MCIT
Modeling, control & information technologies

# A proposal of algorithm for automated chromosomal abnormality detection

https://doi.org/10.31713/MCIT.2021.26

Pysarchuk Oleksii
National Aviation University
Kyiv, Ukraine
PlatinumPA2212@gmail.com

Mironov Yurii
National Aviation University
Kyiv, Ukraine
yuriymironov96@gmail.com

***Abstract* — The article considers the problem of automatic chromosome abnormalities recognition, using images of chromosomes as an input. This paper's scope includes overview of application domain and analysis of existing solutions. A high-level algorithm for chromosome abnormalities recognition automation is proposed, and a proof-of-concept application is built on top of the algorithm.**

***Keywords* — computer vision; object recognition; decision support system.**

## I. Introduction

According to World Health Organization estimates, 295 000 newborns die within 28 days of birth every year, worldwide, due to congenital abnormalities. Also congenital abnormalities can cause long-term disability [1]. Approximately 20% of such abnormalities are caused by chromosomal and genetic conditions [2]. Modern reproductive medicine makes it possible to detect these abnormalities before and during pregnancy by means of cytogenetic diagnostics [3].

However, currently such diagnostics is conducted manually or partially automated, which leaves the diagnostics result prone to errors due to human factor. Moreover, such a sophisticated procedure takes significant time and effort to conduct. These problems may be solved by process automation on a larger scale. This article intends to briefly overview application domain, analyze existing solutions and propose a high-level algorithm for automatic detection of chromosome abnormalities.

## II. Related Work

The problem of chromosome abnormalities detection is known, but is not covered by an extensive amount of research papers. The current state of problem leaves room for further research. Some of the most prominent researches concerning the topic have been considered.

"Development and evaluation of automated systems for detection and classification of banded chromosomes: Current status and future perspectives" by Xingwei Wang and Bin Zheng has been a starting point for the current research. It reviews possible approaches to the problem of chromosome pathologies recognition [4]. This article is more of a theoretical research, it it underlines the importance of artificial intelligence and neural networks in chromosome abnormalities recognition.

"Deep Learning for Medical Image Processing: Overview, Challenges and Future" by Muhammad Imran Razzak, Saeeda Naz and Ahmad Zaib considers possible approaches to the problem of medical images recognition and processing [5]. The main point of the paper is about healthcare is much different from majority of other application domains due to high risk, cost of error and complexity/diversity of data. However, according to authors, all these challenges might be faced by means of modern Machine Learning (ML) algorithms. Authors also state that Computer Vision (CV) is an crucial tool for medical data processing, since a large part of medical data is visual by its nature (e.g. microscope photos or X-Ray images).

Authors mention peculiarities of ML and deep learning that might make its usage more difficult in medical domain. The following challenges are listed:

- Input data standardization. Any medical sub-domain is powered by a variety of hardware and software, and they do not necessary have any convention about data format. This will complicate gathering a comprehensive dataset;
- Privacy issues. Medical data is sensitive, so gathering a dataset will be further perplexed due to the need to anonymize data;
- Dataset size issue. Specific medical data is sparse, therefore challenging to gather. Moreover, corner-cases, such as rare diseases, are even more sparse yet crucial for neural network efficiency;

Authors point out that the key to success of ML-powered systems is in cooperation among international healthcare organizations and providers of healthcare solutions, since this will help to develop uniform standards for future hardware and software.

All the issues about ML-powered solutions mentioned by the authors are present and relevant in domain of reproductive medicine. Therefore, usage of ML for implementing chromosome pathology detection is debatable.

"End-to-end chromosome karyotyping with data augmentation using GAN" by Yirui Wu, Yisheng Yue, Xiao Tan, Wei Wang and Tong Lu proposes a chromosome recognition method powered by Generative Adversarial Network (GAN) [6]. Authors review main challenges of chromosome recognition, such as arbitrary shapes and contour overlapping, and claim that traditional data recognition algorithms are

inefficient and hard to use for such tasks. They propose usage of deep learning neural networks. However, such an approach requires an arranged dataset of significant size. There are multiple reasons why gathering a dataset like this is a challenging task due to multiple reasons such as privacy and data heterogeneity.

