0.66	7.68	4.20	20.32	1.47	0.47	0.70	-0.53	0.85	0.97	0.99
	5	8.85	1.76	0.75	0.75	321.64	33.36	0.80	0.22	2.55	0.73	0.69	0.46	9.65	4.97	21.51	1.96	1.76	1.20	-0.27	0.74	0.92	0.98
	6	5.39	0.61	0.88	0.88	527.70	60.16	0.31	0.56	1.44	0.88	0.87	0.75	5.63	3.52	15.25	1.18	0.61	0.69	-0.45	0.75	0.97	0.99
	7	4.75	0.53	0.88	0.88	442.32	52.57	0.26	0.61	1.29	0.90	0.89	0.78	4.96	3.32	13.65	1.05	0.53	0.62	-0.48	0.74	0.97	0.99
	8	10.45	0.98	0.88	0.88	458.22	42.62	0.53	0.27	2.36	0.80	0.78	0.51	10.85	5.24	26.53	1.89	0.98	0.98	-0.41	0.84	0.95	0.99
	9	8.76	0.82	0.89	0.89	633.94	66.04	0.47	0.32	2.22	0.81	0.80	0.56	9.10	4.63	22.46	1.79	0.82	0.91	-0.43	0.84	0.95	0.99
	10	5.68	0.84	0.85	0.85	490.60	57.63	0.41	0.52	1.60	0.86	0.84	0.72	6.04	3.66	15.79	1.29	0.84	0.81	-0.40	0.74	0.96	0.99
SECOND CROMOSOME	1	3.81	0.50	0.86	0.86	336.07	40.70	0.26	0.65	1.15	0.91	0.90	0.80	4.01	3.05	10.81	0.96	0.50	0.60	-0.44	0.69	0.97	0.99
	2	5.72	0.59	0.89	0.89	408.57	50.18	0.33	0.50	1.61	0.87	0.86	0.70	5.96	3.71	15.44	1.32	0.59	0.72	-0.45	0.78	0.97	0.99
	3	4.81	0.58	0.87	0.87	389.49	48.63	0.32	0.58	1.36	0.88	0.87	0.76	5.05	3.38	13.54	1.12	0.58	0.70	-0.45	0.74	0.97	0.99
	4	6.11	0.71	0.87	0.87	395.55	49.38	0.38	0.47	1.69	0.86	0.84	0.68	6.40	3.85	16.48	1.37	0.71	0.80	-0.44	0.78	0.96	0.99
	5	7.69	1.07	0.84	0.84	412.00	47.68	0.55	0.34	2.16	0.80	0.78	0.57	8.19	4.44	19.63	1.70	1.07	0.99	-0.37	0.79	0.94	0.98
	6	6.58	0.80	0.87	0.87	513.13	58.84	0.43	0.44	1.84	0.84	0.82	0.66	6.92	3.97	17.54	1.48	0.80	0.86	-0.42	0.79	0.96	0.99
	7	7.87	0.56	0.91	0.91	314.00	37.55	0.38	0.33	2.04	0.84	0.83	0.56	8.07	4.52	19.66	1.68	0.56	0.81	-0.47	0.85	0.96	0.99
	8	4.77	0.33	0.93	0.93	470.34	55.23	0.20	0.63	1.18	0.92	0.91	0.79	4.89	3.29	13.84	1.00	0.33	0.53	-0.54	0.77	0.98	1.00
	9	6.41	0.81	0.87	0.87	521.76	60.05	0.41	0.46	1.76	0.85	0.83	0.68	6.75	3.90	17.40	1.42	0.81	0.84	-0.40	0.77	0.96	0.99
	10	7.14	0.85	0.87	0.87	536.92	61.88	0.46	0.43	1.85	0.83	0.81	0.65	7.50	4.12	19.34	1.49	0.85	0.88	-0.41	0.79	0.95	0.99

For the linearly non separable case, this modification results in a soft margin classifier that allows but penalizes errors by introducing a new set of variables $\varepsilon_{i=1}^{l}$ as the measurement of violation of the constraints

$$Min = \frac{1}{2} w^T w + c\sum_{i=1}^{l} z_i, \; y_i(w, x_i + b) \geq 1 - z_i \quad (5)$$

