UNIVERSIDADE D
COIMBRA

Manuel Francisco da Silva Lúcio Gonçalves

SEGMENTAÇÃO AUTOMÁTICA DE CROMOSSOMAS EM IMAGENS MICROSCÓPICAS DE CARIÓTIPOS

Dissertação no âmbito do Mestrado Integrado em Engenharia Biomédica, orientada pelo Professor Doutor Francisco José Santiago Fernandes Amado Caramelo, pela Professora Doutora Joana Barbosa de Melo e pela Professora Doutora Ilda Patrícia Ribeiro e apresentada ao Departamento de Física da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Setembro de 2022

• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Manuel Francisco da Silva Lúcio Gonçalves

Segmentação automática de cromossomas em imagens microscópicas de cariótipos

Dissertação apresentada à Universidade de Coimbra
para cumprimento dos requisitos necessários à obtenção do grau
de Mestre em Engenharia Biomédica.

Orientadores:
Professor Doutor Francisco José Santiago Fernandes Amado Caramelo (iCBR)
Professora Doutora Joana Barbosa de Melo (iCBR)
Professora Doutora Ilda Patrícia Ribeiro (iCBR)

Coimbra, 2022

Este trabalho foi desenvolvido em colaboração com:

Coimbra Institute for Clinical and Biomedical Research

Faculdade de Ciências e Tecnologias da Universidade de Coimbra

"*Pursue excellence. And success will come chasing you.*"

Ranchoddas Chanchad

Agradecimentos

Esta dissertação é o culminar da minha jornada no Mestrado Integrado em Engenharia Biomédica. Se há coisa que aprendi ao longo destes anos, é que enquanto indivíduos não somos nada sem as relações interpessoais que criamos e cuidamos ao longo da vida. Justamente, começo por agradecer a todos que, de uma forma ou outra, contribuíram para o meu percurso académico e para a minha formação pessoal.

Ao Professor Doutor Francisco Caramelo, manifesto o seu papel preponderante ao longo do desenvolvimento desta dissertação. Agradeço-lhe imensamente por toda a disponibilidade e orientação fornecidas. Sem dúvida que o seu espírito crítico e a sua forma de enfrentar adversidades são atributos que espero levar comigo. O rigor científico associado à boa disposição motivaram-me sempre a continuar e a querer mais. Um grande obrigado por todo apoio. Esta dissertação não seria possível sem as suas palavras de incentivo e os seus comentários científicos oportunos.

À Professora Doutora Joana Barbosa de Melo, expresso a minha profunda gratidão pela colaboração ao longo deste projeto. Agradeço-lhe por todo o suporte científico e apoio técnico que me dispensou, reconhecendo o seu extraordinário valor na investigação, ensino e produção científica.

À Professora Doutora Ilda Ribeiro, devo-lhe os meus agradecimentos pelo desafio lançado e pela partilha de conhecimentos.

À Doutora Ana Jardim, agradeço a disponibilidade e entusiasmo com que me recebeu. Um grande obrigado por todas as noções e princípios que me permitiu adquirir, sem os quais não seria possível realizar esta dissertação.

Agradeço a toda a minha família, por todos os valores transmitidos e por toda a paciência com que lidam comigo. À minha irmã, obrigado por todos os conselhos, és o meu maior

exemplo e o meu maior orgulho. Aos meus tios, Rafa e Anita, a vossa alegria é a minha fonte de inspiração para o modo como enfrento os meus problemas. Aos meus primos, Miguel e Pedro, estão sempre no meu pensamento e tento sempre ser mais e melhor por vocês. Ao meu pai, devo-lhe grande parte do meu pensamento crítico e espírito de resolução de problemas. Aos meus padrinhos, João e Luísa, que me ajudaram em tudo o que puderam e que nunca me falharam.

À Filó, a minha namorada e maior confidente. Foste determinante na minha jornada. Estás sempre presente e com o maior dos sorrisos, sem ti tudo seria mais difícil. Obrigado por todos os momentos que me proporcionaste e pela paciência que tiveste comigo. Conheces-me melhor do que eu me conheço a mim próprio.

Agradeço ao Miguel, o meu companheiro de curso desde o primeiro dia, à Inês, a irmã que eu não pedi mas sem a qual já não passo, e à Botinas, a minha amiga mais bem disposta de sempre. A quantidade de aprendizagens que me proporcionaram é astronómica.

À Diana, a minha melhor amiga, ao Joe, o meu pequeno grande amigo, e à Mariana, a minha açoriana. Os momentos que passámos são extremamente valiosos, levo-os comigo bem guardados. Espero ter a oportunidade de criarmos ainda mais memórias.

Aos meus amigos da terra, que me conhecem há muitos anos e que sabem que marcaram o meu percurso. Crescer ao vosso lado tem sido uma viagem estonteante. Espero que o destino ainda esteja bem longe e que aproveitemos o caminho durante muitos mais anos.

Agradeço a todos os demais que conheci em Coimbra e com quem partilhei muitas vivências. Aos meus amigos, à minha família de praxe e a toda a comunidade do Departamento de Física, estudar convosco tornou esta cidade numa segunda casa.

Por último, agradeço à pessoa mais importante na minha vida. Agradeço à minha mãe. É impossível pôr em palavras todo o carinho e preocupação com que cuidas de mim, assim como será impossível eu alguma vez retribuir-te uma pequena fração do amor que me proporcionas. Nos momentos de maior desmotivação, foi por ti que continuei, por isso dedico-te esta dissertação.

Coimbra, 20 de setembro de 2022

MANUEL FRANCISCO DA SILVA LÚCIO GONÇALVES

Resumo

O estudo do cariótipo representa a análise do conjunto de cromossomas de um indivíduo, conferindo à citogenética convencional um papel crucial para o diagnóstico e prognóstico de diferentes tipos de cancro, assim como para a monitorização de doenças residuais.

A automação da análise do cariótipo tem como objetivo produzir um cariograma devidamente anotado e pronto para ser examinado pelo especialista. Desenvolver uma solução automática para detetar os cromossomas permite tornar o processo de cariotipagem mais rápido e exato. A segmentação de cromossomas ainda não foi aplicada em larga escala a nível clínico devido aos desafios da segmentação de *clusters* de cromossomas. Além disso, a falta de volume de *datasets* clínicos ou *datasets* sintéticos fotorrealistas é um obstáculo à automatização deste processo.

Este projeto propõe a automatização da segmentação de cromossomas em imagens celulares de cromossomas com padrão de bandas G. Utilizando um procedimento "*Cut, Paste and Learn*", é apresentada uma ferramenta para gerar imagens sintéticas com elevada variedade morfológica, ainda que baseadas em estruturas celulares reais da citogenética clínica. Foi usado o algoritmo YOLOv5 para a deteção de cromossomas com padrão de bandas G em imagens reais.

Na fase "*Cut*", foi obtido um *dataset* de estruturas celulares reais, composto por 115896 cromossomas, 180 núcleos interfásicos e 6024 objetos ruidosos. Na fase "*Paste*", foram obtidas 10795 imagens sintéticas, utilizando-se um método de *blending* proposto para a suavização das sobreposições das estruturas celulares. Na fase "*Learn*", obteve-se um valor de mAP@0.5 igual a 0.989 para o grupo de validação, composto por 1080 imagens celulares sintéticas. Este modelo foi testado

no *dataset* do LCG-FMUC, constituído por 171 microfotografias de células de cromossomas em prófase ou metafase. No total, foram segmentados 7708 cromossomas dos 7861 cromossomas existentes no *dataset* real, representando um sucesso de 98.05% na segmentação de cromossomas.

O trabalho desenvolvido disponibiliza uma ferramenta que automatiza a aquisição de imagens sintéticas fotorrealistas. Além disso, o algoritmo obtido é capaz de detetar cromossomas em microfotografias de células de cromossomas em prófase ou metafase com alta exatidão e velocidade.

Palavras-chave: Segmentação de Cromossomas; Cariótipo; *Machine Learning*; YOLOv5; Visão Computacional.

Abstract

Since the study of the karyotype entails the examination of a person's unique set of chromosomes, conventional cytogenetics plays a significant part in the diagnosis and prognosis of several types of cancer as well as in the surveillance of residual diseases.

The automated karyotype analysis seeks to create a karyogram that is properly annotated and prepared for the expert's review. The chromosomal detection method can be automated to increase speed and accuracy of karyotyping. Due to the difficulties in segmenting chromosomal clusters, chromosome segmentation has not yet been widely used in clinical settings. Additionally, a barrier to automating this procedure is the lack of clinical datasets or photorealistic synthetic datasets.

This research proposes an algorithm to perform the segmentation of chromosomes in microphotographies of chromosomes exhibiting a G-band pattern. By means of a "Cut, Paste, and Learn"procedure a method to generate synthetic images with great morphological variability is described, while still being based on actual cell structures of the cytogenetic clinic. The algorithm YOLOv5 was used to detect chromosomes with a G-band pattern in real images.

A dataset of real cell structures was produced during the "Cut" phase, including 115896 chromosomes, 180 nucleoli and 6024 noisy objects. Using a proposed blending technique for smoothing the overlapping of cell structures, 10795 synthetic images were created during the "Paste" phase. During the "Learn" phase a value of mAP@0.5 equal to 0.989 was obtained for the validation group, which included 1080 synthetic cell images. The LCG-FMUC dataset, which consists of 171 microphotographs of cells with prophase or metaphase chromosomes, was used to test the YOLOv5 model. Overall, from the 7861 chromosomes that were

contained in the experimental dataset, 7708 chromosomes were correctly segmented which accounts for 98.05% of success in chromosome segmentation.

This project proposes a tool that automates the generation of synthetic photorealistic images. Additionally, the developed algorithm can quickly and accurately identify prophase or metaphase chromosomes in microphotographs of cells.

Keywords: Chromosome Segmentation; Karyotype; Machine Learning; YOLOv5; Computer Vision.

Conteúdo

Siglas

A adenina.

aCGH Hibridização Genómica Comparativa em *array*.

AP *Average Precision*.

ARMS *Adaptive Receptive field Multi-Scale network*.

AUC *Area Under the Curve*.

bbox *Bounding Box*.

C citosina.

CNN Rede Neuronal Convolucional.

DA *Data Augmentation*.

DAPI *4',6'-diamino-2phenylindole*.

DL *deep learning*.

DNA ácido desoxirribonucleico.

DNN *deep neural networks*.

F1 *F1-score*.

FISH *Fluorescence in Situ Hybridization*.

FN Falsos Negativos.

FP Falsos Positivos.

G guanina.

IOU *Intersection Over Union*.

ISCN *International System for Human Cytogenomic Nomenclature*.

JSON *JavaScript Object Notation*.

LCG-FMUC Laboratório de Citogenética e Genómica da FMUC.

LMC Leucemia Mieloide Crónica.

mAP *mean Average Precision*.

Mask R-CNN *Region-based Convolutional Neural Network*.

NGS *Next Generation Sequencing*.

NORs Regiões Organizadoras do Nucléolo.

P *Precision*.

R *Recall*.

RGB *Red-Green-Blue*.

SKY Cariotipagem Espetral.

T timina.

TIF *Tagged Image File*.

TN Verdadeiros Negativos.

TP Verdadeiros Positivos.

TTA *Test Time Augmentation*.

TXT Ficheiro de Texto Simples.

YOLOv5 *You Only Look Once*.

Lista de Figuras

Lista de Tabelas

1

Introdução

1.1 Motivação e Contexto

O cromossoma humano é composto por ácido desoxirribonucleico (DNA) e proteínas. O DNA é responsável pela herança genética de cada indivíduo e por todas as caraterísticas a ele aliadas. O estudo do cariótipo representa a análise do conjunto de cromossomas de um indivíduo, sendo bastante importante na deteção de alterações genéticas relacionadas com o número de cromossomas e a estrutura dos mesmos [1]. Apesar do grande avanço nas técnicas de medida genética a nível do par de bases, como a tecnologia *Next Generation Sequencing* (NGS), a análise do cariótipo é aceite como uma técnica *gold standard* para a análise genética, constituindo o método mais compreensivo para a caracterização de cromossomas [2]. Este é um exame crucial para o diagnóstico e prognóstico de diferentes tipos de cancro, assim como para a monitorização de doenças residuais [3].

Para se obter um cariograma, isto é, a representação gráfica do cariótipo (Figura 1.1), os cromossomas são analisados em prófase ou metafase [1] e tratados num laboratório de citogenética de acordo com a técnica escolhida (por exemplo, bandagem G, C ou R) [2]. O cariograma é obtido através da identificação e classificação de cada cromossoma com base em caraterísticas como o tamanho, a posição do centrómero e o padrão de bandas. Esta análise cromossómica é feita em laboratórios de diagnóstico, com recurso a microscópios óticos e a citogeneticistas

[1] Apesar dos cromossomas serem analisados no estado prófase-metafase, para facilitar a leitura do documento, utilizou-se, na maior parte das vezes, a metafase ao longo do texto para se referir à fase da divisão celular dos cromossomas inseridos em imagens celulares relativas à citogenética convencional.

experientes que conseguem identificar visualmente, e de forma autónoma, os diferentes cromossomas [4]. O processo manual de obtenção do cariograma é demorado e requer muita atenção por parte do citogeneticista, resultando num maior tempo de espera para o paciente que, possivelmente, tem alguma alteração genética e necessita de um diagnóstico rápido. Ao longo deste processo, o especialista tem de analisar centenas de imagens e escolher aquelas que possuem cromossomas em metafase. De seguida, o citogeneticista analisa cada imagem selecionada de forma a contar e a identificar cada cromossoma. Este processo manual é ainda realizado em muitos laboratórios, sendo necessárias cerca de uma a duas semanas para realizar o diagnóstico de cada indivíduo [5].

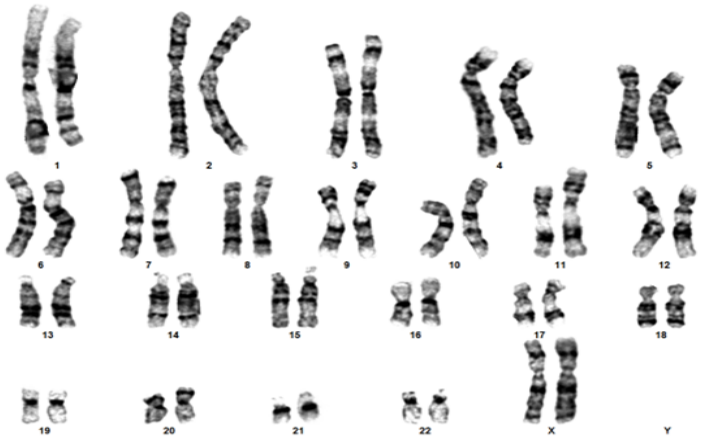


Figura 1.1: Exemplo de um cariograma. Imagem adaptada do *dataset* do Laboratório de Citogenética e Genómica da FMUC (LCG-FMUC).

Atualmente, recorre-se à fotografia digital das imagens cromossómicas em metafase para posterior identificação e classificação de cromossomas. Isto permite a aplicação de algoritmos de segmentação e classificação que auxiliem o citogeneticista [6]. Alguns laboratórios, como o Laboratório de Citogenética e Genómica da FMUC (LCG-FMUC), onde decorreu este projeto, já utilizam *software* com ferramentas para este propósito (por exemplo, o CytoVision [7]). Contudo, o processo de cariotipagem semiautomático é ainda lento e dispendioso em termos de recursos humanos, estando suscetível a erros. O citogeneticista apresenta-se como a principal fonte de erro para a obtenção do cariótipo, contribuindo para uma certa subjetividade que não deveria existir na análise de cariótipos [4]. Fatores como o cansaço, o nível de experiência e até o estado

anímico do citogeneticista são fatores preponderantes que podem afetar a exatidão e a fidelidade do cariótipo.

A automação na análise do cariótipo tem como objetivo descrever por completo o cariótipo de uma célula em metafase e produzir um cariograma devidamente anotado e pronto para ser lido pelo especialista de citogenética clínica [8]. Desde a década de 1980, vários investigadores têm abordado esta temática com recurso a algoritmos computacionais para análise de imagens cromossómicas, incluindo as várias etapas subentendidas na cariotipagem: identificação de células metafásicas de boa qualidade, segmentação de cromossomas, classificação dos diferentes cromossomas e deteção de padrões de bandas anormais [9]. O produto final de um sistema automático de cariotipagem representa um indivíduo, isto é, a composição genética do organismo que se está a analisar. As consequências da análise do seu cariograma podem ser fundamentais para a vida do mesmo. Assim, e devido à complexidade da tarefa a nível de visão computacional, um sistema totalmente automático continua uma utopia, sendo mais sensato melhorar as várias etapas do processo em vez de se investigar o sistema por inteiro. A imagem inicial retirada do laboratório contém, normalmente, objetos que não são cromossomas, nomeadamente *clusters* de cromossomas e outros ruídos (por exemplo, citoplasma ou partículas dos corantes usados). De forma a ser possível obter a produção fiel de um cariograma, primeiramente é necessário remover todo o ruído e isolar cada cromossoma. Esta etapa constitui a segmentação automática dos cromossomas em imagens microscópicas de cariótipos [8].

A segmentação de cromossomas é considerada a segunda etapa do *pipeline* da geração de um cariograma [4]. Apesar de já existirem bastantes estudos acerca de algoritmos capazes de realizar esta tarefa , nenhum destes é capaz de realizar com total exatidão a segmentação de cromossomas, sendo este um problema multifatorial. Desenvolver uma solução automática para detetar os cromossomas a partir de uma imagem celular em metafase tem então duas vantagens pilar: ser um processo mais rápido e ser um processo mais exato e objetivo [10]. Depois desta etapa ser otimizada e automatizada, a extração de *features* para posterior classificação dos cromossomas será menos errónea, tornando todo o processo mais fiel à realidade.

A segmentação automática de cariótipos ainda não foi aplicada em larga escala a

nível clínico, devido aos desafios da segmentação de *clusters* de cromossomas [11]. A estrutura não rígida dos cromossomas causa uma grande diversidade morfológica, contribuindo para o mau desemepnho dos métodos de segmentação, sejam estes baseados em *features* geométricas - definidas *à priori* - ou em modelos de *deep learning*, devido à dificuldade existente no *labelling* dos grupos de treino [12]. Os cromossomas sobrepostos ou a tocarem-se são então o maior entrave à tarefa de segmentação automática. Para se estudar o processo de segmentação é necessária a extração e análise de centenas ou milhares de imagens cromossómicas, o que se revela como outro impedimento ao estudo da automatização deste processo. A pouca disponibilidade de *datasets* clínicos para reprodução de algoritmos de segmentação desenvolvidos anteriormente ou para a exploração de novas técnicas é ainda um obstáculo à resolução desta temática [4].

Tendo em vista os problemas anteriormente mencionados e a importância da análise do cariótipo humano, a intenção deste projeto é automatizar o processo da segmentação de cromossomas em imagens de células em metafase e analisar as falhas dos algoritmos apresentados até à data. A metodologia proposta pretende analisar a fotografia retirada em laboratório, através do pré-processamento da imagem e segmentação dos cromossomas, preparando-os para a fase de cariotipagem seguinte. Isto é, a extração automática de *features* e posterior classificação. Numa primeira instância, o trabalho desenvolvido disponibiliza uma ferramenta para a geração de imagens metafásicas sintéticas a partir de imagens laboratoriais, visando combater a falta de *datasets* clínicos para o estudo desta temática. Relativamente à segmentação de cromossomas, é utilizada a rede neuronal *You Only Look Once* (YOLOv5), que através da estratégia de localização de objetos, fornece um modelo capaz de individualizar os diversos cromossomas em imagens reais provenientes do LCG-FMUC.

1.2 Objetivos

O propósito deste projeto é o desenvolvimento de uma técnica de segmentação automática de cromossomas em imagens celulares em metafase que supere as dificuldades dos atuais algoritmos. Para esta finalidade foram usadas imagens do Laboratório de Citogenética e Genómica da FMUC (LCG-FMUC).

Assim, foram definidos os seguintes quatro objetivos:

1. Levantamento e análise dos algoritmos de segmentação de cromossomas presentes na literatura. Escrutínio dos métodos aplicados e balanço de vantagens e desvantagens, tendo em conta os *datasets* utilizados para avaliar e validar o desempenho dos mesmos;

2. Disponibilização de um novo *dataset* clínico para replicação de resultados por outros investigadores - uso de técnicas de *Data Augmentation* (DA);

3. Desenvolvimento de um modelo de segmentação que seja capaz de ter como *input* uma imagem celular em metafase e que devolva os cromossomas individualizados, ultrapassando os obstáculos presentes na literatura - uso de técnicas de *deep learning* (DL);

4. Validação quantitativa e qualitativa dos resultados do algoritmo proposto para o *dataset* clínico proveniente do LCG-FMUC.

1.3 Estrutura da Dissertação

Esta dissertação encontra-se dividida em seis capítulos. O Capítulo 1, Introdução, serve para introduzir a temática explorada neste projeto e inclui três secções. Nestas são evidenciadas o contexto e motivação para o trabalho desenvolvido, salientando-se os objetivos do mesmo e a forma como a dissertação se encontra organizada.

O Capítulo 2, Estado da Arte, abrange os conceitos gerais necessários para a análise da temática. Este capítulo engloba também três secções. Primeiro, são introduzidos os conceitos biológicos relativos ao cromossoma, assim como as técnicas de citogenética associadas à visualização de cromossomas. De seguida, é apresentado o processo de cariotipagem e os problemas associados à segmentação de cromossomas. Posteriormente, são sumariados os principais algoritmos desenvolvidos até à data para a segmentação de cromossomas, estando estes divididos em métodos heurísticos e métodos de aprendizagem. Finalmente, faz-se uma análise comparativa entre os vários métodos. Este capítulo é o resultado de uma extensa análise da literatura e é essencial para que o leitor consiga abranger a dimensão e a complexidade da temática em discussão.

A metodologia proposta nesta dissertação é abordada no Capítulo 3, Materiais e Métodos. Na primeira secção, são apresentados os materiais usados, onde se explicitam os diversos ambientes de programação (*software*), assim como as caraterísticas das máquinas onde foram processados os algoritmos (*hardware*). Nesta fase, o *dataset* clínico proveniente do LCG-FMUC é apresentado e analisado. Na segunda secção, são discriminadas as diversas etapas do trabalho desenvolvido, organizadas por subcapítulos, onde são explorados os métodos empregues (pré-processamento do *dataset* clínico, aquisição de *labels* para o modelo YOLOv5, obtenção de imagens sintéticas e implementação do modelo de segmentação).

No Capítulo 4, Resultados e Discussão, estão dispostos os resultados da segmentação automática de cromossomas usando os métodos explicitados no capítulo anterior. A estrutura deste capítulo segue a linha de raciocínio do capítulo 3. Para cada fase do trabalho desenvolvido são apresentados e discutidos os respetivos resultados, encontrando-se estes divididos por secções. Recorre-se a exemplos de imagens representativas e métricas estatísticas para a análise gráfica e quantitativa dos resultados, respetivamente.

O Capítulo 5, Conclusão, contém as conclusões gerais do projeto. Os objetivos traçados inicialmente são analisados e refere-se o cumprimento ou não dos mesmos.

Por último, no Capítulo 6, Limitações e Trabalho Futuro, referem-se as limitações do algoritmo desenvolvido e apresentam-se estratégias e sugestões para trabalhos futuros dentro desta temática.

2

Estado da Arte

2.1 O Cromossoma

A citogenética é um ramo da genética, cujo começo é geralmente atribuído a Walther Flemming que publicou as primeiras ilustrações de cromossomas humanos em 1882. Só em 1924 é que surgiu o termo 'cariótipo', por Levitsky [13], referindo-se a este como um grupo ordenado de cromossomas. Mais de 20 anos depois, em 1959, Lejeune *et al.* [14] provaram que a Síndrome de Down está relacionada com um cromossoma extra. Nesse mesmo ano, foram descritas outras três síndromes relacionadas com os cromossomas (Turner, Klinefelter e disfunção sexual) [15]. Em 1960, Nowell e Hungerford [16] evidenciaram a presença de um pequeno cromossoma em pacientes com Leucemia Mieloide Crónica (LMC), demonstrando-se, pela primeira vez, a associação entre cromossomas e o cancro.

Na segunda metade do século XX, a citogenética começou a afirmar-se como uma área emergente, depois de Caspersson *et al.* [17], em 1968, terem colocado a hipótese de que a quinacrina mostarda (um corante) se ligava preferencialmente a resíduos de guanina e, por isso, as regiões cromossómicas ricas em ligações citosina-guanina produziam estrias mais claras, enquanto as regiões adenina-timina ficavam mais escuras. Desta forma, os autores produziram um padrão de bandas único para cada par de cromossomas, tornando possível a sua identificação. No entanto, o método utilizado por estes investigadores apresentava algumas desvantagens, nomeadamente o custo do material utilizado, sendo posteriormente melhorado por Drets e Shaw [18], que usaram um método que envolvia a coloração de Giemsa. Este método possui, por sua vez, várias modificações que facilitaram a aplicação da citogenética clínica

em larga escala desde a década de 1970 [15].

A partir da descoberta do padrão de bandas, seguiu-se a anotação de diversas anomalias cromossómicas e outras síndromes, tais como aneuploidias, deleções, microdeleções, translocações, inversões, inserções, mosaicismo e inúmeros rearranjos cromossómicos. As décadas que seguiram a 1970 viram surgir uma coleção de anomalias citogenéticas associadas com a neoplasia [15]. Em 1983, Felix Mitelman publicou o primeiro volume que serviu de catálogo para essas anomalias [19]. Atualmente, existe uma base de dados online com cerca de 70000 entradas [20].

Com a identificação de regiões cada vez mais pequenas do cariótipo humano, os genes começaram a ser mapeados dentro dos cromossomas a um maior ritmo. Tal foi feito recorrendo a sondas, surgindo o campo da citogenética molecular, onde se usam técnicas como *Fluorescence in Situ Hybridization* (FISH), Cariotipagem Espetral (SKY) e a Hibridização Genómica Comparativa em *array* (aCGH) [21].

Atualmente, mais de um milhão de análise citogenéticas são feitas anualmente, em mais de 400 laboratórios, a nível global. Segundo Wang *et al.* [22], as alterações cromossómicas são responsáveis por mais de 50% dos abortos espontâneos, nados mortos e mortes prematuras. Os pacientes que realizam este tipo de exames citogenéticos são, maioritariamente, mulheres grávidas com idades superiores aos 35 anos, crianças com dificuldades cognitivas e casais com problemas de fertilidade [15]. Devido ao fenótipo patogénico associado a alterações cromossómicas, é essencial que o cariótipo seja analisado de forma rápida, eficaz e exata.

2.1.1 Estrutura

O ácido desoxirribonucleico (DNA) é a matéria-prima do ser humano, influenciando todos os elementos da estrutura e função do nosso organismo [23]. Um cromossoma é composto por uma única molécula de DNA e pelas suas proteínas associadas. Os cromossomas humanos encontram-se no núcleo de todas as células, com exceção dos glóbulos vermelhos adultos [24]. Normalmente, cada célula humana apresenta 23 pares de cromossomas distintos, sendo que cada um é constituído por vários genes. Estes são as unidades funcionais de informação genética e são compostos

por sequências lineares de nucleótidos que codificam proteínas. Durante a divisão celular, a informação genética contida nos cromossomas é copiada e distribuída para as células-filhas [25]. A análise do cariótipo humano utiliza células em divisão celular, sendo, por isso, relevante explanar os conceitos associados a este processo.

A molécula de DNA, representada na Figura 2.1, é composta por uma estrutura em dupla hélice, com duas longas cadeias entrelaçadas constituídas por unidades repetitivas, os nucleótidos. Por sua vez, cada nucleótido é constituído por uma molécula de desoxirribose, um grupo fosfato e uma das quatro bases azotadas: adenina (A), guanina (G), citosina (C) ou timina (T). A ligação entre as duas cadeias de DNA é feita por complementaridade de bases A-T e C-G , enquanto a desoxirribose e o grupo fosfato formam a estrutura exterior da dupla hélice. A sequência de bases azotadas é crítica para a funcionalidade do DNA, uma vez que a informação genética é determinada pela ordem destas mesmas bases [24].

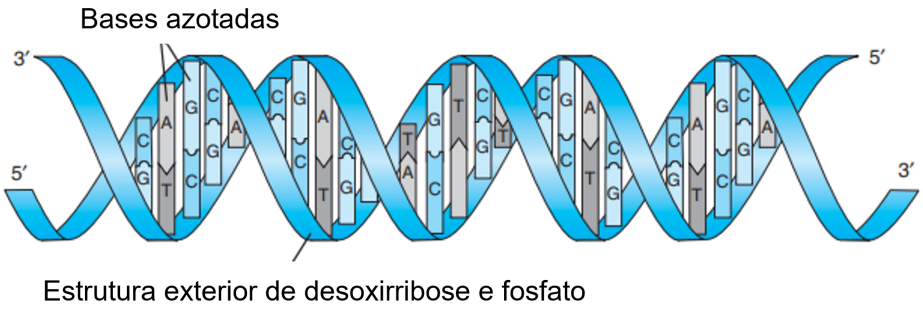


Figura 2.1: Esquema da estrutura do DNA. Imagem adaptada de Buckingham *et al.* [25].

Numa única célula humana diploide, se uma molécula de DNA fosse totalmente esticada, teria um comprimento que chegaria aos dois metros [26]. Assim sendo, esta molécula tem de ser condensada de forma a caber dentro do núcleo celular. Para isso, existem vários níveis de organização, onde os cromossomas representam a última fase de compactação, tornando-se visíveis ao microscópio ótico durante a divisão celular. A representação dos vários níveis de compactação encontra-se representada na Figura 2.2.

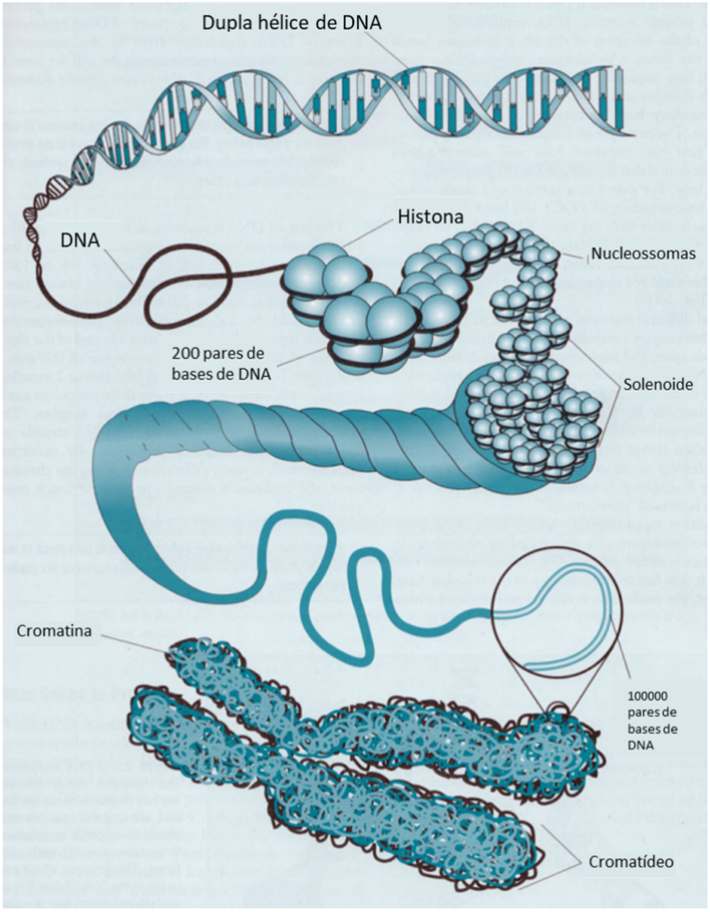


Figura 2.2: Esquema da condensação do DNA num cromossoma. Imagem adaptada de Gersen *et al.* [24].

