Diagnóstico de Leucemia Linfóide Auxiliado por Computador

DANIELA MAYUMI USHIZIMA †

Orientador:

PROF.DR. LUCIANO DA FONTOURA COSTA

Tese apresentada ao Instituto de Física de São Carlos da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Doutor em Ciências, sub-área: Física Computacional.

USP/IFSC/SBI

8-2-001631

USP - São Carlos
Outubro de 2004

†Este trabalho contou com apoio financeiro da FAPESP.

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

Ushizima, Daniela Mayumi
Diagnóstico de leucemia linfóide auxiliado por computador. São Carlos, 2004.
189 p.

Tese (Doutorado) Instituto de Física de São Carlos, 2004.

Orientador: Prof. Dr. Luciano da Fontoura Costa

1. Visão computacional. 2. Reconhecimento de padrões. 3. Mineração de dados. I. Título.

IFSC-USP SERVIÇO DE BIBLIOTECA

UNIVERSIDADE
DE SÃO PAULO
Instituto de Física de São Carlos

Av. Trabalhador São-carlense, 400
Caixa Postal 369
13560-970 - São Carlos - SP - Brasil

MEMBROS DA COMISSÃO JULGADORA DA TESE DE DOUTORADO DE DANIELA MAYUMI USHIZIMA APRESENTADA AO INSTITUTO DE FÍSICA DE SÃO CARLOS, UNIVERSIDADE DE SÃO PAULO, EM 06/10/2004.

COMISSÃO JULGADORA:

Prof. Dr. Luciano da Fontoura Costa (Orientador e Presidente) – IFSC/USP

Prof. Dr. Alexandre Xavier Falcão – UNICAMP

Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho – ICMC/USP

Prof. Dr. Roberto Marcondes César Júnior – IME/USP

Prof. Dr. Alberto Tannus – IFSC/USP

Dedico este trabalho
à menina dos meus olhos:
Iris.

Agradecimentos

À iluminação divina para que eu confie e siga no caminho do bem.

Ao Prof. Dr. Luciano da F. Costa pela orientação e incentivo ao desenvolvimento desse projeto.

À minha família, Ricardo e Iris, pelo amor e apoio, que me ajudam a conduzir o trabalho.

Aos meus amigos Andrea, Bárbara, Bruno, Consul, Edson, Erbe, Fátima, Fabbri, Jander, Julia, LeandroS, Luis, Marconi, Marquinho, Odemir, Regina, Rodrigo, Silvia, dentre outros, pela amizade solícita.

Aos meus pais, Antonio e Lourdes, pelo exemplo de dedicação e amor, como inspiração para toda a vida. Obrigada pela confiança e carinho.

Aos meus irmãos, Thiago, Thales e Larissa pela motivação que me dão de trabalhar para estarmos todos juntos, sempre, mais uma vez.

A todos os funcionários do IFSC, faxineiras, porteiros, bibliotecárias, secretárias, técnicos e professores pelo auxílio, carinho e respeito no dia a dia.

Conteúdo

Resumo xvii

Abstract xviii

1 Introdução 1

1.1 História do diagnóstico 1

1.2 Diagnóstico e cibernética 4

2 O sangue humano 7

2.1 Leucócitos - produção e função 9

2.1.1 Neutrófilos 10

2.1.2 Eosinófilos 11

2.1.3 Basófilos 12

2.1.4 Monócitos 12

2.1.5 Linfócitos 13

2.2 Linfoma e leucemia linfoproliferativa crônica 15

2.2.1 Leucemia linfóide crônica (LLC) 19

2.2.2 Leucemia prolinfocítica (LLP) 21

2.2.3 Leucemia de células cabeludas - LCC 23

2.2.4 Outras neoplasias 24

2.3 Métodos de coloração de esfregaços 26

3 Reconhecimento de Padrões 29

3.1 Processamento de imagens 30

3.1.1 Quantização 31

3.1.2 Filtragem 32

IFSC-USP SERVIÇO DE BIBLIOTECA
INFORMAÇÃO

3.2 Segmentação 36
3.2.1 Limiarização 37
3.2.2 Segmentação por Borda 45
3.2.3 Segmentação por cor 47
3.3 Extração de características 49
3.3.1 Forma 49
3.3.2 Cor 51
3.3.3 Textura 52
3.4 Aprendizado 56
3.5 Classificação 58
3.5.1 Densidade Normal 61
3.5.2 Teoria de decisão bayesiana 62
3.5.3 Estimação de parâmetros 65
3.5.4 Regra dos K-vizinhos mais próximos 66
3.6 Estimação de erro 67

4 Seleção de características **71**
4.1 Métodos de busca 72
4.1.1 Exaustivo 72
4.1.2 FSS - Forward Sequential Selection 73
4.1.3 BSS - Backward Sequential Selection 73
4.2 Abordagem *filter* 74
4.2.1 Euclidiana 74
4.2.2 Mahalanobis 75
4.2.3 Distância entre os centróides das classes 75
4.2.4 Distância média entre os elementos das classes 76
4.3 Coeficiente de correlação 76
4.4 Abordagem *wrapper* 76
4.5 Mineração de dados 77

5 Resultados **81**
5.1 Ferramentas propostas 82
5.1.1 LEUKO 82
5.1.2 Ferramenta para seleção de características 85

5.2 Resultados experimentais 87
5.3 *In vitro* para *in silico* 88
5.3.1 Micrografias 88
5.3.2 Pré-processamento 89
5.3.3 Segmentação 89
5.4 Diagnóstico através de medidas de forma 100
5.5 Estudo de hierarquia nos dados de leucócitos 100
5.6 Inclusão de medidas de textura 101
5.7 Inclusão de medidas de cor 102
5.8 Aplicação de seleção de atributos 106
5.9 Mineração de dados 107

6 Conclusão e trabalhos futuros **111**
6.1 Segmentação 112
6.2 Atributos da imagem 113
6.3 Classificadores 114
6.4 Seleção de características 115
6.5 Trabalhos futuros 116

A Imagens de leucócitos coletadas **119**

B Métodos de limiarização **127**

C Outros processamentos sobre os leucócitos **129**
C.1 Retinex 129
C.2 Difusão anisotrópica 131
C.3 Segmentação baseada em evolução de curvas 135

D Manual dos programas **145**
D.1 Leuko 145
D.2 Seleção de características 147
D.3 Weka 148
D.3.1 Naive Bayes 152
D.3.2 Árvore de decisão 153
D.3.3 Rede neural 154

D.3.4 K-médias 155
D.3.5 Naive Bayes 157
D.3.6 Árvore de decisão 158
D.3.7 Rede neural 159
D.3.8 K-médias 160

E Artigos publicados sobre este projeto de doutorado **161**

Bibliografia **163**

Lista de Figuras

2.1 Desenho esquemático de fagócitos: **(a)** neutrófilo, **(b)** eosinófilo, **(c)** basófilo e **(d)** monócito (fonte: [1]). 9

2.2 Desenho esquemático de linfócitos reativos [1] 16

2.3 Micrografias de linfócitos reativos; (a)linfócito normal; (b-d)linfócito atípico Downey tipo I; (e-g)linfócito atípico Downey tipo II; (h-j)linfócito atípico Downey tipo III [2] 17

2.4 Linfomas e subdivisões 19

2.5 Micrografias de **(a)** LLC, **(b)** prolinfócito, **(c)** *Hairy cell.* 20

2.6 Linfócito normal (esquerda) e célula cabeluda (direita). 23

2.7 Exemplos de células de Sézary 25

3.1 (a) Imagem colorida , (b) sua versão em nível de cinza e (c) imagem nível de cinza quantizada e equalizada segundo algoritmo de Haralick [3] 32

3.2 Segmentação do núcleo de leucócito usando limiar em histograma de intensidades. a) imagem em níveis de cinza; b) histograma da imagem; c) imagem limiarizada para valor do pixel menor que 135. . . 38

3.3 Limiarização entrópica: (a)imagem original em nível de cinza; (b) funções de entropia relativa; (c) e (d) resultados da limiarização usando limiar obtido por Kapur e Sahoo, respectivamente. 43

3.4 Limiarização pelo método LogCum: (a) imagem original em nível de cinza; (b) histograma normalizado de (a); (c) acumulada de (b); (d) logaritmo da acumulada apresentada em (c); (e) função critério f, resultado da multiplicação de (c) por (d); (f) resultado da limiarização. 45

3.5 Segmentação do núcleo de basófilo pelo LogCum usando informação do canal G; (c) acumulada de (b); (d) logaritmo da acumulada apresentada em (c); (e) função critério f, resultado da multiplicação de (c) por (d); (f) resultado da limiarização. 46
3.6 Bordas do leucócito da Fig.3.2 para diferentes técnicas - à esquerda: Sobel, ao centro: Prewitt e 'a direita: Canny - descontinuidades do contorno e ruído 47
3.7 Reconhecimento estatístico de padrões segundo Jain et al [4]. 59
3.8 Distribuição de probabilidade: paramétrica e não-paramétrica 60

4.1 Passos de um sistema de descoberta de conhecimento em base dados [5]. 78

5.1 Diagrama de blocos do programa *Leuko*: blocos cinzas representam processos que podem ser executados apenas uma vez. 83
5.2 Interface do programa *Leuko*: sistema para reconhecimento de leucócitos desenvolvido em Delphi 4.0. 84
5.3 Diagrama de blocos da ferramenta para seleção de característica. . . . 86
5.4 Ferramenta para seleção de característica de propósito geral desenvolvida em Delphi 4.0. 86
5.5 Histograma do canal G de imagens de leucócitos normais: valores menores que 100 indicam pixels do núcleo. 90
5.6 Imagens de vários tipos de leucócitos 93
5.7 Segmentação do núcleo segundo Katz: limiar fixo no histograma do canal G (T=100). 94
5.8 Segmentação do núcleo limiar adaptativo (busca no intervalo [100, 200]). 95
5.9 Limiarização dos prolinfócitos: (a)canal G do prolinfócito; (b)histograma do canal G, mostrando que o mínimo calculado pelo algoritmo adaptativo é 100 e não 150. 96
5.10 Área do núcleo (em pixels) de 938 leucócitos de 29 pacientes diferentes. 98
5.11 Exemplo de segmentação por cor: **(a)** imagem original, **(b)** imagem segmentada usando espaço de cor RGB, **(c)** imagem segmentada usando espaço de cor CIELab. 99

5.12 **(a)** Citoplasma do basófilo. **(b)** Citoplasma do eosinófilo. **(c)** Citoplasma do linfócito. **(d)** Citoplasma do monócito. **(e)** Citoplasma do neutrófilo. 103

5.13 **(a)** Núcleo do basófilo. **(b)** Núcleo do eosinófilo. **(c)** Núcleo do linfócito. **(d)** Núcleo do monócito. **(e)** Núcleo do neutrófilo. 104

A.1 **(a,c,e)** Micrografia do basófilo. **(b,d,f)** Basófilo segmentado. 120

A.2 **(a,c,e)** Micrografia do eosinófilo. **(b,d,f)** Eosinófilo segmentado. . . 121

A.3 **(a,c,e)** Micrografia do linfócito. **(b,d,f)** Linfócito segmentado. . . . 122

A.4 **(a,c,e)** Micrografia do monócito. **(b,d,f)** Monócito segmentado. . . 123

A.5 **(a,c,e)** Micrografia do neutrófilo. **(b,d,f)** Neutrófilo segmentado. . . 124

A.6 **(a,c,e)** Micrografia de LLC. **(b,d,f)** LLC segmentada. 125

B.1 Exemplos comparativos de limiarização utilizando imagens do Matlab. 127

B.2 Exemplos comparativos de limiarização para diferentes tipos de leucócitos: (a)linfócito; (b)neutrófilo; (c)monócito; (d)eosinófilo; (e)basófilo. . . . 128

C.1 Imagem de leucócito (acima) e resultado do retinex (abaixo) para 1 ciclo. 130

C.2 **(a,b,c)** Processamento iterativo usando retinex: a)4x, b)8x e c)128x. 130

C.3 Análise por PCC para comparar segmentação da imagem original e filtrada, considerando 16 imagens de leucócitos segmentadas manualmente: **(a)** núcleo e **(b)** citoplasma 132

C.4 **(a,c,e)** Micrografias. **(b,d,f)** Resultado após filtragem usando o método de difusão anisotrópica. 133

C.5 **(a,c,e)** Micrografias. **(b,d,f)** Resultado após filtragem usando o método de difusão anisotrópica. 134

C.6 **(a,c,e)** Micrografias de neutrófilos. **(b,d,f)** Segmentações por evolução de curva. 136

C.7 **(a,c,e)** Micrografias de linfócitos. **(b,d,f)** Segmentações por evolução de curva. 137

C.8 **(a,c,e)** Micrografias de monócitos. **(b,d,f)** Segmentações por evolução de curva. 138

C.9 **(a,c,e)** Micrografias de eosinófilos. **(b,d,f)** Segmentações por evolução de curva. 139
C.10 **(a,c,e)** Micrografias de basófilos. **(b,d,f)** Segmentações por evolução de curva. 140
C.11 **(a,c,e)** Micrografias de LLC. **(b,d,f)** Segmentações por evolução de curva. 141
C.12 **(a,c,e)** Micrografias de Hairy cells. **(b,d,f)** Segmentações por evolução de curva. 142
C.13 **(a,c,e)** Micrografias de prolinfócitos. **(b,d,f)** Segmentações por evolução de curva. 143

D.1 Interface de Leuko e suas funcionalidades 146

Lista de Tabelas

2.1 Leucócitos: Número total 4.000 - 11.000/mm^3 10
2.2 Características do Neutrófilo, segundo [1]. 11
2.3 Características do Eosinófilo maduro segundo [1] 11
2.4 Características do Basófilo maduro segundo [1] 12
2.5 Características do monócito maduro segundo [1] 13
2.6 Características do linfócito maduro segundo [1] 15
2.7 Características da LLC segundo [1] 22
2.8 Características da LLP segundo [1] 22
2.9 Características da LCC segundo [1] 24
2.10 Características da Sézary segundo [1] 26

3.1 Métodos para estimativa de erro 68

4.1 Estatística como componente principal na mineração de dados [6] . . 77

5.1 Pixels corretamente classificados: média, desvio padrão e mediana para algoritmos de limiarização no processamento de 20 imagens de leucócitos. 92
5.2 Tempo médio, desvio padrão, mediana e máximo de CPU em segundos para 20 imagens de leucócitos. 92
5.3 Matriz de confusão dos leucócitos normais utilizando apenas atributos de forma e avaliação de taxa de acerto através do método de resubstituição. 101
5.4 Parâmetros de textura para citoplasma: avaliação da escala mais apropriada e taxa de acerto usando método da resubstituição. 102
5.5 Atributos extraídos dos leucócitos. 105

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

5.6 Matriz de confusão dos atributos mais relevantes (Id segundo Tabela 5.5): 1,5,7,15,21,34,35,37,41,46,53,54 106
5.7 Matriz de confusão dos atributos mais relevantes (Id segundo Tabela 5.5): 1,5,7,15,21,34,35,37,41,46,53,54 107
5.8 Base de imagens de leucócitos: número de amostras para treinamento e teste do classificador. 107

D.1 Exemplo de uma matriz de confusão relacionando o número de amostras com o resultado de classificação, dadas as classes C1 e C2 150
D.2 Precisão detalhada por classe considerando as 10 características mais relevantes 150
D.3 Precisão detalhada por classe considerando 62 características. 152
D.4 Matriz de confusão. 152
D.5 Precisão detalhada por classe considerando 62 características. 153
D.6 Matriz de confusão 153
D.7 Precisão detalhada por classe considerando 62 características. 154
D.8 Matriz de confusão 154
D.9 Desempenho do k-médias para os diferentes agrupamentos 155
D.10 Classes para agrupamentos considerando 62 características. 155
D.11 Número do cluster e correspondente conforme código do leucócito . . 155
D.12 Precisão detalhada por classe considerando as 10 características mais relevantes 157
D.13 Matriz de confusão 157
D.14 Precisão detalhada por classe considerando as 10 características mais relevantes. 158
D.15 Matriz de confusão 158
D.16 Precisão detalhada por classe considerando as 10 características mais relevantes. 159
D.17 Matriz de confusão. 159
D.18 Desempenho do k-médias para os diferentes agrupamentos 160
D.19 Classes para agrupamentos considerando as 10 características mais relevantes. 160
D.20 Número do cluster e correspondente conforme código do leucócito . . 160

Resumo

O presente trabalho de doutorado visa estudar o diagnóstico de leucemias por meio de processamento das imagens de microscópio óptico de transmissão, em colaboração com médicos hematologistas do HC-FMRP-USP e sob supervisão do Prof. Dr. Marco Zago. Como nem todas as leucemias podem ser diagnosticadas por meio de parâmetros visuais, apenas os casos de leucemia linfóide serão considerados, uma vez que esses são casos onde as células podem ser classificadas visualmente com precisão. A análise citológica é feita por especialistas humanos, cotidianamente em casos de contagem do número de leucócitos e se limitam à avaliação de um número reduzido de amostras pois é uma tarefa repetitiva, minuciosa e especializada. Com a automação desse processo, há possibilidade de maior número de análises de imagens, com geração de informações estatísticas a respeito das células presentes em amostras de sangue. O reconhecimento automático da célula envolve três etapas básicas: a segmentação da imagem, a extração de características e a classificação. A técnica de reconhecimento de padrões adotada para segmentação das imagens de esfregaços de sangue utiliza aprendizagem supervisionada por cor no espaço RGB, gerando imagens binárias contendo as diferentes regiões de interesse: núcleo, citoplasma, fundo e hemácia. O usuário pode treinar o classificador para uma imagem de esfregaço de sangue periférico, segmentar, filtrar e processar várias medidas das ROIs, particularmente do núcleo e citoplasma, considerando parâmetros de forma, textura e cor. A contribuição desse projeto está na elaboração de programas de interface amigável tanto para reconhecimento de padrões quanto para seleção de característica e mineração de dados. O programa de reconhecimento de padrões é baseado em casos de leucócitos normais, de leucemia linfóide crônica, prolinfocítica e tricoleucemia. Para desenvolvimento do programa de reconhecimento de padrões foi necessária uma grande base de dados, que hoje conta com aproximadamente 1.439 imagens, onde cerca de 1.058 são de leucócitos normais e cerca de 381 de leucêmicos.

Abstract

The current PhD project deals with the leukemia diagnosis using optical microscope image processing, in collaboration with hematologists from HC-FMRP-USP and under supervision of Prof. Dr. Marco Zago. Only specific leukemia cases can be diagnosed through visual parameter, therefore only lymphoid leukemias have been considered since these are the cases in which the cells can be classified visually with accuracy. Expert humans have dedicated to leukocyte differential count daily, limiting the analysis to a reduced number of samples since this task is time-consuming and painstaking. The automation of this process would allow to analyze many more images with wide statistical information about each blood smear slide. The cell recognition can be divided in three steps: the image segmentation, the feature extraction and the classification. The pattern recognition technique to color segment the images utilizes supervised learning in the RGB color space, generating binary images containing the different regions of interest: nucleus, cytoplasm, background and red blood cells. The user can train the classifier to segment a particular image, to filter the resulted image, to process and to extract several cell attributes. Particularly, we are interested in the nucleus and the cytoplasm in terms of the shape, size, color and texture. We have developed softwares with user-friendly interface both to pattern recognition and feature selection/datamining. The pattern recognition system is dedicated to recognize normal leukocytes and lymphoid leukemias as chronic lymphocitic, prolymphocitic and Hairy cells. For developing the pattern recognition system, we collected a large database, which contains approximately 1,439 images nowadays, being around 1,058 normal leukocytes and around 381 abnormal lymphocytes.

Capítulo 1

Introdução

"Um pensamento sutil ainda que errado, pode gerar indagações proveitosas que possam estabelecer verdades de grande valor. (Isaac Asimov)"

1.1 História do diagnóstico

Nos Pireneus, França, encontrou-se a notícia mais antiga (provavelmente dezessete mil anos atrás) dos curandeiro-feiticeiros, os profissionais da saúde da antiguidade, que tinham o "poder" de afastar as moléstias; eram os responsáveis pela interpretação dos fatos "incompreensíveis" que levavam o indivíduo a doenças, ou seja, faziam o diagnóstico. Desde o Período Neolítico existiam cirurgiões hábeis, como mostram crânios trepanados e cicatrizados, conservados no Museu do Homem em Paris.

A base essencial da medicina surge com as primeiras civilizações históricas do vale do Nilo e do Eufrates, no início do terceiro milênio e, em seguida, as do Mediterrâneo Oriental, com a tradição dos Vedas, proveniente da Índia Sagrada, os tijolos de Nínive[1] , os papiros do Egito, as tradições bíblicas e a brilhante cultura das Ilhas Egéias e da Magna Grécia. Hamurabi, rei da Babilônia (XXII a.C.), elaborou uma minuciosa legislação, pela primeira vez, prevendo por lei, as responsabilidades do médico e do cirurgião (classe inferior que tinham seus atos controlados pelas leis duras do código de Hamurabi). Na China Antiga (1.000 a.C.), acreditava-se que o sangue continha a alma. Os reis egípcios se banhavam em sangue para cura de

[1]Nínive foi uma das capitais da Assíria durante o governo de Assurbanipal (668 a 626 a.C.), quem dedicou seu longo reinado a embelezar a cidade, nela erguendo uma biblioteca de tijolos. Muitos desses tijolos ainda existem e grande parte deles tratava de assuntos médicos.

elefantíase [7].

Hipócrates (460-370 a.C.), que afastou da medicina os constrangimentos da especulação filosófica e a superstição, introduzindo seus princípios éticos e científicos, distinguindo a medicina como uma ciência e profissão de grande dignidade. Foi Hipócrates[2] quem primeirou empregou a expressão carcinoma para designar tumores malignos, diagnosticando-o como um derrame de humores orgânicos. Houve quem sustentasse existir uma correlação entre o câncer e restos embrionários isolados. Sua terapêutica já contava com uso de medidas dietéticas, doses controladas de arsênico e até mesmo de cauterização.

Com a queda do Império Romano (séc.V), a religião surge como elemento agregador dos inúmeros reinos bárbaros formados após sucessivas invasões. A cultura greco-romana quase desaparece nos períodos mais turbulentos do feudalismo (séc.V a XV), mas permaneceu latente nos mosteiros. O pensamento medieval é marcado pela tentativa de conciliar razão e fé. A ciência medieval valoriza o conhecimento teórico em detrimento das atividades práticas. Ainda que no século XI tenham surgido as primeiras universidades, a especulação filosófica-teológica se desenvolve até o Renascimento. Os árabes foram uma exceção na tradição científica medieval pois em seu movimento expansionista, conheceram a cultura grega e iniciaram sua divulgação por meio de traduções e da criação de centros de estudo. Pensadores como Al Kindi, Alfarabi, Avicena e Averrois transmitem os conhecimentos dos antigos no campo das ciências, são os introdutores dos algarismos arábicos no Ocidente e criadores da álgebra. Na alquimia, sistematizam os fatos observados durante gerações e aceleram a passagem do ocultismo para o estudo racional. Na astronomia, aperfeiçoam os métodos trigonométricos para o cálculo das órbitas dos planetas e na medicina, transmitem as obras de Hipócrates e Galeno, além de organizarem um trabalho original de sistematização.

Na Idade Moderna, a ciência não é mais a serva da teologia, deixa de ser um saber contemplativo, formal e finalista, para que, ligada à técnica, possa servir à burguesia. Surge o método experimental, fundamentado na observação, no delineamento de hipóteses, adoção de controles e interpretação de fatos para permitir generalização. Esse período pode ser resumido pela frase consagrada de Eduardo Prado Coelho na obra Estruturalismo: *com Copérnico, o homem deixou de estar no centro do*

[2] http://www.vidamedica.com.br/mat68.pdf

universo. Com Darwin, o homem deixou de ser o centro do reino animal. Com Marx, o homem deixou de ser o centro da história (que, aliás, não possui um centro). Com Freud, o homem deixou de ser o centro de si mesmo.

Os avanços da medicina foram acompanhados pelos avanços na tecnologia voltada para o diagnóstico, que atualmente conta com equipamentos sofisticados, como é o caso do citômetro de fluxo. Esse aparelho realiza análise hematológica a partir de amostras em tubo de ensaio. O sangue é aspirado, passa por câmaras onde diferentes células são contadas utilizando tecnologia a laser. O primeiro citômetro de fluxo a ser instalado no Brasil foi em 1988 na Fundação Osvaldo Cruz, no Rio de Janeiro, assim, consiste em um equipamento recentemente incluso na rotina laboratorial de um número restrito de laboratórios, principalmente porque seu custo é de cerca de US$ 100.000,00.

Para se fazer a contagem diferencial de leucócitos no citômetro de fluxo é necessária a marcação com anticorpos monoclonais, o que depende da célula de interesse. O custo de um frasco de anticorpo monoclonal é de cerca US$ 230,00, de alto custo pois são substâncias importadas. Cada frasco tem 2 ml e são utilizados 5 ml de cada anticorpo monoclonal em cada teste. O citômetro é utilizado para diversas finalidades num laboratório como o Laboratório de Hematologia da Faculdade de Medicina de Ribeirão Preto (FMRP - USP). As principais delas são o diagnóstico de casos de leucemia aguda, de doenças linfoproliferativas crônicas e de linfomas não-Hodgkin em fase leucêmica, diagnóstico da hemoglobinúria paroxística noturna, contagem de linfócitos T CD4 e CD8 em pacientes com AIDS, estudo de sub-populações linfocitárias em pacientes com imunodeficiência primária, contagem de células CD34+ no sangue periférico e na medula óssea em pacientes ou doadores que participam de transplante de células-tronco hematopoéticas, além do estudo de doença residual mínima em pacientes que estão em tratamento quimioterápico para leucemias agudas.

A porcentagem de classificação correta do citômetro de fluxo pode ser de 90%, uma taxa de acerto ótima e privilegiada pelo grande número de células analisadas por amostra (10.000, 50.000 ou até mais dependendo do procedimento). A empresa Becton-Dickinson é a fabricante dos citômetros mais usados em clínica (FACSCan e FACSCalibur). A porcentagem de erro de um citômetro pode aumentar no procedimento conhecido como *cell sorting*, que envolve um equipamento

conectado ao citômetro para separação das diferentes sub-populações de células mediante parâmetros pré-selecionados. A rapidez em que as células são submetidas na câmara de fluxo também influencia a taxa de erro.

1.2 Diagnóstico e cibernética

Cibernética é uma palavra que se origina do grego *kibernetiké* (timoneiro; o que governa o timão da embarcação; o homem do leme, aquele que dirige ou regula qualquer coisa; guia, chefe). Norbert Wiener (1894 – 1964) foi o matemático norte-americano considerado pai desta disciplina e cunhou o termo para designar a área de pesquisa que explica o estudo das funções humanas de controle e dos sistemas mecânicos e eletrônicos que se destinam a substituí-los. A visão cibernética dedica-se ao estudo de visão de máquina utilizando a visão biológica como modelo, embora sistemas artificiais estejam longe de se equiparar ao potencial da visão humana [8], que integra informações, processando-as paralelamente.

Um exemplo de aplicação de visão artificial é o desenvolvimento de sistemas automáticos de processamento e análise de imagens microscópicas, as quais são importantes em procedimentos laboratoriais de rotina que visam diagnosticar doenças que cursam com alterações morfológicas das células hematopoéticas, dentre elas as leucemias e estados infecciosos. A variabilidade dessas células é muito grande, ainda que sendo do mesmo tipo. Segundo os testes de proficiência em diagnóstico por humanos, realizados pelo Colégio Americano de Patologia, a taxa de acerto para linfócitos, por exemplo, pode variar entre 79,3% a 92,3% [1], embora o acerto entre os profissionais mais experientes esteja em torno de 95%.

O corrente projeto propõe modelos computacionais, contudo se salienta sobre a fragilidade desses sistemas em virtude da complexidade e variabilidade das células, propriedades intrínsecas de sistemas biológicos. Diagnósticos visuais despendem tempo e recursos humanos de alta qualificação, portanto a automação da análise citológica tem potencial no auxílio ao diagnóstico laboratorial de várias doenças, proporcionando maior objetividade, uma vez que permite quantificar os atributos celulares. A motivação desse trabalho foi analisar e classificar células sangüíneas periféricas normais e leucemias linfocíticas crônicas, prolinfocítica e tricoleucemia para fins de auxiliar o diagnóstico por computador a partir de micrografias digitalizadas.

Com a automação desse processo, há possibilidade de analisar um grande número de imagens com geração de melhores estatísticas a respeito das células presentes nas amostras de sangue quando em comparação com a visão humana.

O reconhecimento da célula envolve basicamente três etapas: a segmentação da região de interesse (ex: o leucócito), a extração de características (área do núcleo, densidade de cromatina, razão núcleo-citoplasma, etc.) e a classificação ou diagnóstico (ex: linfócito anormal com características de tricoleucemia). Os atributos dos leucócitos exprimem informação sobre a forma, cor e textura, utilizando métodos clássicos e modernos para cálculo destas propriedades. Métodos de seleção de características foram implementados para investigação de quais dos atributos eram os mais promissores para descrição das classes de padrões em questão, entretanto o programa implementado independe do problema, podendo ser usado como ferramenta de mineração de dados genérica.

Neste projeto de doutorado desenvolveu-se um programa para reconhecimento de leucócitos de esfregaços de sangue periférico, cuja taxa de acerto foi baseada na classificação prévia por hematologistas, que os diferenciou em leucócitos normais (basófilo, eosinófilo, linfócito, monócito e neutrófilo) e anormais (leucemias prolinfocítica, linfóide crônica e tricoleucemia). Esse trabalho vem sendo desenvolvido em colaboração com médicos hematologistas do HC-FMRP-USP, sob co-orientação do professor Dr. Marco A. Zago e ampla ajuda dos alunos de doutorado: Sérgio Martins, Rodrigo Tocantins Calado e Edgar G. Rizzatti. Através do Laboratório de Hematologia da FMRP, imagens foram coletadas, além de informações biológicas sobre os tipos de células existentes, as patologias e os diagnósticos, com a ressalva de que nem todas as leucemias são diagnosticadas através de parâmetros visuais. Particularmente, as leucemias linfoproliferativas crônicas podem ser diagnosticadas através de análise visual. Dentre elas, objetivou-se abordar as seguintes patologias: leucemia linfóide crônica (LLC), leucemia prolinfocítica e leucemia de células cabeludas ou tricoleucemia (*Hairy cell*). A LLC foi particularmente estudada por ser a mais comum das leucemias linfóides, representado 30% de todos os casos, cuja incidência é de cerca de 1.500 novos casos por ano no Brasil, particularmente dois casos por semana no Laboratório de Hematologia da FMRP-USP.

Segundo resultados obtidos utilizando o software projetado para uma base de imagens com 938 células, contendo os 5 tipos de leucócitos normais e a anomalia

LLC, foi possível diagnosticar corretamente 100% dos basófilos, 86% dos eosinófilos, 92% dos linfócitos, 81% dos monócitos, 98% dos neutrófilos e 86% das células de LLC segundo artigo recém publicado [9].

A primeira parte dessa qualificação se dedicará à descrição de metodologias, incluindo o Capítulo 2, que trata da identificação e discriminação das células em esfregaços de sangue, o Capítulo 3 que traça um panorama sobre o métodos em reconhecimento de padrões e o Capítulo 4, que apresenta conceitos sobre seleção de características e delineia alguns conceitos importantes a respeito de mineração de dados. O Capítulo 5 fornece os resultados produzidos durante o programa de doutorado, incluindo as ferramentas implementadas e os métodos descobertos e/ou testados. O Capítulo 6 enfatiza a discussão e conclusões e delineia os trabalhos futuros. Os apêndices complementam o trabalho com ilustrações, manuais e os artigos elaborados durante esse programa de doutorado.

Capítulo 2

O sangue humano

"O sangue, por si só, move as rodas da história. (Benito Mussolini)"

Em todo o mundo, amostras de sangue são analisadas visualmente, sendo que os exames mais comumente solicitados para o diagnóstico dos distúrbios hematológicos são o hemograma e o esfregaço sanguíneo [10]. A necessidade de padronização para o diagnóstico fez surgir testes de proficiência em laboratório clínico, como os ministrados pelo Colégio Americano de Patologia (CAP) [1] para controle de qualidade dos laboratórios, providenciando uma lista de atributos importantes a serem observados nas células. Este teste é considerado padrão de referência mundial para classificação em laboratório. As informações deste capítulo são baseadas em entrevistas com hematologistas, em endereços na internet e livros, especialmente os atlas de hematologia [1, 2], amplamente utilizados para elaboração das tabelas apresentadas neste capítulo.

Os elementos figurados do sangue periférico compreendem: as células completas, ou seja, os leucócitos (células brancas), compostas de citoplasma, núcleo e organelas; as células anucleadas e desprovidas de organelas citoplasmáticas, como os eritrócitos (hemácias) e finalmente as plaquetas, que constituem pequenos fragmentos citoplasmáticos de células da medula óssea, os megacariócitos. Estas células possuem funções bastante diversas. Por exemplo, os eritrócitos são células envolvidas primordialmente no transporte de gases: transporta oxigênio do pulmão aos tecidos e dióxido de carbono dos tecidos aos pulmões.

Os elementos figurados do sangue[1] (leucócitos, eritrócitos e plaquetas) e seus

[1]Links interessantes:

precursores são tradicionalmente agrupados em duas linhagens [11]:

A **linhagem mielóide** inclui células produzidas exclusivamente na medula óssea dos adultos, circulam por um tempo limitado, sendo em seguida destruídas ou migrando para os tecidos para não mais retornar à circulação. As células da linhagem mielóide são:

- eritrócitos e seus precursores;
- plaquetas e seus precursores;
- granulócitos (ou seja, neutrófilos, eosinófilos e basófilos) e seus precursores;
- monócitos e seus precursores;

A **linhagem linfóide** inclui células que são produzidas primariamente na medula óssea e timo, onde estas células se dividem espontaneamente independente de estímulo de antígenos, além de povoar numerosos órgãos linfóides secundários. Nestes órgãos linfóides secundários elas se multiplicam especialmente sob estímulo antigênico.

As principais células da linhagem linfóide ("imunócitos") são os linfócitos, seus precursores e plasmócitos. Estas células, contrariamente às células mielóides, podem "re-circular", ou seja, podem abandonar a circulação e depois voltar ao sangue, podendo ter uma sobrevida de meses ou anos.

Existem, pois, na medula óssea células precursoras indiferenciadas (células tronco) da linhagem mielóide, que são capazes de se diferenciar e dar origem aos precursores de eritrócitos, granulócitos, monócitos e plaquetas. Os granulócitos estão particularmente envolvidos em processos inflamatórios. Existem também células precursoras da linhagem linfóide que são capazes de se diferenciar em linfócitos do tipo T, linfócitos do tipo B e outras células ligadas aos fenômenos de imunidade.

---

http://www.mc.vanderbilt.edu
http://www.meds.com

2.1 Leucócitos - produção e função

Podem ser divididos em dois grandes grupos, quanto a sua função: fagócitos e linfócitos. Os fagócitos, que incluem os neutrófilos, eosinófilos, basófilos e monócitos, são células com capacidade de fagocitar material estranho (como bactérias ou fungos) ou ainda células mortas ou lesadas. Normalmente existem no sangue periférico apenas fagócitos maduros e linfócitos, na seguinte ordem decrescente de frequência: neutrófilos, linfócitos, monócitos, eosinófilos e basófilos [12, 2, 10]. A Fig.2.5 mostra o desenho esquemático de fagócitos e a Tabela 2.1, a distribuição normal de leucócitos no sangue.

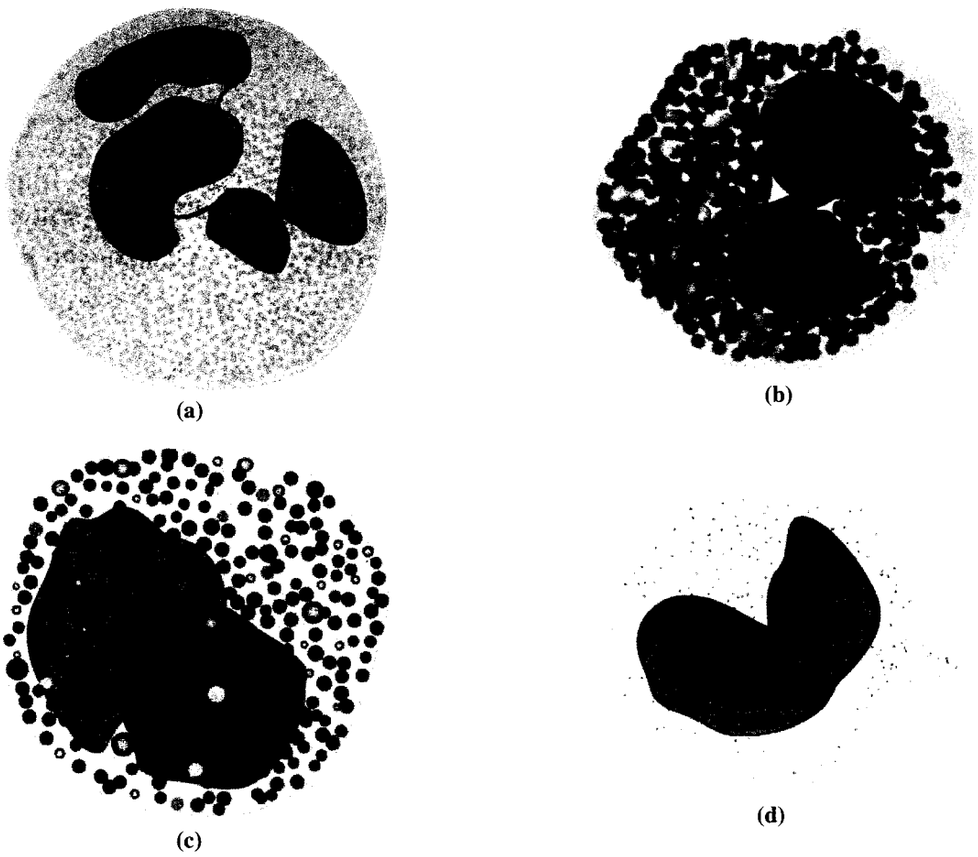


Figura 2.1: Desenho esquemático de fagócitos: **(a)** neutrófilo, **(b)** eosinófilo, **(c)** basófilo e **(d)** monócito (fonte: [1]).