For minimizing the equation (4) term variable C is used to minimize the second term of equation (4) VC dimension is used which controls empirical risk therefore to minimize the classical risk lagrangians method used in Equation (5). Then, (4) can be written as (5) where $\Lambda = (\lambda, \ldots, \lambda_l), D = y_i y_j x_i x_j$ for binary classification and the decision function (1) can be changed as

$$f(x) = R\sum_{i=1} y_i \lambda_i K(x, x_i) + b^* \quad (6)$$

For single SVM K is a kernel function used. By the experience and to guarantee a fair comparison, we are using the same second-order polynomial kernel in all SVMs for our experiments.

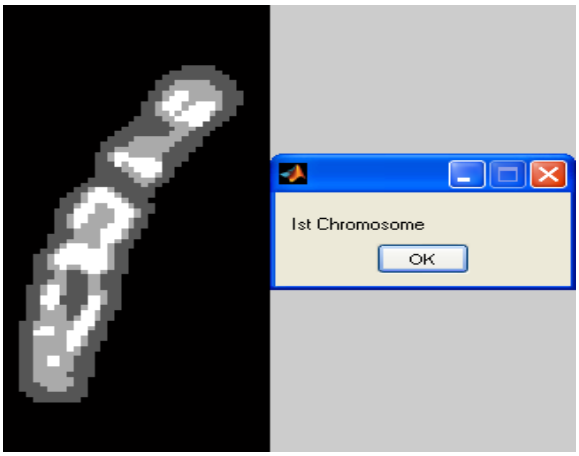

Figure- (a)

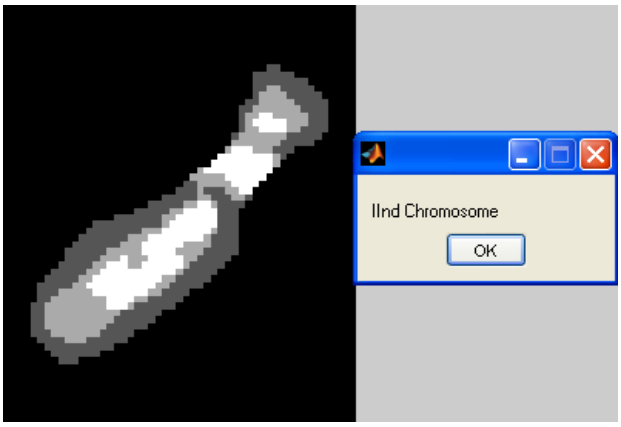

Figure- (b)

Figure-10. Classifier output (a) First chromosomes (b) Second Chromosomes

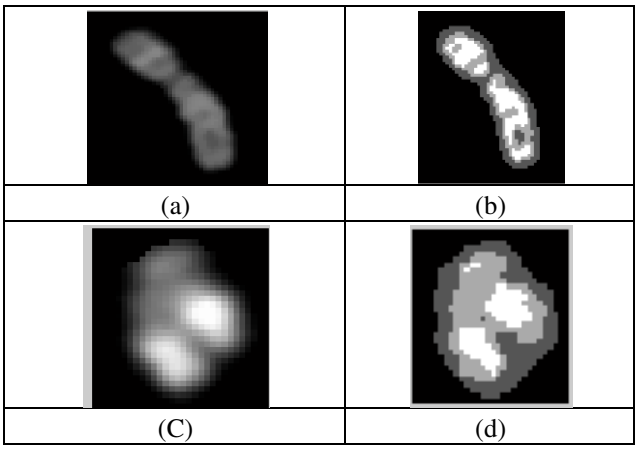

Figure-11. (a) Original image of X chromosome (b) Segmented image (c) Original image of Y chromosome (d) Segmented image.

