Automated human chromosome segmentation and feature extraction: Current trends and prospects [version 1; peer review: 1 approved with reservations]

Umaya Bhashini Balagalla¹, Jagath Samarabandu², Akila Subasinghe¹

¹Department of Electrical and Electronic Engineering, University of Sri Jayewardenepura, Nugegoda, 10250, Sri Lanka
²Department of Electrical and Computer Engineering, Faculty of Engineering, Western University, London, Ontario, N6A 3K7, Canada

Abstract
Automated human chromosome segmentation and feature extraction aim to improve the overall quality of genetic disorder diagnosis by addressing the limitations of tedious manual processes such as expertise dependence, time-inefficiency, observer variability and fatigue errors. Nevertheless, significant differences caused by staining methods, chromosome damage which may occur during imaging, cell and staining debris, inhomogeneity, weak boundaries, morphological variations, premature sister chromatid separation, as well as the presence of overlapping, touching, di-centric and bent chromosomes pose challenges in automated human chromosome segmentation and feature extraction. This review paper extensively discusses how the approaches presented in literature have addressed these challenges, and their strengths and limitations. Human chromosome segmentation algorithms are presented under four broad categories; thresholding, clustering, active contours and convex-concave points-based methods. Chromosome feature extraction methods are discussed under two main categories based on banding-pattern and geometry. In addition, new insights for the improvement of fully automated karyotyping are provided.

Keywords
active contours, banding information, clustering, convex-concave, di-centric, geometric features, learning architectures, overlapping, thresholding, touching

This article is included in the Bioinformatics gateway.
Introduction
The structural and numerical abnormalities in the chromosomes cause genetic disorders. Therefore, chromosome analysis is a key method in diagnosing genetic disorders. Standard method of chromosome analysis is known as karyotyping. Conventional karyotyping classifies a metaphase image into 24 classes manually. Karyotyping requires expertise which may subject to inter-observer variability and intra-observer variability. In addition, it is a time-consuming and tedious process which may result in fatigue errors.

To address the aforementioned limitations of conventional manual karyotyping, automated human chromosome segmentation and feature extraction was introduced. Automated processes are repeatable because of the computer-based performance, hence no observer variabilities occur. Expertise is not required to handle chromosome segmentation and feature extraction algorithms. Moreover, these algorithms address the time in-efficiency of tedious manual process.

The automated karyotyping process consists of four main steps; image pre-processing, chromosome segmentation, chromosome feature extraction and clinical evaluation. This review paper extensively discusses two main steps among them; chromosome segmentation and chromosome feature extraction. Figure 1 illustrates the scope of the paper. Sections II and IV provide a comprehensive explanation of the approaches towards automated chromosome segmentation and chromosome feature extraction respectively. Finally, section V provides suggestions for the future improvement of automated human chromosome segmentation and feature extraction.

Background of karyotyping
Each chromosome in cell nuclei contains two sister chromatids which separate from each other in cell division. During metaphase, two chromatids attach to each other only at centromere, as shown in Figure 2. These chromatids contain unique sub-structures of genes which can be visualized by staining. Condensed areas of the chromosomes appear in dark bands while less condensed areas appear in light bands. Therefore, each chromatid produces a unique banding pattern after staining according to its sub-structures.

Different types of staining such as G-banding, R-banding, C-banding, Q-banding, NOR banding and T-banding are available for chromosome staining. 4,6-diamidino-2-phenylindole (DAPI) is a fluorescent stain which is commonly used for chromosome staining. The reason for this common use is, the possibility to avoid the need of preparation of multiple duplicate cell samples.

After the staining of chromosomes, a visible photographic representation of them is produced by using different imaging techniques such as light microscopy, fluorescent microscopy, electron microscopy and coherent x-ray diffraction imaging. However, producing original data might be difficult and inefficient in some research work. As a solution, publicly available databases of visible photographic representations of the stained chromosomes can be used. Edinburgh, Philadelphia, Copenhagen, BioImlab, Lisbon-K1 (LK1) and M-FISH are few examples for publicly available databases which can be accessed by the researchers.

The visualized stained chromosomes are paired and ordered manually to produce the karyotype during the conventional karyotyping. This is a time-consuming, tedious process and requires expertise to produce the karyotype accurately. In addition, fatigue errors and observer variabilities may occur due to the repetitive process. To address these limitations of inefficient manual karyotyping, semi-automated karyotyping was introduced in late 1970s. Later, the development of the computer processors and digital image processing paved the path towards fully-automated human chromosome segmentation and feature extraction. However, there are multiple challenges in automating this process as explained in the next sub-section.

Challenges in automated segmentation and feature extraction of chromosomes
The banding patterns which appear in the chromosome images depend on the method of staining. Therefore, developing general algorithms applicable to any type of staining is challenging. The majority of the chromosome image segmentation techniques are limited to the images produced by one staining method. It limits the application of these chromosome segmentation algorithms in practical scenario.