Os granulócitos são produzidos a partir de células imaturas denominadas mielo-

Tabela 2.1: Leucócitos: Número total 4.000 - 11.000/mm^3

Leucócito	Porcentagem	Números absolutos
neutrófilos	57 - 67%	2.280 - 7.370/mm^3*
linfócitos	25 - 33%	1.000 - 3.630/mm^3
monócitos	3 - 7%	120 - 770/mm^3
eosinófilos	1 - 3%	40 - 330/mm^3
basófilos	0 - 1%	0 - 110/mm^3

* MÍNIMO = 4.000 x 57% = 2.280, MÁXIMO = 11.000 x 67% = 7.370

blastos. Estas células amadurecem e multiplicam-se passando por estágios denominados: mieloblastos, promielócitos, mielócitos, metamielócitos e, então, célula madura, que é o estágio em que a célula é geralmente encontrada no sangue periférico.

2.1.1 Neutrófilos

Os neutrófilos maduros são subdivididos, segundo a forma do seu núcleo, em segmentados (os mais abundantes) e bastonetes. A percentagem habitual dos dois tipos de neutrófilos é de:

- neutrófilos segmentados = 54 - 62%
- neutrófilos bastonetes = 3 - 5%
- total neutrófilos maduros = 57 - 67%

Em condições normais, a proporção de bastonetes/segmentados varia de 1/11 a 1/21 (3/62 = 1/21; 5/54 = 1/11).Os neutrófilos possuem cromatina com aglutinações (grossa) e o citoplasma apresenta grânulos lilases (característica específica ou típica) predominantes. O núcleo dos bastonetes pode assumir diversas formas: de banda, de salsicha, em S, C, U e lobulado, podendo se dobrar sobre si mesmo. Neutrófilos segmentados possuem núcleo dividido em 2-5 lóbulos ligados por filamentos sem cromatina interna. A presença de um núcleo claramente segmentado permite distingui-los dos monócitos e linfócitos e a natureza dos grânulos os segrega dos outros granulócitos (eosinófilo e basófilo). Apresenta-se a caracterização dos neutrófilos detalhadamente na Tabela 2.2.

Tabela 2.2: Características do Neutrófilo, segundo [1].

	Neutrófilo bastão	Neutrófilo segmentado
Tamanho da célula	$10 - 18\mu$m diâmetro	$10 - 15\mu$m diâmetro
Razão núcleo:citoplasma	1 : 1.5 a 1 : 2	1 : 3
Forma nuclear	banda estreita alongada, com espessura uniforme, lóbulo nuclear único	2-5 lóbulos nucleares (filamento conectando os lóbulos: 0.5μm, menor que $1/3$ da largura do lóbulo)
Posição nuclear	central ou acêntrico	central ou acêntrico
Cor nuclear e cromatina	cromatina granular áspera azul-púrpura escuro	cromatina granular áspera azul-púrpura escuro
Nucléolo	inexistente	inexistente
Cor e quantidade de citoplasma	rosa abundante	rosa abundante
Grânulos citoplasmáticos	específicos, rosa-violeta fino	específicos, rosa-violeta fino

2.1.2 Eosinófilos

Os eosinófilos são leucócitos que se coram com eosina de modo que apresentem grânulos avermelhados. Suas características são listadas na Tabela 2.3. Características chave na identificação:

Tabela 2.3: Características do Eosinófilo maduro segundo [1]

Tamanho da célula	10-15μm
Razão núcleo:citoplasma	1 : 2 a 1 : 3
Forma nuclear	redondo a oval
Posição nuclear	segmentado, com 2-3 lóbulos conectados por um fino filamento de cromatina
Cromatina	densa e compacta
Nucléolo	inexistente
Cor e quantidade de citoplasma	azul pálido e abundante
Grânulos citoplasmáticos	Presença de grânulos esféricos, alta rugosidade, coloração uniforme de laranja a vermelho

2.1.3 Basófilos

Estes são os menores granulócitos circulantes (diâmetro 10-14 μm) [11]. O núcleo é parcialmente visível, obscuro a oculto por grânulos metacromáticos de avermelhados a pretos, composto de cromatina muito densa com formato lobulado incompleto, na forma de uma folha. Os grânulos são menos numerosos que aqueles do neutrófilo e menos empacotados que aqueles do eosinófilo, mas diferente dos grânulos dos eosinófilos, cobrem o núcleo. As características do basófilo podem ser resumidas pela Tabela 2.4.

Tabela 2.4: Características do Basófilo maduro segundo [1]

Tamanho da célula	$10 - 15\mu$m
Razão núcleo:citoplasma	1 : 2 a 1 : 3
Forma nuclear	redondo a oval
Posição nuclear	segmentado, frequentemente obscuro pelos grânulos citoplasmáticos
Cor nuclear e cromatina	cromatina granular áspera azul-púrpura escuro, com aglutinação em alguns pontos
Nucléolo	inexistente
Cor e quantidade de citoplasma	azul pálido abundante
Grânulos citoplasmáticos	Presença de grânulos de diferentes tamanhos e formas densamente corados; os grânulos são específicos e em geral maiores que os do neutrófilo e do mesmo tamanho dos grânulos do eosinófilo

2.1.4 Monócitos

O núcleo se apresenta em geral identado[2], em forma de C ou redondo. A cromatina é mais frouxa, solta, irregular (heterogênea) que o núcleo dos linfócitos típicos. "Assemelha-se a um pintura feita com bucha enquanto os linfócitos parecem ter o núcleo pintado com um pincel de ponta fina", afirma Dr. Rodrigo T. Calado, colaborador nesse projeto de doutorado. A presença freqüente de vacúolos e a cor menos basofílica (acinzentada, com similaridade a "vidro fosco") do citoplasma do monócito podem ser critérios mais diferenciadores dos linfócitos, em alguns casos. A forma da célula é redonda e o citoplasma pode conter grânulos lilases ínfimos

[2]identado = que possui identação, assemelhando-se a rim; forma de feijão

(material fagocitado), podendo se assemelhar a pseudópodes. A Tabela 2.5 sintetiza as características dos monócitos.

Tabela 2.5: Características do monócito maduro segundo [1]

Tamanho da célula	12-20μm
Razão núcleo:citoplasma	4:1 a 2:1
Forma nuclear	redondo a oval, identado
Posição nuclear	central
Cor nuclear e cromatina	Vermelho-púrpura pálido, cromatina mínima, fina, cheias de fios
Nucléolo	inexistente
Cor e quantidade de citoplasma	azul-cinza, vacúolos espalhados
Grânulos citoplasmáticos	pode conter grânulos lilases ínfimos

2.1.5 Linfócitos

Os linfócitos são células do sistema imunitário, formados no tecido linfóide do corpo, isto é, nos nódulos linfáticos, timo, baço, amígdalas, placas de Peyer e algumas vezes na medula. São geralmente homogêneos, mas variam numa faixa de aparências morfológicas [1]. Podem ser pequenos (7 a 9 μm), mas formas mais maduras são frequentemente encontradas (10 a 18 μm). O núcleo contém cromatina homogênea densa que se cora fortemente de azul-púrpura. O citoplasma é escasso, agranular, lateralizado e cora-se de azul celeste pálido. Em formas maiores, o citoplasma pode ser abundante e ocasionalmente contém um pouco de grânulos vermelho-violeta. Estas células aparecem aproximadamente 25-45% na contagem total de leucócitos.

O linfócito imaturo pode ser encontrado no sangue periférico e as principais características que o diferencia do maduro são: o núcleo apresenta uma combinação de cromatina condensada e agrupada, azul-púrpura com paracromatina[3] vermelho-púrpura, citoplasma claramente basofílico[4], podendo apresentar nucléolos, embora pouco distintos [2].

O linfócito grande granular (LGG) é um linfócito normal maior que os linfócitos normais típicos, com citoplasma levemente basofílico, vários grânulos no citoplasma, forma do núcleo mais irregular que o linfócito típico e a cromatina medianamente

[3]paracromatina = cromatina clara e escura.

[4]coloração do citoplasma que varia de cinza-azulado a púrpura

homogênea. Dos linfócitos normais circulantes, 10-15% são linfócitos grandes granulares (células NK[5]).

Linfócitos "maduros"do sangue periférico ou dos tecidos podem ser estimulados a se multiplicar quando entram em contato com antígenos específicos e assumir formas variantes, cuja nomenclatura pode aparecer como linfócito atípico ou linfócito reacional ou linfócito ativado[6]. O desenho esquemático da Fig.2.2 e as micrografias da Fig.2.3 ilustram os linfócitos reacionais. Vários fatores podem desencadear a ativação do linfócito, como viroses, etc.

Linfócitos reacionais são caracterizados por uma variabilidade de aparências morfológicas ainda maior numa mesma lâmina. O tamanho destes linfócitos varia em $10 - 25\mu$m, apresentam aumento do núcleo e do citoplasma, em geral formas mais irregulares (menos redondas ou ovais). A morfologia dos linfócitos pode ser alterada tanto em resposta a estímulos imunológicos quanto infecções e doenças virais, toxoplasmose, reações alérgicas e outras doenças, embora pacientes normais e/ou com hipersensibilidade a drogas também possam apresentar linfócitos alterados. Os linfócitos reacionais são divididos em 3 tipos, segundo Downey (1923). Essa categorização é considerada obsoleta por muitos investigadores e de pouca importância, sendo tema de muita controvérsia entre os hematologistas, porém uma descrição útil didaticamente. Essas células são linfócitos maduros altamente diferenciados e diferem de seus correspondentes imaturos de 2 formas: a)linfócitos atípicos contém cromatina nuclear madura vermelho-púrpura em contraste com formas imaturas que apresentam cromatina nuclear imatura com coloração pálida com grandes quantidades de paracromatina e b) linfócitos atípicos contém citoplasma abundante, não escasso como nas formas imaturas [2].

As variações morfológicas mais comuns dos linfócitos (em relação aos linfócitos típicos) são listadas a seguir, porém têm conotação exclusivamente morfológica, sem significar uma alteração funcional específica ou um fator etiológico [11]:

- núcleos com formas variáveis;
- quantidade de citoplasma aumentada;

[5]NK = abreviatura de Natural Killer ou célula assassina.

[6]O linfócito reativo pode ser considerado uma célula anormal pois há diferenciação em função dos antígenos, contudo não significa leucemizada

Tabela 2.6: Características do linfócito maduro segundo [1]

Tamanho da célula	7-18μm (a maioria das formas com $9-12\mu$m de diâmetro)
Razão núcleo:citoplasma	4:1
Forma nuclear	redondo ou identado
Posição nuclear	central, com citoplasma concentrado em um dos lados ou redondo
Cor nuclear e cromatina	homogênea, rugosa, azul-púrpura
Nucléolo	usualmente ausente, raramente encontrado em formas maduras
Cor e quantidade de citoplasma	azul celeste claro, quantidade escassa a moderada
Grânulos citoplasmáticos	pode conter poucos grânulos vermelho-violeta ou, no caso de linfócito atípico, vários grânulos

- razão núcleo-citoplasmática aumentada;
- basofilia do citoplasma aumentada, podendo ser pálida a moderada;
- perda da homogeneidade do citoplasma;
- vacúolos citoplasmáticos;
- aparecimento de um centrossomo[7] distinto com marcação pálida;
- presença de grânulos azurofílicos grandes;
- redução da condensação de cromatina (frouxa e heterogênea);
- aumento do tamanho nuclear;
- aparecimento de nucléolos;

2.2 Linfoma e leucemia linfoproliferativa crônica

Linfoma, ou linfadenoma, (Fig.2.4) é cada uma das várias patologias benignas e malignas (uma das quatro maiores incidências de câncer) nas quais ocorre proliferação

[7]centrossomo = área pálida do citoplasma, aparecendo pouco corada ou não corada, sem grânulos e situada próxima do centro da célula, correspondendo, em nível ultra-estrutural, a vesículas e membranas do aparato de Golgi.

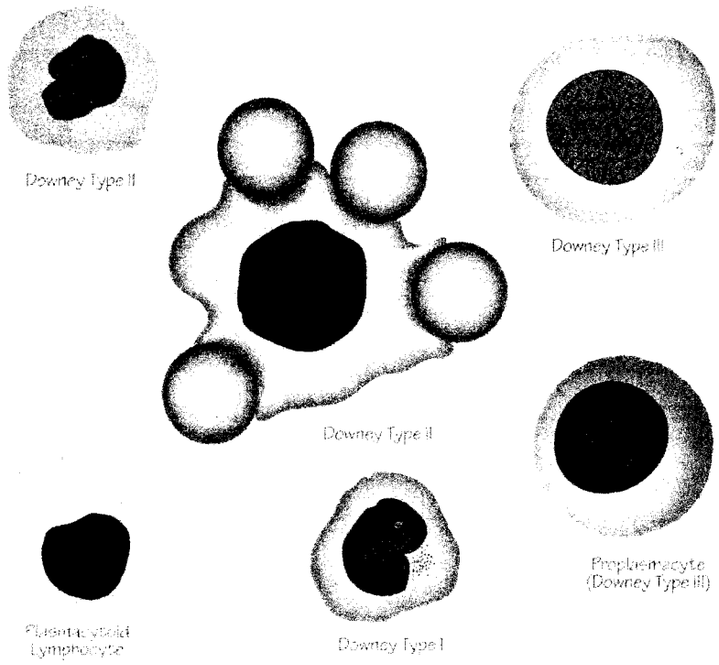


Figura 2.2: Desenho esquemático de linfócitos reativos [1]

de linfócitos histiocitários, ou seja, originados nos tecidos linfóides. Os linfomas ocorrem em decorrência a um erro no processo de produção de linfócitos no organismo, resultando em uma célula anormal que pode se acumular por dois motivos: a) eles podem se duplicar mais rápido que células normais e/ou (b) eles podem viver mais que células normais. São classificados em Hodgkin[8] (LH) e não Hodgkin (LNH), pela presença ou não das células de Reed-Sternberg[9].

Em 1996, mais de 52.000 novos casos de LNH foram diagnosticados nos Estados Unidos e mais de 23.000 pessoas morreram da doença. As várias formas de LNH podem compartilhar características morfológicas e imunológicas, com muitas outras disfunções. Abaixo sumarizam-se algumas características de LH e LNH:

- LH: uma desordem maligna na qual há aumento progressivo (mas sem dor)

[8]Thomas Hodgkin, 1832: cirurgião inglês que descreveu processos que acometiam linfonodos, os quais aumentavam muito de volume e levavam o paciente à morte; havia comprometimento de mediastino e baço.

[9]Sternberg (Áustria) e Dorothy Reed (EUA), 1902: descreveram na histologia dos linfomas de Hodgkin uma célula característica, que ficou conhecida como célula de Reed-Sternberg.

do tecido linfático seguido de aumento do baço e do fígado. Esse linfoma é o mais estudado e os avanços corroboraram para a melhoria e rapidez no diagnóstico e tratamento da doença, cuja cura é possível em 80% dos pacientes,

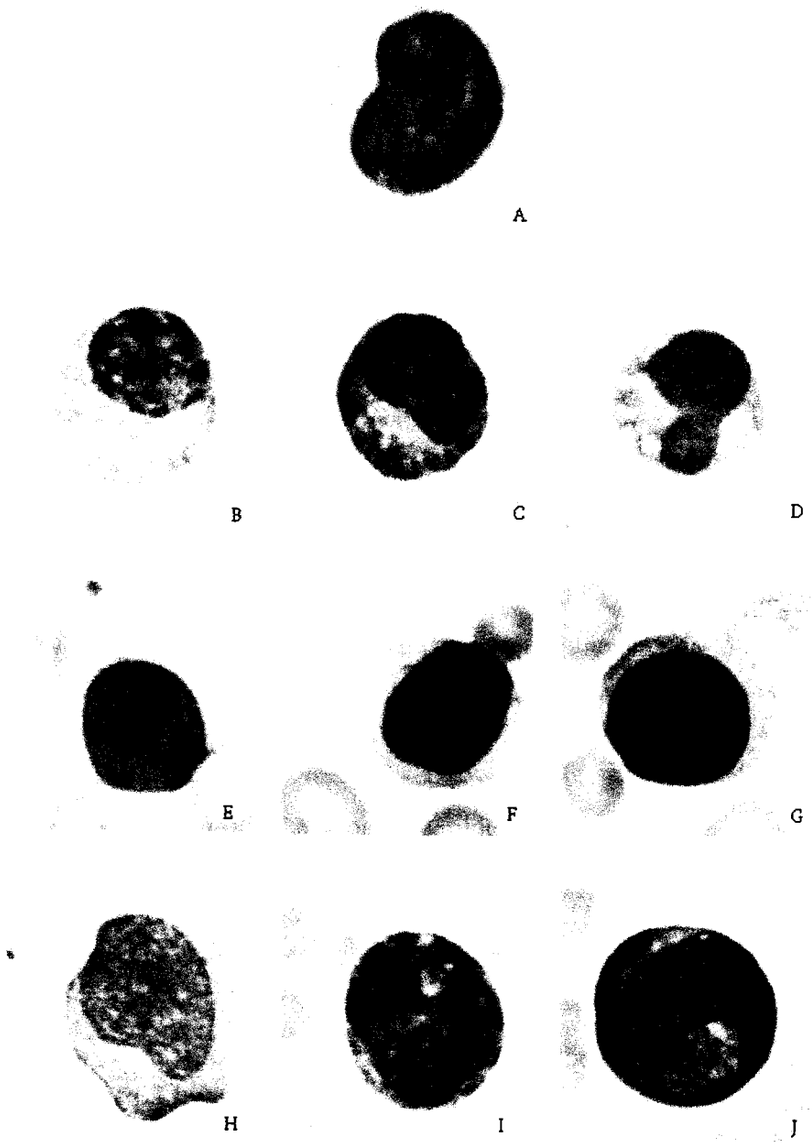


Figura 2.3: Micrografias de linfócitos reativos; (a)linfócito normal; (b-d)linfócito atípico Downey tipo I; (e-g)linfócito atípico Downey tipo II; (h-j)linfócito atípico Downey tipo III [2]

aproximadamente.

- LNH: um grupo de câncer do sistema linfóide, incluindo leucemia linfocítica aguda, linfoma de célula B, linfoma de Burkitt, linfoma da célula difusa, linfoma folicular, linfoma de célula grande imunoblástica, linfoma linfoblástico, linfoma da célula do manto, mycosis fungoides (Sézary), desordem linfoproliferativa pós-transplante, linfoma da célula pequena clivada e linfoma de células T.

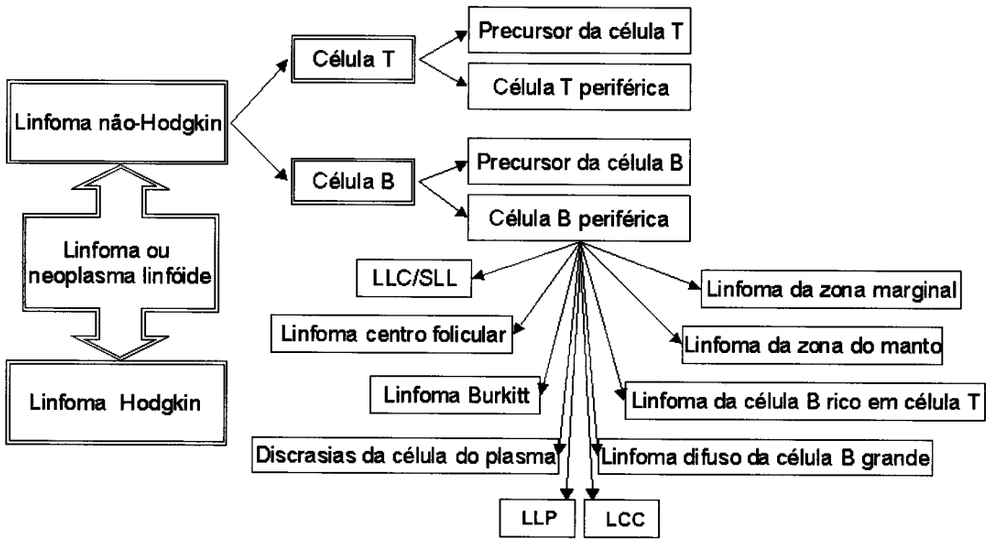
A subclassificação dos LNH representou durante muitos anos um problema para os patologistas, que consideravam apenas o tamanho e classe da célula. Atualmente, sabe-se que existem mais de 40 categorias diferentes de LNH, de diferentes origens celulares. Em 1994, um grupo de patologistas propôs uma classificação consensual dos linfomas, o sistema REAL (Revised European-American Lymphoma), que subdivide os linfomas de acordo com a linhagem da célula (B ou T) e seu grau de maturidade, considera marcadores de superfície e citogenética. Existem outros sistemas de classificação como o proposto pela OMS (Organização Mundial de Saúde), pela FAB (French-American-British cooperative group).

O tratamento de LNH dependo do estágio patológico e tipo histológico pode ser radioterapia local para adenopatia sintomática, quimioterapia com agente único ou combinada , o que depende do estágio de evolução da doença. Se em Estágio I e II: radioterapia isolada senao se em Estágio III e IV: maioria dos casos acomete idosos. A doença é incurável, com evolução lenta 25%. Tratamento com Quimioterapia (clorambucil, fludarabina, COP, CHOP).

Leucemia crônica é aquela cuja progressão se dá lentamente e leucemias linfoproliferativas crônicas são especificamente aquelas em que há proliferação de linfócitos com semelhanças com os linfoblastos (linfócito imaturo). O diagnóstico através da análise visual de esfregaços de sangue periférico é possível em casos de leucemias linfoproliferativas crônicas.

Essa tese tratará somente de neoplasmas das células B do sangue periférico, especificamente a leucemia linfóide crônica (LLC), prolinfocítica e tricoleucemia (LCC), sendo todos casos de desordem linfoproliferativa crônica. O diagnóstico por meio de análise visual da morfologia se torna impossível em casos onde as células neoplásicas possuem grande capacidade de proliferação e são pleomorfas, umas com variantes

Figura 2.4: Linfomas e subdivisões



fonte: http://family.georgetown.edu/welchjj/netscut/heme_onc/lymphoma.html

blastóides e prolinfocitóides, simultaneamente como é o caso do linfoma da zona do manto.

Através da coleta dessas informações, delineou-se o escopo da tese corrente. Segundo dados casuísticos fornecidos pelo Laboratório de Hematologia do Hospital das Clínicas de Ribeirão Preto (HCRP), a leucemia linfóide crônica é o linfoma mais freqüente, inclusive com ocorrência muito maior que os outros linfomas.

2.2.1 Leucemia linfóide crônica (LLC)

A LLC e o linfoma linfocítico ou de células pequenas[10] são basicamente a mesma doença, cuja incidência maior é em adultos do sexo masculino (duas vezes mais freqüente que em mulheres), geralmente acima dos 50-60 anos. As leucemias de células pequenas compartilham várias características morfológicas com a LLC, inclusive indistintos quanto à expressão de alguns antígenos de superfície. Há proliferação de linfócitos de aspecto maduro, bem diferenciado, que infiltram linfonodos, baço (polpa branca e vermelha), fígado e a medula óssea. Se há células neoplásicas

[10]também conhecido como linfoma linfocítico ou de células pequenas, que corresponde à ocorrência da leucemia linfóide crônica no tecido enquanto que a LLC ocorre no sangue periférico.

no sangue circulante (linfocitose de 4.000 células por mm^3 ou maior) fala-se em LLC. Se não, o diagnóstico é linfoma não-Hodgkin linfocítico (considerado de baixo grau de malignidade). A LLC é a leucemia mais comum de adultos nos países ocidentais (mais rara no Oriente). Já o linfoma linfocítico constitui apenas 4% dos linfomas não-Hodgkin [12, 2, 10].

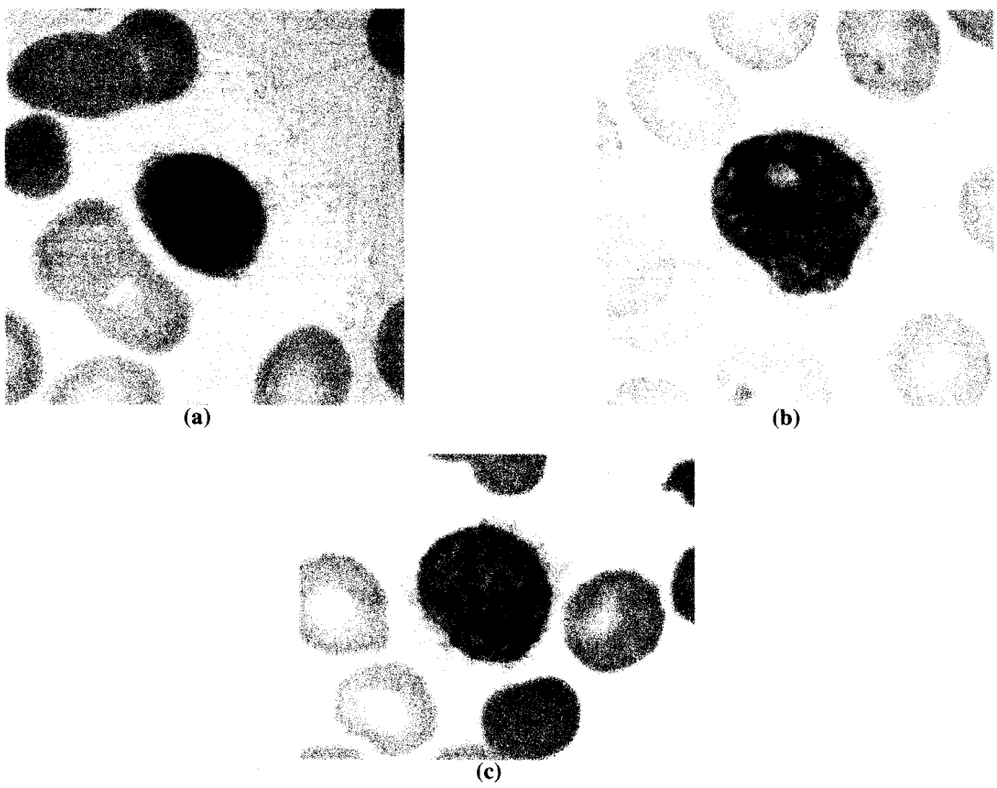


Figura 2.5: Micrografias de **(a)** LLC, **(b)** prolinfócito, **(c)** *Hairy cell.*

Clinicamente, a doença é indolente e muitos pacientes são assintomáticos. Sintomas incluem fadiga, perda de apetite e de peso. Aumento de volume dos linfonodos (linfadenopatia ou linfadenomegalia) e hepatoesplenomegalia ocorrem em 50-60% dos casos. A contagem de células no sangue periférico pode ser normal ou chegar a 200.000 por mm^3. O curso e prognóstico são muito variáveis. Tipicamente a sobrevida é de 4 a 6 anos, mas pode ser de 10 anos ou mais. Um fator de piora prognóstica é a transformação da LLC em tipos mais agressivos, como a transformação prolinfocítica (15 a 30%) ou para linfoma difuso de grandes células B (chamada síndrome

de Richter, 10%).

A leucemia linfóide crônica-B é a mais comum das leucemias linfóides crônicas e representa cerca de 30% de todas as leucemias. A sua incidência é de 5 casos para cada 100.000 habitantes por ano, o que representa cerca de 1.500 casos novos por ano no Brasil. Levando em conta o fato de que a sobrevida média gira em torno de 10 anos, são cerca de 15.000 pacientes com LLC no Brasil atualmente.

No Laboratório de Hematologia da FMRP-USP, faz-se o diagnóstico de aproximadamente dois casos por semana de LLC. Isto representa não só Ribeirão Preto, mas também amostras de sangue e medula óssea que vêm de Araraquara, Barretos, Santos, Brasília, Uberaba, etc. O diagnóstico é feito por um ou mais médicos hematologistas (também pode ser feito por um médico patologista clínico) e a análise morfológica no microscópio *é essencial* para o diagnóstico. O exame complementar é a imunofenotipagem e, às vezes, a biologia molecular. Outros laboratórios no país também fazem este tipo de diagnóstico, mas são poucos os que contam com o hematologista e a imunofenotipagem, por conta do custo. Acredita-se que cerca de 30 a 50 laboratórios no país estejam aptos a fazer este diagnóstico.

Os linfócitos de LLC são células maduras bem diferenciadas e podem se assemelhar muito a linfócitos normais, embora geralmente sejam ligeiramente maiores (LLC) ou pequenos (linfoma linfocítico), com núcleo grande em relação ao citoplasma (algumas vezes o citoplasma é imperceptível), coloração de cromatina alterada, com aparecimento freqüente de "ranhuras[11]"no núcleo. Citoplasma basofílico, com aspecto granuloso. Grande proliferação e aglutinação de células. Contagem absoluta de linfócitos deve ser maior que 5.000 células por μl, mantida por um período maior ou igual a 4 semanas. Exames de medula óssea não são necessários para o diagnóstico, mas podem ajudar no estadiamento e na avaliação do prognóstico. Mais detalhes são listados na Tabela 2.7.

2.2.2 Leucemia prolinfocítica (LLP)

Prolinfócitos (Tabela 2.8) podem ser encontrados em casos de leucemia prolinfócita ou mesmo de LLC. O diagnóstico é determinado pelo número de achados de LLC

[11]ranhuras = pontos mais claros do núcleo, que aparecem consecutivamente, dando a impressão de ter sido arranhado; muitas vezes possue uma direção preferencial.

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

Tabela 2.7: Características da LLC segundo [1]

Tamanho da célula	8-30μm
Razão núcleo:citoplasma	7:1 a 3:1
Forma nuclear	redondo a oval, podendo ser irregular
Posição nuclear	central
Cor nuclear e cromatina	cromatina compacta e áspera.
Nucléolo	inexistente
Cor e quantidade de citoplasma	quantidade escassa e basofílico
Grânulos citoplasmáticos	inexistente

Tabela 2.8: Características da LLP segundo [1]

Tamanho da célula	10-18μm
Razão núcleo:citoplasma	5:1 a 3:1
Forma nuclear	redondo a oval, podendo ser ligeiramente identado
Posição nuclear	central
Cor nuclear e cromatina	algumas vezes condensada, mais áspera que nos linfoblastos e mais aberta que nos linfócitos
Nucléolo	tipicamente único e proeminente; raramente apresenta 2 nucléolos
Cor e quantidade de citoplasma	quantidade moderada, corado com azul homogeneamente
Grânulos citoplasmáticos	pode conter uns poucos grânulos azurofílicos

em relação ao de prolinfócitos. Prolinfócitos são células maiores que as LLC e têm um nucléolo central proeminente, cromatina agrupada e quantidade de citoplasma basofílico de moderado a abundante. A LLP é caracterizada por um número maior que 50% destas células no sangue ou na medula. A maioria destes casos ocorre em adultos mais velhos, com evolução indolente da doença, embora incurável com os regimes de quimioterapia correntes. Indivíduos com LLP apresentam leucocitose e esplenomegalia[12].

[12]esplenomegalia = aumento do baço observado em várias desordens do sangue [2].

2.2.3 Leucemia de células cabeludas - LCC

A leucemia de células cabeludas é também conhecida como reticuloendoteliose leucêmica, leucemia de células pilosas, tricoleucemia ou *Hairy cell.* Esse nome se dá devido a suas projeções que se assemelham a cabelos, dando uma aparência de flagelado (Tabela 2.9). Tem diâmetro grande ($15-30\mu$m) [2] com núcleo redondo a oval, mas ocasionalmente pode apresentar identação. É uma neoplasia rara causada por proliferação de linfócitos grandes maduros e diferenciados. Pacientes usualmente apresentam esplenomegalia e pancitopenia[13]. A LCC é uma das formas mais tratáveis de doença linfoproliferativa crônica, onde as células neoplásicas são linfócitos B monoclonais com projeções do citoplasma em forma de cabelos. São frequentemente confundidos com linfócitos reativos, prolinfócitos, células de linfoma, células de Sézary, monócitos e monócitos imaturos [1]. A LCC ocorre em $2-5\%$ de todos os casos de leucemia, com maior frequência em homens (3-5 vezes maior que em mulheres).

Figura 2.6: Linfócito normal (esquerda) e célula cabeluda (direita).

[13]pancitopenia = redução no número de eritrócitos, leucócitos e plaquetas no sangue periférico [2].

Tabela 2.9: Características da LCC segundo [1]

Tamanho da célula	12-20μl
Razão núcleo:citoplasma	4:1 a 2:1
Forma nuclear	redondo a oval
Posição nuclear	núcleo com forma ovóide a reniforme, podendo ser excêntrico
Cor nuclear e cromatina	cromatina homogênea difusa, fina a suavemente rugosa granular áspera azul-púrpura escuro, com aglutinação em alguns pontos
Nucléolo	pode apresentar um ou mais nucléolos pequenos
Cor e quantidade de citoplasma	irregular, moderado a abundante, azul pálido a cinza-azulado com delicadas projeções citoplasmáticas semelhantes a cabelos;
Grânulos citoplasmáticos	ocasionalmente, podem conter grânulos azurofílicos finos

Na Fig. 2.6, a célula cabeluda[14] é um pouco maior que o linfócito normal, com cromatina mais fina que o linfócito, um nucléolo indistinto e citoplasma abundante e com "cabelos". Este exemplo não exibe as características de um linfócito reativo, como alargamento nuclear e citoplasma aumentado.

2.2.4 Outras neoplasias

Existem outras neoplasias com grande potencial para automatização da análise visual. Dentre elas:

Linfoma centro folicular (linfoma da célula clivada ou linfosarcoma): tumor maligno originado e localizado nos nódulos linfáticos, baço e tecido linfóide extranodal. Há aparecimento de linfócitos ligeiramente maiores que os normais ($10 - 20\mu$m de diâmetro), com pouco citoplasma, basofílico e oval a angulador. O núcleo aparece centralmente na célula, é clivado, com contorno nuclear com identações frequentes, dobramentos, angulações. A cromatina é aberta, apresenta aglutinações com um nucléolo ocasional. Estes tumores ocorrem em adultos e concorrem com 40% dos casos de LNH nos Estados Unidos [1].

[14] fonte:http://www.cap.org

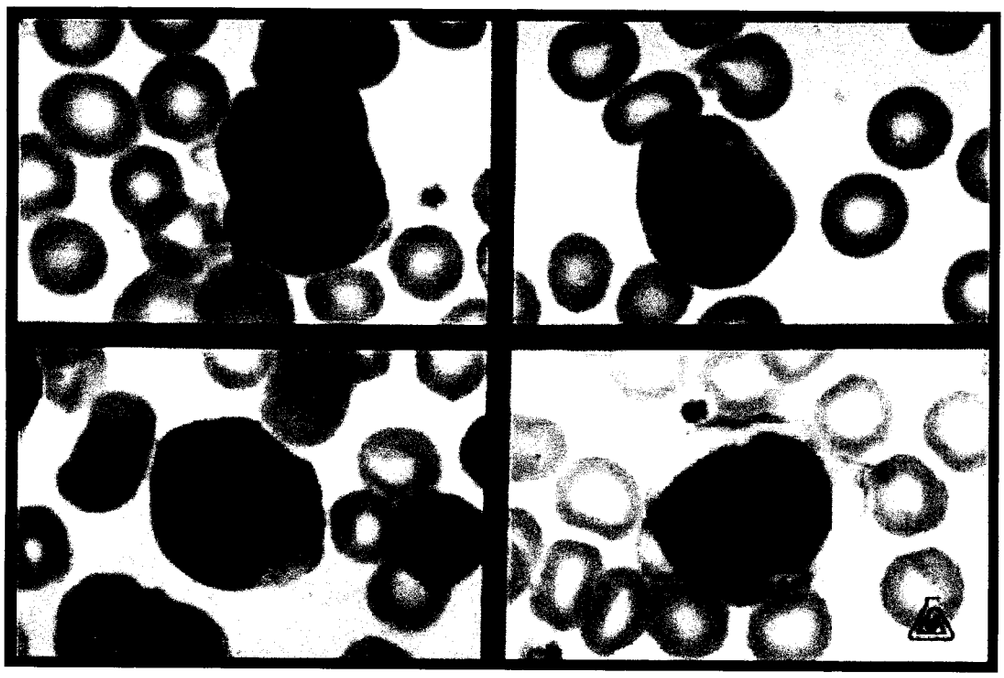


Figura 2.7: Exemplos de células de Sézary

Síndrome de Sézary (Mycosis Fungoides): também conhecida como linfoma de células T cutâneo, pois a doença afeta primariamente a pele, podendo acometer nódulos linfáticos, medula e sangue. Estas células tumorais são normalmente pequenas com núcleo marcadamente convoluto, entalhado, com dobras, dando o aspecto cerebriforme (Tabela 2.10). Células maiores podem ser encontradas apresentando os mesmos aspectos. A Fig. 2.7 mostra células[15] onde o padrão de cromatina é hipercromático com nucléolos não aparentes. O citoplasma é pálido e escasso. Em contraste com as células de Sézary, monócitos imaturos tem mais citoplasma com um núcleo dobrado redondo a oval e cromatina fina. Embora os núcleos de monócitos possam ter dobras delicadas, eles geralmente não exibem as convolução marcadas das células de Sézary.