Um cromossoma é constituído por dois cromatídeos irmãos, cada um contendo uma dupla hélice condensada de DNA, tal como explicado anteriormente. As diferentes áreas funcionais dos cromossomas são o centrómero, o telómero e as Regiões Organizadoras do Nucléolo (NORs) [24]. O centrómero é uma constrição visível dos cromossomas, onde os dois cromatídeos irmãos estão unidos, sendo essencial para a sobrevivência do cromossoma durante a divisão celular. Os cromossomas humanos são classificados com base na posição do centrómero (Figura 2.3): metacêntrico, se o centrómero estiver no meio do cromossoma; submetacênctrico, se o centrómero estiver entre o meio e a ponta do cromossoma; acrocêntrico, se o centrómero estiver numa das pontas do cromossoma [6].

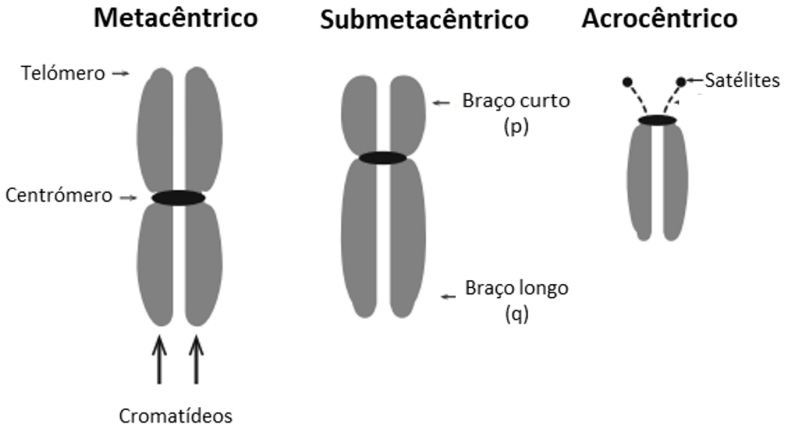


Figura 2.3: Classificação de cromossomas com base na posição do centrómero. Imagem adaptada de Subasinghe Arachchige *et al.* [6].

2.1.2 Divisão Celular

Para se entender o funcionamento da citogenética, é necessário compreender o processo de divisão celular. Existem dois tipos de divisão: a mitose e a meiose. A mitose acontece no processo de divisão de células somáticas, enquanto a meiose acontece nos gâmetas. Muitas das alterações citogenéticas resultam de erros destes dois mecanismos.

A mitose pode ser divida em quatro fases [27]: na prófase, os cromossomas vêm do seu máximo de elongação, não estando visíveis como estruturas distintas ao microscópio ótico - é aqui que começam a condensar-se e a tornarem-se visíveis. Na metafase, o fuso mitótico está completo, os centríolos dividem-se e movem-se para polos opostos, e os cromossomas alinham-se na placa equatorial - é de salientar que, na metafase, os cromossomas estão no seu máximo de condensação, sendo usados tradicionalmente nesta fase para estudos citogenéticos. Seguidamente, na anáfase, os centrómeros dividem-se longitudinalmente e os cromatídeos irmãos separam-se, migrando para polos opostos. Na última fase, a telófase, os cromossomas desenrolam-se, tornando-se outra vez indistintos, com a reconstrução da membrana nuclear. A telófase é usualmente seguida da citocinese, que corresponde à divisão do citoplasma para se obterem duas células separadas.

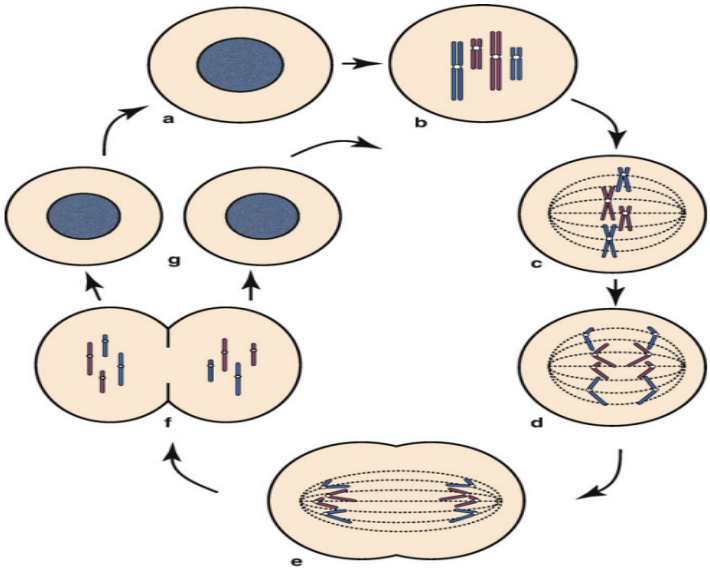


Figura 2.4: Esquema representativo da divisão celular (mitose) de dois pares de cromossomas: (**a**) Interfase, (**b**) Prófase, (**c**) metafase, (**d**) Anáfase, (**e**) Telófase, (**f**) Citocinese, (**g**) Interfase do próximo ciclo celular. Imagem adaptada de Gersen *et al.* [24].

Muitas das vezes é normal mencionar-se também a prómetafase, que corresponde a um período intermédio entre a prófase e metafase, onde a membrana nuclear desaparece e as fibras mitóticas começam a aparecer.

2.1.3 Visualização

Nos últimos anos, a robustez da tecnologia tem proporcionado um diagnóstico das alterações genéticas cada vez mais preciso a nível de alterações congénitas ou adquiridas. Para isso, existem diversas técnicas de citogenética que podem ser aplicadas, podendo ser agrupadas em citogenética convencional e abordagens moleculares.

A cariotipagem ou cariótipo convencional em padrão de bandas é reconhecido como *gold standard* para o diagnóstico e prognóstico genético [28]. O padrão de bandas cromossómicas permite verificar alterações do genoma que envolvam deleções, microdeleções, translocações, inversões e inserções, como já referido anteriormente. A técnica de padrão de bandas permite a obtenção de um cariograma que, após análise, pode associar alterações cromossómicas a síndromes clínicas ou a tumores. Apesar de ser uma técnica trabalhosa, com a possibilidade

de não deteção de alterações cromossómicas submicroscópicas e onde são necessárias cerca de duas semanas para se obterem resultados, esta técnica é aplicada em grande escala [29]. No laboratório de citogenética do LCG-FMUC são analisados cerca de 500 casos clínicos por ano.

Para superar as limitações da análise do padrão de bandas, técnicas de citogenética molecular, como FISH, SKY e aCGH, surgiram como ferramentas de diagnóstico alternativas e/ou complementares. Essas técnicas são amplamente utilizadas como coadjuvantes da citogenética tradicional para identificar alterações cromossómicas, especialmente a técnica aCGH (para informações adicionais consultar o Anexo A). No entanto, a técnica aCGH não é capaz de detetar anomalias equilibradas no cariótipo, tais como translocações recíprocas, que podem afetar um gene crítico para uma determinada função do organismo. Assim, a cariotipagem continua a ser o método preferencial para a análise do cariótipo em algumas situações clínicas.

Tal como mencionado anteriormente, os cromossomas são distinguíveis à luz de um microscópio ótico apenas quando se encontram devidamente corados e durante a divisão celular, principalmente, na metafase. Os cromossomas em metafase podem ser obtidos de amostras que contenham células em divisão espontânea ou culturas quimicamente induzidas para reprodução *in vitro*. Exemplos de amostras utilizadas são a medula óssea, nódulos linfáticos, biópsias a tecidos tumorais, fluido amniótico, vilosidades coriónicas e linfócitos do sangue periférico. A forma como a amostra é colhida e subsequentemente tratada vai influenciar a qualidade das metafases que vão ser posteriormente analisadas e existem também várias formas de cultura [30]. No caso do LCG-FMUC, são utilizadas amostras de sangue, trofoblasto ou pele, podendo ser utilizado um estimulante mitótico ou agentes que parem a divisão celular no estado prófase-metafase. A nível de cultura celular, pode ser necessário até uma semana para se obterem boas imagens celulares.

Entrando em mais detalhe na citologia convencional, através da coloração dos cromossomas é possível identificar o número e morfologia dos 23 pares de cromossomas do corpo humano. As principais técnicas de coloração de padrão de bandas designam-se de: G - Giemsa; R - *Reverse*; *C* - Heterocromatina Constitutiva, e DAPI - 4',6'-diamino-2phenylindole [31]. Na Figura 2.5 é possível observar um esquema de diferentes padrões de banda.

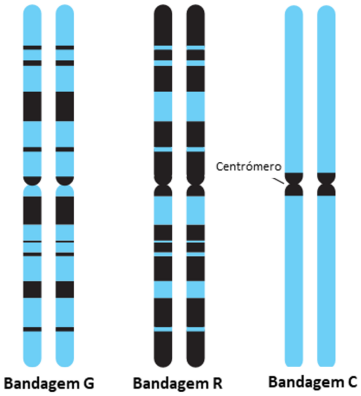


Figura 2.5: Esquema de diferentes tipos de padrões de banda. Imagem adaptada de Buckingham *et al.* [32].

As bandas cromossómicas facilitam a deteção de pequenas deleções, inserções, inversões e outras alterações em localizações cromossómicas distintas [33]. Para cumprir este propósito, o padrão de bandas G, que é o mais utilizado e reproduzível, foi ordenado em regiões, compreendendo bandas e subbandas, tal como pode ser observado na Figura 2.6. Regiões de heterocromatina, que tendem a ser ricas em adenina e timina, incorporam mais Giemsa e por isso aparecem como bandas escuras. Regiões de eucromatina, que tendem a ser ricas em guanina e citosina, incorporam menos Giemsa, aparecendo como bandas claras [32]. Com esta técnica é possível fazer-se a cariotipagem, isto é, a identificação, classificação e apresentação dos 23 pares de cromossomas numa única imagem, o cariograma [27].

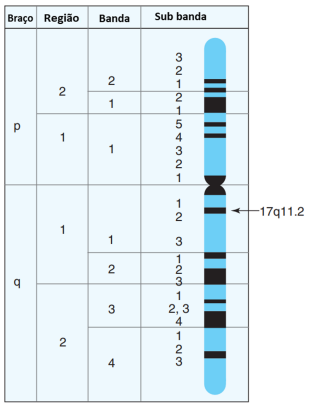


Figura 2.6: Esquema do padrão de bandas G para o cromossoma 17, com a respetiva divisão em bandas e sub-bandas. Imagem adaptada de Buckingham *et al.* [32].

A fase da divisão celular em que os cromossomas são fixados é deveras importante para a posterior análise. Da prófase para metafase há a diminuição da resolução de bandas, resultante do aumento da condensação do material genético - existe uma relação inversamente proporcional. Tal como se pode observar na Figura 2.7, existe a condensação progressiva do material genético, acompanhada da diminuição de bandas visíveis. Apesar de uma maior resolução ser mais vantajosa em termos de informação genética visível ao microscópio ótico, a probabilidade da existência de cromossomas sobrepostos aumenta, o que também aumenta a complexidade da análise da imagem metafásica. Assim, dependendo da amostra, da qualidade da técnica e do citogeneticista envolvidos na cariotipagem, é possível obterem-se diferentes níveis de resolução de bandas e, por isso, diferentes cariogramas.

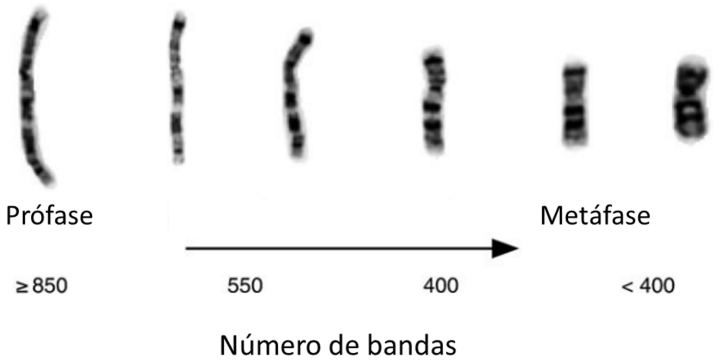


Figura 2.7: Diferença de resolução de bandas para o mesmo cromossoma. Células fixadas no final da prófase podem apresentar mais de 850 bandas (alta resolução de bandas), enquanto no final da metafase, devido à condensação do material genético, podem apresentar menos de 400 bandas (baixa resolução de bandas). Imagem adaptada de Bangs *et al.* (2005), Metaphase Chromosome Preparation from Cultured Peripheral Blood Cells. Current Protocols in Human Genetics, 45: 4.1.1-4.1.19.

Devido a esta subjetividade no processo de cariotipagem, no LCG-FMUC é feito um controlo de qualidade através de duas linhas de montagem, onde para cada caso é feita a análise com diferentes materiais - desde as amostras, às câmaras usadas ou ao citogeneticista. Quanto à validação de cada cariograma, entre outros critérios, é necessário que cada cromossoma seja observado pelo menos em duas metafases diferentes, sendo que quantas mais vezes o cromossoma for observado,

maior é a confiança de que o padrão analisado é fidedigno. Conclui-se, então, que é necessário um grande investimento para formar especialistas capazes de produzir estes cariogramas com alta fidelidade.

2.2 Segmentação Automática de Cromossomas

Atualmente, os laboratórios de citogenética estão repletos de tecnologia, garantindo algumas vantagens chave para a prática laboratorial: automatização de tarefas, maior eficácia e redução de custos; aumento na velocidade, exatidão e reprodutibilidade de tarefas; e a prática de tarefas que não podem ser feitas manualmente [34]. Apesar desta automatização, a natureza manual do processo de análise de cromossomas ainda é uma realidade. A análise de cromossomas requer um grande volume de trabalho por parte do especialista, algo que não se coaduna com o avanço tecnológico [35].

A análise do cariótipo humano requer especialistas com bastante experiência para separar os cromossomas de uma imagem microscópica e ordená-los de acordo com os critérios definidos pelo *International System for Human Cytogenomic Nomenclature* (ISCN) [36]. Dependendo não só da experiência dos especialistas, a inspeção visual das imagens microscópicas é também influenciada por fatores como fadiga e diminuição da atenção, podendo levar a resultados erróneos e/ou enviesados [12].

Na maior parte dos casos, a cariotipagem é ainda um processo totalmente manual, o que acarreta um peso subjetivo no levantamento do diagnóstico, podendo haver divergência de opinião no diagnóstico da mesma imagem por parte de diferentes especialistas [37]. Como a procura deste serviço é frequentemente elevada, o processo de cariotipagem verifica-se lento, laborioso e propenso a erros [10]. De acordo com Matta *et al.* [5], são necessários cerca de cinco dias para se avaliar as amostras de um só indivíduo, sendo que no LCG-FMUC a análise final do cariótipo de um indivíduo pode ser feita num só dia, considerando um citogeneticista com elevada experiência. Por estes motivos, é necessário encontrar uma solução completamente (ou tanto quanto possível) automática para a cariotipagem.

Nas imagens obtidas por padrão de bandas, os 46 cromossomas estão aleatoriamente

dispostos, tal como mostrado na Figura 2.8. A automatização da cariotipagem procura descrever, por completo e de forma autónoma, o cariótipo de um organismo, obtendo o cariograma anotado de uma célula em metafase. O citogeneticista, que representa o especialista capaz de detetar e interpretar anomalias cromossómicas, é responsável por elaborar o cariograma.

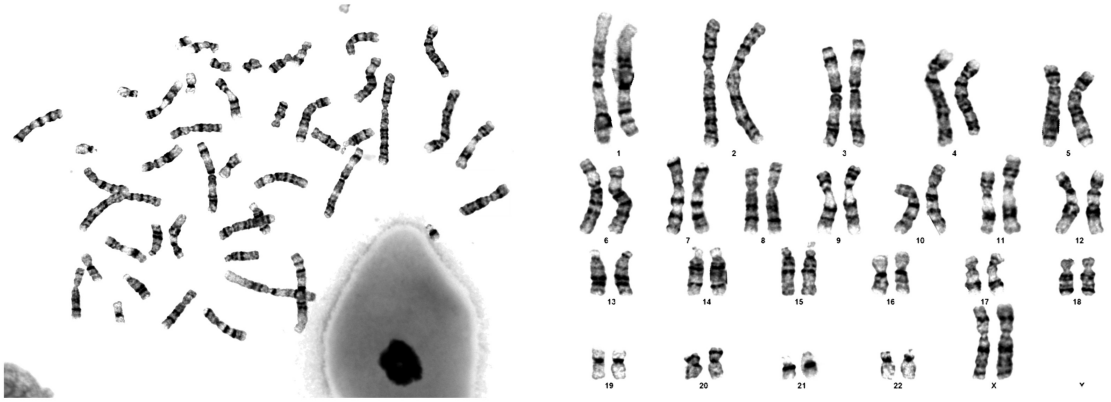


Figura 2.8: Exemplos de uma imagem celular de cromossomas em metafase captada pelo microscópio (à esquerda) e do respetivo cariograma (à direita). Imagens aleatoriamente escolhidas do *dataset* do LCG-FMUC.

Este processo da geração do cariograma segue os quatro passos representados na Figura 2.9 [9]: seleção da imagem a analisar; segmentação dos cromossomas presentes na imagem selecionada; extração das *features* de cada cromossoma segmentado; e a classificação de cada cromossoma numa das 24 classes cromossómicas. Para assegurar um cariograma fidedigno, a imagem celular escolhida para a segmentação e classificação de cromossomas deve ter o máximo de cromossomas individualizados [38]. Os cromossomas devem apresentar padrões de bandas distinguíveis e orientações o mais retas possível, evitando-se, ao máximo, *clusters* de cromossomas. O maior problema reside na segmentação de *clusters* de cromossomas, sejam estes compostos por cromossomas a tocarem-se ou parcial/totalmente sobrepostos.

Tendo em conta o tipo de amostra usada, o especialista necessita de um número diferente de imagens microscópicas. Segundo Arora *et al.* [4], em amostras de medula óssea são precisas 20 imagens de cromossomas em metafase. Apesar de não ser possível atribuir um número mínimo de metafases a analisar para se obter um

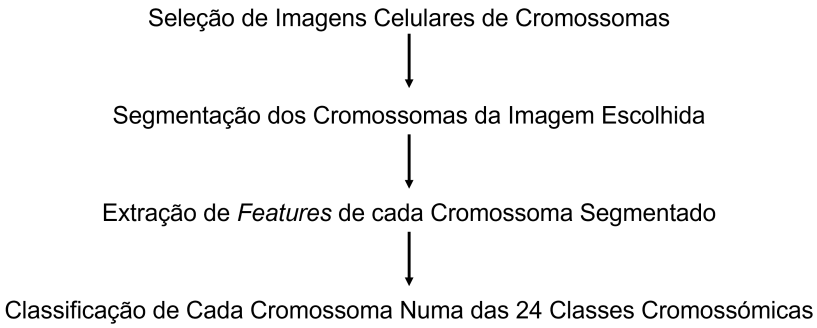


Figura 2.9: *Flowchart* da obtenção automática de um cariograma. Imagem original.

cariograma fidedigno, - uma vez que cada diagnóstico apresenta uma variedade morfológica diferente -, no LCG-FMUC são analisadas, no mínimo, 10 metafases nos casos mais simples, mas podem ser necessárias pelo menos 50 ou mais imagens celulares (por exemplo, no caso de deteção de mosaicismo). Assim sendo, o citogeneticista tem de analisar centenas de imagens diariamente para selecionar aquelas com as melhores metafases para os seguintes passos de cariotipagem [10].

A seleção manual deste tipo de imagens pelo citogeneticista é laboriosa, e, ao mesmo tempo, aumenta o risco de potenciais erros de observação [38]. Assim, a seleção da imagem microscópica de cromossomas em metafase, que constitui o primeiro passo no *flowchart* da geração de um cariograma, é uma tarefa que pode ser automatizada. Apesar de não ser o foco desta dissertação, a seleção de imagens com boa qualidade de cromossomas é importante para o sucesso da segmentação e posterior classificação de cromossomas. Por outro lado, dependendo do tipo de amostra, um sistema automático para a seleção de metafases não é imprescindível. Por exemplo, no caso de amostras de fluído amniótico ou sangue periférico, o contributo de um sistema destes não é tão relevante, dado que as metafases são, geralmente, de boa qualidade.

A segunda etapa na cariotipagem é a segmentação dos cromossomas, cujo objetivo final consiste na atribuição de um conjunto de píxeis a um único cromossoma, distinguindo-o por completo do seu ambiente envolvente [8]. O desempenho da segmentação vai influenciar diretamente a exatidão do cariograma obtido.

2.2.1 Obstáculos

As imagens resultantes do padrão de bandas G tendem a ter ruído de fundo devido ao procedimento de coloração com Giemsa, nomeadamente a existência de bolhas e problemas de contraste. Além do ruído de fundo, as imagens contêm também outros objetos sem ser cromossomas. Devido à estrutura não rígida dos cromossomas, estes podem ainda estar organizados em *clusters*, sejam estes constituídos por cromossomas a tocarem-se ou sobrepostos. Assim, com vista a segmentar os cromossomas, é necessário eliminar da imagem o ruído de fundo, objetos irrelevantes que não sejam cromossomas, e separar *clusters* de cromossomas, de forma a ser possível a individualização dos mesmos. Só depois se pode proceder à extração de *features* para a classificação final. Todo este processo de segmentação é ainda uma problemática em aberto [39].

A segmentação automática de cromossomas é, então, uma área ativa de pesquisa e investigação [40], podendo ser subdividida em duas fases: identificação de objetos, i.e., cromossomas individuais, *clusters* e outros objetos (Figura 2.10); e a separação de *clusters* (Figura 2.11). No caso dos *clusters*, apesar da região sobreposta não ser clinicamente relevante, dada a incerteza do padrão de bandas nessa região, as restantes partes dos cromossomas, que não estão sobrepostas, podem ajudar no diagnóstico, uma vez que podem ter uma boa resolução de bandas. Assim, é importante haver a identificação e separação destas regiões - quanto mais informação o citogeneticista puder obter de uma imagem, melhor e mais válido será o cariograma.

Figura 2.10: Tipos de estruturas cromossómicas presentes nas imagens microscópicas da citogenética convencional do *dataset* do LCG-FMUC. O nível de complexidade da estrutura cromossómica aumenta da esquerda para a direita e de cima para baixo. Imagem original.

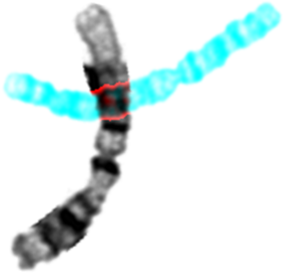


Figura 2.11: Exemplo da separação de cromossomas num *cluster* constituído por dois cromossomas. Imagem original.

Ao longo das últimas três décadas, muitos investigadores tentaram automatizar a segmentação de cromossomas [39]. Existem várias abordagens, que serão exploradas na próxima secção, mas nenhuma é capaz de fornecer resultados totalmente fidedignos, encontrando-se sempre limitações nos algoritmos desenvolvidos. As imagens celulares de cromossomas contêm também artefactos dependentes da preparação das amostras ou da fotografia microscópica. Nenhum tipo de processamento pode compensar uma imagem pobre devido a uma má configuração do microscópio ou a uma má preparação dos *slides* [34]. Para automatizar a cariotipagem, centenas ou milhares de imagens de cromossomas têm de ser analisadas e rotuladas por especialistas de forma a ser averiguada a exatidão dos métodos aplicados, o que requer ao citogeneticista dispensar tempo para tal. Estes fatores, associados à presença de *clusters* de cromossomas, aumentam a dificuldade da automatização da cariotipagem [4].

Desenvolver uma solução automática para a deteção de cromossomas tem dois impactos principais: processamento e análise mais rápido do diagnóstico, reduzindo o tempo de espera do paciente e a carga de trabalho do especialista; e redução do erro associado à cariotipagem, que pode levar a uma interpretação errónea do cariótipo humano [10]. Segundo Huang *et al.* [12], os três principais problemas associados à segmentação automática de cromossomas são:

• A natureza intrínseca não rígida dos cromossomas causa a diversidade morfológica e a incerteza nos *clusters*, que é a razão fundamental para o baixo desempenho de métodos de segmentação baseados em características geométricas. Também causa grande dificuldade nos métodos de aprendizagem, ao nível do *labelling* dos *datasets*

usados para implementar modelos de DL.

• A análise do cariótipo clínico exige que cada cromossoma seja completamente segmentado. O uso de cariogramas produzidos por cromossomas segmentados incorretamente pode levar os médicos a um diagnóstico erróneo. Por exemplo, um mau diagnóstico pré-natal pode levar à interrupção da gravidez erroneamente.

• Existência de poucos *datasets* clínicos públicos disponíveis para se reproduzirem resultados de segmentação de *clusters* propostos na literatura e para posterior comparação com novos métodos de segmentação de cromossomas.

2.3 Algoritmos de Segmentação de Cromossomas

A automatização de qualquer processo de cariotipagem que permita retirar tempo de trabalho ao citogeneticista é vantajosa, pois irá facilitar-lhe o trabalho e melhorar o resultado final. Em termos de obtenção do cariograma, no LCG-FMUC é utilizado o *software Cytovision*, tornando a tarefa semiautomática. Este programa fornece ferramentas ao técnico que lhe permitem arrastar e marcar cromossomas de forma intuitiva, mas ainda dependentes do especialista. Este *software* oferece ainda uma sugestão de cariograma para cada imagem celular, infelizmente sempre errónea, com sucesso inversamente proporcional à complexidade da imagem. Assim, a intervenção e experiência do especialista são imprescindíveis.

A tarefa de identificar e isolar cromossomas direitos e simples é relativamente fácil. No entanto, a complexidade da segmentação aumenta à medida que o número de sobreposições ou cromossomas adjacentes aumenta. A estrutura geométrica complexa de *clusters* de cromossomas torna o problema mais árduo.

Os principais algoritmos propostos na literatura e estudados ao longo do desenvolvimento desta dissertação encontram-se expostos nas secções seguintes. De acordo com a sua natureza, os métodos de segmentação de cromossomas podem ser classificados em duas categorias: métodos heurísticos e métodos de aprendizagem [41].

2.3.1 Métodos Heurísticos

Os métodos heurísticos abordam a segmentação de cromossomas com algoritmos e metodologias de visão computacional. Exemplos de técnicas utilizadas são a análise do histograma, *thresholding*, operações morfológicas, árvores de decisão, *active countour* ou a análise de caraterísticas geométricas. O principal problema destas técnicas é a dependência de parâmetros fixados *à priori*, que se tornam enviesados de acordo com o *dataset* usado na elaboração do algoritmo. Por outro lado, a implementação deste tipo de métodos costuma ser mais fácil, uma vez que não necessita de um grande conjunto de dados, e os métodos a ser usados já se encontram bem definidos no paradigma computacional.

Ji *et al.* [42] apresentaram um procedimento de segmentação de cromossomas baseado em regras e caraterísticas geométricas dos mesmos. Este algoritmo apresentou uma técnica para prever o número de cromossomas presentes numa imagem, o que serviu como controlo de qualidade para o resultado da segmentação. Apesar do bom desempenho do algoritmo, este método não foi testado em imagens com padrão de bandas G e, por vezes, as caraterísticas geométricas escolhidas são inconsistentes, variando de *dataset* para *dataset*.

Karvelis *et al.* [43] usaram uma metodologia baseada numa segmentação pelo método *Watershed*. Usando um *dataset* público composto por 200 imagens celulares marcadas com DAPI, atingiram uma exatidão de 82.4%. Contudo, esta base de dados já não se encontra *online*, sendo que não é possível analisar as imagens, nem reproduzir esses resultados.

Muitos algoritmos propostos usam as caraterísticas geométricas dos cromossomas para detetar *clusters*, através de sistemas de regras organizadas em árvores de decisão. Posteriormente, estes algoritmos analisam pontos de curvatura dos *clusters* para obter pontos de corte candidatos (*cutting points*). Por fim, traçam linhas de corte (*cutting lines*) entre esses pontos para extrair os cromossomas individualizados [44]–[47]. Estes algoritmos utilizam vários métodos intermédios, tais como *thinning*, *active contour* ou a triangulação de Delaunay. Saiyod *et al.* [48] comprovaram a eficácia da deteção de contornos em estruturas cromossómicas através de um algoritmo que usa *floodfill*, erosão e deteção *canny edge*.

Existem numerosos trabalhos realizados acerca de *thresholding* para segmentação de cromossomas. Andrade *et al.* [49] propuseram uma metodologia de segmentação baseada em *thresholding fuzzy*-adaptativo, sendo um bom exemplo da aplicação desta técnica. Subasinghe *et al.* [50] e Bashmail *et al.* [35] fizeram uso do *thresholding* de Otsu, e aplicaram outros métodos, como por exemplo filtros da mediana e gaussiano, respetivamente.

Existem também algumas ferramentas com interfaces gráficas para ajudar na segmentação de cromossomas, baseadas em métodos heurísticos. Exemplos disso são os trabalhos desenvolvidos por Uttamatanin *et al.* [38] e Altinordu *et al.* [51], que desenvolveram o *Metasel* e o *KaryoType*, respetivamente. Enquanto este último devolve métricas relacionadas com a estrutura geométrica dos cromossomas (por exemplo, o índice centromérico), o *Metasel* consegue classificar as estruturas em quatro classes (cromossoma individualizado direito, cromossoma individualizado dobrado, *clusters* e outros objetos). Apesar de serem ferramentas que podem apresentar algumas funcionalidades intuitivas para o especialista, ambas dependem do utilizador tendo, por isso, um erro significativo associado.

Wu *et al.* [52] elaboraram uma metodologia de segmentação de cromossomas baseada em características (forma, contraste ou a coloração uniforme dentro do cromossoma), com uma fase de separação de *clusters*. Para isso, o algoritmo usa uma forma elíptica para identificar cromossomas individualizados dando uso a um *threshold* definido empiricamente para descartar as elipses que não correspondem a cromossomas. Este trabalho diferenciou-se da restante literatura, pois tentou trabalhar com casos de sobre e subsegmentação.

Altinsoy *et al.* [39] propuseram um algoritmo que remove o *background* ruidoso com base num método adaptativo, através da avaliação do histograma da imagem. Posteriormente, usando uma metodologia baseada em regras relacionadas com caraterísticas geométricas, classifica os objetos em cromossomas, *clusters* ou objetos irrelevantes. Por fim, separa os *clusters* com base numa transformação geodésica de distância. Apesar deste trabalho ser o método heurístico mais consistente, de entre os analisados à data, este algoritmo requer que o utilizador trace manualmente o esqueleto de cromossomas sobrepostos.

2.3.2 Métodos de Aprendizagem

O *deep learning* (DL), um subconjunto do *machine learning*, utiliza um nível hierárquico de redes neuronais artificiais designadas de *deep neural networks* (DNN) para fazer o processo de aprendizagem. As DNNs têm a capacidade de aprender as *features* de grandes bases de dados devido à sua arquitetura em camadas. Nos últimos anos, tem surgido uma grande tendência do uso de DNNs para segmentar cromossomas. Em geral, estes métodos de aprendizagem, devido à sua conceptualização, não dependem de parâmetros empíricos definidos pelo ser humano. Assim, e como a morfologia das imagens de células em metafase é tão variada, estes métodos de aprendizagem apresentam-se como vantajosos face aos métodos heurísticos. Contudo, um entrave à aplicação de DNNs é a falta de dados rotulados. O principal problema com a segmentação de cromossomas por DL é a escassez de *labels* em imagens da citogenética clínica.

Xie *et al.* [53] propuseram uma rede neuronal designada de *Region-based Convolutional Neural Network* (Mask R-CNN) para segmentação semântica de cromossomas, isto é, individualização de cromossomas píxel a píxel. Apesar de se ter obtido uma exatidão de 91.673%, o *dataset* no qual o modelo obtido foi testado não é público e não se sabe o nível de complexidade das imagens. Huang *et al.* [12] testaram o desempenho de quatro redes neuronais na segmentação de *clusters* de cromossomas: Mask R-CNN, PathNet, Yolact e D2Det. Os autores pré treinaram os modelos no *dataset* COCO [54] e fizeram *fine-tuning* no *dataset* usado para treinar os modelos, concluindo que a Mask R-CNN obteve o melhor conjunto de métricas usadas. Feng *et al.* [55] propuseram uma melhoria à Mask R-CNN que introduziu informação relativa à orientação (designada de Mask Oriented R-CNN) para produzir segmentação de cromossomas de imagens celulares. Os resultados sugerem que este método é melhor do que o *baseline* Mask R-CNN.

