

Correlation-based feature selection and classification via regression of segmented chromosomes using geometric features

Tanvi Arora¹ · Renu Dhir¹

Received: 3 February 2016 / Accepted: 18 July 2016 / Published online: 29 July 2016
© International Federation for Medical and Biological Engineering 2016

Abstract The genetic defects in the humans are uncovered by studying the chromosomes, as they are the genetic information carriers. They are non-rigid objects and they appear in different orientations when they are imaged. To find out the genetic defects, the chromosomes are pre-processed so that they are not touching, overlapping, and bent, and the noise is also discarded. The presence of bends, overlaps, or touches makes it difficult to uncover the genetic abnormalities. So there is a need for development of an efficient technique to classify the segmented chromosomes into different types and then pre-process them in order to correct their orientation. In this work, a hybrid classification technique based upon correlation-based feature selection and classification via regression approach, which will classify the segmented chromosomes into five categories viz; straight, overlapping, bent, touching, or noise is presented. The performance evaluation has been done using 1592 segmented chromosomes from Advance Digital Imaging Research data set. The over-all accuracy of 94.78 % has been obtained for the five class problem. The performance of the proposed classifier has been compared with Bayes Net, Naïve Bayes, Radial Bias Feed Forward Network, and k-nearest-neighbour classifiers. Based upon this categorization, different pre-processing techniques will be applied to correct the orientation of the chromosomes.

Keywords Genetic defects · Chromosomes · Feature extraction · Classification · Feature selection

✉ Tanvi Arora
tanviverma@rediffmail.com

¹ Department of Computer Science and Engineering,
Dr. B.R Ambedkar National Institute of Technology,
Jalandhar, Punjab, India

1 Introduction

A healthy human has 46 chromosomes [29] out of which 22 chromosomes are paired and 23rd and 24th are sex-determining chromosomes, which may be either XX or XY for male and female, respectively. The chromosomes are thin thread-like structures and carry the genetic information of an individual. They are made up of DNA (deoxyribonucleic acid) molecules that are tightly bound around the specific type of proteins termed as histones [1, 3]. The DNA carries the instructions for protein synthesis. Any alteration either in number of chromosomes or the structure of chromosomes results in a condition termed as genetic defect. In order to know the cause of the genetic defects, the chromosomes are imaged and studied.

The chromosomes are generally imaged during metaphase of cell division, as during that phase they appear the longest. For the purpose of chromosome imaging, the cell division phase is inhibited by treating it with colchicines or colcemid [9] as shown in Fig. 1. They are imaged during the metaphase of cell division, and the images thus generated are termed as metaspread images. The chromosomes are non-rigid objects, so they are present in different orientations in the metaspread images [28]. They are extracted from the metaspread images, and then they are classified into 24 classes and arranged in decreasing order of size to form a karyogram for the purpose of analysis [4] as depicted in Fig. 2. The karyograms are analysed for the purpose of uncovering the genetic defects.

To uncover the genetic defects, the features are extracted from the individual segmented chromosomes from the selected metaspread image [13]. But the segmented objects from the metaspread image contain objects in different orientations. These objects may be bent, overlapping, touching, or may be noise [4] as shown in Fig. 3. In order

Fig. 1 Metaphase chromosome imaging process

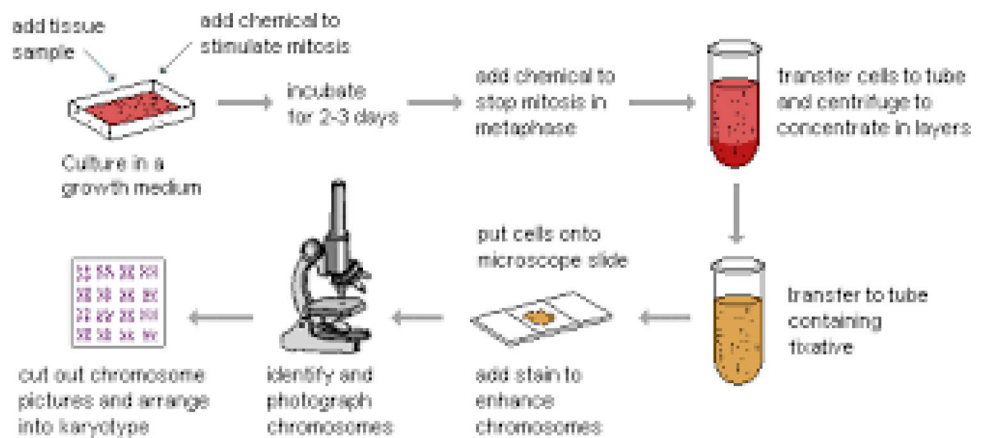


Fig. 2 Karyogram generated by a human expert

to extract the features and uncover the defects, these bent chromosomes need to be straightened, the touching and overlapping chromosomes need to be further segmented to create individual straightened chromosomes, and the noise needs to be removed.

In order to straighten the bent chromosomes, disentangle the touching or overlapping chromosomes and remove the noise, these segmented objects need to be classified into five classes viz; straight chromosome, bent chromosome, touching chromosome, overlapping chromosome, and noise. Classifying the segmented objects into these five categories will help in correcting their orientation by assigning them the appropriate pre-processing technique.

Over the years, several pre-processing techniques have been proposed for correcting the orientation of the segmented chromosomes. In the last few years, some studies have been carried out by researchers for the purpose of classification of the segmented objects from the metaphase images. Thereby improving the quality of the metaphase images, which are used for the generation of karyograms. The karyograms are further used to diagnose the genetic defects. In the following part, the previous works are being presented.

