

Prediction of Centromere Location in Human Chromosome Using Convolutional Neural Networks

Ajdin Vatreš¹, Naris Pojskić², Edin Kadrić¹

¹ *University of Sarajevo - Faculty of Mechanical Engineering, Vilsonovo setaliste 9, Sarajevo, Bosnia and Herzegovina*

² *University of Sarajevo – Institute for Genetic Engineering and Biotechnology, Zmaja od Bosne 8, Sarajevo, Bosnia and Herzegovina*

Abstract – Accurate determination of chromosome centromere location is of high importance in cytogenetics, particularly in karyotyping, chromosome classification and determination of exposure to genotoxic environmental effects. This study investigates the ability of CNN to accurately predict the human chromosome centromere location and the effect centering chromosomes in images, by predicted centromere location, has on classification accuracy. Dataset, used to train and test CNN models, contained 8283 annotated individual chromosome images. Prior to performing centromere detection, followed by chromosome classification, the individual chromosome images are preprocessed using sequence of filtering algorithms. The CNN model achieved an average error of 0.5586 and 0.4543 in predicting x and y coordinates of centromere location, respectively. The achieved classification accuracy of randomly oriented and centered chromosomes in images, is 71.10 and 96.73%, respectively. Achieved increase in chromosome classification accuracy of 25.63% highlights importance of chromosome centromere detection, importance of positional variation removal, and high performance of CNN in prediction of centromere location and chromosome classification.

Keywords – Convolutional network, chromosome centromere detection, chromosome classification, image filtering.

1. Introduction

Cytogenetics is a branch of biology created by connecting the sciences of cytology and genetics. The research focus of cytogenetics is on those structures that are in direct relationship with heredity using the methods of cytology and genetics.

As part of cytogenetics, chromosomes, their morphology, structure, number, location of genes, pathology, as well as their function as carriers of hereditary factors are investigated [1]. In addition, cytogenetics studies the behavior of chromosomes during cell division (mitosis and meiosis), as well as the factors that influence changes in chromosomes [2].

Although a number of different techniques have been developed over the years, as part of conventional (routine) cytogenetics, the G-banding technique using trypsin and Giemsa has become one of the most widely accepted in the field. The pattern of stripes created on the chromosome using this technique enables detection of various microscopically visible aberrations, such as translocations, inversions, deletions, and duplications [2].

The centromere represents a special region within the chromosome structure composed of specialized chromatin. The centromere divides the chromosomes into two arms, the short p and the long q arm. In chromosome images the centromere appears as a narrow (constricted) region, hence it is described as the primary constriction of a chromosome. The centromere performs a very important function during cell division. Namely, the centromere is the location of the attachment of sister chromatids and represents the origin for the formation of the kinetochore of the dividing spindle [3].

DOI: 10.18421/TEM123-02

<https://doi.org/10.18421/TEM123-02>


Corresponding author: Ajdin Vatreš,
University of Sarajevo - Faculty of Mechanical Engineering, Vilsonovo setaliste 9, Sarajevo, Bosnia and Herzegovina
Email: ajdin.vatres@mef.unsa.ba

Received: 25 April 2023.

Revised: 13 June 2023.

Accepted: 24 July 2023.

Published: 28 August 2023.

 © 2023 Ajdin Vatreš, Naris Pojskić & Edin Kadrić; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

The appearance and size of this constriction can vary drastically depending on the species in question, but centromeres can be grouped into point, regional, and holocentric. Human centromeres, like centromeres of other primates, are relatively large regional centromeres, and there is only one region of this type on chromosomes, as long as cells and/or chromosomes are in a normal state [4], [5]. As such, they can be recognized on images of human chromosomes as a specific visual location, that is, a specific region within the image. In addition to its biological function, the specificity of the centromere is very useful in chromosome classification, as certain types of chromosomes are characterized by a different position of the centromere. Contrary to the suggestion given by its name, the centromere is located in the center of only some chromosomes, the so-called metacentric chromosome group. In human chromosomes, metacentric, submetacentric, and acrocentric chromosomes can thus be distinguished [4], [5]. Figure 1 shows schematically the differences between the types of these chromosomes, together with the indicated basic parts of the chromosome. Metacentric chromosomes are characterized by the centromere, which is located almost in the middle of the chromosome and divides it into two approximately equal parts, the short arm is the same length as the long arm, $p=q$. Submetacentric chromosomes are characterized by a centromere located submedially. As a consequence, the upper arms are shorter than the lower ones ($p<q$). Acrocentric chromosomes are characterized by the centromere, which is located at the very end of the chromatid, so the short arms are hardly noticeable, while the long arms of the chromosome are well expressed ($p\ll q$). Acrocentric chromosomes often contain structures called satellites, in the form of small appendages separated from the chromosome by a stalk [6].

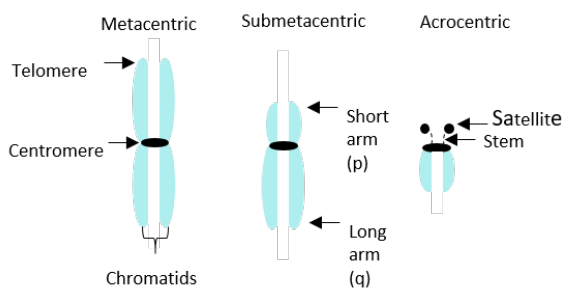


Figure 1. Differences between chromosome types [7]

According to the previously presented data, it can be concluded that the automatic detection of a chromosome centromere location would enable easier and more accurate classification of chromosomes.