Chen *et al.* [56] e Hu *et al.* [57] implementaram a UNet para segmentação semântica de *clusters* de cromossomas. Saleh *et al.* [58] vieram aumentar o desempenho da UNet, através da aplicação de uma técnica designada *Test Time Augmentation* (TTA). Song *et al.* [59] sugeriram uma nova Rede Neuronal Convolucional (CNN) baseada na estrutura da Unet, designada de *Compact Seg-UNet*, conseguindo atingir

melhores métricas do que a *baseline*. Mei *et al.* [41] apresentou uma metodologia que adaptou a NestedUnet à aprendizagem de *features* em multiescala para detetar regiões cromossómicas sobrepostas, conseguindo uma exatidão de 99.9776% num *dataset* artificial.

Andrade *et al.* [10] sugerem um algoritmo para a classificação de estruturas em cromossomas individualizados e *clusters*. Para isso, testaram vários modelos de DL: VGG16, VGG19, Inception_v3, MobileNet, Xception, MiniVGG e Sharma Model. O melhor desempenho foi obtida com a arquitetura VGG. Os autores sugeriram ainda um método de deteção de falsos positivos e negativos, o que também aumentou o desempenho do algoritmo.

Wang *et al.* [22] usaram a rede *Adaptive Receptive field Multi-Scale network* (ARMS) que extrai, de forma adaptativa, *features* multi-escala e consegue lidar com falta de informação semântica nas imagens, atingindo um alto desempenho na segmentação de cromossomas sobrepostos. O trabalho destes investigadores focou-se na segmentação detalhada píxel a píxel das regiões sobrepostas de cromossomas com diferentes tamanhos, identificando três tipos de cenários de *clusters*: cromossomas com as pontas a tocarem-se; cromossomas cruzados; e cromossomas sobrepostos quase na sua totalidade. O modelo obtido da ARMS teve uma exatidão de 99.99% num *dataset* artificial. Lin *et al.* [11], [60] utilizaram um modelo obtido a partir da *ResNeXt Weakly-Supervised Learning* (WLS) para a classificação de objetos em cromossomas individualizados, a tocarem-se, sobrepostos ou a tocarem-se e sobrepostos. Atingiu uma exatidão de 94.47% num *dataset* clínico privado.

Bai *et al.* [61] usaram uma metodologia dividida em três fases, dando uso à UNet e ao YOLOv3. Primeiramente, a UNet remove o ruído de fundo, tal como núcleos interfásicos e outras interferências. Seguidamente, o YOLOv3 deteta e extrai cada cromossoma. Finalmente, a UNet é novamente aplicada de forma a extrair os píxeis relativos aos cromossomas detetados pelo YOLOv3. O YOLOv3 é um algoritmo de deteção a uma fase, onde a deteção de um alvo se apresenta como um problema de regressão. Este modelo não requer processos complicados para gerar limites candidatos para os objetos em causa, neste caso os cromossomas. Além disso, a sua rede multi-escala permite detetar objetos de várias dimensões, conseguindo detetar

os mais variados cromossomas. O algoritmo foi testado em imagens reais de clínica, compostas por 2300 cromossomas, obtendo-se uma exatidão de 99.3%. Comparado a outros modelos de *deep learning*, esta metodologia é a mais fidedigna, resolvendo a questão de segmentação *end to end*.

Tendo em conta a complexidade em rotular grandes quantidades de imagens de células em metafase, a maioria dos autores mencionados nesta subsecção deparou-se com falta de dados. Para contornar o problema, esses autores tiveram de aplicar estratégias de *Data Augmentation* (DA) a fim de conseguirem milhares de cromossomas anotados. Inclusivamente, Sharma *et al.* [62] recorreu a *crowndsourcing* de forma a obterem grandes quantidades de *labels*.

A maioria dos investigadores recorreu, direta ou indiretamente, a transformações *affine*, como rotação, translação, *flipping* horizontal ou vertical, redimensionamento da escala da imagem ou recortes [12], [22], [41], [58], [61]. No entanto, existem outras estratégias capazes de gerar imagens sintéticas com mais informação para treinar os modelos de DL.

Chen *et al.* [56] sintetizaram *clusters* de cromossomas, através da colagem de cromossomas. A cada colagem, a região sobreposta é submetida a duas estratégias: primeiro, utiliza o somatório das intensidades dos pixeis sobrepostos; segundo, utiliza um *merge*, com pesos aleatórios para cada um dos dois cromossomas. Song *et al.* [59] também enveredaram pelo mesmo caminho, mas modificando a opacidade das regiões sobrepostas.

Xie *et al.* [53] utilizaram o método "*Cut and Paste*", que consiste em colar vários elementos de uma imagem de citogenética clínica. Estes investigadores, a fim de criarem o seu *dataset*, deram uso a objetos irrelevantes, tipos de *background* e cromossomas provenientes de imagens reais. Primeiro, recortaram esses elementos, e depois colaram-nos, escolhendo ter 48 cromossomas e 2-6 objetos irrelevantes por imagem. Para aumentar a robustez do modelo treino usado, recorreram a três padrões de *blending* diferentes (*blur* gaussiano, *blur* de movimento e *blur* de caixa).

2.3.3 Análise Comparativa

Os métodos heurísticos podem alcançar bons resultados de segmentação, mas são muito sensíveis à forma geométrica e às regiões sobrepostas dos cromossomas. Além disso, como estes métodos não consideram caraterísticas não designadas *à priori*, o desempenho e a aplicabilidade destes métodos são limitados e difíceis de analisar em grandes *datasets*.

Os métodos de aprendizagem aplicam técnicas de DL para extrair potenciais informações das imagens que possam ser úteis para segmentação de cromossomas. Estes métodos, após serem implementados, podem conduzir a segmentação, de forma independentemente e autónoma. No entanto, a arquitetura dos algoritmos usados não consegue compensar a falta de grandes quantidades de dados no que diz respeito a imagens realistas de cromossomas em metafase. Apesar da maioria dos modelos presentes na literatura mostrarem métricas altamente satisfatórias, a qualidade dos *datasets* usados no treino e teste desses modelos apresenta-se sempre como insuficiente. A maioria dos *datasets* contêm imagens, onde cada uma inclui um só *cluster* de apenas dois cromossomas parcialmente sobrepostos, amenizando a complexidade morfológica de uma imagem celular cromossómica. Para além disso, há alguns autores que não disponibilizam o *dataset* teste, e, por isso, torna-se impossível avaliar a veracidade do desempenho do modelo em causa.

Há ainda autores que tentam abordar a segmentação de cromossomas com modelos híbridos, juntando métodos heurísticos e métodos de aprendizagem. Exemplo disso é o trabalho desenvolvido por Cao *et al.* [63] e Huang *et al.* [64], que tentam tirar partido das vantagens dos dois tipos de métodos. Ao dividirem as suas metodologias por fases, conseguem, por exemplo, usar algoritmos de regressão logística para identificar *clusters* e apenas usar modelos de DL para segmentar as sobreposições. Os autores referem que apenas ao aplicarem a rede neuronal a uma fase específica do algoritmo, em vez de a aplicarem a toda a imagem celular, conseguiram obter resultados melhores e mais rápidos.

Tendo em conta as limitações referidas anteriormente, foi elaborada uma tabela que resume os algoritmos analisados, explicitando algumas caraterísticas envolvidas nesses estudos, tais como o *dataset* usado ou a métrica com maior

relevância (consultar Anexo B). Além disso, analisando o artigo de revisão elaborado por Sathyan *et al.* [37], onde se explorou em grande detalhe a literatura associada à automatização da cariotipagem, verificou-se que uma das direções fundamentais para automatizar a cariotipagem é o desenvolvimento de algoritmos automáticos de segmentação. Seja baseado em métodos heurísticos ou de aprendizagem, um algoritmo de segmentação de cromossomas totalmente automático irá resolver a necessidade de intervenção humana na separação de cromossomas sobrepostos ou a tocarem-se. Para isso, é também necessária a criação de ferramentas capazes de gerar dados (neste caso, imagens de células em metafase rotuladas) de forma fácil e sistemática.

3

Materiais e Métodos

Este capítulo descreve os materiais usados nesta dissertação, e apresenta também o raciocínio que levou à escolha dos métodos aplicados na segmentação de cromossomas. Por forma a cumprir os objetivos mencionados no Capítulo 1, foram utilizadas imagens de células em metafase provenientes do LCG-FMUC, recorrendo-se a métodos de visão computacional (por exemplo, *Data Augmentation* e *Deep Learning*) programados na linguagem Python. O *dataset* artificial fotorrealista obtido, assim como o código elaborado, encontram-se publicados *online* no GitHub [1].

Este terceiro capítulo encontra-se dividido em duas secções: na Secção 3.1, Materiais, é descrito o *dataset* disponibilizado pelo LCG-FMUC, bem como o *software* e o *hardware* usados para aplicar a metodologia proposta; e na Secção 3.2, Métodos, apresentam-se as várias fases do desenvolvimento da parte prática desta dissertação.

3.1 Materiais

3.1.1 *Dataset* do LCG-FMUC

Este estudo só foi possível graças à colaboração com o Laboratório de Citogenética e Genómica da FMUC (LCG-FMUC), que providenciou as imagens necessárias para a aplicação dos métodos de segmentação e processamento de imagem. Uma vez que o LCG-FMUC recorre ao *software* Cytovision [7], foi neste realizada uma tentativa inicial de extrair os ficheiros automaticamente do sistema. Contudo, as imagens

[1] `https://github.com/manuelfslg/Chromosome_Segmentation_ManuelGoncalves`

recolhidas encontram-se encriptadas tornando-as impossíveis de serem analisadas fora desse mesmo *software*.

Posto isto, foi pedido aos especialistas do laboratório para retirarem, manualmente, imagens do Cytovision. Este processo compreendia abrir e guardar cada ficheiro de imagem celular em metafase e o correspondente cariograma. Tal resultou em 171 pares de imagens, sendo cada par composto por dois ficheiros de formato *Tagged Image File* (TIF): uma imagem celular dos cromossomas em metafase e uma imagem do respetivo cariograma anotado (Figuras 3.1 e 3.2, respetivamente).

Todas as imagens fornecidas contêm: *clusters* de cromossomas, ruído referente à pigmentação ou outros artefactos devido ao sistema de imagem utilizado, e ainda objetos não cromossómicos, como por exemplo núcleos interfásicos. É também de salientar que a maioria das imagens contêm algum pós-processamento resultante do *software* Cytovision. Este processamento é observado pelas cores vermelho e verde, que significam contornos de ruído ou contornos de cromossomas que estavam muito afastados do campo de visão da câmara, respetivamente (Figura 3.1). No Anexo D podem ser observados exemplos deste *dataset*.

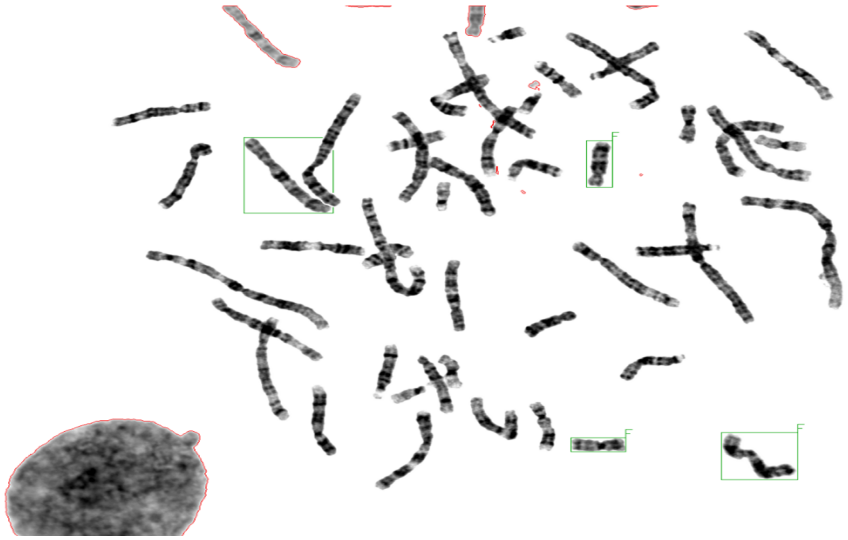


Figura 3.1: Exemplo de uma imagem celular de cromossomas em metafase obtida pelo LCG-FMUC. Imagem escolhida aleatoriamente do *dataset* do LCG-FMUC.

O *dataset* inicial é, então, constituído por 342 imagens com caraterísticas diferentes entre elas. As imagens assumem larguras de 744 a 1376 píxeis e alturas de 568 a 1024 píxeis, retratando-se desta forma diferentes escalas nas microfotografias realizadas

Figura 3.2: Exemplo de um cariograma obtido pelo LCG-FMUC. Imagem escolhida aleatoriamente do *dataset* do LCG-FMUC.

no microscópio ótico. Relativamente à resolução das imagens, conclui-se que existe também uma grande variedade, estando esta intervalada entre 84 e 394 kB. Através de uma inspeção visual do *dataset*, infere-se ainda que os cromossomas retratados nas imagens apresentam diferentes resoluções ao nível do padrão de bandas (Figura 3.3), conferindo a este *dataset* uma maior variabilidade em termos da fase de fixação de cromossomas no estado prófase-metafase.

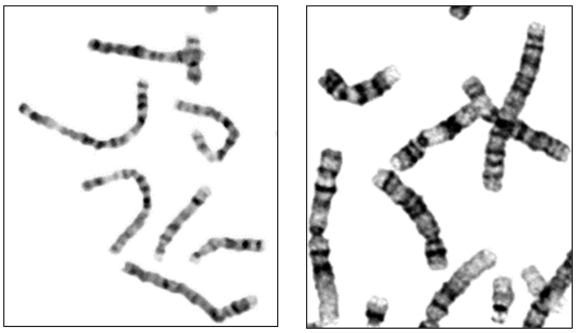


Figura 3.3: Diferença de resolução de bandas em cromossomas entre dois recortes de imagens celulares provenientes do *dataset* do LCG-FMUC. Na imagem da esquerda observa-se uma maior resolução de bandas, devido à descondensação do material genético. Na imagem à direita observa-se uma menor resolução de bandas, devido à condensação do material genético. Imagem original.

Dos 171 cariogramas obtidos, 66 correspondem ao género masculino e 105 ao género feminino. Apesar das imagens terem sido escolhidas de forma aleatória pelo especialista, a diferença de casos deve-se, provavelmente, a um maior número de

exames de citogenética convencional realizados por mulheres em estado pré-natal.

3.1.2 *Software*

3.1.2.1 Python

Ao longo do desenvolvimento deste trabalho foi usada a versão Python 3.8.5 para o processamento de toda a análise e segmentação de cromossomas a partir do *dataset* fornecido pelo LCG-FMUC. A linguagem Python foi escolhida devido à sua popularidade no campo da visão computacional, à sua sintaxe intuitiva e devido também ao facto de ser uma linguagem *open source*, existindo uma grande documentação e bibliotecas disponíveis *online*.

3.1.2.2 Spyder e Google Colaboratory

Relativamente ao ambiente de desenvolvimento, foram usados dois ambientes de programação distintos, um local (recorrendo ao ambiente de desenvolvimento integrado Spyder) e um na *cloud* (recorrendo ao *notebook* Colab). Todo o código foi desenvolvido e testado localmente no Spyder 4.2.3 [2]. Uma vez que foram utilizados grandes números de imagens e modelos de segmentação de DL, o serviço gratuito fornecido pelo Google Colaboratory [3] permitiu processar código na *cloud* e, em simultâneo, implementar novas metodologias na máquina local. Além disso, o Colab permite ao utilizador utilizar um acelerador GPU, sendo possível obter resultados mais rápidos no caso de modelos de DL [65]. A Google Drive foi utilizada para gerir ficheiros entre a máquina local e o Colab.

3.1.2.3 LabelMe

O *software* LabelMe [4] [66] foi criado pelo laboratório de *Computer Science and Artificial Inteligence* do MIT, e consiste numa ferramenta simples de anotação de imagens com fim à criação de *datasets* para investigação em visão computacional. O programa é gratuito e pode ser usado *online* ou localmente. Através deste *software* intuitivo tornou-se possível o *labelling* de cromossomas a partir de cariogramas no *dataset* fornecido pelo LCG-FMUC (Figura 3.4).

[2] https://www.spyder-ide.org/
[3] https://colab.research.google.com/
[4] http://labelme.csail.mit.edu/Release3.0/

Figura 3.4: *Software* LabelMe [66] utilizado para anotação de cromossomas em cariogramas. Imagem original.

3.1.2.4 Git e GitHub

O Git [5] é um sistema que permite registar o histórico de edições de qualquer tipo de arquivo, enquanto o GitHub [6] é uma plataforma de hospedagem e controlo de arquivos que funciona em *cloud*. Estas duas ferramentas foram utilizadas para controlar o desenvolvimento do código e para disponibilizar não só o *dataset* artificial fotorrealista, mas também as metodologias de segmentação propostas nesta dissertação. O objetivo desta partilha é facilitar a abordagem desta temática, através da reprodução e comparação de resultados. O GitHub também foi utilizado para a visualização de *datasets* e a implementação de algoritmos de outros investigadores da mesma área.

3.1.3 Hardware

O computador utilizado para desenvolver e implementar a metodologia proposta está equipado com um processador Intel Core i5-10300H CPU @ 2.50GHz, 2496 Mhz, 4 Núcleos, 8 Processadores Lógicos e 16 GB RAM. No Google Colaboratory, o computador virtual gerado tem os seguintes recursos: NVIDIA Tesla K80 GPU e 12 GB RAM.

[5] https://git-scm.com/
[6] https://github.com/

3.2 Métodos

A metodologia descrita nesta secção pretende colmatar as dificuldades relativas aos algoritmos de segmentação de cromossomas estudados e apresentados no Capítulo 2, e ao mesmo tempo satisfazer os objetivos desta dissertação.

Para isso, é seguida a linha de pensamento de Dwibedicut *et al.* [67], através da aplicação de um procedimento "*Cut, Paste and Learn*", como se encontra esquematizado na Figura 3.5. Os autores mostraram que o fotorrealismo de colagens de imagens de objetos em imagens de *background* permite criar grandes *datasets*. Por sua vez, estes *datasets* sintéticos culminam na melhoria do desempenho de modelos de deteção de objetos. Além disso, depois de se investigarem os vários métodos de DA, concluiu-se que a técnica mais fotorrealista para se obter um elevado número de imagens de células em metafase é conseguida através de colagem de estruturas celulares. Assim sendo, torna-se possível gerar imagens sintéticas com elevada variedade morfológica, mas baseadas em estruturas celulares reais da citogenética clínica. Contudo, é necessário contornar os artefactos gerados na colagem de objetos e que são formados, principalmente, nas bordas dos objetos que são colados. Caso estes artefactos não sejam corrigidos, o desempenho do modelo de treino pode ser negativamente afetado.

Desta forma, é sugerida, na Secção 3.2.2, uma abordagem à fase "*Cut*", onde é sistematizada a estratégia para recortar e individualizar estruturas celulares a partir de imagens reais do *dataset* do LCG-FMUC. Seguidamente, na Secção 3.2.3, é proposta uma ferramenta capaz de criar automaticamente imagens de células em metafase sintéticas, representando a fase "*Paste*". Por último, na Secção 3.2.4, dá-se uso ao modelo YOLOv5 para a deteção de cromossomas em imagens reais do *dataset* do LCG-FMUC. Além disso, é feito um pré-processamento inicial a este mesmo *dataset*, descrito na Secção 3.2.1.

3.2.1 Pré-processamento do *Dataset* do LCG-FMUC

O *dataset* fornecido pelo LCG-FMUC, caraterizado na Secção 3.1 contém algum processamento, tal como explicado anteriormente. Com a finalidade de utilizar as imagens de células em metafase para testar o algoritmo proposto, foi necessário fazer

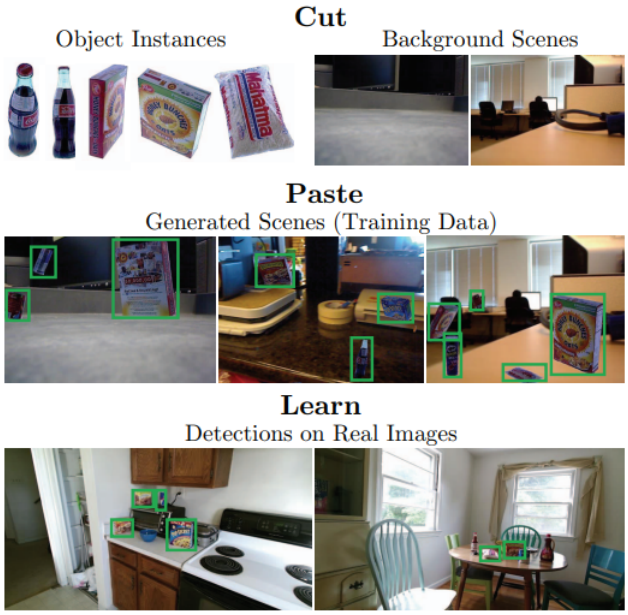


Figura 3.5: Esquema da metodologia *Cut, Paste and Learn* usada para a segmentação de cromossomas em imagens de células em metafase. Imagem adaptada de Dwibedicut *et al.* [67].

um pré-processamento destas mesmas imagens.

Ao longo da metodologia foi usada a biblioteca *OpenCV* que permite, entre outras funcionalidades, ler ficheiros TIF através da função *imread*, retornando uma matriz representativa da imagem com três dimensões: altura x largura x canais de cor, onde os canais de cor correspondem a vermelho, verde e azul - imagem *Red-Green-Blue* (RGB). Cada píxel da imagem é representado pela junção dos valores de cada canal de cor respetivos a esse píxel. Por sua vez, cada canal de cor é representado por valores inteiros entre 0 e 255, inclusive, onde 0 corresponde à ausência total da cor em questão e 255 corresponde à intensidade máxima desse canal.

A fim de eliminar as anotações a vermelho presentes nestas imagens RGB, substituíram-se todos os pixéis cujo canal *Red* estivesse no seu máximo de intensidade (igual a 255) e os correspondentes canais *Blue* e *Green* estivessem no mínimo de intensidade (igual 0) por pixéis brancos (correspondentes a ter todos os canais no máximo de intensidade). O mesmo raciocínio foi aplicado para as anotações a verde.

Após a eliminação das cores presentes nas imagens celulares, estas foram convertidas

para uma escala de cinzentos, de forma a minimizar o tamanho dos ficheiros a serem processados pelos métodos de visão computacional propostos. Tal foi possível, dado que as microfotografias retiradas ao microscópio estão originalmente nesta escala de cinzentos. Nesta escala, todos os canais de cor têm a mesma intensidade para o mesmo píxel, sendo desnecessário trabalhar com matrizes a três dimensões. A conversão de escalas traduziu-se na representação das imagens do *dataset* do LCG-FMUC por matrizes a duas dimensões: altura x largura.

Em alguns casos, as anotações a vermelho e verde sobrepunham-se a estruturas celulares, como cromossomas ou núcleos interfásicos. Assim, a remoção de pixéis a cores e respetiva substituição por pixéis brancos levou à criação de buracos nessas estruturas, não correspondendo à realidade das microfotografias retiradas em laboratório. A fim de resolver esta incoerência, foi utilizado um filtro de caixa que atribui a cada píxel o valor médio de intensidades dos pixéis vizinhos (exemplo demonstrativo na Figura 3.6) [68]. Este filtro foi aplicado apenas aos pixéis que foram modificados anteriormente, com um *kernel* de 3x3 que engloba os 8 pixéis vizinhos por cada iteração (Equação 3.1). Definiram-se, empiricamente, três iterações deste filtro por cada píxel modificado.

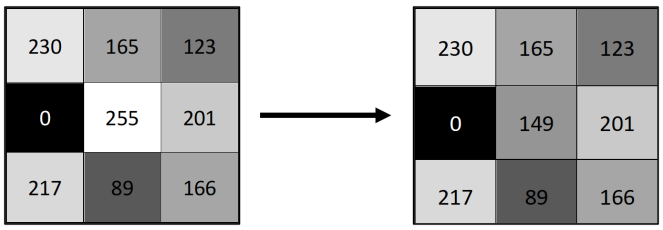


Figura 3.6: Esquema do filtro de caixa utilizado para suavização dos píxeis substituídos no *dataset* do LCG-FMUC. Imagem original.

$$f\,(i,j) = \frac{1}{8}\sum_{l=-1}^{1}\sum_{k=-1}^{1} f\,(i+l, j+k) \times w\,(l,k) \tag{3.1}$$

$$,\ onde \qquad w\,(l,k) = \begin{cases} 1, & se \quad l \neq 0 \quad ou \quad k \neq 0 \\ 0, & se \quad l = 0 \quad e \quad k = 0 \end{cases}$$

Este pré-processamento permitiu preparar as imagens de células em metafase para a fase de teste do algoritmo de segmentação proposto. O código em Python referente a este pré-processamento encontra-se no *script* "*dataset_preparation.py*" disponível no GitHub.

3.2.2 Aquisição de Imagens de Estruturas Celulares - "*Cut*"

Um dos entraves à investigação de algoritmos de segmentação automática de cromossomas é a falta de dados para testar e treinar modelos de segmentação. Dado que esta dissertação aborda esta temática recorrendo a métodos de aprendizagem, tornou-se necessária a criação de uma estratégia para a obtenção de estruturas celulares e respetivas *labels* a partir de imagens reais. Esta secção apresenta a fase "*Cut*" da metodologia proposta.

A ferramenta LabelMe [66] foi utilizada para fazer o *labelling* de estruturas celulares. Para cada imagem rotulada foi gerado um ficheiro *JavaScript Object Notation* (JSON) contendo as *labels* respetivas a cada estrutura celular. Neste ficheiro, cada estrutura está associada a um nome, a um grupo de identificação e às coordenadas da caixa retangular - *Bounding Box* (bbox) - que isola a estrutura celular em causa. Como o *dataset* do LCG-FMUC é composto por imagens celulares e pelos respetivos cariogramas, foi possível obter *labels* de três tipos de estruturas celulares (Figura 3.7):

- cromossomas, rotulados a partir dos cariogramas;
- núcleos interfásicos, rotulados a partir das imagens celulares;
- objetos ruidosos, rotulados a partir das imagens celulares.

É importante que os limites das bboxes estejam o mais próximo das extremidades da estrutura a ser rotulada para um melhor desempenho dos modelos de DL. Quanto à marcação da *label* em si, esta é feita por dois pontos que correspondem aos dois vértices opostos da caixa retangular que isola a estrutura celular a ser anotada. É obrigatório que o primeiro ponto a ser marcado seja o vértice superior esquerdo e o segundo ponto corresponda ao vértice inferior direito (Figura 3.8). Esta forma de

marcação das bboxes é essencial para o passo seguinte de recorte das estruturas a partir das imagens rotuladas.

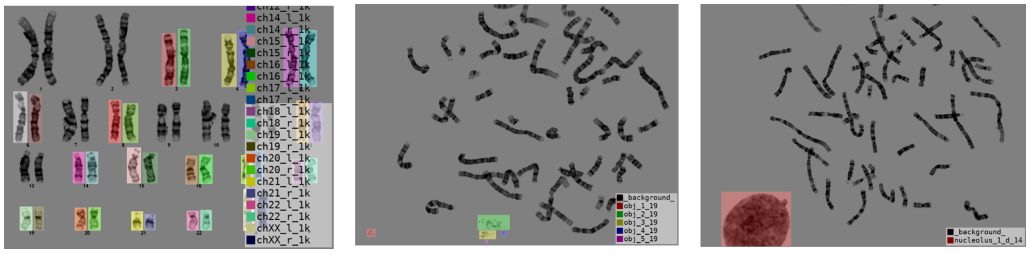


Figura 3.7: Exemplo do *labelling* de estruturas celulares no LabelMe. À esquerda, observam-se *labels* de cromossomas num cariograma. No meio e à direita, observam-se *labels* de objetos ruidosos e núcleos interfásicos, numa imagem celular, respetivamente. Imagem original.

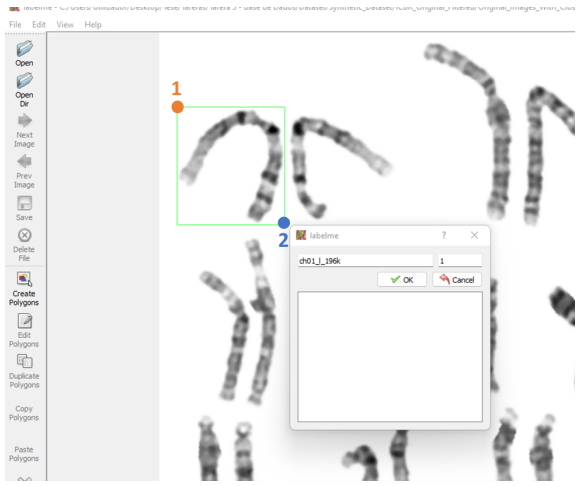


Figura 3.8: Ordem de marcação dois vértices da bbox de cada *label*, através do *software* LabelMe. Imagem original.

Foi também definida uma nomenclatura para as *labels* extraídas a partir do LabelMe, de forma a garantir uma fácil identificação da origem da estrutura rotulada. Assim, todas as *labels* indicam o tipo de estrutura (cromossoma, nucléolo ou objeto ruidoso) e o nome da imagem do *dataset* do LCG-FMUC a que pertencem. Em relação aos cromossomas, as suas *labels* apresentam também a classe cromossómica em que se inserem (1 a 24, onde 23 e 24 correspondem aos cromossomas sexuais X e Y, respetivamente) e se são o cromossoma da direita ou esquerda no respetivo cariograma. Quanto aos núcleos interfásicos, as suas *labels* contêm a posição em que se encontram na imagem celular (por exemplo, nas margens delimitadoras da imagem, nos cantos da imagem ou no meio da imagem).

Todas estas informações são de elevada importância para a geração automática de imagens sintéticas fotorrealistas a partir de estruturas celulares reais.

Visto que os cariogramas do *dataset* do LCG-FMUC resultam de casos clínicos reais, existem cromossomas que tiveram de ser individualizados a partir de *clusters*. Este tipo de cromossomas, assim como os seus cromossomas homólogos, não foram rotulados. Dado que o objetivo do *labelling* e recorte destas estruturas celulares é a posterior geração de imagens sintéticas, é necessário garantir que os cromossomas estão bem individualizados. Portanto, o *labelling* das imagens do *dataset* do LCG-FMUC foi restringido por três tipos de situações que se encontram na Figura 3.9: impossibilidade de individualização da estrutura celular; existência de um cromossoma individualizado a partir de *cluster*; e o cromossoma homólogo ao cromossoma a ser rotulado resultar de um recorte de *cluster*.

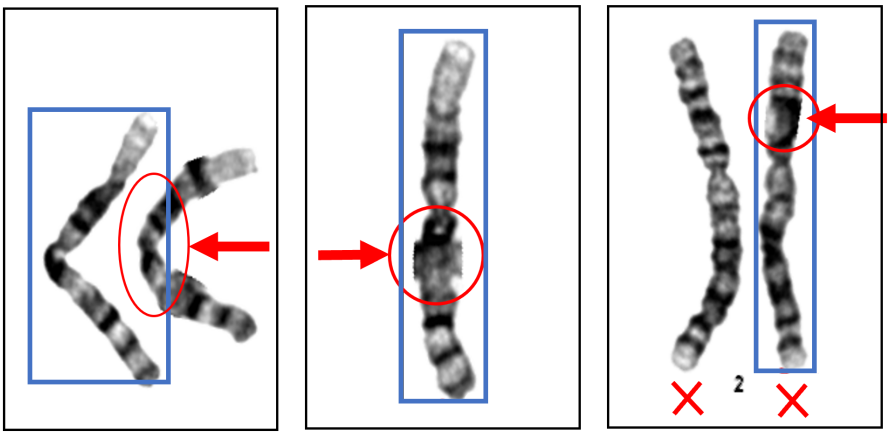


Figura 3.9: Restrições no *labelling* de estruturas celulares no LabelMe. A azul estão marcadas as *labels* restrita por uma das seguintes condições: Impossibilidade de individualização da estrutura celular (à esquerda); existência de um cromossoma individualizado a partir de *cluster* (no meio); e o cromossoma homólogo ao cromossoma a ser rotulado resultar de um recorte de *cluster* (à direita). Imagem original.