Linfoma de células T do adulto: núcleo em trevo.

[15]Fonte:http://www.cap.org/apps/docs/committees/hematology/97he/43.html

Tabela 2.10: Características da Sézary segundo [1]

Tamanho da célula	8-20μm
Razão núcleo:citoplasma	7:1 a 3:1
Forma nuclear	redondo a oval, podendo ser irregular; forma irregular, com variados graus de dobramento (cerebriforme)
Posição nuclear	central
Cor nuclear e cromatina	hipercromática com agrupamentos densos em células pequenas; células maiores podem apresentar cromatina mais aberta e dispersa.
Nucléolo	inexistente
Cor e quantidade de citoplasma	quantidade escassa, pálida e basofílico
Grânulos citoplasmáticos	agranular, podendo ocasionalmente apresentar vacúolos que circundam o núcleo

2.3 Métodos de coloração de esfregaços

Coloração de Romanovsky se refere ao protótipo das muitas colorações que usam eosina e azul de metileno para esfregaços de sangue e parasitas de malária. A coloração de Romanowsky é usada amplamente em laboratórios de hematologia para rotinas de coloração de esfregaços de sangue periférico. O nome da técnica provém do nome do seu descobridor, Romanowsky, que em 1890 observou que uma mistura de azul de metileno com eosina permite a coloração diferenciada de regiões com diferentes alcalinidades.

Exemplos de variações da coloração de Romanowsky são:

- Giemsa: solução contendo azuro II-eosina, azuro II, glicerina e metanol; usado para corar parasitas protozoários tais como *Plasmodium spp.* e *Trypanossoma ssp.*, e *Chlamydia spp.* para coloração diferencial de esfregaços de sangue e para corpos de inclusão viral. Elementos corados apresentam cor de rosa a púrpura a azul.

- Leishman: mistura de azul de metileno e eosina para coloração de esfregaço de sangue e certos parasitas.

- Wright: mistura de eosina e azul de metileno, usado para demonstrar corpúsculos sangüíneos e parasitas de malária.

O RNA se cora intensamente em azul pelos corantes básicos (azul de metileno, azuros) e a hemoglobina se cora de rosa pelos corantes ácidos (eosina). O citoplasma basofílico tem tonalidade azul intensa em virtude de seu elevado conteúdo de RNA, enquanto que o citoplasma ortocromático é rosa em virtude de seu conteúdo de hemoglobina.

O maior problema associado à coloração tipo Romanowsky para sangue periférico é a grande variação nas propriedades do corante em diferentes lâminas. Pequenas variações nas cores e intensidades dos leucócitos corados podem ser toleradas quando observada por um técnico altamente treinado, uma vez que decisões de classificação são feitas com base em amplo conhecimento prévio [2].

Capítulo 3

Reconhecimento de Padrões

"A criatividade consiste apenas em perceber o que já está lá. Sabia que os sapatos direito e esquerdo só foram inventados há pouco mais de um século? (Bernice Fitz-Gibbon) "

Padrões estão ligados a "regularidades", são entidades vagamente definidas que podem ser nomeadas ou reconhecidas, como por exemplo, uma impressão digital, um caractere, uma face, um sinal de fala, etc. Reconhecimento de padrões é a categorização do padrão, isto é, o ato de usar atributos do padrão para classificação.

As informações disponíveis são as *características* ou medidas das populações (*classes*) para essa tomada de decisão que não são completamente precisas, fazendo com que os procedimentos de decisão sejam estatísticos por natureza. Dado que haja diferenças entre as populações, considera-se que essas populações têm descrições (*modelos*) diferentes. Antes que se extraia propriedades dos indivíduos (*objetos*), é necessário um *pré-processamento* para simplificar operações subseqüentes sem perda de informação relevante. Então, atributos podem ser extraídos e passados ao *classificador* que avalia a evidência apresentada e dá a decisão final a respeito da classe a que pertence um objeto. O classificador supervisionado depende do conhecimento prévio dos objetos, isto é, das *amostras de treinamento.* Em geral, cada amostra é formada por mais de uma medida (*dimensão do problema*), formando um vetor de características que representa um ponto no espaço de medidas. A precisão de estimativas no espaço de características dependerá dos *limites de decisão*, que determinarão as regras de separação entre as classes e da *generalização* do classificador

frente a novos padrões [13].

No contexto de processamento de imagens, o padrão é geralmente dado por uma representação mais compacta, contendo atributos que juntos são supostos poder sintetizar o que o padrão tem de similar a outros padrões da mesma classe e de dissimilar em relação a padrões de classes diferentes. O reconhecimento de padrões envolve uma vasta coleção de problemas cujas soluções empregam técnicas importadas de outras áreas, como teoria de autômatos, teoria de decisão, teoria da informação, probabilidade, estatística, lógica, etc.

Uma vez que classificar é basicamente a tarefa de recuperar o modelo que gerou o padrão, técnicas de classificação diferentes se adequam melhor a diferentes modelos, portanto processos de decisões são fundamentalmente específicos para certas tarefas, ou seja, criar um dispositivo de reconhecimento de padrões de propósito geral é impossível até o momento.

O objeto de interesse neste corrente projeto é a imagem, cuja manipulação pode ser dividida em três etapas dependendo da entrada ou saída da etapa:

1. processamento: imagem $\rightsquigarrow$ imagem (Seções 3.1 - 3.2);

2. análise: imagem $\rightsquigarrow$ medidas (Seção 3.3);

3. entendimento: medidas $\rightsquigarrow$ descrição de alto nível (Seções 3.4 - 3.6);

As próximas seções serão dedicadas à explicação dos vários passos no reconhecimento de padrões, começando com o processamento de imagens na Seção 3.1. Embora conceitualmente parte do processamento de imagem [14], a segmentação será descrita separadamente na Seção 3.2. Os atributos das imagens de leucócitos serão detalhados na Seção 3.3.

3.1 Processamento de imagens

O uso de técnicas de processamento para melhoria da imagem e/ou simplificação da informação para etapas posteriores é importante na pesquisa em imagens médicas. O processamento de imagem denota basicamente as seguintes manipulações [15]:

- melhoria da imagem, através de operações que melhorem a aparência para um visualizador humano ou operações que convertam uma imagem para um formato mais apropriado para o tratamento digital;
- restauração da imagem, que visa a modificação da imagem para compensar os defeitos do sistema que produziu a imagem;
- modificação geométrica da imagem, como magnificação, redução, rotação, etc.

Um exemplo de processamento é a redução do número de níveis de cinza (ou cor) possíveis na imagem como forma de simplificar a informação contida na imagem. Dá-se o nome de quantização a esse processo, o qual será discutido na seção a seguir.

3.1.1 Quantização

A eficiência e desempenho de algoritmos que consideram os níveis de cinza (ou cores) das imagens podem requerer uma redução drástica no número de níveis de cinza (ou cores) a ser usada na representação. Existem duas formas de se discretizar uma imagem: amostrando os pontos espacialmente ou o nível de cinza (ou cor) de cada pixel [16]. Geralmente, reduz-se o número de níveis de cinza ou cores através da quantização do espaço, usando uma paleta pré-definida (quantização estática) ou por agrupamento e/ou segmentação espacial (quantização dinâmica). Formalmente, suponha que C seja o espaço de cores e P, um subconjunto de C chamado espaço de quantização dado na Eq.3.1.

$$P = c_1, c_2, ..., c_i, ..., c_n \mid c_i \in C, n \ll \|C\| \tag{3.1}$$

Uma função Q que mapeia cada cor em C a cada elemento em P é chamada de quantizador e é definida na Eq.3.2:

$$Q : C \rightarrow P \tag{3.2}$$

Equalização de histograma

Variações na luz, lente, filme e digitalizador normalmente introduzem transformações monotônicas em imagens em níveis de cinza. A normalização de imagens através

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

de quantização com equiprobabilidade[1] é um tipo de equalização de histograma que garante que imagens que sejam uma transformação monotônica de uma outra produzam os mesmos resultados por apresentarem as mesmas características. O algoritmo é apresentado detalhadamente por Haralick [3].

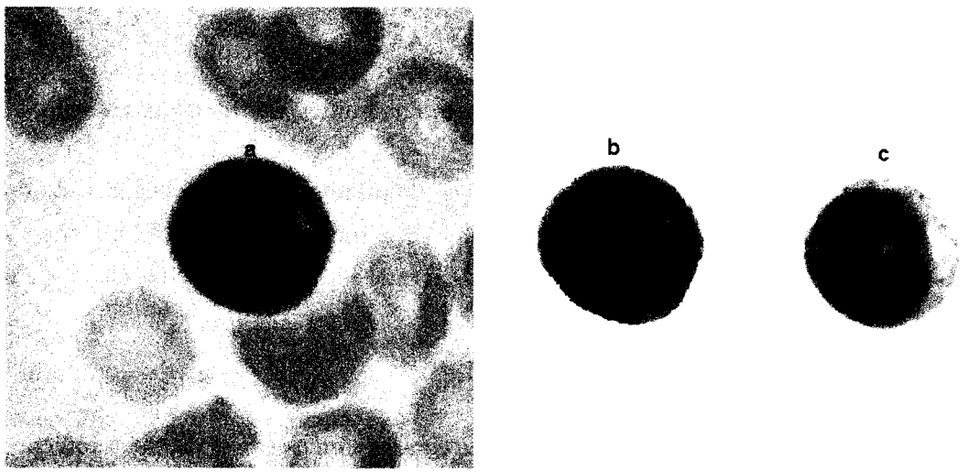


Figura 3.1: (a) Imagem colorida , (b) sua versão em nível de cinza e (c) imagem nível de cinza quantizada e equalizada segundo algoritmo de Haralick [3]

Esse algoritmo foi utilizado no estudo de textura do núcleo e do citoplasma dos leucócitos nessa tese. A Fig.3.1 ilustra o processamento que se realizou para obter atributos de textura. Após a etapa de segmentação a partir da imagem colorida (Fig.3.1.a), tem-se a separação do leucócito do resto da imagem e suas respectivas regiões (núcleo e citoplasma). Então, considera-se a versão em níveis de cinza (Fig.3.1.b) dessas regiões e aplica-se a equalização (Fig.3.1.c) para o estudo de textura usando matrizes de coocorrência, conforme proposto por Haralick [3, 17]. A extração de medidas de textura será detalhada na seção 3.3.3.

Além da quantização, processamentos por meio de filtros permitem melhorar a imagem pra extração de informação posteriormente.

3.1.2 Filtragem

Em um processo de filtragem, cada pixel da imagem resultante é uma geralmente uma função desse pixel e de seus pixels adjacentes. O objetivo é obter uma imagem

[1]Tradução de *equal-probability quantizing*.

resultante mais adequada que a original, reduzindo-se o ruído e/ou preparando a imagem para processamentos seguintes. Através da filtragem, características interessantes da imagem, como bordas podem ser ressaltadas ou peculiaridades indesejáveis suprimidas, como o excesso de textura.

Pode-se dividir os filtros em lineares e não lineares. Filtros lineares são aqueles que preservam a estrutura do espaço vetorial no espaço de imagens, ou seja, são invariantes perante a multiplicação por um escalar e a soma. Para filtros lineares, a implementação é uma convolução, ou seja, uma soma ponderada dos pixels da vizinhança do pixel em questão. Exemplos de filtros lineares são: média, gradiente e o laplaciano. Um dos filtros não lineares mais importantes é o filtro de mediana, usado para eliminar ruído sem introduzir níveis de cinza que não existam na imagem original. Outros exemplos de filtros não lineares são o Sobel, filtros morfológicos e difusão anisotrópica. Alguma referências interessantes são: [16, 18, 15, 19]. A seguir, descreve-se os filtros utilizados nessa tese.

Morfologia matemática binária

A teoria de morfologia matemática permite filtrar imagens binárias e em nível de cinza para realçar, segmentar, detectar bordas e/ou esqueletos, afinar, avaliar granularidade, etc. Tem por princípio básico extrair informação da geometria e topologia da imagem, dado um conjunto de pontos completamente definido, chamado elemento estruturante, cuja forma e tamanho são responsáveis por gerar transformações na imagem. As operações morfológicas devem respeitar princípios fundamentais como invariância à translação, invariância à mudança de escala e considerar o conhecimento local, ou seja, o conhecimento da região onde será aplicado o elemento estruturante é suficiente para executar a transformação.

As operações básicas da morfologia matemática são [20]:

a) erosão: a erosão de um conjunto de pixels X (imagem binária) por um elemento estruturante B resulta no conjunto Y, ou seja, o conjunto de translações de B sobre X onde todos os elementos de B devem estar em X, onde B_x indica o elemento estruturante centrado em x.

$$Y = x \in X : B_x \subset X$$

b) dilatação: a dilatação é a operação inversa da erosão, definida pelo conjunto das translações de B sobre X sempre que ao menos um elemento de B estiver em X. Essa operação se caracteriza por preencher todos os espaços da imagem com tamanho inferior ao elemento estruturante da imagem original, podendo também ligar elementos isolados.

$$Y = x \in X : B_x \cap X \neq \varnothing$$

c) abertura: é uma operação morfológica definida a partir da erosão, seguida de uma dilatação.

d) fechamento: é a operação inversa da abertura, definida a partir de uma operação de dilatação seguida de uma erosão.

Diferente da ordem adotada nessa tese, as operações morfológicas foram realizadas na fase pós-segmentação (Seção 3.2). Sucessivas operações de erosão permitiram eliminar plaquetas, regiões normalmente segmentadas como fazendo parte do núcleo devido à semelhança em termos de cor. Entretanto, a seção a seguir trata de um filtro a ser aplicado antes do processo de segmentação pois, em teoria, garante maior homogeneização das regiões, preservando bordas.

Difusão anisotrópica

A suavização de altas frequências mais comum é aquela feita por filtros gaussianos com objetivo de homogeneizar a imagem. Assim, através da convolução da imagem original ($I_o(x, y)$) com uma gaussiana com parâmetro de escala t (Eq. 3.3), é possível gerar imagens com diferentes resoluções em termos de detalhes. Esta abordagem tem um incoveniente principal: é difícil obter a localização precisa das bordas semanticamente significativas. Foram testados filtros gaussianos para o problema de reconhecimento de leucócitos, mas a perda de bordas compromete o reconhecimento dos leucócitos, tanto considerando o contorno do núcleo quanto o contorno do citoplasma.

$$I(x, y, t) = I_o(x, y) * G(x, y; t) \tag{3.3}$$

Aplicando a definição de processos de condução de calor, define-se o filtro não-linear conhecido como difusão anisotrópica, usado para redução de ruídos na imagem

ou mesmo de detalhes (ex: textura) que não devem ser considerados no processo de segmentação da imagem. Segundo Perona e Malik [21], foi Koenderik quem verificou que a convolução do sinal com as gaussianas definida em várias escalas é semelhante a solução da equação do calor com o sinal original como condição inicial. Algoritmos de difusão suavizam regiões homogêneas de acordo com uma equação diferencial parcial mostrada em Eq.3.4:

$$I(x,y,t) = \Delta I = I_{xx} + I_{yy} \tag{3.4}$$

com condição inicial $I(x,y,0) = I_o(x,y)$, que é a imagem original. Mudando do paradigma de espaço-escala linear para a abordagem usando difusão anisotrópica, a suavização é privilegiada dentro das regiões ao invés das bordas, portanto se define um coeficiente de condução ($c(x,y,t)$) variável em função da borda da imagem. Escolhe-se $c(x,y,t)$ como uma função do gradiente. É evidente que não se sabe onde estão as bordas a priori, porém o gradiente da imagem é uma estimativa suficiente para "proteger" a borda no processo de suavização por difusão anisotrópica. A imagem final, dada a escala t, é obtida segundo equação Eq.3.5.

$$I(x,y,t) = div(c(x,y,t)\nabla I) = c(x,y,t)\Delta I + \nabla c.\nabla I \tag{3.5}$$

Na Eq.3.5, div é o divergente e ∇ e Δ são os operadores de gradiente e laplaciano, respectivamente.

Retinex

Em visitas a grupos de pesquisa em visão computacional na Itália, particularmente o Instituto de Tecnologia de Informação e Multimedia (ITIM) de Milão, micrografias de leucócitos foram filtradas utilizando o algoritmo de normalização de imagens RGB Retinex [22]. Foi sugerida a aplicação desse algoritmo como forma de generalizar os treinamentos para segmentação de leucócitos bem como tornar o sistema de análise de imagem mais independente do tipo de imagem de treinamento.

O retinex foi proposto por Land em 1964 [23], cujo nome provém da fusão das palavras retina e córtex, que propõe que, no sistema visual, os registros de luz obtidos simultaneamente em 3 comprimentos de onda diferentes são comparados para compor a cor da superfície. Esta comparação é independente do comprimento

de onda da composição da luz incidente e, portanto, não considera a intensidade relativa às luzes de diferentes comprimentos de onda.

Land sugeriu um algoritmo[2] que simula a constância de cores em conjunto com outro pesquisador, McCann [22]. Calcula-se a razão entre o logaritmo de luz de um dado comprimento de onda refletida por uma superfície (o numerador) pela média da luz desse comprimento de onda, refletido pela vizinhança (o denominador). O processo é feito independentemente três vezes para os três comprimentos de onda (R,G e B). Essa técnica explora a distribuição espacial (vizinhança) das cores, separando-se o efeito da iluminação do reflexo. Em geral, comporta-se como um passa-alta: ressalta bordas salientes e exerce influência a distâncias arbitrárias, característica algumas vezes inconveniente pois pode ressaltar ruído. Essa técnica simula os operadores biológicos do sistema visual, que somam, diferenciam e retificam sinais de entrada para obter interações espaciais, comparando o pixel com sua vizinhança. A luminosidade máxima é considerada como valor de referência para comparação com outros pixels da imagem.

O Apêndice C mostra alguns resultados utilizando o algoritmo retinex para leucócitos, embora não tenha sido adequado para leucócitos por alterar o citoplasma com perdas. Portanto não foi adotado no software desenvolvido.

3.2 Segmentação

Segmentação de uma imagem é a designação que se dá à separação ou divisão da imagem em regiões significativas segundo algum atributo, que tem forte correlação com objetos ou regiões do mundo real contidas na imagem. A segmentação é uma etapa essencial na análise de imagem, um passo chave no sentido de organizar o arranjo de pixels da imagem em regiões que correspondem a entidades significativas na cena [24]. O objetivo da segmentação pode ser a separação completa do objeto ou parcial, quando as regiões não correspondem diretamente aos objetos da imagem.

O atributo mais básico para segmentação é a luminosidade, segundo a qual diversos problemas de segmentação podem ser resolvidos com sucesso usando somente processamentos simples como a limiarização. Neste caso, a imagem geralmente con-

[2] http://www.cs.sfu.ca/ contour/publications/IST-2000

siste de objetos contrastantes localizados num fundo uniforme. Tal processamento é independente de contexto. Assim, nenhum modelo relacionado ao objeto é usado e nenhum conhecimento sobre o resultado de segmentação esperado é suposto.

Os métodos de segmentação podem ser divididos em três grupos, de acordo com a representação da característica considerada [14]:

- segmentação baseada em **conhecimento global** da imagem ou de suas partes, onde o conhecimento é representado pelo histograma das características da imagem. Ex: limiarização;
- segmentação baseada em **borda** da imagem, onde somente o contorno é extraído;
- segmentação baseada em **região**, onde propriedades (cor, textura, etc.) de cada região independente das outras são consideradas na segmentação dos objetos;

Em geral, a segmentação por borda e por região usam critérios **locais** de pré-processamento, localizando mudanças abruptas nas intensidades ou cores da imagem. Empregou-se o termo *segmentação* indiscriminadamente nessa tese para os resultados de métodos que visam separar objetos, tenham eles segmentado a imagem com sucesso ou requerido pós-processamentos.

3.2.1 Limiarização

A limiarização da imagem em níveis de cinza é o processo de segmentação mais simples, barato e rápido, sendo o método mais antigo ainda largamente utilizado em aplicações simples. Isso se deve ao fato de que muitos objetos ou regiões da imagem são caracterizadas por uma refletividade ou absorção de luz constantes em suas superfícies.

Nos casos em que a limiarização é útil na segmentação do objeto de interesse, o valor de limiar é crucial para uma segmentação bem sucedida; esta seleção pode ser determinada pelo usuário interativamente ou através de algum método.

O limiar é global quando determinado a partir da imagem inteira ou local quando dependente da posição na imagem, técnica interessante em casos de iluminação não

uniforme. Métodos mais complexos de detecção de limiar são baseados em análise da forma do histograma. Um exemplo é o método que procura mínimos locais entre dois ou mais picos de histogramas bimodais, trimodais, etc. Estes métodos baseados na forma do histograma são bastante sensíveis ao ruído e frequentemente exigem etapas de pré-processamento para suavização do histograma.

Também é possível considerar a limiarização por banda (*band-thresholding*), onde subconjuntos determinados por intervalos de níveis de cinza contém os pixels do objeto. Sonka et al [14] argumentam a utilidade dessa abordagem na segmentação de imagens microscópicas de células do sangue, onde um intervalo particular de níveis de cinza representa o citoplasma, o fundo é mais claro e o núcleo da célula mais escuro. Contudo, a variação de luminosidade entre as imagens impede que esses limiares sejam calculados automaticamente com precisão.

A Fig.3.2 apresenta a imagem de um leucócito, o histograma geralmente bimodal e a segmentação do núcleo obtida por limiarização global da imagem através do histograma para limiar na intensidade 135, respectivamente.

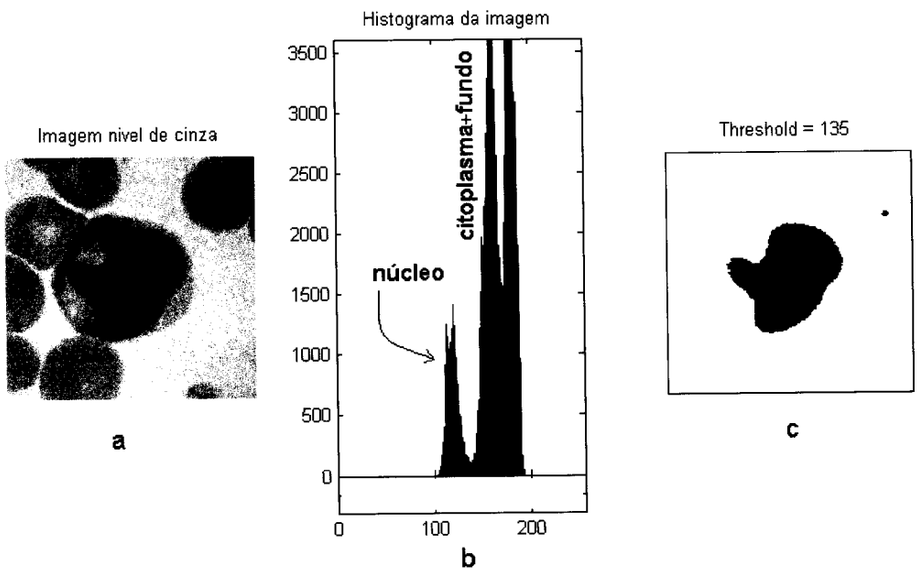


Figura 3.2: Segmentação do núcleo de leucócito usando limiar em histograma de intensidades. a) imagem em níveis de cinza; b) histograma da imagem; c) imagem limiarizada para valor do pixel menor que 135.

Quando o histograma é bimodal e a informação do objeto de interesse se con-

centra em torno de uma das modas, um único limiar é suficiente para separação do objeto de interesse (ex: o núcleo da célula) do fundo, porém a limiarização pode ser impossível se o contraste entre o objeto de interesse e o fundo não for grande. A dificuldade de se encontrar um limiar ótimo pode ser superada se o valor do limiar for aquele que maximiza a variância entre objetos e fundo. Existe uma ampla variedade de algoritmos para determinação de limiar devido à popularidade dessa ferramenta na segmentação de imagens [14].

A seleção iterativa de limiar é considerada ótima, contudo altamente custosa sob o ponto de vista computacional. Quando há conhecimento sobre as propriedades da imagem e/ou da forma do histograma a priori, métodos mais simples podem ser adotados, obtendo-se resultados interessantes. A seguir, apresentaremos técnicas que utilizam o conhecimento prévio que se tem das imagens das células do sangue.

Limiarização do canal G

Depois de tentar a limiarização por banda, em imagens em níveis de cinza, para dividir imagens em núcleo, citoplasma e fundo como proposto no trabalho de Sonka et al [14], Katz [25] observou que esse tipo de segmentação era efetiva somente para um único glóbulo branco na imagem ou mesmo para um número pequeno. Outro transtorno foi que os níveis de cinza do citoplasma e fundo (particularmente na presença de hemácias) tornam a técnica quase sem uso quando considerando um amplo conjunto de imagens.

Ainda que usando limiarização para segmentação, Katz testou os canais de cromaticidade RGB e verificou que o canal G limiarizado permitia a segmentação da região nuclear. A banda G carrega grande parte da informação do material nuclear; essa região se caracteriza pela cor púrpura, ou seja, variações de magenta, que é a combinação de valores altos para o canal R e B e valores baixos para G. Foi verificado que um valor de limiar fixo igual a 100 numa escala de 0-255 dava uma discriminação aceitável entre os pixels nucleares e não-nucleares. Testou-se esse método com sucesso em imagens coletadas que apresentassem um histograma bimodal, com informação do núcleo concentrada nos valores de níveis de cinza inferiores a 100.

Os resultados utilizando limiarização pelo canal G foram importantes pois per-

mitiram delimitar a região da imagem que contém o leucócito uma vez que o núcleo sempre é circundado pelo citoplasma da célula. Assim, a partir de imagens com dimensão de 760 por 480 pixels, foi possível redimensionar a imagem para casos de até 200 por 200 pixels.

O limiar fixo apresenta limitações pois a intensidade luminosa pode variar deslocando o histograma para a esquerda (menos luminosidade) ou direita (mais luminosidade). Portanto, o próximo passo foi encontrar um algoritmo de limiarização que se adaptasse à imagem, retornando o melhor limiar em função do histograma da matriz referente ao canal G de micrografias coloridas.

Limiarização por Entropia

A aplicação da medida de entropia na área de processamento de imagens provém do conceito introduzido por Shannon na formulação da teoria da informação. Clausius em 1854, atribuiu o termo grego entropia (evolução ou transformação), para entender a natureza do calor e a irreversibilidade em sistemas termodinâmicos, baseando-se no trabalho de Carnot (1824) e Kelvin (1850) [26]. Esse trabalho resultou na formulação da segunda lei da termodinâmica ($dS = dq/T \geq 0$) na qual dq representa o fluxo de calor e T a temperatura, que resumidamente afirma que a entropia (S) sempre cresce em sistemas reais [27].

A entropia é uma propriedade macroscópica de sistemas (tal como volume e tamanho). No entanto, Boltzmann procurou uma explicação estatística reduzindo o conceito de entropia para níveis moleculares e chegou à forma $S = klnW$ (onde W é o número de estados possíveis do sistema e k é a constante[3] de Boltzmann). Shannon (1948)[4], aproveitando a idéia de Boltzmann, foi quem definiu uma medida de quantidade de informação dada pela ocorrência de um símbolo em uma mensagem. Assim, quanto mais raro um símbolo em uma mensagem, maior seria sua contribuição para a quantidade de informação passada pela mensagem [26].

A entropia de Shannon é a medida de informação contida em uma mensagem em oposição a porção da mensagem que é estritamente determinada. Na teoria de informação, entropia é conceitualmente a quantidade real de informação em um pedaço de dado. Dados inteiramente aleatórios em byte tem uma entropia próxima a

[3] k é a constante de Boltzman e vale $1.380658 * 10^{-23} Joules/Kelvin$

[4] http://planetmath.org/encyclopedia/ShannonsTheoremEntropy.html

infinito, uma vez que o próximo caractere é desconhecido. Uma cadeia de caracteres composta somente pela letra A tem entropia zero, pois o próximo caractere sempre será um A. A entropia de um texto em inglês está por volta de 1.5 bits por caractere. A taxa de entropia de uma fonte de dados é calculada por meio do número médio de bits por símbolo necessário para codificá-lo.

Definindo W como a multiplicidade de um sistema e pensando esse sistema como imagens, define-se a Eq.3.6:

$$W = \frac{N!}{\prod_i N_i!} \tag{3.6}$$

Seja N o número de pixels em uma imagem, $N!$ o número de imagens possíveis com N pixels e $N_i!$ o número de possíveis configurações dos pixels com nível de cinza i em uma imagem, W indica todas as imagens possíveis com N pixels para um número de pixels com valor i, N_i. Para se perceber o acréscimo de informação (I) de uma maneira aditiva e não multiplicativa, utiliza-se o logaritmo da medida W multiplicado por uma constante k, definida a seguir.

$$I = k * ln(W) \tag{3.7}$$

Segundo Shannon, a entropia de informação (S) indica a quantidade de informação por símbolo, cuja fórmula origina-se da substituição da Eq.3.6 em 3.7, então $S = I/N$, segundo formulação da Eq.3.8:

$$S = k\frac{1}{N}ln(\frac{N!}{\prod_i N_i!}) = k\frac{1}{N}(ln(N!) - \sum_i ln(Ni!) \tag{3.8}$$

Usando a aproximação de Stirling para $N!$ com N suficientemente grande, $ln(N!) \approx N(ln(N) - 1)$, chega-se na Eq. 3.9:

$$S = -k\sum_i \frac{N_i}{N} * ln(\frac{N_i}{N}) \tag{3.9}$$

Note que $p_i = N_i/N$ é a probabilidade de que o pixel i seja encontrado na

imagem, portanto pode-se escrever 3.9 como:

$$S = -k \sum_i p_i * ln(p_i) \tag{3.10}$$

Para a teoria da informação, a constante k pode ser arbitrariamente escolhida, então se adota k como a quantidade de bits, fixando a quantidade de informação ($I = N$) com $W = 2^n$ configurações possíveis. Daí, $k = 1/ln(2)$ ser usado para medir a quantidade de informação contida em uma imagem em bits, que pode ser escrito como 3.11:

$$S = -\sum_i p_i * log_2(p_i) \tag{3.11}$$

Métodos entrópicos não determinam o limiar ótimo para a binarização através da forma do histograma, contudo usa o histograma como uma aproximação da distribuição de probabilidades. O cálculo de entropia foi proposto por Pun, posteriormente Kapur et al [28] evoluiu as considerações teóricas e práticas do método inicial proposto e desenvolveu um eficiente método de binarização entrópica: a soma de entropias, que utiliza a propriedade de adição da entropia. Se consideramos a imagem digital como um sistema que pode ser decomposto em dois subsistemas A e B estatisticamente independentes, a entropia do sistema apresenta a propriedade da aditividade de forma que:

$$S = S_A + S_B \tag{3.12}$$

O histograma normalizado dá a frequência relativa à ocorrência dos níveis de cinza para o conjunto total de pixels da imagem. Os pixels da imagem podem pertencer ao "objeto"e ao "fundo", duas classes estatisticamente independentes, componentes de um único sistema (a imagem digital), cujos eventos são os valores de luminância dos pixels. Assim, para o número de níveis de cinza considerados, ou seja, para n eventos define-se entropia conforme Eq.3.11. Para um sistema composto por dois subsistemas A e B, dependentes do parâmetro t (t=*thresholding* ou limiar), define-se dois conjuntos de probabilidades normalizados:

$$p_A \ : \ \frac{p_1}{P_t}, \frac{p_2}{P_t}, ..., \frac{p_t}{P_t} \tag{3.13}$$

$$p_B \ : \ \frac{p_{t+1}}{1-P_t}, \frac{p_{t+2}}{1-P_t}, ..., \frac{p_n}{1-P_t} \tag{3.14}$$

onde $P_t = \sum_{i=1}^{t} p_i$ e t é o limiar que "separa"a imagem em 2 classes.

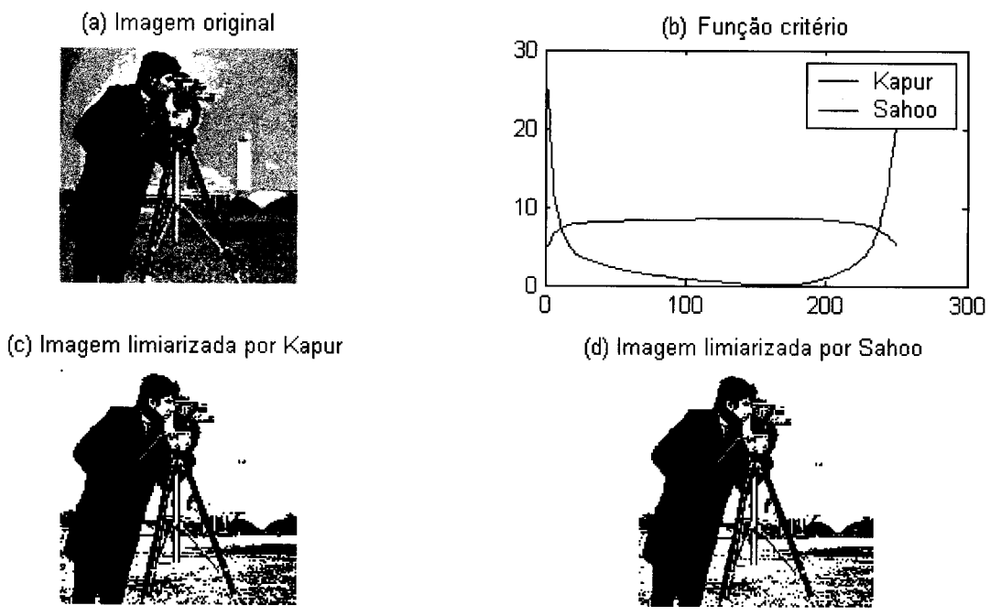


Figura 3.3: Limiarização entrópica: (a)imagem original em nível de cinza; (b) funções de entropia relativa; (c) e (d) resultados da limiarização usando limiar obtido por Kapur e Sahoo, respectivamente.

Podemos escrever $S_t = -\sum_{i=1}^{t} p_i log p_i$ de modo análogo à Eq.3.11, que depende de t e é definida como a **função critério** para seleção do nível ótimo para a binarização das imagens. O valor discreto de t que maximiza a função $S(t)$ é considerado o nível ótimo de binarização. Uma outra abordagem, proposta por Sahoo [27], seria tomar como limiar o valor que minimiza a diferença quadrática da entropia dos 2

subsistemas para diferentes valores de t, conforme Eq.3.15.

$$S(t) = (S_A - S_B)^2 \tag{3.15}$$

A Fig.3.3 ilustra resultados do algoritmo de determinação de limiar segundo função critério definida por Kapur e Sahoo. O resultado da binarização é semelhante para ambos os métodos adotados e a avaliação sobre a eficiência desses métodos, para a imagem do fotógrafo, é subjetiva, uma vez que depende do contexto de utilização. Por exemplo, na avaliação de células, deseja-se o núcleo, portanto medidas quanto a taxa de acerto dos pixels do núcleo pode ser utilizada para avaliação quantitativa [29].

Limiarização pelo Logaritmo da Acumulada (LogCum)

Uma abordagem interessante para determinação do limiar automaticamente foi notado durante o desenvolvimento do corrente projeto de pesquisa. Através de uma técnica não supervisionada, considera-se a forma do histograma para limiarização global.

Diferente dos algoritmos baseados em entropia, o algoritmo original é calculado a partir de uma transformação da função de distribuição acumulada (P) mostrada na Eq.3.16, gerando a função critério (f). Determina-se o limiar ótimo (T) minimizando-se o valor de f conforme a Eq.3.17, onde g indica os níveis de cinza possíveis.

$$f(g) = log(P(g))^{P(g)} = P(g).log(P(g)) \tag{3.16}$$

$$T = argmin(f(g)) \tag{3.17}$$

Esse algoritmo ainda não possui fundamentação matemática e surgiu, curiosamente, a partir de uma modificação no cálculo do algoritmo de Kapur. Embora os resultados sejam comparáveis, o método proposto é de menor custo computacional e mais robusto conforme testado para várias imagens. Aqui mostramos um teste usando uma imagem em níveis de cinza (Fig. 3.4) e a informação do canal G de uma imagem colorida (Fig.3.5).

