Towards automated detection of dicentric chromosomes in metaphase images

by

Sárah Galloway

Thesis presented in partial fulfilment of the requirement for the degree of Master of Science (Applied Mathematics)in the Faculty of Science at Stellenbosch University

Supervisor : Dr J. Coetzer

Co-supervisor : Dr N. Muller

March 2021

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Student number	Signature
S Galloway	1 March 2021
Initials and surname	Date

Abstract

The aim of the proposed research is to investigate methods to identify objects of interest and classify dicentric and normal chromosomes in metaphase images using suitable digital image processing techniques. Dicentric chromosomes are abnormal chromosomes with two centromeres (instead of one) created by a variety of processes, including irradiation. When a chromosome is exposed to radiation, two chromosome segments, each with a centromere may join together resulting in a dicentric chromosome. An acentric fragment, i.e. a partial chromosome with no centromere, is also formed. The first stage of the proposed system is geared towards the detection of objects of interest, i.e. isolated normal and isolated dicentric chromosomes, as well as acentric fragments and clusters of overlapping chromosomes. The last stage of the proposed system is geared towards the classification of isolated chromosomes as either normal or dicentric. The proposed system *automatically* detects objects of interest not associated with dirt. The classification of the aforementioned objects into isolated and clustered chromosomes, as well as acentric fragments, is conducted *manually*, while the automation of this stage is reserved for future work. The proposed system subsequently *automatically* categorises isolated chromosomes as either normal or dicentric. It is demonstrated that the system correctly detects and classifies a significant number of the aforementioned chromosomes within metaphase images provided by iThemba LABS.

Uitreksel

Die doel van die voorgestelde navorsing is om metodes te ondersoek wat voorwerpe van belang in metafase-beelde identifiseer en disentriese en normale chromosome met behulp van geskikte beeldverwerkingstegnieke klassifiseer. Disentriese chromosome is abnormale chromosome met twee sentromere (in plaas van een) wat deur verskeie prosesse, insluitende bestraling, geskep word. Wanneer 'n chromosoom aan bestraling blootgestel word, kan twee chromosoomsegmente, elk met 'n sentromeer, saamgevoeg word wat 'n disentriese chromosoom tot gevolg het. 'n Asentriese fragment, dit wil sê 'n gedeeltelike chromosoom sonder 'n sentromeer, word ook gevorm. Die eerste fase van die voorgestelde stelsel is op die opsporing van voorwerpe van belang gerig, dit wil sê geïsoleerde normale en geïsoleerde disentriese chromosome, sowel as asentriese fragmente en groepe van oorvleulende chromosome. Die laaste fase van die voorgestelde stelsel is op die klassifikasie van geïsoleerde chromosome as normaal of disentries gerig. Die voorgestelde stelsel bespeur *outomaties* voorwerpe van belang wat nie met vuilheid verband hou nie. Die klassifikasie van bogenoemde voorwerpe as geïsoleerde en gegroepeerde chromosome, asook asentriese fragmente, word met die *hand* gedoen, terwyl die outomatisering van hierdie fase vir toekomstige werk gereserveer is. Die voorgestelde stelsel kategoriseer vervolgens geïsoleerde chromosome *outomaties* as normaal of disentries. Daar word aangetoon dat die stelsel 'n beduidende aantal van bogenoemde chromosome korrek opspoor en klassifiseer binne die metafase-beelde wat deur iThemba LABS verskaf is.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following people and organisations:

- My supervisor, Dr Hanno Coetzer, for his invaluable insight, guidance, patience, unwavering support, immense knowledge and valuable critiques of this research work. This study would not have been possible without his input.
- Dr Neil Muller, for his co-supervision and fundamental insights that contributed to the quality of the research presented in this thesis.
- Dr Charlot Vandevoorde and her team at iThemba LABS for generating the data set.
- iThemba LABS for their patience and effort regarding the financial assistance of this study.
- My family and close friends, for their unconditional love and support.
- My fiancé, Michael Fouché, for his patience.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND/OR ACRONYMS

ACC	Accuracy
DCE	Discrete curve evolution
DRT	Discrete Radon transform
FNR	False negative rate
FPR	False positive rate
GVF	Gradient vector field
Gy	Gray
HPM	Histogram projection method
ROI	Region of interest
SE	Structuring element
SMAC	Straightening via Medial Axis extraction and Crowdsourcing
SPV	Straightening via Projection Vectors
SVM	Support Vector Machine
TNR	True negative rate
TPR	True positive rate

Nomenclature

$S_{x,y}$	Set of all the pixels in a selected image window with the centre at point (x, y)
g	Degraded image
$\hat{f}$	Restored image
$\bar{f}$	Blurred image
g_{mask}	Image mask
f	Input image
w	Weighted degree of sharpening
h	Histogram
q_k	Number of pixels with grey-scale intensity value k
r_k	Grey-scale intensity value k
ρ	Projection profile
$\Re$	Discrete Radon transform (DRT)
l	Maximum variation associated with connected component
n	Number of boundary points in edge graph
N	Number of neighbouring boundary points
a	Bounding box area of connected component
C	Twice differentiable plane curve
κ	Curvature, that is the magnitude of the change in θ per unit arc length
m_x	The gradient of a straight line
c_x	The y-intercept of a straight line

Chapter 1

Introduction

1.1 Background and motivation

The dicentric chromosome assay is a well-established method used to estimate exposure to ionising radiation. Dicentric chromosomes are considered to be specific to radiation exposure as they are primarily generated by ionising radiation and only a few radiomimetic drugs. As a result, the background levels of dicentric chromosomes are low in non-exposed individuals, which makes it possible to assess irradiation doses as low as 0.1 Gy. Due to the aforementioned advantages, this assay is considered to be the gold standard of radiation biodosimetry (Romm *et al.* (2013)).

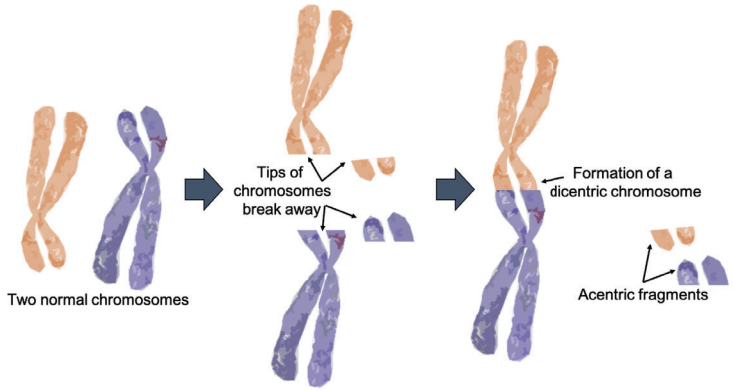


Figure 1.1: Formation of a dicentric chromosome with accompanying acentric fragments.

The formation of a dicentric chromosome is illustrated in Figure 1.1. This aberration involves an interchange between two separate chromosomes. If ionising radiation causes breaks in two chromosomes and the sticky ends are sufficiently close to one another, they may rejoin so that a grossly distorted *dicentric* chromosome, with two centromeres, is produced. The two remaining

fragments that possess no centromeres are referred to as acentric fragments. Since the assay is labour-intensive and time-consuming, automated tools for detecting and classifying metaphase images are of considerable interest to increase throughput for biodosimetry in radiation mass casualty incidents.

1.2 Scope and objectives

Several image processing-based techniques and machine learning-based techniques have been successfully implemented for the purpose of chromosome detection and classification (Jindal *et al.* (2017), Markou *et al.* (2012), Rogan *et al.* (2014) and Shirazi *et al.* (2016)). Due to variations in imaging equipment and differences in methods for treating samples, implementations are often specific to a single laboratory. The quality of the images from a single experiment may also vary significantly as a result of differences in the imaging and biological effects.

The purpose of this research is to investigate the feasibility of image processing techniques for the purpose of detecting and classifying chromosomes in metaphase images produced at iThemba LABS. This may eventually constitute an essential part of a semi-automated system that lessens the burden on human operators.

The scope of this work is limited to an investigation into image segmentation strategies for the purpose of object detection in metaphase images. Said strategies involve the application of a novel binarisation method in order to optimise the number of objects of interest to be detected. After the segmentation process has been completed, the scope of this work is further limited to the classification of *only* the *isolated* chromosomes as either normal or dicentric. The segmented acentric fragments and clusters of chromosomes that are either on top of each other or in close proximity of one another is

manually discarded. The automated classification of these objects therefore falls outside the scope of this thesis.

Chromosome classification is not only a time-consuming process for well-trained experts, but these same experts often disagree on the classification results. It is also important to note that, within the context of this thesis, the ground truth can only be as good as the judgments of the experts that were considered in generating the ground truth for the purpose of this research study. The reported proficiency and contribution to the current state of the art of the systems developed in this thesis should therefore be judged against this background.

1.3 Thesis overview

In this section the reader is provided with a concise overview of the thesis. This overview comprises of (1) the proposed region of interest (ROI) detection protocol, (2) the proposed feature extraction protocol for isolated chromosomes, (3) an outline of the experiments performed in order to assess the efficacy of the systems developed in this thesis, and (4) the experimental results.

1.3.1 System design

A general system composed of three stages is proposed for this research. The proposed system is conceptualised in Figure 1.2. During Stage A all objects of interest are detected, while isolated chromosomes are selected during Stage B. Stage C is employed for the purpose of classifying the isolated chromosomes as either normal or dicentric. For the purpose of this research Stages A and C are automated, while Stage B is manually executed. The automation of Stage B is reserved for future work.

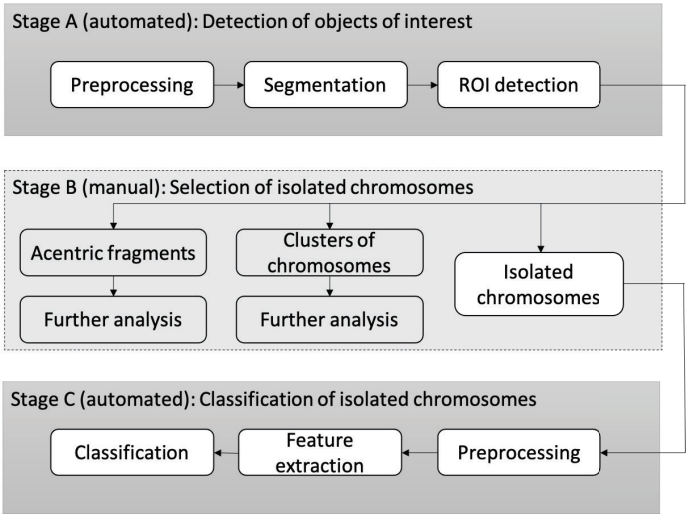


Figure 1.2: Conceptualisation of the chromosome detection, selection and classification protocol proposed in this research.

1.3.2 Data

The experimental data constitutes metaphase images from healthy male and female individuals that were prepared in the radiobiology laboratory at iThemba LABS. This dataset consists of grey-scale images that are captured under a light microscope at a resolution of 1280×1024 pixels. The metaphase images are subsequently exported as TIFF files from the Metafer4 system for further analysis. In order to validate the system proposed in this thesis, a semi-automated protocol is developed in order to generate a ground truth. In generating the ground truth for the acquired metaphase images in the dataset, seven experts from iThemba LABS with varying degrees of experience were utilised.

1.3.3 Image segmentation

In Stage A all objects of interest, that do not constitute dirt, are *automatically* detected. These objects include isolated normal chromosomes, isolated dicentric chromosomes, acentric fragments and clusters of chromosomes. The aforementioned clusters may contain more than one chromosome that are either located on top of each other, or are in very close proximity of each other. Stage A (the generic detection phase) is implemented in three steps:

1. Preprocessing for the purpose of enhancing image quality,
2. segmentation for the purpose of determining possible regions of interest (ROIs) in the metaphase image, and
3. object of interest detection for the purpose of eliminating dirt within the slide image.

In Stage B (the selection phase) the detected objects are *manually* categorised into dirt, acentric fragments, clusters of chromosomes, isolated normal chromosomes and isolated dicentric chromosomes. All acentric fragments, clusters of chromosomes and dirt are manually discarded, while only the isolated chromosomes are retained.

The proposed generic detection phase (Stages A and B) is conceptualised in Figure 1.3.

1.3.4 Feature extraction

In Stage C (the classification phase) the manually selected isolated chromosomes are automatically classified as either normal or dicentric. This stage is implemented in three steps:

1. Preprocessing for the purpose of image analysis,
2. feature extraction for the purpose of maximising the proficiency of the

classifier, followed by

3. categorisation through merging the aforementioned information.

The proposed classification phase (Stage C) is conceptualised in Figure 1.4.

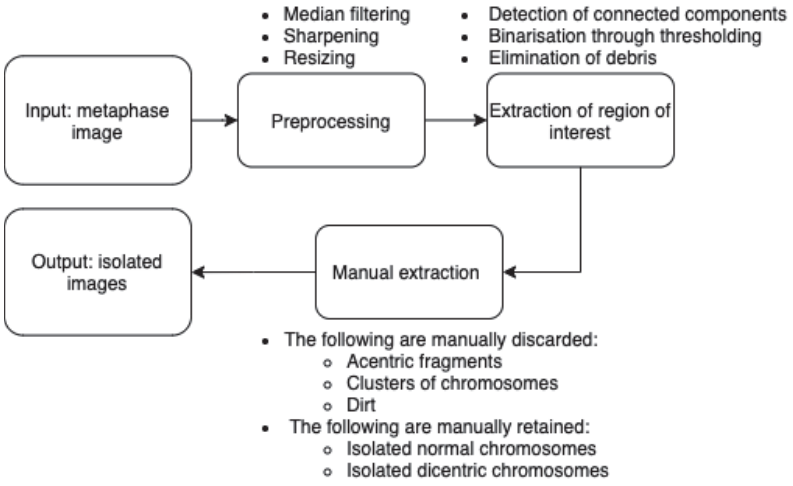


Figure 1.3: Conceptualisation of the proposed generic detection phase (Stages A and B) implemented in this thesis.

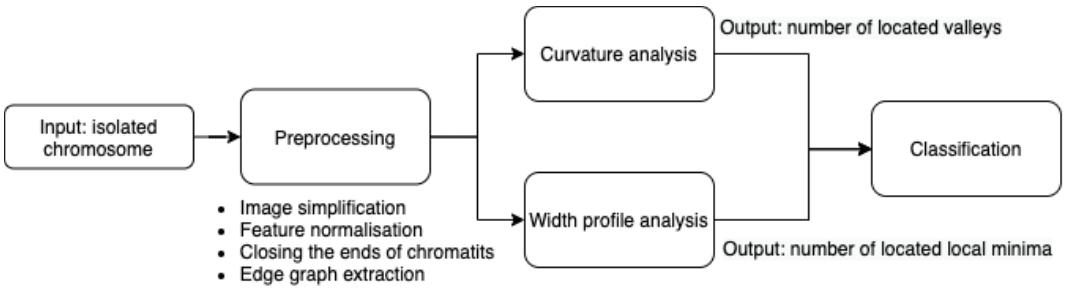


Figure 1.4: Conceptualisation of the proposed classification phase (Stage C)

1.3.5 Abbreviated results

In this thesis experiments are carried out in order to determine the proficiency of the novel strategies proposed to detect regions of interests. In order to evaluate the proposed protocols a ground truth is generated. The detection phase that employs the image segmentation protocol proposed in this thesis

detects isolated objects of interest with a promising true positive rate (TPR) of 90.10% and detects objects of interest which includes cluster of chromosomes with a TPR of 94.37% when compared to the ground truth obtained from experts at iThemba LABS.

In classifying the isolated chromosomes, three strategies were considered, that is (1) width profile analysis, (2) curvature analysis and (3) an aggregated approach that combines width profile analysis and curvature analysis. Accuracies of 37.49%, 87.14% and 89.95% are respectively reported in classifying isolated chromosomes as normal or dicentric. It is therefore concluded that width profile analysis (in isolation) does not perform adequately for the iThemba LABS dataset, and that the proposed *novel* protocol based on curvature analysis is much more robust and proficient. However, the ulilisation of width profile information in *conjunction* with curvature information, may improve the accuracy of the results. When the curvature analysis and aggregated approaches are compared, true positive rates (TPRs) of 87.04% and 81.48% are respectively reported. The overall TPR is deemed an important metric to consider when comparing these two approaches.

1.4 Contributions

The task of manually scoring metaphase images is an extensive process that is very time-consuming. Determining the manual assay requires well trained experts which makes it an extremely costly task. The time-consuming nature of manually scoring metaphase images is eliminated by introducing automated or semi-automated systems. The aforementioned systems can assist experts in cases of large scale radiation exposure. The main contributions made in this thesis are as follows:

A novel chromosome segmentation protocol. A novel automated segmentation protocol for detecting chromosomes in a metaphase image produced at iThemba LABS is proposed by automatically determining the threshold value needed to binarise the image in a heuristic way. Since metaphase images is specific to the laboratory where they are produced, the iThemba LABS dataset did not perform well using well known binarisation methods (like Otsu's method). However, by implementing the proposed heuristic binarisation method, significantly better results are obtained. It is demonstrated in Chapter 4 that the proposed chromosome segmentation and detection protocol successfully segments the chromosomes in a metaphase image.

A novel feature extraction protocol that employs curvature analysis on isolated chromosomes. To the best of our knowledge, the utilisation of curvature analysis to exploit shape information for the purpose of classifying chromosomes is novel and constitutes one of the contributions of this study.

Publication. The proposed system was published in the Proceedings of the 2020 International SAUPEC/RobMech/PRASA Conference (Galloway *et al.* (2020)).

Chapter 2

Literature study

2.1 Introduction

As mentioned in the previous chapter numerous research studies on automated chromosome detection and classification have been conducted. In this chapter a concise overview of relevant existing chromosome detection and classification systems is presented, irrespective of whether these systems are based on metaphase images, karyotyped images or isolated chromosome images. The discussion provided on the aforementioned systems is therefore in some way related to the work presented in this thesis.

The relevant systems are therefore categorised into (1) algorithms proposed for the automated analysis and classification of chromosomes using predominantly *image processing* techniques (see Section 2.2) and (2) algorithms proposed for the analysis and classification of chromosomes using predominantly *machine learning* techniques (see Section 2.3).

Since most existing chromosome detection and classification systems have not been evaluated on the same datasets than those considered in this thesis, it is not possible to directly compare the reported proficiency of these systems to those proposed in this thesis.

2.2 Analysis and classification of chromosomes using predominantly image processing techniques

Moradi *et al.* (2003) tested the proficiency of landmarks such as the location of the centromere, the medial axis of a chromosome, as well as endpoints

and branching points of the medial axis. The medial axis is obtained using morphological thinning. The global minimum of the horizontal and vertical projections of the aligned chromosome indicates the location of the centromere. The proposed strategy was evaluated on 219 single chromosome images provided by the cytogenetic laboratory of the Cancer Institute, University of Tehran. The results for the proposed automated landmark detection protocol is in most cases in complete agreement with the cytogeneticist expert.