Após serem gerados os ficheiros JSON para todas as imagens do *dataset* do LCG-FMUC, cada estrutura celular rotulada foi individualizada e guardada como ficheiro TIF. Para isso, os ficheiros JSON e TIF foram lidos em simultâneo, sendo que para cada *label* foram usadas as coordenadas da sua bbox para o recorte da estrutura

celular dessa *label*. Os pixéis correspondentes aos valores da matriz da imagem TIF entre as coordenadas dos vértices da bbox foram guardados como uma nova imagem, individualizando-se, desta forma, as estruturas celulares rotuladas a partir da imagem original.

A partir do momento em que a estrutura se tornou individualizada, recorreu-se a uma técnica de DA para gerar novos dados. A técnica usada foi a rotação a duas dimensões, que faz uso da transformação linear dos pixéis a partir de uma matriz rotação segundo a Equação 3.2 [69]. Considerando um ponto *(x,y)* de uma imagem como um vetor 1x2, é possível transformar esse vetor num outro conjunto de coordenadas através da multiplicação por uma matriz de transformação 2x2. Essa matriz rotação é definida pela Equação 3.3, onde θ corresponde ao ângulo de rotação relativamente ao eixo XX.

$$v = A_{rot} v_0 \tag{3.2}$$

$$A_{rot} = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix} \tag{3.3}$$

De acordo com Huang *et al.* [12], o *trade-off* encontrado entre o volume de imagens gerado e o aumento do desempenho do modelo treino de DL para segmentação de cromossomas é de 15°. Assim, as estruturas individualizadas (cromossomas e objetos ruidosos) sofreram transformações sucessivas de 15°, isto é, dos 0°aos 345°. A nível de computação deste método, recorreu-se às funções *getRotationMatrix2D* e *warpAffine* para se obter a matriz rotação e fazer a rotação das imagens, respetivamente. Foi possível obter 24 imagens por cada cromossoma e objeto ruidoso rotulado.

No caso dos núcleos interfásicos, estes aparecem a maioria das vezes nas extremidades das imagens celulares, isto é, não estão totalmente expostos na imagem. Assim, devido à fase "*Paste*" usada para gerar imagens sintéticas (fase exposta na secção seguinte), estas estruturas só podem ser colocadas em lugares específicos de uma imagem sintética para se ir de encontro ao fotorrealismo pretendido. Por exemplo, na Figura 3.10 observa-se parte de um nucléolo no canto

superior direito. Este nucléolo só poderá ser colocado num do quatro cantos da imagem, após sofrer rotação, sem que seja perdida coerência a nível do que é uma imagem celular em metafase. Portanto, os núcleos interfásicos sofreram rotação apenas quatro vezes, com um ângulo de 90°, dado que só podem ser colocados em uma das quatro margens da imagem.

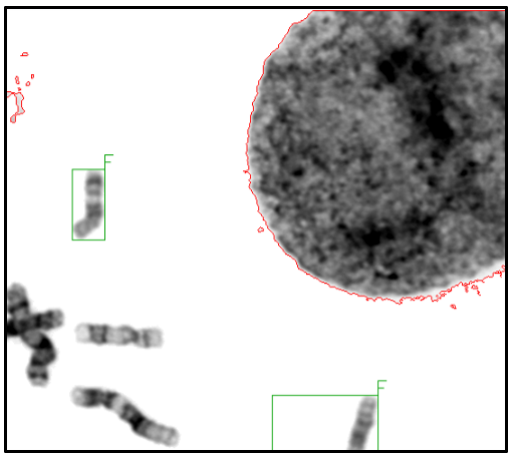


Figura 3.10: Imagem representativa das limitações na rotação dos núcleos interfásicos. Neste exemplo, o nucléolo está situado no canto superior direito da imagem. Para garantir fotorrealismo, este nucléolo só poderá ser colado noutras imagens caso sofra rotações de 90°e seja colado num dos quatro cantos da nova imagem. Imagem escolhida aleatoriamente do *dataset* do LCG-FMUC.

A metodologia detalhada passo a passo, para a extração, individualização e rotação das estruturas celulares mencionadas encontra-se apresentada no Anexo E.

3.2.3 Obtenção de imagens de células em metafase Sintéticas - "*Paste*"

Posteriormente à obtenção de imagens de estruturas celulares reais, automatizou-se a geração de imagens sintéticas (fase "*Paste*"). Tendo em conta que o objetivo da criação destas imagens é sistematizar a obtenção de dados para a segmentação de cromossomas, é importante que as imagens sintetizadas apresentem uma grande variabilidade morfológica e que sejam fotorrealistas. Para isso, procedeu-se à colagem automática das várias estruturas celulares numa nova imagem, recorrendo-se ao *script* "synthetic_methapase_generator.py", disponível no GitHub.

Para cada imagem sintetizada, os seguintes parâmetros são gerados de forma aleatoriamente uniforme dentro de um intervalo de valores definido:

- Número de cromossomas: 4 - 46 cromossomas;
- Número de núcleos interfásicos: 0 - 1 núcleos interfásicos;
- Número de objetos ruidosos: 0 - 50 objetos ruidosos;
- Altura da imagem: 500 - 1024 pixéis;
- Largura da imagem: 700 - 1376 pixéis;

Além destes parâmetros, todas as estruturas são aleatoriamente escolhidas a partir dos recortes da fase "*Cut*", o que confere uma variedade imensa no processo de colagem. Assim sendo, tornou-se possível a sintetização de milhares de imagens diferentes que podem contribuir para melhorar o treino de modelos de DL para a segmentação de cromossomas.

Tendo em mente que este processo de colagem é iniciado com uma tela em branco, foi possível gerar máscaras das imagens sintetizadas, píxel a píxel, anotando-se o tipo de estrutura em causa. Com este processo foi possível gerar não só dados para segmentação por deteção de objetos, mas como também criar máscaras, isto é, a *ground truth* das imagens para fins de segmentação semântica.

O processo de colagem em si mesmo é composto por três fases: colagem de núcleos interfásicos, colagem de cromossomas e colagem de objetos ruidosos. A base comum às três fases do processo de colagem corresponde à substituição dos pixéis da nova imagem por pixéis das estruturas que tenham intensidade inferior a 255, isto é, todos os pixéis que não sejam brancos são copiados e sobrepostos na nova imagem. Além disso, a posição escolhida para a colagem das estruturas foi também definida de forma aleatória, seguindo uma distribuição uniforme. As coordenadas do primeiro e último píxel a serem colados são guardados como a bbox da estrutura celular, servindo como a sua *label* na nova imagem sintética.

Na primeira fase da colagem, caso o número de núcleos interfásicos definido aleatoriamente seja igual a 1, é escolhido um nucléolo entre as imagens de todos os núcleos interfásicos individualizados para ser inserido na imagem. Posteriormente, o algoritmo analisa o nome do ficheiro escolhido e cola o nucléolo de acordo com a posição definida no seu nome.

Na segunda fase da colagem, as estruturas a serem colocadas na imagem sintética são os cromossomas, cuja quantidade é definida aleatoriamente entre 4 e 46. Em primeiro lugar, as classes cromossómicas são definidas aleatoriamente. Exemplificando, se o número de cromossomas a colar for igual a 10, são escolhidos cromossomas de 10 classes cromossómicas diferentes de entre as 24 possíveis. Caso haja mais do que 23 cromossomas a serem colados na mesma imagem, começam a ser escolhidos os respetivos homólogos. Em segundo lugar, cada um desses cromossomas é escolhido, de forma aleatória, dos recortes realizados anteriormente. Em terceiro lugar, o cromossoma é colado na posição definida aleatoriamente. Por último, em quarto lugar, dada a possibilidade de sobreposição com outras estruturas, é aplicado um método de *blending* de imagens.

O grande problema das sobreposições é o fotorrealismo da colagem. Caso se colassem os cromossomas sem nenhuma alteração posterior, existiria sempre uma margem esbranquiçada à volta do cromossoma que foi colado em último lugar (Figura 3.11). O problema reside, então, nas bordas dos cromossomas aquando da sobreposição.

Figura 3.11: Exemplo de sobreposição sintética de cromossomas através de uma colagem simples e sem aplicação de nenhum método de *blending*. Imagem original.

Para a suavização das sobreposições foi proposto e aplicado um método de *blending* que consiste nas etapas descritas de seguida e representadas graficamente na Figura 3.12:

1. Localização dos pixéis referentes à sobreposição e binarização desta região para obtenção de uma máscara. Neste caso, o processo de binarização corresponde a considerar todos os pixéis desta região como 1 e todos os outros pixéis como 0. A binarização da sobreposição apresenta-se exemplificada pelos pixéis pretos na Figura 3.12.

2. Localização da fronteira da sobreposição, através da criação de um gradiente, o qual corresponde à diferença entre as operações morfológicas básicas de dilatação e de erosão. A erosão permite erodir as fronteiras de um objeto: considerando um *kernel* com uma determinada conetividade (isto é, o número de pixéis vizinhos), o píxel a ser erodido da imagem binária original é considerado como 1 caso todos os pixéis da sua conetividade também sejam 1. A dilatação faz o inverso, aumentando as fronteiras de um objeto: um píxel é considerado 1, caso pelo menos um dos pixéis na sua conetividade sejam também 1. Para aplicação deste método, recorreu se à função *morphologyEx* com um *kernel* de 9x9. A fronteira obtida está representada graficamente a verde na Figura 3.12.

3. Localização dos pixéis do cromossoma a ser colado, cuja intensidade seja superior a 200, ou seja, a margem esbranquiçada. Com este passo, pretendeu guardar-se a posição dos pixéis da estrutura já existente que está a ser sobreposta. Recorreu-se a um método simples de binarização com um *threshold* igual a 200 para se obter a máscara pretendida. Esta máscara está representada pelos pixéis vermelhos na Figura 3.12.

4. Localização do esqueleto do cromossoma a ser colado. Este esqueleto permitiu obter apenas os pixéis das margens longitudinais da sobreposição, excluindo os pixéis das bordas transversais ao cromossoma a ser colado. Para isso recorreu-se à erosão com um *kernel* de 5x5 e 3 iterações, através da função *erode*. O esqueleto pode ser observado na Figura 3.12.

5. Aplicação de 10 iterações do filtro de caixa definido pela Equação 3.1 aos pixéis localizados nos passos anteriores: primeiro, aplicação desse filtro aos pixéis das margens da sobreposição correspondentes ao cromossoma a ser colado (representados a verde na Figura 3.12); e segundo, aplicação desse filtro aos pixéis da sobreposição pertences à estrutura celular a ser sobreposta (representados a vermelho na Figura 3.12).

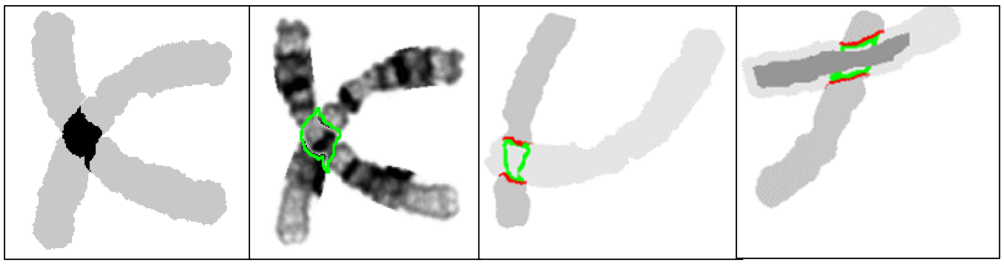


Figura 3.12: Esquematização do método de *blending* proposto. Da esquerda para a direita: localização dos pixéis referentes à sobreposição; localização da fronteira da sobreposição; localização dos pixéis do cromossoma a ser colado, cuja intensidade seja superior a 200; localização do esqueleto do cromossoma a ser colado. Imagem original.

Na terceira e última fase deste processo de colagem, são colados os objetos ruidosos. Apesar da posição da colagem destas estruturas ser aleatória, o algoritmo não permite que se sobreponham a outras estruturas. Esta escolha tem dois motivos: primeiramente, nas imagens observadas da citogenética clínica, por norma, os objetos ruidosos não se sobrepõem aos cromossomas nem aos núcleos interfásicos, espalhando-se nas proximidades destas estruturas celulares; segundamente, o gasto computacional do método de *blending* proposto é proporcional ao número de estruturas sobrepostas, salvaguardando-se, desta forma, a sintetização demorada de imagens celulares. Caso não seja encontrada uma posição aleatória cujo objeto ruidoso não se sobreponha a outra estrutura, o algoritmo gera outra posição aleatória – este processo é repetido para 50 iterações, no final das quais o objeto não é utilizado na colagem.

À medida que o processo de colagem decorre, as *labels* das estruturas são atualizadas e guardadas em formato JSON. No final da colagem, é anotado o número de cromossomas, o número de núcleos interfásicos, o número de objetos ruidosos, o número de *clusters* e as dimensões da imagem. Estas informações permitem caraterizar, de forma pormenorizada, o *dataset* sintético obtido. Tendo em conta o processamento em torno da metodologia proposta para a fase "*Paste*", as imagens levam algum tempo a ser sintetizadas - por exemplo, um *dataset* de

1000 imagens demora cerca de 10 dias a ser gerado.

3.2.4 YOLOv5 - "*Learn*"

Após a obtenção de um *dataset* com 10795 imagens de células em metafase sintéticas, seguiu-se a fase "*Learn*" que consistiu no uso de modelos de DL para a segmentação de cromossomas em imagens de células em metafase. Nesta secção é apresentada, em primeiro lugar, a rede neuronal YOLOv5 que foi utilizada para treinar um modelo capaz de detetar cromossomas. De seguida, descrevem-se os passos necessários para implementar e treinar este modelo. Por fim, são referidas as métricas usadas para avaliar o modelo obtido.

3.2.4.1 YOLOv5

A família de redes neuronais YOLO consiste em detetores de objetos a uma fase, baseados numa rede de deteção de objetos por região. As redes YOLO abordam a deteção de objetos como se de um problema de regressão se tratasse, resultando numa alta velocidade de processamento. Diferentes de outros algoritmos de deteção de objetos, como a R-CNN ou a Faster R-CNN, as redes YOLO analisam a imagem uma única vez - daí o nome "*You Only Look Once*". Devido a esta característica, as redes YOLO são capazes de conseguir uma velocidade de deteção muito maior do que outras técnicas de deteção de objetos, sem perderem alta exatidão. O YOLOv5 é a versão mais recente e documentada da família YOLO, sendo considerado o pináculo do desenvolvimento da deteção de objetos [70].

O YOLOv5 é um projeto de código aberto que usa PyTorch e inclui uma variedade de modelos e algoritmos de identificação de objetos. Estes modelos utilizam a arquitetura do YOLOv4 e foram pré-treinados no *dataset* COCO (*Common Objects in Context*) [54]. O YOLOv5 é então uma coleção de modelos de deteção de objetos com recursos fáceis para treinar modelos, aplicar técnicas - como *Test Time Augmentation* (TTA) - e estudar os hiperparâmetros associados aos vários modelos. Como o YOLOv5 não implementa ou desenvolve nenhuma abordagem exclusiva, o documento formal não pôde ser divulgado pela Ultralytics [7].

Relativamente à estrutura deste detetor, o YOLOv5 é composto por três

[7] https://ultralytics.com/

componentes principais: o *backbone*, o *neck* e a *head* (ou deteção) [71]. O *backbone* corresponde à CNN CSPDarknet53 que coleta e processa as *features* presentes nas imagens, e é composto por dois blocos: uma camada base de convolução e um bloco de Cross Stage Partial. O primeiro bloco é responsável pela geração do mapa de *features*, enquanto o segundo bloco divide este mapa em duas partes, utilizando uma hierarquia de Cross-Stage. O *neck* é a parte da arquitetura onde acontece a agregação de features. Ele coleta mapas de *features* das diferentes fases do *backbone*, depois mistura-os e combina-os para prepará-los para a próxima etapa de deteção. A rede de agregação de *features* usada é a PANet. Entre o *backbone* e o *neck* está ainda um bloco, designado de Spatial Pyramid Poolin, que separa as *features* mais importantes. A *head* corresponde ao processo de deteção e formulação das bboxes e as respetivas classes. O diagrama da arquitetura do YOLOv5 está esquematizado na figura 3.13.

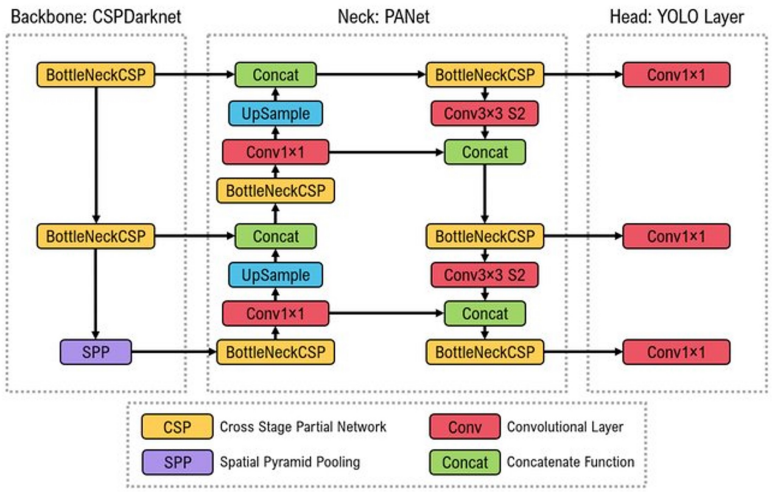


Figura 3.13: Arquitetura do YOLOv5. Imagem adaptada de Katsamenis*et al.* (2022). TraCon: A novel dataset for real-time traffic cones detection using deep learning. 10.48550/arXiv.2205.11830..

Em termos de metodologia, o primeiro passo que as redes neuronais YOLO executam é a divisão da imagem a analisar numa grade de S por S células (nas versões mais recentes esse tamanho é de 19×19). Cada uma destas células é responsável por realizar a deteção de cinco bboxes, pois podem existir mais do que um objeto nesta célula. Por sua vez, cada bbox é responsável por analisar um pedaço da imagem e extrair informações da região, contendo três atributos:

1. Probabilidade de confiança: atributo com a probabilidade de existir um objeto na determinada bbox;

2. Coordenadas: contém a localização da bbox - a localização no YOLO é representada com posição central do objeto, a sua altura e a sua largura;

3. Probabilidade da classe: atributo com a probabilidade de ser determinado um objeto.

Finalizada a previsão das probabilidades de cada bbox, é depois necessário decidir quais são as bboxes que realmente possuem um objeto. Para tomar esta decisão, na sua última etapa de deteção, o YOLO realiza o processo de supressão não máxima. O YOLO utiliza ainda âncoras, que são retângulos de tamanhos pré-definidos. Estes retângulos são utilizados para que as bboxes previstas possuam uma maior relação com as *labels* dos objetos. Estas âncoras possuem dimensões próximas aos tamanhos dos objetos a serem identificados, sendo criadas durante o processo de treino da rede neuronal. A partir das coordenadas das bboxes dos objetos detetados, são selecionadas as âncoras de cada objeto, que serão posteriormente redimensionadas para a proporção dos objetos identificados e utilizadas como saída da rede neural YOLO. Assim, a rede neuronal YOLO não prevê o tamanho final do objeto, apenas ajusta o tamanho da âncora mais próxima ao tamanho do objeto.

Como resultado final, o YOLOv5 apresenta a imagem que foi dada como *input* e as *labels* das previsões, tal como representado na Figura 3.14. Cada *label* está associada a uma probabilidade de confiança do algoritmo - quanto mais alta for esta probabilidade, maior será a probabilidade da *label* corresponder à realidade.

3.2.4.2 Implementação de Modelos YOLOv5

Relativamente à implementação do YOLOv5, esta foi feita em duas fases. Primeiro, foi necessário converter as *labels* para um formato específico requerido pelo modelo. Seguidamente, foram treinados modelos pré-treinados no *dataset* COCO. O código base para a implementação do YOLOv5 encontra-se disponível no GitHub da Ultralytics e está escrito maioritariamente em Python.

Para treinar o YOLOv5, é necessário ter ficheiros de imagem e os respetivos ficheiros

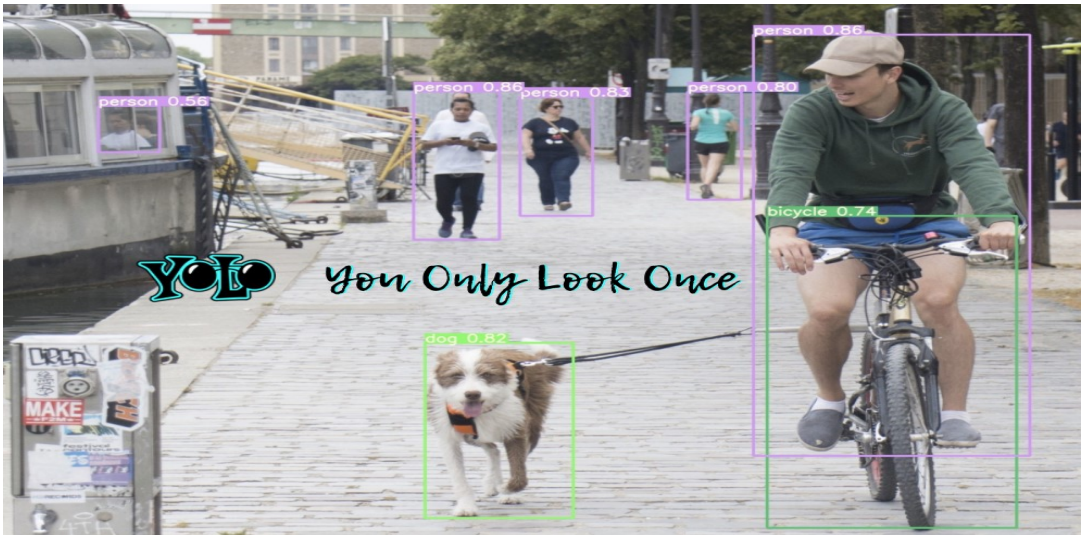


Figura 3.14: Exemplo do resultado final da deteção pelas redes YOLO. O algoritmo devolve a imagem inicial com as *labels* previstas, a classe atribuída e a probabilidade de confiança. Imagem retirada a 8 de agosto de 2022, de `https://docs.ovh.com/sg/en/publiccloud/ai/notebooks/yolov5-example/`.

das suas *labels* com a mesma nomenclatura. Por sua vez, as *labels* têm de estar no formato YOLO. Isto significa que a cada imagem tem de estar associado um Ficheiro de Texto Simples (TXT), que obedece aos seguintes requisitos:

1. Cada linha contém um objeto. Neste caso, o ficheiro TXT irá conter uma estrutura celular por linha;

2. Cada objeto é definido por cinco campos, representados na Figura 3.15, obrigatoriamente pela seguinte ordem: classe, abcissa do centro da bbox, ordenada do centro da bbox, largura da bbox e altura da bbox;

3. As coordenadas da bbox têm ainda de estar normalizadas em relação à imagem onde está inserida, tal como apresentado na Figura 3.16);

4. As classes são indexadas a partir do zero.

Figura 3.15: Estrutura da *label* de um objeto no formato YOLO. ID_{grupo} é a classe, x_{norm} é a abcissa do centro da bbox, y_{norm} é a ordenada do centro da bbox, $width_{norm}$ é a largura da bbox e $height_{norm}$ é altura da bbox. Imagem Original.

Figura 3.16: Esquematização das coordenadas de uma *label* no formato YOLO. Imagem retirada a 10 de agosto de 2022, de `https://blog.paperspace.com/train-yolov5-custom-data/`.

Posto isto, foi necessário converter os ficheiros JSON para o formato YOLO. Para esta finalidade, foi utilizada a função *get_YOLO_dataset*, que fez esta conversão de forma automática para as milhares de imagens sintetizadas pelo método *"Paste"*. Relativamente à classe dos objetos, os cromossomas passaram a ser representados pelo grupo de identificação 0, os núcleos interfásicos por 1 e os objetos ruidosos por 2. Em relação às coordenadas da bbox, procedeu-se ao cálculo do centro da *label* (equações 3.4 e 3.5), à normalização das coordenadas desse centro (equações 3.6 e 3.7), e também à normalização das dimensões da *label* (equações 3.8 e 3.9), relativamente às dimensões da imagem onde a *label* está inserida.

$$x = x_i + \frac{x_f - x_i}{2} \tag{3.4}$$

$$y = y_i + \frac{y_f - y_i}{2} \tag{3.5}$$

$$x_{norm} = \frac{x}{width} \tag{3.6}$$

$$y_{norm} = \frac{y}{height} \tag{3.7}$$

$$width_{norm} = \frac{width}{image_{width}} \tag{3.8}$$

$$height_{norm} = \frac{height}{image_{height}} \tag{3.9}$$

Depois da conversão das *labels* para o formato YOLO, foram obtidos os grupos de treino e validação, a partir do *script* *execute_yolo*, que faz uso da função *train_test_split* da biblioteca *scikit-learn*. Os grupos treino e validação foram divididos numa proporção 90/10. Não se utilizou validação cruzada, uma vez que os modelos demoram bastante tempo a serem treinados e o *dataset* utilizado continha um grande volume de dados.

Seguidamente, foi necessário organizar as diretorias das duas pastas que continham as imagens e os ficheiros TXT. Para isso foi escrito um ficheiro YAML que continha o caminho das diretorias para os ficheiros de treino e de validação. Apesar de parecer trivial, este passo é estritamente necessário para treinar modelos YOLO com *datasets* customizados.

Por fim, fez-se uso do Google Colab e do acelerador de GPU disponibilizado para treinar o modelo YOLO. Uma vez que já existem modelos pré-treinados (detalhados na Figura 3.17) num dos *datasets* mais reconhecidos a nível de deteção de objetos - o *dataset* COCO [54] - foram escolhidos os modelos YOLOv5s, YOLOv5m e YOLOv5l para fazer *fine-tuning* no *dataset* sintético criado anteriormente. Por sua vez, estes modelos foram treinados com *datasets* de 200, 1000 e 10000 imagens, respetivamente. Assim, foi possível comparar o desempenho destes modelos, assim como o resultado de se usarem diferentes volumes de dados para treinar modelos de aprendizagem de deteção de objetos. Um exemplo de *script* usado para treinar um destes modelos está em *YOLOv5_example.py*.

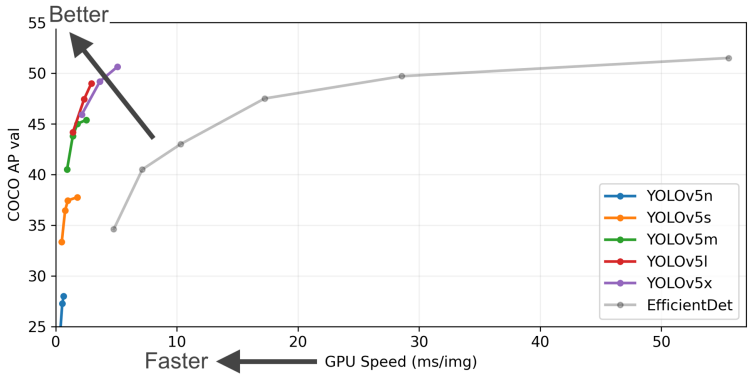


Figura 3.17: Comparação dos diferentes modelos pré-treinados do YOLOv5. Gráfico da *Average Precision* em função do tempo de computação pela GPU, para cada modelo. Imagem retirada a 10 de agosto de 2022, de `https://github.com/ultralytics/yolov5`.

3.2.4.3 Métricas de Avaliação

Para avaliar os modelos obtidos pelo YOLOv5, utilizaram-se as seguintes métricas, usuais no campo da deteção de objetos: *Intersection Over Union* (IOU), *Precision* (P), *Recall* (R), *F1-score* (F1), curva P-R, *Average Precision* (AP) e *mean Average Precision* (mAP) [72].

Na deteção de objetos, um modelo pode prever a existência de uma determinada classe de objeto, sendo as previsões positivas ou negativas. Uma previsão correta acontece quando a classe detetada pelo modelo corresponde à realidade e uma previsão incorreta acontece quando a classe detetada pelo modelo não corresponde à realidade. Desta forma podemos definir Verdadeiros Positivos (TP) e Falsos Positivos (FP), respetivamente. Quanto à inexistência de *labels* em certos locais onde há objetos, podem ser consideradas de Falsos Negativos (FN), no caso de existir um objeto e o modelo não o considerar, ou então podem existir Verdadeiros Negativos (TN), isto é, realmente não existirem objetos em determinados locais da imagem e o modelo não os identificar. Quanto à exatidão de um modelo de deteção de objetos, esta vai depender de vários parâmetros, nomeadamente de um *threshold* para a IOU e de um valor de confiança dado pelo modelo.

O rácio *Intersection Over Union* (IOU) é usado como um *threshold* para a determinação de Verdadeiros Positivos (TP) e Falsos Positivos (FP). Este rácio IOU corresponde ao valor da sobreposição das bboxes da previsão do modelo e da *label* verdadeira do objeto. Graficamente, pode ser representada pela Figura 3.18, onde o numerador é a área de interseção e o denominador é a área de união da previsão e da *label*.

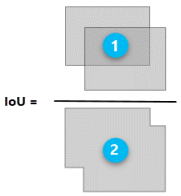


Figura 3.18: Esquematização da métrica *Intersection Over Union* (IOU). 1 corresponde à área de interseção. 2 corresponde à área de união. Imagem retirada a 12 de agosto de 2022, de `https://pro.arcgis.com/en/pro-app/latest/tool-reference/image-analyst/how-compute-accuracy-for-object-detection-works.htm`.

Depois de definido o *dataset* para a IOU, é possível obter a matriz confusão, que agrega os valores de TP, FP, TN e FN. Posteriormente, definem-se as métricas de exatidão do modelo:

1. A *Precision* (P) é a razão entre verdadeiros positivos e o número total de previsões feitas (Equação 3.10).

$$Precision = \frac{TP}{TP + FP} \tag{3.10}$$

2. A *Recall* (R) é a razão entre o número de verdadeiros positivos e o número total de objetos (Equação 3.11).

$$Recall = \frac{TP}{TP + FN} \tag{3.11}$$

3. O *F1-score* (F1) é uma média ponderada da precisão e da recall, cujo valor varia entre 0 e 1, correspondendo 1 ao valor máximo de exatidão (Equação 3.12).

$$F1_{score} = \frac{Precision \times Recall}{\frac{Precision+Recall}{2}} \tag{3.12}$$

4. A curva P-R é um gráfico da *Precision* em função da *Recall*. Um modelo com maior exatidão é aquele que consegue manter valores elevados de *Precision* quando a *Recall* aumenta (Figura 3.19).

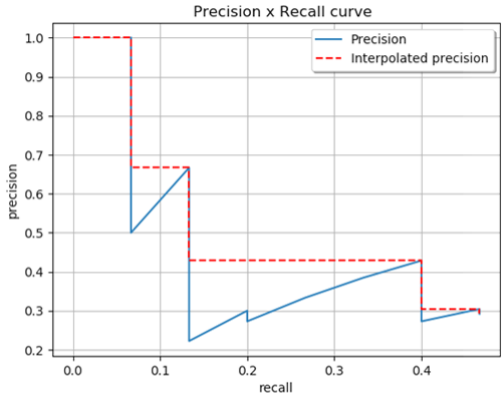


Figura 3.19: Gráfico da curva P-R. Imagem retirada a 12 de agosto de 2022, de `https://pro.arcgis.com/en/pro-app/latest/tool-reference/image-analyst/how-compute-accuracy-for-object-detection-works.htm`.

5. A *Average Precision* (AP) é a precisão ao longo de todos os valores de *Recall*, entre 0 e 1. Esta métrica pode ser interpretada como a área debaixo - *Area Under the Curve* (AUC) - da curva P-R, resultante de uma interpolação dos pontos desta curva.