Centromere location, as well as the lengths of the p and q arms, represent useful classification features. Using only these features human chromosomes can be grouped into seven subgroups of the Denver system, and when combined with chromosome band patterns it is possible to determine all 24 classes. Besides the possibility for use as a feature, the detected centromere location can be used to center chromosome images in a meaningful way. This should make classification using convolutional network easier, as it removes large amount of unnecessary variation, carrying no information about chromosome class. In addition, the predicted location could serve as a feature in analysis of various chromosomal abnormalities presence.

This study introduces a model that employs convolutional neural networks to automatically detect the location of centromeres. We demonstrate how this approach enhances the accuracy of chromosome classification. The structure of the paper is as follows: in section 2. previous research, the current state of the literature in this field is presented, in section 3. methodology, information about the used data set is given, as well as the steps needed for image preprocessing, and the convolutional neural network used for centromere location detection. Further in section 4. results and discussion, an analysis and discussion of the estimated centromere location with performance measures is presented. Finally, under conclusion section, the main findings of this study are discussed.

2. Previous Research

The chromosome centromere location is estimated in [8] based on the chromosome width profile, defined by its edges and projected onto the medial axis. As chromosomes are highly bent structures, a skeletonization algorithm is used in combination with fourth-degree polynomial interpolation to detect the medial axis. Since the polynomial line is less sensitive to sporadic changes, this approach enables the reduction of noise influence on the quality of the detected axis. The condition that chromosomes have the narrowest width in the location of their centromere, is not sufficient for accurate detection of centromere location. Hence, an additional condition is introduced, namely, at each point representing the potential location of the centromere, the edges of the chromosomes must be concave.

The skeletonization algorithm is used to determine the longitudinal axis of chromosomes in [9]. Based on the determined medial axis, the chromosomes are rotated in the same direction. Prior to the application of the skeletonization algorithm, the contrast of the chromosome images is adjusted using adaptive histogram equalization.

The images are also converted from RGB format to binary using the standard luminance method and a threshold value determined for the used data set. The size of the binary chromosome surface area is used to divide the chromosomes into two groups: the first representing the metacentric and submetacentric chromosomes, and the second the acrocentric ones. Finally, the location of the centromere is determined as the location of the global minimum of the binary image histogram, transformed with its Gray Level Mask (GLM). The division into groups enabled two different histogram constructions procedures. Histogram for the first group is created using 60%, and for the second group, 70% of total image surface. Surface areas are measured from the image center in both groups.

Prior to performing centromere detection, the individual chromosome images undergo a conversion process from grayscale to binary format in [10]. This transformation is performed by the Otsu algorithm. The obtained binary image is then used to determine the chromosome contour, which is then transformed into a smooth contour using the Gradient Vector Flow (GVF) algorithm. The newly created contour is then divided into two approximately symmetrical parts by a longitudinal line, created by the Discrete Curve Evolution (DCE) algorithm. DCE evolved the original complex curve into the simplest polygon through the removal of the least important vertices. The authors empirically showed that 6 vertex points provide satisfactory results. The resulting medial lines are shortened at the ends by 10%, in order to avoid the influence of possible bifurcations. The thickness profile is generated by an algorithm based on Laplace's equation. Chromosome thickness profile is used to detect the centromere location. Solving the fictitious heat flow equation from one contour partition to another, with grayscale values as weighting factors, the chromosome profile thickness lines are created. All locations with minimum thickness are used as inputs to a Support Vector Machine (SVM) classifier to select the best point among candidates.

Centromere location detection is performed in [11] using the Bending Potential Ratio (BPR) algorithm, introduced in [12]. This algorithm is based on the determination of the chromosome skeleton using a measure of importance called BPR that takes into account the contour segment to which a certain branch of the potential skeleton belongs. BPR uses global and local properties of the observed contour to remove irrelevant branches of the skeleton. The obtained chromosome skeleton represents a set of all potential centromere positions (one or two). The distance between the points designated as potential centromeres is used as a boundary for deciding whether there are one or two centromeres.

This distance is required to be greater than 10% of the average length of all chromosomes in the cell. Besides this requirement, the authors propose use of an additional requirement. The ratio of the Euclidean distance of the closest points on the contour around the first potential centromere, to the second potential centromere needs to be greater than 1.05.

The centromere location is determined in [13] based on the selection of points from a set of all points that belong to the medial (central) line of the chromosome. Two methods are used to determine the central line of the chromosome. Which method is used depends on the ratio of the chromosome surface and the surface of the bounding box. For long and straight chromosomes, this ratio is close to 1, and the line parallel to the long side of the bounding box is used as their medial axis. If the chromosome has a ratio below the threshold value, the algorithm presented in [14] is used to determine the skeleton. All determined points of the skeletons are considered as potential centromere locations. In this study a measure called Equivalent Width (EV) is defined. EV represents the product of the inverted gray color intensity at a particular skeleton location and the Euclidean distance between two points positioned at the ends of a vertical line drawn from that same skeleton point. EV is used to reduce noise influence on the measured chromosome width. Finally, if the difference between the value of the trend and the value of the EV curve itself, is within some empirically determined limits, the point is considered as the centromere location.