Markou *et al.* (2012) proposed an automated methodology for human chromosome classification based on basic image preprocessing techniques, morphological operations and Support Vector Machines (SVMs). The proposed strategy employs established image segmentation techniques in order to obtain individual chromosomes from an already karyotyped image. Karyotyping is the process of analysing a metaphase image of chromosomes by identifying and organising the chromosomes into their type, the number of chromosomes and abnormalities. Chromosome localisation is achieved using a median filter and contrast enhancement, followed by global thresholding to obtain a binary image. Morphological operations (closing and opening techniques) are implemented after a binary image is obtained. Standardisation of each chromosome is achieved by straightening out the chromosome and conducting a medial axis computation using morphological thinning and pruning. Curve smoothing is subsequently employed to improve the estimated accuracy of the curve's derivative. Polynomial extrapolation is applied at the two ends of the chromosome so that the medial axis extends the full length of the chromosome. Feature extraction is finally employed and the feature vector is classified using a SVM. The proposed strategy was evaluated on images captured at the Laboratory of Molecular Biology, 1st Department of Obstetrics and Gynecology, General Regional Hospital Papageorgiou, Thessaloniki (Greece).

For the purpose of classifying dicentric chromosomes in a metaphase image,

Rogan *et al.* (2014) employ (1) a gradient vector field (GVF), (2) discrete curve evolution (DCE) and (3) a histogram projection method (HPM) as feature extraction techniques. The GVF is used to produce a descriptive outline of the input chromosome after which DCE is applied to obtain the minimum polygon, which is subsequently pruned to create the centreline of the chromosome. The HPM constitutes a popular technique that is often used for the purpose of feature extraction. In classifying the centromere location in chromosome images a sensitivity of 85% and a specificity of 94% are reported.

Shirazi *et al.* (2016) presented a novel automated chromosome segmentation algorithm that involves the following techniques: image preprocessing, segmentation, and feature extraction, followed by the localisation of the centromere. The input chromosome image is first smoothed by a median filter, after which adaptive histogram equalisation is applied for contrast enhancement. An active contour model is employed for the purpose of moving the contours towards the outline of an object. Subsequently, Canny edge detection is used to obtain the prominent contours of the object. The chromosome image is rotated by determining the polynomial of degree one that best fits the medial axis, followed by boundary tracing in order to obtain the coordinates of the object. The proposed detection of the centromere location is carried out by searching for a local minimum in the width profile of a chromosome image. This is achieved by determining the distance from the upper to the lower part of the chromosome orthogonal to the medial axis. The authors investigate 311 chromosome images which are manually extracted from 80 different metaphase images. In order to test the proposed chromosome detection technique the authors created four different classes representing normal chromosomes, dicentric chromosomes, tricentric chromosomes and chromosome fragments, respectively. Detection accuracies of 96.7%, 87.6%, 60% and 93.8% are reported for the respective classes.

Jindal *et al.* (2017) proposed a chromosome classification strategy that involves preprocessing images utilising a straightening method and classifying chromosomes via Siamese networks. Straightening of the chromosomes is performed using Straightening via Medial Axis extraction and Crowdsourcing (SMAC) and Straightening via Projection Vectors (SPV). The use of SMAC is motivated by the fact that its goal is to extract the centreline of a chromosome using morphological thinning. The centreline extraction through morphological thinning forms the basis of this procedure, yet it is prone to error and thus has a crowd source component from untrained individuals to assist the algorithm in choosing the correct centreline. Crowdsourcing via Medial Axis is achieved when multiple individuals draw a centreline on each chromosome which is validated by spammers identification and consensus. When SVP is applied, a global minimum is found in the horizontal projection of a chromosome. The chromosome image is subsequently split into two subimages that are rotated and recombined to yield the straightened chromosome image. The authors concluded that SVP is superior to SMAC. Classification of chromosome images is achieved using a Siamese network and comprises of twin neural networks which determines whether input images are similar or dissimilar. The proposed SMAC in conjunction with a Siamese network, and SPV in conjunction with a Siamese network, are evaluated on 209 chromosome images resulting in accuracies of 80.4% and 85.2% respectively.

2.3 Analysis and classification of chromosomes using predominantly machine learning techniques

The research published by Sharma *et al.* (2017) aims to alleviate the work and cognitive load of domain experts when performing segmentation and classification tasks in the karyotyping process. They propose a method for automatic segmentation and classification of chromosomes for healthy patients

using a combination of crowdsourcing, preprocessing and deep learning. The automatic segmentation process utilises non-expert crowd sources to segment chromosome boundaries from metaphase images after which they are extracted for further processing. The ensuing preprocessing method entails chromosome length normalisation and the straightening of bent chromosomes which is motivated by a marked classification accuracy improvement in the deep neural network. A custom built, yet traditional, convolutional network is consequently used for classification purposes and consists of a standard categorical cross-entropy loss function which is optimised with stochastic gradient descent. As their objective was not to replace the domain expert, but to alleviate the associated cognitive burden of segmenting and karyotyping chromosomes, their crowdsourcing method shows promise. After stringent filtering and consensus steps, non-experts were able to identify on average 68.5% of available chromosomes per image. The efficacy of their automatic classification task also showed encouraging results despite relatively sparse training data, reporting a classification accuracy of 86.7% when used in conjunction with their preprocessing steps.

Sharma *et al.* (2018) proposed an end-to-end trainable Super-Xception network for automatic chromosome classification in low-resolution images. A Super-Xception network consists of two sub-networks, that is a super-resolution network and a classification network. The convolutional super-resolution layers enhance the resolution and recover the textural detail of the low-resolution images. The classification network uses layers of the Xception network to learn the feature representation of the images, while a softmax layer is employed to assign the labels. Each 50×50 grey-scale input image is first rescaled to the required resolution of 227×227 through interpolation. Length normalisation is then applied to every chromosome in the dataset. The Super-Xception network is subsequently applied to each of the chromosome images. The proposed

system was evaluated on the Bioimage Chromosome Classification dataset. By employing the proposed protocol for the purpose of classifying chromosome images, the accuracy of existing models was increased by 2.91%, resulting in an accuracy of 92.36%. Qin *et al.* (2019) developed a deep learning method to hasten the diagnosis procedure of karyotyping in abnormal diagnosis. Their novel approach, named Varifocal-Net, consists of a three step process towards classifying chromosome types and polarities simultaneously. Their first step consists of feeding preprocessed chromosome images into independent global and local feature learning networks. By leveraging the zoom capabilities of cameras they locate discriminative local regions and extract features on both the global and local scale. The second stage, built with two multi-layer perceptron classifiers, predicts from the two-scaled extracted features the chromosome's type and polarity. Their third stage adopts a dispatch strategy to allow assignment of each chromosome to a type, based on its predicted probabilities. Their method proves to be successful for clinical practice with healthy and unhealthy chromosomes accurately classified. Their method also allows for diagnosing numerical abnormalities if the number of classified chromosomes is irregular.

2.4 Conclusion

In the next chapter the biological background to the problem at hand is provided.

Chapter 3

Biological background

3.1 Introduction

In this chapter the biological background relevant to this research is briefly outlined. The primary objective of this discussion is to focus the reader's attention on (1) the structure of a chromosome, (2) how dicentric chromosomes are formed and (3) the stage at which a metaphase image is captured. This chapter begins with an introduction to the cell cycle (see Section 3.2), followed by a discussion of the chromosome itself (see Section 3.3).

3.2 The cell cycle

The cell cycle consists of a number of stages in which a cell passes from one cell division to the next. This process is illustrated in Figure 3.1. The cell cycle is divided into two main phases: (1) the interphase where the cell grows, duplicates DNA and prepares for division and (2) the mitotic phase where the division process of one cell into two identical daughter cells occurs (University of Leicester (2017)). The division of the cell nucleus is known as mitosis. Mitosis is divided into five stages: (1) the prophase, (2) the prometaphase, (3) the metaphase, (4) the anaphase and (5) the telophase.

Prophase: The nucleus envelope, that separates all the genetic material from the rest of the cell, begins to disperse. The sister chromotids begin to coil more tightly with the aid of proteins and become visible under a light microscope.

Prometaphase: The chromosomes migrate to the center of the cell. This region is identified as the metaphase plate. The chromosomes continue to

compress. Single spindle fibres wrap on each side of a structure associated with the centromere of each chromosome. This structure can be identified as the kinetochore.

Metaphase: At this stage the chromosomes are maximally condensed where they align themselves along the metaphase plate. This is the optimal stage to see the condensed chromosomes under a light microscope. *This is the critical stage within the context of the research conducted in this thesis.*

Anaphase: The connection between the sister chromotids breaks down, and the microtubules pull the chromosomes toward opposite poles.

Telophase: During this last stage the chromosomes reach the opposite poles and begin to unravel and decondense. Around each set of 46 chromosomes a nucleus envelope starts to form. In the telophase each chromosome has only one chromatid.

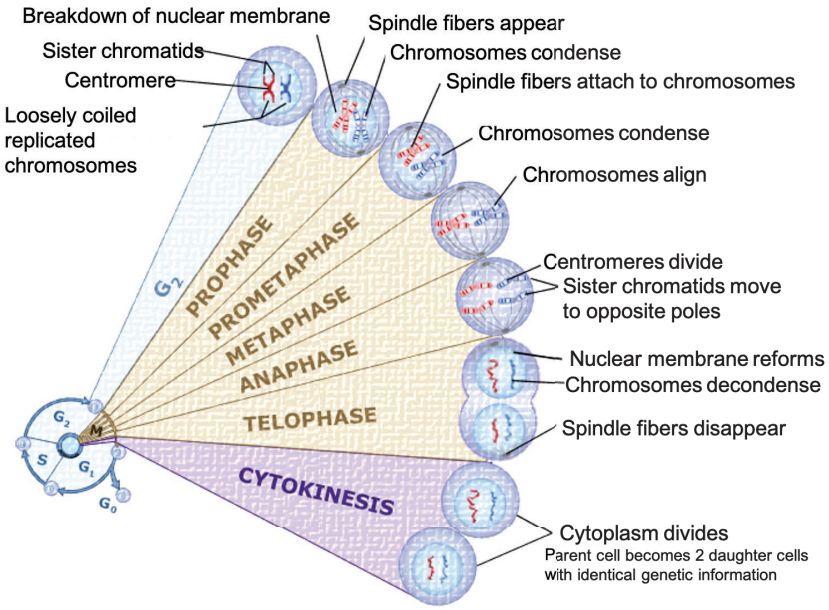


Figure 3.1: Conceptualisation of the five stages of mitosis in the cell cycle, that is (1) prophase, (2) prometaphase, (3) metaphase, (4) anaphase and (5) telophase. University of Leicester (2017)

3.3 The chromosome

The structure of a chromosome is illustrated in Figure 3.2. Chromosomes are packaged tightly only when a cell is going though mitosis. This structure increases the cell's durability when it is divided into two identical daughter cells. Other proteins do not accompany the cellular DNA. Various protein partners form a complex to help package the DNA into such a tiny space. This DNA-protein complex is known as chromatin. Within each cell a folded object forms a characteristic formation, which is called a chromosome. Along with the packaging proteins each chromosome has a single double-stranded DNA piece. As chromosomes undergo cell division they can be viewed through a light microscope. A cell being only a few micrometers wide can contain about two meters of human DNA, and is enabled by DNA packaging which assists in conserving space in cells. The chromatin within chromosomes are packed less tightly during the interphase when cells are not dividing. This important phase allows transcription to occur (O'Connor and Adams (2010)).

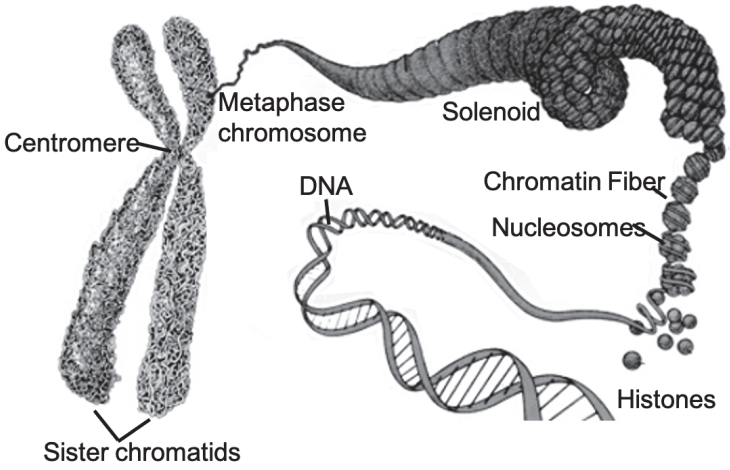


Figure 3.2: The structure of a chromosome. O'Connor and Adams (2010)

As is evident in Figure 3.2, a normal chromosome contains a single centromere where the two sister chromatids are joined. A chromosome with two centromeres is known as a dicentric chromosome which is formed when two chromosome segments, each with a centromere, fuse (Romm *et al.* (2013)). This results in the formation of a dicentric chromosome and acentric fragments (that lack a centromere). The formation of a dicentric chromosome is illustrated in Figure 3.3. (Figure 1.1 is reproduced here to provide context.)

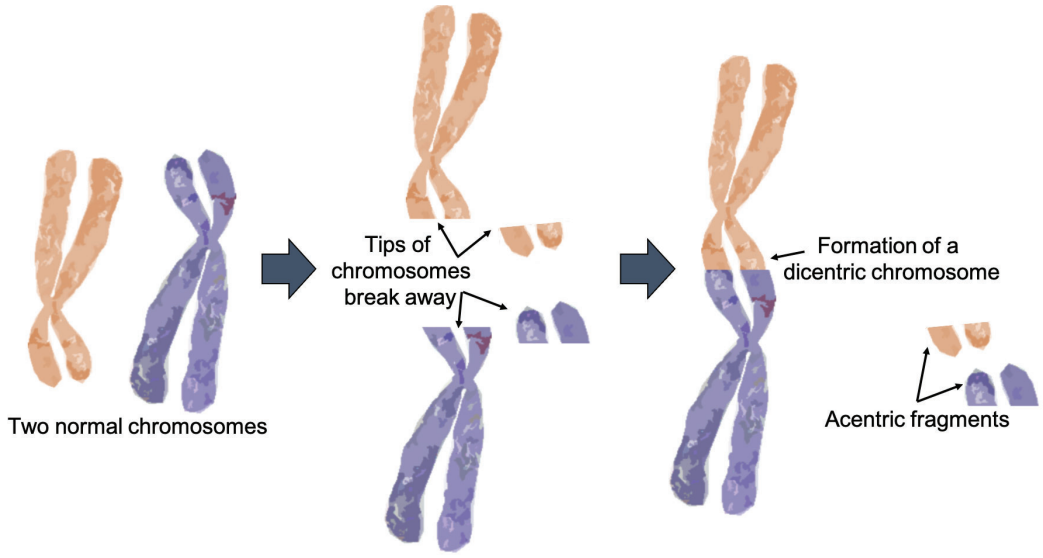


Figure 3.3: (Figure 1.1 is reproduced here to provide context.) Formation of a dicentric chromosome with accompanying acentric fragments.

3.4 Conclusion

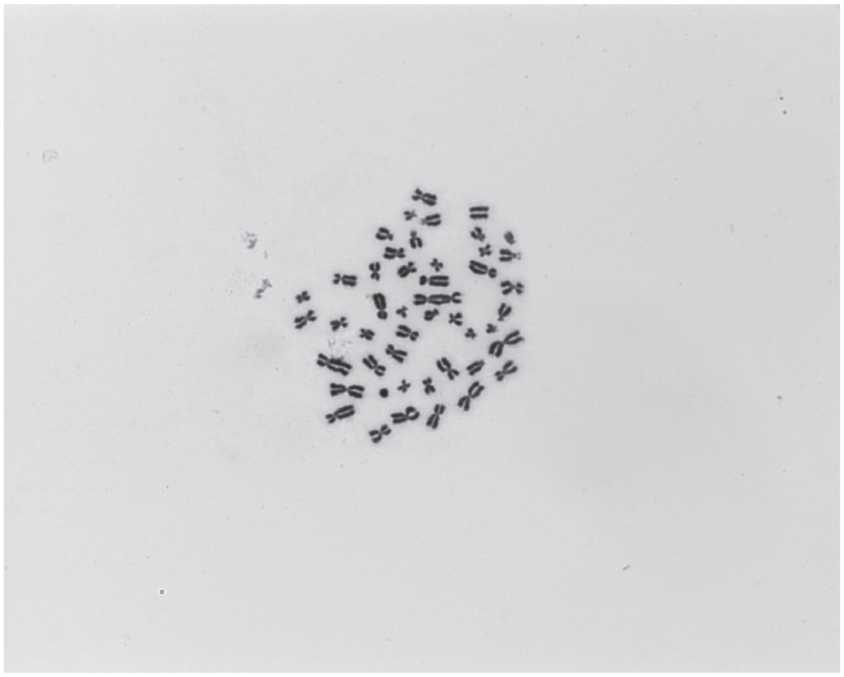
In the next chapter we discuss the novel single chromosome image segmentation protocol developed during the course of this study.

Chapter 4

Image segmentation

4.1 Introduction

For the isolation and extraction of chromosomes in metaphase images, a novel image segmentation strategy is proposed in this chapter. Since the research in this thesis is solely based on the iThemba LABS dataset, the proposed method is specific to the metaphase images produced in their laboratory. For reference purposes, a visual representation of three typical metaphase images obtained from iThemba LABS is shown in Figure 4.1.



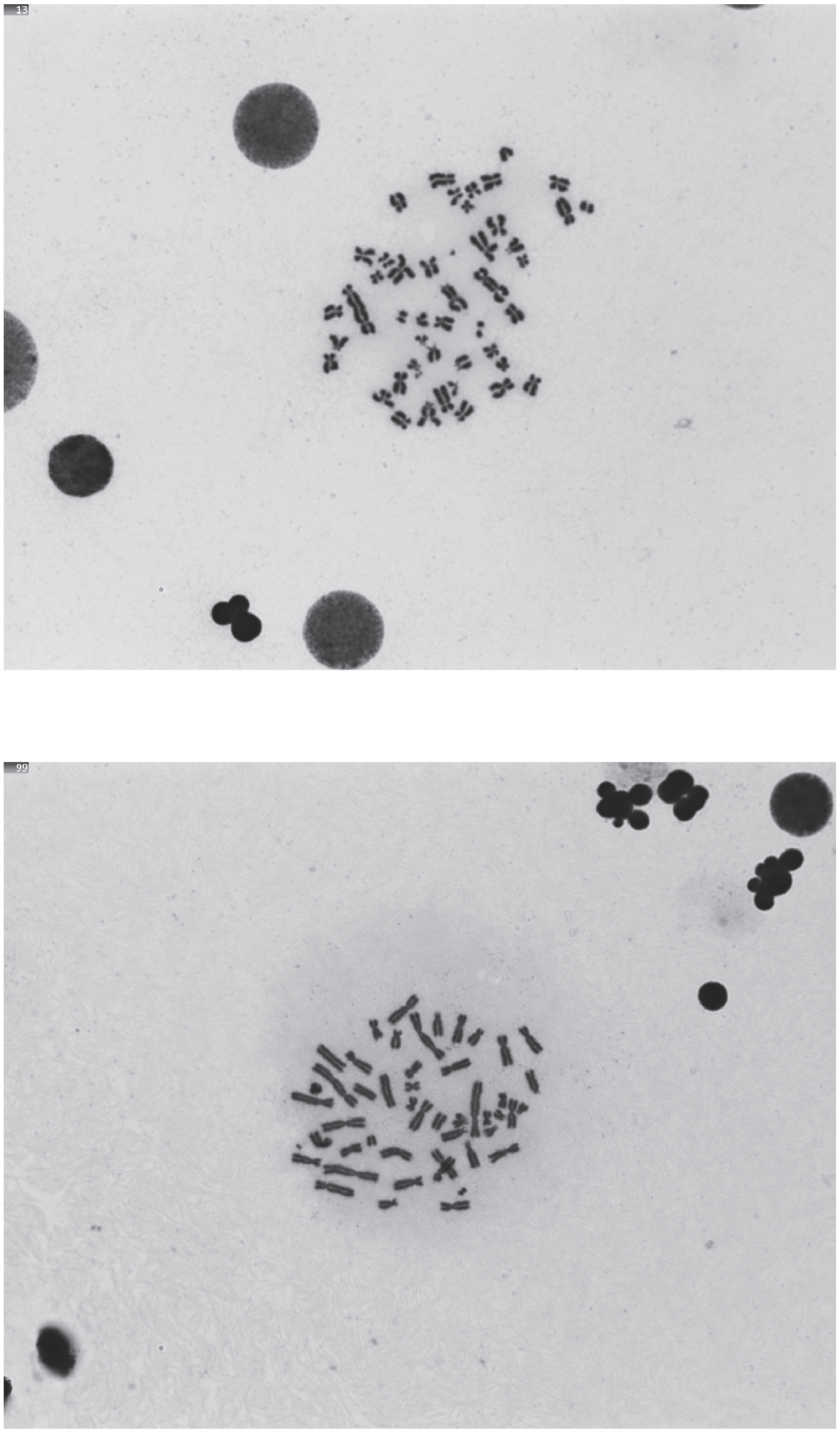


Figure 4.1: A visual representation of typical metaphase images obtained from iThemba LABS.