MetaSel: A metaphase selection tool using a Gaussian-based classification technique [30] was proposed for the selection of the metaphase images. In this study, they have ranked the metaphase images. The images have been ranked considering the type of objects present. The objects have been classified into four categories viz; straight,

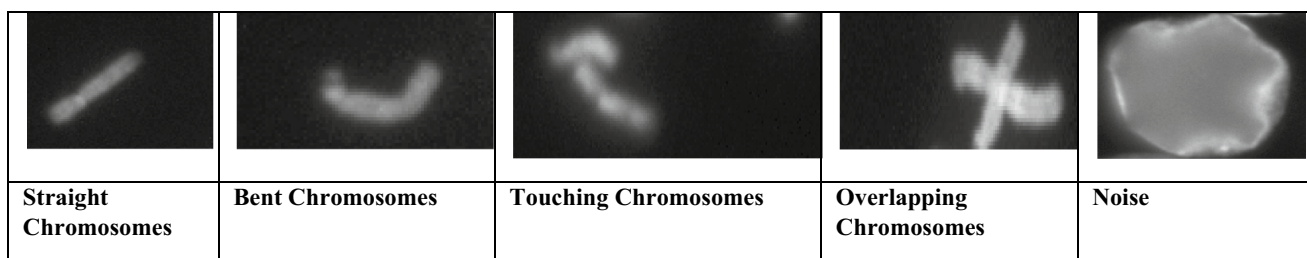


Fig. 3 Sample images of each class

skewed, overlapping or touching, and noise. The Gaussian-based classifier was used considering the four parameters viz; area ratio, rectangle width ratio, max width ratio, and height ratio.

Band classification based on chromosome shapes [31] was proposed to classify the metaspread chromosome images into low band resolution and high band resolution. The low band resolution images are used for finding out numerical abnormalities, whereas high band resolution images are used to find out structural abnormalities. They classified the objects into two classes', viz. individual chromosomes and non-individual chromosomes. The features considered were area ratio, average width, maximum width, and height ratio.

Shape information and dark paths [22] were used to classify the objects of the G-banded chromosome images. The geometric features were used for distinguishing between single chromosomes and cluster of multiple chromosomes. Then dark paths were used to distinguish between the touching and overlapping chromosomes from the cluster of multiple chromosomes. They considered the cross points, end points, and cut points to distinguish between single chromosomes. The dark path based upon the calculation of the curvature has been used to distinguish between the touching and overlapping chromosomes, and artificial neural networks were used for the purpose of classification.

Maximum-likelihood decomposition of overlapping and touching MFISH chromosomes using geometry, size, and colour information [12], in which the segmented chromosomes were classified as cross-shaped chromosomes, T-shaped chromosomes, and I-shaped chromosomes, was proposed. The parameters used were cross points, end points, and cut points. The unsupervised fuzzy logic classifier was used for the purpose of classification.

Automatic identification of overlapping/touching chromosomes in microscopic images using morphological operators [18], in which the segmented chromosomes have been classified as either single chromosomes or as a cluster of touching and overlapping chromosomes, was proposed. They have used size information to distinguish between single chromosome and the cluster of chromosomes. Further, they have counted the number of chromosomes present in the cluster considering the count of the end points.

From the literature review, it has been found that there are very few studies on classifying the segmented objects of the metaspread images into various classes based on geometric features. The dark paths were used to classify the touching and overlapping chromosomes using neural networks [22] with the accuracy of 86 %. The objects of the metaspread images were classified into three classes using geometrical and colour-based information with an accuracy of 89.43 % using unsupervised fuzzy logic [12]. The best performer so far has been the MetaSel [30] that classified

the objects of the metaspread image into four classes with the accuracy of 93.19 % using the Gaussian model.

The proposed work classifies the objects of the metaspread image into five different classes, using geometric features. Further, we will be using feature selection to improve the accuracy of the classifier. The feature selection will be carried out using correlation-based feature selection (CFS) scheme, and the classification of the objects into five classes will be done using classification via regression (CVR) classifier (CVR). The overall classification accuracy of the approach is 94.78 %. So it seems to be better than the previously reported methods.

The remaining paper is organized as follows: Sect. 2 has methods, Sect. 3 has results, Sect. 4 has discussion, and Sect. 5 concludes the paper.

2 Methods

2.1 Image data set used

In order to carry out the proposed work, the ADIR database has been used (www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml). The database contains 200 MFISH images of human metaphase chromosomes of 517×645 pixels. It has images recorded at different wavelengths and for each image a DAPI (4',6-diamidino-2-phenylindole) image is also created. We have taken the DAPI images for the purpose of evaluation. There is a corresponding ground truth for each of the image depicting the number of chromosomes present in the image. It has 124 metaspreads of males, 43 metaspreads of the females, 16 metaspreads have some genetic defects because of which they have only single \times chromosome, and for 17 metaspread images the ground truth information is not available. The 36 images have been randomly selected out of these 200 hundred images, and the individual objects have been extracted from them using region-based active contours. A total of 1592 objects were segmented out from these 36 metaspread images. The segmented objects have been classified into five categories based upon the geometric features as straight chromosomes, bent chromosomes, touching chromosomes, overlapping chromosomes, or noise. The ground truth for the above 1592 objects was assigned by an experienced cytogeneticist.

2.2 Performance evaluation metrics

In order to evaluate the performance of the proposed classification approach, precision, recall, F-measure, and accuracy have been used [27].

Following parameters have been used to measure the performance:

Fig. 4 Proposed algorithm

For each metaphase chromosome image in the dataset D

- a) Segment the objects and place in set R
- b) r_i =Object under study
1. For each r_i in R obtain the geometric features
 - a) x_1 =length
 - b) x_2 =area
 - c) x_3 =circularity
 - d) x_4 =convex area
 - e) x_5 =perimeter
 - f) x_6 = euler number
 - g) x_7 = number of end points
 - h) x_8 =major axis length
 - i) x_9 =extent
 - j) x_{10} =minor axis length
 - k) x_{11} =orientation
 - l) x_{12} =solidity
 - m) x_{13} =number of branch points
 - n) x_{14} =deviation
 - o) x_{15} =eccentricity
 - p) x_{16} =ratio of minor axis to major axis
 - q) x_{17} =equiv diameter
2. S is the set of all the features
3. *Select features using CFS* and place in set S_f
4. For each r_i in R
 - a) Apply CVR based classification scheme using the selected feature set S_f

True positive (TP): relevant object retrieved.

False positive (FP): irrelevant object retrieved.

True negative (TN): irrelevant object and not retrieved.

False negative (FN): relevant object but not retrieved.