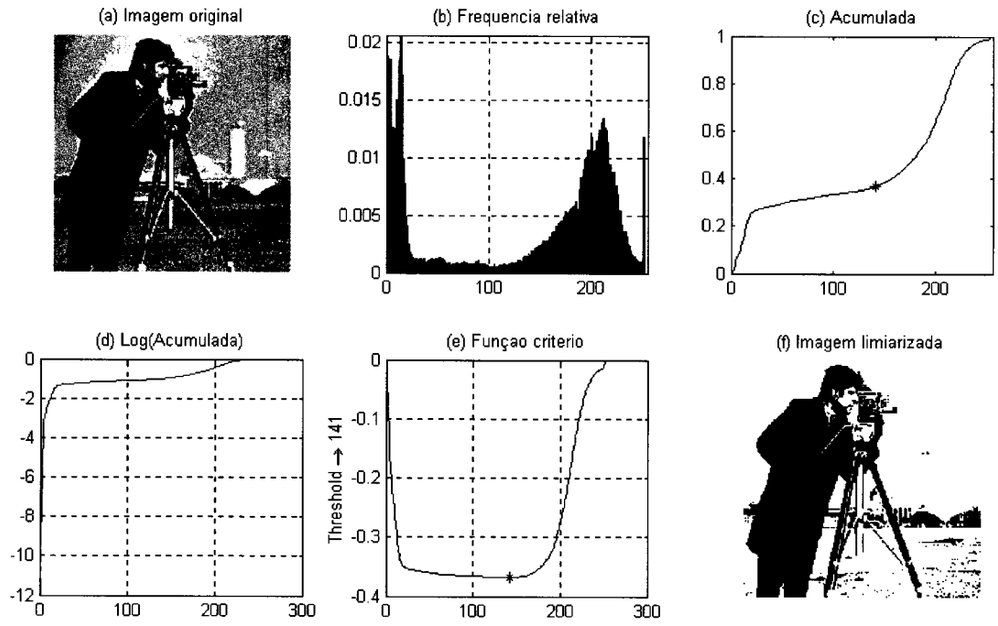


Figura 3.4: Limiarização pelo método LogCum: (a) imagem original em nível de cinza; (b) histograma normalizado de (a); (c) acumulada de (b); (d) logaritmo da acumulada apresentada em (c); (e) função critério f, resultado da multiplicação de (c) por (d); (f) resultado da limiarização.

Tanto a Fig.3.4 quanto a Fig.3.5 mostram exemplos em que a função critério tem um único mínimo e apenas um mínimo foi encontrado para imagens de sangue, o qual foi tomado como valor de limiar, obtendo-se resultados de limiarização satisfatórios. Entretanto mais testes devem ser feitos para se entender melhor o método proposto originalmente. Mais exemplos são mostrados no Apêndice B, ilustrando os métodos entrópicos em comparação com o método proposto.

3.2.2 Segmentação por Borda

A segmentação das bordas da imagem descreve mudanças de funções contínuas usando derivadas, geralmente o gradiente e o laplaciano da imagem [15]. Como a imagem é uma função que depende de duas variáveis, as coordenadas do plano da imagem, os operadores que descrevem as bordas (ou detectores de bordas) são expressos usando derivadas parciais. A borda é uma propriedade ligada a um pixel

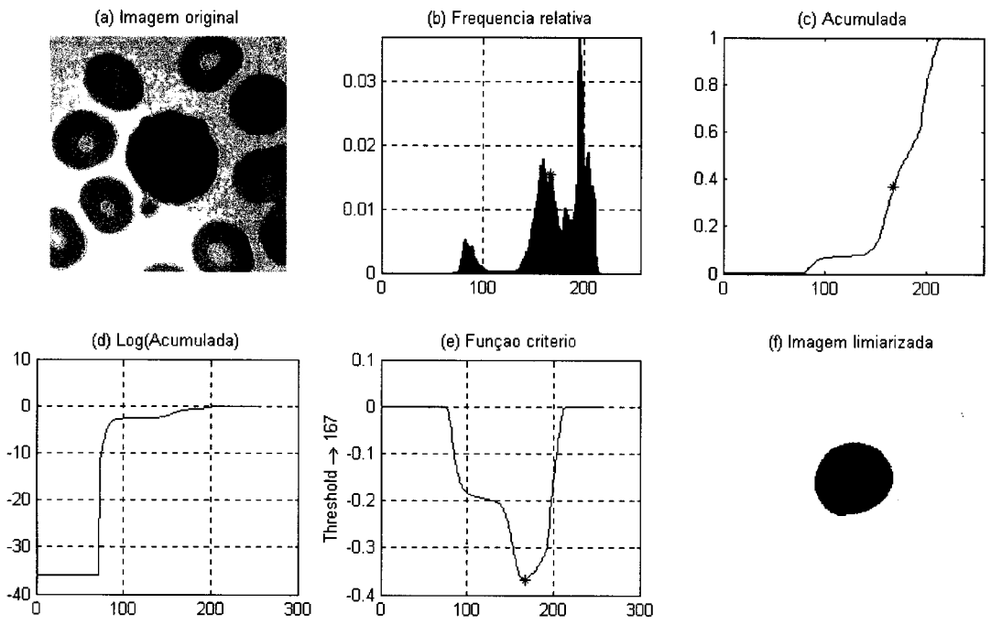


Figura 3.5: Segmentação do núcleo de basófilo pelo LogCum usando informação do canal G; (c) acumulada de (b); (d) logaritmo da acumulada apresentada em (c); (e) função critério f, resultado da multiplicação de (c) por (d); (f) resultado da limiarização.

individual e calculada a partir de uma certa vizinhança deste pixel [14].

Vários detectores podem ser empregados para definição dos limites dos objetos como mostrado na Fig.3.6. O maior problema na análise de bordas dos objetos de interesse é a descontinuidade do contorno e o ruído gerado pelos métodos, como é melhor evidenciado na segmentação do leucócito na Fig.3.6. Geralmente, o resultado da detecção de borda não pode ser usado como resultado de segmentação, assim pós-processamentos devem ser combinados para se ter, no mínimo, uma segmentação parcial significativa. Muitos trabalhos foram dedicados à segmentação de imagens em níveis de cinza por borda aplicadas a leucócitos [30, 25, 31, 32, 33], entretanto os ajustes necessários na fase pós segmentação são muito custosos usando tal abordagem.

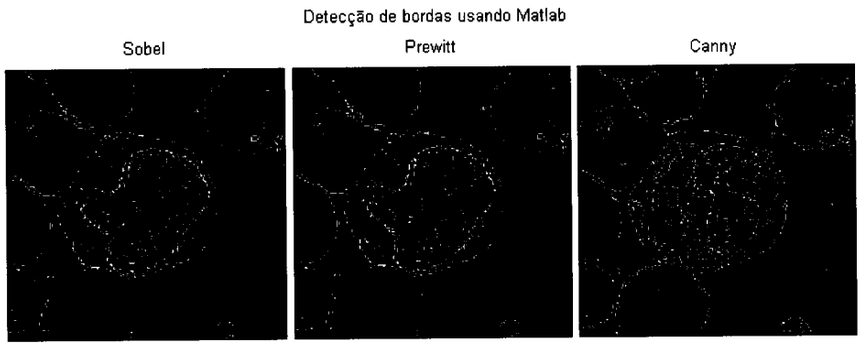


Figura 3.6: Bordas do leucócito da Fig.3.2 para diferentes técnicas - à esquerda: Sobel, ao centro: Prewitt e 'a direita: Canny - descontinuidades do contorno e ruído

3.2.3 Segmentação por cor

A segmentação por cor é um tipo de segmentação por região, que pode ou não considerar conexidade. Como utiliza propriedades da imagem, os pixels vizinhos com amplitudes similares tendem a ser agrupados para formar uma região segmentada [15]. A homogeneidade é o principal critério em algoritmos que visam dividir a imagem em zonas de máxima semelhança. Ao invés de uma escala para amplitude ou nível de cinza, imagens coloridas são segmentadas a partir de vetores contendo as coordenadas no espaço de cor, por exemplo, o RGB. Algoritmos de segmentação de imagens coloridas se desenvolveram muito na década de 90, sendo que na maioria das vezes são uma espécie de extensão dimensional de algoritmos para imagens em nível de cinza. Lucchese e Mitra [19] apresentam um interessante levantamento bibliográfico a respeito de segmentação por cor, que pode ser sintetizado por meio dos comentários a seguir. Sugere-se a divisão dos algoritmos em três grupos de técnicas baseadas em:

a) espaço de características: os pixels da imagem são mapeados em algum espaço de cor, esperando-se que diferentes objetos da imagem se manifestem em diferentes nuvens de pontos. A localização do pixel é perdida, ou seja, negligencia-se a relação espacial entre eles. As técnicas mais populares para segmentação usando espaços de características são por agrupamento (*clustering*) e por histograma. Esses algoritmos operam exclusivamente em alguns espaços de características, resultando em regiões homogêneas com respeito às características representadas nestes espaços.

Contudo, não há garantia de que estas regiões apresentem conectividade na imagem original, o que ocorre principalmente quando utilizando as coordenadas R,G e B; a abordagem por espaço de característica para segmentação de regiões por cor não garante que pixels adjacentes influenciem no julgamento do pixel em questão, embora tal requisito possa ser incorporado ao algoritmo.

b) domínio da imagem: considerando o domínio da imagem, busca-se tanto por homogeneidade das características da região quanto pela conectividade ao mesmo tempo. Essas propriedades são garantidas através dos seguintes métodos:

- subdivisão e fusão das regiões da imagem (técnicas divide-e-funde[5]). Um exemplo é o uso de diagramas de Voronoi para particionar a imagem e mapas de Kohonen para fundir;
- crescimento de regiões, onde a partir de sementes geram-se regiões conectadas, que em seguida são fundidas segundo algum critério de similaridade, por exemplo as distribuições de cor da região;
- técnicas baseadas em borda: ex. funções de gradiente para imagens coloridas, contornos ativos ou *snakes*;
- classificação baseada em redes neurais, que permite alto grau de paralelismo, computacionalmente rápida, mas precisa de fase de aprendizado e conhecimento a priori;

c) física: todos os algoritmos anteriores estão suscetíveis a erros se os objetos da imagem colorida são afetados por variações luminosas, sombreamentos ou sombras, daí estes algoritmos retornarem imagens super segmentadas geralmente. A única maneira de superar esses entraves é analisar como a luz interage com materiais coloridos e introduzir modelos desta interação física nos algoritmos de segmentação.

A respeito dos espaços de cores, estes podem ser divididos basicamente em sistemas baseados em dimensões perceptuais (Lab, HSV, etc.) e não perceptuais de cores (RGB, XYZ, etc.). O sistema Lab da CIE (Comission Internacionale L'Eclairage) é

[5]Tradução de *split-and-merge*.

independente dos dispositivos em que está sendo exibido. Isso já não é verdade para o RGB, que depende do dispositivo em uso. No modelo Lab, L está relacionado à luminância, a e b correspondem a, respectivamente, crominância vermelha/azul e amarelo/azul. Maiores detalhes sobre técnicas de transformação entre espaços de cor, sugere-se [34, 15], além do endereço:
`http://www.dca.fee.unicamp.br/~alexgs/disciplinas/pi/grupo6/lab2/exercicio_4/exercicio_4.html`.

3.3 Extração de características: medindo as ROIs

As características da imagem podem ser divididas em três tipos: baseada na forma, na cor ou na textura. As medidas consideradas a seguir foram adotadas baseadas no estudo do leucócito, conforme descrições apresentadas no capítulo 2.

Através de coleta de dados em estudo bibliográfico e junto aos médicos hematologistas do FMRP-USP, observou-se que tanto parâmetros de forma quanto textura e cor eram importantes no reconhecimento e diferenciação das células do sangue. Nas próximas seções define-se os atributos utilizados no corrente projeto de pesquisa, sendo que a utilização deles e desempenho do classificador só serão discutidos no capítulo de resultados.

3.3.1 Forma

Forma é uma propriedade dos objetos que permite descrevê-los em termos de sua representação externa (contorno) e interna (região) [14]. Dentre os vários atributos de forma existentes [16], aqueles que foram utilizados no corrente projeto de pesquisa são: área (A), área do maior objeto, número de elementos ou partes constituintes, circularidade, descritores topológicos (número de furos do objeto), área dos furos, perímetro (P), número de pontos côncavos, convexos e energia de dobramento do contorno do núcleo. Uma outra medida foi a razão núcleo-citoplasmática (NC), dada pela razão entre a área do núcleo e soma das áreas do núcleo e citoplasma (Eq. 3.18)

$$NC = \frac{A_{nucleo}}{A_{nucleo} + A_{citoplasma}} \tag{3.18}$$

Circularidade

Características que expressem circularidade são interessantes uma vez que os leucócitos são razoavelmente redondos. Portanto a medida de compacidade padrão foi usada e é definida pela Eq.3.19. A compacidade atinge seu mínimo de 4π para regiões que sejam círculos perfeitos. Em geral, células pequenas como os linfócitos e neutrófilos têm valor de circularidade próximos do mínimo, enquanto células maiores, contendo mais protrusões, são menos redondas [35].

Apesar dos esforços para segmentação do citoplasma, os resultados obtidos não foram suficientemente precisos, portanto se calculou a circularidade apenas do núcleo.

$$C = \frac{P^2}{A} \tag{3.19}$$

Contorno

Uma vez determinadas as ROIs, o contorno dos objetos pode ser definido usando uma representação paramétrica, ou seja, definem-se parâmetros para o contorno como ponto inicial (arbitrário), ponto final e sentido de travessia sobre o contorno (horário ou anti-horário) [16]. Durante a extração do contorno é válido considerar se a curva é aberta ou fechada e se o que se deseja é o contorno interno ou externo. Há diversos algoritmos para extração de contornos paramétricos em imagem binária como o famoso código de cadeia ou de Freeman [16]. Um algoritmo bastante útil que segue o contorno, popularmente conhecido como *ceguinho*, é explicado a seguir:

1. varre-se a imagem até que um ponto inicial pertencente ao contorno do objeto seja encontrado, assumindo que o objeto seja preto, o fundo será branco;
2. a partir do ponto inicial, o algoritmo circunda o objeto conectado até que o pixel de início seja revisitado, indicando que a tarefa se completou;
3. define-se o contorno externo de um objeto conectado.

A curvatura é definida a partir do contorno e mede localmente quão abrupta é a curva num dado ponto (analogia com aceleração), consistindo em um conceito

clássico da geometria diferencial de curvas e superfícies. A curvatura é importante na percepção visual, pois ela carrega informações relevantes para o entendimento da forma observada. Pontos de alta curvatura podem ser entendidos como quinas ou extremidades de uma curva. Dada uma curva paramétrica como a evolução de um ponto no espaço bidimensional, definimos a posição de uma partícula no plano como sendo o vetor $\overrightarrow{p}(t) = [x(t), y(t)]$, onde t é um valor real chamado parâmetro que dá cada ponto da curva. Assim, podemos escrever a curvatura de um ponto como na Eq. 3.20:

$$k(t) = \frac{x'(t)y''(t) - x''(t)y'(t)}{[x'(t)^2 + y'(t)^2]^{3/2}} \tag{3.20}$$

Considerando a curvatura, muitas propriedades geométricas podem ser extraídas como pontos de convexidade e de concavidade, pontos de inflexão, segmentos de reta, assim como a energia de dobramento (B), que traduz a rugosidade do contorno e é representada na Eq. 3.21, onde P é o número de pixels do contorno.

$$B = \frac{1}{P}\sum_{t=1}^{P} k(t)^2 \tag{3.21}$$

3.3.2 Cor

A cor pode ser usada como critério para segmentação da imagem, conforme descrito na Seção 3.2.3, bem como um atributo das ROI da imagem. As imagens coloridas são originalmente mapeadas nos canais RGB, ou seja, os canais de cromaticidade que codificam cada pixel da imagem. Partindo dos histogramas de primeira ordem dos canais RGB tomados separadamente, é possível calcular características densitométricas, que expressam a intensidade das regiões. O conceito de histograma pode ser estendido para descrever a distribuição de probabilidade para cada um dos canais RGB. O histograma de cor é uma propriedade extraída dos canais cromáticos e é comumente utilizada [36].

$$P(b) = \frac{N(b)}{M} \tag{3.22}$$

$$S_M = \overline{b} = \sum_{b=0}^{255} b * P(b) \tag{3.23}$$

$$S_D = [\sum_{b=0}^{255} (b - \overline{b})^2]^1/2 \tag{3.24}$$

$$S_S = \frac{1}{\sigma^3} \sum_{b=0}^{255} (b - \overline{b})^3 P(b) \tag{3.25}$$

$$S_k = \frac{1}{\sigma^4} \sum_{b=0}^{255} (b - \overline{b})^4 P(b) - 3 \tag{3.26}$$

$$S_N = \frac{1}{\sigma^3} \sum_{b=0}^{255} [P(b)]^2 \tag{3.27}$$

$$S_E = -\frac{1}{\sigma^3} \sum_{b=0}^{255} P(b) * log_2[P(b)] \tag{3.28}$$

A extração de características se dá por meio de 6 descrições quantitativas da forma do histograma [15] de cada canal de cromaticidade, calculadas segundo a Eq.3.22, portanto 18 medidas diferentes. Algumas das estatísticas do histograma de cor incluem a média (Eq.3.23), desvio padrão (Eq.3.24), coeficiente de assimetria[6] (Eq.3.25), curtose[7] (Eq.3.26), energia (Eq.3.27) e entropia (Eq.3.28) de cada canal RGB, onde P é o histograma normalizado do canal, b é um nível de amplitude em $0 \leq b \leq 255$, N é o número de pixels de amplitude b e M é o número total de pixels. Song [37] utiliza medidas semelhantes na descrição de leucócitos em imagens de esfregaços de sangue.

3.3.3 Textura

Apesar de ser uma característica intrínseca em imagens naturais, não existe uma definição formal e precisa do que seja textura. Uma abordagem aceitável é descrever textura como um padrão repetitivo de variações locais na intensidade da imagem que sejam tênues demais para serem distintas como objetos separados. Descrições

[6] tradução de *skewness*.

[7] tradução de *kurtosis*; sinônimo de coeficiente de achatamento.

texturais têm sido largamente utilizadas em segmentação de imagem, análise de forma, estimação de orientação e outras tarefas de inspeção como a classificação de micrografias de células [31, 38].

A textura é comumente usada como um descritor de região em análise de imagem e visão computacional. As principais abordagens usadas para descrever textura são estatística, estrutural e espectral. Técnicas estatísticas caracterizam textura por propriedades estatísticas dos níveis de cinza da imagem. Tipicamente, estas propriedades podem ser calculadas usando histogramas dos níveis de cinza (estatística de primeira ordem) e matrizes de co-ocorrência dos níveis de cinza (GLCM[8]), ou seja, de segunda ordem. O estudo de textura por GLCM foi proposto por Haralick [3], que define a GLCM como a distribuição de probabilidade conjunta entre dois pixels de uma imagem em nível de cinza dada uma certa direção, distância e tamanho de bloco (janela), sendo que o número de níveis de cinza considerado (N_g) dá a dimensão da GLCM. A partir dessa matriz, várias medidas podem ser extraídas informando a respeito do contraste (quantidade de variações locais) e ordem (regularidade entre os valores de pixel dentro de uma janela).

O algoritmo para extração de parâmetros de textura pode ser sumarizado através do seguintes passos, conforme proposto por Haralick [3]:

1. Equalização da imagem [3];
2. Quantização: a quantização nos níveis de cinza reduz a demanda computacional e elimina grãos finos;
3. Filtro de mediana;
4. Cálculo da GLCM dados o parâmetro de distância (d), ângulo (θ) e área de cobertura da imagem ou janela (w);
5. Cálculo das medidas de textura a partir da GLCM normalizada (p): usou-se particularmente a correlação, energia, entropia, homogeneidade e inércia, porém existem outros parâmetros de textura que podem ser calculados [17].

A janela dá o parâmetro de variação de espaço-escala em casos em que se impõe isotropia, ou seja, a GLCM é insensível a rotações de 180 graus e invariante para os

[8]GLCM = acrônimo de *Gray Level Cooccurrence Matrix*.

diferentes valores de $\theta = 0°, 45°, 90°, 135°$ e distância entre pixels constante, o que resulta fazendo-se a média de p calculada para cada um dos ângulos como na Eq. 3.29.

$$p_{iso} = \sum_{\theta=0,45,90,135} \frac{(p_\theta + p_\theta^T)}{8} \tag{3.29}$$

Sejam dois pixels, i e j, da imagem em nível de cinza (I_g), eles podem assumir valores inteiros em $[0, N_g - 1]$. A distribuição de probabilidade conjunta p é indexada pelos pixels de I_g, onde a linha é dada por i e a coluna por j. Assim, a GLCM é construída a partir da imagem relacionando-se a intensidade dos pixels par a par através da contagem do número de transições ($\sharp$) do pixel i para o pixel j. Considere a funcão ρ como a transição resultante de $i = I_g(x, y)$ para $j = I_g(\bar{x}, \bar{y})$ dada alguma distância d e ângulo θ, então p pode ser expresso como a na Eq.3.30:

$$p(i, j|d, \theta) = \sharp\{(i, j) \in I_g \mid j = \rho(i|d, \theta)\} \tag{3.30}$$

onde ρ é a função do primeiro pixel dado um deslocamento d em uma das 4 direções possíveis, cuja representação se dá através das Eq.3.31-3.34.

$$\rho(i|d, \theta = 0^o) \rightarrow \quad \bar{x} = x + d; \quad \bar{y} = y; \tag{3.31}$$

$$\rho(i|d, \theta = 45^o) \rightarrow \quad \bar{x} = x + d; \quad \bar{y} = y + d; \tag{3.32}$$

$$\rho(i|d, \theta = 90^o) \rightarrow \quad \bar{x} = x; \quad \bar{y} = y + d; \tag{3.33}$$

$$\rho(i|d, \theta = 135^o) \rightarrow \quad \bar{x} = x - d; \quad \bar{y} = y + d; \tag{3.34}$$

Ao invés de se usar a textura para segmentação da imagem, é possível apenas extrair atributos de textura para fins de classificação, os quais são capazes de expressar diferenças entre regiões mais rugosas de regiões menos rugosas, por exemplo. Supõe-se a princípio que a informação textural esteja contida na relação espacial média entre os tons de cinza, ou seja, a textura de uma imagem é interpretada em termos da média de um certo parâmetro de textura calculado para várias partições da imagem. Alguns parâmetros texturais [3, 17] foram adotados na corrente tese e calculados a partir de p em função de i e j, ($p_{i,j}$), sendo listados nas Eq.3.35-3.39:

$$\text{Inércia} = \sum_{i,j=0}^{n-1} p_{i,j}(i-j)^2 \tag{3.35}$$

$$\text{Energia} = \sum_{i,j=0}^{n-1} (p_{i,j})^2 \tag{3.36}$$

$$\text{Entropia} = -\sum_{i,j=0}^{n-1} p_{i,j}[log(p_{i,j})] \tag{3.37}$$

$$\text{Correlação} = \frac{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} ij * p_{i,j} - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{3.38}$$

$$\text{Homogeneidade} = \sum_{i,j=0}^{n-1} p_{i,j}/(1+(i-j)^2) \tag{3.39}$$

As medidas de textura podem ser divididas em dois grupos[9] de acordo com o propósito dos pesos nas equações :

- Contraste: essas medidas usam pesos (valores de $(i-j)$) relacionados com a distância à diagonal. Como exemplo, interpreta-se a Eq. 3.35, quando i e j são iguais, a $p_{i,j}$ está na diagonal e $(i-j)=0$. Estes valores representam pixels inteiramente semelhantes a sua vizinhança, portanto o peso é 0 e quanto maior for a diferença, maior o contraste entre os pixels. Ex: inércia e homogeneidade.

- Ordenação: medidas que expressam a regularidade ou ordenação dos valores dos pixels dentro de uma certa janela. Exemplificando, a simples identificação de um valor máximo para $p_{i,j}$ é um indicador de ordenação uma vez que houve a dominação de uma combinação de pixels. Ex: probabilidade máxima, energia, entropia e correlação.

Extraídas as medidas para reconhecimento de padrão, é necessária normalização antes que haja comparação entre os atributos da imagem, que possuem origem diversa (ex: medidas em pixels, em micrômetros, etc). A normalização de características foi feita através da transformação das características para que se tenha a média zero e desvio padrão unitário, conforme mostra a Eq.3.40, onde f é o conjunto

[9]fonte: http://www.ucalgary.ca/ mhallbey/texture/texture_tutorial.html

de amostras e μ e σ representam a média e desvio padrão do atributo j [16].

$$\hat{f}(i,j) = \frac{f(i,j) - \mu_j}{\sigma_j} \tag{3.40}$$

Tendo determinado os atributos para caracterizar a imagem, passa-se à fase de aprendizado.

3.4 Aprendizado

O aprendizado de máquina é uma sub-área da inteligência artificial que lida com programas que adquirem conhecimento por experiência. Aprendizado é o processo através do qual parâmetros são ajustados para que um programa desempenhe uma certa tarefa; informações relevantes sobre os padrões são apresentados, criando uma representação própria para o problema [39].

A estatística descritiva descreve o conjunto de amostras em termos de medidas de tendência central e dispersão. Já a inferência estatística consiste em efetuar generalizações tomando por base amostras estatísticas, ou seja, traçar conclusões a respeito de uma população com base nas amostras [40].

Mas, qual é a diferença entre aprendizado de máquina e estatística? Essas áreas se tangenciam e se fundem quando técnicas de análise de dados são consideradas. Para apontar uma diferença preponderante, embora muito simplificada, tem-se que a estatística busca se dedicar mais ao teste de hipóteses enquanto o aprendizado de máquina à formulação de processos de generalização como uma busca através de hipóteses [41]. Testes estatísticos são usados para validar modelos de aprendizado de máquina e para avaliar algoritmos de aprendizado de máquina. Basicamente, existem dois modos de aprendizado:

a) supervisionado: dado um conjunto de amostras de um conceito, o aprendiz induz um conceito geral que descreve todos os exemplos [42]. Tanto as entradas quanto as saídas são conhecidas, pois o usuário alimenta o programa com amostras rotuladas (pré-classificadas). Assim o padrão de entrada é identificado como um membro de uma classe pré-definida pelos padrões de treinamento. Este tipo de aprendizado pode ser visto como pura inferência indutiva, o qual induz alguma função que se aproxima da realidade sendo aprendida [25].

b) não-supervisionado: quando não há informação do que devem ser as saídas e o aprendizado ocorre por meio de observação e descoberta. Assim, o padrão é associado a um agrupamento tomando como base, critérios de similaridade entre os padrões de treinamento.

O desempenho de algoritmos de aprendizado supervisionado é medido em termos da capacidade de predição das classes, dados exemplos ainda não vistos. Existem vários métodos de avaliação de aprendizado, os quais medem a capacidade de generalização dos classificadores, característica imprescindível para a viabilização de qualquer sistema de reconhecimento de padrões.

Na fase de treinamento do classificador, ele "aprende"a classificar as amostras de teste. A qualidade e quantidade de exemplos utilizados durante o treinamento influenciam na habilidade de generalização, ou seja, seu desempenho ao classificar padrões de teste que ainda não foram apresentados. Os termos a seguir são frequentemente utilizados para se referir às deficiências na capacidade de generalização do classificador [41, 43]:

a) ajuste demasiado (*overfitting*): quando o classificador se adapta a detalhes específicos do conjunto de treinamento devido a um número exagerado de parâmetros livres, pode ocorrer uma redução da taxa de acerto devido ao ajuste demasiado. Os principais problemas de um classificador baseado em uma superfície de decisão mais complexa do que o necessário são: **(i)** o custo computacional, que é proporcional à complexidade do sistema classificador; **(ii)** um classificador muito complexo possui menor poder de generalização e a complexidade da superfície de decisão é proporcional ao número de parâmetros livres que ele possui. Contornando esse inconveniente, o classificador deve ser o mais simples possível, focando mais nas regularidades dos dados e menos em peculiaridades (em geral, ruídos). O aumento do número de padrões de treinamento, incluindo padrões que representem o máximo de variações possíveis, também pode melhorar os resultados.

b) treinamento demasiado (*overtraining*): quando o conjunto de treinamento contém muitos exemplos e com pequena variação intra-classe, a capacidade de generalização do classificador é comprometida. Uma solução é utilizar dados realmente representativos da variabilidade dos padrões.

c) maldição da dimensionalidade (*curse of dimensionality*): problemas ocor-

rem quando os graus de liberdade do sistema são relativamente maiores que o número de amostras para treinamento, ou seja, o número de amostras de treinamento requeridos para que um classificador tenha um bom desempenho é uma função exponencial da dimensão do espaço de características. O relacionamento exato entre a probabilidade de erro, o número das amostras de treinamento, o número de características e os parâmetros reais das densidades condicionais de classe é difícil de ser estabelecido. Contudo, aceita-se geralmente que a razão entre o número de amostras por classe (n) e o número de atributos, ou dimensionalidade (d), seja no mínimo dez (Eq. 3.41), conforme proposto por [43]. É uma boa prática no projeto de um classificador, embora quanto mais complexo o classificador, maior deve ser essa razão para evitar a maldição da dimensionalidade.

$$\frac{n}{d} > 10 \tag{3.41}$$

d) fenômeno de pico (*peaking phenomena*): é um comportamento que o classificador assume quando o número de características que descrevem o problema é grande em comparação ao conjunto de dados. Desse modo, conforme o número de características aumenta para um tamanho de amostras fixo, a taxa de acerto do classificador decai [43].

3.5 Classificação

Nesta seção serão descritas informações gerais sobre dois classificadores estatísticos (classificador bayesiano e k-vizinhos) implementados durante o doutorado, muito embora outros métodos tenham sido usados como árvore de decisão e redes neurais, dentro do ambiente Weka. Os Classificadores estatísticos operam em duas fases:

1. aprendizado: é durante essa fase que se disponibiliza as amostras, contendo os atributos apropriados para representar os padrões de entrada. Se o classificador for supervisionado, ele é treinado a particionar o espaço de características;

2. classificação ou teste: o classificador relaciona o padrão de entrada a uma das classes sob consideração, baseado-se nas características medidas.

Conforme apresentado anteriormente neste capítulo, o classificador possui como entrada um padrão e, como saída, o rótulo que identifica a que classe tal padrão

pertence. Classificadores de aprendizado supervisionado precisam ser treinados utilizando um conjunto de amostras. Dentre os métodos de reconhecimento de padrões mais conhecidos, destacam-se [43, 13]:

a) Estatísticos: mínima distância ao protótipo, k-vizinhos mais próximos, k-médias, classificador bayesiano, etc.

b) Não-estatísticos: redes neurais, classificadores sintáticos, árvore de decisão e baseados em lógica fuzzy;

c) Multi-classificadores: ou métodos híbridos que são múltiplos classificadores usados em conjunto.

Entre as várias abordagens (Fig.3.7) para reconhecimento de padrões, a abordagem estatística tem sido a mais estudada e utilizada. No reconhecimento estatístico de padrões, foca-se nas propriedades dos padrões (geralmente expressos por densidades de probabilidades).

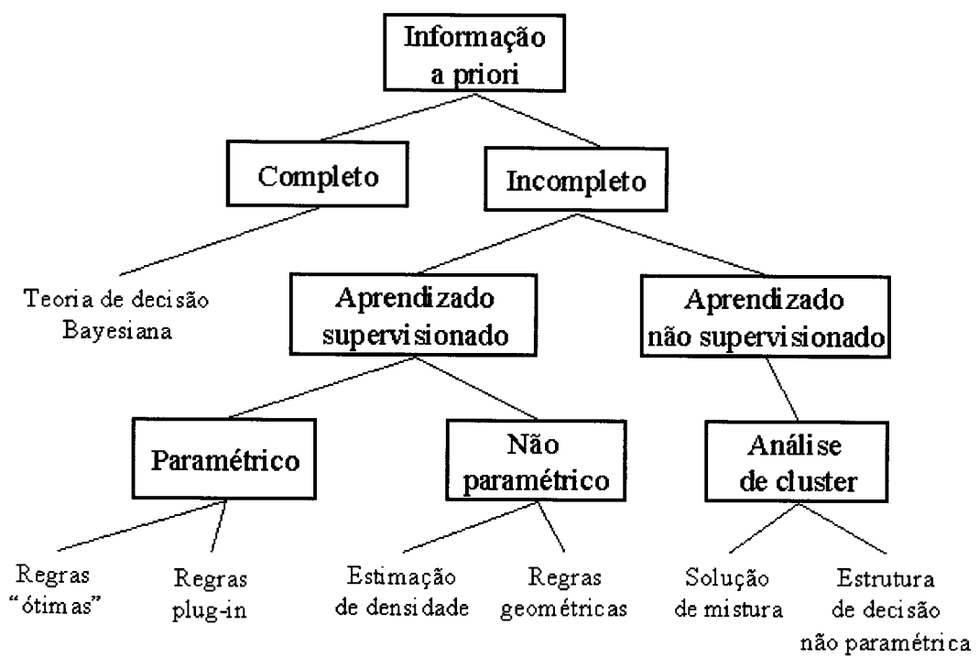


Figura 3.7: Reconhecimento estatístico de padrões segundo Jain et al [4].

Um padrão pode ser associado a um vetor x contendo n medidas também designadas como sendo observações, variáveis, características ou atributos. Dado um conjunto de padrões de treinamento, é necessário estabelecer limites de decisão no espaço de características a partir das c diferentes classes $(\omega_1, \omega_2, ..., \omega_c)$, idealmente ocupando regiões compactas e disjuntas em um espaço n-dimensional.

Refletindo o conhecimento a priori sobre as classes, temos as **probabilidades a priori** $P(\omega_i)$, representadas simplesmente por P_i. A soma de P_i para todas as classes é 1.

$$\sum_{i=1}^{c} P_i = 1$$

Define-se $p(x|\omega_i)$ como a **função densidade de probabilidade multivariada** de x quando se sabe que x pertence à classe ω_i para $i = 1, ..., c$. A **função densidade de probabilidade local** de x é definida por:

$$p(x) = \sum_{i=1}^{c} P_i.p(x|\omega_i)$$

Esses conceitos consistem nos elementos básicos do problema de classificação e são suficientes para descrever ou modelar uma grande gama de mecanismos de geração de padrões [44].

Complementando o conceito de distribuição de probabilidade (Fig.3.8), é possível usar estimação paramétrica (uma natureza matemática da função densidade é assumida) ou não-paramétrica [16]. A técnica paramétrica é ótima se a densidade condicional das classes é gaussiana, embora isso raramente ocorra. Estatísticas assumindo esta distribuição podem ser robustas, portanto é interessante considerar técnicas não-paramétricas somente se os dados se afastam da normalidade.

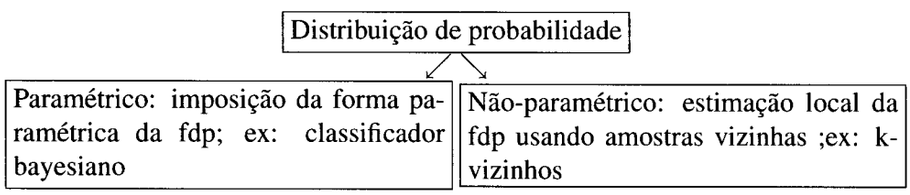


Figura 3.8: Distribuição de probabilidade: paramétrica e não-paramétrica

O corrente trabalho considera a classificação supervisionada, uma vez que se tem conhecimento prévio do número de classes possíveis de células. As classes

foram determinadas pelas distribuições de probabilidade dos padrões pertencentes a cada um dos tipos de leucócitos. Os parâmetros como média e desvio padrão foram estimados, a partir do conjunto de amostras disponível, para cálculo de fdp's gaussianas, que serão definidas na próxima seção.

3.5.1 Densidade Normal

Um dos resultados mais importantes da teoria de probabilidade é o *Teorema Central do Limite*, que afirma que qualquer variável que seja uma soma de várias variáveis aleatórias pode ser aproximada por um distribuição gaussiana. Assim, qualquer que seja a distribuição da variável de interesse para grandes amostras, a distribuição das médias amostrais[10] será aproximadamente normalmente distribuída, e tenderá a uma distribuição normal à medida que o tamanho da amostra crescer. Então podemos ter uma variável original com uma distribuição muito diferente da normal (pode até mesmo ser discreta), mas se tomarmos várias amostras grandes desta distribuição, e então fizermos um histograma das médias amostrais, a forma se parecerá como uma curva normal. Portanto a fdp normal ou gaussiana é muito importante, tanto por razões teóricas quanto práticas. A definição matemática da distribuição normal unidimensional é dada na Eq.3.42 e a d-dimensional na Eq.3.43.

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}.e^{-0,5(\frac{x-\mu}{\sigma})^2} \tag{3.42}$$

$$p(\overrightarrow{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}.e^{-0,5(\overrightarrow{x}-\overrightarrow{\mu})'\sum^{-1}(\overrightarrow{x}-\overrightarrow{\mu})} \tag{3.43}$$

onde μ, σ indicam respectivamente a média e o desvio padrão e e d o comprimento de $\overrightarrow{x}$, $\sum$ indica a matriz de covariância e $|\sum|$ é seu determinante. O vetor de média dá o centro das classes e a matriz de covariância indica a dispersão das amostras em torno da média. É comum o uso do símbolo $\mathtt{N}(\mu, \sum)$ para denotar a função densidade de probabilidade gaussiana multivariada.

Quando um conjunto de valores tem uma tendência central suficientemente forte, é significativo caracterizar o conjunto utilizando o conceito de média, que estima

[10]Maiores informações: http://www.est.ufpr.br/ silvia/CE001/node38.html

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

o valor central dos pontos amostrados. A covariância é outro conceito útil para caracterizar a variabilidade deste conjunto de valores.

Embora diversos trabalhos suponham que as variáveis aleatórias do problema têm comportamento gaussiano, uma minoria deles relata sobre testes de aderência dessas variáveis ao modelo proposto através de testes da hipótese gaussiana em termos da divergência entre os dados observados e os esperados. Um exemplo de método largamente aplicado é o qui-quadrado (χ^2) [40].