The image segmentation strategy will follow a principled approach to extract isolated normal as well as dicentric chromosomes. The procedure can be partitioned into the following three stages:

1. Preprocessing of metaphase images (see Section 4.2).
2. Extraction of regions of interest (ROIs) (see Section 4.3).
3. Manual extraction of isolated chromosomes from ROIs (see Section 4.4).

The proposed chromosome detection protocol is graphically conceptualised in Figure 4.2 and discussed in detail in the remainder of this chapter.

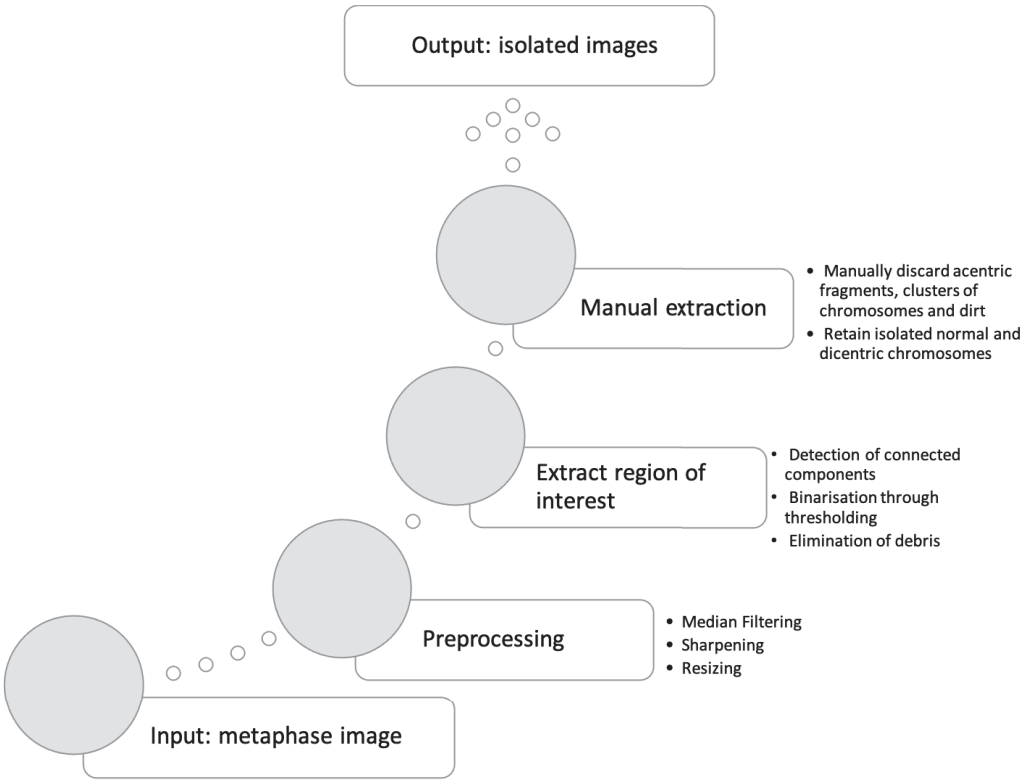


Figure 4.2: Conceptualisation of the protocol for detecting isolated normal and dicentric chromosomes as proposed in this research.

4.2 Preprocessing of metaphase images

Before extracting the ROIs, the metaphase images need to be first tailored and standardised using various preprocessing techniques. The primary purpose of the preprocessing procedure is to manipulate the grey-scale metaphase images provided by iThemba LABS in such a way that they better capture sharp details and remove unwanted distortions. Using specific image processing methods, a systematic approach to reduce noise and blurriness in the input images is applied, after which the images are resized. As previously stated, these preprocessing steps are required for the successful extraction of ROIs from metaphase images and are discussed below.

In order to remove any potential noise in metaphase images, some form of smoothing technique is required. A median filter is chosen as a well suited order-statistical filter for suppressing the effect of salt and pepper noise in metaphase images since it ensures a minimal definition loss when compared to alternative linear smoothing filters of the same size. When a median filter is applied to an image, the image is smoothed by replacing the value of every pixel with the median of the intensity levels in its neighbourhood. The process can be formulated as follows,

$$\hat{f}(x,y) = \text{median}_{(s,t)\in S_{x,y}}\{g(s,t)\},$$

where $S_{x,y}$ represents the set of all the pixels in a selected image window of size 3×3, with the center of the set indicated by the point (x,y). The function takes the median of the original image $g(x,y)$ in the area designated by $S_{x,y}$. The value of the restored image $\hat{f}$ at point (x,y) is the computed arithmetic median using the pixels in the designated region $S_{x,y}$ (Gonzalez *et al.* (2010)).

A proven process to sharpen images consists of subtracting the features obtained by employing a smoothed or unsharpened mask from the original

image. This process, called unsharp masking (Gonzalez *et al.* (2010)), proceeds as follows:

1. Blur the original image as denoted by $\bar{f}(x, y)$.
2. The mask is obtained by subtracting the blurred image $\bar{f}(x, y)$ from the original image $f(x, y)$,

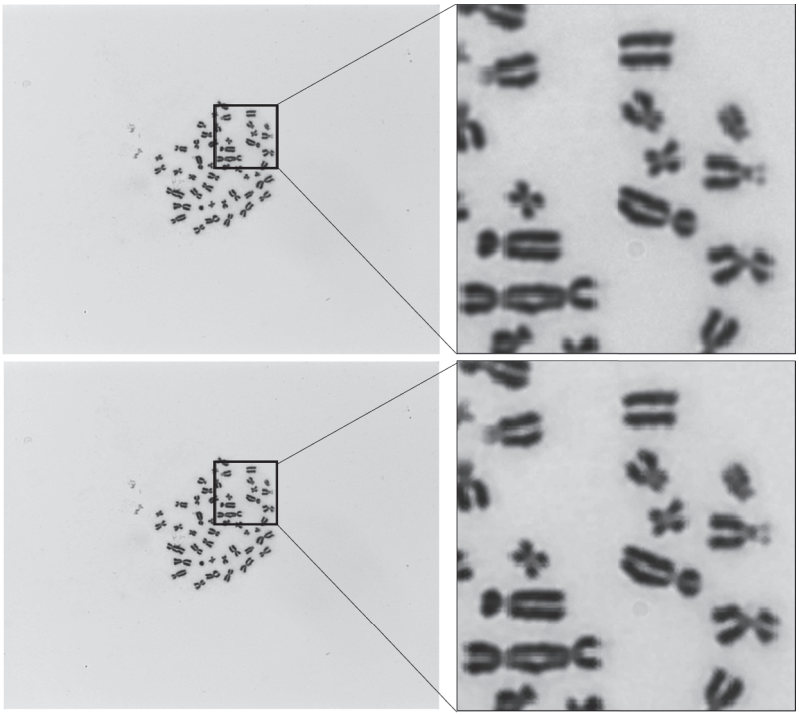
$$g_{\text{mask}}(x, y) = f(x, y) - \bar{f}(x, y).$$

3. The sharpened image is obtained by adding a weighted portion, w ($w \geq 0$), of the acquired mask to the original image,

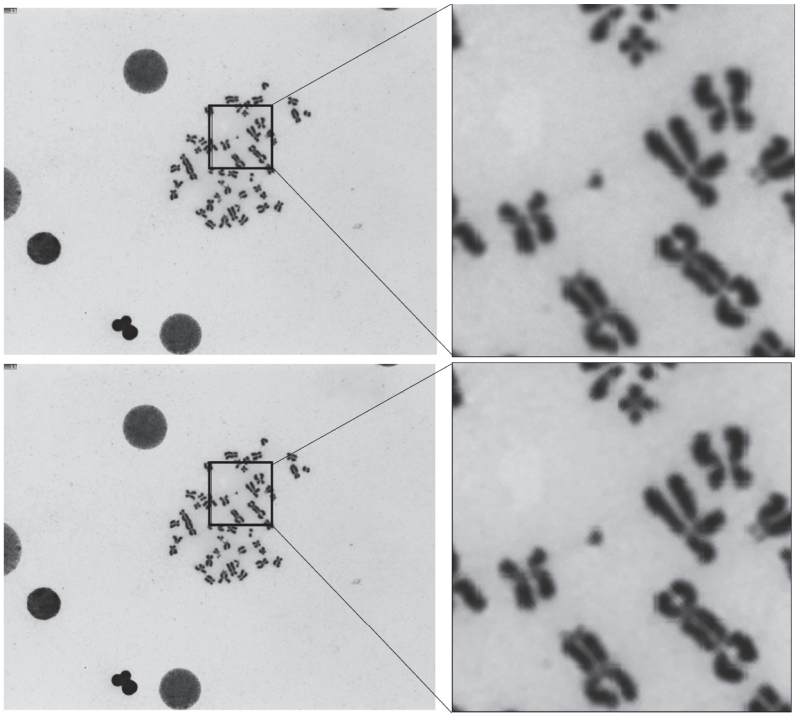
$$g(x, y) = f(x, y) + w * g_{\text{mask}}(x, y),$$

where $*$ denotes multiplication. The weight w therefore determines the degree of sharpening. In general when images are sharpened by employing a weight of $0 \leq w < 1$ the process is referred to as de-emphasising since the sharpness effect is mitigated. When images are sharpened by employing a weight of $w = 1$ the process is referred to as *unsharp masking* whilst the process is referred to as *highboost filtering* when $w > 1$ (Gonzalez *et al.* (2010)). The proposed image sharpening strategy employs *unsharp masking* by setting $w = 1$.

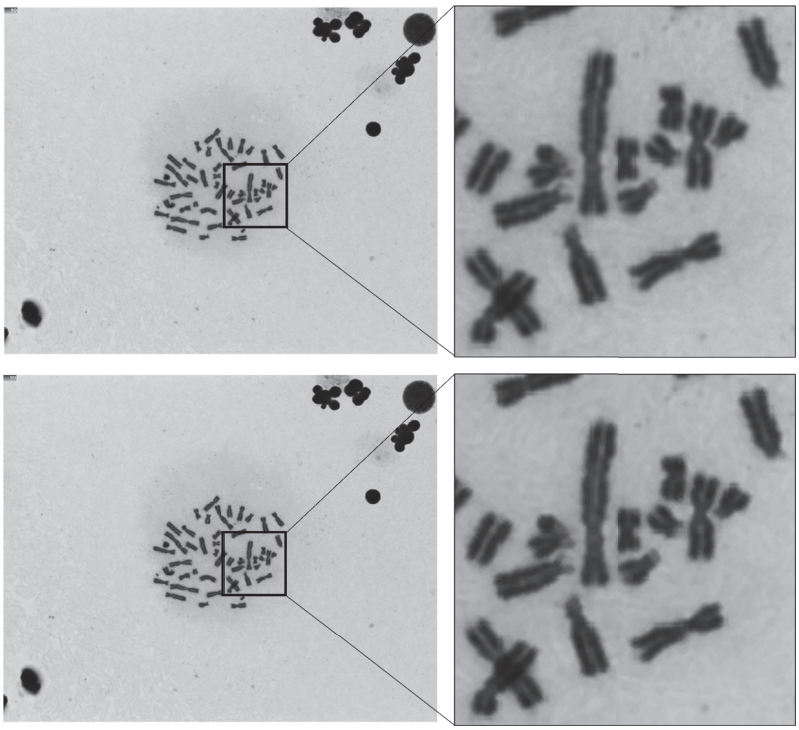
The preprocessing of an input metaphase image is therefore achieved by applying a *median filter* of size 3×3, and with the weight specified as $w = 1$ for the purpose of achieving unsharp masking. These steps ensure that sufficient noise and non-prominent edges are removed while preserving critical features typically associated with the curved edges of chromosomes. For reference purposes, the effect of the proposed preprocessing strategy when applied to three of the metaphase images from the iThemba LABS dataset is illustrated in Figure 4.3.



(a)



(b)



(c)

Figure 4.3: Preprocessing. (**Top**) Input metaphase images. (**Bottom**) Smoothed and sharpened versions of the corresponding images on the top after the application of the median filter and unsharp masking.

4.3 Extraction of ROIs

After performing the necessary preprocessing steps, the regions of interest (ROIs) can be determined. As is evident in Figure 4.1, the prominent chromosomes are found near the center of the metaphase images. Other objects are however also present. Keeping in mind that the objective is to detect single isolated chromosomes, all the possible objects found in a typical metaphase image are categorised as follows:

1. Isolated normal chromosomes (see Figure 4.4 (a)).
2. Isolated dicentric chromosomes (see Figure 4.4 (b)).
3. Acentric fragments (see Figure 4.4 (c)).
4. Clusters of chromosomes (see Figure 4.4 (d) and (e)).

5. Dirt (see Figure 4.4 (f)).

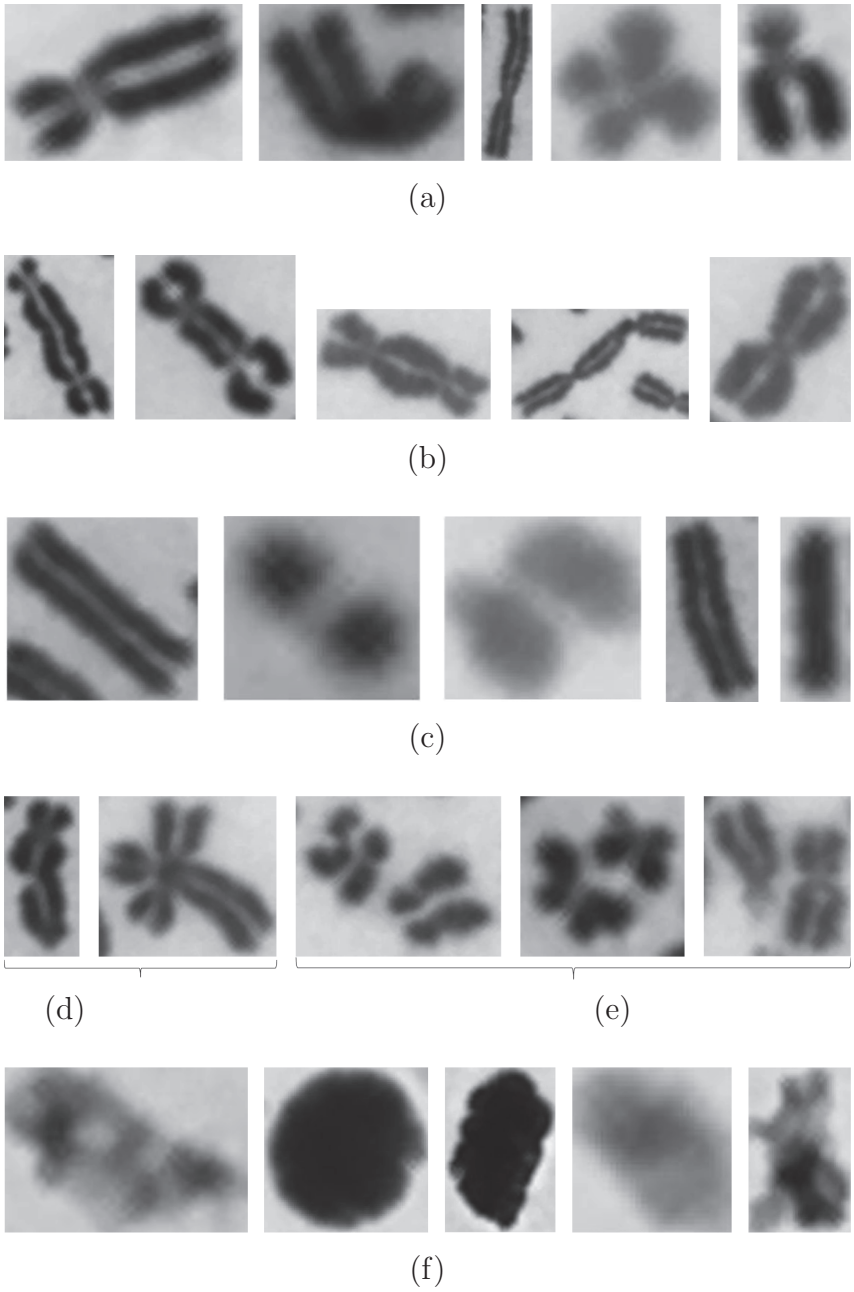


Figure 4.4: Visual representation of objects of interest (see (a) to (e)) and dirt (see (f)). (a) Isolated normal chromosome. (b) Isolated dicentric chromosome. (c) Acentric fragments. (d) Clusters of overlapping chromosomes. (e) Clusters of chromosomes in close proximity of one another. (f) Dirt.

Note that the aforementioned clusters may contain more than one chromosome that can either be located on top of one another (see Figure 4.4 (d)), or reside in very close proximity to one another (see Figure 4.4 (e)).

Each candidate ROI is evaluated in order to determine whether it constitutes a viable object of interest. In order to assist in determining these viable objects, the following information is considered: (1) prior knowledge such as the fact that the expected number of chromosomes within a normal cell is 46 and (2) the fact that certain artifacts such as noise (a component that is too small) or dirt (a component that is too large) (see Figure 4.4 (f)) may be miss-classified as an ROI. The steps in defining the ROI within the context of the iThemba LABS dataset constitutes the following manageable tasks:

1. The detection of connected components (see Section 4.3.1).
2. The elimination of debris (see Section 4.3.2).
3. The extraction of ROIs and final segmentation (see Section 4.3.3).