6. A *mean Average Precision* (mAP) corresponde à média da AP para múltiplos intervalos de IOU. Normalmente é utilizado a métrica $mAP@[0.5 : 0.05 : 0.95]$ que significa a média dos valores de AP para valores de IOU compreendidos entre 0.5 e 0.95, e intervalados de 0.05. Caso a deteção de objetos seja multi-classe, a métrica mAP é ainda ponderada entre as várias classes do modelo.

Depois da avaliação dos modelos de treino e validação obtidos através das métricas mencionadas acima, o melhor modelo foi testado no *dataset* pré-processado do LCG-FMUC. Neste caso, além da avaliação quantitativa, os resultados da deteção de cromossomas foram analisados graficamente e qualitativamente, a fim de se avaliar a deteção de cromossomas em *clusters* com vários níveis de complexidade.

4

Resultados e Discussão

Este capítulo expõe os resultados obtidos ao longo da aplicação da metodologia apresentada previamente, fazendo uso do *dataset* do LCG-FMUC. Estes resultados são oportunamente acompanhados pela sua discussão e análise.

4.1 Pré-processamento

4.1.1 Resultados

O início da metodologia desta dissertação passou pela análise das imagens fornecidas pelo LCG-FMUC. Estas imagens estavam agregadas aos pares, isto é, a cada imagem celular com cromossomas em metafase estava associada uma imagem do respetivo cariograma. Contudo, como explicado anteriormente, as imagens já tinham sofrido algum processamento pelo *software* utilizado pelo laboratório.

Como pode ser visto na Figura 4.1, existem anotações a vermelho e verde que correspondem a contornos de ruído e a cromossomas, respetivamente, e que não estavam no campo de visão da câmara que tirou a microfotografia. A figura 4.1 será utilizada como exemplo, escolhido aleatoriamente, para demonstrar os efeitos do pré-processamento aplicado a todas as imagens do *dataset* do LCG-FMUC.

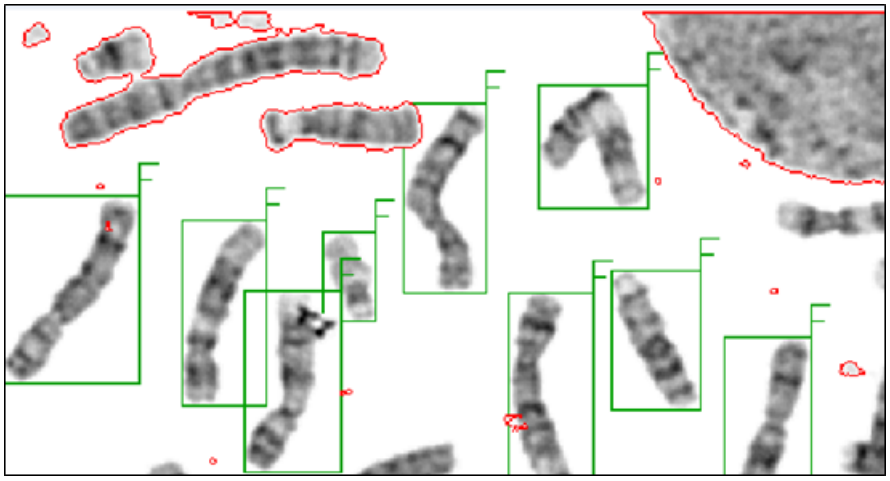


Figura 4.1: Recorte de uma imagem celular em metafase aleatoriamente escolhida do *dataset* fornecido pelo LCG-FMUC.

Através da substituição dos pixéis a cores por pixéis brancos, foi possível remover estas anotações. Na Figura 4.2 apresenta-se o resultado desta substituição, que apesar de remover as anotações, acabou por criar artefactos devidos à sobreposição destes pixéis com estruturas celulares.

Figura 4.2: Resultado da substituição das anotações do *software* Cytovision por pixéis brancos e respetiva conversão da imagem para a escala de cinzentos. Imagem original.

De seguida, uma vez que a imagem só possui pixéis na escala dos cinzentos, mas ainda está representada como uma imagem RGB - isto é, por uma matriz com os três canais de cor - procedeu-se à sua conversão para uma matriz que apenas contenha um canal de cor. No caso da Figura 4.2, houve uma compressão da imagem de

189 kB para 130 kB que não afetou a qualidade da imagem. Esta conversão foi importante para reduzir o gasto computacional e o tempo de processamento dos algoritmos que iriam depois analisar estas mesmas imagens.

Por fim, para colmatar os artefactos mencionados acima, que se traduzem em buracos nas estruturas celulares, foi aplicado um filtro de caixa. Este filtro permitiu suavizar as imagens no caso das estruturas celulares que continham pixéis brancos, tornando-se totalmente indistinguíveis (Figura 4.3). O resultado da aplicação de três iterações deste filtro pode ser observado na Figura 4.4.

Figura 4.3: Resultado final do pré-processamento. Imagem original.

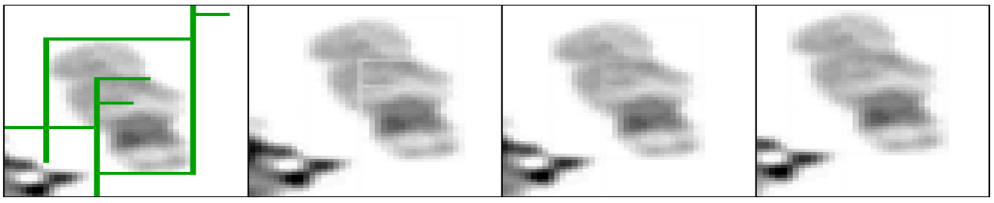


Figura 4.4: Resultado em detalhe da aplicação de três iterações do filtro de caixa com um *kernel* 3x3. À esquerda observa-se o recorte de uma imagem do *dataset* do LCG-FMUC. De seguida, apresentam-se os resultados de três iterações sucessivas, respetivamente. Imagem original.

4.1.2 Discussão

Uma vez que não foi possível obter imagens diretamente do *software* do Cytovision, o especialista da citogenética teve de retirar manualmente as imagens,

uma por uma, do computador do LCG-FMUC e guardá-las em formato TIF. As imagens do *software* estavam encriptadas, o que significa que não era possível serem abertas fora do Cytovision. O processo moroso de obtenção destas imagens pelo citogeneticista levou à obtenção de um *dataset* relativamente pequeno quando comparado ao tamanho típico de *datasets* aplicados a técnicas de DL. Apesar do número reduzido de dados, estes foram essenciais para todo o procedimento que se seguiu.

O *dataset* do LCG-FMUC é composto por 342 imagens, das quais existem imagens celulares e imagens dos respetivos cariogramas. Apesar de se ter realizado pré-processamento aos dois tipos de imagem, a remoção das anotações e a aplicação do filtro de caixa foram aplicados apenas às imagens celulares. Quanto à conversão de escalas, as imagens de cariogramas também sofreram a conversão para a escala dos cinzentos, uma vez que estas imagens também estavam representadas por uma matriz com os três canais de cor. Em geral, o tamanho dos ficheiros TIF do *dataset* passou de estar compreendido no intervalo [84-394] kB para o intervalo [64-283] kB, havendo uma média de compressão de 25.7% por ficheiro TIF, sem perda de qualidade da imagem.

Relativamente ao filtro de caixa utilizado para retificar os artefactos gerados pela substituição dos pixéis das anotações, foram estudadas outras opções, tais como o filtro gaussiano. Contudo, pela sua simplicidade em termos de definição de parâmetros e pela rapidez computacional, o filtro de caixa foi o método escolhido. Em termos do tamanho do *kernel* e do número de iterações, foram testados diferentes tamanhos - como por exemplo um *kernel* de conetividade 4 - e diferentes números de iterações. Analisando-se os resultados, concluiu-se que o *kernel* 3x3 aplicado ao longo de três iterações correspondia ao *trade-off* entre fotorrealismo e a eficácia para suavizar os buracos criados nas estruturas celulares.

4.2 *Dataset* de Estruturas Celulares Individualizadas

4.2.1 Resultados

A primeira fase da metodologia que levou à criação de um modelo de segmentação de cromossomas foi a fase *"Cut"*. Esta fase consistiu na criação de *labels* no *software* LabelMe, a partir das imagens pré-processadas do *dataset* do LCG-FMUC. Para cada imagem, seja ela uma imagem celular ou uma imagem do cariograma, foi obtido um ficheiro JSON. Este ficheiro JSON continha as várias *labels* criadas na respetiva imagem TIF, onde cada *label* é caraterizada pelo nome, pelo grupo de identificação da estrutura em causa e pelas coordenadas da bbox que delimita a estrutura celular.

Os tipos de estruturas celulares que foram rotuladas a partir de imagens reais foram cromossomas, núcleos interfásicos e objetos ruidosos, e que de seguida serão explorados individualmente.

1. Cromossomas

   No total foram rotulados 4829 cromossomas bem individualizados a partir das imagens dos cariogramas. Tendo em conta que não foram rotulados cromossomas cujo cromossoma homólogo não se apresentava bem individualizado, o resultado corresponde a 2302 pares de cromossomas autossómicos, 81 pares de cromossomas homólogos X e 63 cromossomas Y. Para cada imagem de cariograma foram obtidos resultados semelhantes à Figura 4.5. O número de cromossomas obtidos no LabelMe encontra-se caraterizado por classe cromossómica no gráfico da Figura 4.6.

   Posteriormente, cada cromossoma foi recortado com base no ficheiro JSON relativo à imagem do cariograma. Além disso, foi aplicada uma rotação intervalada de 15°, entre os 0°e os 345°, obtendo-se, desta forma, 24 imagens diferentes para cada cromossoma rotulado. A Figura 4.7 exemplifica este processo de individualização e rotação de cromossomas.

   Por fim, os cromossomas foram guardados como imagem TIF e organizados

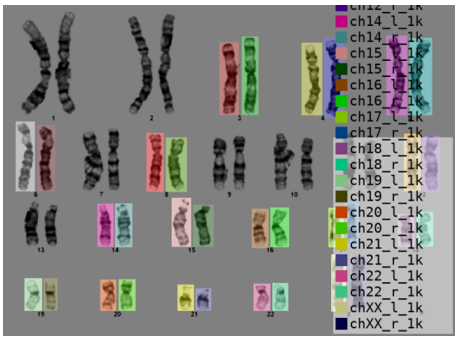


Figura 4.5: Exemplo de um cariograma anotado no LabelMe. Apresentam-se as *labels* obtidas depois das restrições aplicadas, assim como a nomenclatura usada. Imagem original.

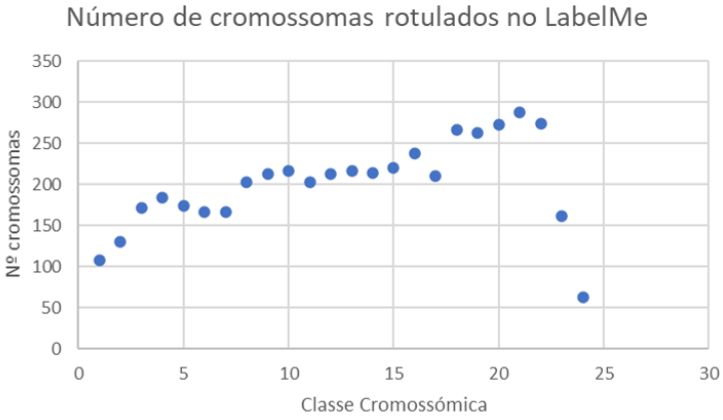


Figura 4.6: Gráfico do número de cromossomas rotulados no LabelMe em função da classe cromossómica. Imagem original.

Figura 4.7: Exemplo da individualização e rotação de um cromossoma a partir do respetivo cariograma. Apresentam-se cinco rotações de 15°. Imagem original.

Classe Cromossómica	Número de Cromossomas LabelMe	Número de Cromossomas DA
1	108	2592
2	130	3120
3	172	4128
4	184	4416
5	174	4176
6	166	3984
7	166	3984
8	202	4848
9	212	5088
10	216	5184
11	202	4848
12	212	5088
13	216	5184
14	214	5136
15	220	5280
16	238	5712
17	210	5040
18	266	6384
19	262	6288
20	272	6528
21	288	6912
22	274	6576
X	162	3888
Y	63	1512
Total	4829	115896

Tabela 4.1: Número de cromossomas individualizados. Tabela original.

pela sua classe cromossómica, isto é, da pasta 1 à pasta 24, onde a pasta 23 e 24 correspondem aos cromossomas sexuais X e Y, respetivamente. No final da aplicação desta metodologia foram obtidos 115896 cromossomas, quantificados por classe cromossómica na Tabela 4.1. Além disso, foi aplicada uma função de validação quanto aos cromossomas homólogos, por forma a confirmar se tinha ocorrido algum erro durante o processo de individualização.

2. <u>núcleos interfásicos</u>

   Através do LabelMe foram rotulados 45 núcleos interfásicos a partir das imagens celulares de cromossomas em metafase. A maior preocupação no *labelling* destas estruturas foi a sua posição. Tal como foi explicado anteriormente, no *dataset* do LCG-FMUC, os núcleos interfásicos aparecem

maioritariamente nas margens da imagem, estando expostos apenas parcialmente. Assim, na fase seguinte de colagem destas estruturas, e de forma a se obterem imagens fotorrealistas, a colagem dos núcleos interfásicos vai ser determinada pela posição que foram extraídas e rotacionadas.

Por este motivo, os núcleos interfásicos apenas puderam sofrer rotações de 90º, correspondendo a um fator de aumentação de 4 vezes. A Figura 4.8 apresenta um exemplo de um nucléolo que está posicionado na base da imagem celular e, por isso, apenas poderá ser colada numa imagem sintética caso esteja posicionada à mesma na base da imagem, numa das margens laterais (esquerda ou direita) ou então no topo da imagem. Assim, depois da individualização de estruturas, os núcleos interfásicos estão individualizados em 180 imagens TIF diferentes entre si.

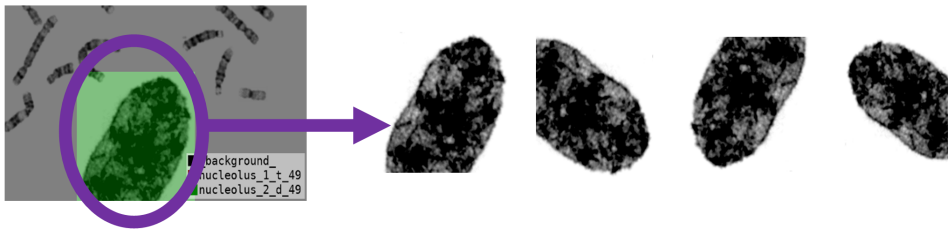


Figura 4.8: Exemplo da individualização e rotação de um nucléolo a partir da respetiva imagem celular. Apresentam-se quatro rotações de 90°. Imagem original.

3. Objetos Ruidosos

As últimas estruturas celulares a serem individualizadas foram objetos ruidosos, que correspondem a restos de pigmentação ou outros artefactos que não sejam nem núcleos interfásicos nem cromossomas. No total, foram rotulados 251 objetos ruidosos, com elevada variabilidade morfológica. Em termos de DA, estas estruturas também sofreram um aumento de 24 vezes o volume de dados inicial, devido à rotação de 15°entre 0°e 345°, resultando num total de 6024 objetos ruidosos. As estruturas foram rotuladas a partir de imagens celulares de cromossomas em metafase, como é apresentado na Figura 4.9.

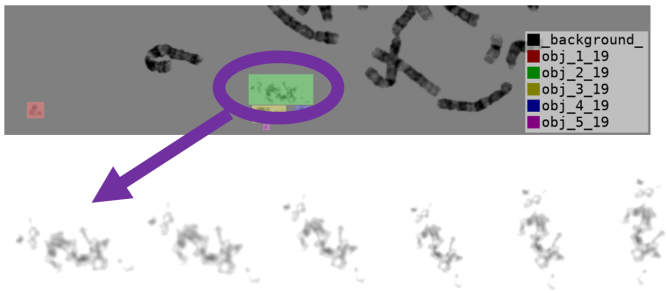


Figura 4.9: Exemplo da individualização e rotação de um objeto ruidoso a partir da respetiva imagem celular. Apresentam-se cinco rotações de 15°. Imagem original.

4.2.2 Discussão

A metodologia aplicada na fase "*Cut*" pode ser dividida em dois blocos. Primeiro, foi necessário recorrer-se ao LabelMe. Posteriormente, foi utilizado a linguagem Python. Em relação ao *labelling* das estruturas celulares no LabelMe, esta tarefa foi repetitiva e demorada, uma vez que foi fundamental a correta identificação de estruturas, bem como a sua nomenclatura e atribuição de um grupo de identificação. No total, foram realizadas 4585 *labels* (4289 cromossomas, 45 núcleos interfásicos e 251 objetos ruidosos). Quanto à fase de extração e individualização destas *labels*, automatizou-se este processo. Desta forma, basta seguir a metodologia descrita no Anexo E para se obter mais estruturas celulares respetivas à temática em questão.

Relativamente às restrições implementadas no *labelling* dos cromossomas, é percetível que estas restrições tenham tido um grande impacto nas classes cromossómicas que apresentam cromossomas mais longos (como por exemplo, o cromossoma 1 e 2). Quanto ao cromossoma sexual Y, era de esperar um menor número de *labels*, uma vez que no *dataset* do LCG-FMUC apenas existem 66 cariogramas associados ao género masculino. Tal confirmou-se.

No que diz respeito à nomenclatura atribuída a cada estrutura celular, esta foi exigente, mas confere a este *dataset* de estruturas celulares um grande rigor. O único *dataset* presente na literatura que contém estruturas celulares individualizadas é o *dataset* CRCN-NE [1] obtido por Andrade *et al.* [10]. Este *dataset* contém estruturas individualizadas de 74 imagens de células em metafase,

[1] `https://zenodo.org/record/3229434#.YyERQaTMJD-`

o que significa que o *dataset* proposto ultrapassa o CRCN-NE em termos de volume de dados. Além disso, a resolução das imagens é muito melhor, assim como a sua estrutura organizacional. No *dataset* CRCN-NE, apenas existe a divisão de estruturas celulares entre cromossomas e não-cromossomas, não havendo a identificação da classe cromossómica.

O *dataset* obtido pela metodologia apresentada pode ser usado por outros investigadores para testarem algoritmos de segmentação de estruturas celulares ou, inclusive, para algoritmos de classificação. Apesar desta dissertação se focar na fase de segmentação na geração automática do cariograma, os cromossomas individualizados nesta fase "*Cut*" podem ser usados para desenvolver algoritmos de extração de *features* e classificação de cromossomas.

4.3 *Dataset* Sintético de imagens de células em metafase

4.3.1 Resultados

Seguidamente à criação de um *dataset* com estruturas celulares reais, procedeu-se à fase *"Paste"*. Fez-se, então, a automatização da sintetização de imagens fotorrealistas através de imagens da citogenética convencional. Para isso, partindo de uma tela branca, foram colados núcleos interfásicos, cromossomas e objetos ruidosos, respeitando esta ordem de colagem.

Tal como mencionado na Secção 3.2.3, o principal obstáculo neste processo de colagem foi a sobreposição de estruturas celulares, nomeadamente cromossomas sobre cromossomas ou cromossomas sobre núcleos interfásicos. De forma a ultrapassar os artefactos gerados nas margens das sobreposições, foi proposto um método de *blending*. Este método fez uso de operações morfológicas básicas de erosão e dilatação, depois de obtidas máscaras dos pixéis relativos às sobreposições pelo método de binarização. Com a aplicação deste algoritmo de *blending* foram alcançados *clusters* fotorrealistas, mostrados na Figura 4.10.

O processo de colagem demorou entre 10 e 20 minutos por imagem, dependendo do

Figura 4.10: Exemplos de *clusters* sintéticos. Imagem original.

número de sobreposições existentes – o tempo de processamento e número de vezes que o método de *blending* é usado apresentam uma razão diretamente proporcional. No final da fase *"Paste"*, cada imagem sintética tem um ficheiro JSON associado que carateriza esta imagem através dos seguintes campos: o nome da imagem, o número de cromossomas, o número de *clusters*, o número de núcleos interfásicos, o número de objetos ruidosos e as bboxes que correspondem às *labels* das estruturas celulares presentes nessa imagem sintética. Por exemplo, para a Figura 4.11, o respetivo ficheiro JSON indica que existem 14 cromossomas, 3 *clusters*, 1 nucléolo e 3 objetos ruidosos. As bboxes são definidas do mesmo modo que no LabelMe, onde são usadas as coordenadas dos vértices opostos da caixa retangular delimitadora da estrutura celular.

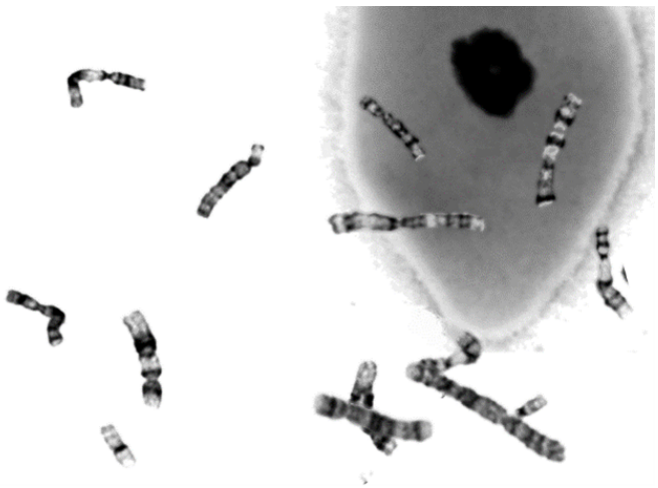


Figura 4.11: Exemplo de uma imagem celular sintetizada pelo algoritmo proposto. Imagem original.

Durante o processo de colagem, foram também criadas máscaras, píxel a píxel, relativas às sobreposições, à diferenciação das estruturas coladas e aos pixéis que sofreram o método de *blending*. Uma vez que o objetivo deste *dataset* era ser o mais caraterizado possível, foram guardadas as estruturas celulares usadas para criar cada imagem. Todas estas máscaras, assim como a imagem com as *labels* estão demonstradas na Figura 4.12.

No total foram obtidas 10795 imagens sintéticas que contêm 274691 cromossomas (dos quais estão inseridos em 78428 *clusters*), 5437 núcleos interfásicos e 132899 objetos ruidosos. Em termos das dimensões das imagens obtidas, observa-se na Figura 4.13 que tanto a altura como a largura das imagens obtidas estão quase uniformemente dispersas no intervalo de valores pré-definido para esses parâmetros. Em relação à resolução das imagens, estas estão intervaladas entre os 19kB e os 206kB, dependendo da complexidade da imagem.

4.3.2 Discussão

Por intermédio da aplicação da fase *"Paste"* obteve-se um *dataset* sintético com mais de 10000 imagens. Apesar de ser sintético, na medida em que as imagens não são reais, todas as estruturas celulares presentes nestas imagens resultam de recortes de microfotografias da citogenética. Assim sendo, existe um fotorrealismo inerente a este *dataset* no que toca às estruturas celulares, sejam elas cromossomas ou núcleos interfásicos. Este fotorrealismo é também conseguido devido à forma como as estruturas interagem umas com as outras – o método de *blending* aplicado permitiu uma suavização da imagem no caso das sobreposições, contribuindo para um resultado final que não se encontra longe da realidade. O ruído aplicado nas imagens também foi retirado das imagens do laboratório de citogenética, o que também contribui para tornar as imagens o mais parecidas às microfotografias do *dataset* do LCG-FMUC.

Em relação aos parâmetros aleatórios usados para gerar uma elevada variabilidade de imagens, definiu-se o número de cromossomas entre 4 e 46, por forma a obterem-se imagens tanto simples como complexas, sendo que apenas as de 46 cromossomas se assemelham, por completo, à realidade. Para o número de núcleos interfásicos, foi definido, no máximo, um nucléolo por imagem, seguindo a

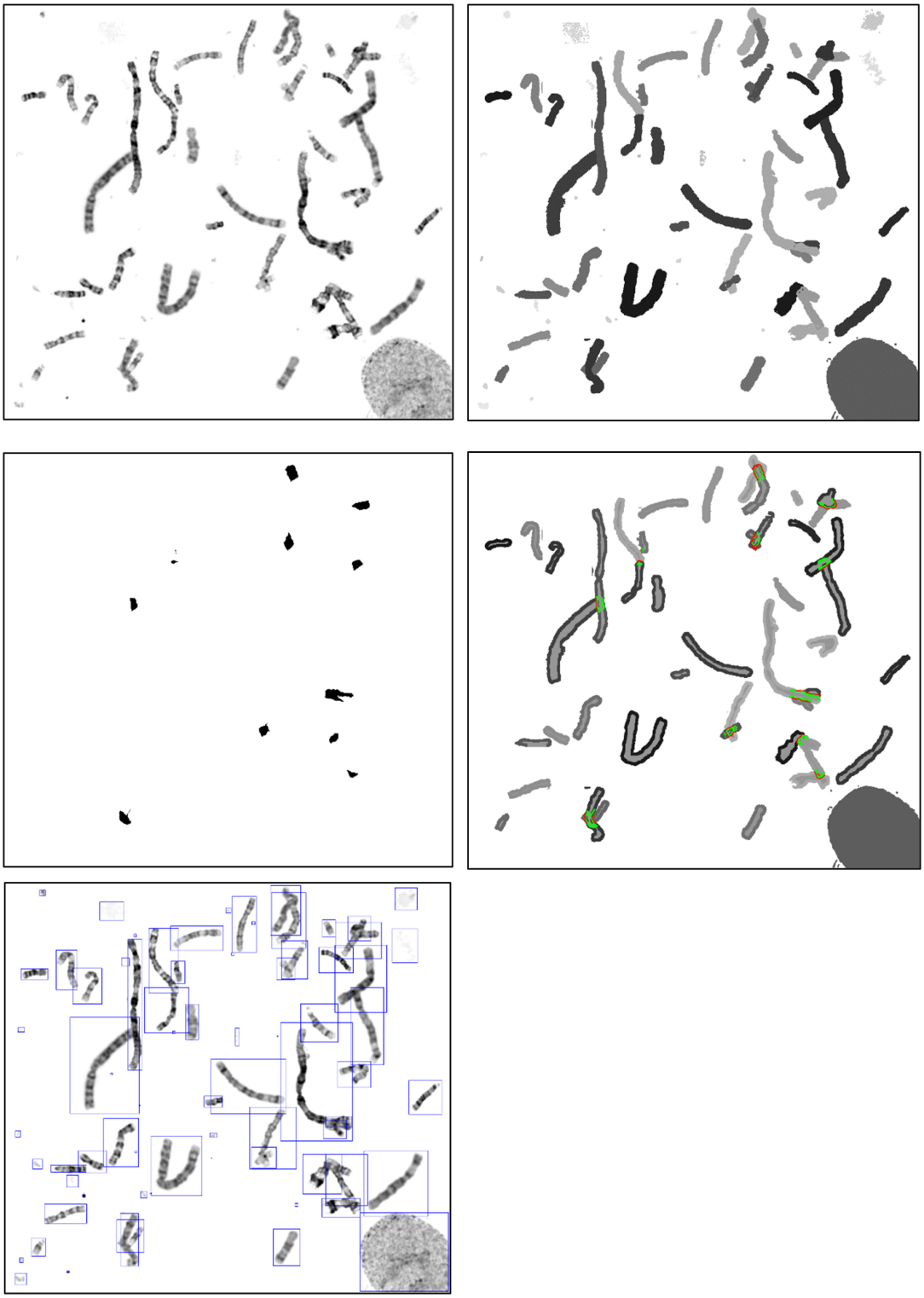


Figura 4.12: Imagem celular sintética e respetivas máscaras sintetizadas para fins de segmentação. No canto superior esquerdo, apresenta-se a imagem celular sintética. No canto superior direito, observa-se a máscara de cada estrutura. Imediatamente abaixo, encontra-se a máscara dos pixéis que sofreram o método de *blending*. Ao seu lado esquerdo, apresenta-se a máscara relativa às sobreposições. No canto inferior, esquerdo observam-se as *labels* de cada estrutura celular. Imagem original.

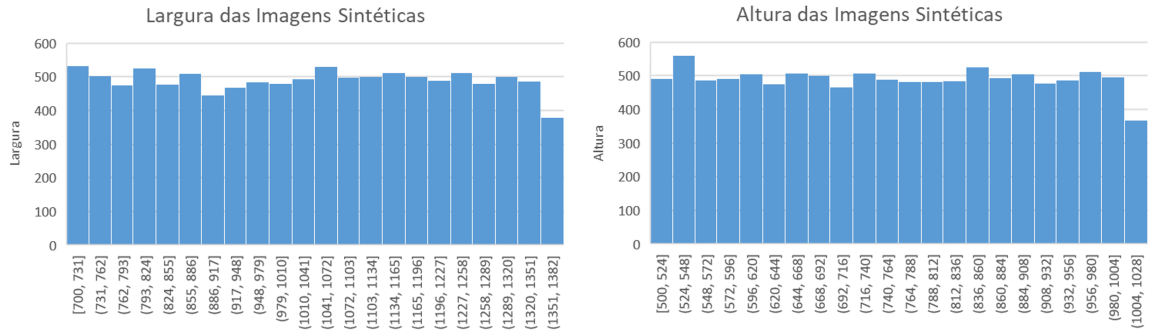


Figura 4.13: Distribuição das dimensões (largura e altura) das imagens do *dataset* sintético. Imagem original.

tendência das imagens do LCG-FMUC. Quanto aos objetos ruidosos, utilizou-se um máximo de 50 objetos ruidosos e um mínimo de 0, de forma a se obterem imagens de complexidade diferente. Relativamente às dimensões das imagens, o intervalo de larguras e alturas das imagens sintéticas foi definido tendo em conta os máximos e mínimos das dimensões das imagens obtidas no *dataset* do LCG-FMUC. A definição aleatória e uniforme da posições das colagens atribuiu ainda mais variedade ao *dataset*, podendo ser interpretada como uma operação de translação de DA.

Quanto ao método de *blending* utilizado, existiram várias experiências realizadas a nível das sobreposições. Inicialmente, a colagem direta e carente de estruturas sobrepostas resultou na criação de margens esbranquiçadas – devida aos pixéis das margens longitudinais dos cromossomas. Posteriormente, reduziu-se o *threshold* dos pixéis que eram colados, de forma a diminuir essas margens esbranquiçadas. No entanto, o mesmo artefacto permanecia na imagem, desta vez um pouco mais esbatido. De seguida, foi efetuada a localização dos pixéis sobrepostos e respetiva aplicação do filtro de caixa – apesar deste método resultar numa melhor fusão das estruturas, esta ainda não era fotorrealista. Nas sobreposições reais de cromossomas, o único *blending* que existe é a nível das margens longitudinais do cromossoma superior, assim como no caso dos pixéis deste cromossoma serem muito claros comparativamente aos pixéis do cromossoma inferior. Portanto, tentou-se ao máximo isolar estas margens longitudinais aplicando-se um filtro de

caixa a duas fases. Relativamente ao tipo de filtro usado, recorreu-se ao filtro de caixa pelos mesmos motivos indicados na secção 4.1.2.

Comparativamente aos trabalhos de Chen *et al.* [56] e Song *et al.* [59], o método de *blending* usado permite obter sobreposições mais próximas da realidade da citogenética. Em Chen *et al.* [56], utilizou-se o somatório das intensidades dos pixéis sobrepostos e a atribuição de pesos aleatórios a cada uma dessas intensidades. Estes métodos produzem regiões sobrepostas que se destacam muito do resto do *cluster*, enquanto na realidade da citogenética há uma fusão quase indistinta ao olho nu. Em Song *et al.* [59], usou-se a opacidade da sobreposição, mantendo-se mais nítido o cromossoma de cima. A razão para esta colagem deveu-se a um motivo bastante generalista por parte dos autores, de onde estes concluíram que as sobreposições opacas são muito mais comuns do que sobreposições translúcidas. Contudo, observando imagens de citogenética, é notório que nas sobreposições existe uma opacidade do cromossomas de cima, com exceção das suas margens longitudinais, que são translúcidas e resultantes de uma fusão com os pixéis do cromossoma de baixo. Este efeito é conseguido pelo método proposto.

