ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Automated Pattern Recognition For Multispectral Chromosome Anlaysis Using Statistical Classifier And Fuzzy Inference Engine.

Mousami Munot[1]*, and Alwin Anuse[2].

[1]Dept. of E&TC, Pune Institute of Computer Technology, Pune, India.
[2]Dept. of E&TC, Maharashtra Institute of Technology, Pune, India.

ABSTRACT

Automated chromosome analysis is an essential task in cytogenetics and has therefore been an important pattern recognition problem. Traditional chromosome analysis is limited to grayscale images, but recently a multispectral imaging technique, Multicolor Fluorescence In Situ Hybridization (MFISH) has gained wide popularity in automated karyotyping systems. In MFISH image, each class of chromosomes binds with a unique combination of fluorophores. The existing MFISH classification methods incorporate a multivariate approach leading to space and time complexities. This paper proposes a novel and efficient univariate statistical approach based on Bayesian classifier for the classification of MFISH chromosomes. The proposed statistical classifier achieves an average overall classification accuracy of 98% (1.3 million operations per image) as compared to 92 % (240 million operations per image) with the previously reported methods. The proposed algorithm leads to substantial reduction in the computational complexity and also achieves significantly higher classification accuracy thereby outperforming the previously reported MFISH classification techniques. A fuzzy inference engine is also proposed to exploit the fuzziness in the derived feature vectors and provide a comparative classification accuracy of 97 % which is also higher than the earlier reported approaches.

Keywords: Karyotyping, metaphase, chromosomes, Bayesian classifier and fuzzy classifier

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Cytogenetics is the study of the genetic makeup of the cells. Chromosomes are the genetic information carriers of the body [1]. They play an important role in cytogeneticanalysis for the diagnosis of the genetic disorders. There are 46 chromosomes in the nuclei of normal eukaryote human cells with 22 pairs of autosomes and a pair of sex chromosomes (XX for a female, XY for a male) [2]. Karyotype image has all chromosomes in a cell graphically arranged (aligned, paired and arranged in their decreasing order of size) according to an international system for cytogenetic nomenclature (ISCN) [3].It is a useful tool to detect deviations from normal cell structure since the abnormal cells may have an excess or a deficit of chromosomes [4, 5].The manual karyotyping requires visual inspection, is tedious, lengthy, laborious, prone to human errors and an expensive procedure. It is a complex and time-consuming operation, as it requires meticulous attention to details and well-trained personnel [3]. Hence, many attempts have been made to automate the process of karyotyping. Within last few decades, Automated Karyotyping Systems [AKS] has offered countless clinical advantages such as interactive and graphical environment, fast examination of the samples, quality printing, self explanatory, better interpretation and storage of the information in a database for future analysis [6].

The progress of cytogenetics has become significantly faster since the discovery of chromosome banding and molecular Fluorescence In Situ Hybridization (FISH)technology[7].In mid 1990s, Multicolor-FISH (MFISH), a multispectral combinatorial labeling technique was developed. This colour karyotyping technique made the analysis of chromosome images easier, not only for visual inspection of the images by humans, but also for computer analysis of the images [1]. The central idea in MFISH is that each chromosome is labeled by a unique combination of five fluors and each chromosome class appears as a distinct color. At least five distinguishable fluors are needed for combinatorial labeling to uniquely identify all 24 chromosome types as the number of useful combinations of N fluors is2^N-1. Fig. 1(a)-(e) shows a typical MFISH dataset of six images where each image is the response of the chromosome to one particular fluor i.e. spectrum: Aqua, Far Red, Green, Red and Gold. Each class of chromosome absorbs a unique combination of dye and enables the chromosome classification. Fig. 1(f) depicts a DAPI (4',6- Diamidino - 2- pheylindole) MFISH image. DAPI is a counter staining dye that is absorbed by all the 24 classes of chromosomes because it attaches to DNA. Thus the chromosomes of all the classes are visible in DAPI image, whereas only selective classes are sensitive to each of the other 5 dyes (Aqua, Green, Gold, Red and Far-Red) [8]. Fig. 1(g) shows the karyotyped image. As indicated in fig 1(h), in an M-FISH image each pixel is represented as a five-dimensional vector, with each element in the vector representing the intensity of that fluorophore (dye) at that pixel. The resulting feature vector for the indicated example is : [Aqua, Green, Gold, Red, Far-red] = [68,205,120,98,10].

MFISH offers certain obvious advantages [9]:

- The task of chromosome classification is greatly simplified as only spectral information of the chromosome enables the karyotyping. Extraction of chromosome features like centromere location, banding pattern and other complicated features is no longer required.
- Even small translocations are easily identified.
- Touching and overlapping chromosomes can be easily disentangled.
- Classification can be performed independently of segmentation. The classification information can be used to attain more accurate segmentation which will in turn yield more accurate classification.

ISSN: 0975-8585

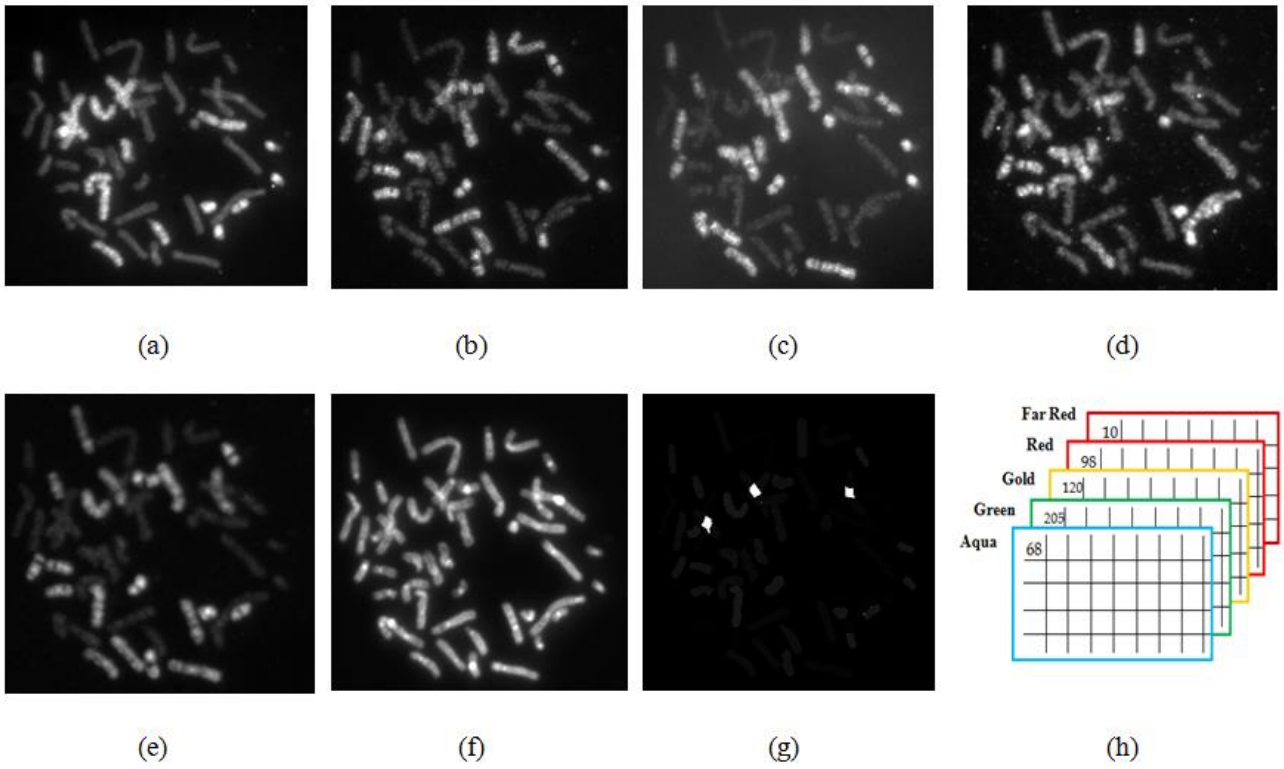


Fig 1: MFISH Images V150141 (Vysis Probe) (a) Spectrum Aqua (b) Spectrum Green (c) Spectrum Gold(d) Spectrum Red (e) Spectrum Far Red (f) DAPI (g) Karyotyped Image (h) Feature Vector [Aqua, Green, Gold, Red, Far-red] = [68,205,120,98,10]