4.3.1 Detection of connected components

The detection of connected components requires a heuristic approach towards automatically estimating a suitable global binarisation threshold. Once an optimal global binarisation threshold has been determined, debris are removed from the resulting binary images and suitable ROIs are extracted. The objective of the binarisation threshold is to create binary images in such a way that the targeted foreground objects are separated from the background. The objective is therefore to set pixel values within the grey-scale metaphase input image $f(x, y)$ that exceed a fixed threshold to a binary value of 1 (rendered white), whilst the remainder of the pixels are assigned a binary value of 0 (rendered black). Once a binary image has been produced, the connected components can be determined. Formally a connected component can be described as a set of 8-connected white (1-valued) pixels within a binary image.

As reference, Figure 4.5 shows the conceptualisation of the proposed protocol for detecting connected components as well as all other necessary steps to produce the final binary image from a metaphase image.

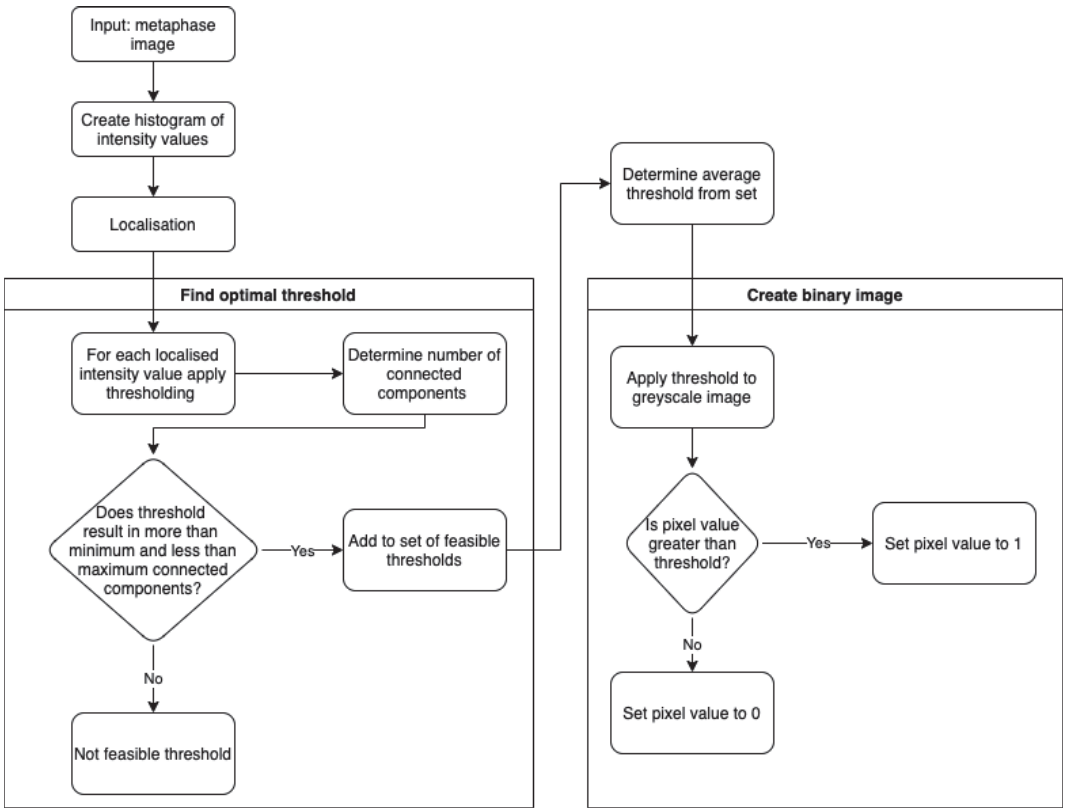


Figure 4.5: Conceptualisation of the novel heuristic binarisation method proposed in this thesis.

As indicated in Figure 4.5, a histogram of the intensity values of the image in question is first constructed before a threshold is applied. Given a metaphase image with intensity levels (pixel values) ranging between 0 and 255, a histogram can be constructed using the function $h(r_k) = q_k$ where r_k is the kth intensity in the given range whilst q_k is the number of pixels in the grey-scale image with intensity r_k.

From a created histogram, various thresholds in the range $[0, 255]$ can be tested for binarisation purposes, but due to computational costs this proves to be impractical. Through observation, however, it has been noted that lower valued intensities tend to contain objects of interest while higher valued intensities contain background elements. To facilitate these observations and constraints, a localisation approach is adopted to find thresholds within a subset of the full range. In order to obtain the localised set, the local minimum is calculated using the following equation,

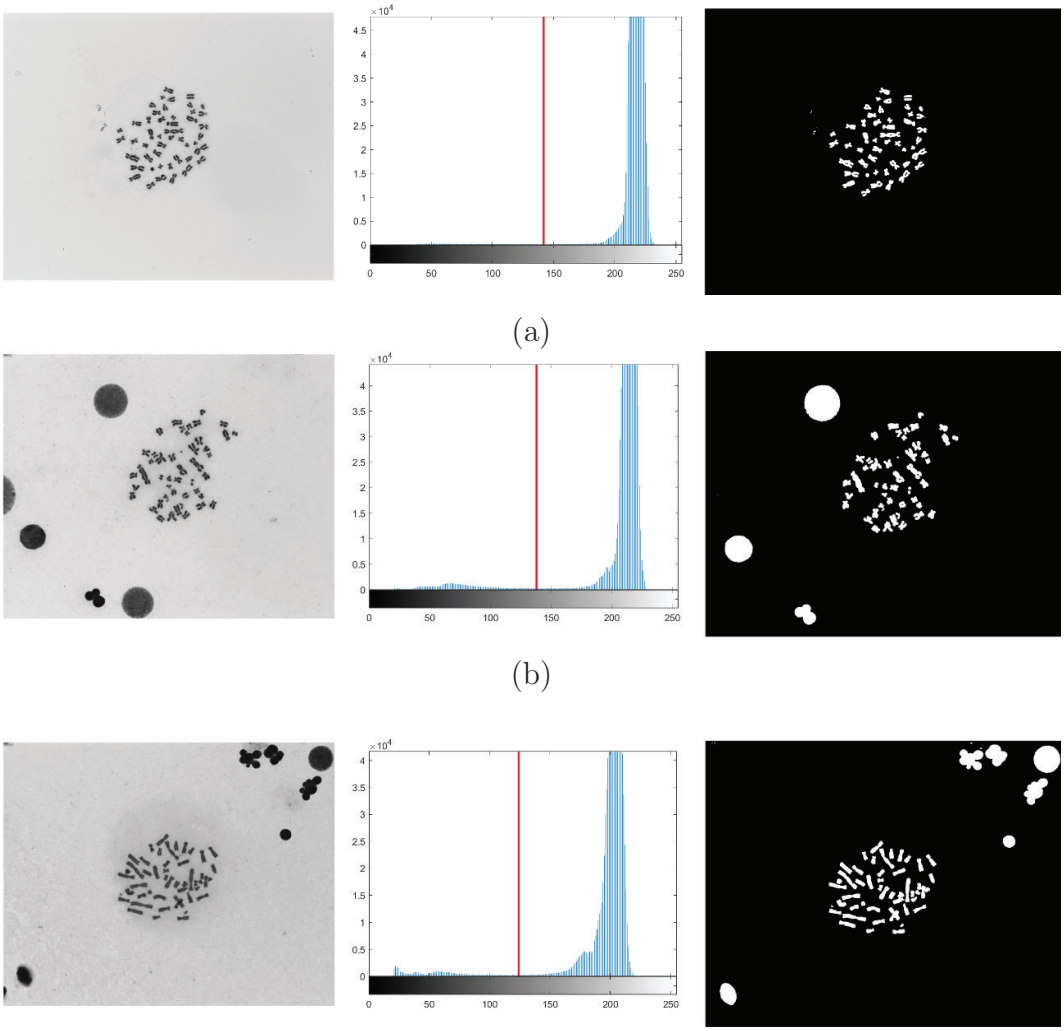
$$\text{localmin}(h(r_j) = q_j), \text{ where } j \in [100, 180].$$

Through observations, the interval $[100, 180]$ is deemed appropriate.

Recall that a threshold of a specific localised intensity value will produce a number of connected components within a binary image. A feasible set of thresholds can subsequently be determined. More specifically, when the application of a threshold results in a total number of connected components that is greater than the minimum and less than the maximum number of allotted objects per image, the threshold is deemed feasible. If the total number of connected components do no meet these criteria, the threshold is discarded. This process is iteratively executed for the entire set of localised intensity values until the set of all feasible thresholds is obtained. Finally the most suitable global binarisation threshold is obtained by taking the average of all the feasible thresholds.

In order to obtain the desired binary image, the most suitable global threshold is applied to the grey-scale metaphase image in question, separating the foreground objects from the background. The principles involved in the novel binarisation method proposed in this thesis is graphically conceptualised in Figure 4.6. It is important to note that the proposed binarisation method

produces significantly superior results to that of the well-known Otsu method (Otsu (1979)).



(a)

(b)

(c)

Figure 4.6: (**Left**) A grey-scale metaphase image. (**Centre**) The histogram of the image on the left, where the location of the appropriate threshold value as determined by the novel binarisation method developed in this thesis is denoted by the red line. (**Right**) The binary image obtained after the appropriate threshold is applied to the image on the left.

4.3.2 Elimination of debris

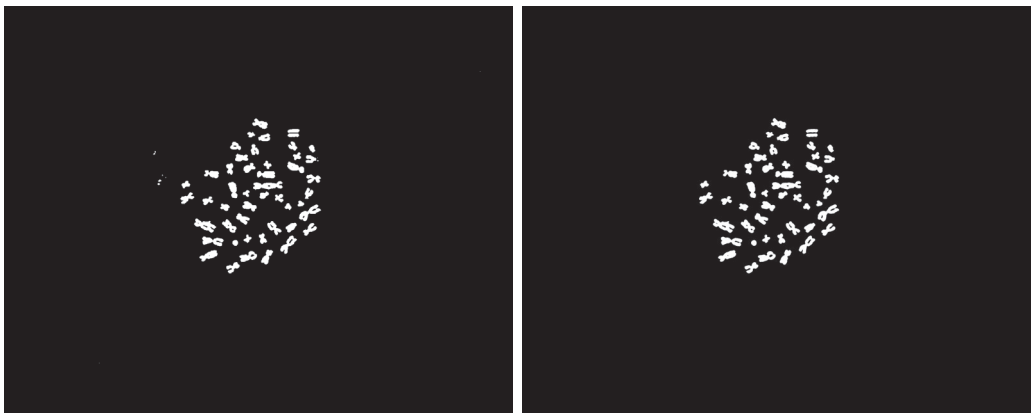
By considering the binary image produced using the strategy outlined in Section 4.3.1, each connected component is assessed to determine whether it is deemed too small (noise) or too large (dirt) by counting the number of pixels in the connected component in question. Therefore, if the connected component falls within this category it is tagged as debris and eliminated. The debris are subsequently deemed part of the background of the binary image in question (see Figure 4.7).

4.3.3 Extraction of ROIs and final segmentation

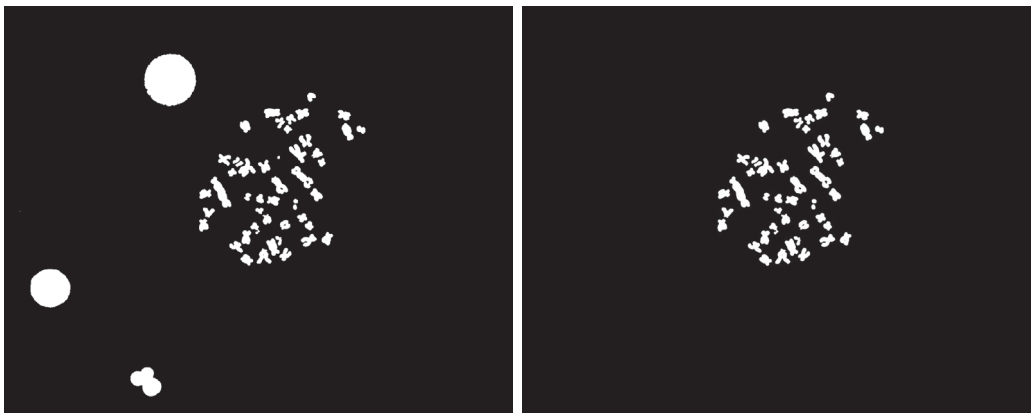
The remaining connected components that are deemed *not* to be debris are each individually extracted to form a set of ROIs that are suitable for further analysis. In addition to this, the coordinates of the ROIs are stored to assist in generating the ground truth as outlined in Section 6.3. This is illustrated in Figure 4.8.

4.4 Manual extraction of isolated chromosomes

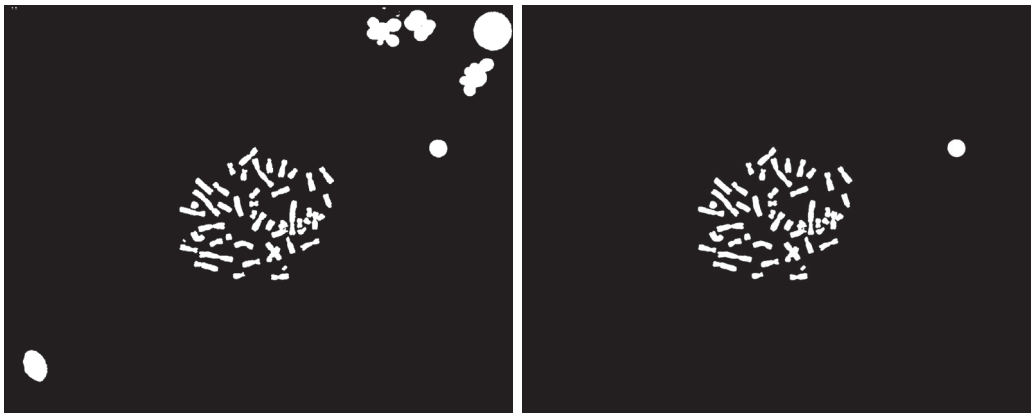
Given the set of individual ROIs (see Figure 4.8 (right)), each ROI is *manually* categorised into dirt, acentric fragments, clusters of chromosomes and isolated chromosomes (see Figure 4.4). All acentric fragments, clusters of chromosomes and dirt are *manually* discarded, while *only* the isolated chromosomes are retained. The automated classification of acentric fragments and clusters of chromosomes falls outside the scope of this research study. The automated categorisation of the isolated chromosomes as either normal or dicentric is however investigated in the remainder of this thesis. It is also important to note that, in Figure 4.8, the process of extracting the ROIs is performed *automatically* from the relevant binary image on the left.



(a)

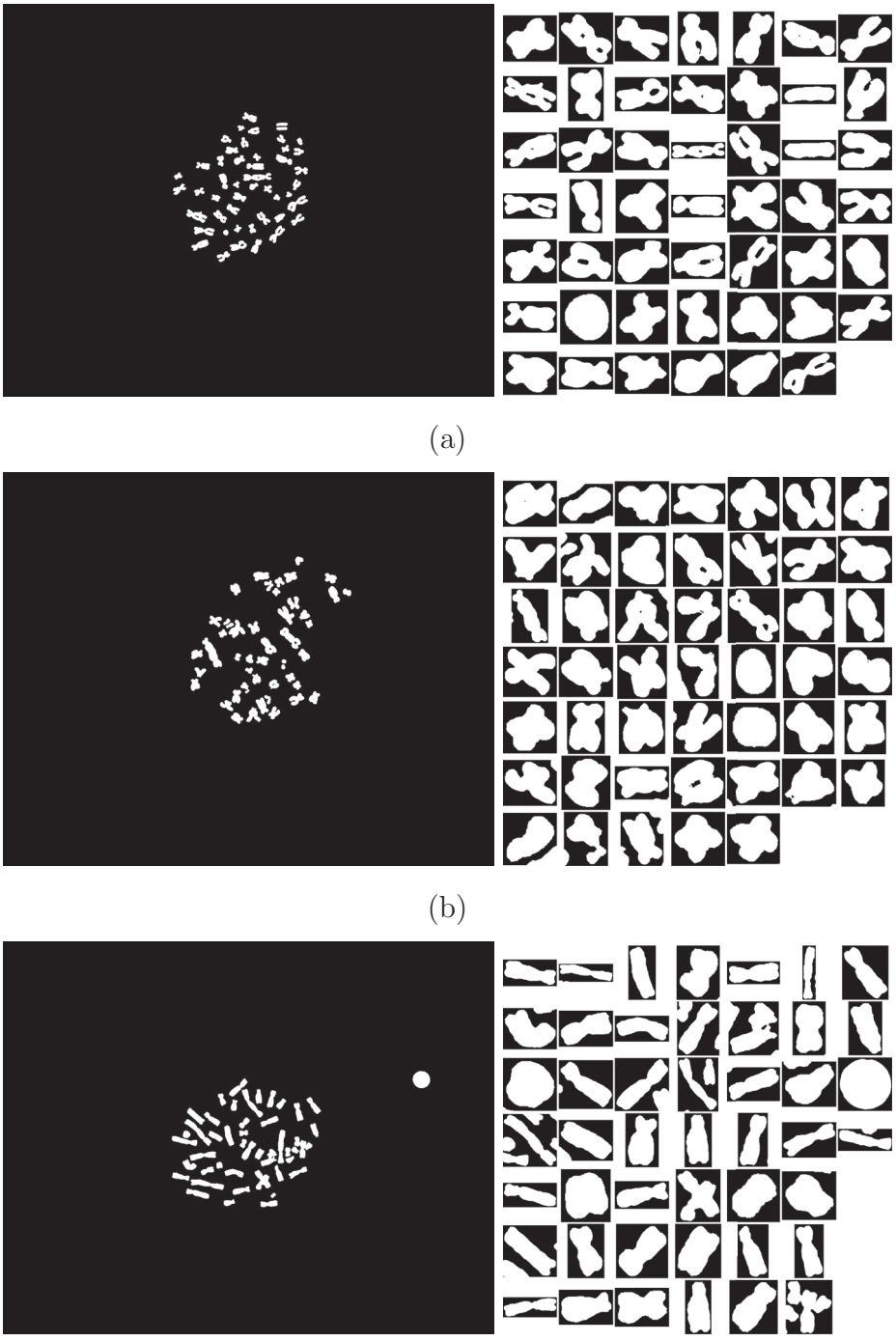


(b)



(c)

Figure 4.7: (**Left**) Binary image obtained using the protocol outlined in Section 4.3.1. (**Right**) Final binary image after the removal of debris.



(a)

(b)

(c)

Figure 4.8: (**Left**) Final binary image obtained using the protocol outlined in Section 4.3.2. (**Right**) Extracted ROIs.

4.5 Concluding remarks

In this chapter a segmentation protocol was developed, which facilitates the automatic detection of a region of interest (ROI) that encloses normal chromosomes, dicentric chromosomes, acentric fragments and clusters of chromosomes. The aforementioned protocol employs several preprocessing techniques and a novel binarisation algorithm, which is followed by the manual extraction of isolated normal and dicentric chromosomes. The proposed feature extraction protocol is discussed in detail in the next chapter.

Chapter 5

Feature extraction

5.1 Introduction

The automated detection of dicentric chromosomes within metaphase images can provide valuable information towards determining the extent of exposure to radiation. In this chapter novel strategies for classifying chromosomes as either normal or dicentric are investigated. The objective of these strategies is to put into place a protocol that uses shape information from isolated chromosomes (as obtained through the protocol outlined in the previous chapter) to automatically classify the chromosomes as normal or dicentric based on the extracted features.