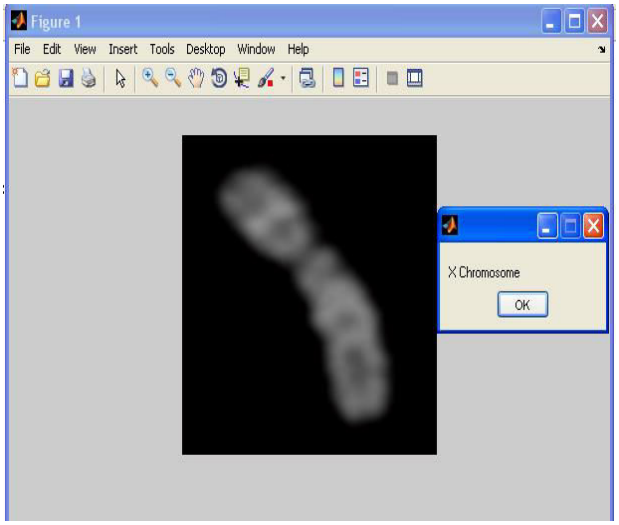

Figure-12. X chromosome classification.

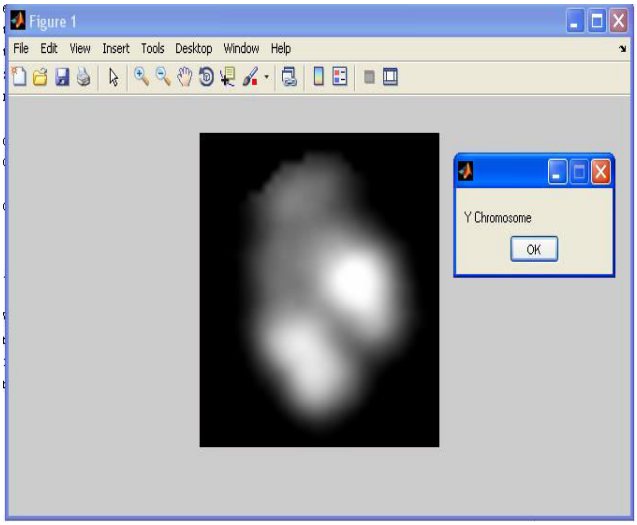

Figure-13. Y chromosome classification.

Table-3. GLCM features of X chromosome and Y chromosome.

Chromosome	X-Chromosome				Y-Chromosome			
Description	Image 1	Image 2	Image 3	Image 4	Image 1	Image 2	Image 3	Image 4
autoc	4.95	4.85	5.92	4.99	13.34	13.46	13.57	13.38
contr	0.69	0.73	0.69	0.69	1.91	1.81	1.99	1.97
corrm	0.84	0.74	0.84	0.84	0.80	0.77	0.79	0.80
corrp	0.84	0.82	0.84	0.84	0.80	0.82	0.79	0.77
cprom	279.16	275.16	260.16	278.16	511.58	540.58	550.58	520.58
cshad	38.26	37.26	38.26	38.26	31.07	33.07	31.07	30.07
dissi	0.36	0.36	0.36	0.36	0.90	0.90	0.90	0.90
energ	0.51	0.51	0.51	0.51	0.14	0.14	0.14	0.14
entro	1.55	1.85	1.45	1.61	2.96	2.82	2.60	2.74
homom:	0.87	0.87	0.87	0.87	0.69	0.68	0.79	0.63
homop	0.85	0.85	0.85	0.85	0.65	0.65	0.65	0.65
maxpr	0.71	0.71	0.71	0.71	0.36	0.36	0.31	0.38
sosvh	5.24	5.24	5.24	5.24	14.16	14.16	14.16	14.16
savgh	3.53	3.53	3.53	3.53	14.16	6.19	6.19	6.19
svarh	13.23	13.59	13.29	13.61	14.16	32.39	32.26	32.39
senth	1.25	1.25	1.25	1.25	14.16	2.26	2.26	2.26
dvarh	0.69	0.69	0.69	0.69	14.16	1.91	1.91	1.91
denth	0.76	0.76	0.76	0.76	14.16	1.27	1.27	1.27
inf1h	-0.40	-0.40	-0.40	-0.40	14.16	-0.22	-0.28	-0.22
inf2h	0.73	0.73	0.73	0.73	14.16	0.69	0.78	0.75
indnc	0.96	0.97	0.95	0.97	14.16	0.96	0.97	0.95
idmnc	0.99	0.98	0.97	0.98	14.16	0.10	0.95	0.94

3. RESULTS AND DISCUSSIONS

SVM classifier has been used for fast and accurate resuts compared with other methods. SVM determines highest decision value, this method isgives good quality result when the data set is large. In this SVM techniques data are divided into two sets one is training whereas the other one is a test set which is shown in Figure-10.