MFISH imaging is thus proved to be extremely useful in cytogenetics, attracting many researchers to contribute in development of AKS. First paper on MFISH technique was published in 1996 by *Speicher et al.*and further in 1996, he proposed semi-automated image analysis consisting of Segmentation, thresholding and classification stages. MFISH Image analysis was fully automated by *Elis et al.* by modeling the task as a 5 feature 24 class pattern recognition problem[8]. *Schwartzkopf et al.*[10] developed a segmentation algorithm for MFISH images that minimizes the entropy of classified pixels. The method successfully decomposed the clusters of touching and overlapping chromosomes but has poor computationally complexity [9]. *Schwartzkopf et al.* [1] further exploited maximum likelihood technique to combine segmentation and classification into robust chromosome identification system. *Samp at et al.*[11] explored pixel by pixel classification achieving 95 % accuracy and further proposed supervised parametric and non -parametric classification of chromosome images [8]. *Wang et al.* [12] proposed a novel approach based on Fuzzy C-means clustering to achieve improved classification accuracy. *Karvelis et al*. [13, 14]reported watershed based method for segmenting the chromosomes with overall accuracy of 82.4%. *Choi et al.*[15, 16] developed joint segmentation of MFISH images and demonstrated the significance of feature normalization, leading to overall pixel classification accuracy improvement by 20%[17]. They also proposed a maximum likelihood decomposition of overlapping and touching MFISH chromosomes using geometry, size, and color information [18]. *Choi et al*. [19]explains the algorithm for removal of non-flat background to improve the classification accuracy by a factor of 10.

MFISH technology has seen major advancements in last few decades leading to significant contributions in the development of highly efficient AKS. Despite the fact that MFISH technology is a boon for the AKS, it has inherent limitations as reported by *Lee et al.* [20]. In certain situations it may lead to misclassifications and produce erroneous interpretations. The huge cost involved in the hybridization process, restricts its routine usage in genetics laboratories, making the wide variety of MFISH databases publically unavailable for analysis, experimentation and further research. Another very important limitation which seems to have received comparatively less attention in the literature is the computational complexity of the overall segmentation and classification process.

The problem of MFISH image classification is formulated as a five feature 24 class pattern recognition problem. Processing a multispectral set of five images increases the overall computational complexity by five times as compared to gray scale imaging. The MFISH image analysis and classification process therefore requires wide assortment of calculations. It is equivalent to handling a multidimensional vector (as depicted in fig. 1(h); working in five dimensions) corresponding to a single pixel as compared to just one intensity value per pixel in gray image. MFISH image analysis therefore leads to substantial space and time complexities and crucial memory handling issues. All the approaches reported in the literature use multivariate analysis techniques for classification which increases the computational complexity of the classifiers. As reported by *Schwartzkopf et al*. [1], in multivariate analysis, a 5-element mean vector is to be multiplied with a 5X5 covariance matrix thus a total of 30 multiplications and 24 additions are required for each pixel and class. Since there are 24 classes and the image size used is 517X645, there are nearly 240 million multiplications/additions for each image. The code takes around 2.5 minutes on 167-MHz Sun workstation to accomplish both segmentation and classification.

Recently, Hua *et al*. [21] introduced embedded M-FISH image coding (EMIC) to accommodate the explosive growth of digital image data and to reduce the storage, transfer and computational cost. Goienetxea *et al.*[22] presented an image analysis pipeline of banded human chromosomes for automated karyotyping. With such topical expansion in the field of AKS, it is extremely important to address the complexity issues in MFISH image analysis. This problem has received comparatively less attention by the researchers and is a major hindrance in the development of AKS.

This paper in itially uses the methods reported in the literature for the preprocessing of the MFISH image and further addresses the issue of computational complexity by formulating a novel classification technique which uses univariate analysis to reduce the number of computations and the processing time. An efficient univariate statistical approach based on Bayesian classifier for the classification of MFISH chromosomes is explored. The proposed algorithm leads to substantial reduction in the computational complexity and also achieves significantly higher classification accuracy thereby outperforming the previously reported MFISH classification techniques. A fuzzy inference engine is also proposed to exploit the fuzziness in the derived feature vectors and provide comparative classification accuracy.

PROPOSED APPROACH

The basic steps in MFISH chromosome image analysis include Image Enhancement, Feature Normalization, Feature Extraction, Segmentation and Classification. The following section briefs the initial processing steps that are already reported in the literature for the sake of comprehensiveness and further details the novel univariate based classification approach.

Image Enhancement:

Microscope images of biological specimens of often contain non-flat background surfaces. The removal of the non-flat background surface is called background correction and is commonly performed as a preprocessing step [17]. The background intensity 'b' is mostly affected by the auto-fluorescence of the slide, the dc offset of the CCD, unattached free fluorescent molecules, the intensity of the defocused objects from out of depth of field. All these factors contribute to the non-flat intensity elevation of the background[23].The algorithm described by *Choi et al.*[23]is used for background correction. Chromosome pixels need to be separated from the background pixels for further processing. The separation between Background $(I_b(k))$ and Chromosomes $(I_c(k))$ is obtained by edge detection using Laplacian of Gaussian (LoG) followed by basic morphological operations like dilation, erosion.

Feature Normalization:

Background correction significantly reduces the variations in the feature vector but there still exists intensity variations within a chromosome, among the chromosomes in a channel and between the various channels. This sometimes leads to some classification error [17]. Given a channel, chromosomes that are supposed to be bright in that channel must ideally have similar intensity levels among them, but often chromosome intensities considerably differ as some are much brighter than others. Often a chromosome with a certain intensity level on one channel appears on another channel with a significantly lower or higher

intensity. This inconsistency causes classification errors since the pattern of a feature vector y becomes inconsistent. Feature normalization is therefore required so as to minimize the difference of the sample distributions for all the images.

Feature normalization approach as proposed by Choi *et al.*[17] is used for normalizing the features before the classification of the MFISH images. Parameter vector ϑ contains $(\mu_1, \mu_2, \sigma_1, \sigma_2, P(w_1), P(w_2))$ where $P(w_1)$ and $P(w_2)$ are mixing parameters (w_1=intensity due to non-fluorophore;w_2= intensity due to fluorophore). Given a bimodal marginal density function and its parameters, the normalization process should cause $y \in w_1$ and $y \in w_2$ to fall within certain ranges and the decision boundary between w_1 and w_2 to lie at a certain point. The unknown parameters can be extracted using Expectation–Maximization (EM) algorithm in which the initial estimate is provided using K-means clustering algorithm. Then the parameter vectors for each class $\vartheta_i = (\mu_i, \sigma_i, P(w_i))$ can be estimated iteratively using the following equations (expressed in general terms)

$$P(w_i) = \frac{1}{N}\sum_{j=1}^{N} P(w_i|z_j,\theta) \tag{2.1}$$

$$\mu_i = \frac{\sum_{j=1}^{N} p(w_i|z_j,\theta) z_j}{\sum_{j=1}^{N} p(w_i|z_j,\theta)} \tag{2.2}$$

$$\sigma_i = \frac{\sum_{j=1}^{N} p(w_i|z_j,\theta)(z_j-\mu_i)(z_j-\mu_i)^T}{\sum_{j=1}^{N} p(w_i|z_j,\theta)} \tag{2.3}$$

Where,
N= number of unlabeled samples drawn independently from the mixture density of c classes
i= 1,…, c; c is number of classes

$$P(w_i|z_j,\theta) = \frac{p(z_j|w_i,\theta_i)P(w_i)}{\sum_{l=1}^{c} p(z_j|w_l,\theta_l)P(w_l)} \tag{2.4}$$