De acordo com a literatura encontrada, o *dataset* obtido destaca-se dos *datasets* disponibilizados para reprodução de resultados de segmentação de cromossomas (consultar Anexo C). Comparativamente aos *datasets DeepFish* e *Overllaping Chromosome Instance Segmentation*, o *dataset* proposto nesta dissertação apresenta imagens celulares fotorrealistas, em vez de imagens compostas apenas por um *cluster*, seja ele composto por dois ou mais cromossomas. Além disso, apresenta 274691 cromossomas na sua totalidade, quando comparado com 26868 e 80494 cromossomas apresentados pelos *datasets* previamente mencionados, respetivamente. No que diz respeito aos *clusters*, estes *datasets* apresentam 13434 e 29180 *clusters*, respetivamente, enquanto o *dataset* proposto apresenta 78428 *clusters*. Relativamente aos *datasets CRCN-NE* e *Bioimlab*, que resultam de imagens reais de citogenética, o *dataset* proposto ultrapassa-os em termos de volume de dados – apesar de não estarem quantificados os cromossomas nem os *clusters* presentes, esses *datasets* são compostos unicamente por 74 e 162 imagens, respetivamente. No caso do *dataset* do *CRCN-NE*, a resolução das imagens é de

baixa qualidade e os cromossomas não apresentam padrão de bandas, estando apenas coloridos homogeneamente por Giemsa. No caso do *dataset* do *Bioimlab*, foi utilizado o padrão de bandas Q, que atualmente está em desuso. Quanto aos outros *datasets* referidos na literatura como públicos, mas cujo acesso não foi possível encontrar, a qualidade das imagens não pode ser aferida. Ainda assim, em termos de volume de dados e de caraterização da informação, o *dataset* proposto supera os *datasets DGMU*, *ChromSeg* e *Copenhagen e Saravejo*.

Deve-se ainda acrescentar que o *dataset* proposto fornece máscaras, píxel a píxel, das estruturas celulares presentes em cada imagem e das sobreposições. No caso das estruturas celulares, cada uma tem uma intensidade diferente (indicada no respetivo ficheiro JSON), enquanto para as sobreposições é utilizada um binarização da imagem com uma intensidade igual a 0 para os pixéis sobrepostos. Assim, este *dataset* também pode ser utilizado para fins de segmentação semântica.

4.4 Modelo de Segmentação de Cromossomas

4.4.1 Resultados

A última fase da metodologia corresponde à fase *"Learn"*, onde foi usada a rede neuronal YOLOv5 para segmentar os cromossomas. Primeiramente, procedeu-se à conversão do formato das *labels* para o formato YOLO. De seguida, usaram-se *datasets* de volumes diferentes associados a diferentes modelos YOLOv5 pré-treinados no *dataset* COCO. Por fim, escolheu-se o melhor modelo com base em métricas de exatidão dos respetivos grupos validação, sendo subsequentemente testado nas imagens de células em metafase do *dataset* do LCG-FMUC.

Tal como mencionado na Secção 3.2.4.2, por forma a implementar o YOLOv5 é necessário converter o grupo de identificação das estruturas celulares, assim como as coordenadas da bbox de cada *label*. Enquanto o grupo de identificação precisa de ser indentado a partir do algarismo zero, as bboxes são representadas pelas coordenadas do seu centro geométrico e pelas suas dimensões, ambas normalizadas em relação às dimensões da imagem onde a bbox está inserida. O resultado da conversão do formato das *labels* resultou na mudança de um ficheiro JSON para um ficheiro

TXT mais simplificado, como mostra a Figura 4.14. Para aferir a validação desta conversão, as novas *labels* foram visualizadas e comparadas com as *labels* do *dataset* sintético (representadas a cor verde e azul na Figura 4.14, respetivamente). Como pode ser observado, apenas estão presentes as *labels* dos cromossomas – a função desenvolvida para a conversão do formato das bboxes permite selecionar quais as estruturas celulares de interesse. Nesta dissertação, a segmentação de cromossomas foi encarada como uma classificação a uma classe, descartando-se, por isso, as classes do nucéolos e objetos ruidosos para treino dos modelos YOLOv5.

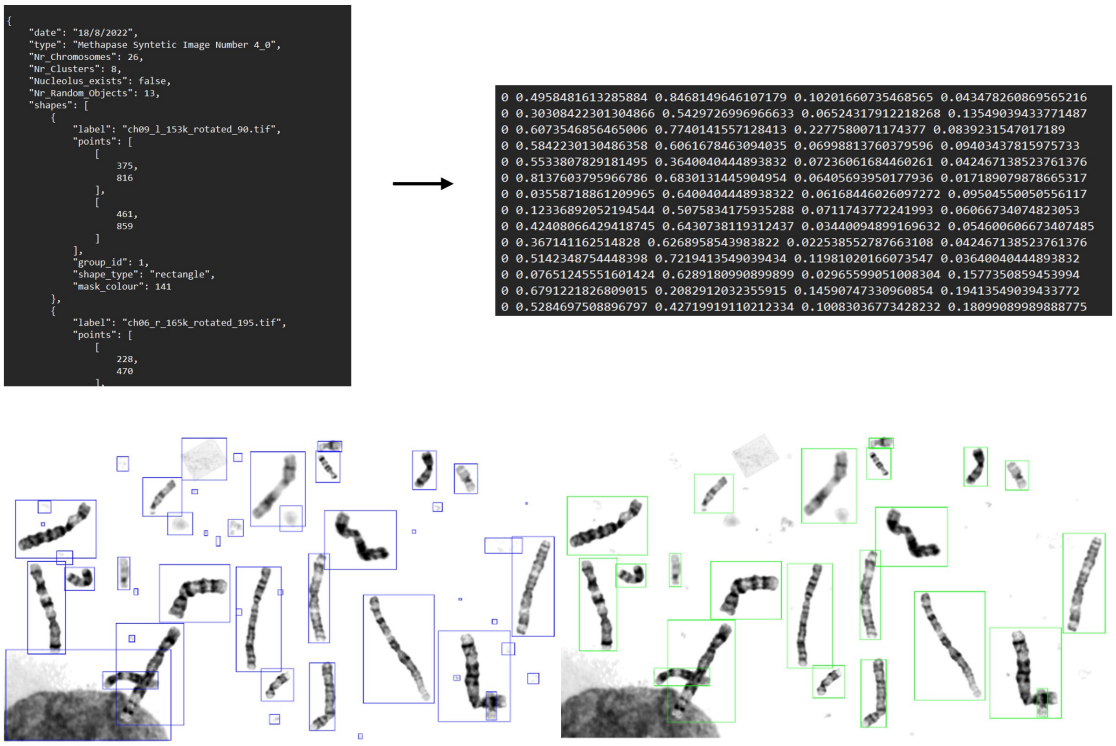


Figura 4.14: Resultado ilustrativo da conversão de *labels* para o formato YOLO. Acima, apresentam-se recortes dos ficheiros JSON e TXT, que mostram a diferença de estrutura para guardar as *labels*. Abaixo, observa-se a validação desta conversão - a azul estão as *labels* de todas as estruturas no formato anterior e a verde estão as *labels* dos cromossomas no formato YOLO. Imagem original.

De seguida, foram criados quatro *datasets* de grupos treino e validação, usados para treinar três modelos do YOLOv5. Para cada um destes *datasets* foi utilizada uma proporção 90/10 relativamente à divisão do grupo treino e validação. Todos os

datasets são provenientes do *dataset* sintético criado anteriormente.

Inicialmente, foram escolhidas, de forma aleatória, 200 imagens, que resultaram num grupo treino de 180 imagens e um grupo validação de 20 imagens. O modelo YOLOv5 pré-treinado YOLOv5s (*small*), que corresponde ao modelo pequeno, foi treinado neste *dataset*. Este modelo demorou 15 minutos a ser treinado no Google Colab para 100 épocas. Atingiu-se uma mAP@0.5 de 0.939 e uma mAP@0.5 : 0.05 : 0.95 de 0.632.

Seguidamente, foram escolhidas aleatoriamente 1000 imagens, resultando num grupo treino de 900 imagens e um grupo validação de 100 imagens. Analogamente, o modelo utilizado para treinar este *dataset* foi o YOLOv5s. Este modelo demorou 70 minutos a ser treinado no Google Colab para 100 épocas. Atingiu-se uma mAP@0.5 de 0.979 e uma mAP@0.5 : 0.05 : 0.95 de 0.749.

Aós a aplicação no modelo YOLOv5 pequeno, utilizou-se o mesmo *dataset* de 1000 imagens para treinar o modelo pré-treinado YOLOv5m (*medium*) - que corresponde ao modelo médio. Este modelo demorou 90 minutos a ser treinado no Google Colab para 100 épocas. Atingiu-se uma mAP@0.5 de 0.985 e uma mAP@0.5 : 0.05 : 0.95 de 0.789.

Posteriormente, foram escolhidas aleatoriamente 5000 imagens, resultando num grupo treino de 4500 imagens e um grupo validação de 500 imagens. Desta vez, o modelo utilizado para treinar este *dataset* foi o YOLOv5l (*large*) - correspondente ao modelo grande. Este modelo demorou 250 minutos a ser treinado no Google Colab para 40 épocas. Atingiu-se uma mAP@0.5 de 0.983 e uma mAP@0.5 : 0.05 : 0.95 de 0.713.

Por último, foram seguidas as seguintes medidas aconselhadas pela Ultralytics, o criador do YOLOv5, de forma a se conseguir atingir o melhor resultado possível ao treinar modelos da família YOLO. Estas medidas são referentes aos dados usados no grupo treino, e referem que:

- Devem existir pelo menos 1500 imagens por classe;
- Devem existir mais de 10000 objetos rotulados por classe;

- Deve haver uma grande variabilidade dentro do *dataset* relativamente às imagens usadas;
- Tem de existir consistência nas *labels* marcadas para cada classe;
- Deve ser assegurado o mínimo de espaço entre a bbox e as extremidades das estruturas celulares;
- Cerca de 0 a 10% do *dataset* deve ser composto por imagens de *background* que não contenham nenhuma classe, de forma a reduzir a quantidade de Falsos Positivos.

Com estas recomendações em mente, foi utilizado o *dataset* total de 10795 imagens, resultando num grupo treino de 9715 imagens e um grupo validação de 1080 imagens. Utilizaram-se ainda 1000 imagens de *background*, criadas apenas com núcleos interfásicos e objetos ruidosos. Estas imagens foram inseridas no grupo de treino, resultando num total de 10715 imagens. Para este treino, o modelo utilizado foi o YOLOv5l (*large*). Este modelo demorou cerca de 14 horas a ser treinado no Google Colab para 75 épocas. Atingiu-se uma mAP@0.5 de 0.989 e uma mAP@0.5 : 0.05 : 0.95 de 0.771.

A análise extensiva da progressão dos cinco modelos obtidos durante o seu treino está apresentada no Anexo F. Os gráficos presentes nesse anexo traduzem a evolução das métricas *Box Loss* e *Objectness Loss* para os grupos treino e validação. A *Box Loss* representa o quão bem o algoritmo consegue localizar o centro de um objeto e quão bem a bbox cobre esse mesmo objeto. Por sua vez, a *Objectness Loss* apresenta uma medida de probabilidade de que um objeto existe numa certa região de interesse. Ambas as métricas devem decrescer ao longo do treino do modelo para se obterem bons resultados [73]. Além disso, também são apresentados os gráficos respetivos à *mean Average Precision* para os grupos validação. Estas métricas já foram explanadas previamente, contudo salienta-se, novamente, que estas devem aumentar ao longo do treino de um modelo. Um resumo das métricas mAP@0.5 e mAP@0.5 : 0.05 : 0.95 está detalhado na Tabela 4.2.

Tendo em conta as métricas de avaliação, o quinto modelo obtido, respetivo ao YOLOv5l e ao *dataset* total de 10795 imagens, foi o modelo com melhor

Modelo	**Imagens no Dataset**	**mAP**@0.5	**mAP**@0.5 : 0.05 : 0.95
YOLOv5s	200	0.939	0.632
	1000	0.979	0.749
YOLOv5m	1000	0.985	0.789
YOLOv5l	5000	0.983	0.713
	10795	0.989	0.771

Tabela 4.2: Métricas mAP@0.5 e mAP@0.5 : 0.05 : 0.95 para os grupos validação dos modelos treinados. Tabela original.

desempenho no respetivo grupo validação. Assim sendo, foi este o modelo escolhido para ser testado nas imagens de células em metafase do *dataset* do LCG-FMUC. Neste *dataset* existem 171 imagens da citogenética convencional, previamente pré-processadas (Secção 3.2.1), onde estão inseridos 7861 cromossomas. O resultado da deteção deste modelo está exemplificado na Figura 4.15, onde se observa a identificação de cromossomas com uma bbox e a respetiva confiança de previsão.

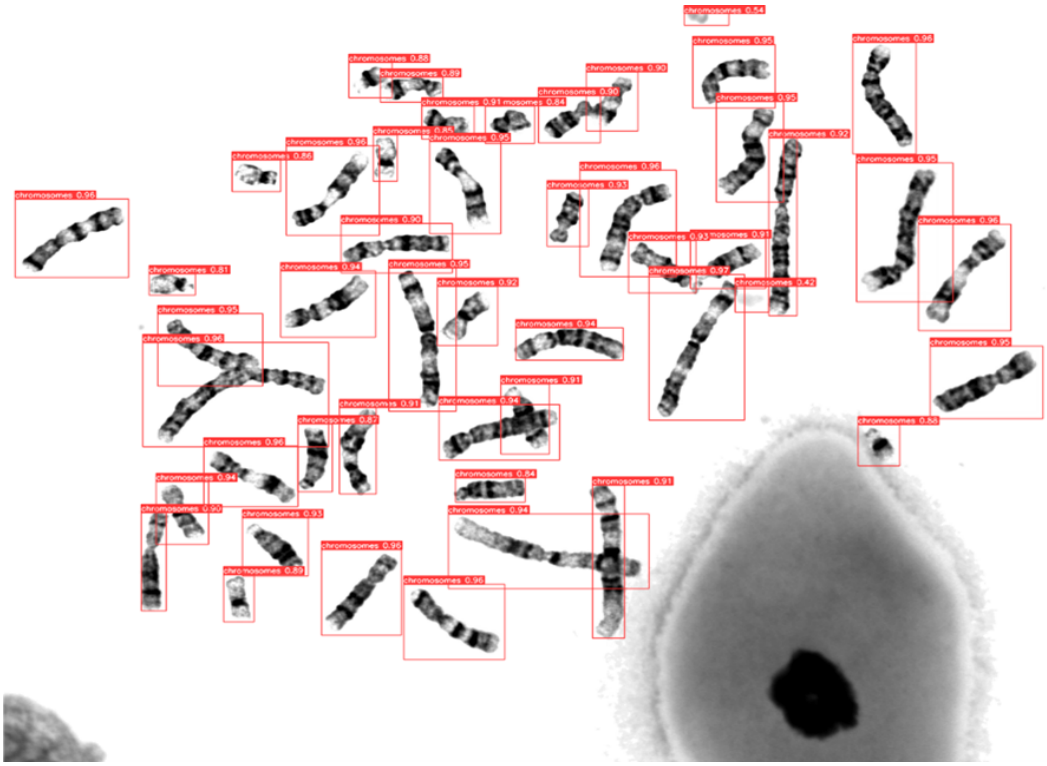


Figura 4.15: Resultado da deteção de cromossomas pelo modelo YOLOv5l treinado no *dataset* sintético de 10795 imagens. Imagem escolhida ao acaso do *dataset* obtido.

Uma vez que o modelo obtido atribui valores de confiança às *labels* obtidas, analisou-

se o resultado da deteção para diferentes valores de confiança. Por exemplo, previsões com níveis de confiança de 0.25, 0.70 e 0.90 resultam em diferentes *labels*, como pode ser observado na Figura 4.16. Assim, foram ainda adquiridas métricas para valores de confiança entre 0.25 e 0.90, intervalados num valor de 0.05. Os resultados estão detalhados na Tabela 4.3 e apresentam-se sob a forma de gráfico na Figura 4.17.

Valores de Confiança	**TP**	**FP**	**Número Total de previsões**	***Recall***	***Precision***	***F1 Score***
0.25	7772	644	8416	0.9887	0.9235	0.9550
0.30	7760	611	8371	0.9872	0.9270	0.9561
0.35	7748	580	8328	0.9856	0.9304	0.9572
0.40	7739	544	8283	0.9845	0.9343	0.9587
0.45	7728	514	8242	0.9831	0.9376	0.9598
0.50	7739	471	8210	0.9845	0.9426	0.9631
0.55	7747	418	8165	0.9855	0.9488	0.9668
0.60	7749	363	8112	0.9858	0.9553	0.9703
0.65	7742	309	8051	0.9849	0.9616	0.9731
0.70	7708	249	7957	0.9805	0.9687	0.9746
0.75	7612	186	7798	0.9683	0.9761	0.9722
0.80	7260	130	7390	0.9235	0.9824	0.9521
0.85	6274	64	6338	0.7981	0.9899	0.8837
0.90	3997	25	4022	0.5085	0.9938	0.6727

Tabela 4.3: Métricas de avaliação de desempenho para diferentes valores de confiança do modelo obtido para segmentação de cromossomas. Tabela original.

Relativamente à identificação de cromossomas inseridos em *clusters*, é notória a capacidade de deteção deste modelo em diferentes níveis de complexidade de estruturas cromossómicas. O algoritmo é capaz de detetar cromossomas individualizados e direitos, cromossomas dobrados, cromossomas a tocarem-se, cromossomas inseridos em *clusters* de dois cromossomas em forma de T, '+' ou 'x', cromossomas sobrepostos mais do que uma vez num só *cluster* e cromossomas em *clusters* de três cromossomas (Figura 4.18). Além disso, a deteção de cromossomas ainda consegue ser efetuada em estruturas super complexas, por exemplo, cromossomas em *clusters* de quatro cromossomas e cromossomas inseridos em estruturas cromossómicas compostas por vários *clusters* (Figura 4.19).

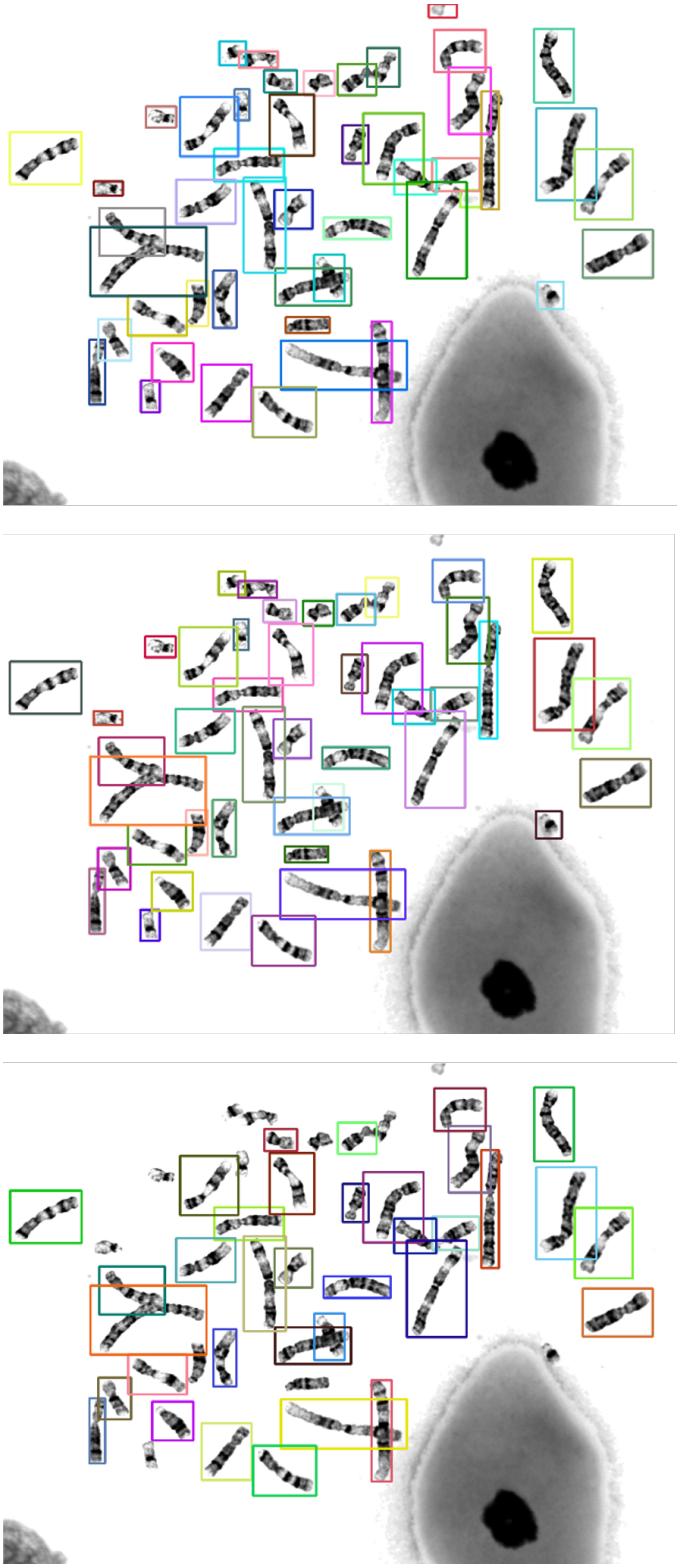


Figura 4.16: Exemplos da aplicação de diferentes valores de confiança para a deteção de cromossomas na mesma imagem celular, escolhida aleatoriamente do *dataset* do LCG-FMUC. Os valores de confiança usados para obter estes resultados foram 25% (acima), 70% (meio) e 90% (abaixo). Imagem original.

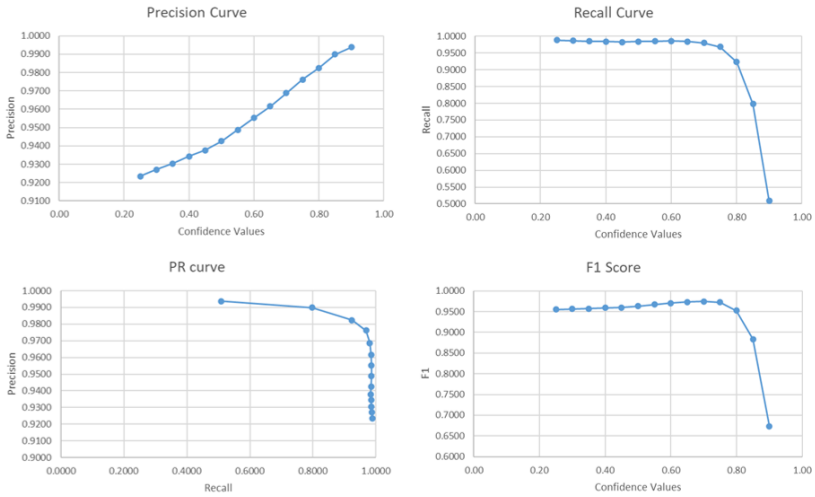


Figura 4.17: Métricas de avaliação de desempenho para a segmentação do *dataset* do LCG-FMUC, através do modelo obtido pelo YOLOv5l treinado no *dataset* sintético de 10795 imagens. Imagem original.

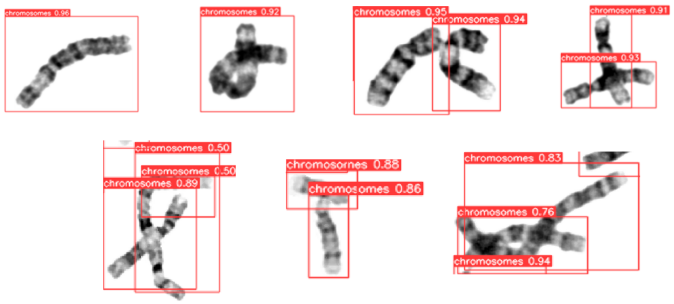


Figura 4.18: Resultado da deteção de cromossomas em *clusters* compostos, no máximo, por três cromossomas. A complexidade da estrutura aumenta da esquerda para a direita e de cima para baixo. Imagem original.

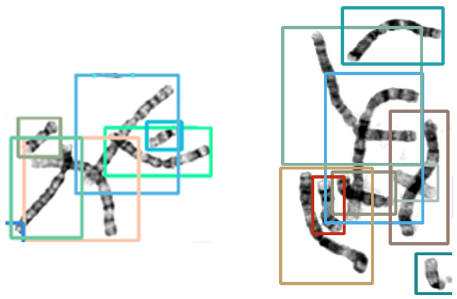


Figura 4.19: Resultado da deteção de cromossomas em estruturas compostas, no mínimo, por quatro cromossomas. À esquerda, encontra-se um estrutura cromossómica com cinco cromossomas. À direita, a estrutura é composta por nove cromossomas. Imagem original.

4.4.2 Discussão

O primeiro passo para a implementação dos modelos YOLOv5 passou pela conversão das *labels* presentes no formato JSON do *dataset* sintético, para o formato YOLO. Esta conversão foi puramente matemática, na medida em que houve apenas transformação de coordenadas da bbox, e permitiu adquirir os ficheiros TXT necessários para treinar os modelos pretendidos. De forma a verificar se a conversão foi bem aplicada, fez-se a validação gráfica da mesma, através da comparação das *labels* antes e depois dessa mesma conversão. É de notar que, nesta passagem de ficheiros JSON para ficheiros TXT, foi incluída uma opção na ferramenta usada para tal, que possibilita a escolha das estruturas celulares que se querem inserir como classes na fase de treino e deteção por parte do modelo YOLOv5. Uma vez que o objetivo essencial deste algoritmo é a segmentação de cromossomas, decidiu-se, para simplificar o problema de deteção, que se iria apenas usar a classe dos cromossomas. Assim, as *labels* dos núcleos interfásicos e dos objetos ruidosos foram descartadas e obteve-se um "novo" *dataset*, onde estão apenas presentes as imagens TIF do *dataset* sintético e os ficheiros TXT respetivos às *labels* dos cromossomas.

Posteriormente, realizou-se a escolha aleatória de imagens para se obterem diferentes volumes de dados – obtiveram-se quatro *datasets* distintos de 200, 1000, 5000 e 10795 imagens celulares sintéticas. Estes *datasets* foram usados para treinar diferentes modelos, por forma a aferir-se qual a influência do tamanho do *dataset* no desempenho do modelo em causa. Da mesma forma, foram utilizados diferentes modelos pré-treinados do YOLOv5 que serviram de base para serem treinados em *datasets* diferentes, permitindo analisar-se o impacto destes modelos pré-treinados.

Quanto à escolha dos parâmetros utilizados para treinar e comparar os modelos, nomeadamente o número de épocas escolhidas, escolheu-se 100 para os modelos mais rápidos (YOLOv5s e YOLOv5m) e 40 ou 75 para os modelos mais lentos (YOLOv5l). Esta escolha foi feita com base na convergência das métricas de evolução no seu *plateau*, mas também no tempo de processamento dos modelos aplicados. Caso o número de épocas fosse maior, esperar-se-ia uma convergência mais lenta das métricas de desempenho a partir das 100, 40 e 75 épocas,

respetivamente, aumentando-se a extensão do *plateau* das mesmas. Este aumento de épocas poderia incorrer no erro de *overfitting*, que corresponde ao sobre-ajuste do modelo obtido aos dados treino, tornando-se menos eficaz quando aplicado a outros dados. Quanto à diferença do número de épocas definidas para os modelos mais lentos (YOLOv5l), esta deve-se principalmente ao tempo de treino dos mesmos. Uma vez que estes modelos demoraram entre 5 a 11 minutos a ser treinados por época, e os recursos disponibilizados no Google Colab são limitados por tempo de uso continuado da GPU atribuída ao utilizador em questão, fez-se a escolha do número de épocas com base na evolução da convergência das métricas de desempenho. Tendo em conta os gráficos destas métricas, apresentados no Anexo F, verifica-se que existe a convergência nos seus *plateaus*. Apesar de não ter sido aplicada nenhuma técnica que prevenisse *overfitting*, é possível inferir-se que as métricas *Box Loss* e *Objectness Loss* estão gradualmente a decrescer em todos os modelos. Concluiu-se, desta forma, que foi evitado *overfitting* dos dados. Contudo, a aplicação de uma técnica, como o *Early Stopping*, é benéfica para o treino de modelos de DL.

Analisando-se as métricas mAP@0.5 e mAP@0.5 : 0.05 : 0.95 obtidas para os três tipos de modelos, conclui-se que o modelo YOLOv5s é aquele que apresenta um menor desempenho. Para o mesmo tamanho de *dataset* de 1000 imagens, utilizando os modelos YOLOv5s e YOLOv5m, observa-se a melhoria do desempenho de 0.979 para 0.85 e de 0.749 para 0.789 nas métricas mAP@0.5 e mAP@0.5 : 0.05 : 0.95, respetivamente. Quanto à comparação entre o modelo YOLOv5m e YOLOv5l, a métrica de mAP@0.5 manteve-se relativamente constante, mas a mAP@0.5 : 0.05 : 0.95 foi bastante superior no modelo médio. Todavia, esta análise entre os modelos médio e grande não pode ser concluída, uma vez que não se manteve constante o volume de imagens do *dataset* utilizado. A acrescentar, a análise entre os modelos usados não é perfeita, tendo em conta que o número de épocas usado não foi mantido constante.

Relativamente ao tamanho do *dataset*, concluiu-se que o desempenho dos modelos é positivamente afetado quando se aumenta o número de imagens. Comparando o resultado do YOLOv5m treinado com o *dataset* de 1000 imagens ao resultado do YOLOv5l treinado com o *dataset* de 10795, apesar de se ter obtido uma métrica

mAP@0.5 : 0.05 : 0.95 inferior (0.789 e 0.771), o YOLOv5l tem um grupo de validação 10 vezes superior ao grupo de validação do YOLOv5m. Por este motivo, o modelo YOLOv5l que foi treinado no maior *dataset* foi aquele selecionado para ser testado no *dataset* do LCG-FMUC, uma vez que indica um bom desempenho do algoritmo quando comparado com um número elevado de imagens com diferentes níveis de complexidade morfológica. Salienta-se ainda o facto de que, apesar de se ter proposto uma rotina para gerar *datasets* de elevada dimensão, obtiveram-se, no máximo, 10795 imagens, sendo este o maior *dataset* usado para treinar os modelos YOLOv5. Uma vez que a geração automática de cromossomas está associada a um tempo de execução discutido anteriormente (10 a 20 minutos por imagem), para se obter um *dataset* de, por exemplo, 100000 imagens, seriam necessários mais de 1000 dias (considerando que se usa apenas um computador e não se utiliza processamento em paralelo). Assim, o *dataset* de 10795 imagens foi o possível de ser obtido dentro do intervalo em que esta dissertação foi desenvolvida e para os recursos disponíveis. Caso se tivesse obtido um *dataset* de maior dimensão, era de esperar uma melhoria nas métricas de desempenho, acompanhadas de um maior tempo associado ao treino por época. Além disso, seria de esperar um maior *plateau* para um número de épocas igual aos dos outros modelos.

A aplicação do modelo acima mencionado no grupo teste, constituído por 171 imagens celulares de cromossomas em metafase da citogenética, apresenta resultados bastante satisfatórios e que traduzem a excelente capacidade de deteção por parte do YOLOv5. A segmentação de cromossomas é assegurada por este algoritmo, sendo necessário estabelecer um *threshold* do valor de confiança das previsões atribuídas pelo mesmo. Com essa finalidade, analisaram-se as métricas *Recall*, *Precision* e *F1-score* para diferentes valores de confiança. Estes valores foram limitados por 0.25 e 0.90, tendo em conta que abaixo desse limite inferior, o número de previsões aumenta de tal forma que começa a classificar muitos falsos positivos, e que acima desse limite superior, o número de previsões decresce na mesma proporção, obtendo-se valores muito reduzidos de *Recall*.