Precision: It is the fraction of the retrieved objects that are relevant.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

Recall: It is the fraction of relevant objects that are successfully retrieved. It measures how complete the results are.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

F-Measure: It is the performance measure for the positive class. It is the harmonic mean of precision and recall.

$$\text{F - measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Accuracy: It is the overall performance measure of the classifier. Better the accuracy better the performance of the classifier.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

Confusion matrix: The confusion matrix depicts the predicted and actual values obtained by the classifier.

The above performance evaluation metrics are used to compare the proposed method with other state of the art classifiers such as Bayes Net, Naïve Bayes, RBF Network, kNN by considering all the features and the selected set of reduced features. Finally, the results of the proposed approach are compared with the techniques as proposed by Castleman [12], Moallem et al. [22], and MetaSel by Uttamatin [30].

2.3 Proposed method

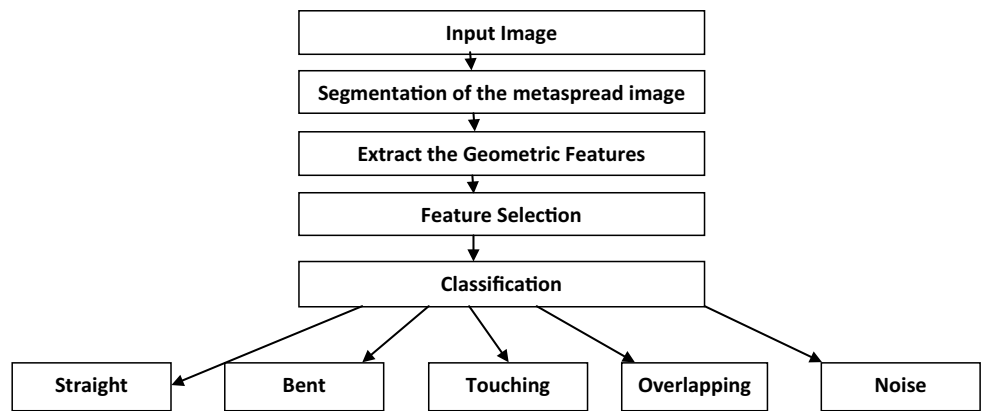
Aims of the proposed method are to categorize the segmented chromosomes from the metaspread image into five categories considering the geometric features, to evaluate the contribution of various geometric features in classifying the segmented objects into five distinct classes, to select the features by feature selection technique and enhance the accuracy of the classifier, and to compare and contrast the performance of various classifiers for classification of the metaspread objects into five different classes.

In order to implement the proposed technique, the algorithm and flowchart of the proposed method are illustrated in Figs. 4 and 5. The various steps of the proposed method are described as follows:

1. Segmentation of the objects from metaspread images

The objects for the purpose of classification have been extracted from the MFISH images of the ADIR data set.

Fig. 5 Proposed approach for the classification system



The segmentation has been carried out using region-based active contours [5]. The active contours based approach has been used as the conventional approaches cannot segment the metaspread images efficiently [6]. The segmentation technique uses the local intensity values of the nearby regions of the objects and finds the approximate intensity values along both sides of the contour. The technique works with the grey-scale images. This model retrieves the intensity of pixels in the nearby regions. It is capable of segmenting the images even in the presence of intensity inhomogeneity and can perform quite well for images of weak boundaries [7]. The method has been implemented using MATLAB 2014.

2. Geometric feature extraction

The automatic region-based active contour segmentation results in the generation of different orientations of the chromosomes as they are non-rigid objects and noise. In order to generate the karyogram from these segmented objects, the noisy objects needs to be discarded and the disoriented chromosomes need to be pre-processed so that they can be analysed for the purpose of uncovering the genetic defects. In order to pre-process the segmented chromosomes they need to be categorized into different classes based upon their orientation. In order to classify them, geometric features are extracted, based upon which the segmented objects are classified into five different classes, viz. straight chromosomes, bent chromosomes, touching chromosomes, overlapping chromosomes, and noise. Based upon the different approaches as studied in the literature [19, 20, 23, 25, 32] a set of 17 geometric features as listed in Table 1 were determined and extracted. The feature extraction has been carried out using the MATLAB 2014 software. The features extracted are further normalized so that they have a unit variance and zero mean value, this process has been carried out using the Waikato environment for knowledge analysis (WEKA tool).

3. Selection of features

A total of 17 geometric features have been extracted to be used for this study. Some features are independent, and few of them are derived features. Out of these 17 features, some might not contribute towards the classification results. Therefore, a feature selection approach has been used to search the combination of those features that have the ability to classify the objects into five classes. The selected features will have high discriminating value and will be quite meaningful for the classification purpose. The redundant and irrelevant features have been removed. In this work, the feature selection has been done using CFS [17]. It is a simple filter-based algorithm. It uses a heuristic function based upon correlation in finding the relevance of the features. It quickly finds out the redundant, relevant, irrelevant, and noisy features. On an average, it may eliminate more than half of the features. In most of the experiments, the classification performance has either been same or has improved by using the reduced feature set as obtained by CFS. There is no requirement of specifying any minimum thresholds or the minimum number of features to be selected; it is a fully automatic algorithm. The importance of the selected features is judged based upon the prediction power of the features and the redundancy associated with them. Those features are chosen that have least inter correlation and more correlation for the class. Following equation illustrates the function that evaluates the subset of features:

$$\text{Merit}_s = \frac{N\overline{p_{ij}}}{\sqrt{N + N(N - 1)\overline{p_{ij}}}} \tag{5}$$

where Merit_s represents the heuristic-based merit of the subset of N features that have been selected in subset named as s , $\overline{p_{ij}}$ is the mean value of the feature class correlation, and $\overline{p_{ij}}$ is the average value of the feature to feature inter correlation.