High resolution images are required to improve the accuracy of the process of segmentation and feature extraction. However, using the visualization techniques which produce high resolution images may damage the chromosomes. Despite the availability of different advanced imaging techniques, highly sensitive and specific methods are required to provide high resolution chromosome images with insignificant damage to the chromosomes. Developing imaging techniques which fulfil the aforementioned requirement, remains a challenge.
Chromosome images are distorted with artifacts such as staining and cell debris.\textsuperscript{28,29} Therefore, additional steps are required to filter the artifacts for accurate chromosome segmentation and feature extraction. However, over-filtering may lead to loss of information. Therefore, filtering the chromosome images while preserving the important information of the chromosomes poses a challenge.

In addition, presence of overlapping, touching and bent chromosomes,\textsuperscript{28} morphological variations, and presence of partial sister-chromatid separation or inter-phase cells\textsuperscript{30} raise a challenging nature in automating the chromosome segmentation and feature extraction. Figure 3 illustrates the appearance of overlapping chromosome clusters, overlapping and touching chromosome clusters, and highly bent chromosomes in the image.

**Review of chromosome segmentation methods**

Chromosome segmentation is the process of extracting the chromosomes from the image. This section extensively discusses chromosome segmentation algorithms under four sub-categories; thresholding, clustering, active contours and convex-concave points-based methods.

**Thresholding-based chromosome segmentation**

Thresholding relies on the intensity differences between the chromosomes and the background. Thresholding works for the chromosome images because chromosomes tend to appear in bright intensities compared to the background of the image.\textsuperscript{31–33}
Common global thresholding techniques which have been used to segment the chromosome images include Otsu’s thresholding and Kapur’s thresholding. Otsu’s thresholding extracts the chromosomes from the image by minimizing the intra-class variance of the pixels within each cluster. Kapur’s global thresholding divides a chromosome image into foreground and background based on the assumption that the chromosomes and the background of the image carry two different probability distributions. The optimal threshold level for Kapur’s thresholding is obtained by maximizing the total entropy. However, global thresholding often fails due to intensity in-homogeneity caused by the staining and...
lighter intensities of smaller chromosomes compared to larger chromosomes. Moreover, global thresholding can detect the chromosome clusters but fails in disentangling the cluster.

Consequently, local thresholding methods were introduced to segment the chromosome images.\textsuperscript{31,32,38,39} Adaptive local thresholding provides better results comparatively.\textsuperscript{35} However, simple thresholding-based techniques create holes inside the chromosomes when pixels inside the chromosomes are lighter than the surrounding pixels and, the boundaries might be cropped due to intensity fading.\textsuperscript{38} Sensitivity of the thresholded image to the quantization error of the digital image is another issue in thresholding-based chromosome segmentation.\textsuperscript{40}

In addition, simple thresholding-based techniques do not work when overlapping chromosomes present in the image because two overlapping chromosomes share a significant area in the histogram.\textsuperscript{17,41,42} When touching chromosomes present in the image, they are detected as a single chromosome. Therefore, the applicability of the thresholding-based segmentation methods is limited. Consequently, when overlapping and touching chromosomes present in the images, human intervention is needed to segregate them.\textsuperscript{43–45}

A detailed review on thresholding-based chromosome segmentation is presented below.

Poletti \textit{et al}.\textsuperscript{46} presented a review on thresholding strategies applied for the chromosome image segmentation. According to this review paper, thresholding is the first step in majority of the chromosome segmentation algorithms. Poletti \textit{et al}. implemented 11 approaches which include thresholding at least in one step. The implemented algorithms are as follows; Otsu’s thresholding, Kapur’s thresholding, improved Sobel with genetic algorithm, K-means clustering on algebraic moments, Fuzzy C-mean clustering, multi-stage adaptive thresholding, local re-thresholding, adaptive thresholding, region-based level sets, multi-thresholding with Particle Swarm and Nelder–Mead optimization, and multi-thresholding with Particle Swarm and expectation-maximization Optimization. The performance of these algorithms were compared with each other using Q-band images. Overall best performance was achieved for adaptive local thresholding and region-based level sets.

Yilmaz \textit{et al}..\textsuperscript{12} introduced a method for accurate segmentation of touching and overlapping chromosomes. To obtain the smooth and circular shape of the chromosomes, morphological closing was applied to the binary image several times. Adaptive local thresholding proposed by Grisan \textit{et al}.\textsuperscript{48} was used to filter the debris. However, a gray level threshold map along with a fixed size Gaussian kernel was used instead of the method of dividing the image into squares as proposed by Grisan \textit{et al}. The method based on watershed transform proposed by Karvelis \textit{et al}.\textsuperscript{47} was used to reconstruct a distance-transformed image. Clusters were identified by considering the average area of 46 chromosomes, convex hull, average thickness of the chromosomes and thickness along the skeleton. Nicholas Howe’s implementation\textsuperscript{18} was used for skeletonization. Concave points and gray-scale geodesic path were considered when separating the touching chromosomes. Overlapping chromosomes were separated by generating candidate combinations. This study achieved 97.8% accuracy with 6,678 chromosomes which was superior to the existing literature.

Grisan \textit{et al}.\textsuperscript{38} introduced a local adaptive thresholding for Q-band chromosome image segmentation. First, the image was divided into tessellation of squares of fixed dimension which include 100 pixels in each. Next, Otsu’s thresholding was applied to each square. After detecting chromosome clusters, medial axis transform, analysis of accumulation points, single chromosome likelihood measure, generation of candidate splits and, generation and exploration of the hypotheses tree were used to disentangle the chromosomes. The dataset used by Grisan \textit{et al}. includes 6,683 chromosomes. Overall accuracy for Q-band image segmentation using this algorithm was 94%. This algorithm achieved 90% accuracy for both touching chromosomes and overlapping chromosomes.