The proposed feature extraction protocol will only focus on classifying *isolated* normal and dicentric chromosomes, and will disregard clusters and acentric fragments. For reference, grey-scale representations of a typical isolated normal and dicentric chromosome cropped from metaphase images are shown in Figure 5.1.

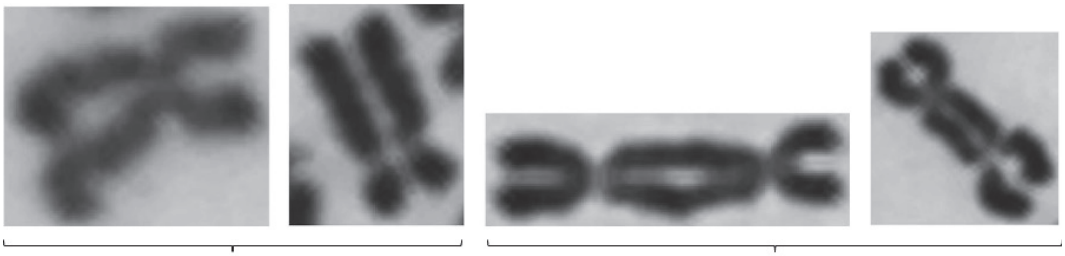


Figure 5.1: Grey-scale representations of isolated chromosomes. (a) Typical normal chromosomes. (b) Typical dicentric chromosomes.

Figure 5.2 depicts the resultant binary images obtained following the proposed image segmentation protocol outlined in the previous chapter.

Figure 5.2: Binary representations of isolated chromosomes after employing the segmentation protocol outlined in Section 4.4. (a) Typical normal chromosomes. (b) Typical dicentric chromosomes.

Recall from Section 3.3 that the thinnest part of a chromosome constitutes a centromere where two sister chromatids are joined together. The important distinguishing feature of isolated normal and dicentric chromosomes is the presence of one or two centromeres respectively. Subsequently, determining the *number* of centromeres associated with a chromosome is a critical step in classifying the chromosome as either normal or dicentric. It is clear from Figure 5.2 that a centromere exhibits two distinct characteristics: (1) It constitutes the thinnest region of a chromosome. (2) The segment of the edge of the chromosome in the region of the centromere is significantly more concave towards the outside than the remainder of the edge. These two fundamental features associated with the location of a centromere within a chromosome can be identified by considering the chromosome's width profile and conducting curvature analysis on the chromosome's edge respectively.

Overall, the feature extraction process can be divided into the following stages:

1. Preprocessing for image analysis (see Section 5.2).

2. Width profile analysis (see Section 5.3).

3. Curvature analysis (see Section 5.4).

The proposed chromosome classification protocol is graphically conceptualised in Figure 5.3 and discussed in detail in the remainder of this chapter.

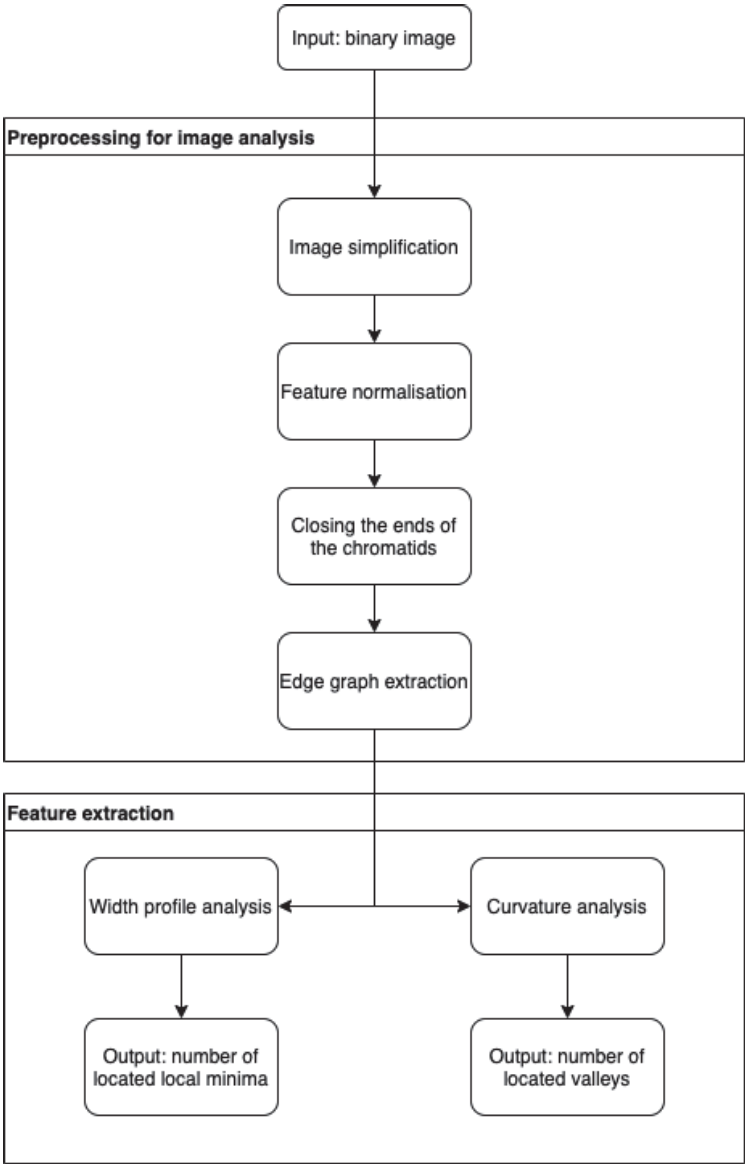


Figure 5.3: Conceptualisation of the novel feature extraction protocol proposed in this thesis.

5.2 Preprocessing for image analysis

The preprocessing protocol proposed in this section is based upon modifying the binary image of an isolated chromosome obtained in Section 4.4. Before feature extraction can be conducted, the binary image is modified in such a way that a normalised edge graph with only the outside boundary of the primary chromosome is obtained. The steps in obtaining the edge graph is divided into the following manageable tasks:

1. Image simplification (see Section 5.2.1).
2. Feature normalisation (see Section 5.2.2).
3. Closing the ends of the chromatids (see Section 5.2.3).
4. Edge graph extraction (see Section 5.2.4).

In order to serve as a visual guide, the proposed preprocessing protocol is graphically conceptualised in Figure 5.4.

5.2.1 Image simplification

In order to modify each binary chromosome image obtained from Section 4.4, morphological transformations need to be applied to isolate, fill and smooth the chromosome in question. These steps are necessary to ensure that a successful feature normalisation and extraction protocol can be applied. In order to obtain the edge graph that is required for width profile analysis and curvature analysis, the borders of the binary image must first be cleared. The fundamental task of border clearing is used to allow only complete objects within an image to be extracted. This step is therefore crucial in ensuring that partial objects (that touch the boundary of the binary image) are removed so that the resulting image only contains a single isolated object associated with a full chromosome.

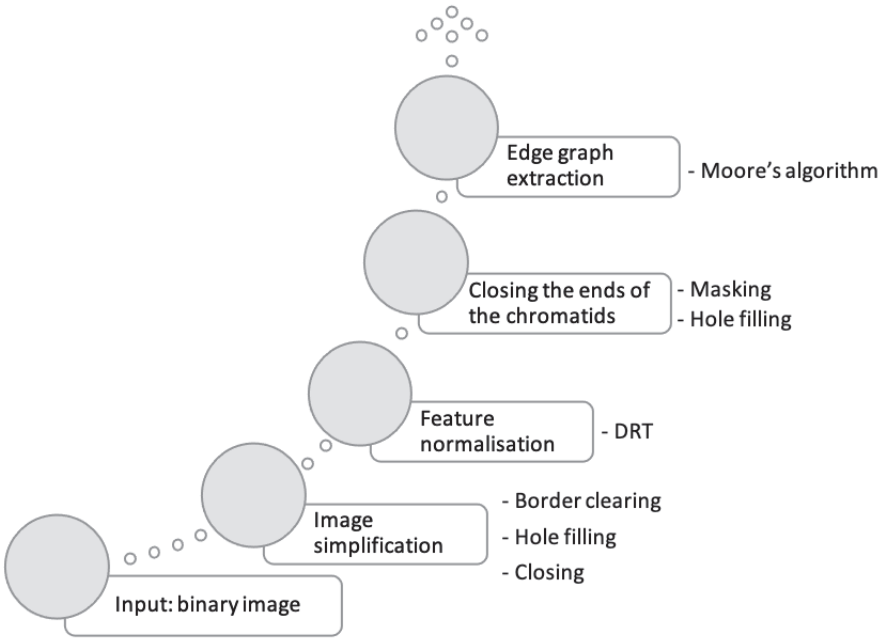


Figure 5.4: Conceptualisation of the preprocessing for image analysis protocol as proposed in this research.

After border clearing, further simplifications on the chromosome object itself may be required. Investigations have shown that irregularities such as holes may be present within isolated objects. In order to remedy this, a commonly employed morphological algorithm, known as hole filling (Gonzalez *et al.* (2010)) is applied to detect and fill holes within the binary images. Holes may be defined as any background region surrounded by a connected component in an image. This step ensures that each chromosome is fully connected and consistent throughout.

As a result of the image capturing process, small gaps and inconsistencies may also be present along the boundaries of the isolated chromosome. A type of smoothing technique known as morphological closing reduces the raggedness of the edges. In general, morphological operations such as closing employs an

appropriately shaped and sized sub-image, known as a structuring element (SE) to determine regions that should be retained or filled along an object boundary in order to render it smoother. By placing and progressively moving a SE along the outside boundary of a connected component, any regions the SE does not fit into, or make contact with are subsequently filled. The reader is referred to page 657 of Gonzalez *et al.* (2010) and page 23 of Beukes (2018) for a more detailed description of the morphological closing operation.

The step by step application of the above-mentioned image simplification protocol is illustrated from left to right in Figure 5.5.

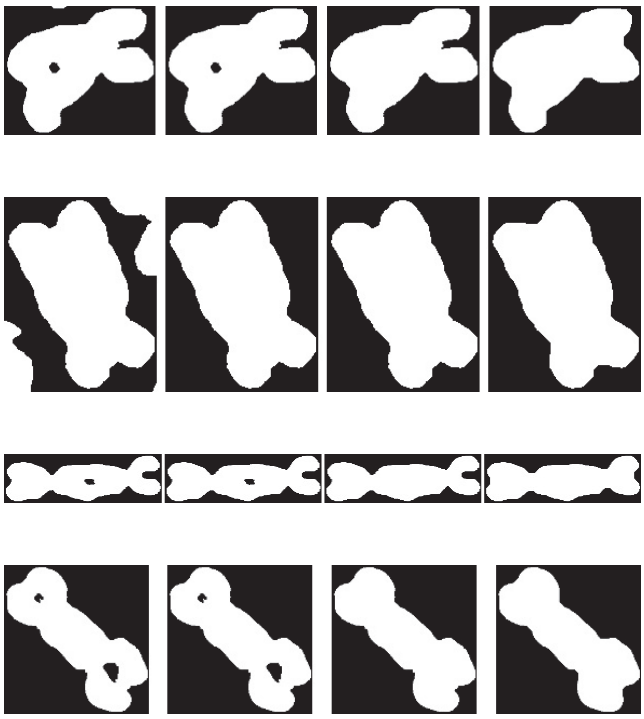


Figure 5.5: A visual representation of the progression of the binary image simplification protocol. **(Column 1)** Original binary input images obtained in Section 4.4. **(Column 2)** Binary image obtained after border clearing is applied. **(Column 3)** Binary image obtained after hole filling is applied. **(Column 4)** Binary image obtained after morphological closing is applied.

5.2.2 Feature normalisation

After the morphological transformations outlined in the previous section have been applied, isolated chromosomes are further modified using feature normalisation techniques. The techniques' primary goal is to ensure rotational invariance between all chromosome objects. This feature normalisation protocol assists in the process of detecting centromeres, since the proposed feature extraction protocol (which is discussed at a later stage) will require objects to be aligned along a specific axis, namely the horizontal axis. By applying a discrete Radon transform (DRT) on an input image, the appropriate alignment for rotational invariance of a chromosome can be determined and enforced.

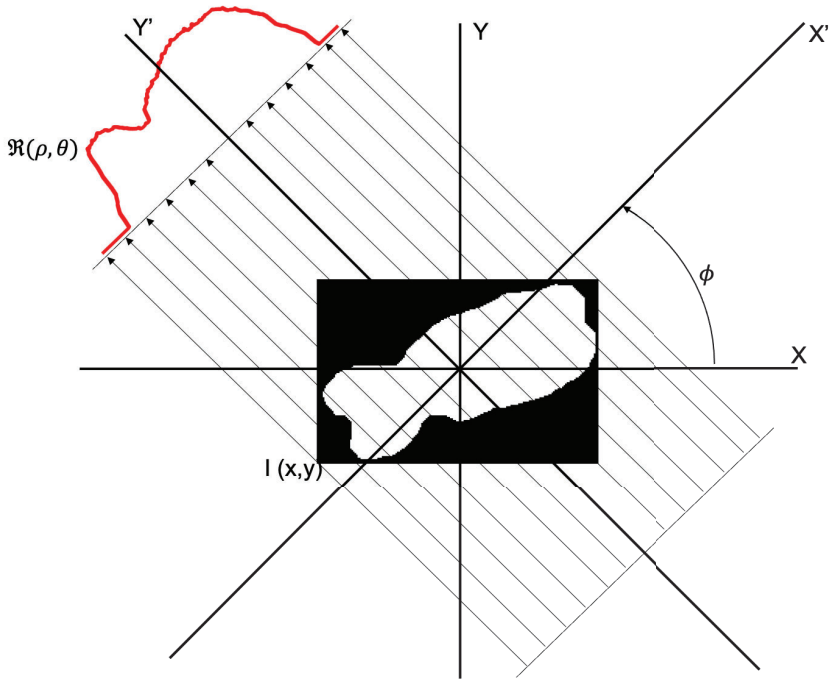


Figure 5.6: The application of the discrete Radon transform to a typical binary isolated chromosome image is conceptualised.

The DRT of an image is obtained when multiple, parallel-beam projections of an image is calculated from equally distributed angles within an interval $\phi \in [0°, 180°)$ (Coetzer (2005)). Specifically, each projection ρ constitutes a vector, of which each component represents a beam-sum. Within the context of the system proposed in this thesis, each projection component approximates the number of chromosome pixels within the relevant beam. The DRT $\Re(\rho, \phi)$ of an input image, $I(x, y)$, therefore constitutes a matrix where each column, ρ, of the DRT represents a projection profile of the input image, acquired from a specified angle, ϕ. The calculation of a projection profile of a typical binary isolated chromosome image from a specific angle ϕ is conceptualised in Figure 5.6.

Using the projection profiles obtained from calculating the DRT we can therefore determine different bounding boxes by considering the properties of two perpendicular projection profiles as will be explained shortly. As Figure 5.7 shows, creating the most compact bounding box around a chromosome encapsulates its rectangular nature and represents the smallest possible area that contains a chromosome.

Figure 5.7: The targeted most compact bounding box manually annotated and superimposed in red onto an isolated normal chromosome.

Therefore the most compact bounding box minimises the number of background (black) pixels relative to the number of foreground (white) pixels. The protocol for calculating the most compact bounding box and in turn determining the appropriate degree of rotation for rendering the chromosomes rotational invariant is outlined below. Formally, let

$$\Re\,'(\rho,\phi) = \begin{cases} 1, & \Re(\rho,\phi) > 0 \\ 0, & \text{elsewhere} \end{cases}$$

define the piecewise function which transforms beam-sums of $\Re$ to a binary function. The total number of non-zero beam-entries from a projection profile can subsequently be calculated as follows,

$$l(\phi) = \sum_{i=1}^{m} \Re\,'(\rho_i,\phi), \quad \text{where } i \in [1,m] \text{ and } \phi \in [0,180).$$

Therefore l represents the maximum variation of the connected component for each angle ϕ and m denotes the total number of beams per angle. Taking into consideration the fact that the area of a rectangle is defined by the product of its dimensions, the most compact bounding box can be determined using the maximum variation l. By considering perpendicular pairs from the various orientations within the interval $[0,180)$, the bounding box area for each possible chromosome orientation can be calculated. Formally this can be written as $\phi_1 \perp \phi_2$ where $\phi_1 \in [0,90)$ and $\phi_2 \in [90,180)$ with the area a defined as

$$a(\alpha) = l(\phi_1) \cdot l(\phi_2) \text{ where } \alpha \in [0,90).$$

Subsequently, finding the minimum area a will result in the most compact bounding box around the object. The parameter α now represents the angle through which the object has to be rotated so that it is aligned horizontally.

In scenarios where more than one value of α is associated with a minimum area a, the average of these values constitutes the required angle of rotation that will ensure horizontal alignment.

Given the fact that the bounding box of an object is orientated along a specific axis, it should be noted that the same area can be produced by rotating the object through 90°. The final step towards determining the degree of rotation is to ascertain whether the major axis is orientated horizontally or vertically after the rotation has been performed. Should the major axis be orientated vertically, the object is simply rotated trough another 90° in order to ensure that it is orientated horizontally. The rotational invariant versions of the images obtained in Section 5.2.1 are depicted in Figure 5.8.

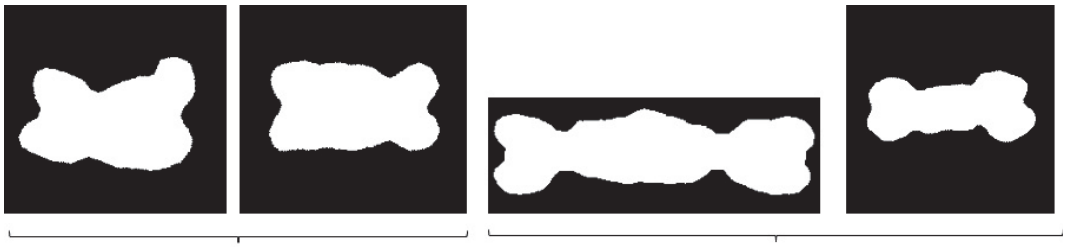


Figure 5.8: Rotational invariant versions of the images obtained in Section 5.2.1 (a) Typical normal chromosomes. (b) Typical dicentric chromosomes.

Note that for some cases where the chromosomes are significantly bent the proposed protocol may fail to guarantee rotational invariance. Straightening of bent chromosomes should rectify this issue (Jindal *et al.* (2017), Markou *et al.* (2012) and Sharma *et al.* (2017)), but this strategy does not fall within the scope of this thesis.

5.2.3 Closing the ends of the chromatids

In order to ensure that the centromere detection protocol proposed later in this thesis is successful, the ends of the chromatids on each side of the chromosome image must be morphologically closed. This step is necessary for the width profile protocol introduced in Section 5.3 as well as for the curvature analysis protocol introduced in Section 5.4 to function properly.