3.5.2 Teoria de decisão bayesiana

A teoria de decisão bayesiana é uma abordagem estatística fundamental ao problema de reconhecimento de padrões, baseada na quantificação de balanços para decidir entre várias classes usando probabilidade e os custos que acompanham tais decisões. Assume-se que o problema de decisão é formulado em termos probabilísticos, que se conhece tais valores e que ele pode ser formalizado tal como procedimentos de senso comum[13].

Sejam x e y dois eventos ou duas variáveis aleatórias, a partir da definição de probabilidade condicional, sabe-se que quando duas variáveis são estatisticamente dependentes, o conhecimento do valor de uma delas conduz a estimativa do valor da outra. Isto é expresso pela seguinte definição de probabilidade condicional de x e y:

$$P(x,y) = \frac{P(x|y)}{P(x)} = \frac{P(y|x)}{P(y)} \tag{3.44}$$

$$P(y) = \sum_{x} P(x,y) \tag{3.45}$$

A expressão da Eq.3.45 é uma instância da *Lei da Probabilidade Total.* Segundo esta lei, um evento (ou variável aleatória) y pode ocorrer de n maneiras diferentes com $x = v_1, x = v_2, ..., x = v_n$ e se esses n sub-eventos são *mutuamente exclusivos*[11], então a probabilidade de y ocorrer é a soma das probabilidades dos sub-eventos y_i. Como as possibilidades são mutuamente exclusivas, segue-se da Lei da Probabilidade Total que $P_y(y)$ é a soma das probabilidades conjuntas $P(x,y)$ sobre todos os valores

[11] mutuamente exclusivos = não podem ocorrer ao mesmo tempo, o que não significa que sejam independentes.

possíveis de x. Assim, através das Eq.3.44 e 3.45 obtém-se:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{3.46}$$

A Eq.3.46, demonstrada em 1763, é chamada de **Regra de Bayes**. Note que o denominador é obtido através da soma do numerador sobre todos os valores possíveis de x e consiste apenas em um fator de normalização necessário para assegurar que a soma dos $P(x|y)$ é igual a 1. A interpretação padrão da fórmula de Bayes é que ela "inverte"a conexão estatística, transformando $P(y|x)$ em $P(x|y)$.

Agora, sejam as duas variáveis aleatórias, x e ω, onde x representa a amostra do padrão e ω a classe. A **probabilidade a priori do padrão** x ocorrer sem levar em consideração nenhuma classe é $p(x)$. Cada padrão x pertence a uma determinada classe ω_i, descrita probabilisticamente por $P(\omega_i)$ ou P_i, que representa a **probabilidade a priori da classe**. Conhecida a **função de probabilidade condicional da classe** ou verossimilhança, $p(x|\omega_i)$, é possível fazer inferências.

Seja ω_i, $i = 1, ..., c$ uma partição do espaço amostral Ω, com $P_i > 0$ e o vento x com $p(x) > 0$, a **função de probabilidade a posteriori**, $p(\omega_i|x)$, usando a fórmula de Bayes (Eq.3.47) pode ser reescrita como:

$$p(\omega_i|x) = \frac{p(x|\omega_i)P_i}{p(x)} \tag{3.47}$$

onde a densidade de probabilidade ou evidência $p(x)$ pode ser expressa por:

$$p(x) = \sum_{i=1}^{n} p(x|\omega_i)P_i \tag{3.48}$$

embora sua função como denominador seja só um fator de escala que pode ser desprezado em etapas de classificação.

Informalmente, a fórmula de Bayes pode ser escrita como:

$$posteriori = \frac{\text{verossimilhança} \times priori}{\text{evidência}} \tag{3.49}$$

A regra de Bayes é a regra de decisão ótima para um classificador, baseada no conhecimento estatístico completo da fonte geradora de padrões, pois ela minimiza a probabilidade média de erro [13, 45]. Este classificador ótimo é frequentemente

usado como comparação para classificadores projetados por outros métodos (subótimos).

Se o classificador bayesiano é ótimo, por que usar outros classificadores? O motivo é que conforme mencionado, o classificador bayesiano só pode ser executado se a probabilidade a priori P_i e a fdp $p(x|\omega_i)$ forem conhecidas, o que em geral não ocorre. Tipicamente, tem-se apenas um conhecimento geral parcial sobre a situação e dados de treinamento [13].

Circundando esse problema, adota-se métodos de estimação na fase de treinamento a partir das amostras, entretanto quando a distribuição das classes possui formas "complicadas"e descontínuas, esses métodos se tornam de alto custo computacional para representações precisas das probabilidades. Geralmente se supõe a distribuição normal (Eq.3.50) como distribuição a priori para a ocorrência dos atributos nos objetos, o que implica em separadores lineares, quadráticos ou hiperbólicos, de acordo com a dimensão dos vetores de características.

$$p(x|\omega_i) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma_i|}}.e^{-0,5(x-\mu_i)'\Sigma^{-1}(x-\mu_i)}, i = 1, ..., c \tag{3.50}$$

onde μ_i e Σ_i são o vetor médio e a matriz de covariância da classe ω_i. Referências importantes sobre o assunto são [46, 13, 43, 47, 45, 41].

Classificador Bayesiano Ingênuo

Apesar dos classificadores bayesianos serem ótimos, não é confiável aplicá-los diretamente uma vez que eles requerem a probabilidade de todas as combinações possíveis das evidências, isto é, $p(x_1, x_2, ..., x_d|\omega_i)$. Contornando esse problema, surge o classificador bayesiano ingênuo[12] que assume independência condicional dos atributos, simplificando as computações drasticamente [48, 41]. Assim, como forma de diminuir o número de probabilidades envolvidas nos cálculos e aumentar a confiabilidade de suas estimativas, a premissa do classificador bayesiano ingênuo é feita de modo que todo atributo é independente dos outros atributos, o que é evidenciado pela

[12] *Naive Bayesian Classifier*: termo cunhado por Konenko em 1990; outros nomes comuns para o método incluem: classificador bayesiano simples e Bayes idiota; *Naïve* = ingênuo, tolo, superficial, com pouca experiência em quão complicada a vida é de modo que confia demais nas pessoas e acredita que coisas boas sempre vão acontecer.

Eq.3.51.

$$p(x_1, x_2, ..., x_n|\omega_i) = \prod_{j=1}^{n} p(x_j|\omega_i) \tag{3.51}$$

Apesar do nome pejorativo, o classificador bayesiano ingênuo funciona muito bem quando testado em bancos de dados reais, particularmente quando combinado com procedimentos de seleção de atributos [41], que serão discutidos no capítulo 4.

3.5.3 Estimação de parâmetros

Supondo que a densidade condicional de classe $p(x|\omega_i)$ é a função densidade normal, o problema passa a ser estimar a média μ_i e a matriz de covariância $\sum_i$. Dois procedimentos comumente considerados são a estimação por máxima verossimilhança e a estimação bayesiana. Embora os resultados obtidos com estes dois procedimentos sejam equivalentes no limite assintótico em dados infinitos de treinamento, as abordagens são conceitualmente diferentes. A estimação por máxima verossimilhança considera os parâmetros como quantidades cujos valores são fixos, embora desconhecidos. A melhor estimativa é aquela que maximiza a probabilidade de obtenção das amostras realmente observadas. Por outro lado, no método bayesiano os parâmetros são variáveis aleatórias tendo alguma distribuição a priori conhecida [13].

A estimativa de máxima verossimilhança é atraente pois quase sempre tem boas propriedades de convergência à medida que o número de amostras de treinamento aumenta, além do que é frequentemente mais simples que outras abordagens, como a bayesiana, por exemplo. Esse foi o método usado no projeto corrente, onde a classe a ser escolhida (ω_k) para maximizar a taxa de acerto é descrita segundo a Eq.3.52.

$$\omega_k = argmax_i\{P(\omega_i|x) = p(x|\omega_i)P_i\}, i = 1, ..., c \tag{3.52}$$

Chamou-se o classificador projetado no corrente trabalho de bayesiano uma vez que as probabilidades a posteriori foram estimadas a partir das probabilidades a priori e minimizou-se a probabilidade de erro por meio da regra de Bayes.

Funções discriminantes ($g_i(x)$, para $i = 1, ..., c$) podem ser encaradas como um tipo de classificador bayesiano. Nestas, o classificador é visto como uma rede que

computa c funções discriminantes e seleciona a categoria correspondente ao maior discriminante. O classificador bayesiano é naturalmente representado desta maneira. No caso da máxima verossimilhança, a função discriminante é $g(x_i) = P(\omega_i|x)$, assim a função discriminante máxima corresponderá à probabilidade a posteriori máxima.

3.5.4 Regra dos K-vizinhos mais próximos

Esse método de classificação não possui processamento na fase de treinamento. Para cada padrão de teste é medida a distância entre ele e todos os padrões de treinamento, verifica-se a quais classes pertencem os K padrões mais próximos e a classificação do padrão teste será a classe predominante entre esses K padrões de treinamento [47].

Um caso particular ocorre quando K, o número de vizinhos, é 1. Neste caso, pode-se demonstrar que o erro máximo cometido pelo classificador nunca é maior do que o dobro do menor erro possível, de acordo com a regra de Bayes, se o número de exemplos classificados disponíveis for infinito [13]. Para K maior, o erro aproxima-se do menor erro possível, mas o fato do número de exemplos disponíveis ser finito contraria esta tendência, pelo que se utiliza geralmente $K = 3$ em uma primeira aproximação, e o valor final é obtido pelo mínimo do erro de generalização.

A principal vantagem desse método é que ele cria uma superfície de decisão que se adapta à distribuição dos dados de treinamento de maneira detalhada, possibilitando a obtenção de boas taxas de acerto quando o conjunto de treinamento é grande ou representa todas as variações possíveis dos dados, principalmente no caso de $K = 1$. Valores maiores de K não produzem bons resultados quando a distribuição dos conjuntos possui poucas sobreposições, pois há uma menor garantia de que os K vizinhos forneçam informação representativa a respeito da verdadeira classe da amostra.

Uma desvantagem deste classificador está na complexidade do teste, visto que para cada padrão de teste é necessário realizar várias medições de distância (a todos os padrões de treinamento) e ordenar os resultados.

3.6 Estimação de erro

O erro de classificação ou simplesmente taxa de erro é geralmente a última etapa de medida do desempenho de um classificador. A taxa de erro de um sistema de reconhecimento deve ser estimada usando-se todas as amostras disponíveis divididas em amostras de treinamento e amostras de teste para evitar resultados tendenciosos, embora nem todas as técnicas considerem esse requisito. O classificador é projetado usando as amostras de treinamento e então avaliado usando as amostras de teste. A porcentagem de amostras de teste mal classificadas é tomada como estimativa da taxa de erro.

Durante o treinamento ou aprendizado de um classificador, é necessário dispor de um conjunto de amostras para ajuste dos parâmetros do classificador ao problema. A habilidade de generalização dos classificadores se refere a seu desempenho ao classificar padrões de teste que não foram utilizados durante o treinamento. Medir o desempenho de um classificador no conjunto de treinamento é possível, mas não é um bom indicador de desempenho nos dados futuros, pois há uma tendência que não necessariamente será encontrada em padrões de teste. Portanto, metodologias devem ser adotadas para se obter o máximo de informação do conjunto de dados disponível sem que essas informações sejam tendenciosas. Para que a estimativa de erro seja confiável é preciso que os conjuntos de treinamento e de teste sejam grandes, além de serem independentes, requisito frequentemente negligenciado [43].

Procedimentos apropriados para lidar com as amostras na fase de treinamento e teste geralmente consideram a divisão do conjunto de amostras em três: treinamento, validação e teste, onde os dados de validação servem para otimizar os parâmetros do classificador. Contudo, cai-se no dilema de como dividir as amostras: quanto maior o número de amostras no treinamento, melhor o classificador e quanto maior o número de amostras para teste, mais precisa a estimativa de erro. Para lidar com esse problema, alguns métodos são utilizados e sintetizados na Tabela 3.1, segundo [43, 46]:

Os métodos *holdout*, *leave-1-out* e *n-fold cross-validation* são versões da abordagem de validação cruzada [43] e uma das principais desvantagens desses métodos, especialmente para conjuntos pequenos de amostras, é que nem todas as amostras disponíveis são usadas para treinar o classificador. Para contornar esse problema, o

Tabela 3.1: Métodos para estimativa de erro

Método	Propriedades	Descrição
Resubstituição	Todas as amostras disponíveis são usadas para treinamento e para teste; conjunto de treinamento e de teste são os mesmos.	Estimativa otimisticamente tendenciosa, especialmente quando a razão do número de amostras para a dimensionalidade é pequena.
Holdout	Amostras são divididas em dois grupos independentes: um conjunto para teste e o resto pra treinamento; sua versão avançada utiliza estratificação, processo que garante que cada classe esteja representada com aproximadamente a mesma proporção em ambos conjuntos.	Estimativa pessimisticamente tendenciosa; diferentes partições darão estimativas diferentes.
Leave-1-out	O classificador é projetado usando $(n-1)$ amostras e avaliado usando a amostra restante, processo repetido n vezes com conjuntos de treinamento diferentes de tamanho $(n-1)$.	Estimativa não tendenciosa, mas apresenta grande variância; alto custo computacional porque n classificadores diferentes têm de ser projetados.
N-fold cross-validation ou validação cruzada com n-cruzamentos ou método da rotação	Amostras são divididas em n subconjuntos de tamanho igual, então cada subconjunto é utilizado para teste e o restante para treinamento, sendo que o desempenho é medido pela média dos resultados de teste. Geralmente os subconjuntos são estratificados antes que a validação cruzada seja realizada. O método mais utilizado de avaliação é o *ten-fold cross-validation* $(n=10)$ estratificado.	Estimativa apresenta-se menos tendenciosa que o *holdout* e é computacionalmente mais rápida que o *leave-1-out*.
Bootstrap	É um método de estimativa que utiliza exemplos com substituição para formar o conjunto de treinamento, k conjuntos de dados com n instâncias são utilizados n vezes a fim de formar um novo conjunto de dados com n instâncias, então estes dados são utilizados como conjunto de treinamento e as instâncias do conjunto de dados original que não ocorrem no novo conjunto de treinamento são utilizadas para teste.	Estimativa pode ter variância menor que o *leave-1-out*, embora computacionalmente mais custoso; pode ser útil em situações onde o conjunto de amostras seja pequeno.

método de *bootstrap* tem sido proposto [49][13].

[13] apud [43]

Capítulo 4

Seleção de características

"Nada é mais difícil, e por isso mais precioso, do que ser capaz de decidir. (Napoleão)"

Seleção de atributos ou seleção de características é uma etapa importante antes do reconhecimento de padrões imprescindível em casos onde a dimensão do espaço de características é muito grande. Esta área vem sendo aplicada no campo de reconhecimento de padrões há décadas, como por exemplo, a estratégia de seleção por busca heurística, introduzida na década de 60 [41].

O objetivo da seleção de atributos é reduzir a dimensionalidade dos vetores $\overrightarrow{x}$ associados aos padrões por meio da seleção de subconjuntos que mantenham a separabilidade entre as classes ou que garantam a maior separabilidade entre elas no espaço de características. Algumas razões pelas quais é desejável efetuar redução de dimensionalidade são a diminuição do custo computacional tanto para extração de medidas quanto para a classificação de muitas variáveis (elimina-se as pouco relevantes). Com a diminuição de redundâncias existentes no conjunto inicial de variáveis, diminui-se o risco de comprometer o desempenho do classificador para um número limitado de padrões de treinamento.

Embora uma profusão de medidas possa ser extraída de imagens, não existe um método geral aplicável a todos os problemas de visão computacional, daí a necessidade de desenvolvimento de ferramentas para refinamento de possibilidades, utilizando técnicas de seleção de características. A seleção automática de características é uma técnica de otimização, onde dado um conjunto de d características, busca-se selecionar um subconjunto de tamanho n ($n < d$) que maximiza uma função critério.

Formalmente, dado um conjunto S de d características, o algoritmo de seleção de características deve encontrar um subconjunto de tamanho $S_i \subsetneqq S$ tal que $|S_i| = n$, em que $|S_i|$ denota a cardinalidade de S_i e J dá a função critério (Eq.4.1).

$$J(S) = max\{J(S_i)\}, para S_i \subsetneqq S, |S_i| = n \tag{4.1}$$

Os algoritmos de busca fornecem as possíveis combinações entre os atributos para se testar diferentes subconjuntos de atributos. São usadas funções critério para que se possa avaliar a separabilidade entre todas as classes de maneira global. Se o critério de separabilidade se basear em distâncias, então o método de busca é do tipo *filter* (ex: a função critério poderia ser a máxima distância média entre as classes). Senão, se o critério se basear em um classificador, então se diz que o método de busca é *wrapper* (ex: a função critério seria a máxima taxa de acerto).

As próximas seções visam descrever unicamente os métodos utilizados durante esse projeto de doutorado. Todos eles foram implementados em software usando Delphi 4.0, com interface gráfica para fácil interação usuário-máquina.

4.1 Métodos de busca

Também conhecidos como processos de geração [50], consistem em métodos para lidar com um conjunto original de d atributos, onde o número total de conjuntos candidatos competindo seja 2^d, ou seja, todas as combinações possíveis. Este é um número bastante grande mesmo para valores não muito altos de d. Há diferentes abordagens para realizar buscas de subconjuntos de medidas relevantes como os processos completos (ou exaustivo), heurísticos (FSS, BSS, etc.) e aleatórios. A condição de parada para adição ou subtração de mais características é dada pela função critério, ou seja, quando a distância média inter-classe começar a diminuir ou a taxa de classificação começar a cair.

4.1.1 Exaustivo

Esse método usa a força bruta para encontrar o melhor subconjunto de n características, onde $n < d$, d é o número total de características. Diz-se força bruta pois se calcula a função critério para todas as combinações possíveis para d eventos, sem

considerar repetição do evento. Daí temos que o número de combinações possíveis para cada subconjunto (S_i) com d características arranjadas n a n é dado pela Eq. 4.2

$$S_i = \binom{d}{n} = \frac{d!}{(d-n)!n!} \tag{4.2}$$

Embora isto seja possível para valores pequenos de d, há alguns problemas óbvios com esta abordagem quando d é muito grande. Mesmo se forem usados os mesmos dados de teste, pode-se obter as características que sejam particularmente ajustadas para aquele conjunto de teste, mas que não sejam as melhores em geral. Pode-se ter que repetir o processo para vários valores de n para se tomar uma decisão acertada [41].

4.1.2 FSS - Forward Sequential Selection

Esse algoritmo heurístico inicia busca com zero atributos, avaliam-se todos os subconjuntos com um atributo através de critérios de separabilidade (distância ou taxa de acerto) e escolhe-se aquele que retorna a maior distância. Então, adiciona-se a esse subconjunto mais uma característica e repete-se o processo, sucessivamente. A nova característica incorporada é aquela que apresenta maior desempenho quando em conjunto com as características previamente selecionadas.

4.1.3 BSS - Backward Sequential Selection

Esse algoritmo heurístico inicia busca com todos os atributos e vai se removendo os atributos um a um, tirando-se aquele que oferece menor separabilidade entre as classes, ou seja, o pior dos atributos para aquela combinação.

4.2 Abordagem *filter*

A distância d é definida como $d : \Im \times \Im \rightarrow \Re^+$ e deve satisfazer as seguintes propriedades para I, J e K em $\Im$:

$$\begin{aligned} d(I,I) &= d(I,I) \rightsquigarrow \text{auto-similaridade} \\ d(I,J) &\geq d(I,I) \rightsquigarrow \text{minimalidade} \\ d(I,J) &= d(J,I) \rightsquigarrow \text{simetria} \\ d(I,K) + d(K,J) &\geq d(I,J) \rightsquigarrow \text{desigualdade triangular} \end{aligned}$$

Toda função que satisfaça essas propriedades é uma métrica e qualquer função que satisfaça somente as três primeiras propriedades é uma medida de similaridade [51]. É importante lembrar que uma distância (ou métrica) é definida somente entre dois elementos (duas classes ou duas amostras). Contudo, para se fazer seleção de características usando distância é preciso definir uma função critério que possa avaliar a separabilidade entre todas as classes de uma maneira global. Isso pode ser feito através de operações simples como a soma, a média ou o ínfimo dos resultados obtidos para todas as combinações entre as classes existentes. Calculam-se todas as distâncias possíveis (d_i) entre as classes(c) combinando-as duas a duas, seguindo uma fórmula da Eq. 4.3 semelhante a Eq. 4.2 apresentada na seção anterior.

$$Comprimento(d_i) = \binom{c}{2} = \frac{c!}{(c-2)!2!} \tag{4.3}$$

Qualquer que seja o critério de separabilidade escolhido, dado um subconjunto de características, a média entre as distâncias entre as classes é dividida pelo desvio padrão mais um ($\sigma+1$) entre as classes , evitando que haja erro por divisão por zero. Seja a distância d usada como critério dada por um escalar, $\vec{d_i}$ é o vetor contendo todas as distâncias possíveis entre duas classes. Seja a média e o desvio padrão, μ e σ, então d pode ser expresso pela Eq. 4.4.

$$d = \frac{\mu(\vec{d_i})}{\sigma(\vec{d_i}) + 1} \tag{4.4}$$

4.2.1 Euclidiana

É a distância mais conhecida e mais utilizada, sendo uma particularidade da métrica de Minkowski. Dadas as amostras $\vec{x}_i$ e $\vec{x}_j$, com variável k indicando um dos atributos

(uma coluna), a distância euclidiana é dada pela Eq. 4.5.

$$d_E = [\sum_{k=1}^{n}(x_{ik} - x_{jk})^2]^{1/2} \tag{4.5}$$

Algumas vezes se considera exclusivamente o somatório da diferença quadrática de $\vec{x}_i$ e $\vec{x}_j$ para diminuir o custo computacional.

4.2.2 Mahalanobis

É uma distância ponderada, onde Σ é a matriz de covariância do vetor aleatório x ou sua estimativa a partir de amostras. A matriz Σ é simétrica definida positiva. Para se poder usar a distância euclidiana, é necessário que as variáveis não sejam correlacionadas, o que é efetivado com a transformação de Mahalanobis, onde dado $\vec{x}$ nas coordenadas originais, esse é substituído por $\vec{z} = \sum^{-1/2} \vec{x}$ com métrica euclidiana [44]. Seguindo a mesma simbologia usada para distância euclidiana, definimos a distância de Mahalanobis através da Eq.4.6

$$d_M = [(\vec{x}_i - \vec{x}_j)^T \sum^{-1}(\vec{x}_i - \vec{x}_j)] \tag{4.6}$$

Vale salientar que muitas vezes se utiliza o quadrado da distância de Mahalanobis ao invés da distância de Mahalanobis.

Muitas outras distâncias podem ser calculadas, embora não sejam aplicados no problema abordado pela tese corrente. Alguns exemplos de outras distâncias são Bhattacharrya [46], Cityblock, Manhattan, Jeffrey-Matusita, etc.

Dentre as diferentes formas de se medir a distância entre várias classes no espaço de características, pode-se citar o caso de classes ou agrupamentos representados por amostras dessas várias classes; a definição de distância se baseia nas amostras disponíveis e é definida em função da distância entre os centróides das classes ou entre cada um dos elementos das classes.

4.2.3 Distância entre os centróides das classes

A distância entre o i-ésimo e o j-ésimo agrupamento ou classe é definida como a distância entre os respectivos centróides dados pelos vetores $\vec{x}_i$ e $\vec{x}_j$, como representado na Eq. 4.7. Esse é o caso de consideração de um protótipo por classe.

$$d_{i,j} = d(\vec{x}_i, \vec{x}_j) \tag{4.7}$$

4.2.4 Distância média entre os elementos das classes

A distância entre duas classes ou agrupamentos é definida como sendo a média das distâncias de todos os pares de elementos, tomados um de cada classe ou agrupamento. É calculada a distância (Eq. 4.8) entre todos os elementos das classes C_i e C_j, ou seja, são considerados vários protótipos por classe.

$$d_{i,j} = \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{m=1}^{n_j} d(\vec{x}_{ik}, \vec{x}_{jm}) \tag{4.8}$$

4.3 Coeficiente de correlação

A seleção de variáveis independentes muitas vezes é feita utilizando-se a matriz de coeficientes de correlação (R) [16, 45] definida na Eq. 4.9, onde x e y são dois atributos, x_i e y_i denotam um conjunto de dados com N elementos, cujos respectivos valores médios correspondem a $\bar{x}$ e $\bar{y}$:

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{4.9}$$

4.4 Abordagem *wrapper*

A utilização de classificadores como função critério para decidir quais serão as características selecionadas é chamada de método *wrapper* de seleção de características. A abordagem *wrapper* para seleção de características foi introduzida em 1994 por John et al. [52] e Kohavi e John, em 1997 [53]. Já o esquema de aprendizado seletivo usando o classificador bayesiano ingênuo foi introduzida por Langley e Sage em 1994 [48].

A idéia do algoritmo *wrapper* é funcionar como um "embrulho"envolvendo o algoritmo de indução, ou seja, esse algoritmo de seleção conduz a busca de um bom subconjunto usando o próprio algoritmo de indução como parte da função critério (ou função de avaliação) [52]. Em consequência, têm-se as características mais relevantes e apropriadas para um certo classificador.

O processo de seleção de características é uma etapa na mineração de dados, onde os dados passam por uma "filtragem", ou seja, o tamanho dos dados é reduzido de forma que apenas as variáveis mais relevantes permaneçam.

4.5 Mineração de dados

Houve um crescimento explosivo na capacidade de gerar e coletar dados nos últimos anos, porém o desenvolvimento de ferramentas para transformar esses dados em conhecimento não ocorreu na mesma velocidade. O método tradicional de transformar dados em conhecimento é aquele baseado em análise manual e interpretação, que é lento, caro e altamente subjetivo. Com o crescimento no volume de dados, a análise manual é impraticável em muitos domínios.

Tabela 4.1: Estatística como componente principal na mineração de dados [6]

Área da estatística	Descrição de atividades
Projeto experimental e amostragem	Como selecionar casos se há liberdade de escolha
Análise exploratória de dados	Geração de hipóteses ao invés de teste de hipóteses
Gráficos estatísticos	Visualização de dados
Modelagem estatística	Técnicas de regressão e classificação
Inferência estatística	Técnicas de estimação e predição

Técnicas avançadas para análise inteligente de dados ainda não atingiram sua maturidade, fazendo com que a estatística (Tabela 4.1) venha vivenciando um período de renascimento. O campo da mineração de dados assim como a estatística, está ligado ao "aprendizado a partir dos dados"ou "transformação dos dados em informação". Porém, os mineradores de dados estão frequentemente mais interessados no entendimento do que na precisão ou predição em si. Estatísticos têm desenvolvido uma grande infra-estrutura (teoria) para suportar seus métodos e uma linguagem (cálculo de probabilidades) para descrever sua abordagem, quantificando as incertezas associadas com a formulação de inferências a partir dos dados. Estes métodos

nos possibilitam descrever as relações entre variáveis para predição, efeitos de quantificação ou sugerindo caminhos causais [6].

A descoberta de conhecimento em dados (KDD) está ligada ao mapeamento de dados de baixo nível em outras formas que sejam mais compactas, mais abstratas e mais úteis; exemplificando, dado um bando de dados de um hospital com imagens de células de doentes, é necessário emprego de KDD para se fornecer estatísticas a respeito da ocorrência de determinada leucemia em pessoas acima de 40 anos. A mineração de dados é a aplicação de algoritmos específicos para extrair padrões dos dados, voltando ao exemplo do hospital, teríamos a classificação das imagens de células para fornecimento estatísticas. A Fig. 4.1 ilustra o processo de KDD e a mineração de dados como um de seus processos.

Mineração de dados é o processo que usa uma variedade de ferramentas para análise de dados para descoberta de padrões e relacionamentos nos dados que podem ser usados para fazer predições válidas. Ela é a componente determinante no processo de descoberta de conhecimento e visa encontrar regras descrevendo propriedades dos dados, ocorrência de padrões, agrupamentos de objetos em grandes bases de dados, normalmente residentes em bancos de dados especializados chamados *data warehouse*.

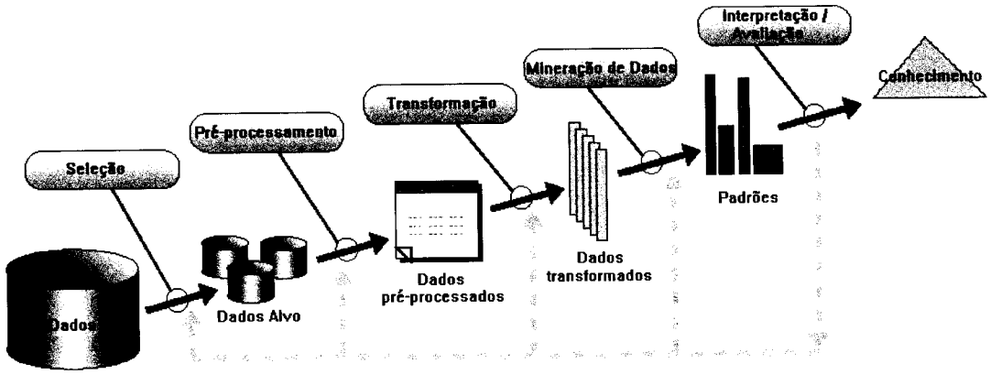


Figura 4.1: Passos de um sistema de descoberta de conhecimento em base dados [5].

Vários autores definem fases para a descoberta de conhecimento [45, 54], cujos principais passos são:

1. Identificação dos objetivos;

2. Construção do banco de dados para mineração: envolvendo coleta de dados, descrição dos dados, seleção, avaliação da qualidade dos dados e limpeza, consolidação e integração, construção de metadados, carregamento dos dados no banco de dados (BD) de mineração, manutenção dos BD de mineração;
3. Exploração dos dados;
4. Preparação dos dados: seleção de atributos, amostragem, construção de novos atributos e transformação dos atributos;
5. Construção do modelo;
6. Avaliação e interpretação: validação do modelo e validação externa;
7. Desenvolvimento do modelo e resultados;
8. Acompanhamento.

Os passos na descoberta de conhecimento para o reconhecimento de leucócitos serão retomados no capítulo de resultados, indicando o grau de evolução do corrente projeto de doutorado até o momento em cada uma das etapas mencionadas.

Capítulo 5

Resultados

"Quando se pode medir o que se fala e expressa-se em números, você sabe alguma coisa sobre isso; mas quando você não pode medi-lo, quando você não pode expressá-lo em números, seu conhecimento é escasso e insatisfatório. (Lord Kelvin)"

Segundo D'Onofrio e Zini [11], instrumentos que desempenham contagem diferencial automática devem ser capazes de contar e classificar os cinco tipos de leucócitos presentes em sangue periférico normal. A precisão é uma propriedade fundamental da análise quantitativa e fundamenta-se na reprodutibilidade dos dados analíticos. Desse modo, a precisão é interpretada através da estimativa de variação quando, repetidas vezes, medidas sejam tomadas na mesma amostra. Tem-se assim um coeficiente de variação (CV). A precisão do método de contagem diferencial de leucócitos normais por análise visual ao microscópio é aceitável somente para as classes mais numerosas, ou seja, neutrófilos e linfócitos (CV = [5,10]%). Para os monócitos, o CV é de 25%, para os eosinófilos é de 40% e para os basófilos é de 95%. Esta imprecisão é atribuída a vários fatores: o número reduzido de células categorizadas manualmente ao microscópio, acarretando um erro de amostragem aumentado; a distribuição irregular de células no esfregaço com excesso de monócitos nas bordas e na cauda e de linfócitos no centro do esfregaço; inconsistência e erros na identificação como resultado da subjetividade, tanto técnica quanto psicológica, em função do treinamento ao qual se submeteu o analista e sua motivação ao realizar tarefa tão repetitiva. Métodos automáticos são geralmente melhores, podendo apresentar coeficientes de variação na faixa de 4% para neutrófilos e linfócitos, 10% para

eosinófilos e monócitos e menos de 50% para basófilos. A análise de imagens digitais tem vantagens e desvantagens quando comparadas com a análise de lâminas de vidro ao microscópio. Enquanto as lâminas simulam a distribuição real das células do sangue para identificação pelo laboratório, a morfologia e as propriedades de coloração podem deteriorar, conduzindo a interpretações errôneas. Enquanto o uso de micrografias para identificação de células tem a desvantagem de mostrar um campo limitado do esfregaço de sangue, a imagem da célula manterá suas propriedades indefinidamente, possibilitando comparações da prática morfológica entre indivíduos e entre laboratórios [1].

As seções que compõem esse capítulo descrevem as ferramentas computacionais projetadas e desenvolvidas e os resultados experimentais obtidos utilizando-as para reconhecer as imagens fornecidas pelo Laboratório de Hematologia de Ribeirão Preto.

5.1 Ferramentas propostas

O corrente trabalho de doutorado aplicou conceitos de representação matemática de imagens digitais, extração de medidas de imagens com objetivo de modelar os leucócitos para realizar seu reconhecimento e classificação. O programa para segmentação, extração de medidas e classificação foi chamado de *Leuko* (Fig. 5.2). Uma ferramenta para mineração de dados (Fig.5.4) que processa vetores de características gerados pelo *Leuko* foi desenvolvida, contudo é independente da aplicação, possibilitando investigação de qualquer tipo de problema que seja representável por vetores de características. Ambos os programas foram desenvolvidos no *Delphi 4.0* para plataforma *Windows*.

5.1.1 LEUKO

Em colaboração com pesquisadores e docentes do Laboratório de Hematologia de Ribeirão Preto, observou-se a utilidade da aplicação de análise quantitativa automática em imagens médicas. Um sistema de classificação automática de leucócitos permitiria quantificações mais objetivas e padronizadas a custo mais baixo. Uma outra utilidade potencial de tal projeto de pesquisa é como módulo educativo para

treinamento. As vantagens desse sistema são auxiliar o estudante a identificar características importantes que devem ser consideradas na análise visual e diminuir a necessidade de manutenção de arquivo pessoal de lâminas uma vez que o programa lida com imagens digitais.

O funcionamento do programa é representado pelo diagrama de blocos da Fig.5.1, onde os blocos acinzentados (2, 6 e 7) indicam a necessidade de execução do processo ao menos uma vez. Ainda que seja desejável que a interação com o usuário seja mínima, na maioria dos casos, a classificação supervisionada em problemas de diagnóstico auxiliado por computador é a opção mais apropriada. A segmentação em *Leuko* se dá através da cor, com classificação supervisionada, embora consideremos essa etapa semi-automática pois algumas imagens precisam de ajuste manual.

Uma motivação ao desenvolver um software com interface amigável (Fig.5.2) para pré-diagnóstico foi quantificar os atributos utilizados pelos humanos e mostrá-los através de números em planilha. A planilha com medidas, mostrada à esquerda da Fig.5.2, permite atribuir valores a características antes consideradas apenas de forma qualitativa, como é o caso de textura e cor. Seria interessante que o nome físico das medidas extraídas fosse substituído por descrições mais intuitivas, de acordo com jargão médico.

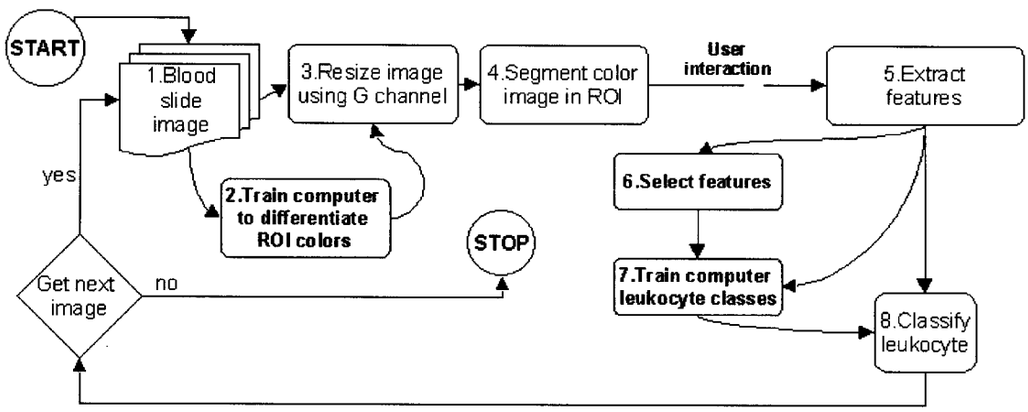


Figura 5.1: Diagrama de blocos do programa *Leuko*: blocos cinzas representam processos que podem ser executados apenas uma vez.