Authors state that generating new data using GAN is a possible solution to these problems.The main idea is to utilize GAN in order to create new images used for learning. Chromosome photos are to be used as an input. These photos should be processed with multiple filters and then chromosomes should be extracted from image.

The proposed approach is promising, however it does not solve some challenges of chromosome pathologies recognition, such as adjacent chromosomes recognition. Moreover, generating new data based on some initial dataset will result in dataset lacking corner-cases, which is crucial for algorithm efficiency.

Having conducted related papers analysis, the following may be concluded:

- The problem of chromosome analysis and pathology detection is widely known and has no applicable solution yet;
- All considered papers suggest using ML-powered approach, yet mention flaws due to specific nature of reproductive medicine data;
- Gathering dataset is a challenging task due to privacy issues, data heterogeneity and risk of insufficient cases coverage;

These problems make it reasonable to look for alternative approaches to the solution of this problem. It is probable that an algorithm that does not rely on a dataset is viable. Therefore, the goal of this paper is to develop a basic algorithm and a proof-of-concept software that could solve the issue.

## III. Proposed Method

A brief recap of application domain research is provided before the proposed algorithm for better understanding and bore context.

### *A. Application Domain Overview*

During the conventional cytogenetic analysis, chromosomes and their abnormalities are detected manually by their shape, size and pattern. Metaphase plate photo serves as an input image for manual cytogenetic analysis (fig. 1. A) [7]. In order to categorize and visually arrange chromosomes, they are manually turned into karyograms (fig. 1. B).

The specialists that conduct this analysis are used to working with raw images captured by microscope and rarely need any reference. However, in order to pinpoint specific regions of chromosomes, international community of cytogenetics has developed a uniform system of chromosome mapping. This system makes use of schematic chromosome images called ideograms [8]. Each chromosome has a schematic representation [8], [9].

Chromosome abnormalities can be categorized as either numerical or structural. Numerical abnormalities are whole chromosomes that are either missing or extra to a normal pair. Structural abnormalities are parts of chromosomes that are missing, duplicated or moved from one chromosome to another [10].

### *B. Expected Input and Output of an Algorithm*

The algorithm is expected to handle chromosome images. As mentioned above, two common ways to depict chromosomes are metaphase plate photo and karyogram. The production-ready version of algorithm should make use of metaphase plate photos, since the ability to process them is a crucial part of chromosome analysis automation. However, a proof-of-concept version makes use of karyograms for simpler recognition.

The expected output of an algorithm is a set of abnormalities and a supposed diagnosis concerning chromosomal diseases. Despite the abnormalities can be reliably associated with diagnosis, it is not the algorithm's responsibility to call final diagnosis. The results of cytogenetic analysis should be forwarded to a doctor that has to make final solution.