Sugapriya \textit{et al}.\textsuperscript{31} introduced a method to extract chromosomes from G-band metaphase images. This algorithm consists of three basic steps; local thresholding, region growing and, edge detection using gradient and Laplacian. The method of calculating the threshold value for local thresholding was as follows. First, the image was divided into sub-images. Next, the global threshold of each sub-image and largest element of the sub image was computed. The product of these two computed values was taken as the local threshold. However, quantitative analysis of the results on accuracy was not presented in the paper.

Di-centric chromosomes can be used to evaluate the radiation exposure. In the light of the above, a modified watershed-based method to segment the overlapped chromosomes was introduced by Li \textit{et al}..\textsuperscript{39} First, the local thresholding was applied and foreground pixels were identified. Next, watershed transform was used to separate the chromosomes. Then, inseparable clusters and debris were filtered by using morphological decision tree. However, this was not a promising
solution because the algorithm failed to segment the chromosomes accurately when considerable sister chromatid separation and non-uniform staining presented.

Stanley et al. introduced a homologous matching algorithm to segment chromosome images. First, the image was split into two using global thresholding. Next, local thresholding was applied to thresholded image for object separation. Finally, each chromosome was detected using five features; size, centromeric index, banding information, profiles along the medial axis and weighted density distribution.

Ji et al. proposed a two-step method to segment chromosomes. In the first step, chromosome image was thresholded by using the main peak of the gray level histogram as threshold. Conservative model was utilized to avoid over segmentation. After this initial segmentation using global thresholding, another step of re-thresholding was performed to classify touching and overlapping chromosomes. In order to do that, the chromosomes were classified into classes depending on size and shape. Single chromosomes were detected by their smaller area. Next, touching chromosomes were detected using concavity and density information. Overlapping chromosomes were separated using three assumptions; at least one node is included in the overlapping cluster, four 'cut points' present around the node, and mean density around the node is higher.

Altinsoy et al. used U-net based Convolution Neural Network (CNN) for chromosome segmentation. First, the image was filtered using a thresholding-based filtering technique. Next, the irrelevant objects were removed by considering the boundary of the image and, shape and size of the objects. Then, watershed transform, adaptive filtering and successive iteration were used to separate the touching chromosomes. To improve the accuracy of classification of single objects and chromosome clusters, 13 parameters were employed. Few of these parameters were not discussed in the previous literature towards classification. The parameters include object size, limited object size, average object size, maximum single chromosome size, ellipse, object pixel intensity, global pixel intensity, skeleton, thickness throughout the skeleton, skeleton size, average thickness, number of branch points, and number of end points. In this approach, cutting points and cutting paths of the clusters were identified using concavity properties and geodesic distance transform, Euclidean distance and, shape of the node. Another merit of the algorithm was the capability in handling multiple overlaps. This algorithm achieved 96.97% accuracy for 40 distorted chromosome images proving that neural networks work better for distorted images. However, applicability of this method might be limited due to the small dataset. Therefore, further research using a large dataset is suggested to verify the conclusion made.

Chromosome image segmentation using clustering-based algorithms

Clustering is one of the simplest and fastest methods of image segmentation.

Clustering-based algorithms can be divided into two as hierarchical algorithms and partitional algorithms. Both types of clustering-based algorithms work with intensity in-homogeneity and, have provided promising results in chromosome segmentation compared to simple thresholding-based algorithms. Hierarchical clustering algorithms provide higher accuracy compared to partitional clustering algorithms. However, the use of hierarchical clustering algorithms is limited due to the computational complexity. K-means and Fuzzy C-Mean (FCM) clustering algorithms, which are easy to implement and are low in computational complexity, have been used in research work to a greater extent. However, Fuzzy clustering performs better in chromosome segmentation compared to K-means clustering. Accuracy of the clustering-based algorithms can be improved by incorporating watershed transform. Clustering-based algorithms are relatively less time and computer complex. However, these algorithms require manual intervention in selecting the seed point to initiate. A slight mistake in selecting the seed point might lead to inaccurate segmentation. In addition, the output is highly sensitive to the number of clusters which is a pre-set parameter, hence expertise is required. Moreover, clustering-based algorithms are highly sensitive to outliers. Therefore, clustering-based algorithms often result in inaccurate chromosome segmentation, specially when stain debris and cell debris present in the image.

This section continues the discussion on clustering-based chromosome segmentation algorithms using the approaches presented in the literature.

Sayed and Hassanien proposed two methods to perform accurate chromosome segmentation when inter-phase cells present in the metaphase chromosome images. The first method is based on FCM and Grey Wolf Optimization. The second method is based on hybrid particle swarm optimization and K-Means clustering. Generic algorithms were used to select the best fit cluster in this algorithm. The experimental results had shown 94% and 95% accuracy for the methods respectively.
Munot et al.\textsuperscript{40} introduced a Random Walker Algorithm (RWA)-based clustering for accurate chromosome segmentation despite the scale and, the presence of chromosome clusters and highly bent chromosomes. Furthermore, this method overcomes the limitation of working only with grayscale images. This algorithm was tested using a publicly available M-FISH database. Its output was compared with the outputs of conventional Otsu’s thresholding and watershed transform. Main fact that limited the application of this algorithm was the need of precise seed point selection to avoid inaccurate segmentation.