A distribuição de probabilidade gaussiana, ponto inicial em estudos de modelagem, foi adotada para modelar as classes, tanto pra segmentação de imagens baseada

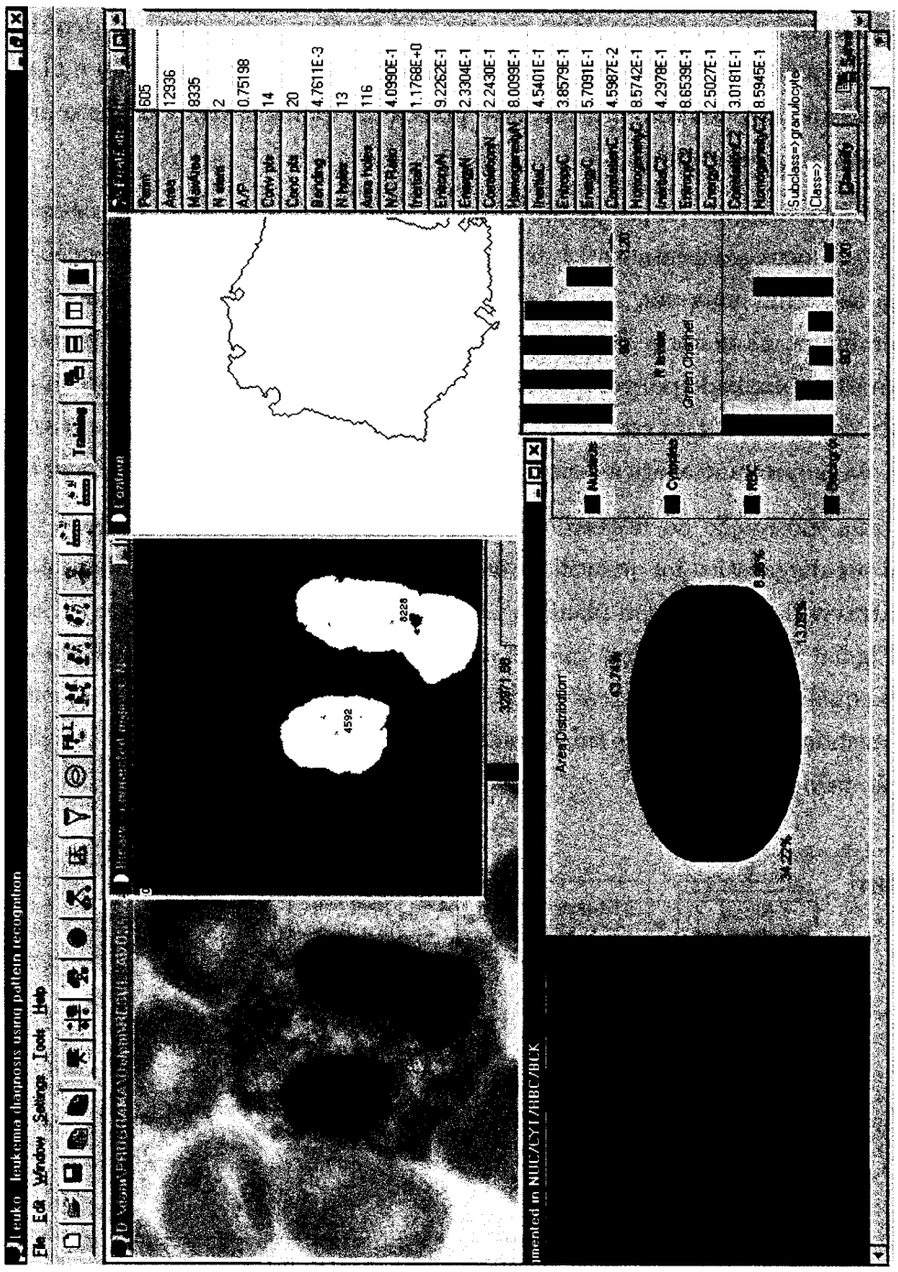


Figura 5.2: Interface do programa *Leuko*: sistema para reconhecimento de leucócitos desenvolvido em Delphi 4.0.

em cor quanto para o reconhecimento de células. Assim, cada classe é representada por uma gaussiana multivariada no espaço de características.

Diversas publicações consideram imagens segmentadas manualmente [37, 55, 56], embora haja quem omita tal intervenção. É importante a determinação de etapas em que a interação com o humano seja imprescindível, como é o caso do programa em desenvolvimento onde legamos ao usuário a definição dos protótipos (amostras classificadas) para que o programa tome decisões futuras, além da disponibilização de ferramentas para ajustes do resultado de segmentação integradas em um único software.

5.1.2 Ferramenta para seleção de características

Embora tenha sido desenvolvida para trabalhar com os arquivos de características gerados pelo programa *Leuko*, a ferramenta de seleção de características (Fig.5.4) é de propósito geral e pode não ser intuitiva para seu uso por médicos. Ela foi desenvolvida para auxiliar projetistas a selecionar quais os atributos mais relevantes para resolução de problemas de reconhecimento de padrões. Assim, pode ser usada para investigação de quaisquer casos que envolvam vetores de características. Para comparar as características, utilizou-se o processo de normalização (Eq.3.40) explicado na seção 3.3.

Esta ferramenta foi importante para mineração de dados, pois permitiu realizar uma triagem e estimar o subconjunto de medidas mais relevantes para caracterização das classes de leucócitos consideradas. Notou-se que tanto parâmetros de forma quanto cor e textura são importantes no processo de diferenciação, conforme se mostrou em publicação sobre esse programa [9].

O programa permite a execução de 14 tipos de técnicas de buscas diferentes nas amostras, dadas todas as combinações possíveis entre abordagens de busca, algoritmo de busca, critério de separabilidade e distância entre pontos. Detalhes sobre o funcionamento do programa são descritos no Apêndice D.

Diversos programas comerciais foram utilizados durante o desenvolvimento das pesquisas e dos softwares. Através do programa Matlab, gráficos, testes e protótipos foram gerados, implementando-se diferentes algoritmos para segmentação das células, além de permitir vislumbrar qual seria a interface mais amigável de acordo com

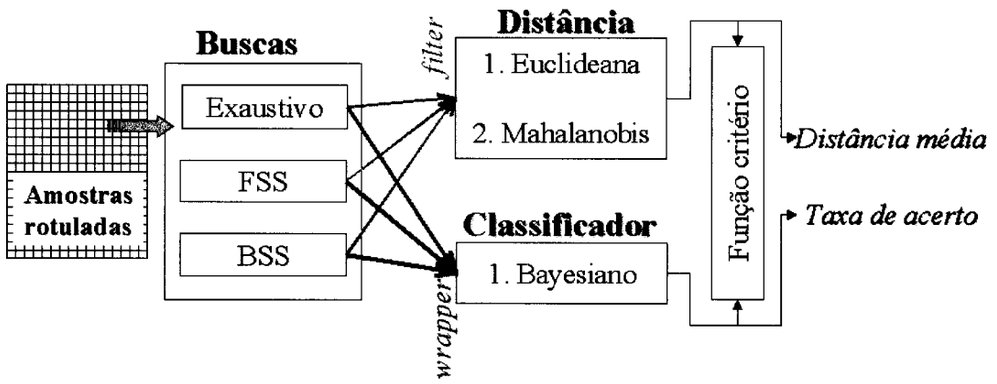


Figura 5.3: Diagrama de blocos da ferramenta para seleção de característica.

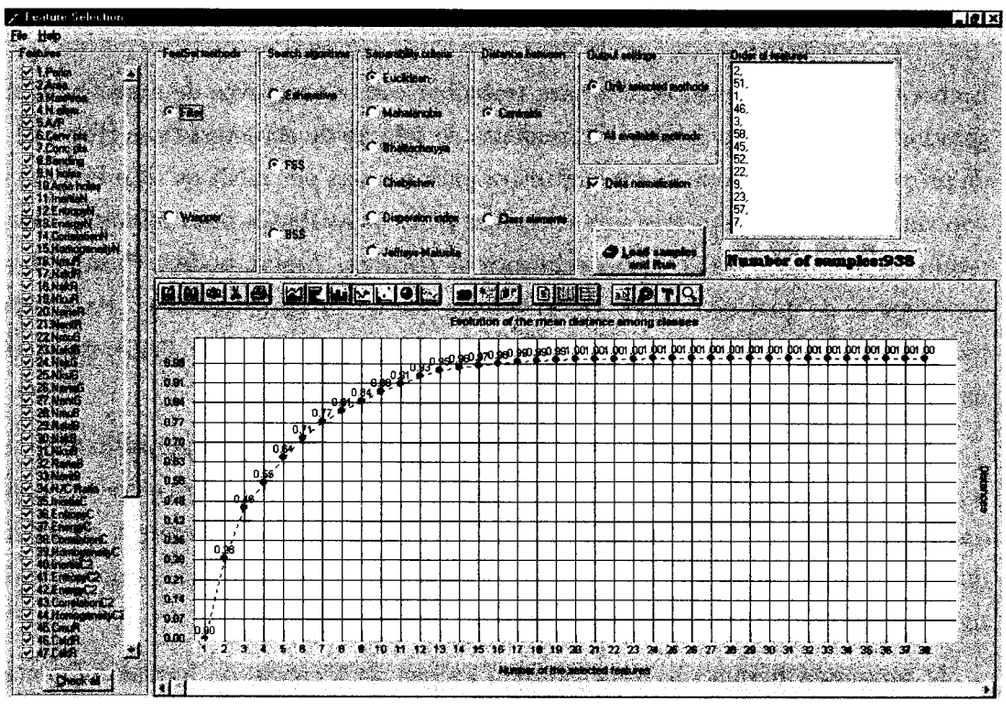


Figura 5.4: Ferramenta para seleção de característica de propósito geral desenvolvida em Delphi 4.0.

critérios sugeridos pelos possíveis usuários do programa. Outro programa particularmente usado na fase final do trabalho foi o Weka, um software para mineração de dados contendo todas as principais implementações de algoritmos de aprendizado de máquina e seleção de características [41]. Graças a essa ferramenta, encontrou-se

uma maneira simples de se testar diversos classificadores sem necessidade de implementar todos eles. Maiores detalhes são apresentados no Apêndice D.

5.2 Resultados experimentais

Experimentos foram conduzidos usando as imagens adquiridas no Laboratório de Hematologia do Hospital das Clínicas de Ribeirão Preto. Foi desenvolvido um programa para processamento, análise e classificação dos leucócitos denominado *Leuko*, cujos resultados foram publicados em [57, 58]. O resumo submetido ao Congresso da Sociedade Brasileira de Microscopia e Microanálise foi selecionado para publicação em versão completa em suplemento da revista Acta Microscópica [59].

Avaliou-se a taxa de acerto no reconhecimento de leucócitos normais utilizando diversas medidas da célula. Primeiramente foram testados apenas atributos de forma e tamanho dos leucócitos e verificou-se ótima separabilidade entre dois grandes grupos: leucócitos mononucleados e polimorfonucleados. Contudo, a subclassificação, em basófilo, eosinófilo, linfócito, monócito e neutrófilo era muito imprecisa [60]. A inclusão de doenças linfoproliferativas utilizando apenas atributos de forma-tamanho levou a uma alta taxa de neoplasias classificadas como linfócitos normais. Alternativamente à abordagem bayesiana de classificação, testou-se a técnica de agrupamento usando modelo superparamagnético, embora os resultados não tenham sido promissores [61].

A dimensão fractal 3-D foi utilizada como medida de textura da cromatina [62], entretanto foi substituída por atributos de textura extraídos das GLCM, mais abrangentes e computacionalmente menos custosos, sendo documentados através de resumo submetido a Conferência Internacional em Bioinformática e Biologia Computacional [63]. Esse resumo foi também selecionado para versão completa com submissão à revista Real-Time Imaging [64], artigo recentemente publicado.

Com um número de atributos grande em comparação ao número de imagens disponíveis, iniciou-se o desenvolvimento de uma ferramenta para seleção de características e investigação de estratégias para busca do subconjunto ideal de atributos para o problema de classificação dos leucócitos [65].

Finalmente, os atributos de cor, apresentados na seção 3.3.2, foram introduzidos

e uma ferramenta consagrada conhecida como Weka passou a ser utilizada para validar as taxas de acerto apresentadas por *Leuko*. Resultados importantes foram registrados através do artigo premiado em *2004 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, que descreve experimentos usando *Leuko* e Weka no reconhecimento de 6 tipos de células diferentes: os 5 tipos de leucócitos normais e a leucemia LLC.

Atualmente, é possível classificar oito diferentes tipos de células como os leucócitos normais maduros basófilo, eosinófilo, linfócito, monócito, neutrófilo e células leucêmicas como as de leucemia linfóide crônica, tricoleucemia e leucemia prolinfocítica.

Vários artigos foram publicados documentando os passos de desenvolvimento do *Leuko*. A publicação de artigos concomitante aos desenvolvimentos permitiu registrar detalhes para reprodutibilidade, como o pré processamento, segmentação, extração de características, seleção de características e classificação. As próximas seções visam descrever o desenvolvimento da pesquisa, começando com a etapa de pré processamento usando limiarização e filtros morfológicos, segmentação por cor até a classificação da célula, salientando as contribuições deste projeto de doutorado. Remete-se o leitor ao Apêndice E para pormenores a respeito do conteúdo dos artigos publicados, cujas principais contribuições serão descritas ao longo deste capítulo.

5.3 *In vitro* para *in silico*

Nesta seção descreveremos os passos desde a captura da imagem ao microscópio até a representação do leucócito para fins de extração de medidas. Várias técnicas de processamento de imagens foram testadas e serão comentadas nesta seção, embora algumas abordagens não tenham sucedido e, portanto, não foram consideradas na implementação do software *Leuko*.

5.3.1 Micrografias

Imagens de esfregaços de sangue periférico corados com Leishman (Sigma Co.) foram capturadas usando uma câmera digital Kodak DC290, 15 pixels por mícron, aproximadamente, zoom: 115 mm, exposição: 1/30 segundos, 2,1 Megapixels, resultando em imagens coloridas de 24-bits, tamanho 720 por 480, resolução 0,06 μm/pixel.

A câmera é acoplada a um microscópio óptico Zeiss com magnificação de objetiva de 100X (imersão em óleo), numerical aperture: 1,25, ocular: 10X. A resolução entre 0,05 e 0,5 μm garante que se possa descrever aspectos de textura de cromatina, segundo Rodenacker [31]. Todas as imagens foram adquiridas usando o mesmo microscópio no intuito de diminuir as variações entre as imagens.

5.3.2 Pré-processamento

Procurou-se normalizar o brilho das imagens usando o algoritmo retinex, discutido na seção 3.1.2, através do qual se teoriza que seria possível generalizar os arquivos de treinamento de cor para segmentação de quaisquer células. Embora normalize o fundo, ou seja, o fundo de diferentes imagens seja bastante parecido após filtragem, a região correspondente ao leucócito sofre perdas irreparáveis. Exemplos são mostrados no Apêndice C, mas o retinex não foi incorporado ao software.

Em busca de homogeneização de regiões com cores semelhantes, testou-se filtros gaussianos, que se mostraram inadequados uma vez que tornam o contorno do leucócito impreciso. O filtro de difusão anisotrópica pareceu visualmente interessante pois diminui detalhes internos à região citoplasmática, como grânulos, e salienta as bordas. Acredita-se que essas características permitiriam ao algoritmo de segmentação melhor separação da região citoplasmática, com menos buracos e mais precisão na região do contorno. Resultados preliminares, mostrados no Apêndice C, indicam que o método de segmentação por cor sem consideração de adjacência não é adequado para ser aliado a esse processamento de imagem. Não se incorporou esse processamento ao software, pois não houve ganho, em termos da segmentação, segundo avaliação utilizando a porcentagem de pixels corretamente classificados. Uma abordagem interessante seria testar métodos de clusterização para segmentação por cor [66] em conjunto com filtros de difusão anisotrópica sobre os canais cromáticos e acromáticos.

5.3.3 Segmentação

Embora se tenha testado filtros para pré-processamento, eles não foram integrados em *Leuko* pois não comprovamos que tais filtros oferecessem melhoria suficiente; isso não significa que os filtros não são bons, mas sim que o método de segmentação por

cor adotado tem várias limitações e sua execução após filtragem foi menos eficiente que a segmetnação das imagens originais.

Dadas as imagens digitais RGB, rastreou-se o leucócito em cada micrografia, que pode conter hemácias, plaquetas, leucócitos e fundo de lâmina. Adotou-se que a melhor maneira de achar o leucócito seria encontrando o núcleo, portanto se passou a utilizar a técnica discutida na Seção 3.2.1: a limiarização visa delimitar e melhorar a segmentação por cor, definindo uma porção específica da imagem e menor que a imagem original. Os leucócitos podem ser segmentados através da análise do histograma do canal G. A Fig.5.5 mostra nitidamente a distribuição bimodal para o canal G da imagem RGB, onde à direita do gráfico se agrupam os pixels pertencentes ao núcleo dos diferentes tipos de leucócitos normais.

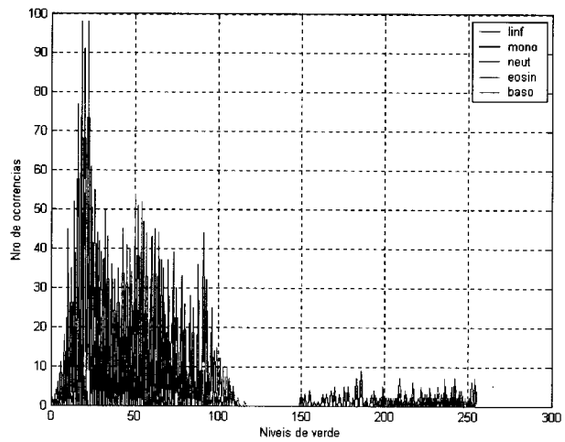


Figura 5.5: Histograma do canal G de imagens de leucócitos normais: valores menores que 100 indicam pixels do núcleo.

Considerando o trabalho desenvolvido por Katz [25], observou-se que a componente verde da imagem é o melhor discriminante entre o púrpura do material nuclear e o resto da imagem (o púrpura claro a cor-de-rosa ou vermelho de citoplasma, o cor-de-rosa das hemácias e o amarelo pálido do fundo). O motivo do canal verde carregar a informação do núcleo deve-se ao fato do púrpura ser variações de magenta, ou seja, combinação de valores altos para o canal R e B e valores baixos ou nulos para G.

O relatório de Katz sugere um limiar $T = 100$ no histograma do canal G, considerando o intervalo de 0 a 255 de modo que todo pixel com valor menor que T é

considerado material pertencente ao núcleo. Buscando um limiar adaptativo, que implicasse em menor dependência da iluminação, observou-se que o limiar poderia ser definido num intervalo conforme se segue:

1. Calcular o histograma do canal G da imagem para caixas de tamanho 2 em um intervalo de 0 a 255;
2. Selecionar o nível de verde, T, cuja taxa de ocorrência seja mínima no intervalo compreendido entre $[100, 200]$;
3. Limiarizar o canal G usando T: dada a imagem do canal G, se o pixel tem valor menor que T, então pertence ao núcleo.

Para idealização de tal algoritmo, testaram-se diversos tamanhos de caixa para o histograma, mas o tamanho 2 foi o que, empiricamente, apresentou melhores resultados. O intervalo de $[100, 200]$ nos níveis de verde foi encontrado através de análise de vários histogramas de imagens de leucócitos.

Independente do método de limiarização adotado, a imagem limiarizada pode conter, além do material nuclear, plaquetas, que também são coradas de púrpura no protocolo de marcação May-Giemsa-Grunwald. Contudo, estas podem ser facilmente removidas através de processos de erosões seguidos de dilatações ou por crescimento de regiões usando a área dos elementos que compõe a imagem como parâmetro de exclusão.

A partir da imagem original colorida (Fig.5.6), obteve-se a Fig.5.7, segundo o algoritmo de Katz e a Fig.5.8 foi obtida usando limiar adaptativo por meio de busca de T no intervalo de $[100, 200]$. Comparando o resultado do algoritmo de Katz com o algoritmo de adaptação pra busca de T num intervalo, umas das contribuições dessa tese, verificou-se que o método proposto é melhor pois acarreta em menor número de buracos espúrios, ou seja, porções de núcleo não detectadas devido a variações na coloração da cromatina (heterocromatina). Essa variações são decorrentes de menor espiralização da cromatina, ficando assim com aspecto mais claro (menor concentração de corante) e menos denso. Apesar da maioria das imagens possuir o mínimo (vale entre picos da distribuição bimodal) de histograma no intervalo proposto, esse não é o caso dos prolinfócitos (Fig.5.9). Nesses casos, tanto o método de Katz quanto o limiar adaptativo não eram eficientes pois o limiar ótimo não era

100 nem era o mínimo do intervalo [100, 200]. Buscando um algoritmo mais genérico para limiarização, testou-se métodos de entropia e observou-se um novo método de limiarização.

Limiarização por entropia e LogCum

A necessidade de se buscar outros algoritmos de limiarização ocorreu principalmente após a inclusão dos prolinfócitos, cujo núcleo pode apresentar cromatina mais frouxa. Assim, os valores de nível de verde correspondentes ao núcleo passam a ser maiores (mais claros).

Visualmente, métodos de limiarização podem desempenhar de maneira bastante semelhante conforme ilustrado nas Fig.3.4 e 3.5 e no Apêndice B, contudo é importante que haja uma avaliação quantitativa [67]. Avaliou-se os métodos de entropia e o LogCum em termos do número de pixels corretamente classificados (PCC) e tempo de execução dos algoritmos, resultados publicados em [29].

Métodos entrópicos envolvem cálculos altamente iterativos, diferente do método proposto LogCum, ainda que o resultado de limiarização seja bastante semelhante (Fig.5.1). O método original proposto apresenta o melhor custo computacional segundo Tabela 5.2.

Tabela 5.1: Pixels corretamente classificados: média, desvio padrão e mediana para algoritmos de limiarização no processamento de 20 imagens de leucócitos.

PCC	**Métodos**		
	LogCum	Kapur	Sahoo
Média	0.93	0.94	0.94
Desvio Padrão	0.04	0.04	0.04
Mediana	0.94	0.95	0.94

Tabela 5.2: Tempo médio, desvio padrão, mediana e máximo de CPU em segundos para 20 imagens de leucócitos.

Tempo de CPU (s)	**Métodos**		
	LogCum	Kapur	Sahoo
Média	0.0135	0.0744	0.0767
Desvio padrão	0.0049	0.005	0.005
Mediana	0.01	0.071	0.08
Máximo	0.02	0.081	0.081

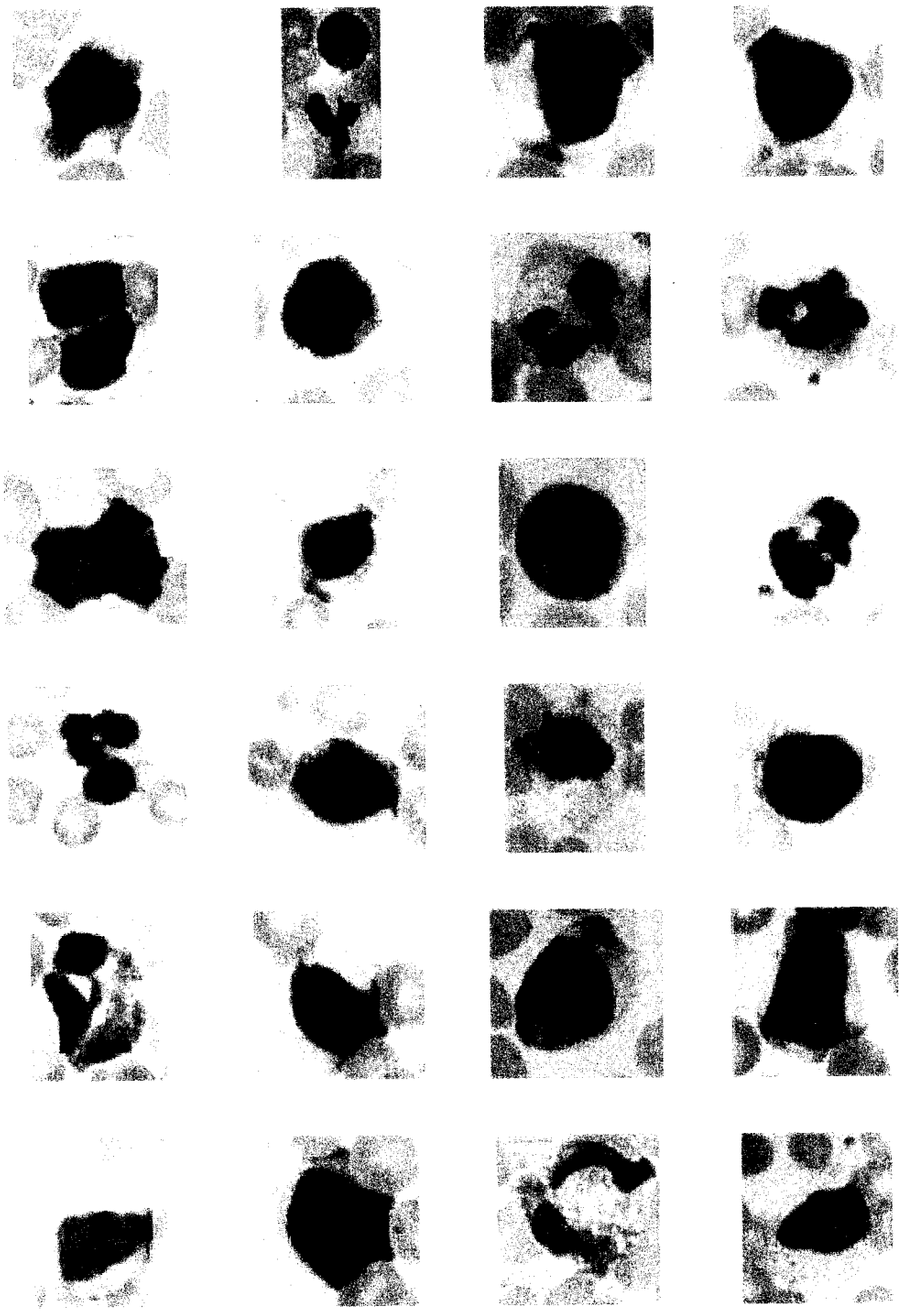


Figura 5.6: Imagens de vários tipos de leucócitos

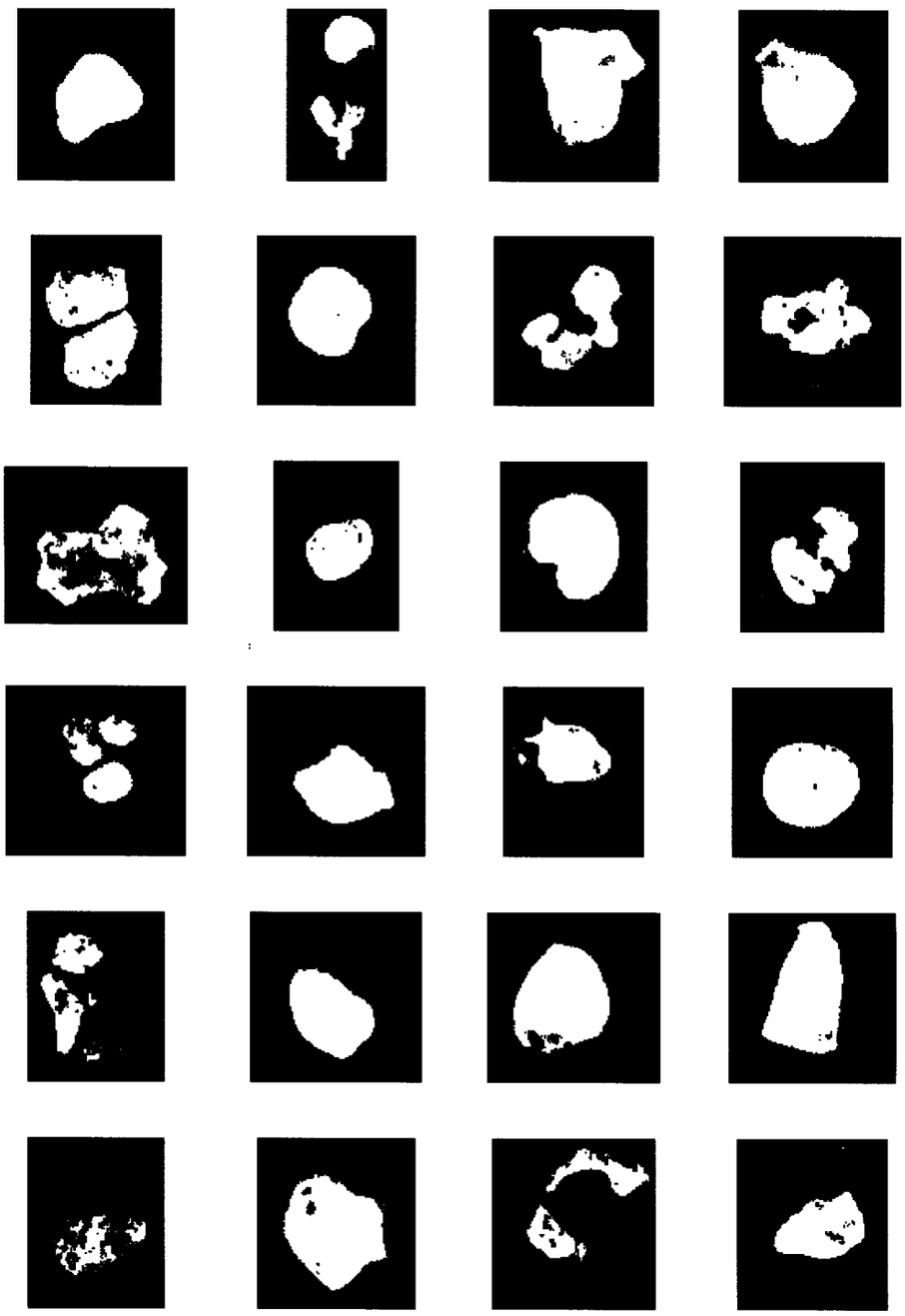


Figura 5.7: Segmentação do núcleo segundo Katz: limiar fixo no histograma do canal G (T=100).

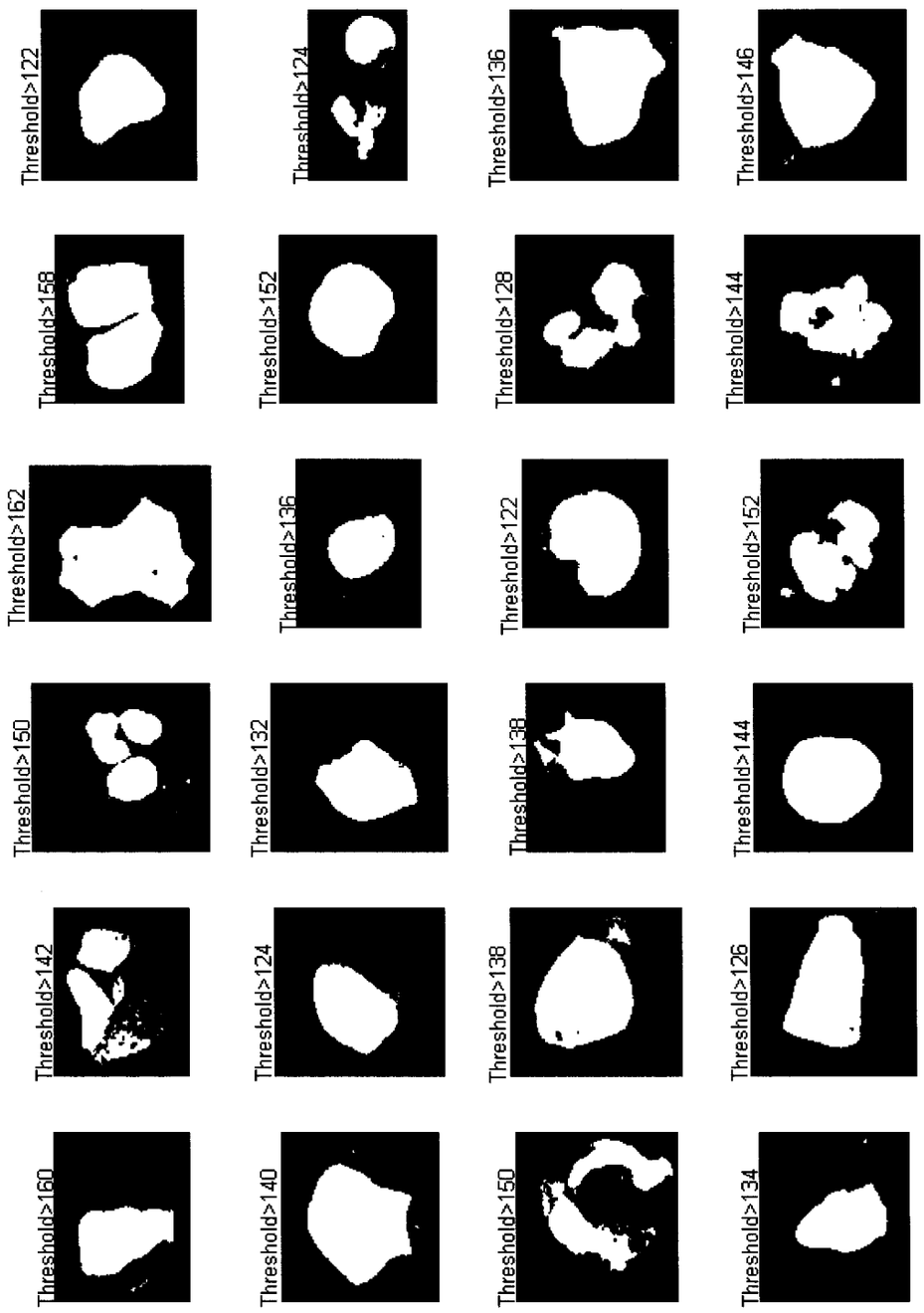


Figura 5.8: Segmentação do núcleo limiar adaptativo (busca no intervalo $[100, 200]$).

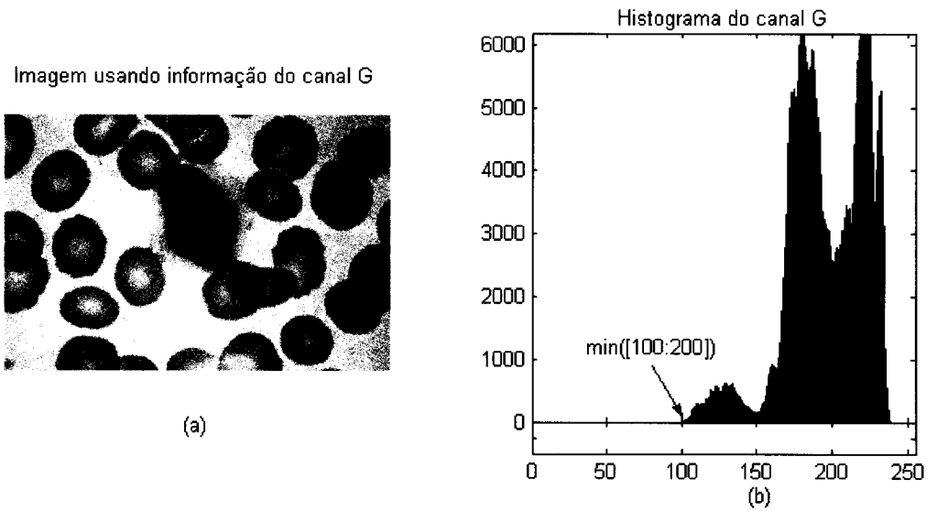


Figura 5.9: Limiarização dos prolinfócitos: (a)canal G do prolinfócito; (b)histograma do canal G, mostrando que o mínimo calculado pelo algoritmo adaptativo é 100 e não 150.

Pós-processamento

Filtros morfológicos como erosão e dilatação foram aplicados após a limiarização pois pode haver presença de plaquetas e espalhamento de pequenos pontos, os quais podem ser eliminados através de erosões da imagem. Para compensar alterações de área dos objetos, operações de dilatação são realizadas em mesmo número das erosões. Uma outra utilidade desses filtros é o pré-processamento voltado a realização de medidas, como utilizar várias erosões de modo a determinar a variação no número de objetos. Uma aplicação interessante seria utilizar essa variação como parâmetro de diferenciação dos núcleos de leucócitos.

Núcleos polimorfonucleados são sensíveis a operações de erosão uma vez que são granulados e apresentam lóbulos ou segmentos, como é o caso dos eosinófilos, neutrófilos e basófilos. Alguns pequenos pontos entre partes de um mesmo núcleo frequentemente indicam presença de finos elongamentos de cromatina que unem partes desse mesmo núcleo. Núcleos lobulados são ligados por fios de cromatina e podem parecer desconectados após a limiarização.

Inclusão de semântica

Após a filtragem, passa-se à busca do material nuclear, ou seja, dos possíveis pedaços de núcleo (bolotas[1]) na imagem binária resultante. Estabelece-se um critério para determinar quais pedaços fazem parte de um mesmo núcleo, segundo os passos seguintes, incluindo contribuições tanto de Katz [25], quanto originais, propostas por esse projeto de pesquisa [64] :

1. Determinar o centróide da bolota, ou seja, (cx, cy), que é a média aritmética dos valores das posições x e y, isto é, $cx = \sum_{i}^{n} x_i/n$ e $cy = \sum_{i}^{n} y_i/n$, onde n é o número de pixels da bolota - manter o (cx, cy) e n para cada bolota da imagem;

2. Consolidar bolotas adjacentes: tratar como uma única bolota quaisquer duas bolotas cuja separação seja menor que um certo valor, dada a distância D máxima entre duas bolotas de um núcleo. Baseando-se no valor sugerido por Katz (8 μm) e dados empíricos obtidos no corrente projeto de pesquisa, assume-se D=7,2 μm, equivalente a 120 pixels. Outra imposição é que o lóbulo tenha área maior que 100 pixels, caso contrário há grandes chances de que seja ruído da limiarização ou uma plaqueta. Assim, dadas duas bolotas ($B1$ e $B2$), a distância entre seus centróides deve ser menor que 120 pixels para serem consideradas parte de um mesmo núcleo (bolota consolidada). O centróide (cx, cy) da bolota consolidada é determinado por:

$$cx = \frac{cx_{B1} * n_{B1} + cx_{B2} * n_{B2}}{n_{B1} + n_{B2}} \tag{5.1}$$

$$cy = \frac{cy_{B1} * n_{B1} + cy_{B2} * n_{B2}}{n_{B1} + n_{B2}} \tag{5.2}$$

O centro da bolota resultante e seu número total de pixels ($n_{B1} + n_{B2}$) são mantidos para as próximas consolidações com quaisquer outras bolotas à distância apropriada. Um problema da consolidação é que dois leucócitos muito próximos podem ser tomados como um único leucócito. A distância entre os centróides, em geral, ultrapassa a distância máxima adotada. Considerando a área máxima da região nuclear, dificilmente dois núcleos de linfócitos (núcleo grande em relação aos polimorfonucleados) seriam consolidados numa mesma bolota (Fig.5.10);

[1]Tradução de *blobs*.