From the previously presented, it can be concluded that the methods for centromere detection include two phases. Firstly, medial (central) line of the chromosome is determined, and secondly, centromere is determined as one point among all medial axis points. The main weakness in the presented approaches is the estimation of the medial axis. Since chromosomes are highly bent structures they do not have a center line that is easily or uniquely determinable. This problem is even more complex if we take into account the noise influence, as well as the low contrast found to be present in the reviewed studies. In addition, some type of measure is always used to determine whether some particular point, from the set of potential centromere locations, is actually the centromere. This measure is usually a modified value of the chromosome width profile. Prevalence of this idea stems from the observation that the centromere represents the primary constriction of the chromosome. However, due to the morphological changes that are present in the chromosomes, as well as the errors, generated in the process of measuring the chromosome width profile, the location of the centromere does not necessarily represent the global minimum of the width profile.

To avoid these shortcomings, in this study a method for detection of useful chromosome features, based on a convolutional neural network model, is presented. The developed neural network is able to detect the pattern of stripes present in the proximity of the centromere location, thus overcoming the shortcomings observed in previously presented approaches. The idea behind this approach is in accordance with the characteristics of the centromere as a region with a special chromatin structure which has different appearance for different chromosome classes.

3. Methodology

In this section we provide an overview of used methodology in this study. Firstly, we present human chromosome image dataset used for training and testing of CNN models, with detailed description of procedures for image acquisition, annotation, resizing, and removal of inadequate images. Since the images contained substantial noise and other impurities that could affect the accuracy of centromere location detection and classification, additional filtering of images was required. We applied sequence of filtering algorithms, namely the Bilateral, CLAHE, and Rolling Ball filters. Following the filtering procedure, we present two CNN models, based on ResNet50, used for prediction of centromere location and chromosome classification. For each CNN model, we present structure, as well as settings for training and testing, and metrics for performance assessment.

3.1. Preparation of the Data set

A convolutional network model for centromere location detection is developed and tested on a dataset consisting of individual human chromosome images. These images are created by extracting images of individual chromosomes from a total of 244 images of the complete metaphases. Among these metaphase cell images, 126 belong to female and 118 to male subjects. The chromosomes in the metaphase cells are distinguished by approximately 400 noticeable stripes (this number refers to the sum of all stripes present in the chromosomes of one metaphase cell). Stripes, created by staining using the G-banding technique, describe the structure of individual chromosomes and represent one of the key features for chromosome classification, as well as centromere detection. All images are saved in ".jpg" format.

As a result of large differences in the preparation of individual slides, as well as in the imaging procedure itself, certain metaphase images could not be used for centromere detection, and are completely removed from the data set.

In addition, certain individual chromosomes are not distinguishable by the necessary characteristics (stripes on the chromosome, changed morphology, etc.) and could not be adequately used for cytogenetic analysis. Such chromosomes are also considered unsuitable for model training, so they are removed from the data set. The cytogeneticist indicated metaphases, as well as individual chromosome images that are not suitable for analysis. Filtering out of inadequate images, from the initial data set, reduced the final count of individual images to 8283. The cytogeneticists annotated all individual chromosome images into one of 24 unique classes, and also determined the location of the centromere for each chromosome image. All images are then scaled to a uniform dimension of 299x299 pixels. The final image data set is then divided in such a manner that 80% of all images are used for model training, and 10% for model validation. The final 10% of the image data set is separated and is never used during the training phase. These images are exclusively used for model testing and performance evaluation.

3.2. Image Filtering

As a consequence of chromosome structure and their biological nature as bent rod-like structures that move freely within space, as well as the process of preparation of slides for microscopy and imaging, the images of metaphase chromosomes contain a number of poor image characteristics. These characteristics include low contrast present in a large number of images, uneven background lighting within the images, and a large amount of noise. In order to reduce these unwanted effects, an image filtering procedure is used, consisting of successive application of three different filters. First, a bilateral filter is used, then Contrast Limited Adaptive Histogram Equalization, i.e., the CLAHE filter, and finally the Rolling Ball filter.

3.2.1. Bilateral Filter

As certain amount of noise is present in all images, it is necessary to reduce it by filtering. The image filtering process is a procedure in which random components (errors), that have changed the values of individual pixels, are removed from the real image (here, the real image is an imaginary ideal image in which there are no random errors). These random components are not information carriers of the observed phenomenon. The centromere location is partially described by the narrowing of the chromosome cross-section, i.e. narrowing of its width in the region of the centromere.

Hence, the denoising process should be executed in a manner that ensures the preservation of chromosome edge integrity, quality, and visibility. Bilateral filtering is an upgrade of the low-pass filter. This simple and fast procedure is non-iterative, and it gives significant focus to preserving the character of the edges present in the filtered image. The basic foundations of the bilateral filter are presented in [15].

Bilateral filtering is a procedure of non-linear weighted averaging of pixel values located within a local region. The weights of this filter depend exclusively on two parameters, spatial distance, as an information carrier about the size of features to be preserved, and intensity distance, as an information carrier about the contrast of features to be preserved. The intensity of bilateral filtering weights is always determined relative to the central pixel of the observed local region [15], [16].

Figure 2 shows the results of individual chromosome image filtering using a bilateral filter with different local region size. From Figure 2 it is possible to notice that, in all cases, the edges of the chromosomes are simultaneously well preserved and smoothed. An area of 25 is selected to prevent blurring of the image, as for larger values blurring starts to occur.