Shen et al.\textsuperscript{53} introduced a novel algorithm to segment di-centric chromosomes based on k-means clustering and the watershed transform. First, minimum circumscribed rectangle was used to identify the curvatures in a single chromosome area. When the ratio between chromosome area to the smallest rectangle was close to 1, the chromosome was straight. For extended type of chromosomes, the long side of the circumscribed rectangle was parallel to the centerline. To extract the centerline of the severely bent chromosomes, the method used by Zhang and Suen was modified.\textsuperscript{56} This modification improved the accuracy by removing the small holes created inside the chromosomes and spurious branches at the end of the sister chromatids.

Manohar et al.\textsuperscript{57} compared the performance of FCM and watershed transform in chromosome segmentation using publicly available M-FISH database. The overall accuracy was 94% and 92% for FCM and watershed transform respectively. However, the dataset of this approach included only 20 images which limited the reliability of the output.

Kanimozhi et al. introduced a novel techniques to segment the chromosome images using quasi-Newton-based K-means clustering which worked for M-FISH images. M-FISH image characterization was performed using expectation-maximization-based hierarchical Bayes model. In this approach, misclassifications were avoided by using contextual-based classification and region merging. The researchers modified the activation function of sigmoid and softmax layer of the AlexNet for the optimum classification of autosomal chromosomes and sex chromosomes. This algorithm outperformed the existing literature with a 6.96% higher accuracy.

Li et al. addressed the low accuracy of FCM in M-FISH images by incorporating spatial and spectral FCM. It provided the advantage of noise reduction by exploiting information from neighboring pixels. In addition, it had the ability to incorporate pixel information across different channels simultaneously. The results obtained from the proposed method were compared with other FCM-based methods using accuracy and false-positive detection. A higher segmentation accuracy and a lower false-positive ratio were obtained for this algorithm, proving that both spatial and spectral data can enhance the performance of FCM-based methods for chromosomal segmentation.\textsuperscript{58}

**Active contours-based chromosome segmentation**

Active contours work with concavities of the chromosomes and create closed curves around the chromosomes. Therefore, active contours are comparatively accurate in segmenting the chromosomes compared to other categories given in this review paper. Active contours have the ability to resolve touching chromosomes; however, considerable improvement in the performance has not been achieved for overlapping chromosomes.\textsuperscript{59}

Active contours can generate closed parametric curves around the chromosomes. However, they fail to extract chromosomes from the image when highly concave boundaries appear. Therefore, poor convergence property of active contours results in inaccurate segmentation specially when highly bent chromosomes present in the image.

Incorporating an external force model such as Gradient Vector Flow (GVF), resolves the aforementioned limitation of active contours. Discrete Cosine Transform (DCT) embedded into GVF is capable in providing a better description of energy levels of chromosomes in the image. Hence, DCT embedded GVF incorporated active contours output accurate segmentation.\textsuperscript{60}

However, active contours suffer from the need of optimum parameter setting. In addition, automated initialization is difficult when artifacts present in the image. Level sets is a promising solution in the cost of time efficiency when minimal artifacts present in the image.

The rest of this section presents a review on related work on active contour-based chromosome segmentation.

Arora\textsuperscript{1} introduced a method to disentangle the overlapping and touching chromosomes irrespective of the staining method. Objects in the image of metaspread chromosomes were extracted using a region-based active contour model incorporating level sets used by Minaee et al.\textsuperscript{43} This algorithm worked properly even when weak boundaries and intensity inhomogeneity presented. However, region-based level sets consume higher execution time and might be inaccurate.
when artifacts present in the image.\cite{12} After segmentation, objects were classified as single chromosomes, chromosome clusters or artifacts considering the area, circularity and length. To separate touching chromosomes from overlapping chromosomes, gradient paths were considered. Over-segmentation was avoided by using region merging. This algorithm outputs 96% and 81% of accuracy for touching chromosomes and overlapping chromosomes respectively.

Hu et al.\cite{42} proposed an algorithm to address the requirement of human intervention for segmentation when partially overlapping chromosomes presented in the image. It was developed using U-net artificial neural network based on active contours. In fact, the model was improved in terms of computational time and storage by reducing the layers and blocks of the network to a minimum. This model was trained using 64% of the data, validated using 16% of the data, and tested on the remaining 20% of the data. The results showed intersection over union accuracy, 94.7% and 88-94% for overlapped regions and non-overlapped regions respectively.

Hu et al.\cite{61} proposed snake active contour-based CNN for chromosome segmentation. Edge-based snake active contour was used for manual annotation in the training process. The seed points were manually provided near to the outline of the chromosomes during the training process. Advantages of using edge-based snake model for training include avoiding over-segmentation of chromosome boundaries and low computer complexity. The dataset of this approach includes 4,184 chromosomes. However, under-fitting and over-fitting were observed due to low resolution of the image, blurry boundaries and natural defects of the snake models.