An initial estimate for Eq. (2.4) is evaluated using K-means clustering and the Eq. (2.1) - Eq. (2.3) are used to update the estimates. This iterative method is called Expectation- Maximization [17].μ_1 and μ_2 are found via the K-means clustering, the value σ_i^2 are estimated from the samples classified to w_1 and w_2. These means and variances are used as initial estimates for Eqn. (2.1) - Eqn. (2.3). Once the parameters are estimated by the EM method, the decision boundary between w_1 and w_2 is found by

$$T = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{2.5}$$

Where,
$A=\sigma_2^2-\sigma_1^2; B=2\sigma_1^2\mu_2-2\sigma_2^2\mu_1; C=\sigma_2\mu_1^2-\sigma_1^2\mu_2^2-2\sigma_2^2 ln(\sigma_2 P(w_1)/\sigma_1 P(w_2))$

Given the Parameter vectors and the decision boundary, the sample distribution is normalized by piece-wise linear transformations. The input intensity r is mapped to the output intensity $s$by:

$$f(r) = \begin{cases} \frac{64}{\mu_1 - min(r)}(r - min(r)) + min(r) & \leq r < \mu_1 \\ \frac{64}{T-\mu_1}(r-\mu_1) + 64 & \mu_1 \leq r < T \\ \frac{64}{\mu_2 - T}(r-T) + 128 & T \leq r < \mu_2 \\ \frac{63}{max(r) - \mu_2}(r-\mu_2) + 192 & \mu_2 \leq r < max(r) \end{cases} \tag{2.6}$$

Where, min (r): minimum intensity level; max(r):maximum intensity level

ISSN: 0975-8585

Feature Extraction and Training:

Distribution of intensity values of chromosome pixels in each channel is assumed to be mixture of two Gaussians- one corresponding to fluorophore class and other corresponding to non-fluorophore class. Parameters like mean, variance and priori probabilities(μ, σ, $P(w_i)$) are extracted from each channel of M-FISH image set. Thus for each M-FISH image set, mean, variance and priors are vectors of size 5 X 2. Using seven datasets, each containing six images, population estimate of mean and standard deviation is obtained for training of classifier.

Joint segmentation and Classification:

Literature reports a multivariate approach for the analysis of M-FISH images where each channel is modeled as a mixture of 24 Gaussians each corresponding to one class. Fig 2(a) depictsa channel comprising of 24 Gaussians having a different mean and variance. The computational complexity of multivariate analysis can be easily seen as mean vector would be of dimension 24 X 5 where 24 corresponding to classes and 5 corresponds to the spectral channels.

This paper proposes a novel and efficient univariate approach for the classification of chromosomes. In the proposed univariate approach, as illustrated in Fig. 2(b) each channel is modeled as a mixture of 2 Gaussians each corresponding to one class, fluorophore and non-fluorophore and having different mean and variance. The ease of computation for univariate analysis can be justified as mean vector would be of dimension 5 X 2 where 5 corresponds to 5 spectral channels and 2 corresponds to fluorophore and non-fluorophore classes.

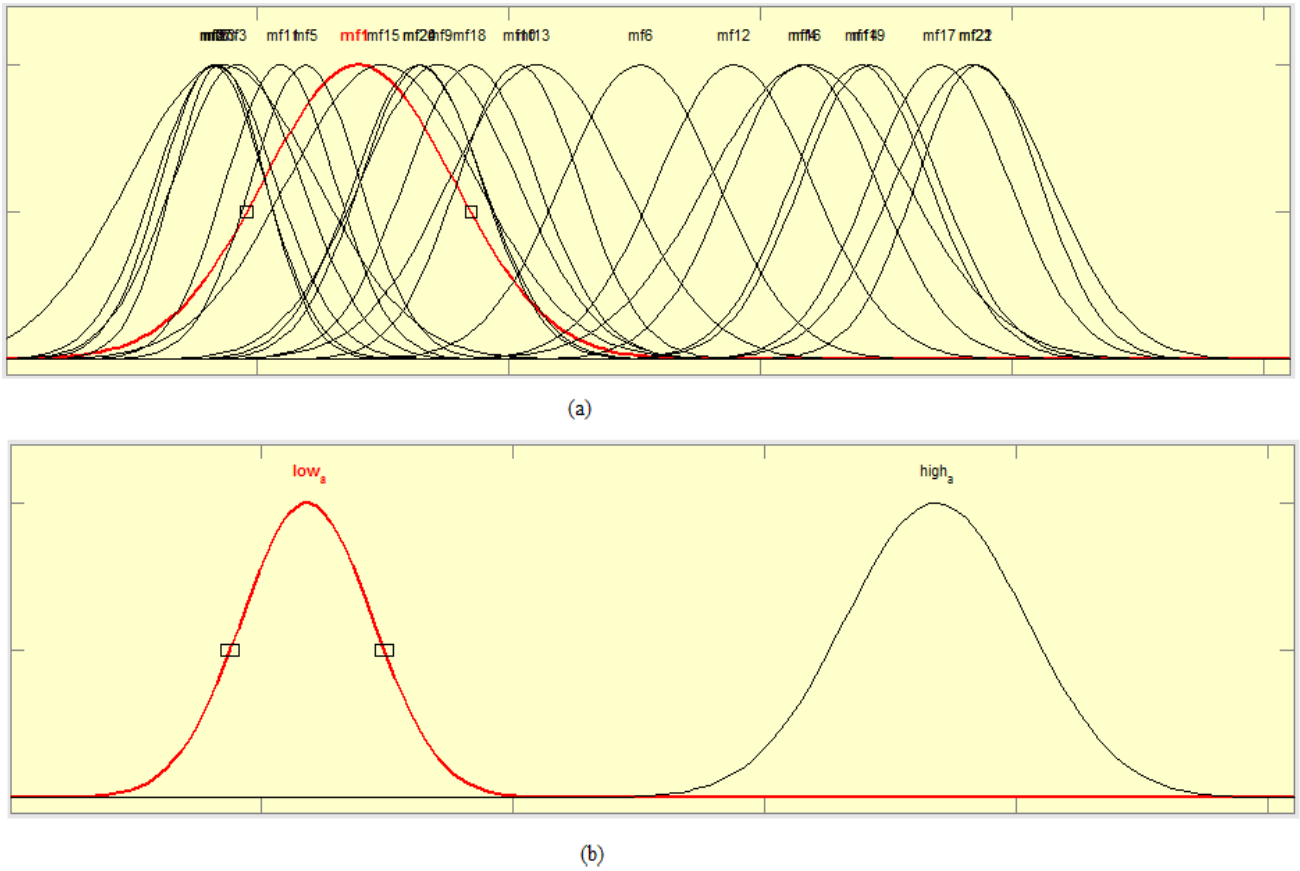


Fig 2: (a) Single channel for Multivariate Analysis (b) Single channel for Univariate Analysis

Classifier is implemented using supervised parametric technique with two different approaches.

(i) Statistical Classifier:

Bayes Theorem

$$P(C_i|x) = \frac{P(x|C_i) * P(C_i)}{P(x)} \quad (2.7)$$

Where,

$P(x) = \sum P(x|C_i)$

i : class

$P(C_i)$: Prior class probabilities

$P(x|C_i)$: Class conditional probability distribution function

$P(C_i|x)$: Posteriori probability

Here, $P(x|C_i)$ and $P(C_i)$ are estimated from training data by fitting a Gaussian Mixture Model to each channel to determine class conditional probability distributions. The two Gaussians of mixture model corresponds to the two classes, fluorophore and non-fluorophore for each of the 5 channels. Thus i = 2 in our case. From Prior class probabilities and priori class conditional probability distribution, two posterior probabilities for each chromosome pixel is calculated. By comparing the two posterior probabilities of each pixel, it is classified to one of the two classes, low and high, in each image individually. Thus, each image is converted into binary image. The feature vector for each pixel consisting of any value between 0-255 is now converted into a feature vector which has only two discrete values 0 and 255.