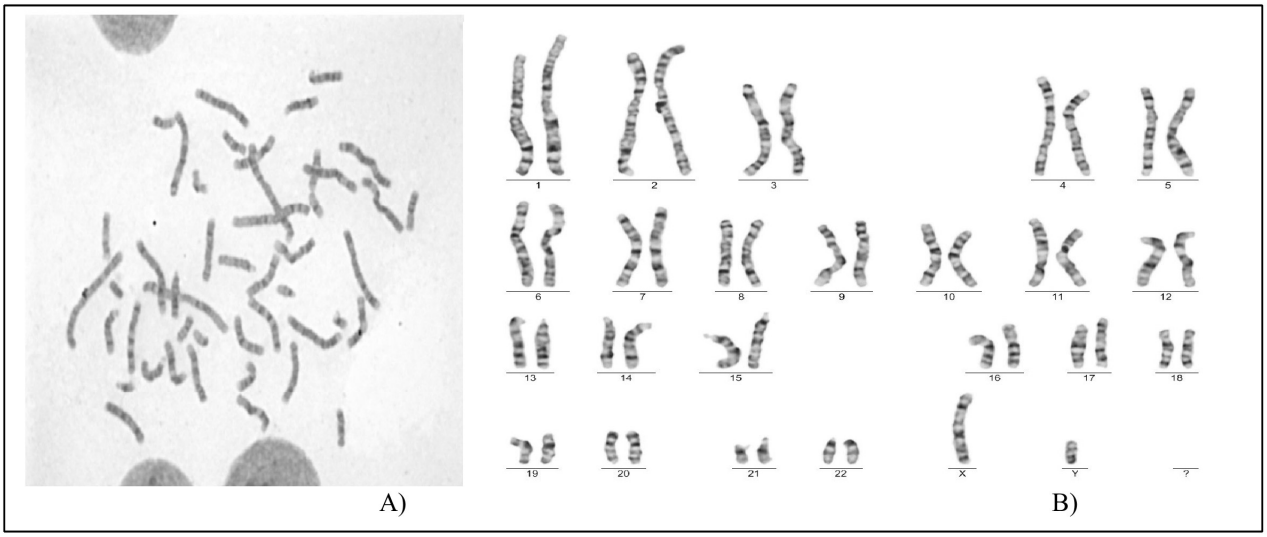


Figure 1. Chromosome images: A) Metaphase plate; B) Karyogram

*C. Algorithm Overview*

Figure 2 depicts the general algorithm flow. High-level algorithm will consist of the following steps:

- Image processing. Detection of both numerical and structural abnormalities depend on recognized chromosomes. The goal of this step is to get input image and extract chromosomes of it. This step also includes chromosome feature extraction and serialization of chromosome data into the custom format. Before attempting to get chromosome data from image, it has to pe preprocessed, removing noises and obstacles;
- Mapping chromosomes to ideograms and detecting abnormalities. This step should be powered by custom data format designed for efficient difference detection on chromosome structure;
- Identifying pathologies. Having recognized chromosomes and detected differences between them and their respective ideograms, it is possible to identify pathologies. The goal of this step is to conduct decision-making process and map differences (e.g. three 21 chromosomes instead of two) and specific diseases (e.g. Down syndrome);