**Chromosome image segmentation based on convex-concave points**

Human metaphase chromosomes can be segmented using curvature function which is built upon the concept of convexity and concavity.\cite{34,62} Centromere is the global minimum point of the image intensity profile\cite{63} which is presented in Figure 4. Concave-convex-based chromosome segmentation algorithms are capable in extracting both touching and overlapping chromosomes. They work with non-rigid chromosome boundaries. However, they often fail when chromosome clusters and highly-bent chromosomes appear in the image. Nevertheless, incorporating convex-concavity theorem to other segmentation categories results in comparatively accurate segmentation.\cite{1,28}

Summary of the selected methods which use the theory of convex-concave points is given in Table 1. Majority of the approaches uses concave-convex points in the algorithms to improve the performance.

In 2018, a new method was implemented by Maidan et al.\cite{65} to address the limitation of segmenting either touching chromosomes or overlapping chromosomes using Giemsa banding images. First, the stained microscopic images were

![Figure 4. Global minimum representation of a chromosome in a intensity profile.](image-url)
binarized. Then, the holes inside the chromosomes were filled using morphological operations to address one limitation of simple thresholding-based segmentation, i.e. creating holes inside the chromosomes when inner pixels appear in light intensities. Then, the contours of the chromosomes were approximated using deformation function. Interesting points, i.e. concave points (convex point is optional), were estimated using the curvature function. Finally, touching and overlapping chromosomes were segregated with 97.09% and 95% accuracy respectively based on interesting points and neighborhood connectivity property.

Table 2 presents a summary on the discussed human metaphase chromosome segmentation categories in this review paper including their strengths and limitations.

Incorporating learning architectures to automate the process of chromosome segmentation, is the current trend. Besides, reducing the blocks and the layers in order to minimize the computer complexity, consequently maximize the time efficiency and storage efficiency is required. Most of the segmentation algorithms are limited to one type of staining method which is a main limitation. In addition, it is important to consider the inter-phase cells present in the image during chromosome segmentation.

**Approaches towards automated feature extraction of chromosomes**

Accurate centromere detection is needed in many applications such as determining the radiation exposure of an individual. Centerline (medial axis) detection aids in detecting the centromere of the chromosome and consequently, the abnormalities in it. Accurate chromosome segmentation is required for precise chromosome feature extraction. There are two major approaches in chromosome feature extraction based on banding information and based on geometric features. Two sub-sections presented hereafter in this section discuss the approaches under the aforementioned categories.
Chromosome feature extraction based on banding information

Banding information of the chromosomes depends on the banding pattern visualized by the staining methods. It can be presented by a density profile as illustrated in Figure 5. These methods are restricted to a particular type of staining. Consequently, the use of these methods is limited.

Wang et al. proposed image enhancement algorithm based on a family of differential wavelet transforms to improve the contrast of the banding pattern of the chromosome images which is distorted due to cell culturing, staining and imaging conditions. However, the objective of this approach was to extract the geometric features including edges. Multi-scale point-wise product was used to improve the accuracy as geometric features correlate in different scales. This approach provided better classification results against three convolution-based methods; adaptive contrast stretch, adaptive contrast enhancement and contrast gain transform. In addition, time efficiency was improved.

Errington et al. proved that considering only the banding pattern results in less accuracy. This approach was tested using Copenhagen, Edinburgh and Philadelphia databases. Higher accuracy was achieved when normalized length and centromeric index were considered together with the banding pattern.

The majority of the chromosome classification algorithms work on straight chromosomes to improve the classification accuracy. Sharma et al. proposed an algorithm based on deep CNN and Residual Neural Networks (ResNet) assuming that the chromosomes were segmented and straightened. Banding patterns appeared in dark and light intensities in the chromosome images were identified using recurrent neural network based on sequence learning. This algorithm was tested using 5,256 chromosomes and compared with traditional deep CNN, ResNet-50, CRNN and attention-based sequence model. The proposed algorithm (Res-CRANN) outperformed the other modules in terms of accuracy and robustness.

Jindal et al. used the information on banding patterns to classify the chromosomes using Siamese Network with a margin value of 0.5. This algorithm surpassed the deep CNN based strong baseline algorithms.

Chromosome feature extraction based on geometric information

Geometrical features including length of the chromosomes, end points etc. are used in the majority of research work towards chromosome feature extraction.

Precisely, estimating the length of the chromosome; i.e. number of pixels along the centerline and, the location of the telomere, i.e. the edge of the chromatid are the two most common geometric features used in centromere detection. In addition, centerline detection is required to estimate the density profile.

One common method of centerline detection found in early literature is skeletonization. However, this method fails because it produces spurious branches at the telomere due to the morphological variations present in the chromosomes. Therefore, applicability of this method is limited. Arachchige et al. addressed this issue by introducing a hybrid approach.

Figure 5. Density profile along the medial axis of an individual chromosome.
algorithm using GVF active contours, Discrete Curve Evolution based skeleton pruning and, morphological thinning. This algorithm was tested using 120 DAPI-stained lymphocyte cell images. The proposed algorithm was compared with gold standard using Mean Absolute Distance and Maximum Absolute Distance. The algorithm was able to detect centromere in bent chromosomes accurately. This approach can be improved by testing it with various cell types and increasing the number of samples in the dataset.