Como seria de esperar, os valores de *Recall* e de *Precision* apresentam tendências inversas nas suas relações com os valores de confiança. À medida que o valor de

confiança do algoritmo aumenta, o número de falsos positivos diminui, o que significa que a *Precision* vai aumentar. Contudo, o número total de previsões irá diminuir, ou seja, os verdadeiros positivos diminuem, diminuindo também a *Recall*. Posto isto, o *F1-score*, que pondera estas duas métricas, apresenta um máximo absoluto para o valor de confiança 0.7. Concluiu-se, então, que o algoritmo apresenta o melhor desempenho para um valor de confiança de 70%. Considerando este valor de confiança, o modelo apresenta um sucesso de 98.05% na segmentação de cromossomas - 7708 dos 7861 cromossomas presentes nas imagens foram corretamente segmentados. O bom desempenho do algoritmo neste grupo de teste é ainda representado pela sua curva *Precision-Recall*, que mantém os valores de *Precision* muito próximos de 1 à medida que a *Recall* aumenta.

Apesar do treino do modelo YOLOv5l com o *dataset* de 10795 imagens ser bastante demorado (cerca de 14 horas), o uso deste modelo para a subsequente fase de inferência e deteção de cromossomas é bastante rápido. Uma vez que o YOLO apresenta uma metodologia *single-shot*, tal como explicado anteriormente, no grupo teste com o *dataset* do LCG-FMUC, o modelo treinado demorou 1079 ms por imagem a segmentar os cromossomas. Conclui-se que o algoritmo apresentado supera o único tempo de processamento encontrado na literatura e referido no Anexo B, referente à segmentação dos cromossomas de uma imagem celular, que corresponde a [2-7] segundos por imagem - algoritmo proposto por Altinsoy *et al.* [39].

Finalmente, infere-se que este algoritmo é capaz de identificar cromossomas em diferentes níveis de complexidade. Através da análise gráfica das imagens do grupo teste, após a deteção, deduz-se que o algoritmo é capaz de segmentar cromossomas a partir de variados níveis de complexidade de estruturas cromossómicas. Todavia, quando a complexidade da estrutura cromossómica aumenta, o algoritmo segmenta os cromossomas apresentando menores valores de confiança. Relativamente às várias formas de *clusters*, determinou-se que a estrutura de *cluster* em "x" se apresenta como a maior dificuldade em termos da segmentação correta de cromossomas por este algoritmo de deteção.

5

Conclusão

Este capítulo apresenta as conclusões referentes ao projeto desenvolvido sobre a segmentação automática de cromossomas em imagens microscópicas de cromossomas. Os quatro objetivos definidos no início desta dissertação e descritos na Secção 1.2 foram consolidados:

1. A primeira fase desta dissertação correspondeu à análise dos conceitos e dos algoritmos referentes à segmentação automática de cromossomas. Para isso, foi estudada uma extensa bibliografia através do levantamento de artigos publicados em revistas ou apresentados em conferências. A revisão de literatura foi sistematizada, utilizando-se os repositórios PubMed, Google Schoolar e IEEE Xplore, assim como algumas *keywords*, nomeadamente "*Automatic Segmentation of Chromosomes*", "*Chromosome Image Segmentation*" e "*Overlapping Chromossomes*". Daí concluiu-se que, de modo geral, o maior entrave ao sucesso dos algoritmos de segmentação é a existência de *clusters* de cromossomas que pressupõem sobreposição e, consequentemente, dificultam a segmentação automática. Depois da análise dos métodos de segmentação e dos *datasets* usados na literatura, é possível concluir-se ainda que, atualmente, os métodos de aprendizagem, principalmente de *Deep Learning*, são a melhor alternativa para abordar a segmentação de *clusters* de cromossomas. Além desta conclusão, ainda se infere que o maior obstáculo face a um método totalmente eficaz é a falta de *datasets* clínicos de grandes dimensões ou *datasets* sintéticos que se assemelhem à realidade da citogenética clínica. Desta forma, foi apontada a necessidade de disponibilização de *datasets* clínicos ou *datasets* artificiais

fotorrealistas para fins de segmentação de cromossomas e subsequente reprodução de resultados.

2. Posteriormente a serem analisados os *datasets* usados na literatura, de forma qualitativa e quantitativa, procedeu-se à criação de uma metodologia capaz de automatizar a sintetização de imagens celulares fotorrealistas de cromossomas em metafase. Para isso, utilizou-se uma metodologia "*Cut and Paste*" onde, num primeiro momento foram obtidas estruturas celulares presentes em imagens da citogenética, e de seguida foram coladas e suavizadas para se assemelharem a imagens reais. Na fase "*Cut*" foi obtido um primeiro *dataset* de estruturas celulares reais, composto por 115896 cromossomas, 180 núcleos interfásicos e 6024 objetos ruidosos. Este *dataset* inclui informações pormenorizadas da classe de cada cromossoma, permitindo o uso deste *dataset* para outros fins, como por exemplo para a classificação automática de cromossomas. Na fase "*Paste*" foram obtidas 10795 imagens sintéticas, utilizando-se um método de *blending* proposto para a suavização das sobreposições das estruturas celulares. Assim, cumpriu-se o objetivo referente à disponibilização de um *dataset* sintético baseado em estruturas celulares reais da clínica, completamente caraterizado em termos de número de cromossomas e de *labels*, e que podem ser usadas tanto para segmentação de objetos como para segmentação semântica.

3. O grande objetivo desta dissertação corresponde à segmentação de cromossomas a partir de imagens celulares de cromossomas em metafase. Para isso, foram implementados modelos YOLOv5, que são capazes de detetar objetos com alta precisão e rapidez. Os modelos pré-treinados do YOLOv5 foram *fine-tuned* com base no *dataset* de imagens sintéticas e concluiu-se que o modelo YOLOv5l apresenta as melhores métricas para a deteção de cromossomas. No grupo de validação, composto por 1080 imagens, este modelo obteve um valor de mAP@0.5 igual a 0.989 e um valor de mAP@0.5 : 0.05 : 0.95 igual a 0.771, podendo afirmar-se que o YOLOv5 é uma boa ferramenta para a deteção de cromossomas no *dataset* sintético.

4. O modelo YOLOv5 obtido para a segmentação de cromossomas foi posteriormente validado no *dataset* do LCG-FMUC, constituído por 171

microfotografias de células de cromossomas em metafase. Analisando as métricas *Precision* e *Recall* obtidas para diferentes valores de confiança do modelo proposto, concluiu-se que para um valor de confiança de predição igual a 70%, o algoritmo apresenta um *F1-score* de 0.9746. Este valor representa a classificação correta de 7708 cromossomas dos 7861 cromossomas existentes no *dataset* do LCG-FMUC, o que traduz um sucesso de 98.05% na segmentação de cromossomas. Concluiu-se ainda que o modelo obtido através do YOLOv5 é capaz de individualizar cromossomas a partir de *clusters* complexos, seja em número (*clusters* de dois, três ou quatro cromossomas) ou em forma (*clusters* em forma "T", "+"ou "x"). Portanto, pode afirmar-se com confiança que foi proposta uma boa ferramenta de individualização de cromossomas a partir de microfotografias celulares da citogenética clínica.

6

Limitações e Trabalho Futuro

Apesar da contribuição desta dissertação para a investigação de metodologias capazes de automatizar a análise do cariótipo humano - principalmente a nível da segmentação de cromossomas -, o trabalho desenvolvido encontra algumas limitações.

Relativamente à análise dos modelos YOLOv5 (utilizados para a deteção de cromossomas), esta deveria ser mais extensa. Na metodologia desenvolvida, poderiam ainda ter sido utilizados diferentes parâmetros no treino dos modelos, como por exemplo o número de épocas. Como referido anteriormente, também poderia ser usada uma técnica para prevenir *overffiting*, tal como a técnica *Early Stopping*.

Quanto à utilização do algoritmo proposto, é manifestada a intenção de validação tanto do *dataset* sintético, como do modelo obtido, por técnicos da citogenética. Como trabalho futuro, os especialistas da área deveriam validar qualitativamente o fotorrealismo do *dataset* proposto, assim como a eficácia da segmentação de cromossomas do algoritmo YOLOv5. Contudo, sendo esta uma metodologia com alguns aspetos subjetivos, seria necessário consultar vários especialistas para se obter uma análise robusta e sem viés.

Uma vez que o objetivo final da segmentação de cromossomas é individualizar os cromossomas píxel a píxel, é sugerido, como trabalho futuro, o uso do *dataset* sintético proposto para o estudo da segmentação semântica de cromossomas. Através do modelo obtido com o YOLOv5, os cromossomas são detetados nas imagens celulares, mas as suas bboxes ainda contêm restícios de estruturas

celulares à sua volta. Sendo que o YOLOv5 consegue detetar, com alta fidelidade, os cromossomas nas microfotografias da citogenética convencional, poderia ser usada uma rede neuronal, como a UNet, para a segmentação semântica desses cromossomas detetados na microfotografia. A partir deste método, poderia ser alcançado um algoritmo *end-to-end* para a segmentação automática de cromossomas.

Bibliografia

[1] G. Carey, "Chapter 8: Chromosomes and Chromosomal Anomalies," em *Human Genetics for the Social Sciences*, Sage Publications, 1998, pp. 131–46.

[2] H. Huang e J. Chen, "Chromosome Bandings," em *Cancer Cytogenetics: Methods and Protocols*, T. S. Wan, ed., Springer, 2017, pp. 59–66, ISBN: 978-1-4939-6703-2.

[3] A. Bernheim, "Cytogenomics of cancers: From chromosome to sequence," *Molecular Oncology*, vol. 4, n.º 4, pp. 309–322, 2010.

[4] T. Arora e R. Dhir, "A review of metaphase chromosome image selection techniques for automatic karyotype generation," *Med Biol Eng Comput*, vol. 54, n.º 8, pp. 1147–1157, 2015.

[5] M. C. D. Matta, M. Dümpelmann, M. M. Lemos-Pinto, T. S. Fernandes e A. Amaral, "Processamento de imagens em Biodosimetria: Influência da qualidade das preparações cromossômicas," *Scientia Plena*, vol. 9, n.º 8, 2013.

[6] A. M. Subasinghe Arachchige, "Human metaphase chromosome analysis using image processing," tese de doutoramento, Electronic Thesis e Dissertation Repository, 2014, p. 2178.

[7] Genetix, *CytoVision® 7.0 The platform for every cytogenetics lab*, `https://www.well.ox.ac.uk/files-library/genetix-cytovision-brochure-2.pdf`, Acedido em: 10/03/2022, 2010.

[8] J. Graham e J. Piper, "Chapter 11: Automatic Karyotype Analysis," em *Chromosome Analysis Protocols*, Humana Press, 1994, pp. 141–185.

[9] X. Wang, B. Zheng, M. Wood, S. Li, W. Chen e H. Liu, "Development and evaluation of automated systems for detection and classification of banded

chromosomes: Current status and future perspectives," *Journal of Physics D: Applied Physics*, vol. 38, pp. 2536–2542, 2005.

[10] M. F. Andrade, L. V. Dias, V. Macario et al., "A study of deep learning approaches for classification and detection chromosomes in metaphase images," *Machine Vision and Applications*, vol. 31, n.º 7-8, 2020.

[11] C. Lin, A. Yin, Q. Wu et al., "Chromosome cluster identification framework based on geometric features and machine learning algorithms," *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020.

[12] R. Huang, C. Lin, A. Yin et al., "A Clinical Dataset and Various Baselines for Chromosome Instance Segmentation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, n.º 1, pp. 31–39, 2022.

[13] G. A. Levitsky, "Materielle Grundlagen der Vererbung," *Kiew: Staatsverlag*, 1924.

[14] J. Lejeune, M. Gautier e R. Turpin, "Étude des chromosomes somatiques de neuf enfants mongoliens," *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, vol. 248, n.º 11, pp. 1721–2, 1959.

[15] S. L. Gersen e M. B. Keagle, "Chapter 1: History of Clinical Cytogenetics," em *The Principles of Clinical cytogenetics*, Springer, 2013, pp. 3–8.

[16] P. Nowell e D. Hungerford, "A minute chromosome in human chronic granulocytic leukemia," *Science*, vol. 132, p. 1497, 1960.

[17] T. Caspersson, S. Farber, G. E. Foley et al., "Chemical differentiation along metaphase chromosomes," *Exp Cell Res*, vol. 49, n.º 1, pp. 219–2, 1968.

[18] M. E. Drets e M. W. Shaw, "Specific Banding Patterns of Human Chromosomes," *Proceedings of the National Academy of Sciences*, vol. 68, n.º 9, pp. 2073–7, 1971.

[19] F. Mitelman, "Catalogue of chromosome aberrations in cancer," *Cytogenet Cell Genet*, vol. 36, n.º 1-2, pp. 1–515, 1983.

[20] F. Mitelman, F. Mertens e B. Johansson, *Mitelman database chromosome aberrations and Gene Fusions in cancer*, `http://precog.iiitd.edu.in/people/anupama`, Acedido em: 17/06/2022, 2022.

[21] M. A. Ferguson-Smith, "History and evolution of Cytogenetics," *Molecular Cytogenetics*, vol. 8, n.º 19, 2015.

[22] G. Wang, H. Liu, X. Yi, J. Zhou e L. Zhang, "ARMS Net: Overlapping chromosome segmentation based on Adaptive Receptive field Multi-Scale network," *Biomedical Signal Processing and Control*, vol. 68, p. 102 811, 2021.

[23] S. K. Veerabhadrappa, P. R. Chandrappa, S. Y. Roodmal, S. J. Shetty, M. S. Gunjiganur S e M. K. Kumbar P, "Karyotyping: Current perspectives in diagnosis of chromosomal disorders," *Sifa Med J*, vol. 3, pp. 35–40, 2016.

[24] S. L. Gersen e M. B. Keagle, "Chapter 2: DNA, Chromosomes, and Cell Division," em *The Principles of Clinical cytogenetics*, Springer, 2013, pp. 9–21.

[25] L. Buckingham e M. L. Flaws, "Chapter 1: DNA," em *Molecular Diagnostics: Fundamentals, Methods, & Clinical Applications*, F.A. Davis, 2007, pp. 1–26.

[26] J. S. Heslop-Harrison, "Comparative Genome Organization in Plants: From Sequence and Markers to Chromatin and Chromosomes," *The Plant Cell*, vol. 12, n.º 5, pp. 617–5, 2000.

[27] M. Javan-Roshtkhari e S. K. Setarehdan, "A New Approach to Automatic Classification of the Curved Chromosomes," em *2007 5th International Symposium on Image and Signal Processing and Analysis*, 2007, pp. 19–24.

[28] A. Moncada e A. Pancrazzi, "Chapter Six - Lab tests for MPN," em *Cellular and Molecular Aspects of Myeloproliferative Neoplasms – Part B*, sér. International Review of Cell and Molecular Biology, N. Bartalucci e L. Galluzzi, eds., vol. 366, Academic Press, 2022, pp. 187–220.

[29] J.-U. Kang, "Overview of Cytogenetic Technologies," *The Korean Journal of Clinical Laboratory Science*, vol. 50, n.º 4, pp. 375–381, 2018.

[30] S. L. Gersen e M. B. Keagle, "Chapter 4: Basic Cytogenetics Laboratory Procedures," em *The Principles of Clinical cytogenetics*, Springer, 2013, pp. 53–66.

[31] T. P. Kannan e B. A. Zilfalil, "Cytogenetics: past, present and future," *The Malaysian journal of medical sciences : MJMS*, vol. 16, n.º 2, pp. 4–9, 2009.

[32] L. Buckingham e M. L. Flaws, "Chapter 8: Chromosomal Structure and Chromosomal Mutations," em *Molecular Diagnostics: Fundamentals, Methods, & Clinical Applications*, F.A. Davis, 2007, pp. 155–172.

[33] I. C. Yilmaz, J. Yang, E. Altinsoy e L. Zhou, "An improved segmentation for raw G-band Chromosome Images," em *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018.

[34] S. L. Gersen e M. B. Keagle, "Chapter 7: Instrumentation in the Cytogenetics Laboratory," em *The Principles of Clinical cytogenetics*, Springer, 2013, pp. 95–109.

[35] R. Bashmail, L. A. Elrefaei e W. Alhalabi, "Automatic segmentation of chromosome cells," *Advances in Intelligent Systems and Computing*, vol. 845, pp. 654–63, 2018.

[36] J. McGowan-Jordan, R. J. Hastings e S. Moore, *ISCN 2020: An International System for Human Cytogenomic Nomenclature (2020)*. Karger, 2020.

[37] R. Remani Sathyan, G. Chandrasekhara Menon, H. S, R. Thampi e J. H. Duraisamy, "Traditional and deep-based techniques for end-to-end automated karyotyping: A Review," *Expert Systems*, vol. 39, n.º 3, 2022.

[38] R. Uttamatanin, P. Yuvapoositanon, A. Intarapanich et al., "Metasel: A metaphase selection tool using a Gaussian-based classification technique," *BMC Bioinformatics*, vol. 14, n.º S13, 2013.

[39] E. Altinsoy, J. Yang e C. Yilmaz, "Fully-automatic raw G-band chromosome image segmentation," *IET Image Processing*, vol. 14, n.º 9, pp. 1920–1928, 2020.

[40] Y. Hernández-Mier, M. A. Nuño-Maganda, S. Polanco-Martagón e M. d. R. García-Chávez, "Machine Learning Classifiers Evaluation for Automatic Karyogram Generation from G-Banded Metaphase Images," *Applied Sciences*, vol. 10, n.º 8, 2020.

[41] L. Mei, Y. Yu, Y. Weng et al., "Adversarial Multiscale Feature Learning for Overlapping Chromosome Segmentation," *CoRR*, vol. abs/2012.11847, 2020.

[42] L. Ji, "Fully automatic chromosome segmentation," *Cytometry*, vol. 17, n.º 3, pp. 196–208, 1994.

[43] P. S. Karvelis, A. T. Tzallas, D. I. Fotiadis e I. Georgiou, "A multichannel watershed-based segmentation method for multispectral chromosome classification," *IEEE Transactions on Medical Imaging*, vol. 27, n.º 5, pp. 697–708, 2008.

[44] S. Minaee, M. Fotouhi e B. H. Khalaj, "A geometric approach to fully automatic chromosome segmentation," em *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, IEEE, 2014, pp. 1–6.

[45] T. Tanvi e R. Dhir, "An Efficient Segmentation Method for Overlapping Chromosome Images," *International Journal of Computer Applications*, vol. 95, n.º 1, pp. 29–32, 2014.

[46] M. V. Munot, J. Mukherjee e M. Joshi, "A novel approach for efficient extrication of overlapping chromosomes in automated karyotyping," *Medical & biological engineering & computing*, vol. 51, n.º 12, pp. 1325–1338, 2013.

[47] H. A. Al-Ameri e W. Al-Hameed, "New algorithm for separation overlapping & touching chromosomes," em *Journal of Physics: Conference Series*, IOP Publishing, vol. 1530, 2020, p. 012 024.

[48] S. Saiyod e P. Wayalun, "A new technique for edge detection of chromosome g-band images for segmentation," em *Advanced Approaches to Intelligent Information and Database Systems*, Springer, 2014, pp. 315–323.

[49] M. F. S. Andrade, F. R. Cordeiro, V. Macário, F. F. Lima, S. F. Hwang e J. C. G. Mendonça, "A Fuzzy-Adaptive Approach to Segment Metaphase Chromosome Images," em *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, 2018, pp. 290–95.

[50] A. Subasinghe, J. Samarabandu, Y. Li et al., "Centromere detection of human metaphase chromosome images using a candidate based method," *F1000Research*, vol. 5, p. 1565, 2016.

[51] F. Altınordu, L. Peruzzi, Y. Yu e X. He, "A tool for the analysis of chromosomes: KaryoType," *Taxon*, vol. 65, n.º 3, pp. 586–592, 2016.

[52] Y. Wu, Y. Yue, X. Tan, W. Wang e T. Lu, "End-To-End Chromosome Karyotyping with Data Augmentation Using GAN," em *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2456–60.

[53] N. Xie, X. Li, K. Li, Y. Yang e H. T. Shen, "Statistical Karyotype Analysis Using CNN and Geometric Optimization," *IEEE Access*, vol. 7, pp. 179 445–53, 2019.

[54] T. Lin, M. Maire, S. J. Belongie et al., "Microsoft COCO: Common Objects in Context," *CoRR*, vol. abs/1405.0312, 2014.

[55] Y. Z. Tao Feng Bin Chen, "Chromosome image segmentation framework based on improved Mask R-CNN," *Journal of Computer Applications*, vol. 40, n.º 11, pp. 3332–39, 2020.

[56] P. Chen, J. Cai e L. Yang, "Chromosome segmentation via data simulation and shape learning," em *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2020, pp. 1637–40.

[57] R. L. Hu, J. Karnowski, R. Fadely e J. Pommier, "Image Segmentation to Distinguish Between Overlapping Human Chromosomes," *CoRR*, vol. abs/1712.07639, 2017.

[58] H. M. Saleh, N. H. Saad e N. A. M. Isa, "Overlapping chromosome segmentation using u-net: Convolutional networks with test time augmentation," *Procedia Computer Science*, vol. 159, pp. 524–533, 2019.

[59] S. Song, T. Bai, Y. Zhao et al., "A new convolutional neural network architecture for automatic segmentation of overlapping human chromosomes," *Neural Processing Letters*, vol. 54, n.º 1, pp. 285–301, 2022.

[60] C. Lin, G. Zhao, A. Yin et al., "A novel chromosome cluster types identification method using ResNeXt WSL model," *Medical Image Analysis*, vol. 69, p. 101 943, 2021.

[61] H. Bai, T. Zhang, C. Lu, W. Chen, F. Xu e Z.-B. Han, "Chromosome extraction based on U-Net and YOLOv3," *IEEE Access*, vol. 8, pp. 178 563–69, 2020.

[62] M. Sharma, O. Saha, A. Sriraman, R. Hebbalaguppe, L. Vig e S. Karande, "Crowdsourcing for chromosome segmentation and deep classification," em *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 34–41.

[63] X. Cao, F. Lan, C.-M. Liu, T.-W. Lam e R. Luo, "ChromSeg: two-stage framework for overlapping chromosome segmentation and reconstruction," em *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2020, pp. 2335–42.

[64] K. Huang, C. Lin, R. Huang et al., "A novel chromosome instance segmentation method based on geometry and Deep Learning," em *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.

[65] S. Mittal e S. Vaishay, "A survey of techniques for optimizing deep learning on GPUs," *Journal of Systems Architecture*, vol. 99, p. 101 635, 2019.

[66] B. C. Russell, A. Torralba, K. P. Murphy e W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, n.º 1, pp. 157–73, 2008.

[67] D. Dwibedi, I. Misra e M. Hebert, "Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection," em *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1310–19.

[68] W. ya Guo, X. fei Wang e X. zhi Xia, "Two-dimensional Otsu's thresholding segmentation method based on grid box filter," *Optik*, vol. 125, n.º 18, pp. 5234–40, 2014.

[69] J. Schloss, *Affine transformations*, `https://www.algorithm-archive.org/contents/affine_transformations/affine_transformations.html`, Acedido em: 04/07/2022.

[70] J. Ieamsaard, S. N. Charoensook e S. Yammen, "Deep learning-based face mask detection using yolov5," em *2021 9th International Electrical Engineering Congress (iEECON)*, IEEE, 2021, pp. 428–31.

[71] A. Bochkovskiy, C. Wang e H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *CoRR*, vol. abs/2004.10934, 2020.

[72] R. Padilla, S. L. Netto e E. A. Da Silva, "A survey on performance metrics for object-detection algorithms," em *2020 international conference on systems, signals and image processing (IWSSIP)*, IEEE, 2020, pp. 237–42.

[73] M. Kasper-Eulaers, N. Hahn, S. Berger, T. Sebulonsen, Ø. Myrland e P. E. Kummervold, "Short Communication: Detecting Heavy Goods Vehicles in Rest Areas in Winter Conditions Using YOLOv5," *Algorithms*, vol. 14, n.º 4, 2021.

Anexos

A

Comparação de Técnicas de Citogenética

As principais técnicas de citogenética, mencionadas nesta dissertação, são a Citogenética Convencional, a *Fluorescence in Situ Hybridization* (FISH), a Cariotipagem Espetral (SKY) e a Hibridização Genómica Comparativa em *array* (aCGH). Para se obter uma ideia geral das diferenças entre estas técnicas, é disponibilizada a tabela abaixo, adaptada de Kang *et al.* [29].

Técnica	Método	Caraterísticas	Aplicação	Vantagens
Citogenética Convencional	Cultura Celular	A coloração gera um padrão de bandas específico para cada cromossoma.	Deteção de alterações cromossómicas numéricas e estruturais.	Diagnóstico amplo de anormalidades ao nível cromossómico.
FISH	Cultura Celular e Hibridização *in situ*	Sondas de DNA são usadas para se ligarem a sequencias específicas de DNA nos cromossomas.	Deteção de todos os tipos de alterações cromossómicas equilibradas e não equilibradas.	É possível fazer-se citogenética ao nível da interfase (procedimentos mais simples).
SKY	Cultura Celular e Hibridização *in situ*	Uso de sondas específicas para colorir todos os cromossomas.	Deteção de rearranjos incluindo alterações cromossómicas complexas.	Rápida caracterização da eucromatina.
aCGH	Técnica Molecular	Hibridização comparativa entre o DNA a ser testado e o DNA referencia.	Deteção e analise de alterações submicroscópicas no DNA.	Deteção de alta resolução e de alta especificidade, permitindo detalhar informação submicroscópicas.

B

Comparação de Algoritmos de Segmentação de Cromossomas

Tendo em conta a quantidade de algoritmos encontrados na literatura, por forma a condensar informação, elaborou-se uma tabela com os principais estudos relativos a segmentação de cromossomas. Assim, tornou-se mais simples perceber as falhas de metodologias anteriormente propostas. A tabela contem os seguintes parâmetros:

- Autor.
- Ano da publicação.
- Tipo de métodos usados (heurísticos, de aprendizagem ou híbridos).
- Objetivo do algoritmo proposto. Pode realizar segmentação da imagem celular, segmentação de *clusters* ou a classificação de objetos em classes (*clusters* e cromossomas individualizados).
- Tipo de imagem. Este parâmetro permite que o leitor perceba se o algoritmo foi feito para imagens de células em metafase ou se foram utilizadas imagens simplistas, como por exemplo imagens que apenas contêm 1 *cluster* com dois cromossomas sobrepostos. Além disso é indicado o padrão de bandas usado para obter as imagens.
- Nível de ruído das imagens usadas. É referido se as imagens usadas têm ruído, como restos de pigmentação ou outros objetos como núcleos interfásicos.
- Informações relativas à privacidade do *dataset* e a sua origem.

- Número de imagens de células em metafase e cromossomas usados no estudo.
- Existência de *clusters* nas imagens.
- Principal métrica de validação do algoritmo.
- Tempo de execução do algoritmo por imagem.

Autores	**Ano da publicação**	**Tipo de Metodologia**	**Objetivo do algoritmo**	**Tipo de Imagem (Padrão de Bandas)**	**Nível de Ruído**	***Dataset* Público (Origem)**	**Número de Imagens (cromossomas)**	**Existência de *clusters* (Número de clusters)**	**Métrica para o grupo teste [%]**	**Tempo de execução [s]**
Karvelis *et al.* [43]	2008	Heurístico	Segmentação da imagem celular	Celular (DAPI)	Imagens reais extraídas do laboratório	Não (-)	200 (-)	Sim (-)	Exatidão = 82.4	-
Uttamatanin *et al.* [38]	2013	Heurístico	Classificação de objetos em classes	Celular (G)	Imagens reais extraídas do laboratório	Não (Instituto Rajanukul, Bangkok)	192 (-)	Sim (-)	Exatidão = [89-99]	0.185
Munot *et al.* [46]	2013	Heurístico	Segmentação de *clusters*	1 cluster de 2 cromossomas (Q)	Imagens sem ruído	Sim (Bioimlab Dataset)	60 (120)	Sim (60)	Exatidão = [75-100]	-
Tanvi *et al.* [45]	2014	Heurístico	Segmentação de *clusters*	1 cluster de 2 cromossomas (Q)	Imagens sem ruído	Sim (Bioimlab Dataset)	162 (342)	Sim (162)	Exatidão = 87.4	-
Minaee *et al.* [44]	2014	Heurístico	Segmentação da imagem celular	-	-	Não (-)	25 (1150)	Sim (62)	Exatidão = 91.9	-
Hu *et al.* [56]	2017	Aprendizagem	Segmentação de *clusters*	1 cluster de 2 cromossomas (DAPI)	Imagens sem ruído	Sim (*DeepFish Dataset*)	13434 (26868)	Sim (13434)	IoU = 94.7	-
Wu *et al.* [52]	2018	Heurístico	Segmentação da imagem celular	Celular (G)	Imagens reais extraídas do laboratório	Não (-)	120 (5474)	Sim (-)	Exatidão = 95.9	-
Bashmail *et al.* [36]	2018	Heurístico	Segmentação da imagem celular	Celular (G)	Imagens reais extraídas do laboratório	Sim (Universidade King Abdulaziz)	130 (6011)	Não	Exatidão = 98.8	-
Xie *et al.* [53]	2019	Aprendizagem	Segmentação da imagem celular	Celular (G)	Imagens com ruído sintetizadas a partir de imagens reais do laboratório	Não (-)	100 (5000)	Sim (-)	AP_{50} = 95.644	-
Saleh *et al.* [57]	2019	Aprendizagem	Segmentação de *clusters*	1 cluster de 2 cromossomas (DAPI)	Imagens sem ruído	Sim (*DeepFish Dataset*)	13434 (26868)	Sim (13434)	IoU = 99.68	-
Altinsoy *et al.* [39]	2020	Heurístico	Segmentação da imagem celular	Celular (G)	Imagens reais extraídas do laboratório	Não (Central South University and Diagens-Hangzhou)	508 (23374)	Sim (5039)	Exatidão = [95-99]	[2-7]
Andrade *et al.* [10]	2020	Aprendizagem	Classificação de objetos em *clusters*	Celular (Giemsa Homogéneo)	Imagens com ruído sintetizadas a partir de imagens reais do laboratório	Sim (RCN-NE Chromosome Dataset)	74 (-)	Sim (-)	Exatidão = 93.19	-
Chen *et al.* [56]	2020	Aprendizagem	Segmentação de *clusters*	1 cluster de 2 cromossomas (Q)	Imagens sem ruído	Sim (Bioimlab Dataset)	117 ()	Sim (26)	Exatidão = 96.2	-

Bai *et al.* [61]	2020	Aprendizagem	Segmentação da imagem celular	Celular (G)	Imagens reais extraídas do laboratório	Não (-)	1300 (27600)	Sim (-)	Exatidão = 99.3	-
Cao *et al.* [63]	2020	Híbrido	Segmentação de *clusters*	1 cluster composto por vários cromossomas (G)	Imagens sem ruído	Sim (Universidade de Hong Kong)	345 (-)	Sim (345)	Exatidão = 90.5	-
Wang *et al.* [22]	2021	Aprendizagem	Segmentação de *clusters*	1 cluster de 2 cromossomas (DAPI)	Imagens sem ruído	Sim (*DeepFish Dataset*)	13434 (26868)	Sim (13434)	Exatidão = 99.99	-
Mei *et al.* [41]	2021	Aprendizagem	Segmentação de *clusters*	1 cluster de 2 cromossomas (DAPI)	Imagens sem ruído	Sim (*DeepFish Dataset*)	13434 (26868)	Sim (13434)	Exatidão = 99.977	0.568
Song *et al.* [59]	2021	Aprendizagem	Segmentação de *clusters*	1 cluster de 2 cromossomas (DAPI)	Imagens sem ruído	Sim (*DeepFish Dataset*)	13434 (26868)	Sim (13434)	F1 score = 95.96	-
Huang *et al.* [64]	2021	Híbrido	Segmentação da imagem celular	Celular (G)	Imagens reais extraídas do laboratório	Sim (ChromSeg Dataset)	162 (7452)	Sim (-)	Exatidão = 97.5	-
Lin *et al.* [60]	2021	Aprendizagem	Classificação de objetos em *clusters*	Celular (G)	Imagens reais extraídas do laboratório	Sim (Chromosome Cluster Identification Dataset)	500 (6592)	Sim (4880)	Exatidão = 94.09	-

C

Comparação de *Datasets* Públicos para Segmentação de Cromossomas

Os *datasets* usados na literatura estudada são importantes para reprodução de resultados e também para testar novos algoritmos de segmentação de cromossomas. Um grande problema, mencionado nesta dissertação, é a falta de *datasets* públicos e fidedignos, na medida em que contenham imagens reais anotadas ou então imagens sintéticas semelhantes às da citogenética clínica. Este anexo pretende resumir os *datasets* mencionados como públicos na literatura analisada. Contudo, como mostra a seguinte tabela, de oito *datasets*, apenas cinco foram encontrados *online*. Esta tabela também permite verificar a falta de complexidade das imagens usadas.