Ex. Pattern [220 40 18 60 178] is converted to [255 0 0 0 255].

Such pattern for every chromosome pixel is compared, pixel-by-pixel, with standard pattern for each class of chromosome and the pixel is classified into one of the 24 classes.

(ii) Fuzzy Inference Engine:

A Sugeno model based fuzzy inference system (FIS) is used as a classifier that uses fuzzy set theory to map inputs (Distribution of five channels) to outputs (24 chromosome classes). In Sugeno FIS, output membership functions are linear or constant. It uses weighted average to compute the crisp output. Sugeno method is computationally efficient and works well with optimization and adaptive techniques. It also has better processing time compared to Mamdani model and is therefore used in the proposed FIS approach. There are five inputs to the FIS (one for each of the five channels) consisting of two Gaussian membership functions. First corresponds to non-fluorophore intensity and other corresponds to fluorophore intensity. Output of FIS is only one that is class of chromosome, consisting of 24 membership functions corresponding to 24 chromosome classes as shown in Fig 3(a).Rules for classification are formulated based with reference of standard Vysis chromosome labeling chart**.** Based on description of the input (Aqua, Green, Gold, Red, Far-red) and output variables (24 chromosome classes), the rule statements are constructed in the rule editor of Matlab.

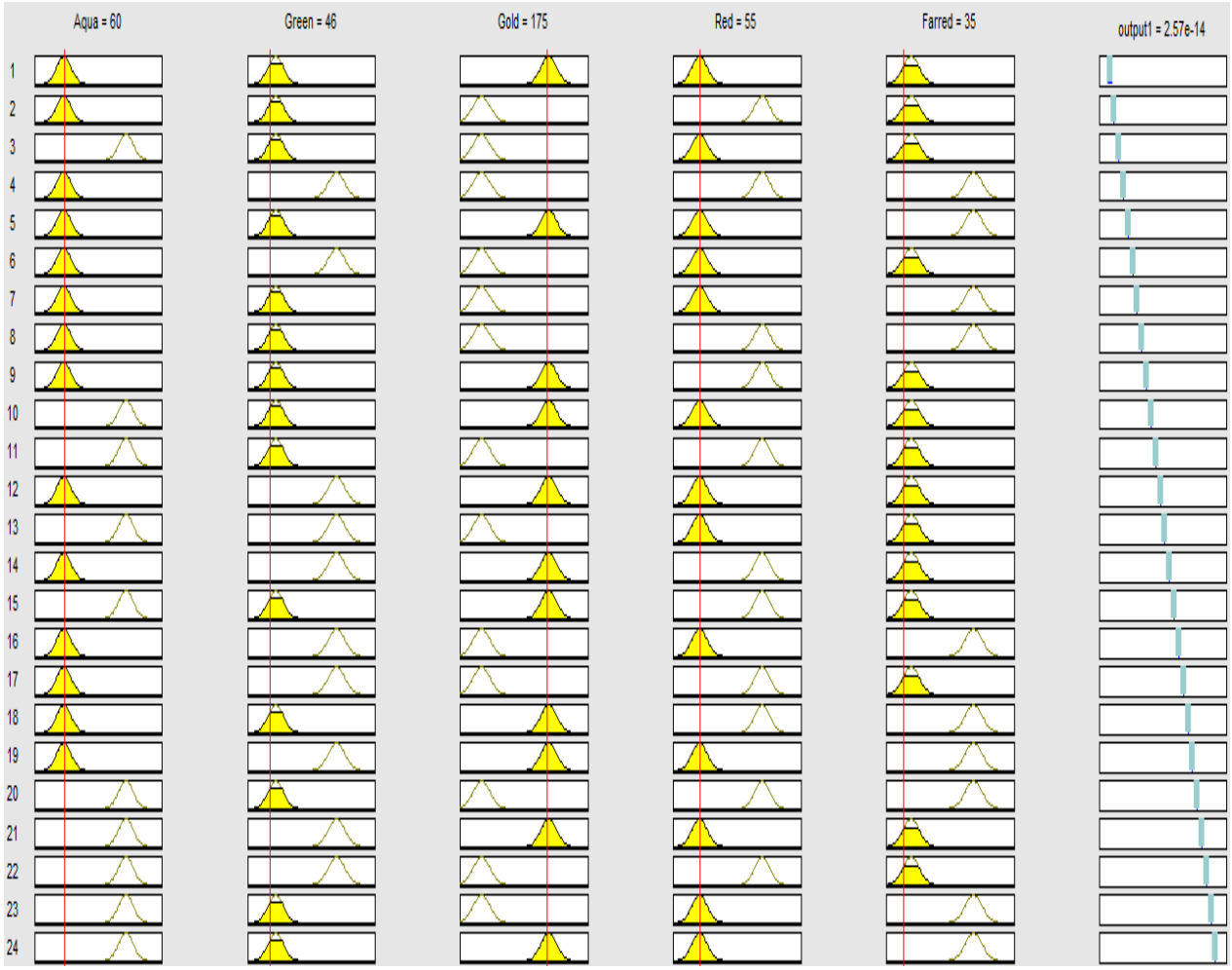
Rules for image classification procedure are as follows:

- For class 1 chromosomes: If (Aqua is low_a) and (Green is low_g) and (Gold is high_y) and (Red is low_r) and (Farred is low_fr) then (output1 is Class1) (1)
- For class 2 chromosomes: If (Aqua is low_a) and (Green is low_g) and (Gold is low_y) and (Red is high_r) and (Farred is low_fr) then (output1 is Class2) (1)

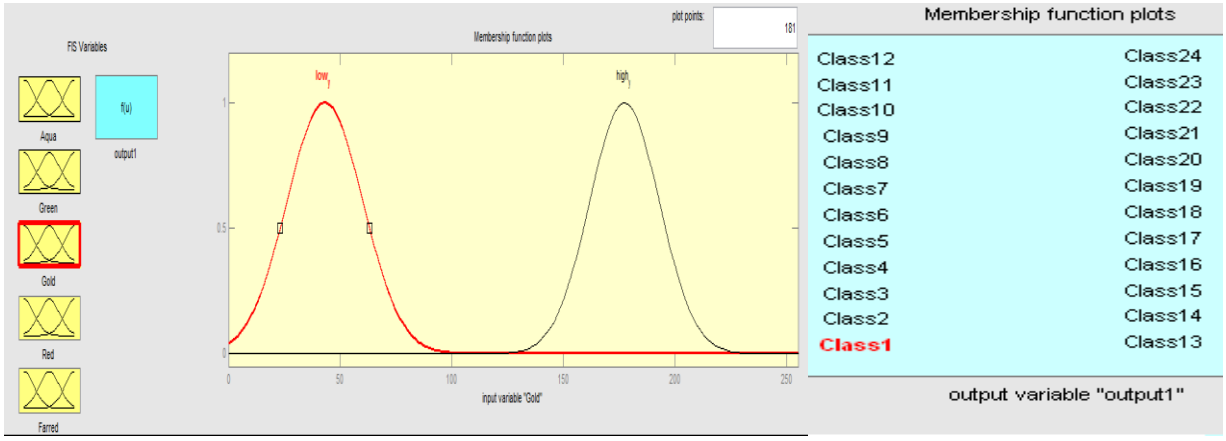
The structure of the rules is same for other classes of chromosomes. Here, as indicated in Fig.3, there are five inputs namely Aqua, Green, Gold, Red and Far-red. The two membership functions of each input are labeled as low and high. So low_a means lower membership function of aqua (non-fluorophore) and high_a means higher membership function of aqua (fluorophore).From standard chromosome labeling chart, for class 1, intensity pattern should be [low, low, high, low, low]. So input membership functions are compared with trained membership functions of classifier and the class of chromosome is estimated. As shown in fig 3(b) the rule base consists of 24 rules. Each rule considers 5 channels which are to be low or high for a particular class as per the standard labeling Vysis chromosome labeling chart. Fig 3(c) indicates 24 output membership functions ranging between 0 and 1. So, each class has a unique number associated with it which

ISSN: 0975-8585

is used in classification. Each pixel that is classified is assigned a particular value between 0 and 1. The class closest to the assigned value is determined as the class of the chromosome pixel.