Generally, polarization is performed after the classification. Poletti et al.35 introduced a method to orient the chromosomes before the classification regardless of the reading orientation. This method reduced feature variance among the chromosomes in the same class and resulted in accurate classification. This research team worked with Q-band images. Using a large dataset of 5,474 chromosomes was a merit of this research. This dataset included amniotic fluid and choroidal villi cells. Robust estimation of the medial axis of the chromosomes was performed in order to extract the features, namely, length of the chromosome, density profile and contour function. It was assumed that the chromosomes were already segmented by using their previous work presented in 2009.38 The results were compared with the manual classification performed by an expert. This system showed 94% accuracy.35 However, manual classification by only one expert was subjective. Therefore, collecting data of manual classification from a group of experts could have justify the results in a better way.

A study was conducted by Vaurijoux et al. to detect di-centric centromeres.78 The results showed 50% misclassification due to the presence of multiple centromeres in the object despite the fast execution of the algorithm. However, it showed 4.35% accuracy than the manual process.40,78

The US patent, US8605981B2, addressed the drawbacks of manual and semi-automated centromere detection techniques, such as false positive detection and errors occur due to non-rigid chromosome boundaries. This algorithm had the capability to detect at least one centromere in each metaphase di-centric chromosome. The centerline was accurately drawn after proper segmentation of metaphase chromosomes. The number of centromeres was counted by identifying the longitudinal cross sections with minimum width by calculating the number of pixels with the same intensity. Accordingly, the percentage of di-centric chromosomes in a population of cells was counted. This was used to calculate the radiation dose using a previously determined dose-response curve from a calibrated source.65

The same research group developed an algorithm with higher accuracy to detect di-centric metaphase chromosomes to determine the radiation exposure. The dataset included overlapped and overspread chromosomes in a cell population. First, the overspread chromosomes were excluded by separating the segmented image into two depending on the area of the detected metaphase chromosome regions. The method of separating single metaphase chromosomes from chromosome clusters introduced by Rizvandi et al., was used to identify chromosome clusters. In order to do that, the algorithm approximated the centerline and counted the co-joints within the blob. The number of co-joints classified the blob into single chromosome or chromosome cluster. Clusters with multiple chromosomes were prone to be detected as false di-centric metaphase chromosomes. Therefore, the same procedure followed by Rizvandi et al.79 was used to separate touching and overlapping chromosomes. Next, the chromosomes were extracted using GVF which was superior to traditional active contour models because it used non-conservative forces in contouring. GVF could detect the concavities, hence the bent chromosomes. The centerline of each long chromosome and short chromosome was detected previously. This method showed 87% accuracy for a data set of 1400 chromosomes. It was a significant improvement compared to their previous attempt which achieved 81% accuracy with 226 chromosomes. The candidate Based Centromere Confidence (CBSS) was used to measure the confidence in each centromere detection. This metric might be helpful in improving time efficiency. In addition, this algorithm worked for both DAPI and Giemsa.

In 2016, Subasinghe et al presented an advanced method that can handle large morphological variations to address the issue of artifacts occurring due to premature sister chromatid separation. Support Vector Machine was used to detect the telomere region. It accurately detected chromosomes with sister chromatid separation. This research work was presented mainly in three steps. First, the chromosomes were segmented and centerlines were extracted. Next, the contour was partitioned and chromosomes with sister chromatid separation were detected. The final step was centromere detection. This method showed 87% accuracy for a data set of 1400 chromosomes. It was a significant improvement compared to their previous attempt which achieved 81% accuracy with 226 chromosomes. The candidate Based Centromere Confidence (CBSS) was used to measure the confidence in each centromere detection. This metric might be helpful in improving time efficiency. In addition, this algorithm worked for both DAPI and Giemsa.

Jahani and Setarehdan73 implemented an algorithm to straighten the chromosomes with an ability to detect single chromosomes in the entire image. Median filter was utilized to reduce the noise in the chromosome images. Next, the
medial axis of the binarized image was extracted based on thinning. The medial axis was extended from both sides, assuming that the next 5 neighboring pixels had the same slope. A perpendicular line to the medial axis at each pixel was drawn. Finally, a vertical chromosome was obtained by making all the perpendicular lines lie parallel to each other. End-points of the medial axis were detected using morphological thinning and two masks. Then, a gray-level mask which varies the intensity level from 0 to 255 gradually, was used between the two end-points. According to the definition, the centromere is the narrowest area in the chromosomes. Accordingly, minimum total number of pixels having the same intensity was selected as the centromere.

Li et al.\textsuperscript{49} detected the centromeres of each isolated chromosome. Consequently, this research group was able to detect dicentric chromosomes. The maximum width of the chromosome was calculated by measuring point-wise inner distances. The areas that exceeded the pre-defined maximum width were removed as outliers. Centromeres were localized using a similar algorithm presented by Subasinghe et al.\textsuperscript{7} They implemented this algorithm using GVF, DCE, IIL, and SVM. Subsequently, di-centric centromeres were separated from monocentric chromosomes using a compound Machine Learning model. This model provided a solution for misclassifications occur when multiple centromeres present in one object. However, the accuracy achieved for Giemsa-stained images was comparatively low due to the presence of chromosomes with severe sister chromatid separation. In addition, staining debris resulted in over-segmentation. To address the aforementioned issues, conditional filters were introduced to the algorithm.\textsuperscript{49}

The introduced algorithms for chromosome extraction by Mohommed,\textsuperscript{31} Jahani and Setarehdan,\textsuperscript{73} and Li in 2016,\textsuperscript{49} require smooth boundaries of chromosomes. However, chromosome boundaries are irregular, so the accuracy of centerline and centromere detection using the aforementioned algorithms should be verified by using irregular boundaries as well.