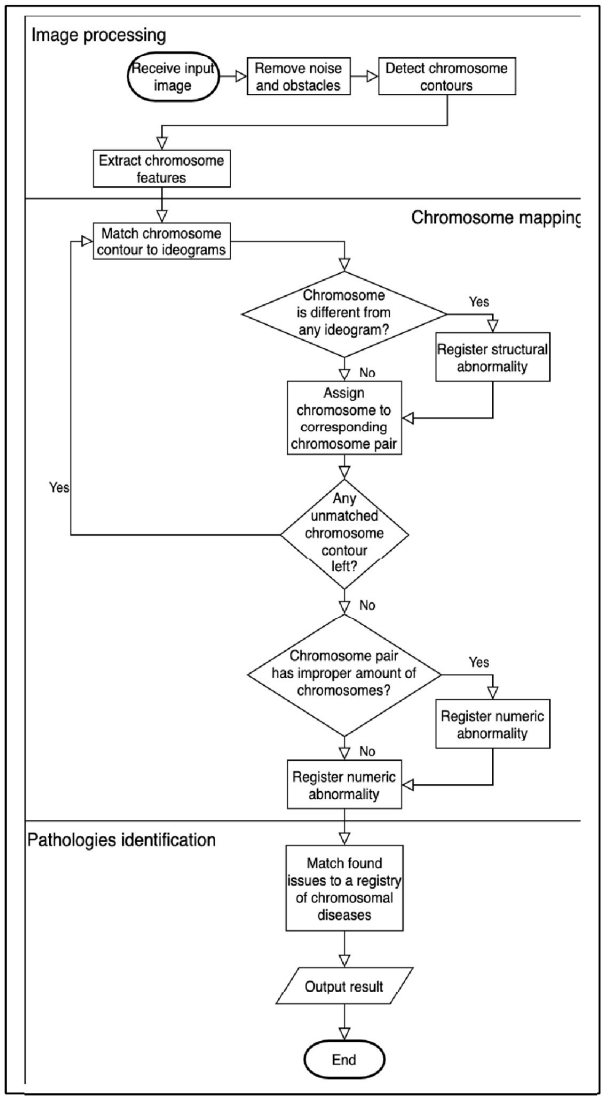


Figure 2. Proposed algorithm flowchart

This algorithm is just a basic framework designed to be broken down into separate steps that can be extended and improved separately. Each of aforementioned steps also can be broken down into separate problems.

*D. Prototype implementation*

Having described an algorithm, it is possible to design a software prototype that implements this algorithm. Python programming language is selected because of its rich infrastructure and set of powerful libraries. OpenCV library is selected for tasks related to image processing. The business-logic core is designed as a library that can be used by a variety of interfaces, and a simple command-line interface has been implemented along with the prototype. The class diagram with application structure is shown on Figure 3.

The software will use karyograms as an input, will have simplified checks for chromosome types and therefore will be capable to detect only numeric pathologies. Due to the nature of karyogram image, it will be possible to assume a chromosome group by its position on image. Later this way of chromosome recognition will be replaced by a proper comparison with ideogram, but detecting chromosome type is out of scope of this article and is a subject for further research - the algorithm will not be complete without it.

Several object detection algorithms have been tested, including Blob Detection [11], Laplacian Edge Detection, Ridge Detection and Canny Edge Detection. The latter proved to be most effective with the image set used for this paper. However, by design they are interchangeable and can easily be replaced if other object detection algorithm will provide better results.

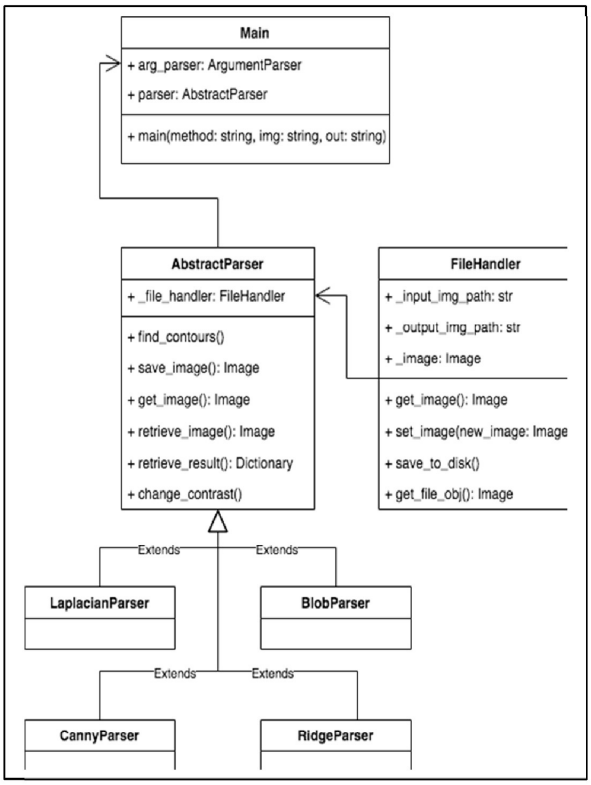


Figure 3. Prototype software class diagram

## IV. Experiment Result

A testing dataset consists of 45 karyotype images, including 20 karyotypes with no detected pathologies and 25 karyotypes with 4 different numeric chromosome abnormalities. General detection precision within this dataset is 93.33% (Table 1). No false negative results have been obtained, however there have been 3 false positive results for karyotypes with no pathologies. Wrong results have been caused by insufficient calibration of contour detection settings.