(a)



(b) (c)

Fig 3: (a) FIS Rule base (b)FIS for chromsome classification (c) Output variable of FIS

RESULTS

The software used for implementation is MATLAB R2011b. The configuration of the machine used for implementation is 2.4GHz Intel i3 processor with 3GB RAM. The standard M-FISH database used for

experimentation is obtained from Advanced Digital Imaging Research (ADIR) Laboratory. The image set contains 200 M-FISH images. The efficiency of the proposed approach is examined on M-FISH image sets (Vysis probe) including normal and abnormal, male and female chromosomes and touching and overlapping chromosomes.

Two separate sets of M-FISH images are used for formulating the result.

Set I- Consists of only normal male chromosome images. Non-touching, touching and overlapping chromosome images are included. This dataset is used for direct comparison of overall classification accuracy of the proposed univariate approach with the results reported in the literature.

Set II- Consists of normal and abnormal, non-touching and touching male and female chromosome. This dataset is used to calculate the accuracy of the classifiers and also to examine the effects of each process such as background correction and feature normalization.

The results of pre-processing feature extraction and classification on the MFISH images are detailed in this section.

Results of Background Correction:

The method used for background correction is a retrospective method in which noise is modeled as a 2D cubic surface as introduced by *Choi et al.* [23]. Fig 4(a) illustrates the non-uniform nature of the background intensity that may give rise to misclassifications. This non flat background surface is modeled as a 2D cubic surface and is removed from the original image to obtain a uniform intensity background. As seen in Fig 4(b), after background correction uniform background illumination is obtained. The original and the processed pixel values are high lightened in Fig 4(c). Intensity value of 81, in the original image is reduced to 12 after background correction. Thus the noise is reduced in the background to a very small and negligible level. Fig 4(d) represents the profile density of the original image (blue color) and background corrected (red color) image. As clearly seen the DC Offset is removed from the original image after background correction.

ISSN: 0975-8585

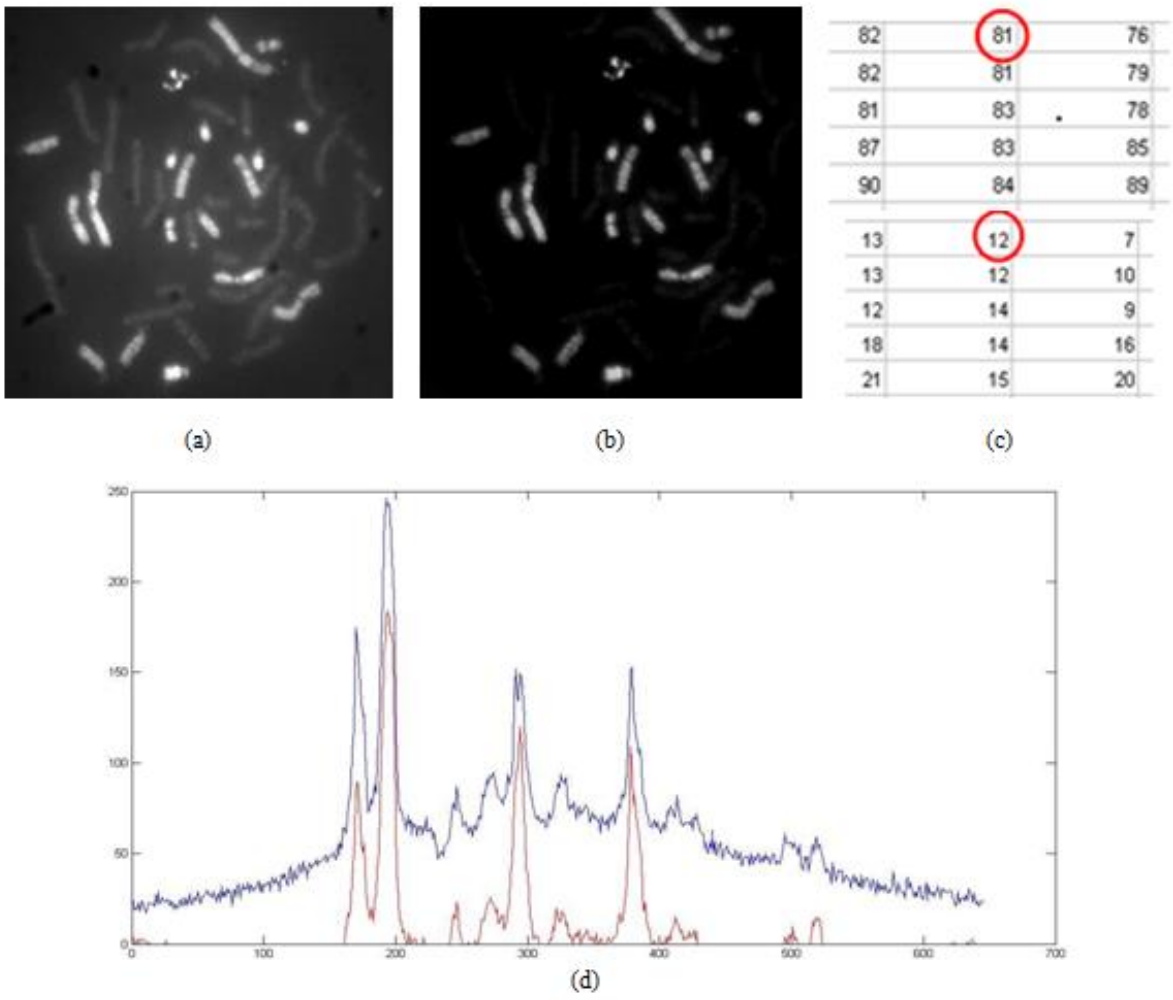


Fig 4: (a) Original Image V1306XYA (b) Background corrected image (c) Pixel values before (top) and after (bottom) background correction (d) Profile density of original image (blue) and background corrected image (red)

Result of Foreground- Background Separation:

Basic edge detection and morphological operations are performed for foreground-background separation. Steps involved in obtaining the basic-segmented image are Edge detection, Median Filtering, Convolution using a 3x3 mask to fill the detected edges and Imfill holes and bridge are used to enhance the foreground-background separated binary image.Fig.5(a) - 5(d) depicts the results of fore-ground background separation.

Result of Feature Normalization:

Fig. 5(e) represents the original data distribution that is normalized by a piecewise linear mapping function indicated in Fig.5(g). Fig. 5(f) depicts the resultant distribution after the normalization of the features. It can be seen that the two classes F (Fluorophore class) and NF (Non- Fluorophore class) are more clearly distinguishable in Fig. 5(f) compared to Fig. 5(e). In original data distribution, intensity values were ranging from 0 to 150. After Normalization, intensity values range from 0 to 255.

Normalization of input data distribution is implemented by estimating a piecewise linear mapping function for that particular distribution as described by Choi *et al*. [17]. Parameters such as mean, variance and prior probabilities of the original data distribution are extracted by Expectation-Maximization Algorithm

ISSN: 0975-8585

and the mapping function is designed with the help of these parameters. Thus, the mapping function is unique for each image.