Madian et al.\textsuperscript{63} proposed a method to detect centromere in extremely bent chromosomes. First, the contours of the chromosomes were identified using GVF in a binarized image. To find the threshold value, the intra-class variance was minimized while increasing the inter-class variance. Curvature function and weighted shortest path calculation were used to identify the centromere. Medial Axis Transform (MAT) provided better results for straight chromosomes. For both straight chromosomes and bent chromosomes, the projection vector performed with more than 90% accuracy in detecting the centromere. The proposed algorithm provided accurate results for highly bent chromosomes comparatively.

To summarize the section, centromere detection is essential for the diagnosis of genetic disorders. Telomere detection and centerline detection lead to accurate centromere detection. However, morphological variations and weak boundaries are challenges during the process. The use of improved algorithms incorporating GVF, has proven promising solutions in feature extraction.

Table 3 presents a summary on selected state-of-the-art approaches\textsuperscript{82} towards chromosome segmentation and feature extraction.

| Publication            | Method                                           |
|------------------------|--------------------------------------------------|
| Wang et al., 2021\textsuperscript{83} | Adaptive Receptive field Multi-Scale network (ARMS Net) |
| Mei et al., 2020\textsuperscript{68} | U-net & Conditional GAN                           |
| Bai et al., 2020\textsuperscript{84} | U-net & YOLOv3                                    |
| Al-Kharraz et al., 2020\textsuperscript{3} | YOLOv2 & VGG19                                   |
| Sale et al., 2019\textsuperscript{4} | U-net                                            |
| Qin et al., 2019\textsuperscript{85} | G-Net & L-Net                                    |
| Somasundaram et al., 2019\textsuperscript{5} | CNN                                           |
| Xie et al., 2019\textsuperscript{86} | ResNet-50                                        |
| Swati et al., 2018\textsuperscript{87} | Super-Xception network                           |
| Sharma & Vig, 2018\textsuperscript{75} | ResNet-50                                        |
| Zhang et al., 2018\textsuperscript{88} | CNN                                             |
| Jindal et al., 2017\textsuperscript{76} | Twin Siamese Network (CNN)                       |
Suggestions for future improvement
Most of the chromosome segmentation and feature extraction algorithms are restricted to a particular type of staining. However, to fulfill the clinical requirement, algorithms which work despite the method of staining is required.

Working with high resolution images in digital image processing always provides better results. However, the methods that provide better visualization of chromosomes may result in significant chromosomal damage, consequently inaccurate diagnosis. Therefore, visualization methods that provide images with high resolution with negligible chromosomal damage are required. Nevertheless, algorithms might be improved using novel image processing concepts to work with low resolution and distorted chromosome images.

One common limitation of the approaches towards chromosome segmentation and feature extraction is the undesirable smaller dataset. Hence, the reliability of these approaches remains low. Therefore, work with large dataset by using image augmentation, adding noise to the image and changing the scale is recommended to improve the diversity of the dataset.

State-of-the-art human chromosome segmentation and feature extraction algorithms incorporate learning architectures. However, time efficiency of these algorithms depends on the complexity of the chromosome image. Precisely, presence of large clusters with overlapping and touching chromosomes reduces the overall time efficiency. New approaches should focus on architectures with ‘parallel run’ to improve the time efficiency as current development of the computer processors supports complex algorithms. In addition, selecting the optimum number of parameters for the learning architecture can minimize the execution time. There are existing reliable models such as AlexNet and ResNet which can be adapted and improved.

The majority of the algorithms fail to disentangle both touching and overlapping chromosomes. In addition, some algorithms are incapable in working with bent chromosomes and di-centric chromosomes. However, algorithms which satisfy the clinical requirements should work with all of these types of chromosomes. Therefore, common approaches towards detecting touching, overlapping, bent and di-centric chromosomes are required. In addition, karyotyping is recommended for the suspicious cases of genetic disorders. Hence, translocations may present in the image which affect the banding pattern. Intelligent algorithms can be developed to address such cases. In addition, active learning may minimize the crowd dependency.

Data availability
No data are associated with this article.