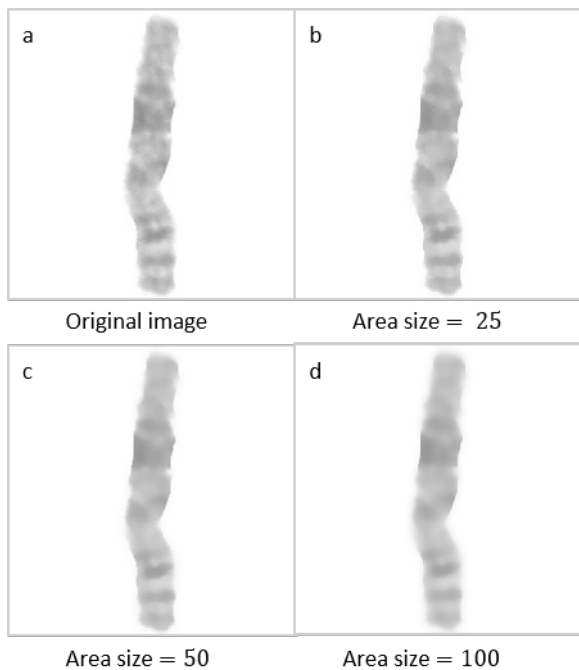


Figure 2. Chromosome images after bilateral filtering

3.2.2. CLAHE

Large differences in contrast are observed between individual chromosome images, both within the same and between images originating from different metaphase cells.

Since the position of the centromere is not exclusively described by the narrowing of the cross-section, but also by the content and arrangement of the stripes around that location, it is necessary to adjust the contrast in such a way that the stripes are uniformly visible within all images. For this purpose, the CLAHE algorithm is used.

The CLAHE algorithm is a modification of adaptive histogram equalization (AHE) introduced in [17], [18]. Although AHE has proven to be an effective algorithm, it has a tendency to cause strong amplification of noise present in images, especially in local regions with relatively uniform pixel values [19].

The prevention of unwanted noise increase is achieved by reducing the size of the histogram. Namely, in methods of this type, the height of the histogram column for a given pixel intensity indicates the strength of the contrast enhancement. Thus, it is sufficient to limit the size of the histogram columns in order to limit the level of contrast enhancement. This is achieved by simply uniformly shifting pixels to lower columns [19].

As adaptive histogram equalization requires that the histogram is determined not for the entire image but for a local region around the observed pixel, in all cases a local region of size 8x8 pixels is selected. Figure 3 shows the results of contrast enhancement using the CLAHE algorithm with different values of the contrast enhancement limitation level. From Figure 3, it can be seen that even for small values of the limitation of the contrast adjustment, there is a good expression of the important stripes of the chromosome. However, for higher values of the contrast adjustment, there is an increase in noise around the edges of the chromosomes. Accordingly, a contrast limitation level of 4 is selected.

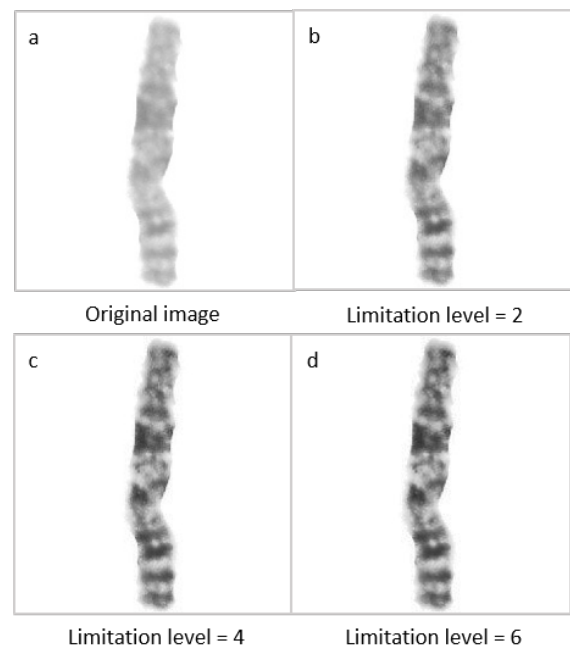


Figure 3. Chromosome images after CLAHE

3.2.3. Rolling Ball

During photography of cells in metaphase, the position and intensity of the microscope light can lead to uneven levels of background illumination in different locations of the same image. Unequal illumination of the image background can have an impact on the intensity level of individual stripes, which are of high importance in centromere location detection.

This unwanted effect can be reduced by specifying a unique uniform background illumination level for each image, which then changes the local pixel value minima and maxima. Determination of the uniform background illumination value, and the removal of the initial one, is in this study achieved by the use of the Rolling Ball algorithm, whose settings are first introduced in [20].

The idea behind the rolling ball algorithm is based on the morphological transformation that occurs when a certain shape, in this study a ball of radius r , slides along an uneven surface. Depending on the size of the selected element, i.e., the sphere (ball) with radius r , certain points of the original surface come into contact with the upper surface of the element. The larger the element that performs the transformation, the greater the number of contact points, and vice versa. When r is small, the sphere enters all the "valise" that represent the local minima and maxima of the original surface, while when r is large, the sphere passes them without contact, describing a smooth surface. Using the presented transformation, a new uniform level of background illumination is generated. The shape of the new uniform illumination surface is determined by the size of r .

Figure 4 a-c shows results of background illumination removal using the Rolling Ball algorithm for different r values, while Figure 4 d-f shows the detected background values.

From Figure 4 it can be seen that with small values of the radius $r \leq 5$, the surface of the sphere touches almost all chromosome points, thus recognizing the complete structure of stripes as background. On the other hand, for values of $r > 20$, the sphere does not enter the "valise", described by local minima and maxima, and is no longer able to detect differences in the background illumination. Finally, a radius value of $r = 20$ is selected as the most suitable, because it balances uneven background illumination removal with chromosome feature preservation.