In order to perform the closing procedure, a specific mask is digitally superimposed onto the rotated binary image of a segmented chromosome for the purpose of filling the ends of the chromatids. This process is automated. In order to create the mask, the centre-point of the most compact bounding box along its horizontal axis is first determined. From this centre-point, a centre-line is drawn horizontally so as to split the image into two sub-images, each roughly containing a chromatid of the chromosome. The mask is created by drawing a line between the leftmost pixels and between the rightmost pixels detected in the respective chromatid sub-images. Figure 5.9 shows the digital superposition of the aforementioned mask onto the rotated binary image of a chromosome.

Figure 5.9: A visual representation of an appropriate mask being digitally superimposed onto rotated binary chromosome images. (a) Typical normal chromosomes. (b) Typical dicentric chromosomes.

The final step involves the application of the morphological hole filling protocol introduced in Section 5.2.1 in order to fill the empty regions. For

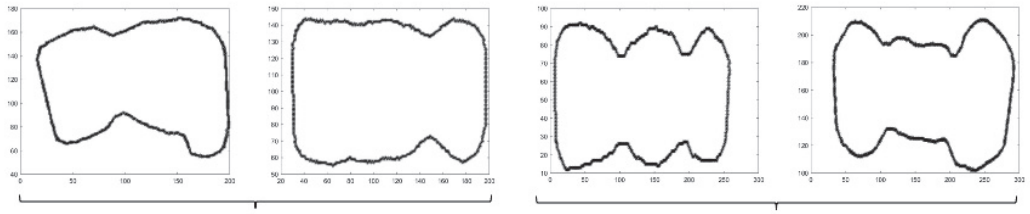
clarity, Figure 5.10 shows the results.

Figure 5.10: A visual representation of the binary chromosome images after the ends of the chromatids have been digitally closed. (a) Typical normal chromosomes. (b) Typical dicentric chromosomes.

Again note that when the chromosomes are bent (not perfectly straight) this protocol may fail.

5.2.4 Edge graph extraction

A commonly used method, known as Moore's algorithm (Moore (1968)), is employed for edge graph extraction. Edge graph extraction produces data points along the edges of objects and proves to reduce the computational burden of processing large amounts of data from an image matrix. This is therefore a necessary and efficient preprocessing step for feature extraction. The boundary associated with the binary image is traced with Moore's algorithm such that the edge graph depicted in Figure 5.11 is obtained. Moore's algorithm (Moore (1968)) operates by finding the left most, lowest pixel on the boundary, after which the chromosome boundary is traced in a clockwise direction. A $2 \times n$ matrix (where n denotes the number of boundary points) containing the x and y coordinates of the chromosome boundary is then obtained. Figure 5.11 shows a visual representation of the plotted x and y coordinates.



(a) (b)

Figure 5.11: A visual representation of the edge graphs associated with isolated chromosome images. (a) Typical normal chromosomes. (b) Typical dicentric chromosomes.

5.3 Width profile analysis

A protocol to extract and detect the most telling features of normal and dicentric chromosomes is now proposed. As stated previously, the number of centromeres a chromosome possesses will determine whether a chromosome is normal or dicentric. A protocol based on the width profile is therefore proposed for the purpose of analysing the edge graph of a chromosome. The aim of the technique is to successfully locate centromeres given an edge graph of a chromosome.

The width profile is determined from the x and y coordinates of the edge graph depicted in Figure 5.11. These coordinates are first segmented into the upper and lower sections of the chromosome by applying a threshold that constitutes the average value of the width of the graph,

$$\text{threshold} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

where y_i is the vertical coordinate of a point on the graph and n is the total number of points in the graph.

The end points of the upper and lower sections is a key concern in identifying the local minima within the width profile. In order to remove

unwanted local minima at the end points of the width profile, 10% of the length is truncated at each end of the upper and lower sections.

Next, smoothing is applied to the upper and lower sections of the edge graphs so as to discard insignificant local minima. In order to achieve this, a robust local regression smoothing technique is implemented, which uses a weighted linear least squares algorithm and a second degree polynomial model. In order to smooth a given data point the weighted neighbouring data points defined within the span is used. In this case a robust weight function is applied, which makes the process unsusceptible to outliers (Cleveland (1979)). The resulting upper and lower sections of the edge graphs as well as their respective smoothed versions are graphically illustrated in Figure 5.12.

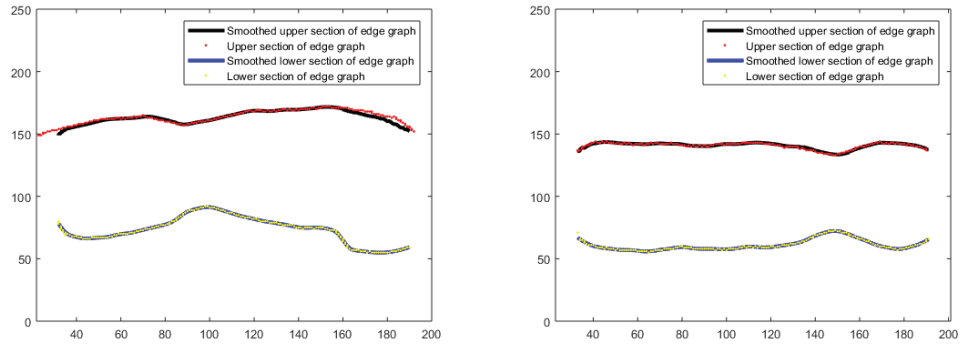
Finally, the distance between the respective y (vertical) coordinates for a corresponding x (horizontal) coordinate associated with the upper and lower edge graphs is calculated so as to obtain the width profile (see Figure 5.13).

The position of the centromere(s) within the width profile is typically associated with a deep (strong) valley. Due to the prevalence of noise in the chromosome edges, a large number of shallow (weak) valleys are often present. In order to mitigate this the prominence of each local minima detected in the width profile is taken into account. The prominence of each local minima measures how the valley stands out with respect to its depth and location relative to other valleys. Recall that the important distinguishing feature of isolated normal and dicentric chromosomes is the presence of one or two centromeres respectively. Subsequently, the two most prominent local minima are extracted whenever more than two local minima are detected.

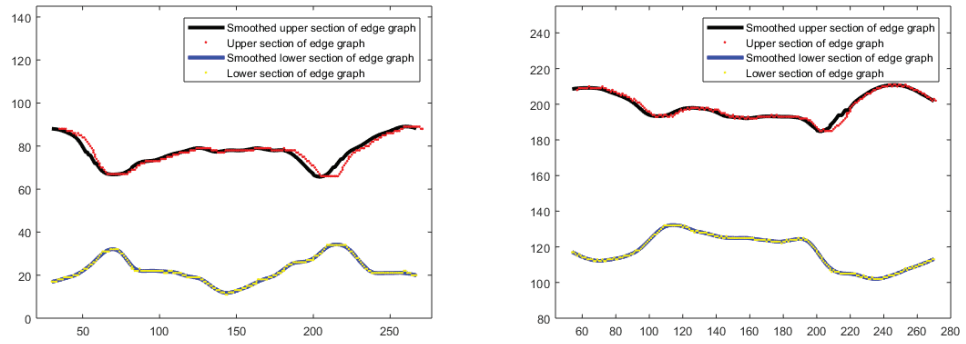
It is clear from Figure 5.13 (a) (right) that the number local minima detected using the width profile analysis protocol proposed in this thesis is not always as expected. This is due to the fact that chromosomes are typically

irregular. In order to improve upon the width profile analysis protocol proposed in this thesis, an investigation into the location of the local minima with respect to one another may be investigated, but this is considered to be part of future work.

The number of the local minima in the width profile will be used in conjunction with curvature analysis to determine whether the isolated chromosome is normal or dicentric (see Section 6.4).



(a)



(b)

Figure 5.12: A visual representation of the upper (red points) and lower (yellow points) sections of the edge graphs obtained through applying an appropriate threshold. The resulting smoothed upper section (black line) and smoothed lower section (blue line) are obtained by applying robust local regression. (a) Typical upper and lower section of edge graph of normal chromosomes. (b) Typical upper and lower section of edge graph of dicentric chromosomes.

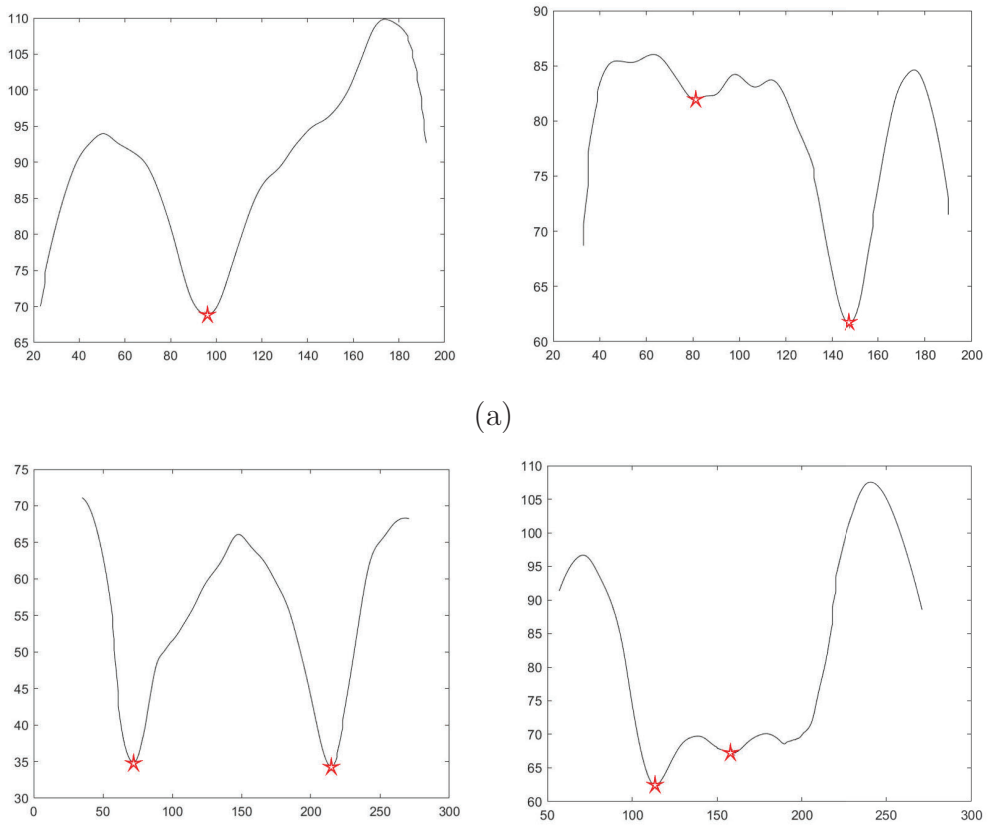


Figure 5.13: A visual representation of the width profile. The most prominent local minima are indicated by the red stars. (a) Typical representation of normal chromosomes. (b) Typical representation of dicentric chromosomes.

5.4 Curvature analysis

In this section a feature extraction protocol that relies on the curvature of a given chromosome is proposed. The aim of the extraction protocol is to locate concave regions in the edge graph obtained in Section 5.2.4 that may be associated with centromeres.

An equation for the curvature of a twice differentiable curve will be derived in Sections 5.4.1 and 5.4.2.

5.4.1 Curvature equation

Let C be a twice differentiable plane curve as depicted in Figure 5.14.

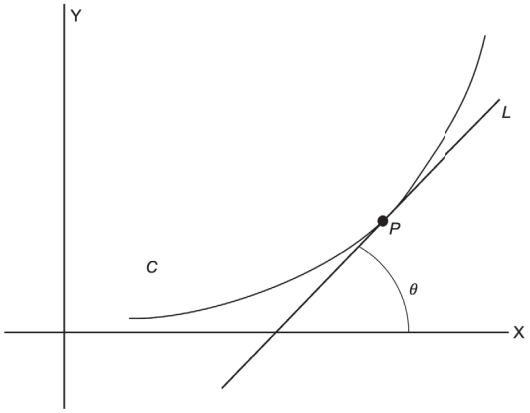


Figure 5.14: Plane curve.

At the point P on the curve a tangent line L is drawn. The angle that L makes with the x-axis is labelled θ. As P moves along the curve, it causes L and θ to change. The magnitude κ of the change in θ per unit arc length is called the curvature and can be expressed as follows,

$$\kappa = \frac{d\theta}{ds}, \tag{5.1}$$

where ds denotes the change in the position of P.

If the curve is a twice differentiable function $y = y(x)$, then the curvature κ can be calculated using the formula

$$\kappa = \frac{\frac{d^2y}{dx^2}}{[1 + (\frac{dy}{dx})^2]^{\frac{3}{2}}}. \tag{5.2}$$

Derivation of Equation 5.2

The angle θ is related to the first derivative by $\tan\theta = \frac{dy}{dx}$. Therefore, it follows that $\theta = \tan^{-1}(\frac{dy}{dx})$. Differentiating with respect to x gives

$$\frac{d\theta}{dx} = \frac{1}{1 + (\frac{dy}{dx})^2}.\frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{\frac{d^2y}{dx^2}}{1 + (\frac{dy}{dx})^2}. \tag{5.3}$$

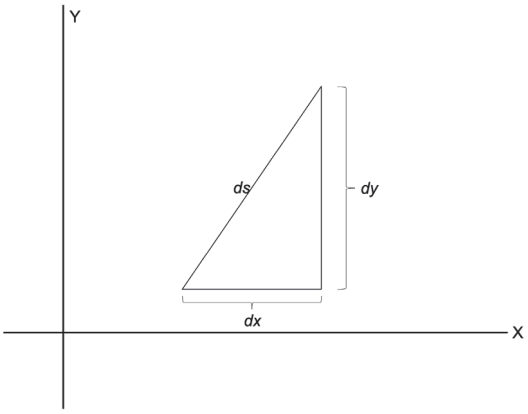


Figure 5.15: The relationship between dx, dy and ds.

From Figure 5.15 it follows that

$$ds = \sqrt{(dx)^2 + (dy)^2}$$

or

$$\frac{ds}{dx} = \frac{1}{dx}\sqrt{(dx)^2 + (dy)^2}$$

which may be expressed as

$$\frac{ds}{dx} = \sqrt{\left(\frac{dx}{dx}\right)^2 + \left(\frac{dy}{dx}\right)^2},$$

i.e.

$$\frac{ds}{dx} = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}. \tag{5.4}$$

From Equation 5.4 the chain rule can be applied. It follows that,

$$\frac{d\theta}{dx} = \frac{d\theta}{ds}.\frac{ds}{dx} = \frac{d\theta}{ds}\sqrt{1 + \left(\frac{dy}{dx}\right)^2} = \frac{\frac{d^2y}{dx^2}}{1 + (\frac{dy}{dx})^2}. \tag{5.5}$$

From Equation 5.5 it follows that

$$\frac{d\theta}{ds} = \frac{\frac{d^2y}{dx^2}}{[1 + (\frac{dy}{dx})^2]\sqrt{1 + \left(\frac{dy}{dx}\right)^2}}.$$

Therefore

$$\kappa = \frac{d\theta}{ds} = \frac{\frac{d^2y}{dx^2}}{[1 + (\frac{dy}{dx})^2]^{\frac{3}{2}}}, \tag{5.6}$$

which completes the derivation of Equation 5.2. ■

5.4.2 Parametric curvature equation

If the curve is specified parametrically by a twice differentiable vector function

$$\boldsymbol{r}(t) = x(t)\boldsymbol{i} + y(t)\boldsymbol{j},$$

then

$$\kappa = \frac{\left(\frac{dx}{dt}\right)\left(\frac{d^2y}{dt^2}\right) - \left(\frac{dy}{dt}\right)\left(\frac{d^2x}{dt^2}\right)}{\left[(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2\right]^{3/2}}. \tag{5.7}$$

Derivation of Equation 5.7

$$\frac{dy}{dx} = \frac{dy}{dt} \Big/ \frac{dx}{dt} = \frac{y'}{x'}$$

From differentiation rules it follows that

$$\frac{d^2y}{dx^2} = \frac{x'y'' - y'x''}{(x')^3}$$

Therefore, from Equation 5.2 and the above expressions for $\frac{dy}{dx}$ and $\frac{d^2y}{dx^2}$, the curvature can also be expressed as,

$$\kappa = \frac{x'y'' - y'x''}{[(x')^2 + (y')^2]^{3/2}}, \tag{5.8}$$

which completes the derivation of Equation 5.7. ■

From Equation 5.8 it follows that the curvature can be negative or positive (see the curvature plot of Figure 5.17). If a curve is traversed in a clockwise fashion and a peak (which is considered convex) is passed, the curvature is positive. When a valley is passed (which is considered concave), the curvature is negative. This is illustrated in Figures 5.16 and 5.17.

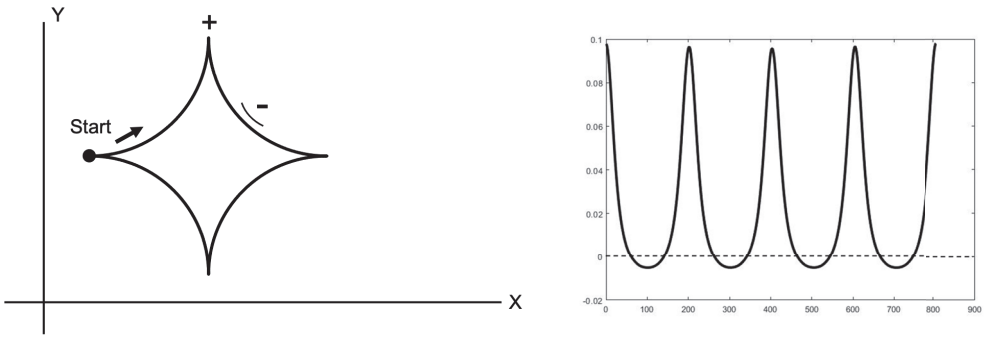


Figure 5.16: A visual representation of curvature analysis using a toy example. **(Left)** Edge graph with 807 points. **(Right)** Corresponding curvature plot using 200 neighbouring points. The concept of neighbouring points and the need for boundary smoothing are discussed in Section 5.4.3

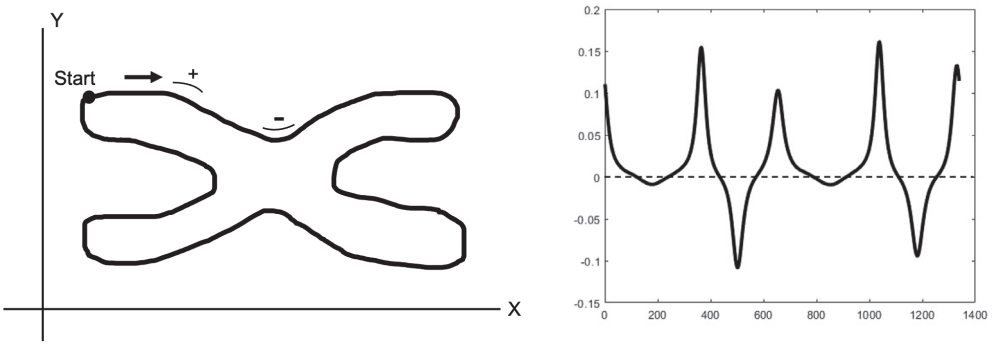


Figure 5.17: A visual representation of curvature analysis using an idealised chromosome. **(Left)** Edge graph with 1339 points. **(Right)** Corresponding curvature plot using 200 neighbouring points. The concept of neigbouring points and the need for boundary smoothing are discussed in Section 5.4.3

5.4.3 Parametrisation and curvature calculation

The chromosome boundary can be viewed as a contiguous curve, with each chromosome boundary point viewed as a parametric function pair $(x_j, y_j),\ j = 1, ..., n$, where n denotes the number of boundary points.