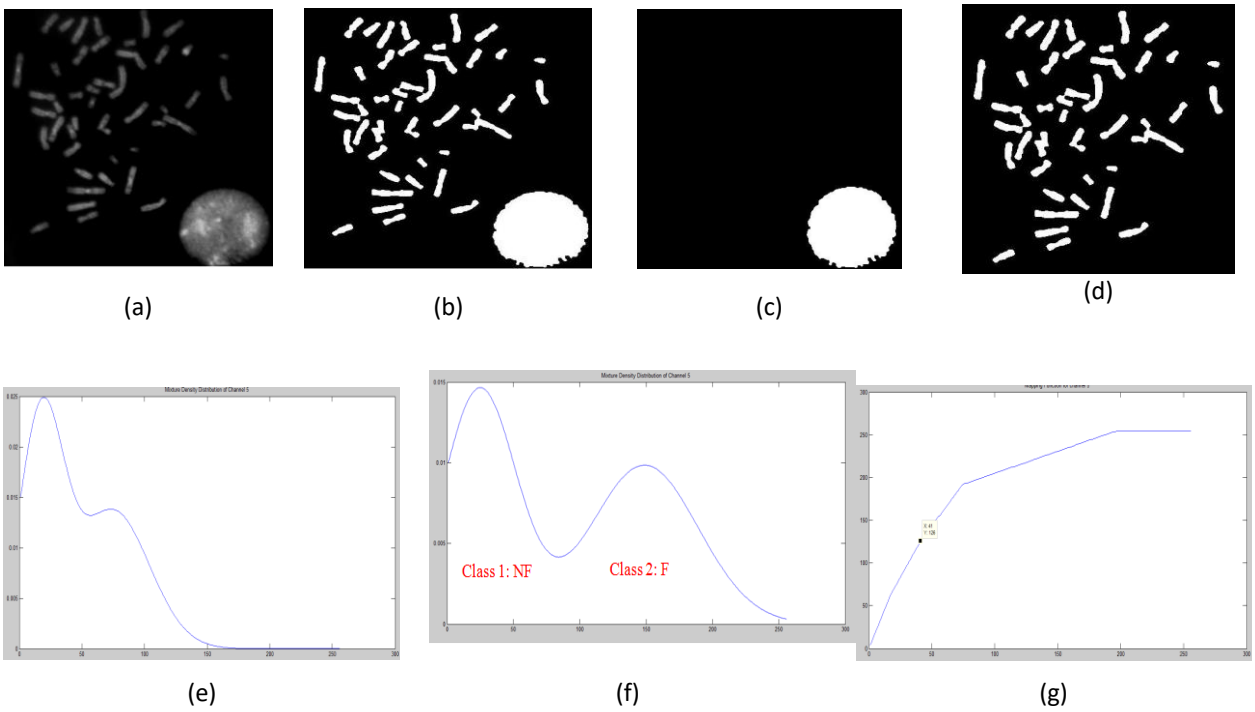


Fig 5: Results of foreground - background separation (V1308XYD) and feature normalization (V290463) ***Foreground -Background Separation:*****(a) Original Image (b) Result after basic morphological operations (c) Erroneous Illumination (d) Final Background-Foreground separated image.** ***Feature Normalization:*** **(e) Original data distribution (f) Data distribution after normalization (g) Piecewise linear mapping function**

Initial parameters of the M-FISH data distribution, mean, variance and prior probabilities are estimated from k-means clustering. As indicated in Table 1, five rows correspond to five spectral channels Aqua, Green, Gold, Red and Far-red. Each of the five images is divided into two classes: non-fluorophore (NF) and fluorophore (F) (Fig.5(f)). Parameters estimated from k-means clustering are updated by EM algorithm. These parameters are further used to obtain the threshold and normalized data. Data distribution parameters may vary drastically for each spectrum hence normalization of the intensities is necessary for classification.

The initial mean value without feature normalization, calculated using K- means clustering for aqua channel is 23.71 which is finally evaluated as 21.7 after application of EM algorithm. After normalization all the five images have intensity values between 0 and 255. The mean value for Aqua, after feature normalization is 59.7 using K means which finally is 60, as calculated with EM algorithm. Table 1depicts a snapshot of the parameters calculated using k means clustering and EM algorithm with and without feature normalization for a test image in the database.

Table 1: Snapshots of the parameters with and without feature normalization

1(a): Snapshots of parameters (without feature normalization)estimated using K-means Clustering

Parameter	Mean		Variance		Prior Probability	
Class / Channel	F	NF	F	NF	F	NF
Aqua	23.7165	133.2394	196.0290	975.2760	0.7635	0.2365
Green	19.8466	110.6518	222.9528	753.4169	0.7680	0.2320
	9.2033	100.4831	175.2986	829.6394	0.6685	0.3315
	10.2449	77.7681	130.6071	482.9319	0.6630	0.3370
	20.5076	82.9098	317.9336	659.1844	0.7611	0.2389

ISSN: 0975-8585

Gold			
Red			
Far Red			

1(b): Snapshots of parameters (without feature normalization) estimated using EM algorithm

Parameter	Mean		Variance		Prior Probability	
Class / Channel	F	NF	F	NF	F	NF
Aqua	21.7017	124.4601	111.7784	1.3822e+03	0.7283	0.2717
Green	18.4745	104.2824	178.7439	999.0162	0.7385	0.2615
Gold	5.8542	91.6061	58.3626	1.1500e+03	0.6081	0.3919
Red	7.6793	71.4883	63.7210	642.7769	0.6031	0.3969
Far Red	21.2445	77.6120	372.1234	918.6887	0.7486	0.2514

1(c): Snapshots of parameters (with feature normalization) estimated using K-means Clustering

Parameter	Mean		Variance		Prior Probability	
Class / Channel	F	NF	F	NF	F	NF
Aqua	59.7447	188.1450	650.1822	931.6759	0.7206	0.2794
Green	52.0378	176.1650	994.3875	1.0015e+03	0.6928	0.3072
Gold	34.0287	178.3285	1.1244e+03	930.6799	0.5496	0.4504
Red	41.1267	176.0082	1.0973e+03	746.6606	0.5477	0.4523
Far Red	33.9806	155.2638	1.0717e+03	1.3448e+03	0.5568	0.4432

1(d): Snapshots of parameters (with feature normalization) estimated using EM algorithm

Parameter	Mean		Variance		Prior Probability	
Class / Channel	F	NF	F	NF	F	NF
Aqua	60.0571	188.1838	677.5403	961.5118	0.7224	0.2776
Green	54.0837	178.2222	1.1377e+03	1.0020e+03	0.7093	0.2907
Gold	35.7946	179.5780	1.2606e+03	891.2528	0.5602	0.4398
Red	43.3311	177.5998	1.2546e+03	691.5165	0.5621	0.4379
Far Red	31.2897	149.0092	1.0509e+03	1.7040e+03	0.5205	0.4795

Results of Classification:

Classifier is implemented with the two approaches, statistical classifier and fuzzy classifier.

ISSN: 0975-8585

Accuracy of the classifier is calculated by comparing the chromosomes image classified with the proposed approach and the manually karyotype image (ground truth) provided in the M-FISH database. The karyotyping results of the developed classifiers and the ground truth are pseudo colored to make them visually appealing. Each class of chromosome in the results and the ground truth is assigned a different color for easy identification and comparison of the results.

(a) (b) (c)

Fig 6: Result of classification (Pseudo Colored Images) (a) Reference Karyotype image as ground truth. (b)Karyotyping using Statistical Classifier(c) Karyotyping using Fuzzy Inference System

Fig.6 depicts the results of classification using both the classifiers. Fig. 6(a) illustrates the ground truth. Results obtained using the proposed statistical classifier and FIS are indicated in Fig. 6(b) and 6(c) respectively. The non uniformities and slight discontinuities in the color of the chromosomes indicate the misclassification. The pixels with distinct colors within a single chromosome obviously represent belongingness to some other chromosome (class) and are therefore said to be misclassified. A comparison between both the developed classifiers is performed. The overall classification accuracy and chromosome classification accuracy is calculated using eq 2.8 and eq 2.9.

$$Chromosome\ Classification\ Accuracy = \frac{Chromosome\ Pixels\ correctly\ classified}{Total\ number\ of\ Chromosome\ pixels} \quad (2.8)$$

$$Overall\ Classification\ Accuracy = \frac{Pixels\ correctly\ classified}{Total\ number\ of\ pixels} \quad (2.9)$$

The correctly classified pixel is determined by using the Karyotyped images (ground truth) in the database. Table 2 provides comparison between two approaches of univariate classifiers. This clearly shows that Statistical classifier has better chromosome classification accuracy than Fuzzy classifier.