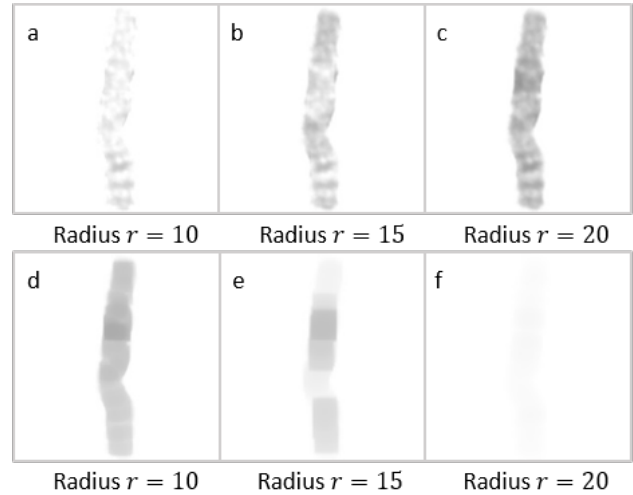


Figure 4. Effects of the Rolling Ball algorithm on Chromosome images

3.3. Convolutional Network for Centromere Location Detection

The convolutional network model used for centromere location prediction consists of a base convolutional network model pre-trained on the image classification task. Figure 5 shows the structure of the convolutional network used for centromere location prediction.

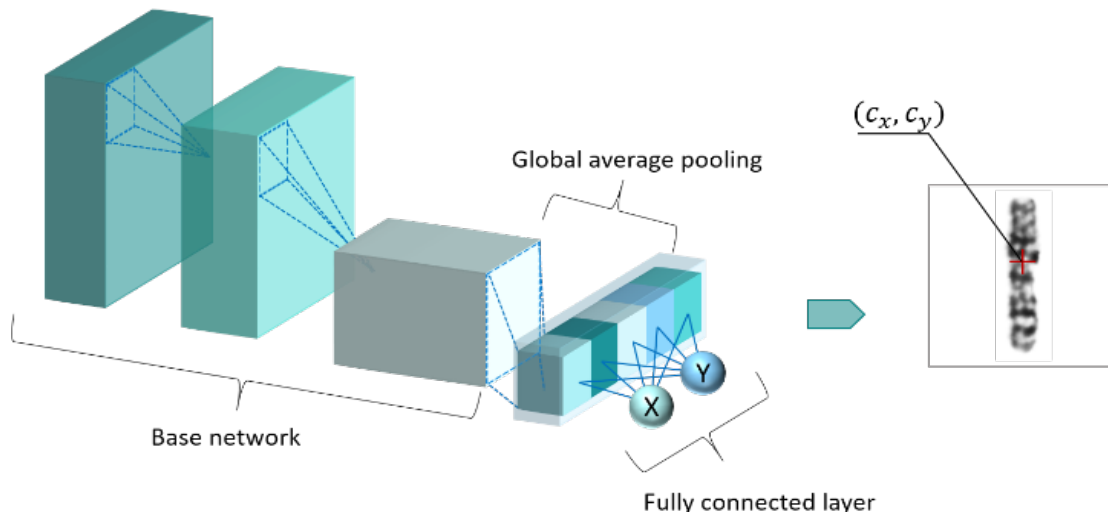


Figure 5. Convolutional network for centromere location prediction

In this model, the final Fully Connected layers are removed and replaced with one layer for Global Average Pooling and finally a Fully Connected layer. Fully Connected layer contains two neurons, without nonlinear activation functions, used to predict the coordinates of the centromere location. The neural network ResNet50, defined in [21], previously trained on the ImageNet data set [22], is used as the base model.

The Adam optimizer defined in [23] is used to determine the network weights. The learning rate is set to 10^{-2} with a reduction factor of 10^{-1} , every time two consecutive epochs resulted in no reduction of the cost function during validation. This procedure is continued until the minimum allowed learning rate value of 10^{-4} is reached. Finally, after 5 consecutive epochs, without reducing the value of the validation cost function, the training procedure is stopped.

As the position of the centromere is defined by two coordinates (x, y) in the 2D space of the chromosome image, it is natural to consider the deviation vector of the predicted point from the actual one, annotated by the cytogeneticist, as a performance measure. Therefore, the Mean Squared Error (MSE) is used as the cost function. MSE can be derived directly from the Euclidean distance, thus it has a sensible graphical interpretation as a cost function.

Let the vector of n predicted values, based on n samples, be denoted by $\hat{\mathbf{C}} = \begin{bmatrix} \hat{c}_x \\ \hat{c}_y \end{bmatrix}$, and let the vector of observed actual centromere coordinate values be denoted by $\mathbf{C} = \begin{bmatrix} c_x \\ c_y \end{bmatrix}$, where c_x represents its x and c_y its y coordinate, then the cost function is defined by the following equation:

$$MSE = \frac{1}{n} \sum_{i=1}^n (C_i - \hat{C}_i)^2 \quad (1)$$

Besides the MSE, the Root Mean Square Error (RMSE) is used as a performance measure. RMSE is defined by the following equation:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (C_i - \hat{C}_i)^2} \quad (2)$$

In the context of the distance between two points, it is intuitive to view the error in the form of absolute deviation. Therefore, an additional error function is used in the form of Mean Absolute Error (MAE), which is defined by equation:

$$MAE = \frac{1}{n} \sum_{i=1}^n |C_i - \hat{C}_i| \quad (3)$$

3.4. Chromosome Classification

The neural network, used for centromere detection is modified and used to estimate the effect of chromosome centering, by predicted centromere location, on classification accuracy. The final fully connected layer is removed and replaced with a fully connected layer with 24 neurons, one for each chromosome class.