WEKA tool has been used for the purpose of feature selection. The CFS attribute evaluator was used with best

Table 1 Geometric features

Type of features	Feature	Description
Spread	Length	It is the distance between the two extreme end points. Let (x_1, y_1) and (x_n, y_n) be the two extreme end points and the length can be calculated as follows: Length = $\sqrt{(x_n - x_1)^2 + (y_n - y_1)^2}$
	Area	It is the number of the pixels in the object that have intensity value equal to one Area = $\sum p_i$ where p_i are the pixels of the object having intensity value = 1
	Convex area	It is the area of the convex hull, where convex hull is the minimum region that is convex and it covers the given region. It is the sum of the pixels in the convex image. Convex Area = $\sum p_i$ where p_i are the pixels of the convex hull
	Perimeter	It is the sum of the distance between the adjoining pixels around the boundary of the region. Perimeter = \sum Distance between adjoining pixels Distance = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
	Equi-diameter	It specifies the diameter of the circle with the same area as the region. It is computed as Equi diameter = $\sqrt{(4 \times \text{Area})/\pi}$
	Major axis length	It is the length of the major axis of the ellipse that has the same normalized second central moments as the region
	Minor axis length	It is the length of the minor axis of the ellipse that has the same normalized second central moments as the region
	Ratio of minor axis to major axis	It is the ratio of minor axis to major axis. Ratio of minor axis to major axis = $\frac{\text{Length of minor axis}}{\text{Length of major axis}}$
	Solidity	It specifies the pixels in the convex hull that are also in the region it is computed as area/convex area Solidity = $\frac{\text{Area}}{\text{Convex area}}$
	Eccentricity	It specifies the eccentricity of the ellipse that has the same second moments as the region. It is the ratio of the distance between the foci of the ellipse and its major axis length. Eccentricity = $\frac{\text{Distance between foci of ellipse}}{\text{Length of major Axis}}$
	Extent	Ratio of number of pixels in the region to the number of pixels in the bounding box. Extent = $\frac{\text{Area}}{\text{Area of bounding box}}$
	Shape	Deviation
Euler Number		It is the number of objects in the region minus the number of holes in those objects Euler number = count of objects in the region – count of holes in those objects
Number of End Points		To calculate the number of end points of an image, the image is first skeletonized. Then the number of end points is calculated as follows: Number of end points = $\sum p_i$ where p_i are the pixels obtained by setting the intermediate pixels to 0
Circularity		The amount of roundness calculated as follows: Circularity = $\frac{4 \times \text{Area} \times \pi}{\text{Perimeter}^2}$
Orientation		It is the angle between the x -axis and the major axis of the ellipse that has the same second moments as the region
	Number of branch points	To calculate the number of branch points of an image, the image is first skeletonized. Then the number of branch points is calculated as follows: No of branch points = $\sum p_i$ where p_i are the pixels that are having four connectivity

first searching method that used forward selection heuristic approach and had the stopping criteria after five iterations if no change in subsets takes place. The merit of each subset of features was evaluated using heuristic function using Eq. 5 as given above. Here in this

study, the subset that has the highest merit of 0.546 was selected. The features of the selected subset are: (1) Convex Area (2) Minor Axis Length (3) Solidity (4) Number of Branch_pts (5) Number of End_pts (6) Deviation (7) Orientation.

4. Classification

Based upon the selected features, the segmented objects have been classified into five classes' viz; straight chromosomes, bent chromosomes, touching chromosomes, overlapping chromosomes, and noise using CVR classifier. The CVR classifier is based upon the model trees [14, 16]; they are a kind of decision trees which have linear regression at the leaf nodes. The model trees are generated by first constructing a simple decision tree, the second stage prunes the tree by replacing the sub-trees by using linear regression.

The CVR has been implemented using the random forest algorithm. In this algorithm, a large number of decision trees are built during training time, in order to classify an object it is given to each of the trees in the forest. Each tree gives its classification which is treated as a vote for that class; the object is assigned to the class that has the maximum number of votes. This algorithm is a powerful tool for predicting, it is based upon the law of large numbers so it does not over fit. The accuracy of the algorithm as a classifier depends upon the random inputs and features [11].

The classification performance of the proposed classifier has been compared with Bayes net, Naïve byes, RBF, and kNN classifiers. The Bayes net classifier is a probability-based graphical model, it represents a collection of variables that are random and their conditional dependencies using a directed acyclic graph [15]. The Naïve Bayes classifier is as probability-based classifier; it is based upon the Bayes theorem. It is quite efficient in predicting the classes of unknown data sets. It assumes that a particular feature of a class is not related to other features [26]. The radial basis function classifier is a type of neural network, it consists of three layers, namely input layer, hidden layer, and output layer. The input layer passes the coordinates of the input vector to each of the nodes of the hidden layer. Each node of the hidden layer produces output based upon the radial basis function. Then each node of the output layer produces an output based upon the inputs received from each of the hidden layers [10, 24]. The k-nearest neighbours (kNN) classifiers are nonparametric methods that are based upon instance-based learning. In this classification scheme, the class membership is assigned based upon the voting of its k-neighbours [2].

The performance of all the classifiers is estimated by performing k-fold cross-validation. In k-fold cross-validation, all the training samples are placed randomly, and they are divided into k groups. In order to estimate the performance of the classifier, k iterations are performed. For all iterations, the classifier is trained with $k - 1$ groups and the kth group is used for the testing purpose. The results of testing are used to find the count of the data objects for which the classifier gave wrong results. Based upon the count of wrongly classified data objects from each iteration, the

error rate of the classifier is calculated. In order to calculate the accurate estimate of the classifiers accuracy, the above k-fold cross-validation is performed several times with different random groups of data objects [8].

3 Results

The work presented in this paper has been tested on 1592 images extracted from 36 randomly selected MFISH images from ADIR data set. To get the best performance of the classifier, the various parameters are tested and tried. The best ones have been taken to carry out the experimental work. The performance evaluation has been carried out using k-fold cross-validation, the results presented in this work are calculated using threefold cross-validation. The training of the classifier was carried out using different percentages of training and testing data sets, for the purpose of result presentation the training of the classifier has been carried using two-third of the data set, and the testing has been carried on the remaining one-third. Five rounds have been taken to record the readings taking different random data samples so that no biasing takes place. The results of the different rounds are averaged to determine the overall accuracy.

In the following sections, the performance of the proposed method is compared with various state of the art classifiers considering the shape-based features, spread-based features, and selected features. Further, the proposed approach has been compared with the works that have been carried out in this field over the period of time.