TABLE I. Test Run Results

| Input Image Info | Number of Images | Test Run Comments | Success Rate |
|---|---|---|---|
| No known pathologies | 20 | False positives have been caused by insufficient contour detection settings calibration | 85% |
| Edwards syndrome | 4 | - | 100% |
| Down syndrome | 14 | - | 100% |
| 22 chromosome trisomy | 3 | - | 100% |
| X chromsome trisomy | 4 | - | 100% |
| Mean success rate | | | 93.33% |

## V. Conclusion

Given paper considers the problem of chromosome abnormality recognition. Manual workflow of chromosomal diagnostics has been overviewed and related papers have been researched. It has been concluded that there are no comprehensive solutions to this problem yet, so an algorithm has been described.

The algorithm is based on extracting features out of chromosomes, matching them to ideograms and deriving diagnosis out of found abnormalities. This algorithm is not powered by neural networks and no dataset is needed for its learning. The way the algorithm is built allows to break the problem into separate parts and improve on them iteratively.

A prototype has been implemented to prove that the algorithm is viable. Currently prototype accepts limited amount of image types and detects only numeric pathologies, but its architecture allows to separately improve on its module, reflecting the flexibility of the algorithm. The prototype run on test dataset showed high success rate.

This paper described the basic framework of the algorithm, granting room for future improvements. The future work concerning the algorithm should be focused around several main areas:

- Raw image recognition. Currently the prototype only supports images that have been manually preprocessed (karyograms). In order to be an efficient decision support tool for cytogeneticist, it should be able to work with raw metaphase plate images, reducing time needed for manual analysis;
- Internal chromosome features processing . In order to detect structural abnormalities, the algorithm should be able to recognize internal chromosomal features. This way it will be able to match them to ideograms and detect pathologies caused by changes inside single chromosome;

## References


[1] World Health Organization, “Congenital anomalies,” 1 December 2020. Retrieved from https://www.who.int/news-room/fact-sheets/detail/congenital-anomalies

[2] A. Mohammadzadeh, S. Akbaroghli, E. A.-Moghadam, N. Mahdieh, R. S. Badv, P. Jamali, et al, “Investigation of Chromosomal Abnormalities and Microdeletion/ Microduplication(s) in Fifty Iranian Patients with Multiple Congenital Anomalies,” Cell J., Autumn 2019. Vol. 21(3), pp. 337–349.

[3] R. Mariluce, “Human molecular cytogenetics: from cells to nucleotides,” Genetics and Molecular Biology. 2014. V. 37.

[4] X. Wang, B. Zheng, M. Wood, S. Li, W. Chen, and H. Liu, “Development and evaluation of automated systems for detection and classification of banded chromosomes: current status and future perspectives,” Journal of Physics D: Applied Physics, 2005. Vol. 38.

[5] M. I. Razzak, S. Naz and A. Zaib, “Deep Learning for Medical Image Processing: Overview, Challenges and the Future,” Deep Learning for Medical Image Processing: Overview, Challenges and the Future, N. Dey, A. Ashour, S Borra, Eds. Springer, 2018. vol. 26.

[6] Y. Wu, Y. Yue, X. Tan, W. Wang and T. Lu, "End-To-End Chromosome Karyotyping with Data Augmentation Using GAN," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 2456–2460.

[7] C. O’Connor, “Karyotyping for Chromosomal Abnormalities,” Nature Education, 2008.

[8] C. O’Connor, “Chromosome Mapping: Idiograms,” Nature Education, 2008.

[9] National Center for Biotechnology Information (US), “Genes and Disease,” 1998. Retrieved from. URL: https://www.ncbi.nlm.nih.gov/books/NBK22266/

[10] Genetic Alliance, The New York-Mid-Atlantic Consortium for Genetic and Newborn Screening Services, “Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals,” Washington (DC), July 8 2009.

[11] K. T. M. Han and B. Uyyanonvara, "A Survey of Blob Detection Algorithms for Biomedical Images," 2016 7th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), 2016, pp. 57–60.