The neural network is trained using the Adam optimizer with a learning rate of 10^{-2} . The learning rate is reduced by a factor of 10^{-1} , up to a minimum value of 10^{-5} , if the validation cost function is not decreased after two consecutive epochs. If the cost function during validation is not reduced after 10 consecutive epochs, the training procedure is stopped.

The network is trained 5 times using different folds of the data set. First, the network is trained on images with chromosomes in their initial random spatial orientation, and then on images with chromosomes centered by their predicted centromere location.

For image classification tasks it is natural to represent the results of the network as a probability distribution describing how likely an image belongs to a particular class. Therefore, the cross entropy denoted by $H(p, q)$, of the Kullback–Leibler divergence is used as a cost function.

Let p be the distribution of observed values and q the distribution of the model predictions then the cost functions $H(p, q)$ is defined by the following equation:

$$H(p, q) = \sum p(n) \log q(n) \quad (4)$$

4. Results and Discussion

In this section the performance of the convolutional network for chromosome centromere location detection is examined. The effect of centering chromosome images based on their centromere location as opposed to using random spatial positioning on classification accuracy of a standard, well known model (ResNet50), is also given.

The training of the models and experimental analyses are performed on a windows laptop with a AMD Ryzen 7 5800H 3.2Ghz 8 core CPU and a 6 GB NVIDIA GeForce RTX 3060 Laptop GPU.

The centromere location detection convolutional neural network model converged to the final weight values after 38 epochs with achieved minimum validation MSE value of 6.3187, and minimum validation MAE value of 1.7632. Graphical presentation of MSE and MAE during training and validation phases of the neural network are shown in Figure 6.

From Figure 6 it can be seen that the values of the cost function of training and validation converge towards the final difference between them, and that there is no increase in the validation cost at any moment. The training cost function is asymptotically approaching zero. Hence, it can be concluded that the model converged to adequate final values of its weights.

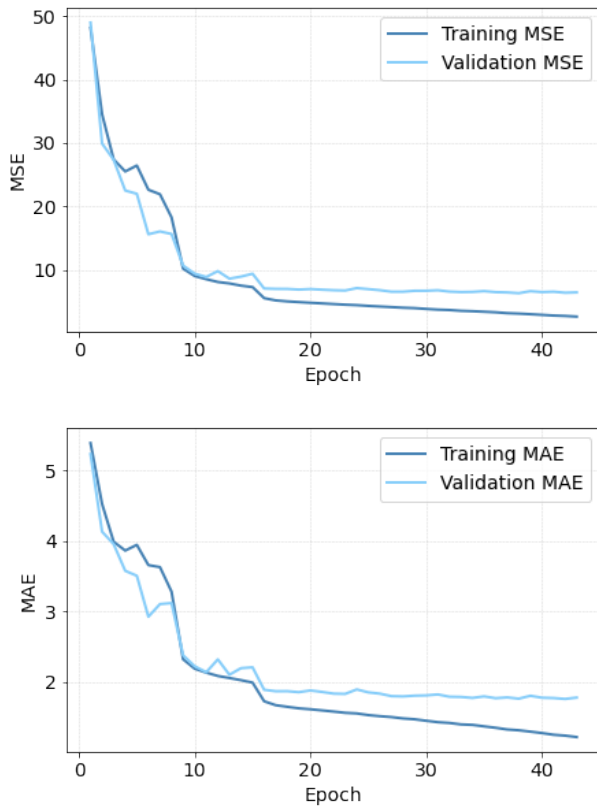


Figure 6. MSE and MAE during model training and validation

After training, the model is tested on an unseen data set, achieving values of 6.6638 for MSE and 1.8357 for MAE. These values are very close to those obtained on the validation set, indicating good generalization ability of the model on unseen test data.

As the centromere location is defined by two coordinates, it is important to test whether the model adequately predicts both the x and y coordinates. The model achieved an average error between the actual x coordinate and the predicted one of 0.5586, with a standard deviation of 0.7091. The average error between the actual y coordinate and the predicted one is 0.4543, with a standard deviation of 0.3346. The maximum achieved prediction error of the x coordinate is 9.7774, and 2.3287 of the y.

Figure 7 shows the box plot of x and y coordinates prediction errors. From Figure 7 it can be seen that the model slightly better predicts centromere location y coordinate than x.

However, it is necessary to keep in mind that the differences refer to pixels. This means that the model never made an error greater than 10 pixels, while on average it is less than 1 pixel.

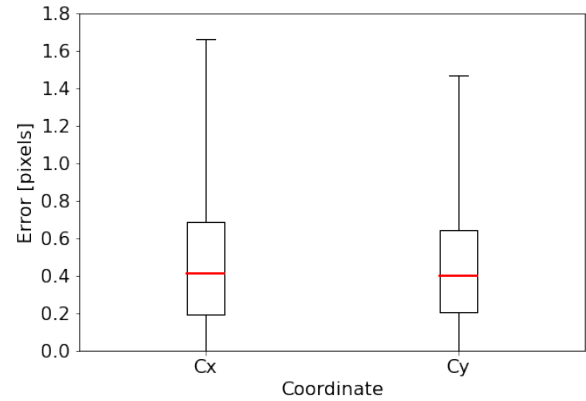


Figure 7. Box plot of centromere location x and y coordinate prediction error

Since different classes of chromosomes are morphologically distinct, both in their size and shape, it is necessary to test the performance of the model on individual classes. Figure 8 shows the values of RMSE and MAE for all individual chromosome classes.