3.1 Classification performance of different geometric features using CVR classifier

The geometric features have been classified into two categories viz; spread-based features and shape-based features. With shape-based features alone, accuracy of 87.1 % was obtained. Then CVR classifier was tested using the spread-based features alone, and overall accuracy of 91.7 % was achieved. But spread-based features and shape-based features individually could not give very good precision for all the categories. Then, both the features were clubbed together and an overall accuracy of 93.21 % was obtained. Table 2 illustrates the results obtained.

3.2 Classification performance with selected features with CVR classifier

The results of classification on selected features of shape showed an accuracy of 78.4 %, whereas the selected features of spread showed an overall accuracy of 90.6 %. The CVR classifier showed the best performance of 94.78 %

Table 2 Classification performance of all geometric features using CVR classifier

Model	Confusion matrix					True positive	False positive	Precision	Recall	F-measure	Accuracy (%)	
	n	tc	sc	oc	bc							
Shape	n	47	0	17	7	7	0.6,025,641	0.0033025	0.9038462	0.6025641	0.7230769	87.1
	tc	0	70	0	0	0	1	0.0118265	0.7954545	1	0.8860759	
	sc	4	1	849	1	69	0.9188312	0.1392216	0.9012739	0.9188312	0.9099678	
	oc	1	11	0	61	3	0.8026316	0.0065963	0.8591549	0.8026316	0.829932	
	bc	0	6	76	2	360	0.8108108	0.0688153	0.8200456	0.8108108	0.815402	
Spread	n	58	0	11	4	5	0.7435897	0.003963	0.90625	0.7435897	0.8169014	91.7
	tc	0	70	0	0	0	1	0.0026281	0.9459459	1	0.9722222	
	sc	5	0	893	1	25	0.9664502	0.0703593	0.95	0.9664502	0.9581545	
	oc	1	2	0	51	22	0.6710526	0.0145119	0.6986301	0.6710526	0.6845638	
	bc	0	2	36	17	389	0.8761261	0.0452962	0.8820862	0.8761261	0.879096	
Spread + shape features	n	59	0	9	6	4	0.75641026	0.004623514	0.89393939	0.75641026	0.81944444	93.21
	tc	0	70	0	0	0	1	0.006570302	0.875	1	0.93333333	
	sc	4	1	894	0	25	0.96753247	0.068862275	0.95106383	0.96753247	0.95922747	
	oc	2	5	0	63	6	0.82894737	0.006596306	0.8630137	0.82894737	0.84563758	
	bc	1	4	37	4	398	0.8963964	0.030487805	0.91916859	0.8963964	0.90763968	

with the selected features of both types. The results are evident that the selected features have good correlation. The summarized results are shown in Table 3.

3.3 Classification performance with other classifiers considering all features

The comparison of the results for all the features is shown in Table 4. The CVR classifier outperforms all other classifiers with an accuracy of 93.21 %, whereas other classifiers have comparatively lesser accuracy. If all the features are considered, the computational cost is too high.

3.4 Classification performance with other classifiers considering selected features

The performance comparison of the selected features is also performed with other classifiers as depicted in Table 5. The CFS–CVR classifier gave an overall accuracy of 94.78 % as compared to other classifiers.

3.5 Comparison with previous works on chromosome classification

Table 6 gives the comparative analysis of the proposed method with other works that used different features for the classification of the chromosomes. As evident from the table, the proposed work has the overall best performance. The performance improvement over the previously reported techniques can be attributed to the consideration of both shape- and spread-based geometric features.

As evident from the above comparisons, the proposed model outperforms all other techniques proposed so far. It is capable of classifying the segmented objects of the metaspread images into five categories with an accuracy of 9.78 % by just using seven features.

4 Discussion

In order to create a karyogram from the metaspread images, the segmented objects need to be pre-processed, as the segmented objects may be either single chromosomes, bent chromosomes, or cluster of touching, or overlapping chromosomes, or may be noise. Each type of objects requires different type of pre-processing in order to retrieve the relevant features and organize the chromosomes in the karyogram or to discard the noise. So a method needs to be developed to classify the segmented objects into the five categories. In this work, shape- and spread-based features have been considered so as to classify the objects into different categories. As it can be seen from the above results that neither the shape-based features alone nor the spread-based features alone can classify the objects into five categories efficiently.

The classification performance of the proposed CVR classifier using the shape features is 87.1 %, and for the spread-based features it is 91.7 %. But when the shape- and spread-based features are combined, the classification accuracy increases to 93.21 %. The major factor for the increase in the classification accuracy is the correlation between the features. When features based upon the correlation

Table 3 Classification performance of selected geometric features using CVR classifier

Model	Features selected	Confusion matrix				True positive	False positive	Precision	Recall	F-Measure	Accuracy (%)	
		n	tc	sc	bc							
Shape	Convex area minor axis length solidity	n	48	0	18	6	6	0.6153846	0.7868852	0.6153846	0.6906475	78.4
		tc	0	70	0	0	1	0.0124836	0.7865169	1	0.8805031	
		sc	9	1	779	1	134	0.8430736	0.8340471	0.8430736	0.8385361	
		oc	2	12	1	56	5	0.7368421	0.8358209	0.7368421	0.7832168	
		bc	2	6	136	4	296	0.6666667	0.6712018	0.6666667	0.6689266	
Spread	Number of branch points number of end points deviation orientation	n	56	0	12	4	6	0.7179487	0.875	0.7179487	0.7887324	90.6
		tc	0	70	0	0	1	0.0059133	0.8860759	1	0.9395973	
		sc	7	0	889	2	26	0.9621212	0.9518201	0.9621212	0.9569429	
		oc	0	2	0	45	29	0.5921053	0.6338028	0.5921053	0.6122449	
		bc	1	7	33	20	383	0.8626126	0.8626126	0.8626126	0.8626126	
Spread + shape features	Considering all above selected features	n	71	0	2	3	2	0.91025641	0.9220779	0.9102564	0.916129	94.78
		tc	0	70	0	0	1	0.0052562	0.8974359	1	0.9459459	
		sc	5	1	896	0	22	0.96969697	0.9593148	0.969697	0.9644779	
		oc	0	3	0	70	3	0.92105263	0.9459459	0.9210526	0.9333333	
		bc	1	4	36	1	402	0.90540541	0.9370629	0.9054054	0.9209622	