References
1. Arora T, Dhir R: A novel approach for segmentation of human metaphase chromosome images using region based active contours. Int. Arab J. Inf. Technol. 2019; 16(1): 132–137.
2. Nirmala Madian KB, Jayanthi, Suresh S: Contour based segmentation of chromosomes in g-band metaphase images. 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2015, pages 943–947. IEEE.
3. Al-Kharraz MS, Elrefaey LA, Fadel MA: Automated system for chromosome karyotyping to recognize the most common numerical abnormalities using deep learning. IEEE Access. 2020; 8: 157727–157747. Publisher Full Text
4. Saleh HM, Saad NH, Isa NAM: Overlapping chromosome segmentation using u-net: Convolutional networks with test time augmentation. Procedia Comput. Sci. 2019; 159: 524–533. Publisher Full Text
5. Somasundaram D: Machine learning approach for homolog chromosome classification. Int. J. Imaging Syst. Technol. 2019; 29(2): 161–167. Publisher Full Text
6. Sharma M, Saha O, Sinraman A, et al.: Crowdsourcing for chromosome segmentation and deep classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops. 2017; pages 34–41.
7. Subasinghe A, Samarabandu, Liy, et al.: Centromere detection of human metaphase chromosome images using a candidate based method. F1000Res. 2016; 5(1565): 1565. Publisher Full Text
8. Murakami Y: Heterochromatin and euchromatin. Encyclopedia of Systems Biology. Dubitzky W, Wolkenhauer O, Yokota H, et al., editors. 2013; pages 881–884.
9. Speicher MR, Carter NP: The new cytogenetics: blurring the boundaries with molecular biology. Nat. Rev. Genet. 2005; 6(10): 782–792. PubMed Abstract | PUBMED | Publisher Full Text
10. Ana Katrina C, Estandarte: A review of the different staining techniques for human metaphase chromosomes. Department of Chemistry, University College London, University of London; 2012.
11. Dolan M: The role of the giemsa stain in cytogenetics. Biotech. Histochem. 2011; 86(2): 94–97. PubMed Abstract | PUBMED | Publisher Full Text
12. Yilmaz IC, Yang J, Altinsoy E, et al.: An improved segmentation for raw g-band chromosome images. 2018 5th International Conference on Systems and Informatics (ICSAI). 2018; pages 944–950. IEEE.
13. Shaffer LG, McGowan-Jordan J, Schmid M, et al.: ISCN 2013: an international system for human cytogenetic nomenclature (2013). Karger Medical and Scientific Publishers; 2013.
14. Dzvanko A, El Sayed A, Meedovic S: A comprehensive review of the multidisciplinarity in karyotyping. Heal. Technol. 2020; 10(1): 101–109. Publisher Full Text
This research paper presents human chromosome segmentation algorithms under four broad categories: thresholding, clustering, active contours and convex-concave points-based methods. Chromosome feature extraction methods are discussed under two main categories based on banding-pattern and geometry. The research paper presents new insights for the improvement of fully automated karyotyping are provided.

A few suggestions/comments to further enhance the quality of the manuscript:

1. The reviews of the previous work presented in the manuscript presents are fairly brief, and they focus on the outcomes and limitations rather than the algorithms used. Some deeper technical discussion of the previous work most related to this study may be included.

2. The comparison tables in the manuscript list the methods used. It is imperative to present an exhaustive review comparing the advantages / limitations / weakness / strengths of the methods.

3. Various publicly available datasets (LK1, BioLab etc) with the varying complexities in segmentation and subsequent feature extraction may be discussed to highlight the choices of the algorithms, research challenges addressed, contribution of the researches.

4. Various performance parameters to measure the success of the segmentation algorithms are popularly reported. Please consider including a discussion on the performance of the reported algorithms in terms of the various performance parameters.

5. Few researchers have successfully addressed extrication of touching and overlapping chromosomes (up to seven overlaps) using novel approaches. Please include a discussion on the same in the manuscript. See Munot, Jayanta & Joshi (2013).
6. Manuscript mentions use of pre-trained network for chromosome segmentation. A few recent references and their findings may also be considered in the manuscript, such as Sun et al. (2021), and Kanimozhi, Balasubramani, & Anuradha (2021).[ref-1-2]

7. Role of deep learning approaches and other aspects or considerations like Transfer learning, regularization hyper parameter tuning may be considered.

8. Literature reports use of background correction and other pre-processing techniques to enhance the segmentation accuracies. MFISH. A discussion of the similar approaches may be incorporated.

9. The discussion section must be dedicated to list and summarize crucial finding of the devised approach, the strengths of the approach, the critical cases where the reported algorithms failed or outperformed, research gap addressed, the major considerations from the perspective of practical deployment and impact of clinical environment.

10. Manuscript needs to be written scientifically and must be presented aptly. Manuscript needs numerous grammatical corrections. Additionally, clarity in the presentation, correct use of English language and improved flow of explanation will be helpful to improve its quality

References
1. Munot MV, Mukherjee J, Joshi M: A novel approach for efficient extrication of overlapping chromosomes in automated karyotyping. *Med Biol Eng Comput*. 2013; 51 (12): 1325-38 PubMed Abstract | Publisher Full Text
2. Sun X, Li J, Ma J, Xu H, et al.: Segmentation of overlapping chromosome images using U-Net with improved dilated convolutions. *Journal of Intelligent & Fuzzy Systems*. 2021; 40 (3): 5653-5668 Publisher Full Text
3. Kanimozhi VS, Balasubramani M, Anuradha R: Hierarchal Bayes model with AlexNet for characterization of M-FISH chromosome images. *Med Biol Eng Comput*. 2021; 59 (7-8): 1529-1544 PubMed Abstract | Publisher Full Text

Is the topic of the review discussed comprehensively in the context of the current literature?  
Yes

Are all factual statements correct and adequately supported by citations?  
Yes

Is the review written in accessible language?  
Partly

Are the conclusions drawn appropriate in the context of the current research literature?  
Yes
Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bio medical imaging, AI, ML, Deep Learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com