The curvature at any point on the curve is given by

$$C(t) = \frac{\dot{x}(t)\ddot{y}(t) - \ddot{x}(t)\dot{y}(t)}{(\dot{x}(t)^2 + \dot{y}(t)^2)^{\frac{2}{3}}}, \tag{5.9}$$

as illustrated in the Section 5.4.2. To assert that a curve has as curvature $\frac{1}{r}$ at an arbitrary point P on the curve, is to say that the curve is turning at a rate of a circle of radius r. The smaller the circle, the tighter the turn and, thus, the greater the curvature. When computing the curvature of the chromosome boundary, the curvature at the valleys near centromeres is typically negative.

Since the angle between neighbouring pixels in digital images is one of the following, $(0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°)$, as illustrated in Figure 5.18, the chromosome boundary has to be smoothed in order to accurately approximate the first and second derivative.

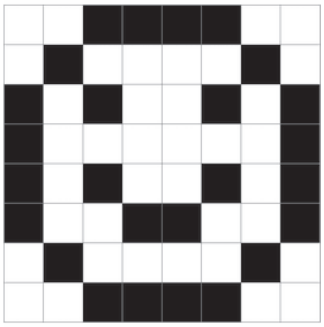


Figure 5.18: Graphical depiction of pixels.

An economic way of achieving the smoothing is to use approximations for the derivatives in Equation 5.9 based on the number of coordinates. Therefore, the least square method provides a suitable technique for accurately approximating the first and second derivatives.
The equation of a straight line is as follows,

$$x(t) = m_x t + c_x, \tag{5.10}$$

where m_x is the gradient of the line and c_x is the intersection of the line with the $x(t)$ axis. Since the equation of the line contains two unknowns, m_x and c_x, two equations are needed to solve them. If more than two points are involved, an over-determined system is obtained, which can be solved by means of the least squares method. The resulting line is then a least squares fit through the points. The over-determined system is given by,

$$\begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_c \\ \vdots & \vdots \\ 1 & t_N \end{bmatrix} \begin{bmatrix} c_x \\ m_x \end{bmatrix} = \begin{bmatrix} x(t_1) \\ x(t_2) \\ \vdots \\ x(t_c) \\ \vdots \\ x(t_N) \end{bmatrix},$$

where N is the number of neighbouring pixels. Solving the above over-determined system gives an approximation of the gradient of the curve at $x(t_c)$. From Equation 5.10 it follows that the first derivative is given by $\dot{x}(t) \approx m_x$. A similar result is used when computing the first derivative of $y(t)$, that is $\dot{y}(t) \approx m_y$.

In order to approximate the second derivative of a smoothed version of $x(t)$ the best parabola, with equation

$$x(t) = a_x t^2 + b_x t + c_x, \tag{5.11}$$

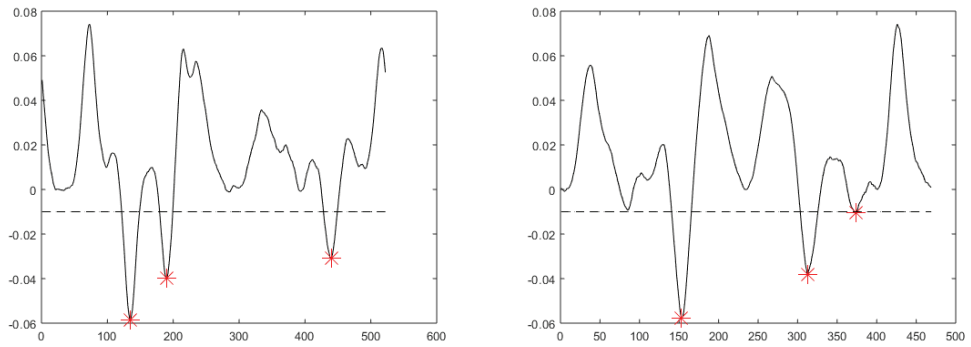
is fitted through the N points, where the resulting over-determined system is given by,

$$\begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_c & t_c^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_N^2 \end{bmatrix} \begin{bmatrix} c_x \\ b_x \\ a_x \end{bmatrix} = \begin{bmatrix} x(t_1) \\ x(t_2) \\ \vdots \\ x(t_c) \\ \vdots \\ x(t_N) \end{bmatrix},$$

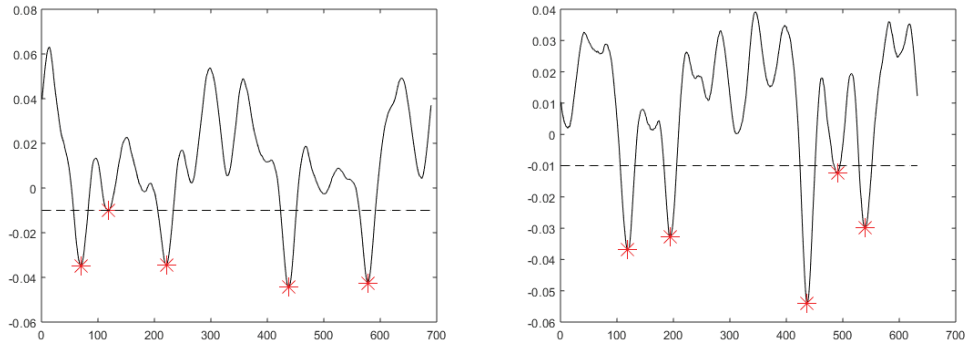
and N is the number of neighbouring pixels. From Equation 5.11 it follows that the second derivative of $x(t)$ is given by $\ddot{x}(t) \approx 2a_x$, which can be easily obtained from the least squares solution. Similar results follow when computing the second derivative of $y(t)$, that is $\ddot{y}(t) \approx 2a_y$.

The next step is to find the points where the centromeres are located, known as extremal points. These points will later contribute to the main objective, which is the classification of an isolated chromosome as either normal or dicentric. When calculating the curvature, it is noted that scaling with respect to the total number of points, n, in the edge graph is required. Therefore, the number of neighbouring points N that is used for a given chromosome is determined as follows, $N = \frac{2}{65} \times n + 32$, which is subsequently

rounded to the nearest integer in intervals of 5. This ensures that large enough curvatures are located at the extremal points to distinguish them from flatter curves, as is illustrated in Figure 5.20. The dashed lines in Figure 5.19 graphically represent the threshold values employed in order to determine the extremal points from the curvature plot.

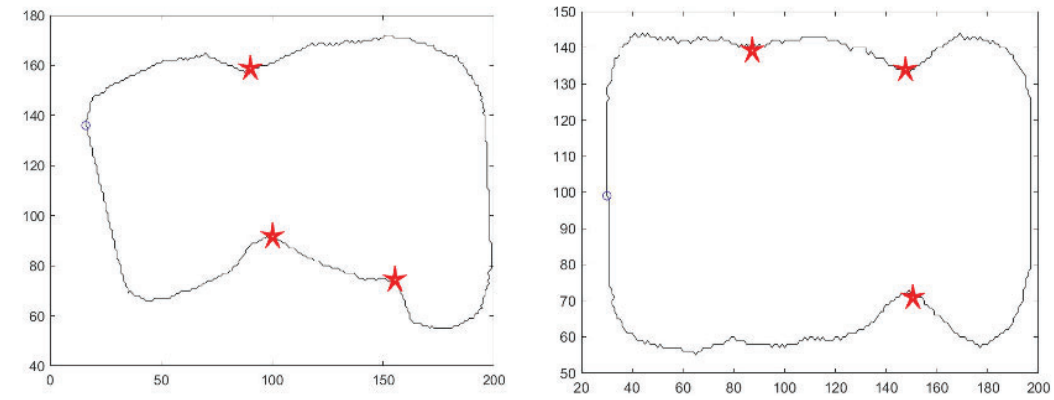


(a)

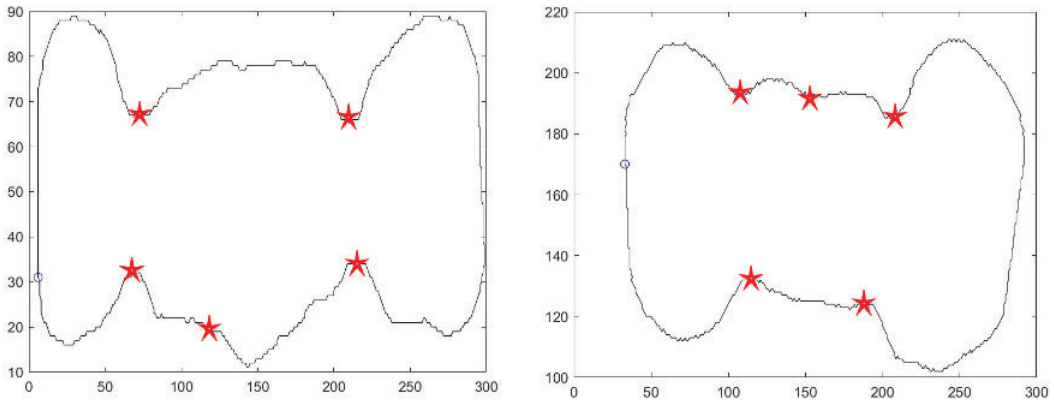


(b)

Figure 5.19: A visual representation of the curvature plot of an isolated chromosome using appropriate smoothing. The resulting local minima which are less than the threshold (dashed line) are indicated by the red stars. (a) Typical representation of normal chromosomes. (b) Typical representation of dicentric chromosomes.



(a)



(b)

Figure 5.20: A visual representation of the located valley points (red stars) on the edge graph that is obtained in Section 5.2.4. (a) Typical representation of normal chromosomes. (b) Typical representation of dicentric chromosomes.

The number of the located valleys in the curvature analysis are then used in conjunction with width profile analysis to determine whether the isolated chromosome is normal or dicentric (see Section 6.4).

5.5 Concluding remarks

In this chapter a suitable feature extraction protocol was proposed in order to obtain valuable shape information. In order to exploit the shape information

the image should be preprocessed and rotated to ensure rotational invarience. Edge graph extraction was employed in order to obtain the data points along the edges of objects. Finally two feature extraction protocols were proposed in order to determine the number of centromeres present for a given chromosome. This was achieved by determining the width profile and conducting curvature analysis associated with the chromosome. In the following chapter a ground truth is generated in order to determine the proficiency of the proposed system.

Chapter 6

Experiments

6.1 Introduction

In this chapter experiments are carried out in order to determine the proficiency of the novel strategies proposed to detect regions of interest (ROIs), as well as the proficiency of the novel strategies proposed to classify isolated chromosomes as normal or dicentric. Recall that due to variations in imaging equipment and differences in methods for treating samples, implementations are often specific to a single laboratory. Therefore the aforementioned experiments are exclusively conducted on and compared to the metaphase images provided by iThemba LABS. This dataset is described in Section 6.2. In order to evaluate the proposed protocols, a ground truth is developed and outlined in Section 6.3. For each individual experiment that is conducted, the experimental protocol and the corresponding results are outlined in Section 6.4.

6.2 Data

The dataset considered in this study was acquired directly from the radiobiology laboratory at iThemba LABS. This dataset consists of grey-scale images that were captured under a light microscope and contains 742 metaphase images from healthy male and female individuals.

In order to obtain the metaphase images, small rectangular glass microscope slides with metaphase spreads were prepared in the radiobiology laboratory at iThemba LABS on which the specimens are mounted. In order to stain the chromosomes in the metaphase spreads, a nucleic acid stain called Giesma stain is used. The Giesma stain joins itself to the parts of

the DNA where high amounts of adenine-thymine is found. The stained specimen is subsequently scanned with the Metafer4 system using a Carl Zeiss AxioImager.Z2 microscope (Metasytems, Germany). The automated metaphase finding module (MSearch) was used at a 10x magnification to detect the metaphases on the microscope slide. Thereafter, the unattended image acquisition module (AutoCapt) was used to capture metaphases at a higher magnification with a 63x/1.40 oil objective. The images were captured at a resolution of 1280×1024 pixels and exported as TIFF files from the Metafer4 system for further analysis. Figure 6.1 shows a typical grey-scale metaphase image.

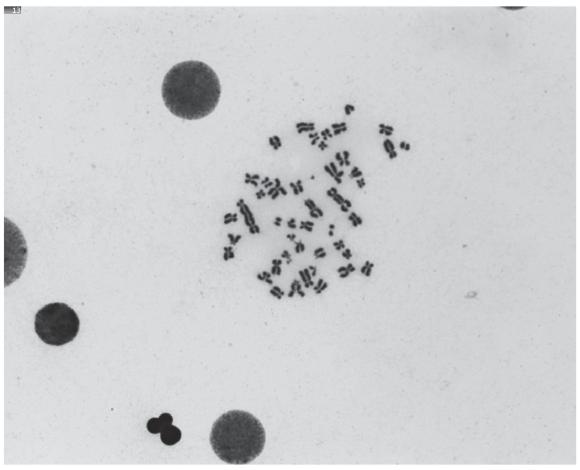


Figure 6.1: A visual representation of a typical grey-scale metaphase image obtained from iThemba LABS.

6.3 Ground truth generation

The process of manually scoring metaphase images is proven to be labour intensive, highly time-consuming and open to human error depending on the expert experience. For these reasons, laboratories all over the world have implemented analytical processes and protocols for dicentric assay. In order

to validate the proposed protocol in the following sections, a protocol for generating a ground truth must first be put into place. In generating the ground truth for the acquired metaphase images in the dataset, seven experts from iThemba LABS with varying degrees of experience were employed. In order to initialise the scoring protocol, each expert is tasked to score 400 random metaphase images out of the total of 742 metaphase images in the dataset. Each expert is therefore given the following instructions:

1. Indicate whether the metaphase image is score-able or has to be discarded.
2. For each chromosome on a metaphase image, indicate the chromosome's category using a coloured dot as follows:
   (a) a red dot for a normal chromosome,
   (b) a cyan dot for a dicentric chromosome, and
   (c) an orange dot for an acentric fragment.

Given the instructions above, the scored images are returned for ground truth generation. Figure 6.2 shows an example of a scored image obtained from the experts.

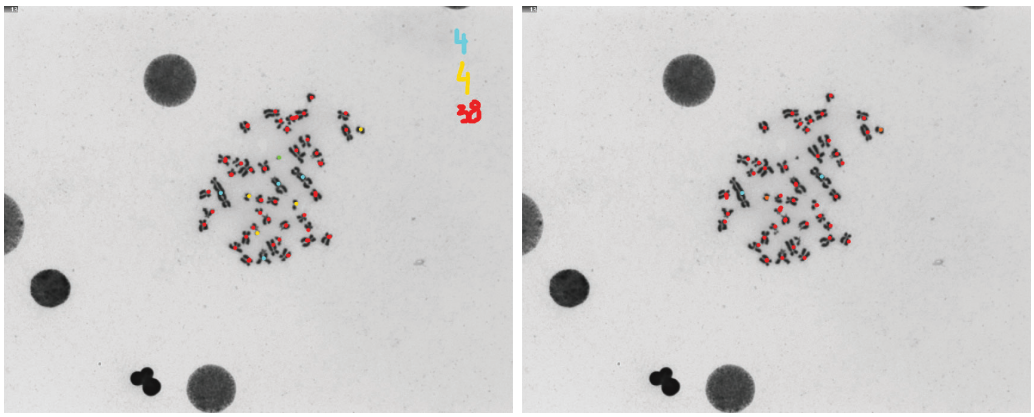


Figure 6.2: A visual illustration of scored metaphase images returned from the experts.

Each expert is ranked according to his/her experience in scoring metaphase images. A scale of 1 to 7 is used where an expert with a ranking of 1 is the most advanced, while an expert with a ranking of 7 is the least experienced. At minimum, each metaphase image is scored by two experts independently. Once the images are scored, the analysis to evaluate the ground truth is started. The proposed ground truth generation protocol can be divided into four main steps:

1. Discarding metaphase images (see Section 6.3.1).
2. Collection of data points (see Section 6.3.2).
3. Labelling using ROIs (see Section 6.3.3).
4. Manual labelling (see Section 6.3.4).

6.3.1 Discarding metaphase images

The first stage in generating the ground truth is to remove metaphase images that are deemed to be not score-able. The experts returned a table with two columns containing the image number and a string stating whether the image should be discarded. In order to determine which metaphase images should be discarded, the tables from the experts are joined according to the image number. The resulting table therefore has three columns: the image number, an array containing the rankings of the experts that scored the metaphase image in question, and an array containing the strings that state whether the metaphase image in question should be discarded. The array of strings are converted into a binary value where the capitalised 'Yes' is converted to 1 and the capitalised 'No' is converted to a 0. Given the prepared table, the array of strings is assessed as follows so as to determine whether the image should be discarded:

1. A majority vote decision is first employed to determine whether the image

should be discarded.

2. In the event of a tie, the scorers' expert ranks are used to swing the vote by considering the minimum average rank per category, since a scorer with a lower rank is deemed to be more experienced.

Using the above-mentioned protocol, 99 images were discarded, leaving 643 images for further analysis. Figure 6.3 shows examples of images that were discarded. Images may be discarded for the following most common reasons:

1. The metaphase spread is not clear enough.
2. The image consists of more than the metaphase spreads.
3. The view to analyse the chromosomes is obstructed.
4. The image is of low quality.

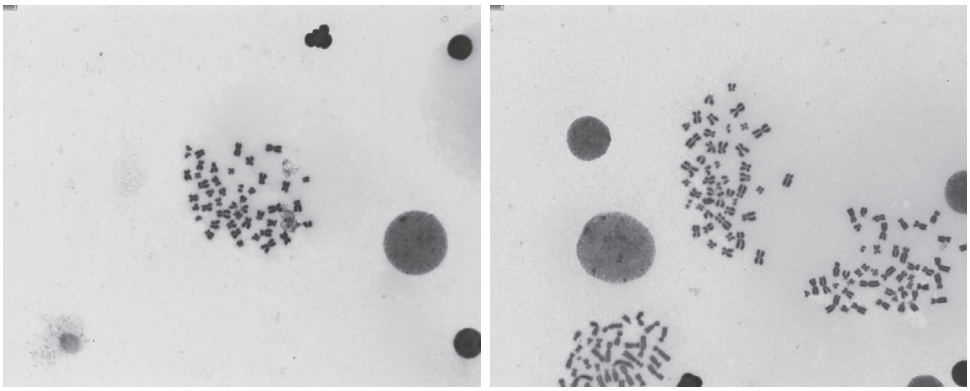


Figure 6.3: A visual illustration of discarded images.

6.3.2 Collection of data points

In order to determine the category of a chromosome in each expert's metaphase image, colour segmentation is applied using the RGB colour model. The RGB colour model is based on the fact that a colour image has three primary spectral components that is red (R), green (G) and blue (B). The colour of a given pixel can be defined as a vector (R,G,B) in 3-dimensional space by specifying

the intensity level ranging between 0 and 255 for each of the red, green and blue components. Colour segmentation for each metaphase image is achieved through the extraction of specific colours associated with a corresponding category. Therefore, enforcing the requirement that $R = 255$ while $G = B = 0$ will extract all the red dots (considered to be normal chromosomes) in the image. The image is subsequently binarised by allocating the pixels in question an assigned value of one (rendered white), while the other pixels are assigned a value of zero (rendered black).