Table 4 Classification performance of all features using different classifiers

Classifier	Confusion matrix					True positive	False positive	Precision	Recall	F-measure	Accuracy (%)	
	<i>N</i>	tc	sc	oc	bc							
Naïve bytes	n	47	0	17	12	2	0.6025641	0.007926024	0.79661017	0.6025641	0.68613139	76.8
	tc	0	52	5	10	3	0.74285714	0.017082786	0.66666667	0.74285714	0.7027027	
	sc	8	1	790	1	124	0.85497835	0.230538922	0.83686441	0.85497835	0.84582441	
	oc	2	14	0	58	2	0.76315789	0.029683377	0.5631068	0.76315789	0.64804469	
	bc	2	11	132	22	277	0.62387387	0.114111498	0.67892157	0.62387387	0.65023474	
Bayes net	n	53	1	7	15	2	0.67948718	0.00660502	0.84126984	0.67948718	0.75177305	78.9
	tc	0	68	0	0	2	0.97142857	0.014454665	0.75555556	0.97142857	0.85	
	sc	10	1	758	1	154	0.82034632	0.151197605	0.88242142	0.82034632	0.85025238	
	oc	0	10	0	61	5	0.80263158	0.025725594	0.61	0.80263158	0.69318182	
	bc	0	10	94	23	317	0.71396396	0.141986063	0.66041667	0.71396396	0.68614719	
Rbf	n	58	1	12	5	2	0.74358974	0.052166225	0.464	0.74358974	0.57142857	78.5
	tc	0	69	0	1	0	0.98571429	0.009590235	0.8625	0.98571429	0.92	
	sc	14	11	819	0	80	0.88636364	0.322033898	0.86029412	0.88636364	0.87313433	
	oc	3	8	1	61	3	0.80263158	0.01642178	0.7625	0.80263158	0.78205128	
	bc	4	13	166	18	243	0.5472973	0.287533512	0.00923788	0.05128205	0.01565558	
Knn	n	39	3	27	6	3	0.5	0.005284016	0.82978723	0.5	0.624	86.1
	tc	0	70	0	0	0	1	0.026938239	0.63063063	1	0.77348066	
	sc	6	1	883	0	34	0.95562771	0.158682635	0.89282103	0.95562771	0.92315734	
	oc	0	12	1	47	16	0.61842105	0.008575198	0.78333333	0.61842105	0.69117647	
	bc	2	25	78	7	332	0.74774775	0.046167247	0.86233766	0.74774775	0.80096502	
CVR	n	59	0	9	6	4	0.75641026	0.004623514	0.89393939	0.75641026	0.81944444	93.21
	tc	0	70	0	0	0	1	0.006570302	0.875	1	0.93333333	
	sc	4	1	894	0	25	0.96753247	0.068862275	0.95106383	0.96753247	0.95922747	
	oc	2	5	0	63	6	0.82894737	0.006596306	0.8630137	0.82894737	0.84563758	
	bc	1	4	37	4	398	0.8963964	0.030487805	0.91916859	0.8963964	0.90763968	

parameter were selected using CFS, the classification accuracy further increased to 94.78 %. Based on a combination of seven shape- and spread-based features, the proposed method is capable of classifying the chromosomes into five different categories.

Taking the combination of selected seven shape- and spread-based features, the performance of the proposed method is compared with other state of the art classifiers. The classification performance was compared with four different classifiers, namely Bayes Net, Naïve Bayes, RBF Network, and kNN by considering all the features and the selected set of reduced features. In both the cases, the proposed classifier outperformed all other classification methods. But certainly by selecting the features using the CFS approach, the classification accuracy of Bayes Net, Naïve Bayes, RBF Network, and kNN substantially increased. The CFS approach helps in selecting the most relevant features; thereby, the accuracy of classification increases.

The classification accuracy of the proposed method is compared with the previous reported works done by others; most of the approaches have classified the segmented

objects as either chromosomes or noise. The objects classified as noise were discarded, and the objects classified as chromosomes were pre-processed before karyogram generation using different approaches as per the discretion of the cytogeneticist. Thus, these approaches rely upon the cytogeneticist for pre-processing the chromosomes before feature extraction and subsequent karyogram generation. Few works have been reported in which the effort has been made to classify the segmented objects into three or four classes. Castleman [12] classified the segmented objects into three classes viz; single chromosomes, touching chromosomes, and overlapping chromosomes using unsupervised fuzzy logic with 89.43 % accuracy. They used geometrical features, colour-based features, and size-based features for the purpose of MFISH chromosomes. Moallem and his team [22] used neural networks to classify the segmented chromosomes into either single chromosome or a cluster of touching or overlapping chromosomes, with an accuracy of 86 % using shape and dark band paths. Their method worked just on G-banded images. Uttamatani and his team [30] used Gaussian model to classify the

Table 5 Classification performance with selected geometric features with other classifiers