3. Recortar uma subimagem: uma ROI é determinada ao redor do centróide da bolota e uma subimagem contendo o leucócito é recortada da imagem original, reduzindo a imagem de 760 x 480 pixels para 400 x 400 pixels, na pior das hipóteses. O valor inicial da largura e altura igual a 400 pixels é aproximadamente 24 μm, o diâmetro da maior célula: o monócito;

4. a subimagem facilita a segmentação por cor pois reduz o número de pixels a ser varrido na imagem original para extração do objeto de interesse: o leucócito.

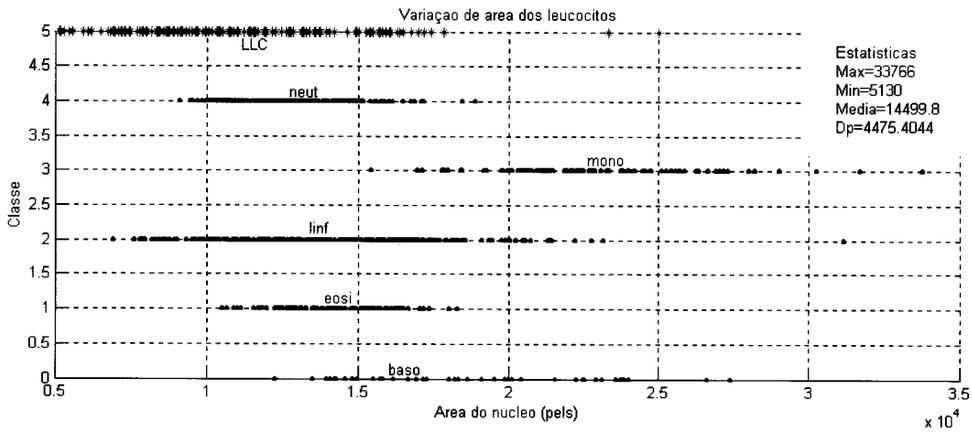


Figura 5.10: Área do núcleo (em pixels) de 938 leucócitos de 29 pacientes diferentes.

Segmentação por cor no espaço RGB

A segmentação por cor pode ser realizada tanto no espaço RGB quanto em outros espaços de cor como os espaços HSI e CIELab, mais intuitivo uma vez que expressam melhor a percepção de cores [14]. Entretanto, o RGB pode gerar resultados melhores. A transformação dos valores do RGB para outros espaços colorimétricos constitui um problema mal formulado, onde distorções podem ocorrer além de que a transformação entre espaços de cores só se justifica quando os resultados são melhores no espaço transformado do que no espaço original RGB. A Fig.5.11 ilustra tal questionamento e consiste em um exemplo onde a segmentação no espaço RGB é mais

apropriada, sendo computacionalmente mais rápida e melhor que a segmentação no espaço CIELab, que apresentou perturbações decorrentes da transformação. Esses testes foram realizados usando classificação supervisionada bayesiana e podem ser realizados usando o software *Leuko*, que permite segmentação tanto no espaço RGB quanto CIELab.

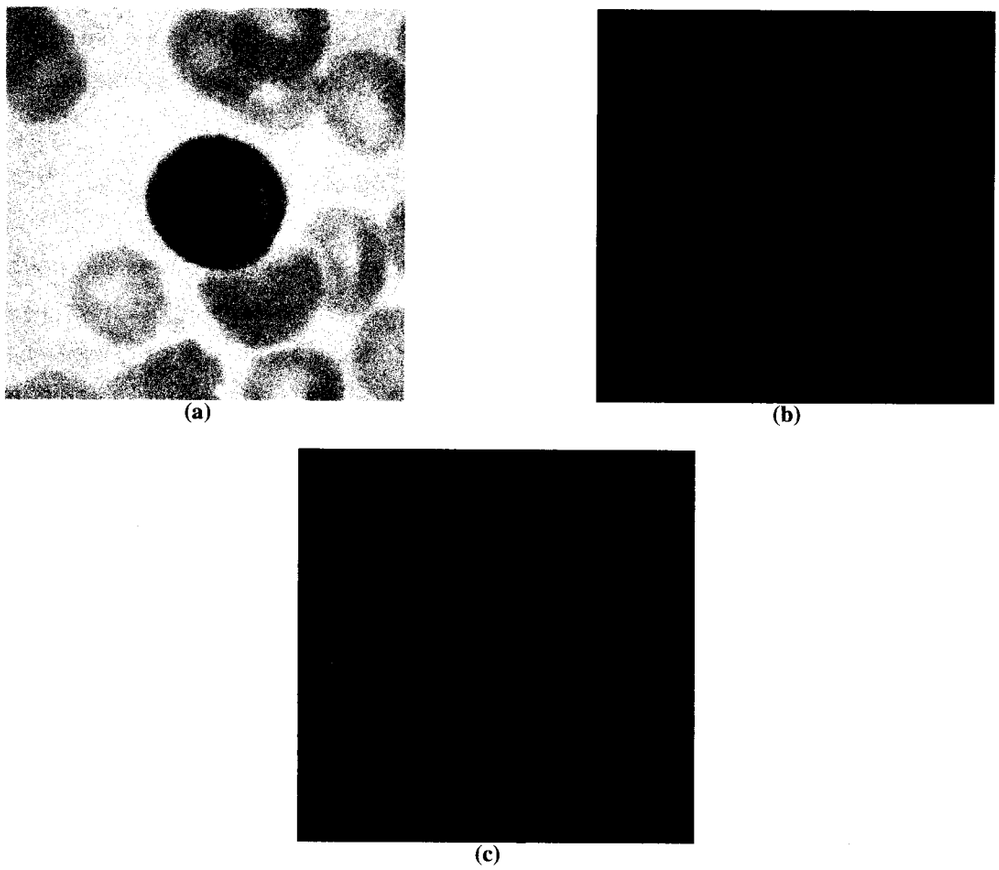


Figura 5.11: Exemplo de segmentação por cor: **(a)** imagem original, **(b)** imagem segmentada usando espaço de cor RGB, **(c)** imagem segmentada usando espaço de cor CIELab.

Outra abordagem de segmentação utilizada foi aquela proposta por Sumengen et al [68], onde a segmentação de imagem utiliza um campo de fluxo baseado na imagem. Métodos de evolução de curva geométrica requerem uma função de borda e um campo vetorial com certas características que são obtidas da própria imagem. Dessa forma, o campo vetorial é utilizado na obtenção tanto da função de borda

quanto para guiar a evolução da curva em direção das bordas do objeto. O campo vetorial é calculado a partir da imagem, considerando características de intensidade, textura e cor. O programa de evolução de curva foi disponibilizado para testes, sendo que resultados preliminares são mostrados no Apêndice C.

Após a segmentação, segue-se a extração de características. Conforme o projeto de doutorado evoluiu, foi-se implementando a extração de diversos atributos, alguns tradicionalmente aceitos como é o caso da razão núcleo-citoplasmática e outros menos utilizados no estudo de leucócitos, como é o caso da energia de dobramento.

5.4 Diagnóstico através de medidas de forma

As características a respeito da forma da célula como diâmetro e área são pontos de partida naturais no reconhecimento, muito embora estas características sejam raramente decisivas na discriminação [31]. Os primeiros resultados de classificação obtidos consideravam apenas atributos da forma dos leucócitos normais e o método de resubstituição para avaliação da taxa de acerto. Inicialmente, optou-se pelo método de resubstituição devido à sua simplicidade de implementação e a capacidade desse algoritmo de lidar com pequeno número de amostras, embora as estimativas de erro sejam otimistas.

Houve alta taxa de erro na diferenciação entre os granulócitos (basófilo, eosinófilo e neutrófilo), ainda que o uso do método da resubstituição levasse a taxas de acerto tendenciosas. A matriz de confusão é mostrada na Tabela 5.3, onde a diagonal indica a taxa de acerto e as outras colunas a taxa de erro (ex: linha 1 representa os basófilos, onde 12 deles foram corretamente classificados, três foram classificados errados como eosinófilos e um como monócito) [60].

5.5 Estudo de hierarquia nos dados de leucócitos

Buscando estudar aspectos da forma de leucócitos sem supor conhecimentos a priori sobre a distribuição das medidas coletadas, investigamos as medidas das células usando a técnica de agrupamento superparamagnética [69]. O algoritmo utilizado se baseia numa analogia física utilizando um sistema magnético. Cada ponto de dado é considerado um spin de Potts e interações ferromagnéticas de curto alcance são

Tabela 5.3: Matriz de confusão dos leucócitos normais utilizando apenas atributos de forma e avaliação de taxa de acerto através do método de resubstituição.

Leucócitos	Baso.	Eosino.	Linf.	Mono.	Neut.	Taxa acerto
Baso.	**12**	3		1		75%
Eosi.	11	**9**			5	36%
Linf.			**25**			100%
Mono.			3	**22**		88%
Neut.	4	3	3		**15**	60%
Total	27	15	31	23	20	116

introduzidas entre os spins. A intensidade destas interações decai com a distância. O sistema é termodinâmico, apresentando diferentes regimes auto-organizáveis regulados pela temperatura.

Apresentou-se uma breve descrição do modelo e resultados preliminares quanto a sua aplicação ao estudo da hierarquia de classes dos leucócitos. Contudo os resultados não foram animadores uma vez que apenas a separação entre mononucleados (linfócito e monócito) e polimorfonucleados (basófilo, eosinófilo e neutrófilo) foi evidente. O estudo publicado em [61] considerou um número reduzido de características (somente três), entretanto estudos posteriores incluindo mais características não melhoraram os resultados. Essa abordagem requer futuras investigações.

5.6 Inclusão de medidas de textura

A caracterização de textura de imagens de sangue em níveis de cinza foi baseada em GLCM. O objetivo foi utilizar regiões pré-segmentadas (núcleos ou citoplasmas) e estudar as características texturais de cada região através das medidas propostas por Haralick [3, 17] para discriminação entre leucócitos normais. As Figuras 5.12 e 5.13 mostram regiões citoplasmáticas e nucleares, respectivamente, pré-segmentadas e processadas (quantização e equalização por equiprobabilidade). O resultado desse processamento será submetido à extração das GLCM e, posteriormente, dos atributos de textura: inércia, entropia, energia, correlação e homogeneidade.

Iniciou-se as investigações a partir do citoplasma como forma de suprir a deficiência de medidas dessa região, uma vez que a segmentação citoplasmática era ruidosa, portanto medidas de forma eram ineficientes. Notou-se que medidas glo-

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

bais de textura podiam fornecer informações essenciais a respeito da granularidade citoplasmática.

Resultados preliminares foram descritos em [65], dedicando-se exclusivamente a caracterização do citoplasma de uma base de imagens com 50 leucócitos (10 imagens por classe de leucócito normal). Dois estudos foram abordados:

1. Avaliação da escala mais apropriada para cada parâmetro: calculou-se cada parâmetro de textura para tamanho de janela (no cálculo da GLCM) variando entre [2,20] para toda a base de imagens. O intervalo [2,20] foi definido baseado no tamanho dos grânulos citoplasmáticos: aproximadamente, 5x5 e 15x15 pixels em linfócito e basófilo, respectivamente. As outras células possuem tamanho de grânulo intermediário.

2. Avaliação da relevância dos parâmetros em termos da taxa de acerto: verificou-se a separabilidade entre os 50 leucócitos em termos de cada parâmetro conforme ilustrado na Tabela 5.4.

Assumimos como janela **pequena** aquela entre [2,5] e como janela **grande** aquela entre [15,20], segundo as definições apresentadas no artigo [65].

Tabela 5.4: Parâmetros de textura para citoplasma: avaliação da escala mais apropriada e taxa de acerto usando método da resubstituição.

Parâmetro	Escala	Acerto(%)
Inércia	2	88
Entropia	19	86
Energia	4	90
Correlação	20	90
Homogeneidade	4	92

5.7 Inclusão de medidas de cor

Embora o número de características já somasse 26 quando incluindo medidas de forma e textura, resolve-se também utilizar medidas de cor uma vez que, segundo relatos de hematologistas, a coloração do leucócito era frequentemente considerada no processo de classificação. O número de atributos passa a ser 62, que são atual-

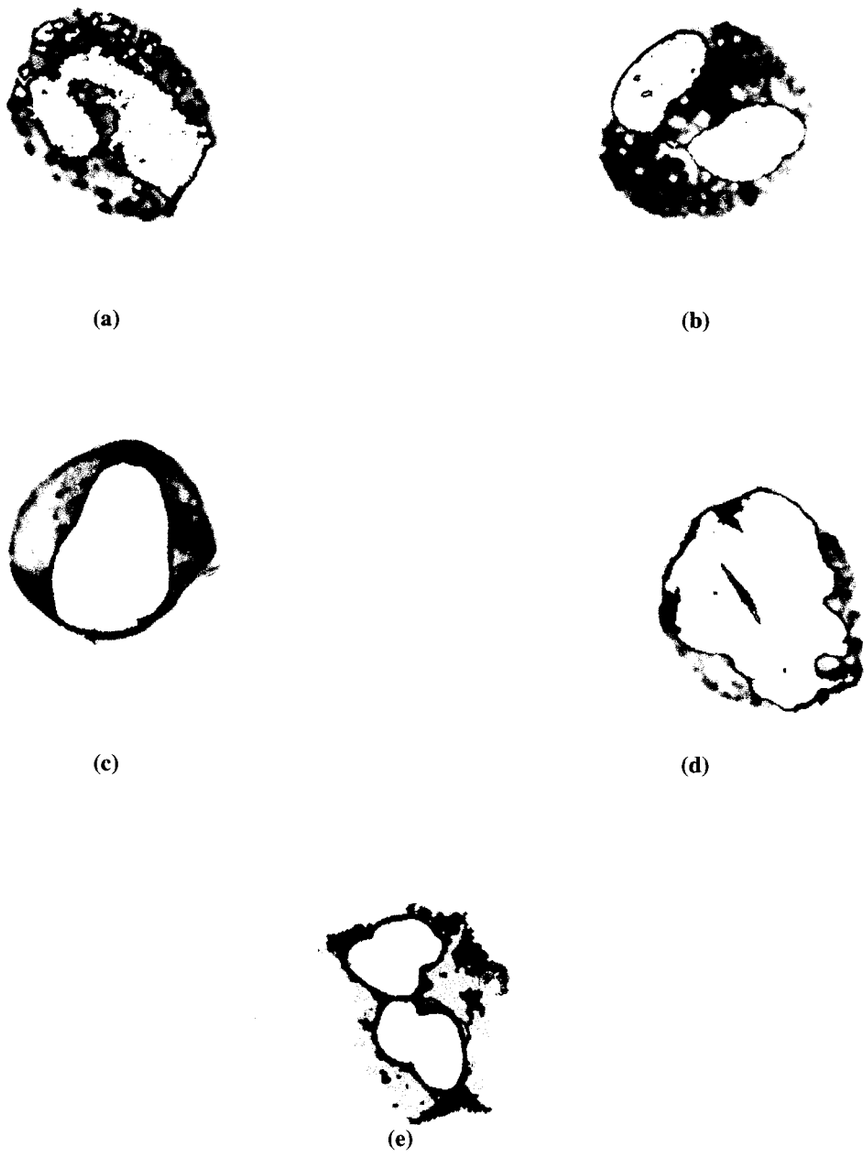


Figura 5.12: **(a)** Citoplasma do basófilo. **(b)** Citoplasma do eosinófilo. **(c)** Citoplasma do linfócito. **(d)** Citoplasma do monócito. **(e)** Citoplasma do neutrófilo.

(a)

(b)

(c)

(d)

(e)

Figura 5.13: **(a)** Núcleo do basófilo. **(b)** Núcleo do eosinófilo. **(c)** Núcleo do linfócito. **(d)** Núcleo do monócito. **(e)** Núcleo do neutrófilo.

mente extraídos usando o programa *Leuko*, os quais aparecem listados na Tabela 5.5.

Tabela 5.5: Atributos extraídos dos leucócitos.

Id	Descrição do atributo	Id	Descrição do atributo
1	N perímetro	32	N energia Blue
2	N área	33	N entropia Blue
3	N max área	34	razão N/C
4	N num elementos	35	C inércia1
5	N circularidade	36	C entropia1
6	N num concavidade	37	C energia1
7	N num convexidade	38	C correlação1
8	N energia de dobramento	39	C homogeneidade1
9	N num furos	40	C inércia2
10	N área furos	41	C entropia2
11	N inércia	42	C energia2
12	N entropia	43	C correlação2
13	N energia	44	C homogeneidade2
14	N correlação	45	C média Red
15	N homogeneidade	46	C dp Red
16	N média Red	47	C coef.assimetria Red
17	N dp Red	48	C curtose Red
18	N coef.assimetria Red	49	C energia Red
19	N curtose Red	50	C entropia Red
20	N energia Red	51	C média Green
21	N entropia Red	52	C dp Green
22	N média Green	53	C coef.assimetria Green
23	N dp Green	54	C curtose Green
24	N coef.assimetria Green	55	C energia Green
25	N curtose Green	56	C entropia Green
26	N energia Green	57	C média Blue
27	N entropia Green	58	C dp Blue
28	N média Blue	59	C coef.assimetria Blue
29	N dp Blue	60	C curtose Blue
30	N coef.assimetria Blue	61	C energia Blue
31	N curtose Blue	62	C entropia Blue

N = núcleo, C = citoplasma

Diversos atributos podem ser extraídos, contudo a alta dimensionalidade precisa ser contornada (veja Seção 3.4), principalmente com a base de imagens disponível, onde o número de amostras é relativamente pequeno. Uma maneira de diminuir o número de atributos é determinar a sua relevância e selecionar os mais significativos.

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

Para isso foi desenvolvida a ferramenta para seleção de características, onde vários *métodos foram implementados, discutidos no Capítulo 4.*

5.8 Aplicação de seleção de atributos

A ferramenta de seleção de atributos foi utilizada tanto para a seleção de escala para estudo de textura do citoplasma quanto para a determinação de um subconjunto de características dos leucócitos para classificação em um dos oito tipos: basófilo, eosinófilo, linfócito, monócito, neutrófilo, leucemia linfóide crônica, leucemia prolinfocítica e tricoleucemia.

O último resultado publicado [9], aliando a variedade de atributos extraídos por *Leuko* à capacidade de seleção dos mais relevantes, pode ser sintetizada pela Tabela 5.6, que mostra a matriz de confusão para seis tipos de leucócitos, considerando o melhor subconjnto de atributos da Tabela 5.5.

Tabela 5.6: Matriz de confusão dos atributos mais relevantes (Id segundo Tabela 5.5): 1,5,7,15,21,34,35,37,41,46,53,54

%	baso	eosi	linf	mono	neut	LLC
baso	100					
eosi		86			14	
linf			92	1	1	6
mono		3	13	81	3	
neut	1	1			98	
LLC		2	8	2	2	86

porcentagem de acerto da classe dada pelo hematologista em relação a classificação de máquina.

Recentemente, resultados incluindo leucemia prolinfocítica e tricoleucemia foram obtidos e são apresentados na Tabela 5.7.

As Tabelas 5.6 e 5.7 foram obtidas utilizando a estimação de erro *holdout*, onde o conjunto de amostras foi dividido aleatoriamente em dois, sendo dois terços para treinamento e um terço para teste [41], conforme se mostra na Tabela 5.8. O classificador utilizado foi o bayesiano ingênuo.

Tabela 5.7: Matriz de confusão dos atributos mais relevantes (Id segundo Tabela 5.5): 1,5,7,15,21,34,35,37,41,46,53,54

%	baso	eosi	linf	mono	neut	LLC	pro	hairy
baso	86				7	7		
eosi		97			3			
linf			71	3	3	9	1	13
mono			3	89	6			3
neut		2		2	95			
LLC			14			82		4
pro			13	13		6	69	
hairy								100

porcentagem de acerto da classe dada pelo hematologista em relação a classificação de máquina.

Tabela 5.8: Base de imagens de leucócitos: número de amostras para treinamento e teste do classificador.

Classes	**Treino**	**Teste**
baso	30	14
eosi	71	35
linf	190	94
mono	72	36
neut	163	82
llc	101	50
pro	31	16
hairy	46	24

5.9 Mineração de dados

Organizando o desenvolvimento do projeto de acordo com as fases indicadas na seção 4.5, as etapas para a descoberta de conhecimento do corrente projeto de pesquisa foram:

a. Definir o problema: no projeto corrente essa etapa caracteriza-se pelo delineamento das células que serão abordadas, ou seja, definem-se os casos tratáveis, uma vez que não há meios de se identificar e diferenciar, através de visão computacional, todos os tipos de células possíveis de se ocorrer no sangue.

b. Construir a base de dados para mineração: em conjunto com o laboratório de hematologia do HCRP, coletou-se 1.439 imagens de microscópio e dados clínicos, incluindo citologia, morfometria e patologia. A base de imagens consiste apenas em

uma coleção de micrografias, segmentações e vetores de características correspondentes organizados segundo sua classificação, contudo seria interessante desenvolver um banco de dados propriamente dito, onde se viabilizasse a recuperação de imagens baseada em outras imagens ou mesmo em um conjunto de características. Sistemas desse tipo vêm sendo amplamente desenvolvidos [70, 51] e são de grande utilidade.

c. Explorar os dados: inicia-se o processo exploratório a partir da extração de medidas das ROI da imagem. No projeto corrente tem-se como ROIs de uma imagem microscópica digitalizada, os leucócitos (incluindo sua subdivisão em núcleo e citoplasma), hemácia e fundo. Medidas relativas a forma, tamanho, cor e textura do leucócito podem ser extraídas e devem passar por uma triagem antes do reconhecimento [47].

d. Preparar os dados para modelagem: alguns critérios devem ser considerados quando comparando medidas de origem diversa, portanto foi necessária normalização conforme explicado na seção 3.3

e. Construir o modelo: desempenhada a exploração inicial do espaço de características, selecionadas as características adequadas e normalizando-as, passa-se a considerar os métodos para implementar a classificação ou reconhecimento propriamente dito. A classificação automática adotada no corrente projeto de pesquisa considera protótipos (amostras) das classes envolvidas (aprendizado supervisionado de classes), com decisão bayesiana. Estimando-se as funções densidade de probabilidade, assumiu-se a hipótese gaussiana de distribuição dos atributos, embora ainda seja necessário validar esse modelo estatístico para o problema de reconhecimento de leucócitos.

f. Avaliar o modelo: a avaliação do modelo é dada pela razão entre acerto do classificador em comparação com a classificação dada pelo hematologista. Como a análise visual está sujeita a erro, somente classificações feitas por hematologistas experientes foram consideradas. As células, cuja classificação o hematologista teve dúvidas, foram descartadas.

g. Disponibilizar modelos e resultados: essa etapa foi concretizada por meio de publicações periódicas a respeito do projeto de pesquisa.

Durante estágio na Universidade da Califórnia, houve possibilidade de construir um banco de dados para as imagens, contendo todas as imagens do projeto, através

do qual buscas de imagem baseadas em similaridade seriam possíveis. Atualmente contamos com um repositório de 1.439 imagens, segmentações e vetores de características correspondentes a grande parte das imagens, além de dados clínicos.

Tanto os dados quanto as imagens foram organizados e poderão ser disponibilizados gratuitamente na internet mediante autorização prévia da agência de fomento (FAPESP).

Capítulo 6

Conclusão e trabalhos futuros

"A mais grave das faltas é não ter consciência de falta alguma. (Einstein)"

Experimentos biológicos frequentemente estão sujeitos a aspectos não controlados, um entrave que pode ser minimizado coletando-se várias amostras de diferentes indivíduos utilizando o mesmo microscópio e a mesma câmera sob mesmo aumento. Seguindo esses preceitos, realizou-se os experimentos usando 1.055 células de 33 indivíduos diferentes; os sujeitos não foram analisados separadamente. Geralmente, os dados experimentais médicos são observacionais, o que torna a questão de inferência mais difícil. A quantificação em detrimento da qualificação das características das células representa uma maneira de permitir a inferência automática.

Argumenta-se que quando há um grande número de atributos candidatos a gerar o classificador, é mais fácil achar padrões envolvendo alguns dos atributos. Métodos de validação (ex: *holdout* e validação cruzada) foram aplicados no intuito de evidenciar se o classificador é bom tanto para o conjunto de características total assim como para subconjuntos.

Quanto a significância dos padrões encontrados, ou seja, em que medida eles são provenientes de diferenças biológicas reais e não artefatos, vale salientar que o processo de segmentação de células adotado implica em uma descrição bastante peculiar e não fidedigna com a realidade biológica. Assim, coube ao projetista identificar não unicamente medidas correlatas com descrições médicas, mas também outras que não são consideradas no processo de classificação manual, além de interpretar os resultados da segmentação de acordo com o conhecimento adquirido com os especialistas

(hematologistas).

Um exemplo é o ilustrado nas segmentações de imagens de LLC (Fig.5.11), onde o afrouxamento de cromatina pode ser detectado e é evidenciado pela segmentação como buracos no núcleo com "cor de citoplasma", embora se saiba que não se trata de citoplasma no núcleo. Outras idiossincrasias impostas pelo método podem ser notadas na segmentação de eosinófilos, onde ocorrem grânulos com "cor de hemácia" no citoplasma, que na realidade são grânulos eosinofílicos, que possuem aspecto avermelhado, característica atribuída, segundo treinamento, como sendo de hemácia.

Sem dúvida, a segmentação determina a precisão de classificação, portanto foi imprescindível que se colecionassem atributos das células que fossem menos suscetíveis a erro quanto possível. Exemplificando, não se considerou o contorno do citoplasma, pois a princípio se sabe que a segmentação do citoplasma é bastante sensível a ruídos, além de que, células adjacentes podem se aglutinar, o que levantaria suspeita de pontos de alta curvatura onde, na verdade, não existem.

A seguir são sintetizados os últimos avanços no corrente projeto de pesquisa, levando em consideração as principais etapas do software em desenvolvimento. A intercomunicação de todas essas funcionalidades voltadas ao estudo das células do sangue objetiva auxiliar o médico no reconhecimento de leucócitos, além de fornecer a pesquisadores mecanismos simples para mineração de dados.

6.1 Segmentação

A pesquisa em métodos de segmentação é vasta e se justifica por si mesma, portanto essa tese se limitou a abordar alguns métodos convencionais, além de propor novos métodos de limiarização para segmentação do núcleo de leucócitos. Através do estudo de métodos de limiarização, desenvolveu-se e testou-se um método original.

Na segmentação da imagem de sangue em ROIs, assume-se que a cor é uma propriedade constante da ROI, portanto é possível mapear cada pixel da imagem no espaço de cor RGB, esperando-se que objetos diferentes se manifestarão como agrupamentos ou nuvens de pontos com coordenadas RGB semelhantes.

É utilizada classificação bayesiana para determinar pertinência do pixel em relação a ROI a que pertence. O modelo de distribuição RGB é dado por gaussianas multi-

variadas. A segmentação da imagem de sangue baseada em espaço de cor negligencia a relação espacial entre as cores, portanto foi necessária a implementação de técnicas de processamento de imagem, como filtros morfológicos e limiarização por área para minimizar o ruído das ROIs. Apesar da segmentação ser automática, ajustes interativos são necessários em alguns casos, os quais podem ser realizados utilizando as ferramentas de processamento de imagem oferecidas por *Leuko*.

Embora trabalhar em espaços como o HSI, CIELab seja mais intuitivo, o RGB foi o melhor espaço de cor para segmentação dos leucócito utilizando o classificador disponível. A transformação dos valores do RGB em outros espaços colorimétricos acarretou distorções indesejáveis quando utilizando o algoritmo de segmentação bayesiana sem exploração de adjacência.

Buscando melhorar os resultados de segmentação, aplicou-se filtros e, empiricamente, notou-se que a filtragem anisotrópica acentuava a característica circular em linfócitos e, em Hairy cells, ajudava a capturar os "cabelos"do citoplasma; em células com contornos mais rugosos, ocorreu apenas uma suavização do contorno sem perdas dos pontos mais importantes. Experimentos preliminares foram realizados, cujos resultados se encontram no Apêndice C.

6.2 Atributos da imagem

Diversos atributos da imagem foram testados, considerando medidas de forma, tamanho, cor e textura. Medidas de forma e tamanho não podem ser consideradas exclusivamente, pois são insuficientes para caracterização dos leucócitos ainda que só para descrição dos casos normais. As medidas de textura foram particularmente estudadas, incluindo uma ampla abordagem multiscala e decisão da melhor escala baseada em resultados de seleção de características. Atributos de textura trouxeram uma melhoria significativa na classificação de células normais, permitindo também a inclusão de neoplasias. A inclusão de medidas de cor permitiu uma melhor descrição do citoplasma, eliminando erros grotescos como confundir granulócitos com mononucleados e diminuição de erros entre os granulócitos e entre os mononucleados.

Algumas características possuem correspondente biológico direto como é o caso da razão núcleo-citoplasmática e circularidade, uma vez que estas são utilizadas pelos próprios hematologistas. Exemplificando, a razão núcleo-citoplasmática do

linfócito normal é menor que a do linfócito com LLC, ou seja, o linfócito normal frequentemente tem mais citoplasma que o de LLC. A circularidade é particularmente interessante para distinguir núcleos de mononucleados de polimorfonucleados. Entretanto, os parâmetros de textura e cor, que levam em consideração estatísticas a respeito de ocorrências de grânulos ou cores, não possuem correspondentes biológicos tão evidentes, ou seja, traduzem quantitativamente aspectos tomados pelos hematologistas somente qualitativamente (ex: citoplasma homogêneo, tênue = baixa entropia média da GLCM).

Para todo o conjunto de medidas extraídas dos leucócitos, mostrou-se que tanto os atributos de forma, quanto os de cor e textura são igualmente importantes através de estudo de seleção de características sobre cada subconjunto de atributos em função de sua origem. Assim, selecionou-se os atributos mais relevantes do subconjunto de medidas de forma, então separadamente selecionou-se os atributos de cor mais relevantes, sendo o mesmo feito com os de textura [9]. Com os melhores atributos de cada subconjunto, realizou-se classificação bayesiana com modelagem gaussiana de classes e estimação de erro considerando os métodos de resubstituição e holdout, conforme discutido na seção 3.6.

6.3 Classificadores

Técnicas bayesianas vêm sendo usadas no campo de reconhecimento de padrões [13], mas somente recentemente tem sido testadas com afinco em conjunto de dados com atributos redundantes [48]. O rótulo "Naive Bayes"infelizmente advém do fato de ser difícil de se usar este método sem se sentir simplista. Contudo, não há nada de ingênuo sobre seu uso em circunstâncias apropriadas [41].

A classificação bayesiana foi inicialmente utilizada na determinação das ROIs da imagem. Alguns testes foram realizados usando o método dos k-vizinhos mais próximos, mas foi comparativamente pior tanto no tempo de execução quanto nos resultados de segmentação.

Além de ser considerado na determinação das ROIs, o classificador bayesiano também foi usado no reconhecimento das células, obtendo-se resultados animadores segundo os artigos anexos. Também se utilizou o ambiente Weka e seu classificador bayesiano para se comparar os resultados de classificação entre as células [9].

Verificou-se que tanto *Leuko* quanto o Weka deram resultados semelhantes quanto à taxa de acerto do classificador diante dos atributos selecionados. No caso da utilização de todos os atributos, mostrou-se que a taxa de acerto diminuiu em relação à classificação utilizando somente o subconjunto mais relevante. Resultados preliminares utilizando outros classificadores no ambiente Weka são ilustrados no Apêndice D.

6.4 Seleção de características

A ferramenta para seleção de características permite avaliar um atributo multiscala [65] de modo a se estimar qual escala permite maior diferenciação entre as classes, além de auxiliar na definição de subconjuntos de características mais relevantes para a classificação de padrões [64, 9].

É difícil de se estabelecer o número ideal de características, mas geralmente se usa ao menos 10 vezes mais amostras de treinamento por classe que o número de características [43]. Como o número de amostras era reduzido, principalmente para os basófilos (veja tabela 5.8), fez-se seleção de características por partes, conforme último artigo completo publicado [9]. Este descreve a busca das 4 características mais relevantes entre as medidas de forma, as 4 mais relevantes entre as medidas de textura e as 4 mais relevantes entre as medidas de cor como método de se manter a proporção de 10 vezes mais amostras que características. Como resultado da seleção de medidas das células, mostrou-se que a separação inter-classes era melhor quando o número de atributos era de aproximadamente 12 em comparação com o uso de todos os atributos.

Para validar os resultados de seleção de características, comparou-se o resultado do algoritmo de seleção do Weka com o do selector de características proposto e verificou-se alta concordância entre os subconjuntos selecionados por cada programa. O subconjunto de atributos mais relevantes considerando a seleção de características por partes foi o que apresentou melhor taxa de acerto na classificação tanto no *Leuko* quanto no Weka.

6.5 Trabalhos futuros

A pesquisa envolve imagens coloridas (RGB) de lâminas de esfregaços de sangue, as quais possuem geralmente variação de luminosidade espúria, artefatos de coloração, etc. Lâminas com problema de coloração são eliminadas pelos hematologistas, mas a variabilidade de iluminação é facilmente superada pelo sistema visual humano, que se adapta à luz. Somente técnicas de processamento de imagem podem mimetizar esta adaptação, diminuindo a variação de iluminação entre as imagens de diferentes lâminas de esfregaço para uma comparação fidedigna entre imagens similares. Maiores investigações devem ser feitas a respeito de filtros e espaços de cor (HSV, CIELab, YCbCr) no intuito de normalizar as imagens.

Testou-se o filtro de difusão anisotrópica, que apresentou resultados visualmente melhores em termos da homogeneização de regiões conexas da imagem. Entretanto, a segmentação usando *Leuko* não mostrou melhora significativa. Acredita-se que as limitações impostas pelo processo de segmentação sejam responsáveis pelo baixo desempenho após filtragem, mas só se poderá afirmar quando outros métodos de segmentação forem testados. A limitação de tempo também inviabilizou maiores testes, pois tais conceitos foram adquiridos somente ao final do programa de doutorado, durante estágio na Universidade da Califórnia. Pretende-se quantificar a melhoria que esse filtro carreta ao processo de segmentação.

A segmentação utilizada é baseada em estimação de funções densidade de probabilidade gaussiana a partir de amostras de cor nos pixels, contudo o processo é sensível a variações de luminosidade, sua generalização é frequentemente limitada a um subconjunto de imagens com propriedades de coloração similares. Por esta razão, nós usamos uma abordagem semi-automática. Outros classificadores poderiam ser usados no processo de classificação dos pixels, além de outros modelos de distribuição das probabilidades.

Os resultados de segmentação por fluxo de região baseado em evolução por curva [68], que considera métodos de conjunto de nível (*level-set*) [71] empregados no problema de reconhecimento de leucócitos tiveram vantagens e desvantagens, pois embora permita que haja segmentação não supervisionada totalmente automática, as segmentações foram pouco satisfatórias para monócitos e eosinófilos, conforme mostrado no Apêndice C. Tais resultados preliminares foram coletados durante

estágio, sob supervisão do Prof. B. S. Manjunath, na Universidade da Califórnia em Santa Bárbara, EUA. Trabalhos prévios em segmentação de leucócitos usando o modelo de propagação de interface estiveram restritos aos autores Nilsson e Heyden [35]. Contudo pouca informação sobre a pesquisa deles encontra-se disponível, provavelmente porque este foi um projeto desenvolvido em colaboração com a empresa *Cellavision* (Suécia) para elaboração de software de reconhecimento de leucócitos, lançado comercialmente em 2001. Complementando as publicações de Nilsson e Heyden, abordou-se o problema como um todo, cobrindo desde a segmentação até a pré-classificação das células, incluindo métodos de avaliação de segmentação. Os testes e desenvolvimentos relativos ao software para leucócitos continuarão.

Várias considerações biológicas, como a concentração de células por lâmina, dados clínicos do paciente como idade, presença de nódulos, etc., poderiam ser consideradas no processo de tomada de decisão para predição de diagnóstico de doenças.

O objetivo inicial do projeto foi de desenvolver uma ferramenta de auxílio no diagnóstico de leucemias linfóides crônicas, contudo esse subconjunto das leucemias engloba diversas neoplasias diferentes cujas diferenças morfológicas não são usadas como critérios exclusivos para o diagnóstico [72]. Assim, os dados clínicos e a imunofenotipificação são muitas vezes indispensáveis.

Quanto aos casos de leucemias, analisou-se apenas a LLC, prolinfócitos e tricoleucemia nesse trabalho de doutorado, entretanto se acredita que outras células possam ser incluídas. Particularmente a tricoleucemia apresentou problemas quanto a sua diferenciação em relação ao linfócito normal pois os "cabelos"do citoplasma são imprescindíveis no processo de análise visual e não são evidentes no processo de segmentação atualmente adotado. Métodos de segmentação foram estudados e sua limitação ocorreu em detrimento de uma abordagem mais ampla ao problema de reconhecimento, ou seja, embora a segmentação seja limitada, várias outras etapas foram tratadas no desenvolvimento de uma ferramenta para auxílio no diagnóstico de leucemia.

