The University of Hong Kong
Department of Computer Science

AI-powered Abnormality Detection for Karyotype Analysis

Project Plan

Supervised by Professor T.W. Lam

Ng Hoi Ying (3035377957)
# TABLE OF CONTENT

1. INTRODUCTION .......................................................... 3
   1.1 BACKGROUND .................................................. 3
   1.2 OBJECTIVE .................................................. 3
   1.3 SCOPE ....................................................... 3

2. PROPOSED METHODOLOGY ........................................... 4
   2.1 STATISTICAL ANALYSIS AND CLASSIFICATION .............. 4
      2.1.1 DATA PREPARATION .................................. 4
      2.1.2 CLASSIFICATION ...................................... 5
   2.2 CONVOLUTIONAL NEURAL NETWORK .......................... 6

3. SCHEDULE AND MILESTONE ........................................ 7

4. CURRENT STATUS ..................................................... 8

5. REFERENCE ........................................................... 8
1. Introduction

1.1 Background

A chromosome is an organized package of DNA found in the nucleus of the cell. Each person normally has 23 pairs of chromosomes in each cell—one of the pair is called sex chromosomes, X and Y. The other 22 pairs are numbered chromosomes, called autosomes [1].

The study of metaphase chromosomes is an important aspect of clinical diagnosis of genetic disorders. Karyotyping is a conventional test to examine chromosomes for identifying chromosomal abnormalities. However, karyotyping analysis is a very tedious and time-consuming task. It takes around 1.5 days for an experienced cytogeneticist to serve one patient. Although chromosome image is a topic undergoing intense study for more than a decade, most research focuses on segmentation and classification. Not many researches have been conducted, from the bioinformatics perspective, to identify and interpret images for structural abnormalities.

1.2 Objective

The demand for identifying chromosomal abnormalities is increasing, especially in mainland China. This project aims to create an automated karyotyping analysis system for abnormality detection at clinical level. This will help to shorten the analysis time, addressing the shortage of experienced cytogeneticists. The end deliverables included an image pre-processing system, a well-trained machine learning model, which identify various genre of structural abnormalities (deletions, duplications, translocations, inversions, insertions, rings, and isochromosome) of the chromosome and its precise location.

1.3 Scope

The scope of this project is limited to a model that detect structural abnormalities on karyogram (segmented, classified, and organized chromosome image), the model does not include segmentation and classification from an unprocessed microscopic image. Furthermore, the project is specifically focused on a comparatively long chromosome, i.e. structural variation > 10MB.
2. Proposed Methodology

As a pilot project, the study and implementation of various models based on different machine learning algorithms and architecture will be conducted to find the optimal solution. The project is currently at a nascent stage and involves learning several new concepts, the methodology proposed in this report is tentative and may be altered, depending on the flow of the project.

2.1 Statistical Analysis and Classification

In this approach, the data undergoes pre-processing, transform the image information into numerical data. The key feature of a chromosome using for abnormality identification is the banding pattern. Banding patterns are patterns of light and dark transverse bands on the chromosome. Each chromosome has a characteristic pattern. By comparing the chromosomes in a pair, including but not limited to the size, intensity level, and banding pattern, it is feasible to detect the abnormalities.

2.1.1 Data preparation

To extract the banding profile, a method proposed in previous research is adopted and modified to better suit our needs and simplify the implementation[2].

**Medial axis extraction.** The approximation medial axis is the ground of the band profile extraction. Zhang-Suen’s two-pass thinning algorithm was chosen. The thinning algorithm works by making successive passes of the image, removing pixels on object borders. This continues until no more pixels can be removed[3]. It is widely used due to its outstanding performance at connectivity and contour noise immunity. Besides that, an existing open-source library (scikit-image) which helps shorten the implementation time is another reason for picking this algorithm.

However, the algorithm is not perfect. This approach requires the medial axis to reach two ends of the chromosome so as to analyse the whole banding pattern. Therefore, a simple algorithm is applied to extend the skeleton by calculate the slope at two ends and add lines with the same slope connecting the skeleton’s end and the chromosome’s end. Another commonly used algorithm on medial axis extraction is the discrete curve algorithm. Despite the advantage it has, it is difficult to implement[4]. A MATLAB library could be found but obsolete.

**Intensity analysis.** The intensity sampled at ¼, ½, and ¾ of width along each cross-section is used to calculate the grey-value representing that cross-section to reduce the effect of noise at the boundary and reduce outliner caused by banding and the image acquisition process[2].
2.1.2 Classification

Classification is a technique to categorize our data into a desired and distinct number of classes where we can assign a label to each class. Classification can be binary or multi-class, as to differentiate the different types of abnormality, a multi-class classifier is chosen. There are several types of commonly used classification algorithms. Here is the comparison of those which cover the requirement of this project:

**Decision Tree.** Decision Tree is a tree-like model for decision making. It produces a sequence of rules that can be used to classify the data.