The experiment is performed two times by replacing the train and test sets. The results displayed an average of 95.89% in chromosome. Table-2 shows the GLCM features of sex chromosome. Figure-10 shows the classified image chromosome I and Chromosome II and also the Figure-14 shows simulation results of classification of chromosomes (class1 and class 2) using SVM.

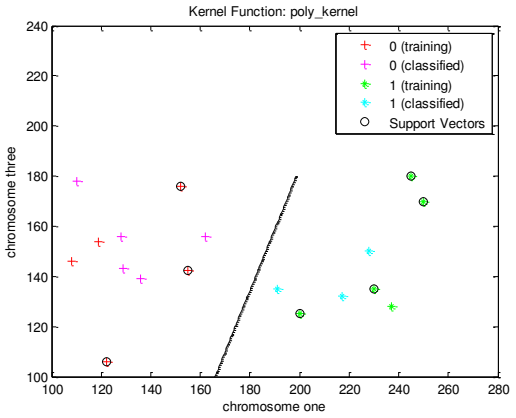

Figure-14. Classification of chromosomes using SVM.

REFERENCES

[1] J. W. Butler, M. K. Butler, and A. Stroud, "Automatic classification of chromosomes," Data Acquisition and Processing in Biology and Medicine, vol. 3, pp. 261-275, 1964.

[2] M. Moradi, S. Setarehdan, and S. Ghaffari, "Automatic Classification of Group E Chromosomes Simulating the Human Expert Method," Proceedings of the 5th Iranian Conference on Intelligent Systems (in persian), Mashad, Iran, pp. 425-432, October 2003.

[3] B. Lerner, "Toward a completely automatic neural network based human chromosome analysis," IEEE Transactions on Systems, Man, and Cybernetics Special issue on Artificail Neural Networks, vol. 28, pp. 544-552, 1998.

[4] A Multichannel Watershed-Based Segmentation Method for Multispectral Chromosome Classification Petros S. Karvelis, Student Member, IEEE, Alexandros T. Tzallas, Student Member, IEEE, Dimitrios I. Fotiadis*, Senior Member, IEEE, and Ioannis Georgiou.

[5] Seyed Alireza seyedin,"Direct Classification Of Human G-Banded Chromosome Images using Support Vector Machines" 2007 9th International Symposium on signal processing and its Application 02/2007.

[6] A Multichannel Watershed-Based Segmentation Method for Multispectral Chromosome Classification Petros S. Karvelis, Student Member, IEEE, Alexandros T. Tzallas, Student Member, IEEE, Dimitrios I. Fotiadis*, Senior Member, IEEE, and Ioannis Georgiou.

[7] P. Mousavi, R. K.Ward, and P. M. Lansdorp, "Feature analysis and classification of chromosome 16 homologs using fluorescence microscopy image," IEEE Can. J. Elect. Comput. Eng., vol. 23, no. 4, pp. 95-98, 1999.].

[8] http://www.metasystems.de/.

[9] Feature Analysis and Centromere Segmentation of Human Chromosome Images Using an Iterative Fuzzy Algorithm.

[10] Z. Kou, L. Ji, and X. Zhang, "Karyotyping of comparative genomic hybridization human metaphases by using support vector machines," Cytometry, vol. 47, pp. 17-23, 2002.

[11] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121-167, June 1998.

[12] Lijiya A, Sreejini K.S., and V.K. Govindan. "M-FISH Chromosome Images Classification by Watershed Based Segmentation Approach", In: Proceeding of Futuristic of Computer Science Engineering and Information Technology-ICCT2012, volume 2, pages 501 – 505, March 2012.

[13] P. Karvelis, A. Likas, and D.I. Fotiadis. "Semi unsupervised M-FISH chromosome image classification", In Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on, pages 1 - 4, November 2010.

[14] Hongbao Cao and Yu-Ping Wang. "Segmentation of M-FISH Images for improved classification of chromosomes with an adaptive fuzzy c-means clustering algorithm", In Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on, pages 1442 – 1445, 30 2011- April 2 2011.