In order to determine the location of the extracted dots, the centres of mass of the connected components in the binary image is calculated using `regionprops` in MATLAB. It therefore returns the x and y coordinates of the centres of mass. These coordinates, in conjunction with the category and ranking of the expert scoring the image, is sorted in an array for further analysis. This process is repeated for all the (R, G, B) values associated with a specific category. Consequently, by enforcing the requirement that $(R, G, B) = (0, 255, 255)$, all the cyan dots (considered to be dicentric chromosomes) in an image are extracted. In the case of acentric fragments multiple shades of orange is extracted, since the different experts did not use the same shade of orange.

6.3.3 Labelling using ROIs

The above-mentioned stored data points containing the location, category and ranking of the experts for each scored chromosome in a metaphase image can now be used in order to generate the ground truth. In this section the focus is on assigning a final label to each ROI obtained during the image segmentation protocol (see Section 4.3.3). In the context of labelling the ROIs it can be assumed that each bounding box will contain one chromosome and can be used to indicate the position of each chromosome in the metaphase image. The stored coordinates from the set of ROIs in a metaphase image can now

be used to determine which data points are in a given bounding box. In the following equation,

$$x_2 \leq x \leq x_1,$$

$$y_2 \leq y \leq y_1,$$

the coordinates (x_1, y_1) and (x_2, y_2) represent the top left and bottom right corners of the bounding box and (x, y) represents the location of the data point in question. The dots within each bounding box are identified, after which analysis is conducted on the identified dots using the category specified by the experts so as to determine the label according to the following criteria:

1. A majority vote decision is first employed to determine the correct label for a given bounding box.
2. In the event of a tie, the scorers' expert ranks are used to swing the vote by considering the minimum average rank per category, since a scorer with a lower rank is deemed to be more experienced.

6.3.4 Manual labelling

Manual labelling is applied to the following objects: (1) all chromosomes and fragments that were not detected during the image segmentation process, (2) all clusters of chromosomes (on top of each other or in very close proximity to each other), and (3) all chromosomes where a bounding box contains another bounding box.

6.4 Experimental protocol and results

In this section the protocol and effectiveness of the proposed image processing based detection and classification experiments are elaborated and reported on. In this study, for the purpose of quantifying the proficiency of the proposed

detection and classification experiments, the following statistical measures are used:

- Number of true positives (TP), that is the number of correctly accepted positives instances.
- Number of false positives (FP), that is the number of incorrectly accepted negatives instances.
- Number of false negatives (FN), that is the number of incorrectly rejected positives instances.
- Number of true negatives (TN), that is the number of correctly rejected negatives instances.

In order to quantify the proficiency of the proposed system the following statistical performance measures are used:

Performance measure	Definition	Relationship
False positive rate (FPR)	$\frac{\text{FP}}{\text{FP+TN}}$	FPR = 1−TNR
False negative rate (FNR)	$\frac{\text{FN}}{\text{FN+TP}}$	FNR = 1−TPR
True positive rate (TPR)	$\frac{\text{TP}}{\text{TP+FN}}$	TPR = 1−FNR
True negative rate (TNR)	$\frac{\text{TN}}{\text{TN+FP}}$	TNR = 1−FPR
Accuracy (ACC)	$\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$	

The above-mentioned performance measures are measured using a range $[0, 1]$ where the first two measures indicates a higher accuracy if they are close to 0 since they represent error rates, while the last three measures represent system proficiency and indicate a higher accuracy when it is close to 1. For reference purposes Table 6.1 shows the confusion matrix with the appropriate statistical measures.

		Actual class		
		Positive	Negative	
Predicted class	Positive	TP	FP	FPR
	Negative	FN	TN	FNR
		TPR	TNR	**ACC**

Table 6.1: Illustration of confusion matrix with the appropriate statistical measures.

6.4.1 Experiment 1: Detection of ROIs in metaphase images

The main objective of the image segmentation protocol is to successfully detect objects of interest. Recall that an object of interest constitutes either a normal chromosome, a dicentric chromosome, an acentric fragment or a cluster of chromosomes. In this experiment a questioned ROI is matched to the corresponding reference bounding box containing a label that was assigned to its location during the ground truth generation phase. The questioned ROI should be detected (accepted) as an object of interest if and only if the corresponding reference contains a label. In scenarios where the corresponding reference has not been assigned a label, the ROI constitutes dirt and should therefore not be detected. The number of true negatives (that is dirt classified as dirt) is not reported on, since dirt is simply discarded during the image segmentation phase. Table 6.2 illustrates the resulting effectiveness of this detection phase.

		Actual class		
		Positive	Negative	
Predicted class	Positive	27611	1384	■
	Negative	1646	■	5.63%
		94.37%	■	■

Table 6.2: Detection of objects of interest. Use Table 6.1 as a reference.

An object is quantified as dirt when it is significantly larger or smaller than the average size of an object in a metaphase image. Consequently,

dirt of a size similar to that of a normal chromosome is detected as an object of interest. Note that a cluster of chromosomes is detected as a single object of interest since it constitutes a single connected component. The single object of interest therefore contains multiple labels. Out of the 27611 detected objects of interest, 365 constitute clusters that contain 993 individual chromosomes. The separation of these chromosomes is challenging and forms part of future work. The detection of a cluster that is associated with chromosomes in close proximity of each other is usually the result of poor image quality. Therefore only 26618 isolated chromosomes and fragments are detected. The separation of these chromosomes may be facilitated by improved image enhancement techniques and forms part of future work. The detection of isolated chromosomes can be further broken down into the respective categories. Table 6.3 illustrates the resulting effectiveness of this detection phase within the context of each category.

Category (isolated)	Total in ground truth	Total detected	TPR
Normal chromosomes	27794	25438	91.52%
Dicentric chromosomes	529	386	72.97%
Acentric chromosomes	934	794	85.01%
Total objects of interest	29257	26618	90.1%

Table 6.3: The effectiveness of the detection phase within the context of each category.

Within the context of detecting *isolated* objects of interest, a TPR of 90.1% is achieved. Within the context of detecting objects of interest, which includes clusters of chromosomes, a TPR of 94.37% is achieved.

6.4.2 Experiment 2: Classification of isolated chromosomes

The main objective of the classification protocol is to classify an isolated chromosome as either normal or dicentric using the extracted features. In

the feature extraction protocol discussed in the previous chapter two main methods are used, namely (1) width profile analysis and (2) curvature analysis. Width profile analysis returns the number of local minima in the width profile for an isolated chromosome, while curvature analysis returns the number of located valleys on the boundary of the chromosome. In this experiment a questioned isolated chromosome is matched to the corresponding reference bounding box containing a label that was assigned to its location during the ground truth generation phase. The questioned isolated chromosome should be accepted as dicentric if and only if the corresponding reference contains the label "dicentric". In classifying isolated chromosomes, three sub-experiments are conducted: (a) classification based only on the local minima obtained through width profile analysis, (b) classification based only on the located valleys obtained through curvature analysis, and (c) classification based on information obtained through *both* width profile analysis *and* curvature analysis.

Experiment 2a: Width profile analysis. In order to classify an isolated chromosome using width profile analysis the following guideline is taken into account: centromeres are typically located within the slimmest region of a chromosome. After a single ROI has been automatically identified and the detected object has been orientated and analysed, the chromosome is classified as dicentric (or accepted as an instance belonging to the positive class) when its width profile has two prominent local minima and therefore contains two centromeres. A chromosome is classified as normal (or rejected as an instance belonging to the negative class) when its width profile has only one (or less than one) prominent local minimum and therefore contains only one centromere. Table 6.4 illustrates the proficiency of the classification protocol using only width profile analysis.

		Actual class: Positive	Actual class: Negative	
Predicted class	Positive	342	16065	63.30%
	Negative	36	9313	9.52%
		90.48%	36.70%	**37.49%**

Table 6.4: Classification of isolated chromosomes using width profile analysis. Use Table 6.1 as a reference.

From the results presented in Table 6.4 it is clear that the proposed width profile analysis-based classification system is more proficient in correctly classifying dicentric chromosomes. It is also noted from the FPR of 63.30% that the proposed system is not proficient in classifying normal chromosomes. This deficiency of the proposed protocol may be due to the fact that chromosomes are typically irregular. In addition to this, isolated chromosomes may often be bent, while in other cases, the concavity associated with a centromere may be very subtle.

Experiment 2b: Curvature analysis. In order to classify an isolated chromosome using curvature analysis the following guideline is taken into account: the boundary segment in the vicinity of a centromere is generally more concave towards the outside of the chromosome than is the case for the rest of the boundary. After a single ROI has been automatically identified and the detected object has been orientated and analysed, the chromosome is classified as dicentric (or accepted as an instance belonging to the positive class) when its curvature analysis yields four or more local valleys and is therefore concluded to contain two centromeres. A chromosome is classified as normal (or rejected as an instance belonging to the negative class) when its curvature analysis yields three or less local valleys and is therefore concluded to contain only one centromere. Table 6.4 illustrates the proficiency of the classification protocol using only curvature analysis.

		Actual class		
		Positive	Negative	
Predicted class	Positive	329	3263	12.86%
	Negative	49	22115	12.96%
		87.04%	87.14%	**87.14%**

Table 6.5: Classification of isolated chromosomes using curvature analysis. Use Table 6.1 as a reference.

From the results presented in Table 6.5 it is clear that the proposed curvature analysis-based classification system is equally proficient in correctly classifying dicentric and normal chromosomes.

Experiment 2c: Width profile analysis and curvature analysis. In this section an experimental protocol is proposed to evaluate the proficiency of an aggregated classification system which employs *both* curvature analysis *and* width profile analysis. After a single ROI has been automatically identified and the detected object has been orientated and analysed, the chromosome is classified as dicentric (or accepted as an instance belonging to the positive class) when its curvature analysis yields four or more local valleys *and* its width profile analysis yields two local minima. A chromosome is classified as normal (or rejected as an instance belonging to the negative class) when its curvature analysis yields three or less local valleys *and* its width profile analysis yields one or less local minima. Table 6.4 illustrates the proficiency of the classification protocol using this aggregated approach (width profile analysis combined with curvature analysis).

		Actual class		
		Positive	Negative	
Predicted class	Positive	308	2519	9.93%
	Negative	70	22859	18.52%
		81.48%	90.07%	**89.95%**

Table 6.6: Classification of isolated chromosomes using a combination of width profile analysis and curvature analysis. Use Table 6.1 as a reference.

From the results presented in Table 6.6 it is clear that the proposed classification system is sufficiently adequate at correctly classifying dicentric chromosomes, while being more proficient in correctly classifying normal chromosomes.

6.5 Discussion

Based on the experiments described in this chapter it was demonstrated that the detection phase which employs the proposed image segmentation protocol detects *isolated* objects of interest with a promising true positive rate (TPR) of 90.1% and detects objects of interest which *includes* clusters of chromosomes with a TPR of 94.37% when compared to the ground truth obtained from experts at iThemba LABS.

In classifying isolated chromosomes as normal or dicentric, experiments using three strategies were conducted, that is (1) width profile analysis, (2) curvature analysis and (3) an aggregated approach that combines width profile analysis and curvature analysis. Accuracies of 37.49%, 87.14% and 89.95% were respectively reported in classifying normal and dicentric chromosomes. It is therefore reasonable to conclude that width profile analysis (in isolation) for chromosome classification did not perform sufficiently well for the dataset from iThemba iThemba LABS, and that the proposed *novel* protocol based on curvature analysis proved to be much more robust and proficient. However, the ulilisation of width profile information in *conjunction* with curvature information, improved the accuracy of the results. When comparing the classification results when curvature analysis (in isolation) and the aggregated (combined) approach are respectively employed, it is worth noting that true positive rates (TPRs) of 87.04% and 81.48% were reported. When considering the fact that the employed dataset is extremely imbalanced, that is, the ratio of dicentric to normal chromosomes is approximately 1:67, the relatively

small difference in the performance of the proposed strategies my be deemed irrelevant. The overall TPR is however deemed an important metric to consider when comparing these two approaches.

The research provided useful insight into detecting and classifying chromosomes within a metaphase image and unfolded new avenues for future research. This is discussed in the final chapter.

Chapter 7

Conclusion and future work

7.1 Conclusion

In this thesis chromosome detection and classification systems were proposed. Firstly, a segmentation protocol which facilitates the automatic detection of a region of interest (ROI) that encloses normal chromosomes, dicentric chromosomes, acentric fragments and clusters of chromosomes was developed. The aforementioned segmentation protocol involves preprocessing techniques and a novel binarisation protocol, which is followed by the manual extraction of isolated normal and dicentric chromosomes. Preprocessing for image analysis was subsequently achieved by modifying each binary chromosome image through the application of morphological transformations in order to isolate, fill and smooth the chromosome in question. This was followed by the application of a discrete Radon transform (DRT), so as to determine the appropriate alignment for the purpose of achieving rotational invariance. Masking was applied to the ends of the chromatids for the purpose of closing the gaps between them. Edge graph extraction was employed in order to obtain the data points along the edges of objects. Finally two feature extraction protocols were proposed in order to determine the number of centromeres present for a given chromosome. This was achieved by determining the width profile and curvature associated with the chromosome. In order to determine the width profile, the chromosome is divided into upper and lower segments. Said segments were subsequently smoothed and subtracted from each other in order to determine the width profile. The number of prominent local minima was calculated in order to determine whether a chromosome has one or two centromeres. In

order to determine the curvature of a chromosome boundary, the least squares method provided a suitable technique for accurately approximating the first and second derivatives. These derivatives were then used to determine the curvature at any given point. In order to locate prominent valleys in the boundary of the chromosome a suitable threshold is employed. The number of located valleys is then used to determine whether the chromosome has one or two centromeres.

In order to estimate the proficiency of the proposed system a ground truth has been generated with the help of experts at iThemba LABS. The proficiency of the novel strategy proposed to detect regions of interest during the detection phase is validated by the detection of objects with a reasonable accuracy when compared to the ground truth. During the phase in which isolated chromosomes are classified as normal or dicentric, three experiments are conducted, namely (1) width profile analysis, (2) curvature analysis, and (3) an aggregated approach that combines width profile analysis and curvature analysis. Width profile analysis (on its own) did not perform well on the iThemba LABS dataset. However, the classification results for curvature analysis (on its own), as well as the aggregated approach, are very promising.

The objectives mentioned in Section 1.2 have therefore been successfully achieved.

7.2 Future work

The research conveyed in this thesis provided valuable insights into numerous aspects relating to the detection and classification of chromosomes in a metaphase image, but due to time constraints, certain avenues were not pursued and should therefore be considered as future work:

1. The image segmentation protocol developed in this thesis relies on the

manual categorisation of the set of individual ROIs into dirt, acentric fragments, clusters of chromosomes and isolated chromosomes. An investigation into the automated categorisation of the set of individual ROIs into dirt, acentric fragments and clusters of chromosomes should be conducted so as to render the classification system *fully* automatic.

2. An in-depth investigation into the separation of clusters of chromosomes, in which the individual chromosomes reside in very close proximity to one another, should be conducted.

3. An in-depth investigation into the automated straightening of bent chromosomes should be conducted in order to improve the feature extraction protocols proposed in this thesis.

4. The relative computational complexity of the strategies adopted in this thesis should be reported on.

5. An investigation into the feasibility of machine learning-based approaches to solving the problem at hand should also be conducted. This is however subject to the availability of a very large set of slide images for the purpose of training and validation.

REFERENCES

Beukes, E. (2018). *Hand vein-based biometric authentication with limited training samples*. Ph.D. thesis, Stellenbosch: Stellenbosch University.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836.

Coetzer, J. (2005). *Off-line signature verification*. Ph.D. thesis, Stellenbosch: University of Stellenbosch.

Galloway, S., Coetzer, J. and Muller, N. (2020). Image processing-based identification of dicentric chromosomes in slide images. In: *2020 International SAUPEC/RobMech/PRASA Conference*, pp. 509–514. IEEE.

Gonzalez, R.C., Woods, R.E. *et al.* (2010). Digital image processing.

Jindal, S., Gupta, G., Yadav, M., Sharma, M. and Vig, L. (2017). Siamese networks for chromosome classification. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 72–81.

Markou, C., Maramis, C., Delopoulos, A., Daiou, C. and Lambropoulos, A. (2012). Automatic chromosome classification using support vector machines. In: *Pattern Recognition: Methods and Applications*, pp. 1–24. iConcept Press.

Moore, G. (1968). Automatic scanning and computer processes for the quantitative analysis of micrographs and equivalent subjects. *Pictorial Pattern Recognition*, pp. 275–362.

Moradi, M., Setarehdan, S. and Ghaffari, S. (2003). Automatic landmark detection on chromosomes' images for feature extraction purposes. In: *3rd*

International Symposium on Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the, vol. 1, pp. 567–570. IEEE.

O'Connor, C.M. and Adams, J.U. (2010). The essentials of cell biology. Available at: `https://www.nature.com/scitable/topicpage/chromosomes-14121320/`

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66.

Qin, Y., Wen, J., Zheng, H., Huang, X., Yang, J., Song, N., Zhu, Y.-M., Wu, L. and Yang, G.-Z. (2019). Varifocal-net: A chromosome classification approach using deep convolutional networks. *IEEE transactions on medical imaging*, vol. 38, no. 11, pp. 2569–2581.

Rogan, P.K., Li, Y., Wickramasinghe, A., Subasinghe, A., Caminsky, N., Khan, W., Samarabandu, J., Wilkins, R., Flegal, F. and Knoll, J.H. (2014). Automating dicentric chromosome detection from cytogenetic biodosimetry data. *Radiation protection dosimetry*, vol. 159, no. 1-4, pp. 95–104.

Romm, H., Ainsbury, E., Barnard, S., Barrios, L., Barquinero, J., Beinke, C., Deperas, M., Gregoire, E., Koivistoinen, A., Lindholm, C., Moquet, J., Oestreicher, U., Puig, R., Rothkamm, K., Sommer, S., Thierens, H., Vandersickel, V., Vral, A. and Wojcik, A. (2013). Automatic scoring of dicentric chromosomes as a tool in large scale radiation accidents. *Mutation research-genetic toxicology and environmental mutagenesis*, vol. 756, no. 1-2, pp. 174–183. ISSN 1383-5718.
Available at: `http://dx.doi.org/10.1016/j.mrgentox.2013.05.013`

Sharma, M., Saha, O., Sriraman, A., Hebbalaguppe, R., Vig, L. and Karande, S. (2017). Crowdsourcing for chromosome segmentation and deep

classification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 34–41.

Sharma, M., Vig, L. *et al.* (2018). Automatic classification of low-resolution chromosomal images. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Shirazi, Z., Zamani, A., Mortazavi, S., Zakeri, F., Dianatpour, M. and Mosleh-Shirazi, M. (2016). Developing an automated cytogenetic imaging system for detection of dicentric chromosomes in biological dosimetry. *Journal of Biomedical Physics and Engineering*.

University of Leicester, G. (2017). The cell cycle, mitosis and meiosis resources.
Available at: `https://www2.le.ac.uk/projects/vgec/highereducation/topics/cellcycle-mitosis-meiosis`