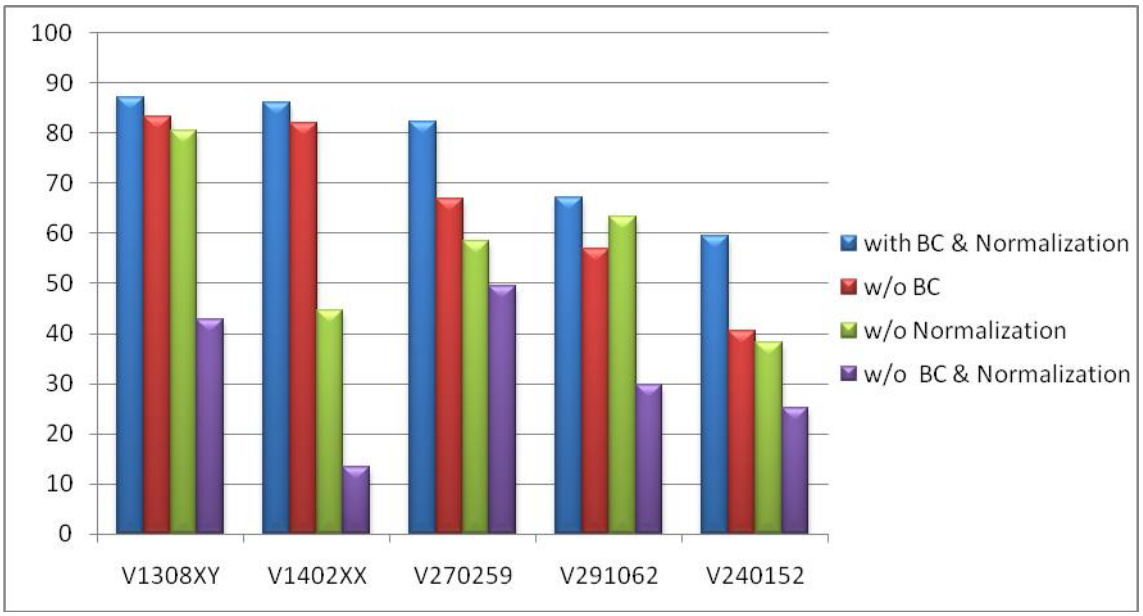
Table 2: Comparison of the classification accuracy of classifiers

Accuracy	Statistical	Fuzzy-Logic
Average Chromosome Classification accuracy	83.44	71.52
Average Overall Classification accuracy	97.32	96.47

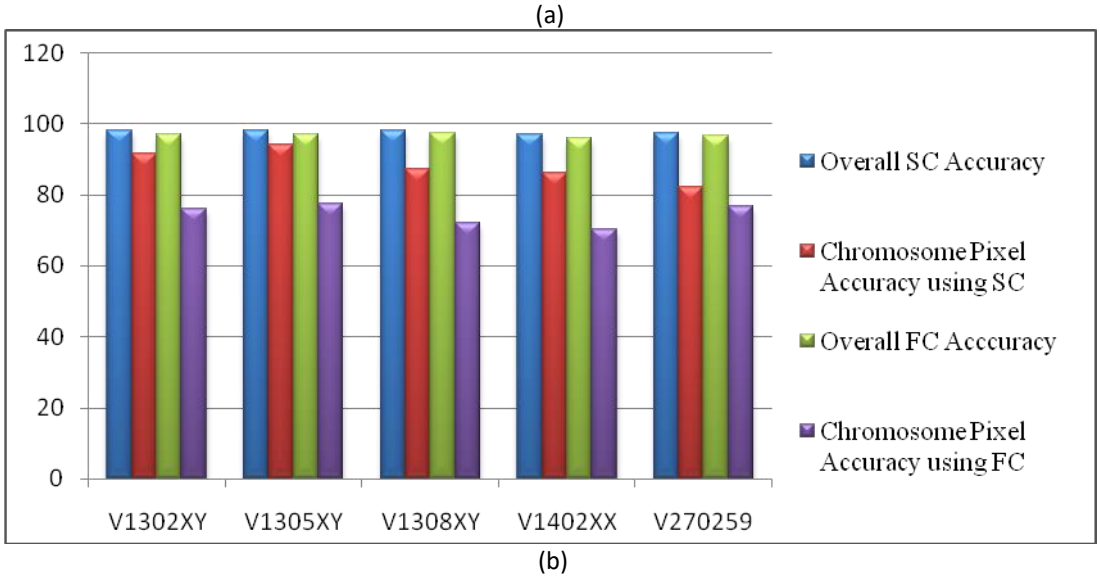
The chromosome classification accuracy of Statistical classifier using Bayes rule is higher than the chromosome classification accuracy of Fuzzy-Logic Classifier by 10.16% on average. The overall classification accuracy of Statistical Classifier using Bayes Rule is **97.32%** whereas overall classification accuracy of Fuzzy-Logic Classifier is **96.47%** on average. The results of both the classifiers are however comparable and acceptable to support and assist (not replace) doctors interpretations.

ISSN: 0975-8585

The effect of preprocessing on the classification is also examined as the part of the experimentation. M-FISH image (Set II) is classified without any preprocessing and the results are then compared to the accuracy with feature normalization and back ground correction. The back correction algorithm which is used for image enhancement removes the non-flat background surface from the images containing elevated background. This processing improves the chromosome classification accuracy by **48.09%** on average as tested on the dataset. The feature normalization algorithm normalizes the data distribution for each feature from grayscale range of 0 to 255. This processing improves the chromosome classification accuracy by **47.37%** on average. The chromosome classification accuracy with no preprocessing is 48.34% on average whereas classification accuracy by using background correction and feature normalization is 81.24%. Thus the efficiency of the classifier increases by **68.06%** on average with background correction and feature normalization. Fig. 7(a) illustrates the effect of background correction and feature normalization and comparison of the classification accuracies of both the classifiers is presented in Fig.7(b).



(a)



(b)

Fig 7: (a) Effect of Background correction (BC) and feature normalization of classification accuracy. (b) Comparison of classification accuracies with Statistical Classifier (SC) and Fuzzy classifier (FC)

DISCUSSIONS AND CONCLUSION

A novel univariate approach for automated classification of MFISH chromosome images is developed to address the issue of computational complexity in multivariate approaches. The designed

classifiers are tested on MFISH dataset - I and dataset - II (114 images) including normal and abnormal, male and female chromosomes. These image sets also include touching and overlapping chromosomes. The classification accuracy of the developed classifiers: *Statistical classifier* and the *fuzzy inference system* based on *univariate approach is* calculated and compared with the previously reported multivariate classifier. Table 2compares the classification accuracy of the proposed approach with the previously reported: maximum distance [MD] and maximum likelihood [ML] methods in literature. It must be emphasized that the results reported in [17] are evaluated using the MFISH dataset-I. The same data set –I is used for examining the efficiency of the proposed approach and to establish a direct benchmark for comparison. The proposed Statistical classifier increases the overall classification accuracy by approximately 5% and the fuzzy inference engine increases the accuracy by 4%.As summarized in table 3,the developed novel statistical and fuzzy classifier use univariate approach for pixel-by-pixel joint segmentation-classification. This approach drastically reduces computational and time complexity compared to multivariate methods reported in the literature. For univariate analysis using statistical classifier, 10 multiplications and 3 additions are required for each pixel. Thus giving rise to approximately 1.3 million multiplications/additions for each image set compared to 240 million multiplications/additions using multivariate approach as reported by Wade *et al*. [1]. Time required for multivariate pixel-by-pixel segmentation-classification was approximately 2.5 minutes which significantly reduces to 1.2 minutes by using univariate pixel-by-pixel segmentation-classification. The developed approach leads to a substantial reduction in the computational complexity (approximately by a factor of 185).