Classifier	Confusion matrix					True positive	False positive	Precision	Recall	F-measure	Accuracy (%)	
	n	tc	Sc	oc	bc							
Naïve byes	n	19	0	17	15	27	0.24358974	0.002642	0.826087	0.2435897	0.3762376	80.5
	tc	0	62	5	3	0	0.88571429	0.0197109	0.673913	0.8857143	0.7654321	
	sc	2	1	848	0	73	0.91774892	0.2110778	0.8574317	0.9177489	0.8865656	
	oc	1	15	0	58	2	0.76315789	0.0211082	0.6444444	0.7631579	0.6987952	
	bc	1	14	119	14	296	0.66666667	0.0888502	0.7437186	0.6666667	0.7030879	
Bayes net	n	51	1	12	12	2	0.65384615	0.003963	0.8947368	0.6538462	0.7555556	85.1
	tc	0	66	0	0	4	0.94285714	0.0177398	0.7096774	0.9428571	0.809816	
	sc	6	2	834	0	82	0.9025974	0.1347305	0.9025974	0.9025974	0.9025974	
	oc	0	12	0	62	2	0.81578947	0.0151715	0.7294118	0.8157895	0.7701863	
	bc	0	12	78	11	343	0.77252252	0.0783972	0.7921478	0.7725225	0.7822121	
RBF	n	45	0	21	7	5	0.5769231	0.007926	0.789474	0.576923	0.666667	86.0
	tc	0	64	5	1	0	0.9142857	0.016426	0.719101	0.914286	0.805031	
	sc	6	5	837	0	76	0.9058442	0.133234	0.903888	0.905844	0.904865	
	oc	4	6	0	63	3	0.8289474	0.007916	0.84	0.828947	0.834437	
	bc	2	14	63	4	361	0.8130631	0.073171	0.811236	0.813063	0.812148	
Knn	n	47	1	21	6	3	0.6025641	0.003963	0.8867925	0.6025641	0.7175573	89.3
	tc	0	70	0	0	0	1	0.021025	0.6862745	1	0.8139535	
	sc	5	5	875	0	39	0.9469697	0.1137725	0.9200841	0.9469697	0.9333333	
	oc	0	10	0	60	6	0.78947368	0.0046174	0.8955224	0.7894737	0.8391608	
	bc	1	16	55	1	371	0.83558559	0.0418118	0.8854415	0.8355856	0.8597914	
CFS–CVR	n	71	0	2	3	2	0.91025641	0.003963	0.9220779	0.9102564	0.916129	94.78
	tc	0	70	0	0	0	1	0.0052562	0.8974359	1	0.9459459	
	sc	5	1	896	0	22	0.96969697	0.0568862	0.9593148	0.969697	0.9644779	
	oc	0	3	0	70	3	0.92105263	0.0026385	0.9459459	0.9210526	0.9333333	
	bc	1	4	36	1	402	0.90540541	0.0235192	0.9370629	0.9054054	0.9209622	

Table 6 Comparison with previous works on chromosome classification

References	Classifier used	Number of classes	Accuracy (%)
Castleman Choi et al. [12]	Unsupervised fuzzy logic	3	89.43
Moallem, Payman et al. [22]	Neural networks	3	86
Uttamatinin, Ravi et al. [30]	Gaussian model	4	93.19
Uttamatinin, Ravi et al. [31]	Gaussian model	2	84.16
Jahani [18]	–	2	–
Proposed method	CFS–CVR	5	94.78

segmented objects into four classes viz; straight chromosomes, bent chromosomes, clusters of chromosomes, and noise using four parameters, with an accuracy of 93.19 %. Their method worked upon just G-banded images. So comparing the proposed work with the work done by others, the proposed work outperforms the previous proposed methods both in performance of classification results and the distinct number of classes generated.

The main contribution of the proposed work will be that it will greatly reduce the dependence upon the cytogeneticist. Based upon this classification approach, different

pre-processing techniques can be automatically applied to correct the orientation of the chromosomes as per the classification results. The proposed method will also speedup the karyogram generation process, which is heavily dependent upon the experienced cytogeneticist for the purpose of feature extraction and classification. Moreover, the biasing in the results will also be removed, which might be there because of human behaviour.

The major limitation of the proposed classifier is that although it classifies all the segmented objects with very good precision, precision of noisy objects is not 1. It is

expected that the precision of these objects should be 1 as false removal or inclusion of noisy objects makes the karyogram generation task difficult.

5 Conclusions

In this work, shape- and spread-based geometric features have been analysed for the classification of the extracted objects from the metaspread images into five different categories. Precision, Recall, F-Measure, and Accuracy have been used to study the performance of the proposed classification scheme after feature selection using CFS. The combination of spread- and shape-based features gave an overall accuracy of 94.78 % using the CFS–CVR classifier. The performance of the classification is compared with five other classifiers, and the proposed classifier outperformed all other classifiers.

The proposed study is different from the previous studies in the following ways: (1) all the existing studies have classified the segmented objects from the metaspread images into at most four classes, but in the proposed study the objects have been classified into five classes. (2) Shape- and spread-based geometric features have been used for the purpose of classification. (3) Precision of more than 89.74 % has been achieved for all the classes. (4) Only seven features have been used for the classification of the five class problem.

Although the proposed work has been obtained with very good precision for all classes, work needs to be done in future to find out the feature set that gives a precision of 1 for the noisy objects, as the false removal or addition of noisy objects limits the task of karyogram generation.

Compliance with ethical standards

Conflict of interest The authors have no conflict of interest.