**Advantage:** Decision tree is simple to understand, interpret, and visualize. The condition can be easily explained by boolean logic. Also, it requires little data preparation comparing to other algorithms, which often require normalisation, dummy data creation, and blank value removal.

**Disadvantage:** Most of the disadvantages can be mitigated, for example, the instability can be improved by using a decision tree within an ensemble. However, one critical disadvantage is that a biased tree could be created due to class domination. In this project, the dataset is not balanced (the number of normal cases is dominated). Balancing the data will lead to a large reduction in dataset size, affect the accuracy of the model. Besides, Decision Tree does not support missing values. Since it is common to have overlapping chromosomes, missing value is not a rare case.

**Random Forest.** Random Forest is based on ensemble learning. It fits multiple decision trees on various sub-samples of datasets.

**Advantage:** The use of multiple trees improve the accuracy of the model and controls over-fitting. It works well with missing value.

**Disadvantage:** Due to the complex design, it requires more computational resources and longer training time. Besides, it is difficult to implement.

**Support Vector Machine (SVM).** A representation of the training data as points in space separated into categories. Prediction is derived based on the side of the gap it fall.

**Advantage:** SVM is effective in high dimension spaces relatively memory-efficient because it uses a subset of training points in the decision function.

**Disadvantage:** The SVM algorithm does not have a good performance in a large and noisy dataset, this two features can be found on the datasets in this project. Also, the algorithm does not provide a probabilistic explanation of the prediction.

Base on the above comparison, the Random Forest is chosen as the prior algorithm.
2.2 Convolutional neural network

Convolutional Neural Network (CNN) is a category of deep neural network that has proven very efficient in areas such as image recognition and classification. It constructs a feature hierarchy by combining the low-level features in a layer-wise fashion to form high-level features. For example, if we’re dealing with images, then low-level features, such as edges and blobs, are extracted from the earlier layers, which are combined to form high-level features – as object shapes like a building, a car, or a dog [6].

There are several commonly used architectures:

**AlexNet(2012).** AlexNet is a deeper and wider version of pervious CNN architect LeNet. It featured convolution layer stacked on top of each other.

**GoogLeNet(2014).** A major breakthrough of GoogLeNet was the development of an Inception Module which greatly reduced the number of parameters in the network (4M, compare to AlexNet with 60M [6]). Besides, this architecture uses Average Pooling instead of Fully Connected layers at the top, eliminating a large amount of trivial parameters.

**VGGNet(2014).** The main contribution of VGGNet was in showing that the depth of network is a critical component for good performance.

**ResNets(2015).** ResNets features special skip connections and a heavy use of batch normalization. ResNets are currently by far state of the art Convolutional Neural Network models and are the default choice for using CNN in practice (as of May 10, 2016) [6].

However, the above architectures mainly focus on image recognition and classification. In this project, comparison of paired chromosomes and identification of unusual cases is the key purpose. Therefore, modification on current architecture or even a whole new design is demanded to serve our goal.
3. Schedule and Milestone

In Phase 1, from August to September, exploratory data analysis has been carried out on the dataset provided to understand the problems. The deliverables are this project plan and the project website.

In Phase 2, from October to January, the 1\textsuperscript{st} approach (statistical analysis and classification approach) and 2\textsuperscript{nd} approach (CNN) will be implemented. The deliverables are the trained model of both approach and the interim report.

In Phase 3, from February to April, since there is a low chance to have the best CNN model in the first trial, therefore 3\textsuperscript{rd} approach also in CNN will be implemented. The deliverables are the final implementation and the final report.

<table>
<thead>
<tr>
<th>Task</th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study and Research</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project plan and web-page</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data cleaning and processing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1\textsuperscript{st} Approach Implementation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Presentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preliminary implementation &amp; interim report</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2\textsuperscript{nd} Approach Implementation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3\textsuperscript{rd} Approach Implementation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organize result and testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finalized implementation &amp; Final Report</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final Presentation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Gantt Chart outlining the proposed timeline. Coloured box indicates tasks; collate indicates the milestones.
4. Current Status

At present the project is on schedule, the data cleaning and processing for the Statistical Analysis Approach is almost finished. A data processing system is created to transform the image information into numerical data that can fit into a classification model. The current task is to finalize the system and manually label abnormal cases. The next step is to train the model with these data with the chosen classification algorithm.

5. Reference