Table 3: Comparison of proposed classifiers with the previously reported methods

3(a): Computational complexity comparison for univariate and multivariate approach

Parameter	Multivariate approach [1]	**Proposed Univariate approach**
Computational Complexity (No. of multiplications/additions)	240 million	**1.3 million**
Time required (minutes)	2.5	**1.2**

3(b): Classification accuracy of the previously reported classifiers [23] and the proposed Statistical and Fuzzy classifiers [MD: Maximum Distance; ML: Maximum Likelihood]

Images	MD Classifier	ML Classifier	**Statistical Classifier**	**Fuzzy Classifier**
V1301XY	90.81	-	97.36	95.91
V1302XY	92.99	-	98.04	97.20
V1303XY	92.77	-	97.69	96.66
V1305XY	94.72	-	98.25	97.08
V1306XY	94.66	-	97.49	96.42
V1308XY	96.28	96.49	98.19	97.36

V1309XY	84.50	86.29	97.76	96.63
V1310XY	84.19	86.90	98.02	97.19
V1311XY	94.50	94.54	98.63	97.69
V1312XY	94.83	95.20	97.49	96.58
V1313XY	94.53	94.88	97.97	96.96
Average	**92.25**	**92.38**	**97.89**	**96.88**

The developed univariate approach for karyotyping the MFISH chromosomes is a successful attempt to address the critical issue of computational complexity in the multivariate analysis of MFISH images. This research identifies this issue and provides a novel solution using statistical classifier and fuzzy inference engine. The developed approach showcase its strength in terms of substantial reduction in the number of computations and the time required and further achieves higher achieves classification accuracy.

ACKNOWLEDGEMENT

This work was supported by Department of Science and Technology, Government of India, under research grant: SR/TP/ETA-15/2009.First author is also thankful to Gulsheen Ahuja, Bhakti Baheti and Ashwini Parode, PICT, Pune for their kind help, untiring determination and valuable efforts in this research.

REFERENCES

[1] Schwartzkopf, W., Bovik, A., Evans, B.: 'Maximum Likelihood Techniques for Joint Segmentation - Classification of Multispectral Chromosome Images',IEEE Trans. Medical Imaging, 2005, 24(12), pp. 1593-1610.

[2] Legrand, B., Chang, C., Ong, S., Neo, S., Palanisamy, N.: 'Chromosome classification using dynamic time warping', Pattern Recognition Letters, 2008, 29, pp. 215-222.

[3] Enea, P., Enriico, G., Alfredo, R.: ' A modular framework for automatic classification of chromosomes in Q-band images', Computer Method and Programs in Biomedicine, 2012, 105(2), pp. 120-130.

[4] Karvelis, P., Fotiadis, D., Georgiou, I., Syrrou, M.: 'A Watershed-Based Segmentation Method for Multispectral Chromosome ImagesClassification', Proc. of the 28th EMBS Ann. Int. Conf. USA, 2006, pp. 3009-3012.

[5] Wang, X., Zheng, B., Wood, M., Li, S., Chen, W., Liu, H.: 'Development and evaluation of automated systems for detection and classification of banded chromosomes: current status and future perspectives', Journal of Physics D: Applied Physics, 2005, 38, pp. 2536–2542.

[6] Munot, M., Joshi, M., Mitra, P.: 'Genetic Algorithm Incorporates with Rough Set Theory: Application to Automated Karyotyping', Indian Int. conf. on artificial intelligence, IICAI 2011, Tumkur.

[7] Wang, Y.,Castleman, K.: 'Normalization of Multicolor Fluorescence In Situ Hybridization (M-FISH) Images for Improving Color Karyotyping', Journal of the Int. Society for Advanced of Cytometry Part A, 2005, 64A (2), pp. 101-109.

[8] Sampat, M., Bovik, A., Aggarwal, K., Castleman, K.: 'Supervised Parametric and Non-parametric Classification of Chromosome Images', The Journal of the Pattern Recognition Society on Pattern Recognition, 2005, 38(8), pp. 1209-1223.

[9] Schwartzkopf, W., Evans, B., Bovik, A.:'Entropy estimation for segmentation of multi-spectral chromosome images', Proc. of 5th IEEE southwest Symposium on Image Analysis and interpretation, 2002, pp. 234-237.

[10] Schwartzkopf, W., Evans, B., Bovik, A.: 'Minimum Entropy segmentation applied to multi - spectral chromosome images', Proc. of International conference on image processing 2001, 2, pp. 865-868

[11] Sampat, M., Castleman, K., Bovik, A.: 'Pixel-by-Pixel Classification of M-FISH Images', Proc. IEEE Int.

Conf. on EMBS/BMES, 2002, 2, pp 999-1000.

[12] Wang, Y., Dandpat, A.: 'Classification of M-fish Images using Fuzzy C-means Clustering Algorithm and Normalization Approaches', Asilomer Conf. on Signals, Systems and Computers, 2004,1, pp. 41- 44.

[13] Karvelis, P., Tzallas, A., Fotiadis, D., Georgiou, I.: 'A Multichannel Watershed-Based Segmentation Method for Multispectral Chromosome Classification', IEEE Trans. Medical Imaging, 2008, 27(5), pp. 697-708.

[14] Karvelis, P., Fotiadis, D., Georgiou, I., Sakaloglou, P.: 'Enhancement of the classification of Multichannel chromosome images using support vector machines', 31st Ann. Int. Conf. of the IEEE EMBS, USA, 2009, pp. 3601-3604.

[15] Choi, H., Castleman, K., Bovik, A.: 'Joint Segmentation of M-FISH Chromosome Images', Proc. IEEE Int. Conf. on EMBS, San Francisco, USA, 2004, pp. 1636-1639.

[16] Choi, H., Castleman, K., Bovik, A.: 'Segmentation and Fuzzy-Logic Classification of M-FISH Chromosome Images', Proc. IEEE Int. Conf. on Image Processing, 2006, pp. 69-72.

[17] Choi, H., Bovik, A., Castleman, K.: 'Feature Normalization via Expectation Maximization and Unsupervised Nonparametric Classification for M-FISH Chromosome Images', IEEE Trans. Medical Imaging, 2008, 27(8), pp. 1107-1119.

[18] Choi, H., Bovik, A., Castleman, K.: 'Maximum likelihood decomposition of overlapping and touching M-FISH chromosomes using geometry, size and color information' Proc. 28th IEEE EMBS Annual Int. Conf., USA, 2006, pp. 3130-3133.

[19] Carothers, A., Piper, J.: Computer aided classification of human chromosomes: A Review. Statist. Comput, 1994, 4 (3), pp. 161-171.

[20] Lee, C., Gisselsson, D., Jin, C., Nordgren, A., Ferguson, D., Blennow, E., Fletcher, J., Morton, C.: 'Limitation of chromosome classification by multicolor karyotyping' Am. J. Hum. Genet. 2001, 68(4), pp. 1043- 1047.

[21] Hua, J., Xiong, Z., Wu, Q., Castleman, K.: 'Wavelet- Based Compression of M-FISH Images, IEEE transactions on Biomedical Engineering, 2005, 52 (5), pp. 890-900.

[22] Goienetxea, I., Barandiaran, I., Jauquicoa, C., Maclair, G., Grana, M.'Image anlayis pipeline for automatic karyotyping', Hybrid artificial intelligent systems, LNCS, 7209, 2012, pp. 392- 403.

[23] Choi, H., Bovik, A., Castleman, K.: 'Color Compensation of Multicolor FISH Images',IEEE Trans. Medical Imaging, 2009, 28(1), pp. 129-136.