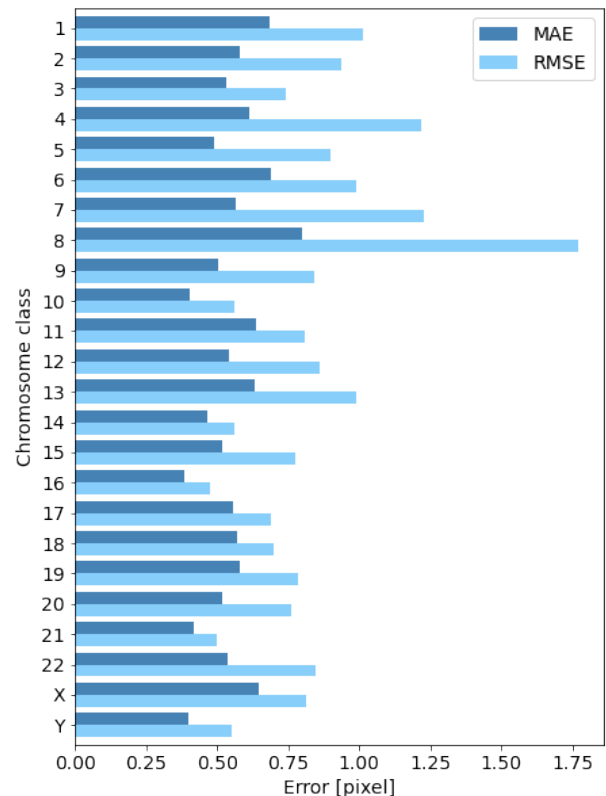


Figure 8. RMSE and MAE for each class

From Figure 8 It can be seen that model does not predict centromere location of all chromosomes with the same accuracy. Slightly less accurate predictions are achieved on classes 8, 7, 4 and 1.

However, these differences are only noticeable in relative comparison of prediction accuracy for different classes. In absolute terms, model mean absolute errors are smaller than 1 pixel, and for most classes, error is around half of a pixel.

Figure 9 shows the chromosomes with the predicted location of the centromere indicated in red, and the actual location in blue.

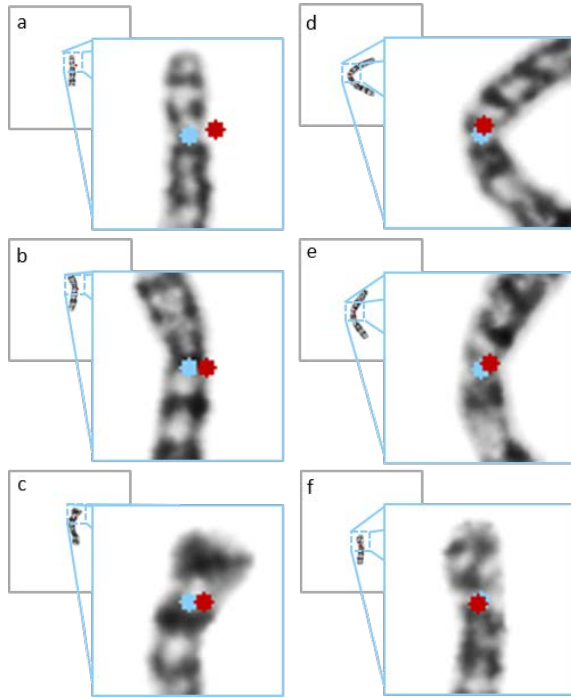


Figure 9. Predicted (red) vs actual (blue) centromere location

Figure 9 a-c shows three chromosomes with the worst predicted x coordinate, while Figure 9 d-f shows chromosomes with the worst prediction of the y coordinate. From Figure 9, it can be inferred that the large difference in the length and width of the chromosome is one of the causes for the poorer prediction of the x coordinate. However, it can be seen from Figure 9 that even the worst positioned chromosomes, the ones that can even be considered outliers, provide a satisfactory estimate of the actual location.

The classification accuracy and cost function is measured on five different folds of the data set, both for images in their random position and for images centered using the predicted centromere location. Table 1 shows values for each fold together with average values.

It can be observed from Table 1 that significant decrease in cost function and increase in classification accuracy is achieved for chromosomes centered to image center by centromere location. The model is only able to achieve 71.10% classification accuracy when chromosomes are distributed in their natural position.

This indicates magnitude of positional variation and its effect on classification accuracy. Comparing results given in Table 1, it can be concluded that chromosome centering can significantly improve model consistency and performance.

Table 1. Classification cost and accuracy

Fold	Random position		Centered position	
	Cost	Accuracy	Cost	Accuracy
1	0.8359	65.38%	0.1301	96.56%
2	0.7862	68.64%	0.1137	96.38%
3	0.6478	72.86%	0.0938	96.92%
4	0.6226	75.39%	0.111	96.26%
5	0.6519	73.22%	0.0897	97.53%
Average:	0.7089	71.10%	0.1077	96.73%

5. Conclusion

In this study we investigated the ability of convolutional neural network model to predict the centromere of human chromosomes location with high accuracy, and the effect of centering chromosome in images by predicted centromere location on classification accuracy.