References

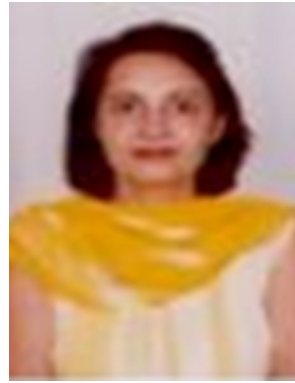
1. Alberts B (2000) Basic genetic mechanism. In: Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD, Grimstone AV (eds) *Molecular biology of the cell*, 5th edn. Garland Publishing Inc, New York, pp 191–234
2. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 46:175–185. doi:10.1080/00031305.1992.10475879
3. Arora T, Dhir R (2014) An efficient segmentation method for overlapping chromosome images. *Int J Comput Appl* 95(1):29–32
4. Arora T, Dhir R (2015) A review of metaphase chromosome image selection techniques for automatic karyotype generation. *Med Biol Eng Comput*. doi:10.1007/s11517-015-1419-z
5. Arora T, Dhir R (2016a) A novel approach for segmentation of human metaphase chromosome images using region based active contours. *Int Arab J Inf Technol*
6. Arora T, Dhir R (2016b) Segmentation approaches for human metaspread chromosome images using level set methods. In: *International conference on mass data analysis of images and signals MDA 2016* in New York
7. Arora T, Dhir R (2016c) Segmentation of human metaspread images using region based active contours. In: *International conference on recent trends in engineering and material science*, Jaipur National University, Jaipur, India, Mar 2016
8. Bengio Y, Grandvalet Y (2004) No unbiased estimator of the variance of K-fold cross-validation. *J Mach Learn Res* 5:1089–1105
9. Bickmore W (2001) Karyotype analysis and chromosome banding. In: Bickmore WA (ed) *Encyclopedia of life sciences*. M R C Human Genetics Unit, Edinburgh
10. Bors AG (1996) Introduction of the radial basis function (RBF) networks. University of Edinburgh, Edinburgh, pp 1–7
11. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. doi:10.1017/CBO9781107415324.004
12. Castleman HC, Bovik AC, Castleman KR (2006) Maximum-likelihood decomposition of overlapping and touching M-FISH chromosomes using geometry, size and color information. In: *Twenty-eighth annual international conference of the IEEE engineering in medicine and society*, New York
13. Devaraj S, Vijaykumar VR, Soundrarajan GR (2013) Leaf biom-etrics based karyotyping of g-band chromosomes. *Int J Hum Genet* 13:131–138
14. Frank E, Wang Y, Inglis S, Holmes G, Witten IH (1998) Using model trees for classification. *Mach Learn* 32(1):63–76
15. Friedman N, Geiger D, Goldszmit M (1997) Bayesian network classifiers. *Mach Learn* 29:131–163. doi:10.1023/a:1007465528199
16. Giraud-carrier C, Vilalta R, Brazdil P (2004) Is combining classifiers with stacking better than selecting the best one ? *Mach Learn* 54:255–273
17. Hall M (1999) Correlation-based feature selection for machine learning. PhD thesis, Department of Computer Science, Waikato University, New Zealand
18. Jahani S, Setarehdan SK, Fatemizadeh E (2011) Automatic identification of overlapping/touching chromosomes in microscopic images using morphological operators. In: *2011 7th Iranian conference on machine vision and image processing*. doi:10.1109/IranianMVIP.2011.6121574
19. Jahani S, Setarehdan SK, Veronica M (2012) An automatic algorithm for identification and straightening images of curved human chromosomes. *Biomed Eng: Appl Basis Commun* 24:1–9. doi:10.1142/S1016237212500469
20. Lerner B, Guterman H, Dinstein I, Romem Y (1995) Medial axis transform-based features and a neural network for human chromosome classification. *Pergamon Pattern Recognit* 28:1673–1683
21. M-FISH database established by advanced digital imaging research. (http://www.adires.com/05/Project/MFISH_DB/MFISH_DB.shtml) [WWW.Document], n.d
22. Moallem P, Karimizadeh A, Yazdchi M (2013) Using shape information and dark paths for automatic recognition of touching and overlapping chromosomes in G-band images. *Int J Image Graph Signal Process* 5:22–28. doi:10.5815/ijgsp.2013.05.03
23. Moradi M, Setarehdan SK (2006) New features for automatic classification of human chromosomes: a feasibility study. *Pattern Recognit Lett* 27:19–28. doi:10.1016/j.patrec.2005.06.011
24. Oyang Y-J, Hwang S-C, Ou Y-Y, Chen C-Y, Chen Z-W (2005) Data classification with radial basis function networks based on a novel kernel density estimation algorithm. *IEEE Trans Neural Netw* 16:225–236. doi:10.1109/TNN.2004.836229
25. Piper J, Granum E (1989) On fully automatic feature measurement for banded chromosome classification. *Cytometry* 10:242–255

25. Rish I (2001) An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial, pp 41–46. doi:[10.1039/b104835j](https://doi.org/10.1039/b104835j)
26. Sharma V, Singh S (2014) CFS–SMO based classification of breast density using multiple texture models. *Med Biol Eng Comput* 52:521–529. doi:[10.1007/s11517-014-1158-6](https://doi.org/10.1007/s11517-014-1158-6)
27. Somasundaram D, Kumar VRV (2014) Separation of overlapped chromosomes and pairing of similar chromosomes for karyotyping analysis. *Measurement* 48:274–281. doi:[10.1016/j.measurement.2013.11.024](https://doi.org/10.1016/j.measurement.2013.11.024)
28. Tjio JH, Levan A (1925) The chromosome number of man. *Genetics* 10:80–85
29. Uttamatin R, Yuvapoositanon P, Intarapanich A, Kaewkamnerd S, Phuksaritanon R, Assawamakin A, Tongsim S (2013) MetaSel: a metaphase selection tool using a Gaussian-based classification technique. *BMC Bioinform* 14:S13. doi:[10.1186/1471-2105-14-S16-S13](https://doi.org/10.1186/1471-2105-14-S16-S13)
30. Uttamatin R, Yuvapoositanon P, Intarapanich A, Kaewkamnerd S, Tongsim S (2013b) Band classification based on chromosome shapes. In: 13th international symposium on communications and information technologies (ISCIT). pp 464–468
31. Van Den Berg H, De France HF, Habbema JDF, Raatgever JW (1981) Automated selection of metaphase cells by quality. *Cytometry* 1:363–368. doi:[10.1002/cyto.990010602](https://doi.org/10.1002/cyto.990010602)



Tanvi Arora Tanvi Arora received her B.Tech. degree in Computer Science and Engineering from Punjab Technical University, Jalandhar, Punjab, India, in the year 2002, M.Tech. degree in Information Technology from Punjab Technical University, Jalandhar, Punjab, India, in the year 2007. She is currently pursuing Ph.D. in Computer Science and Engineering from Dr. B.R Ambedkar National Institute of Technology, Jalandhar, Punjab, India, and working as Associate Pro-

fessor in the department of Computer Science and Engineering at Baddi University of Emerging Sciences and Technology, Baddi, Himachal Pradesh, India. Her teaching and research interests include Image Processing, Pattern Recognition, Machine Learning, Data Mining, and Network Security.



Dr Renu Dhir Dr. Renu Dhir received her B.Tech. degree in Electrical Engineering from Guru Nanak Dev Engineering College, Ludhiana, Punjab, India, in the year 1983, M.Tech. degree in Computer Science and Engineering from TIET Patiala, Punjab, India in the year 1997, and Ph.D. degree in Computer Science and Engineering from Punjabi University, Patiala, Punjab, India, in 2007. She is currently working as Associate Professor at Dr. B.R Ambedkar National Institute of Technology, Jalandhar, Punjab, India. Her teaching and research interests include Image Processing, Pattern Recognition, Machine Learning, and Network Security.