Nome do Dataset	**Acessibilidade**	**Tipo de Imagens**	**Número de Imagens (Número de Cromossomas)**	**Existência de *Clusters* (Número de *Clusters*)**
DeepFish Dataset	Sim [a]	Imagens sem ruído sintetizadas a partir de cromossomas reais. Cada imagem contém 1 *cluster* composto por 2 cromossomas. A coloração usada é DAPI.	13434 (26868)	Sim (13434)
CRCN-NE Chromosomes Dataset	Sim [b]	Imagens reais do laboratório. Cada imagem contém cromossomas, núcleos interfásicos e ruído. Coloração homogéneapor Giemsa.	74 (-)	Sim (-)
Bioimlab Dataset	Sim [c]	Imagens reais do laboratório. Cada imagem contém cromossomas, núcleos interfásicos e ruído. Bandagem Q.	162 (-)	Sim (-)
Overlapping Chromosome Instance Segmentation Dataset	Sim [d]	Imagens sem ruído. Cada imagem contém 1 *cluster* composto por vários cromossomas. Padrão de Bandas G.	29108 (80494)	Sim (29108)
Chromosome *Cluster* Identification Dataset	Sim [e]	Imagens reais do laboratório. Cada imagem contém cromossomas individualizados ou clusters de 1, 2 ou 3 cromossomas. Bandagem G.	500 (6592)	Sim (4880)
DGMU Dataset - Universidade King Abdulaziz	Não	Imagens reais extraídas do laboratório. Padrão de bandas G.	130 (6011)	Não
ChromSeg Dataset	Não	Imagens sem ruído. Cada imagem contém 1 *cluster* composto por vários cromossomas. Bandagem G.	345 (-)	Sim (345)
Copenhagen e Saravejo Dataset	Não	Imagens de cromossomas individualizados. Padrão de bandas G.	- (160)	Não

[a] `https://github.com/jeanpat/DeepFISH`.

[b] `https://zenodo.org/record/3229434#.YxqQkKHMK3B`.

[c] `http://bioimlab.dei.unipd.it/Chromosome%20Data%20Set%204Seg.htm`.

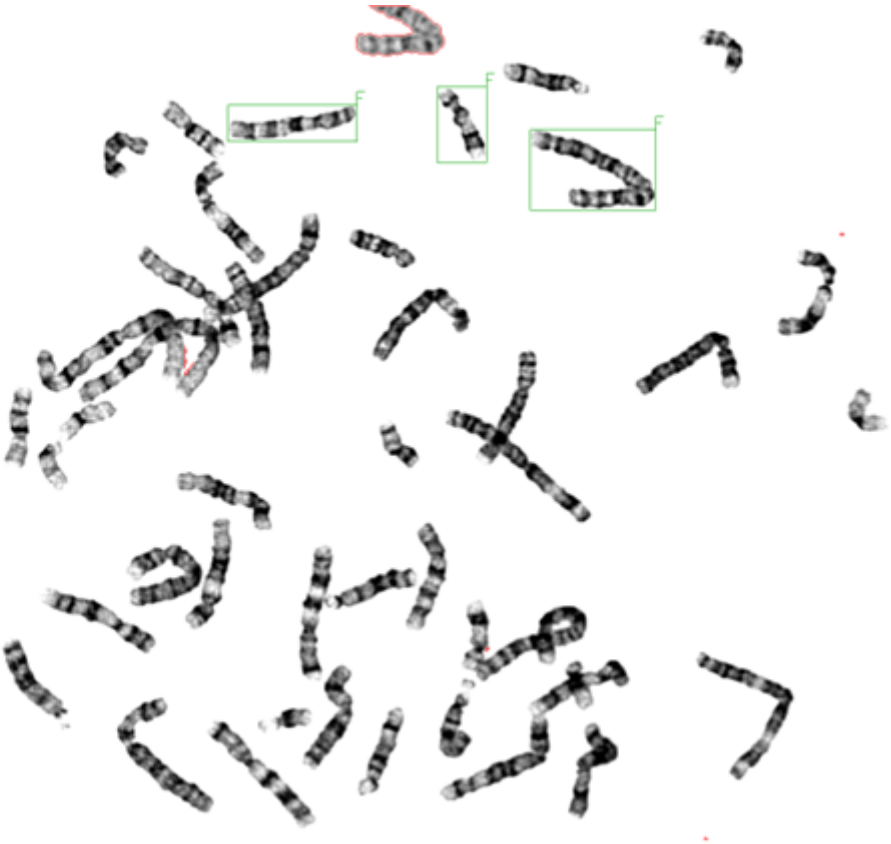
[d] `https://github.com/CloudDataLab/OverlappingChromosomeInstanceSegmentation`.

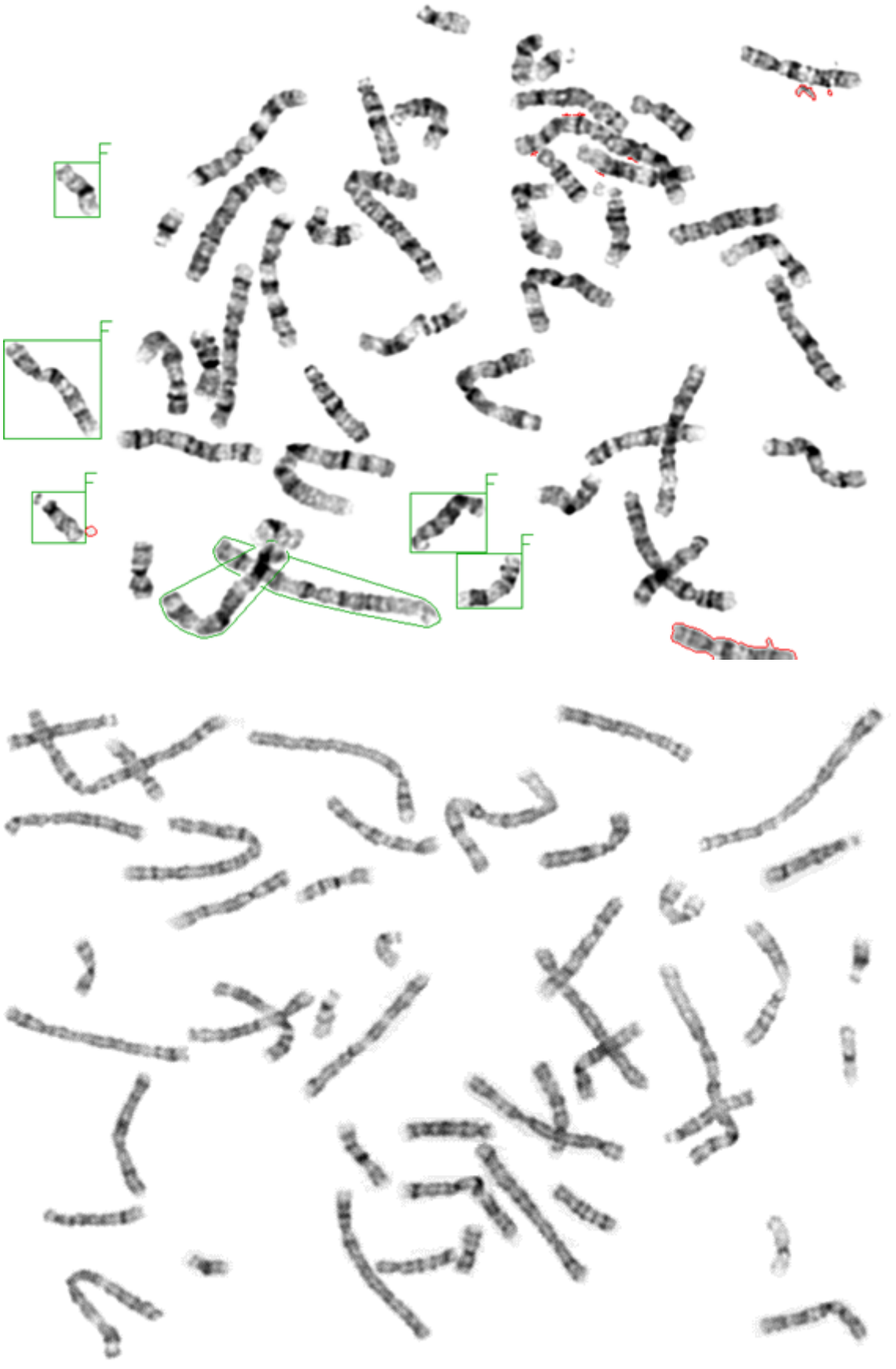
[e] `https://github.com/ChengchuangLin/ChromosomeClusterIdentification`.

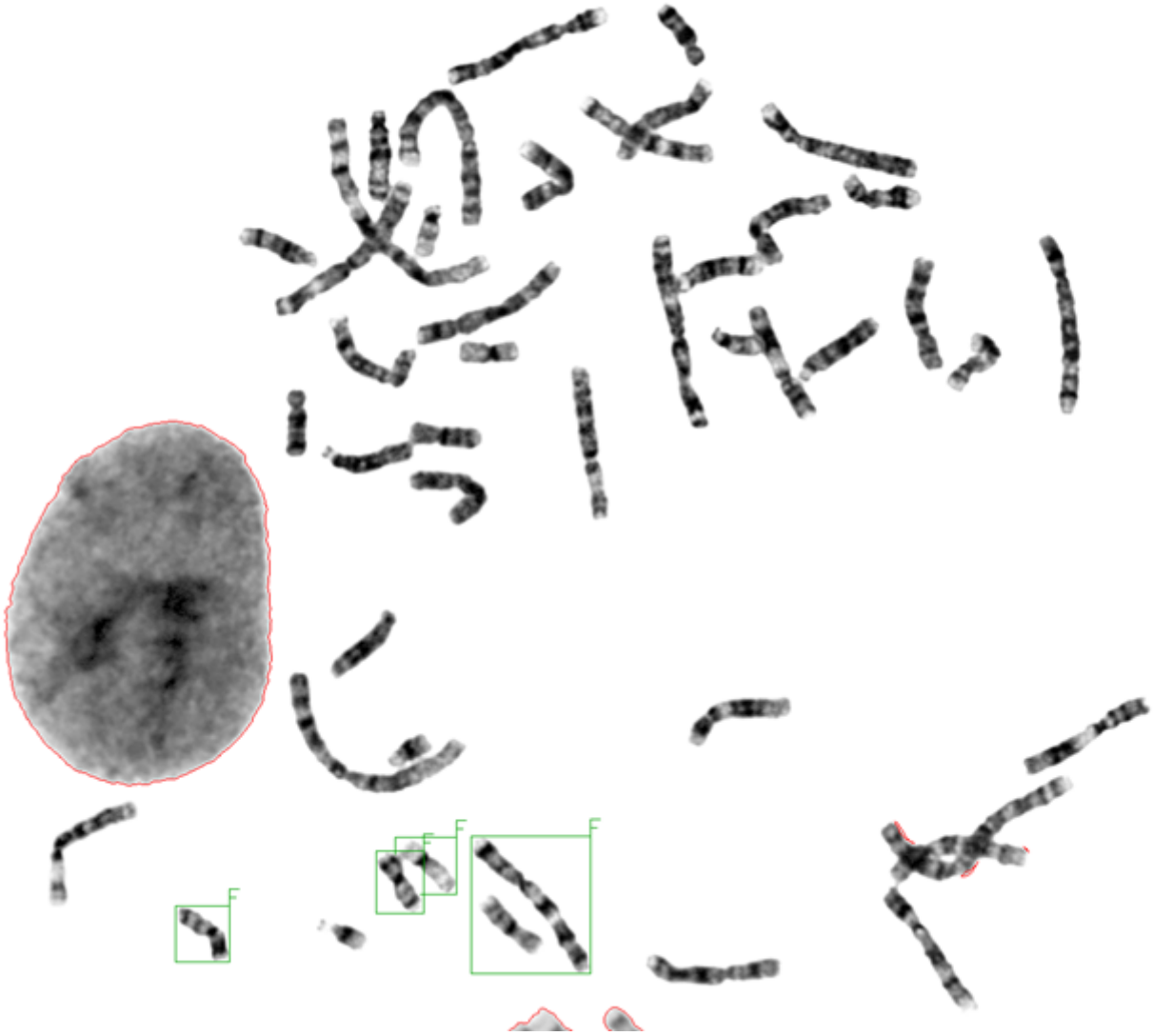
D

Dataset do LCG-FMUC

Neste anexo encontram-se exemplos de imagens fornecidas pelo Laboratório de Citogenética e Genómica da FMUC (LCG-FMUC). O objetivo deste anexo é mostrar a variedade morfológica do *dataset*.








1
2
3
4
5
9
10
11
12
15
16
17
18
19
20
21
22
X

E

Metodologia para a Aquisição de *Labels*

Neste anexo está sistematizado o processo de aquisição de estruturas celulares (cromossomas, núcleos interfásicos e objetos ruidosos) a partir do *dataset* do LCG-FMUC.

1. Aquisição de *labels* no *software* LabelMe.

   (a) Utilização imagens resultantes do pré-processamento da Secção 3.2.1.

   (b) Marcação dos vértices da bbox por ordem: primeiro, o canto superior esquerdo; segundo, o canto inferior direito.

   (c) Extrair estruturas celulares bem individualizadas e com os limites da bbox próximos das extremidades dessas estruturas. Não fazer a *label* de cromossomas que resultem de recortes de *clusters*, assim como os seus cromossomas homólogos.

   (d) O grupo de identificação das estruturas é a seguinte: 1 para cromossomas; 2 para núcleos interfásicos; 3 para objetos ruidosos.

   (e) Na nomenclatura da *label* de cada cromossoma são referidos a classe cromossómica, a posição do cromossoma respetivamente ao seu homólogo no cariograma e o nome do ficheiro do cariograma.

      - Exemplo: *"chr_NUMBER_DIRECTION_FILENAME"*, onde o *NUMBER* é composto por dois dígitos (01, 02, ..., 23, 24) e a

DIRECTION é descrita por *l* ou *r* (*left* ou *right*, respetivamente).

(f) Na nomenclatura da *label* de cada nucléolo são referidos o número por ordem de *labelling* na imagem celular, a posição do nucléolo na imagem celular e o nome do ficheiro da imagem celular.

- Exemplo: *"nucleolus_NUMBER_POSITION_FILENAME"*, onde o a *POSITION* pode ser: *ltc* (*left top corner*), *rtc* (*right top corner*), *ldc* (*left down corner*), *rdc* (*right down corner*), *l* (*left margin*), *u* (*up margin*), *d* (*down margin*), *r* (*right margin*) e *m* (*middle*).

(g) Na nomenclatura da *label* de cada objeto ruidoso são referidos o número por ordem de *labelling* na imagem celular e o nome do ficheiro da imagem celular.

- Exemplo: *"obj_NUMBER_FILENAME"*.

2. Criação de *datasets* a partir do ficheiro JSON gerado anteriormente: usar o comando *labelme_json_to_dataset "label_filename" -o "dataset_filename"* .

   (a) *label_filename* – corresponde ao nome do ficheiro JSON

   (b) *dataset_filename* - coresponde ao nome da imagem TIF

   (c) Colocar o ficheiro JSON e a imagem TIF na diretoria do *dataset* criado.

3. Repetição dos passos 1 e 2 para todas as imagens a serem rotuladas.

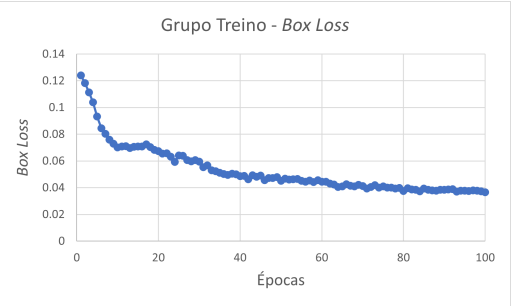
4. Utilização do *script "extract_from_labelme.py"* para recortar as *labels* criadas.

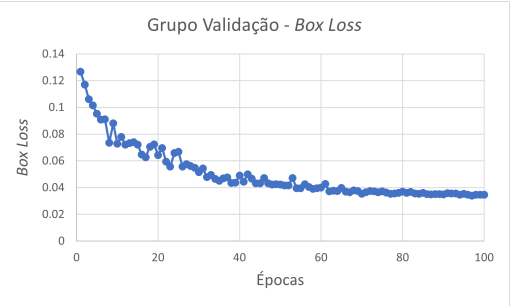
F

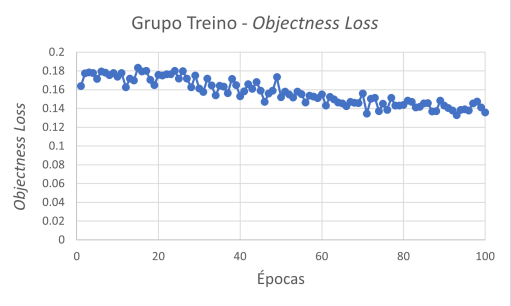
Avaliação do Desempenho dos Modelos YOLOv5

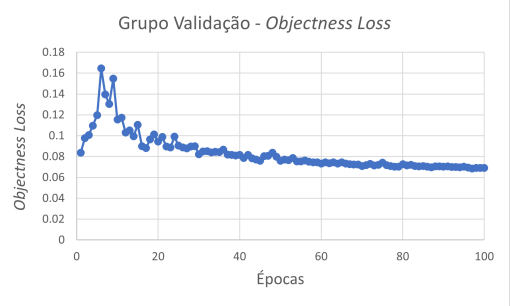
Neste anexo estão apresentados os gráficos das métricas de desempenho *Box Loss*, *Objectness Loss* para os grupos treino e validação, em função do número de épocas usadas. Além disso, para o grupo validação também são apresentados os gráficos da mAP@0.5 e mAP@0.5 : 0.05 : 0.95 em função do número de épocas. Assim, são apresentados seis gráficos para cada um dos cinco modelos treinados:

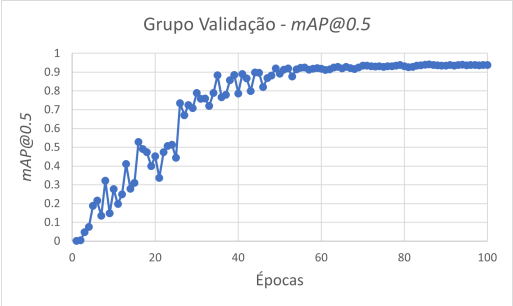
1. Modelo YOLOv5s para o *dataset* de 200 imagens.

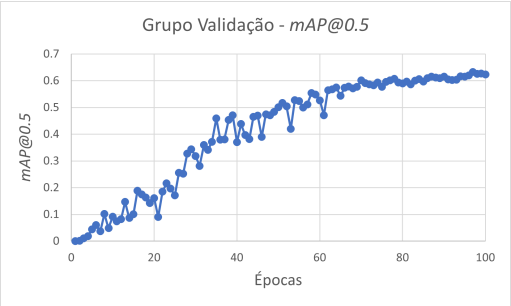




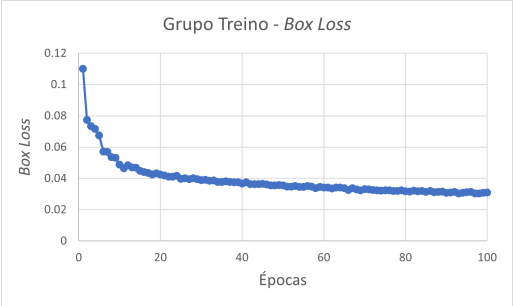


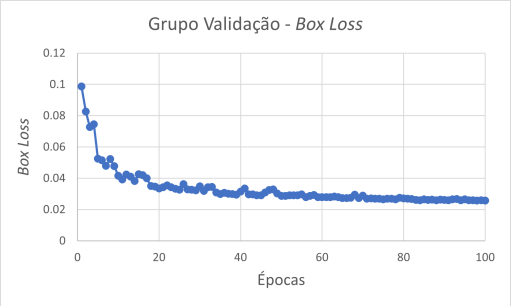


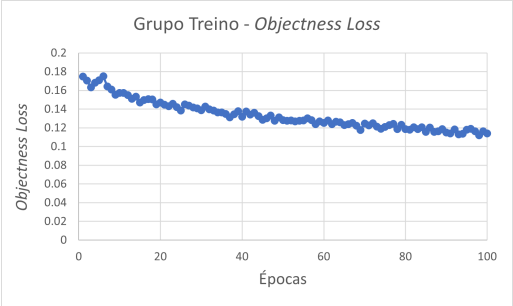


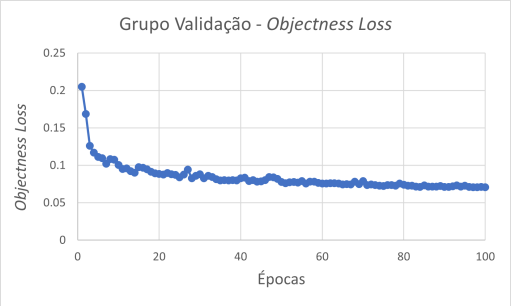


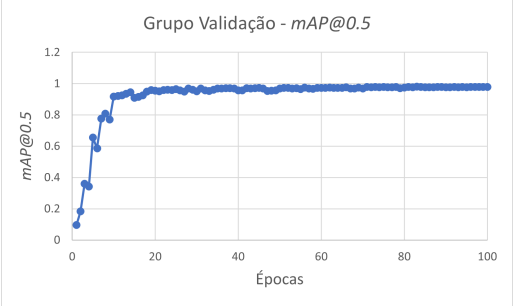
2. Modelo YOLOv5s para o *dataset* de 1000 imagens.

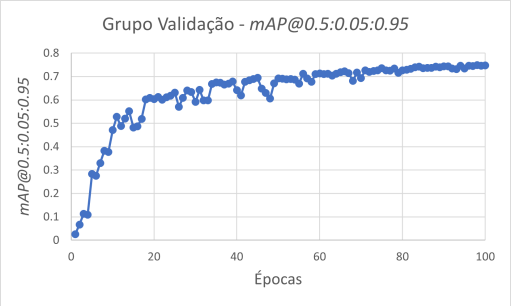




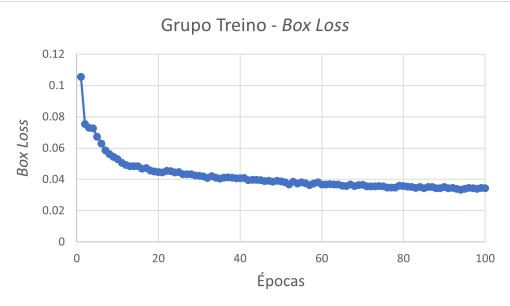


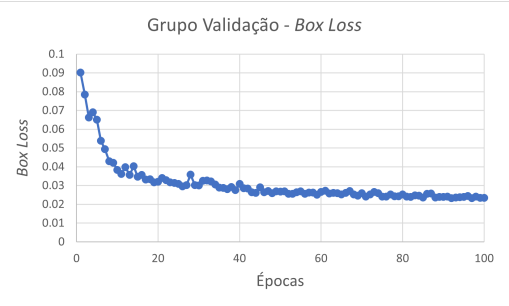


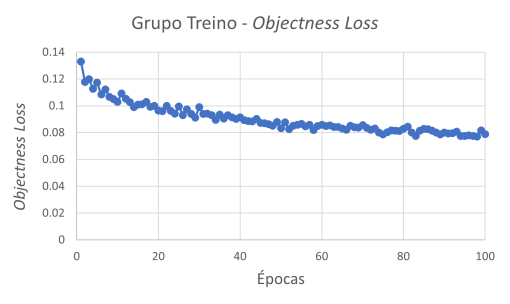


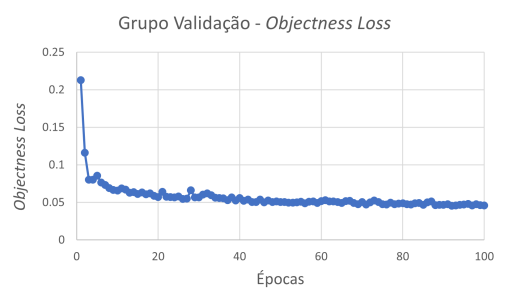


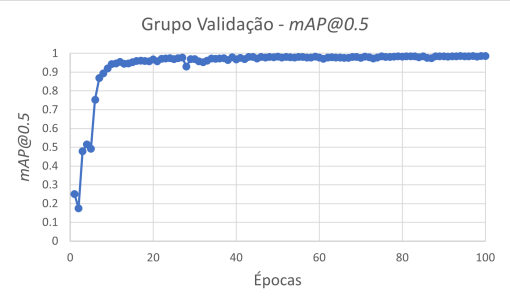
3. Modelo YOLOv5m para o *dataset* de 1000 imagens.

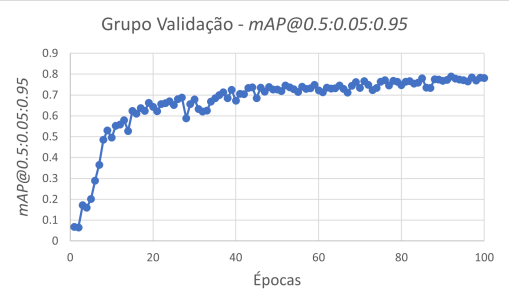




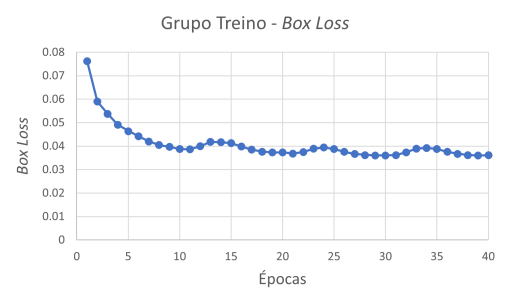


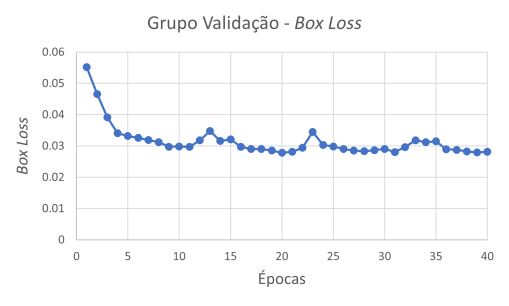


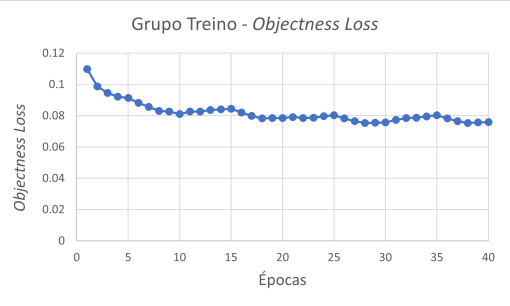


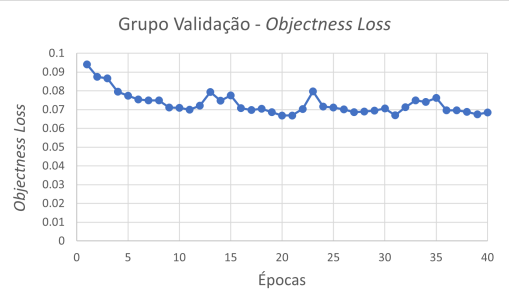


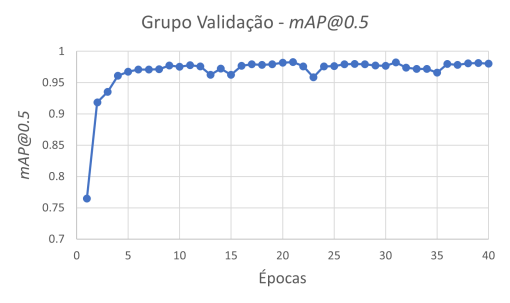
4. Modelo YOLOv5l para o *dataset* de 5000 imagens.

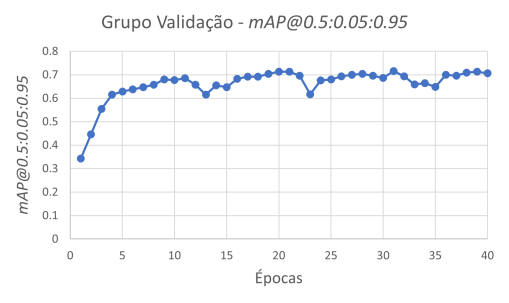




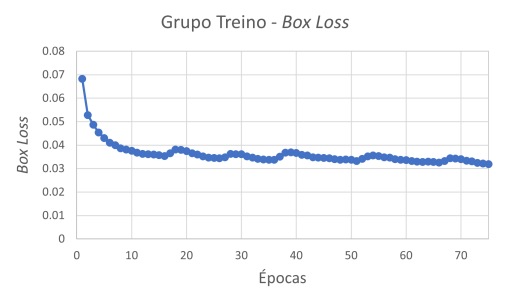


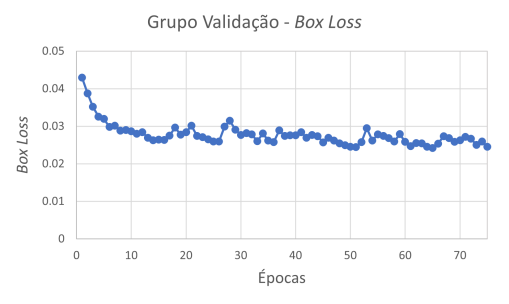


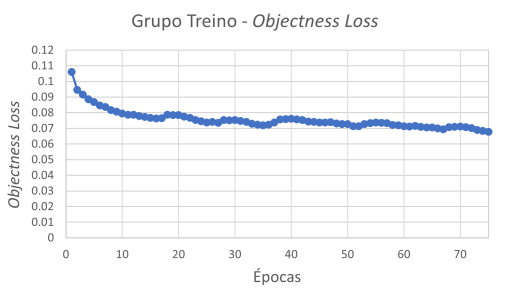


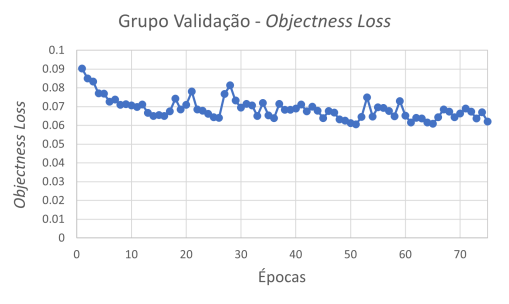


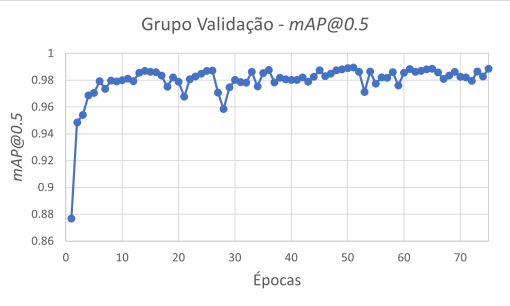
5. Modelo YOLOv5l para o *dataset* de 10795 imagens.

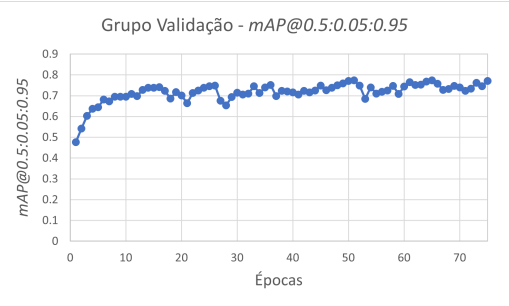









Grupo Validação - mAP@0.5
mAP@0.5
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0
10
20
30
40
50
60
70
Épocas


Grupo Validação - mAP@0.5:0.05:0.95
mAP@0.5:0.05:0.95
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
Épocas