We proposed preprocessing procedure for chromosome images filtering to reduce noise and other impurities present in images. Image filtering provided images of satisfactory quality that are used as inputs to neural network models for centromere location detection and chromosome classification.

The prediction errors of centromere location are very small, both in terms of MSE and MAE, achieving 6.6638 and 1.8357 respectively. Since the centromere is not defined exclusively nor uniquely by one point, but rather as a small region, achieved results are particularly good as the region is always determined. In general, the model showed a tendency to predict the x coordinate of the chromosome centromere slightly poorer but real difference can only be seen on a few outlining examples and both x and y coordinates are predicted well.

The achieved classification accuracy of chromosomes, randomly oriented in the image, is 71.10%. After centering the chromosomes in the image by predicted centromere location, classification accuracy increased to 96.73%. Achieved increase in chromosome classification accuracy of 25.63% indicates high importance of chromosome centromere detection, and importance of positional variation removal, usually present in images.

References:

- [1]. Schulz-Schaeffer, J. (1980). History of Cytogenetics. In Schulz-Schaeffer, J., *Cytogenetics: Plants, Animals, Humans, 1st ed.*, 2-29. New York: Springer.
- [2]. Kannan, T. P., & Zilfalil, B. A. (2009). Cytogenetics: past, present and future. *Malaysian Journal of Medical Sciences*, 16(2), 4-9.
- [3]. O'Connor, C. (2008). Chromosome segregation in mitosis: The role of centromeres. *Nature Education*, 1(1),
- [4]. Stimpson, K. M., & Sullivan, B. A. (2013). Centromere. In Maloy, S., & Hughes, K. (Eds.), *Brenner's Encyclopedia of Genetics, 2nd ed.*, 500-502. Cambridge, Massachusetts: Academic Press.
- [5]. Lennarz, W. J., & Lane, M. D. (Eds.). (2013). *Encyclopedia of Biological Chemistry, 2nd ed.* Cambridge, Massachusetts: Academic Press.
- [6]. Heim, S., & Mitelman, F. (2015). *Cancer Cytogenetics: Chromosomal and Molecular Genetic Aberrations of Tumor Cells, 4th ed.* Hoboken: Wiley-Blackwell.
- [7]. Gersen, S. L., & Keagle, M. B. (2013). *The Principles of Clinical Cytogenetics, 3rd ed.* New York: Springer.
- [8]. Mohammadi, M. R. (2012). Accurate Localization of Chromosome Centromere Based on Concave Points. *Journal of Medical Signals and Sensors*, 2(2), 88-94.
- [9]. Keerthi, V., Remya, R. S., & Sabeena, K. (2016). Automated detection of centromere in G banded chromosomes. In *2016 International Conference on Information Science (ICIS), Kochi, India, 2016*, 83-86. IEEE. Doi: 10.1109/INFOSCI.2016.7845305.
- [10]. Subasinghe A., Samarabandu J., Li Y., Wilkins, R., Flegal, F., Knoll, J. H., & Rogan, P. K. (2016). Centromere Detection of Human Metaphase Chromosome Images using a Candidate Based Method. *bioRxiv preprint*, 1-12.
- [11]. Wadhwa, A. S., Tyagi, N., & Chowdhury, P. R. (2022). Deep Learning based Automatic Detection of Dicentric Chromosome. In *arXiv preprint arXiv:2204.08029v1*, 1-14.
- [12]. Shen, W., Bai, X., Hu, R., Wang, H., & Latecki, L. J. (2011). Skeleton growing and pruning with bending potential ratio. *Pattern Recognition*, 44(2), 196-209. Doi: 10.1016/j.patcog.2010.08.021
- [13]. Shen, X., Qi, Y., Ma, T., Zhou, Z. (2019). A dicentric chromosome identification method based on clustering and watershed algorithm. *Scientific Reports*, 9(1), 2285.
- [14]. Zhang, T. Y., & Suen, C. Y. (1984). A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3), 236-239. Doi: 10.1145/357994.358023
- [15]. Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), Bombay, India, 1998*, 839-846. IEEE. Doi: 10.1109/ICCV.1998.710815
- [16]. Aswatha, S. M., J. Mukhopadhyay, J., & Bhowmick, P. (2011). Image Denoising by Scaled Bilateral Filtering. In *2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, Hubli, India, 2011*, 122-125. IEEE. Doi: 10.1109/NCVPRIPG.2011.33.
- [17]. Hummel, R. (1977). Image enhancement by histogram transformation. *Computer Graphics and Image Processing*, 6(2), 184-195. Doi: 10.1016/S0146-664X(77)80011-7
- [18]. Ketcham, D. J. (1976). Real-time image enhancement techniques. In *Image processing*, 74, 120-125. SPIE. Doi: 10.1117/12.954708
- [19]. Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., & Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3), 355-368. Doi: 10.1016/S0734-189X(87)80186-X
- [20]. Sternberg, S. R. (1983). Biomedical Image Processing. *Computer*, 16(1), 22-34.
- [21]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016*, 770-778. IEEE. doi: 10.1109/CVPR.2016.90.
- [22]. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009*, 248-255. IEEE. doi: 10.1109/CVPR.2009.5206848.
- [23]. Kingma, D. P., & Ba J. (2017). Adam: A Method for Stochastic Optimization. In *arXiv preprint arXiv:1412.6980v9*, 1-15.