Uma ferramenta de reconhecimento de leucócitos foi desenvolvida, contudo bem limitada em relação aos planos audaciosos iniciais, muito embora de grande utilidade para fins didáticos, permitindo quantificação objetiva dos atributos das células. Várias funcionalidades foram agregadas a um único software de forma a facilitar análise e classificação de leucócitos, assim se pretende veicular uma versão *demo* na

internet para testes.

Finalizando esse projeto de doutorado, testou-se outros classificadores utilizando o software Weka, cujos resultados são apresentados no Apêndice D. Futuros desenvolvimentos devem explorar outros métodos de classificação.

"O ser humano não pode deixar de cometer erros; é com os erros que os homens de bom senso aprendem a sabedoria para o futuro.(Plutarco)"

Apêndice A

Imagens de leucócitos coletadas

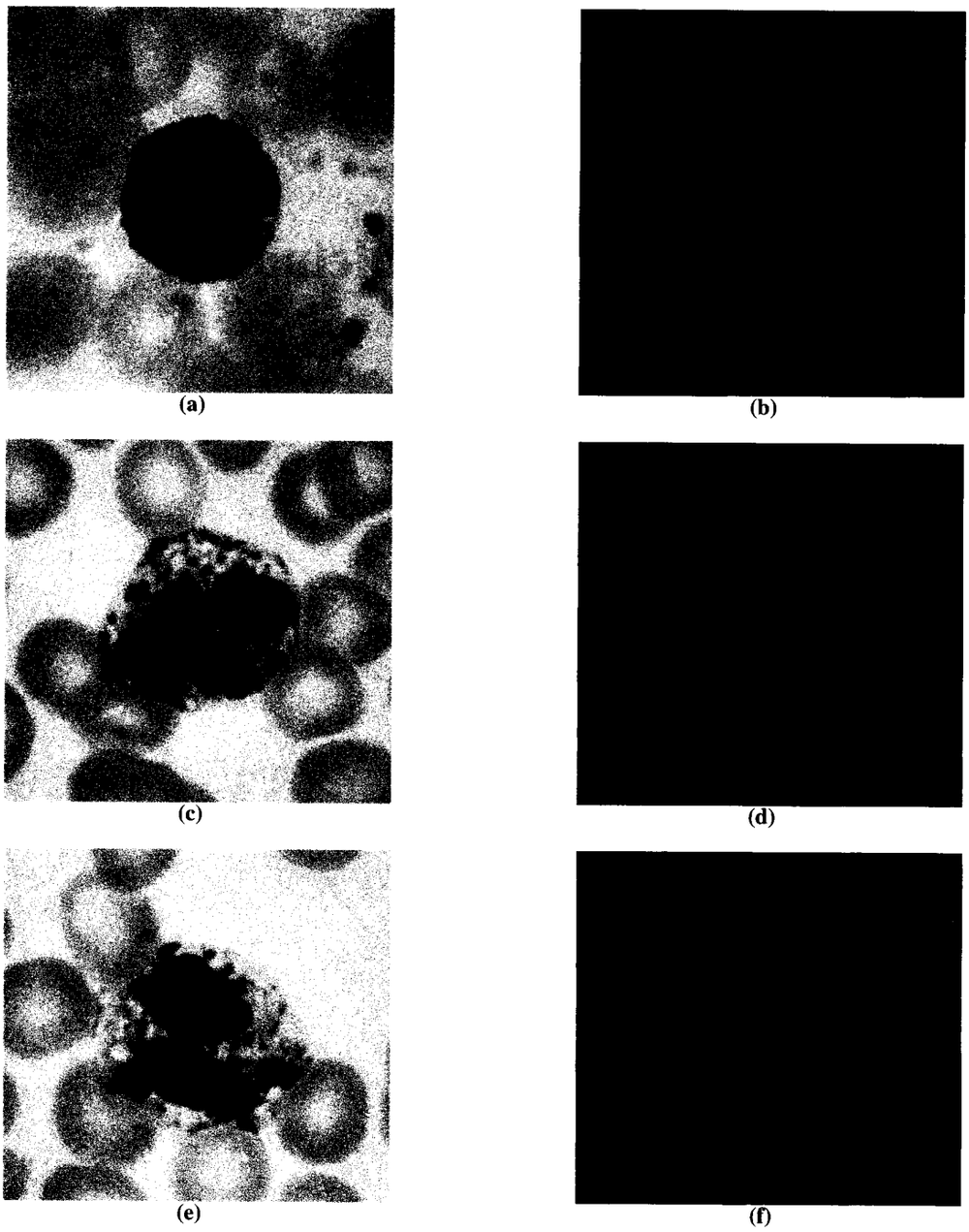


Figura A.1: **(a,c,e)** Micrografia do basófilo. **(b,d,f)** Basófilo segmentado.

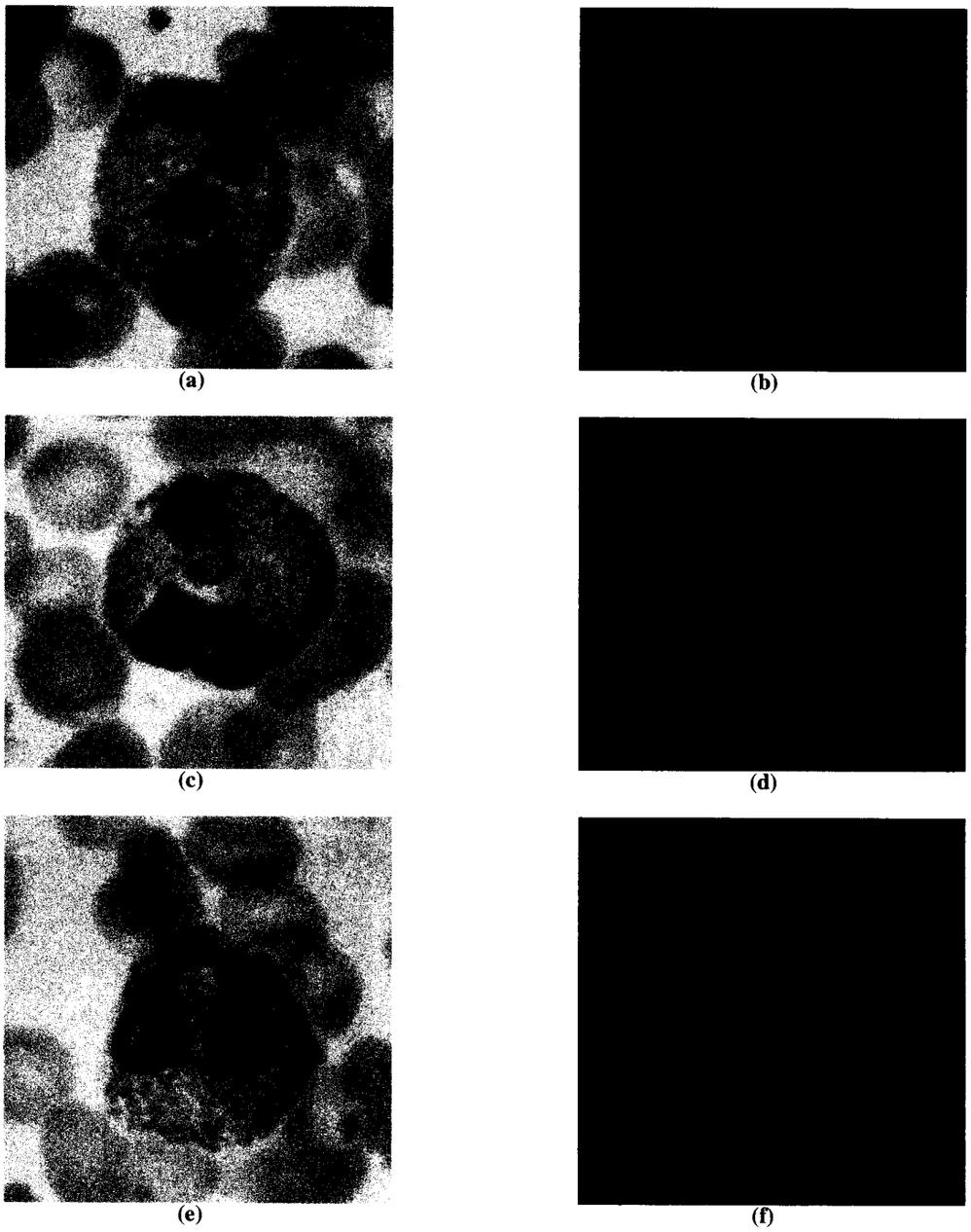


Figura A.2: **(a,c,e)** Micrografia do eosinófilo. **(b,d,f)** Eosinófilo segmentado.

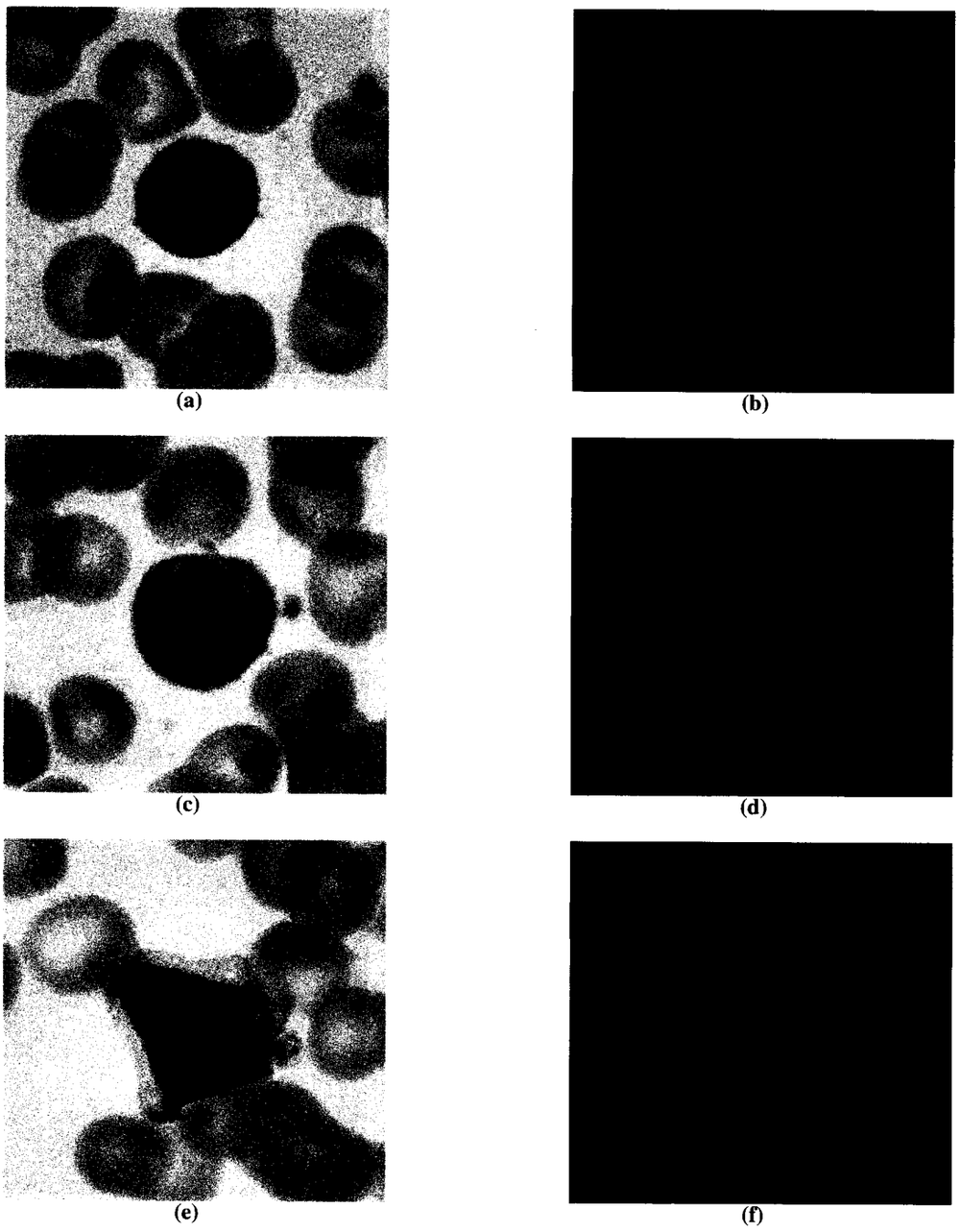


Figura A.3: **(a,c,e)** Micrografia do linfócito. **(b,d,f)** Linfócito segmentado.

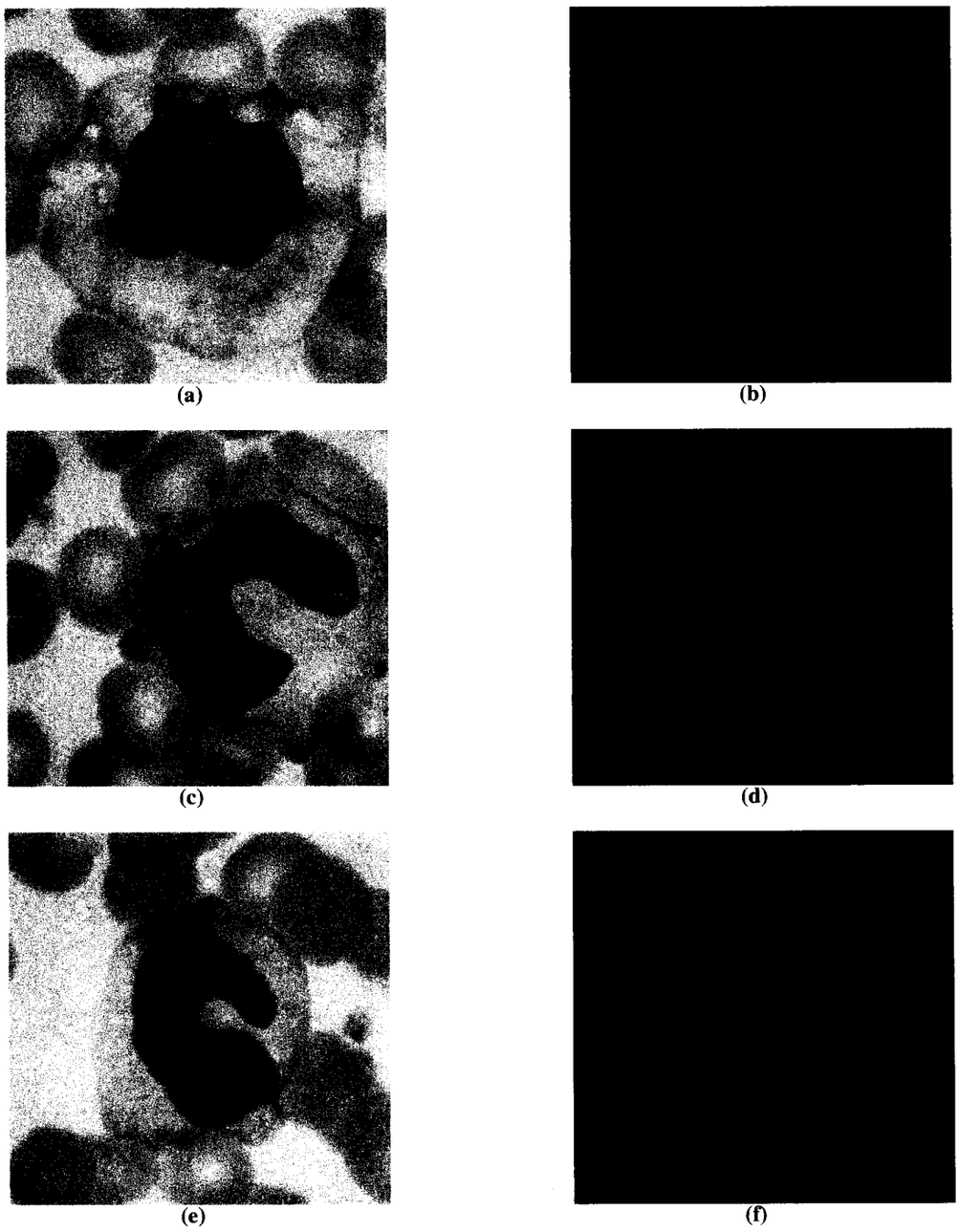


Figura A.4: **(a,c,e)** Micrografia do monócito. **(b,d,f)** Monócito segmentado.

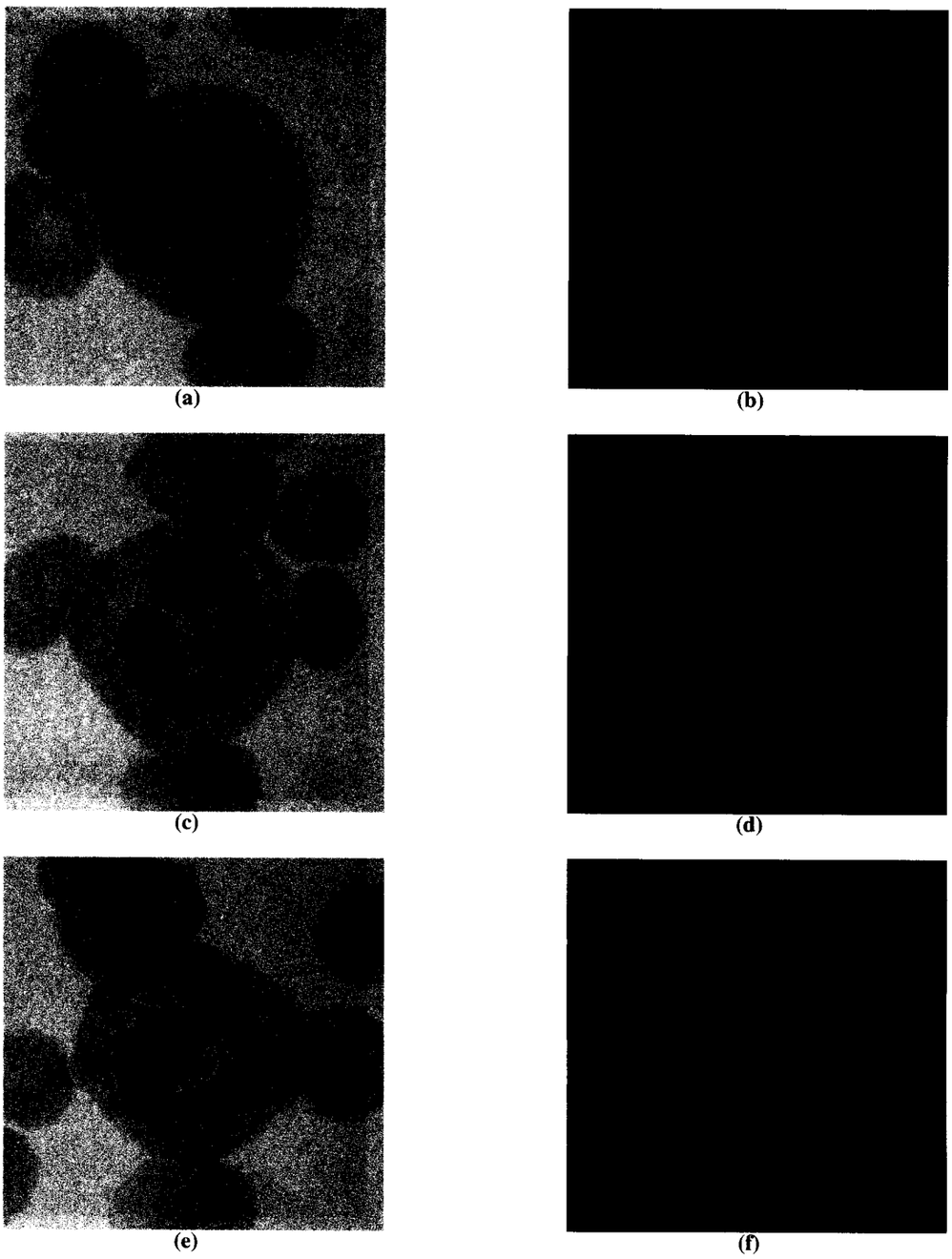


Figura A.5: **(a,c,e)** Micrografia do neutrófilo. **(b,d,f)** Neutrófilo segmentado.

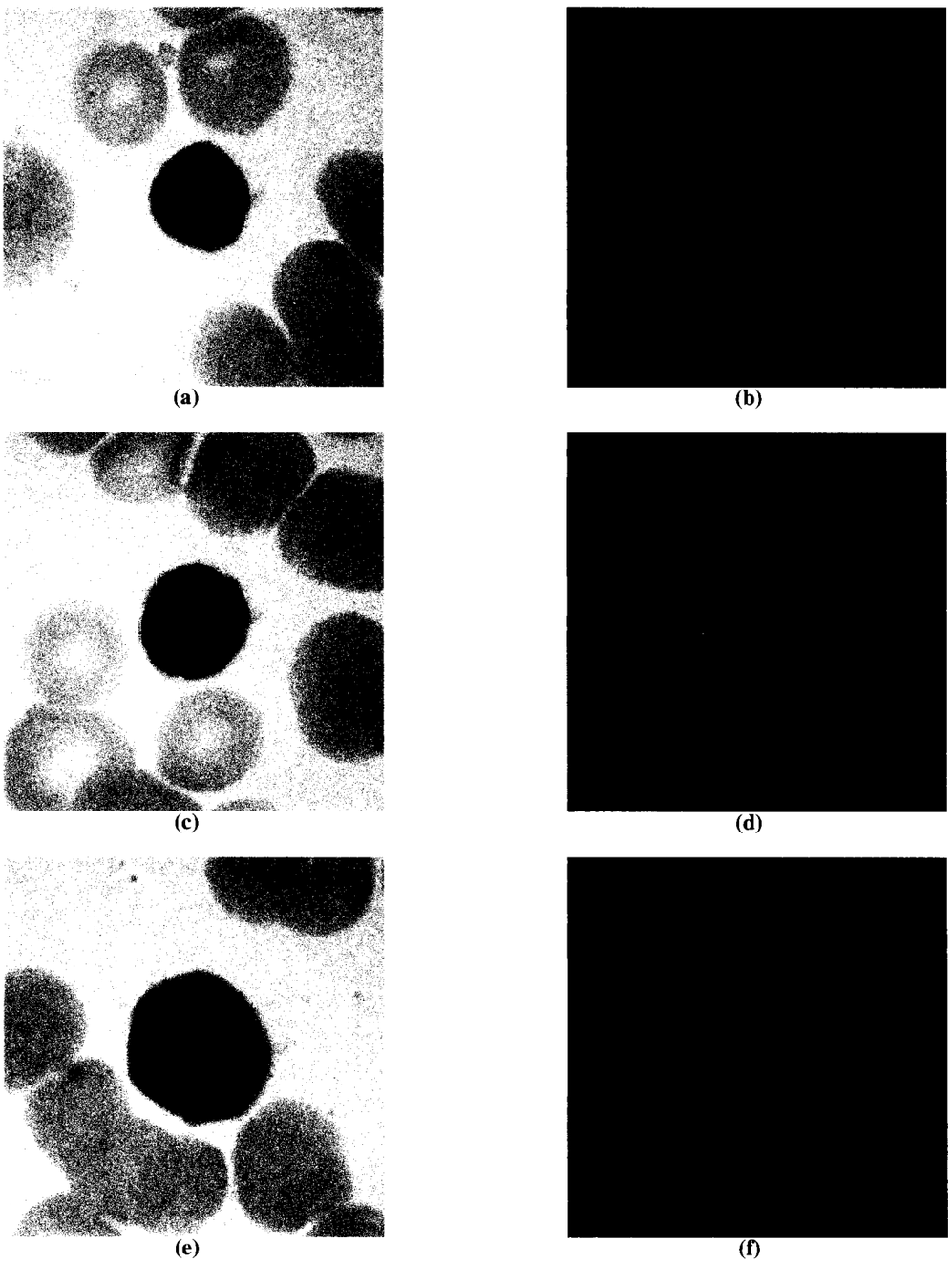


Figura A.6: **(a,c,e)** Micrografia de LLC. **(b,d,f)** LLC segmentada.

Apêndice B

Métodos de limiarização

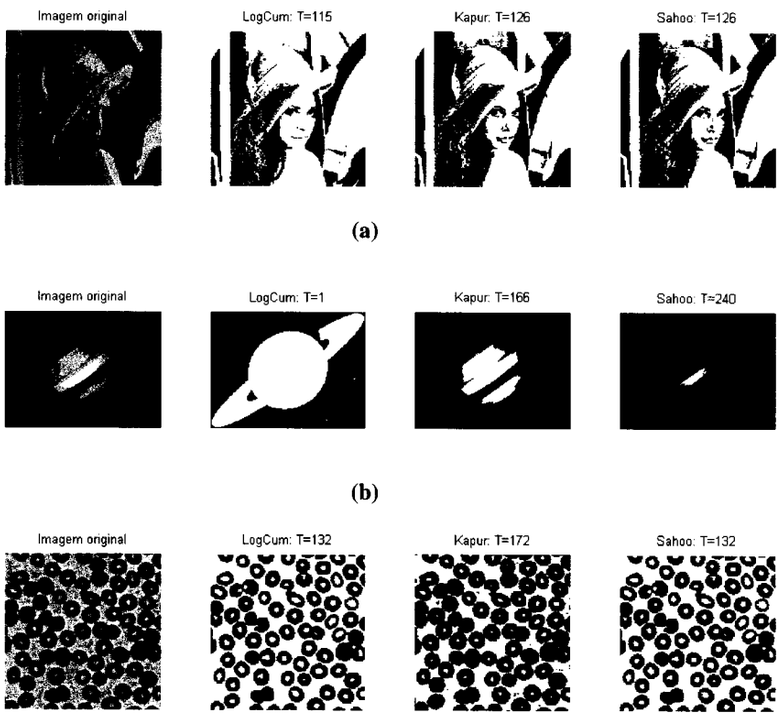


Figura B.1: Exemplos comparativos de limiarização utilizando imagens do Matlab.

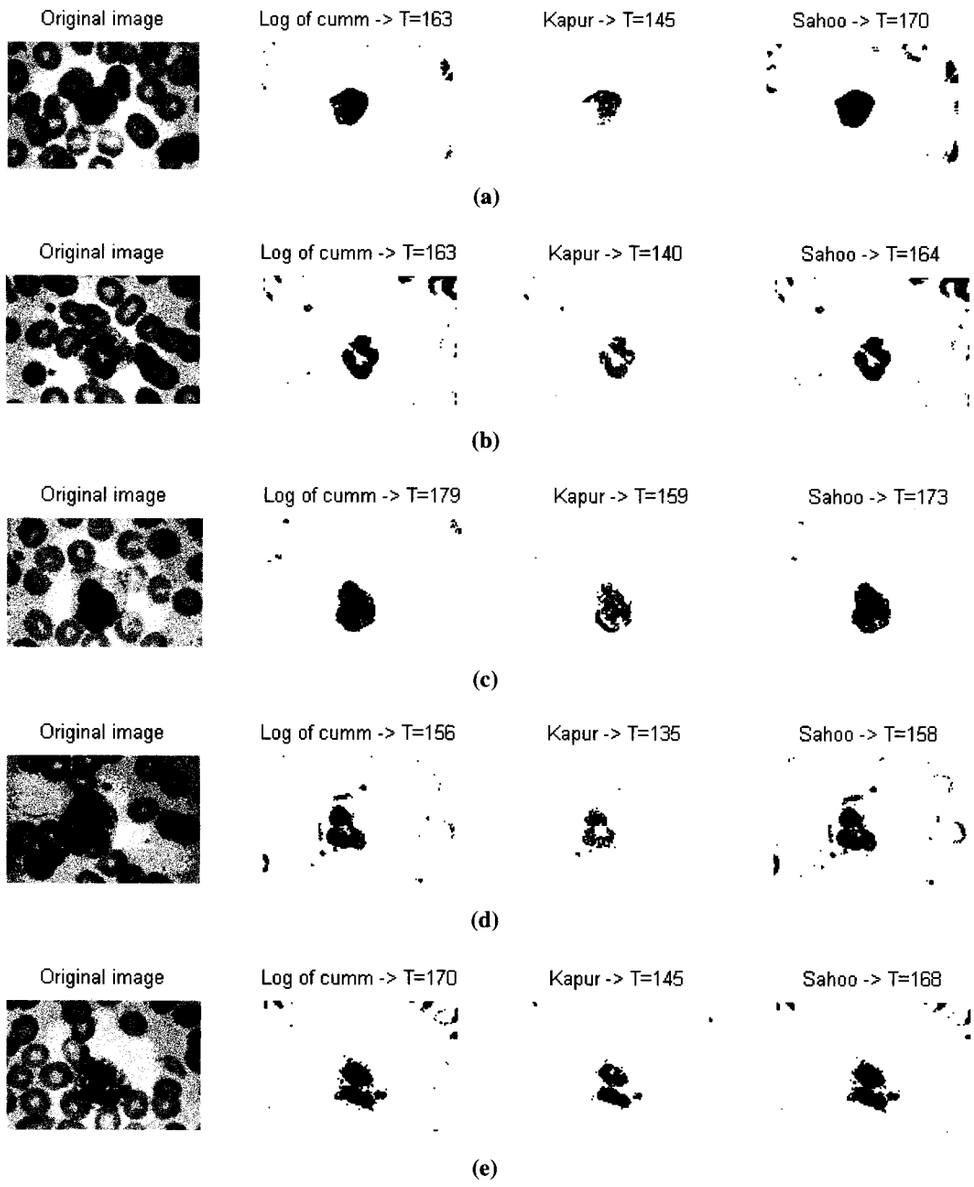


Figura B.2: Exemplos comparativos de limiarização para diferentes tipos de leucócitos: (a)linfócito; (b)neutrófilo; (c)monócito; (d)eosinófilo; (e)basófilo.

Apêndice C

Outros processamentos sobre os leucócitos

C.1 Retinex

Testou-se o algoritmo retinex nas imagens de sangue, cujo resultado é um fundo mais homogêneo quando comparando com outras micrografias de sangue, conforme era esperado. Entretanto perdas significativas ocorreram na imagem do leucócito, onde regiões do citoplasma se confundem com o fundo, conforme mostrado nas Fig.C.1 e Fig. C.2.

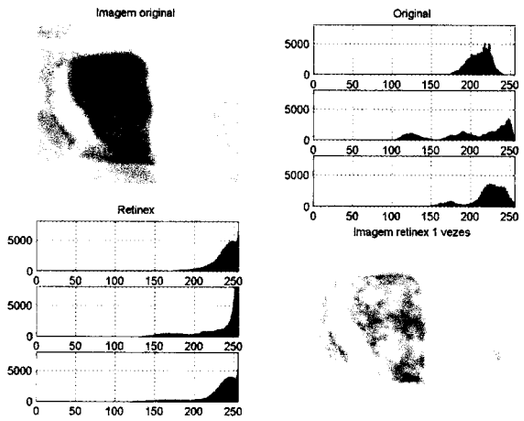


Figura C.1: Imagem de leucócito (acima) e resultado do retinex (abaixo) para 1 ciclo.

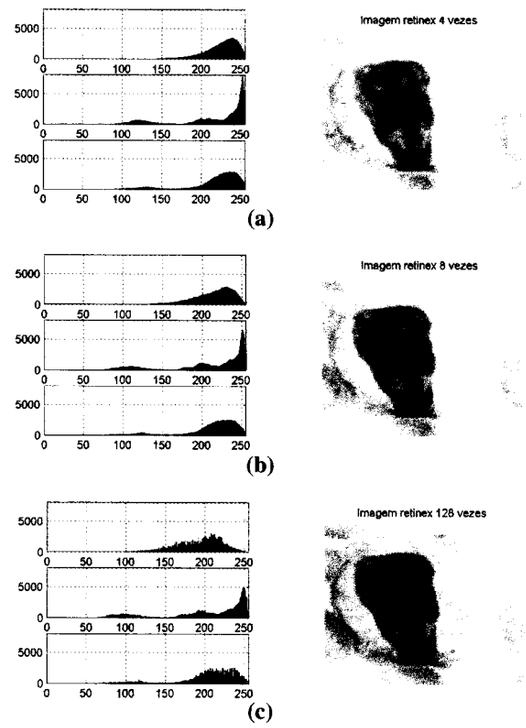


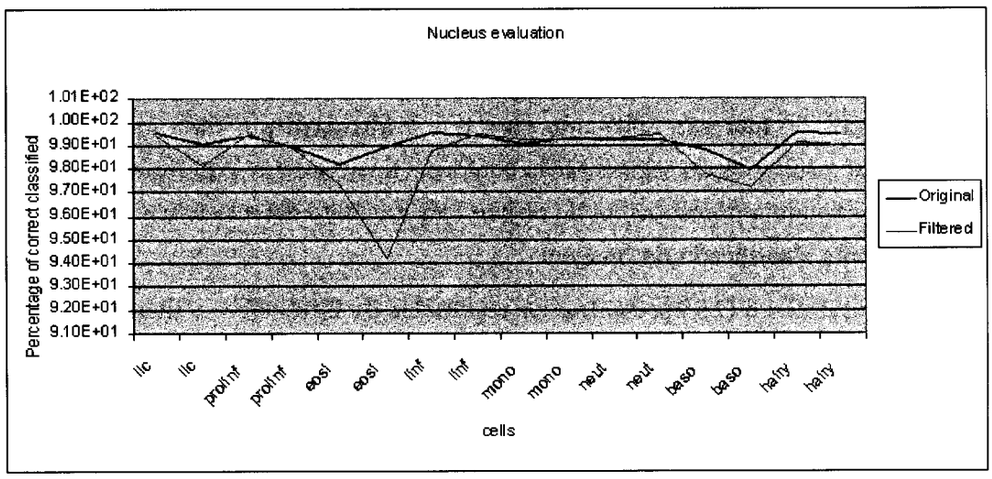
Figura C.2: (a,b,c) Processamento iterativo usando retinex: a)4x, b)8x e c)128x.

C.2 Difusão anisotrópica

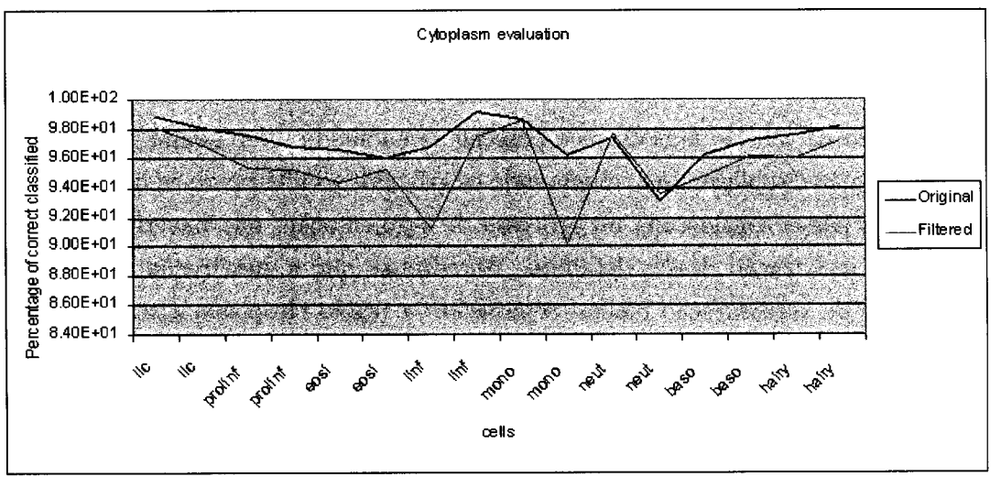
Através de critérios observacionais, julgou-se que a difusão anisotrópica poderia melhorar a segmentação da imagem, principalmente no caso das Hairy-cell onde o citoplasma apresenta estruturas que lembram cabelos, característica essencial para o diagnóstico desse tipo de leucemia. Acredita-se que o algoritmo de segmentação por cor adotado não seja adequado para que a filtragem por difusão anisotrópica melhore os resultados da segmentação conforme ilustrado em experimentos preliminares observando a porcentagem do objeto corretamente detectado e a porcentagem corretamente classificada da imagem.

As figuras C.4 e C.5 apresentam os resultados do filtro sobre a imagem original. Buscando quantificar a qualidade da segmentação bayesiana por cor considerando ou não o processo de filtragem, calculou-se a porcentagem dos pixels corretamente classificados (PCC), ou seja, a soma dos pixels verdadeiro-positivo e falso-positivo divididos pelo número total de pixels da imagem. Os gráficos da Fig.C.3 ilustram o desempenho do processo de segmentação do núcleo e do citoplasma, dadas as versões filtrada e não filtrada da imagem, em termos para 16 imagens, 2 imagens representando cada uma das 8 classes de leucócitos abordados.

O filtro de difusão anisotrópica pareceu visualmente interessante pois homogeneiza a região citoplasmática, eliminando os grânulos e salientando as bordas. Como a textura não é considerada no processo de segmentação, não haveria perdas por conta da suavização desses detalhes. Acredita-se que a suavização permitirá ao algoritmo de segmentação melhor separação da região citoplasmática, com menos buracos e mais precisão na região do contorno. Os resultados apresentados a seguir indicam que o método de segmentação por cor sem consideração de adjacência não é adequado para ser aliado a esse processamento de imagem. Não se incorporou esse processamento ao software, pois não houve ganho, em termos da segmentação, segundo avaliação utilizando a porcentagem de pixels corretamente classificados.



(a)



(b)

Figura C.3: Análise por PCC para comparar segmentação da imagem original e filtrada, considerando 16 imagens de leucócitos segmentadas manualmente: **(a)** núcleo e **(b)** citoplasma .

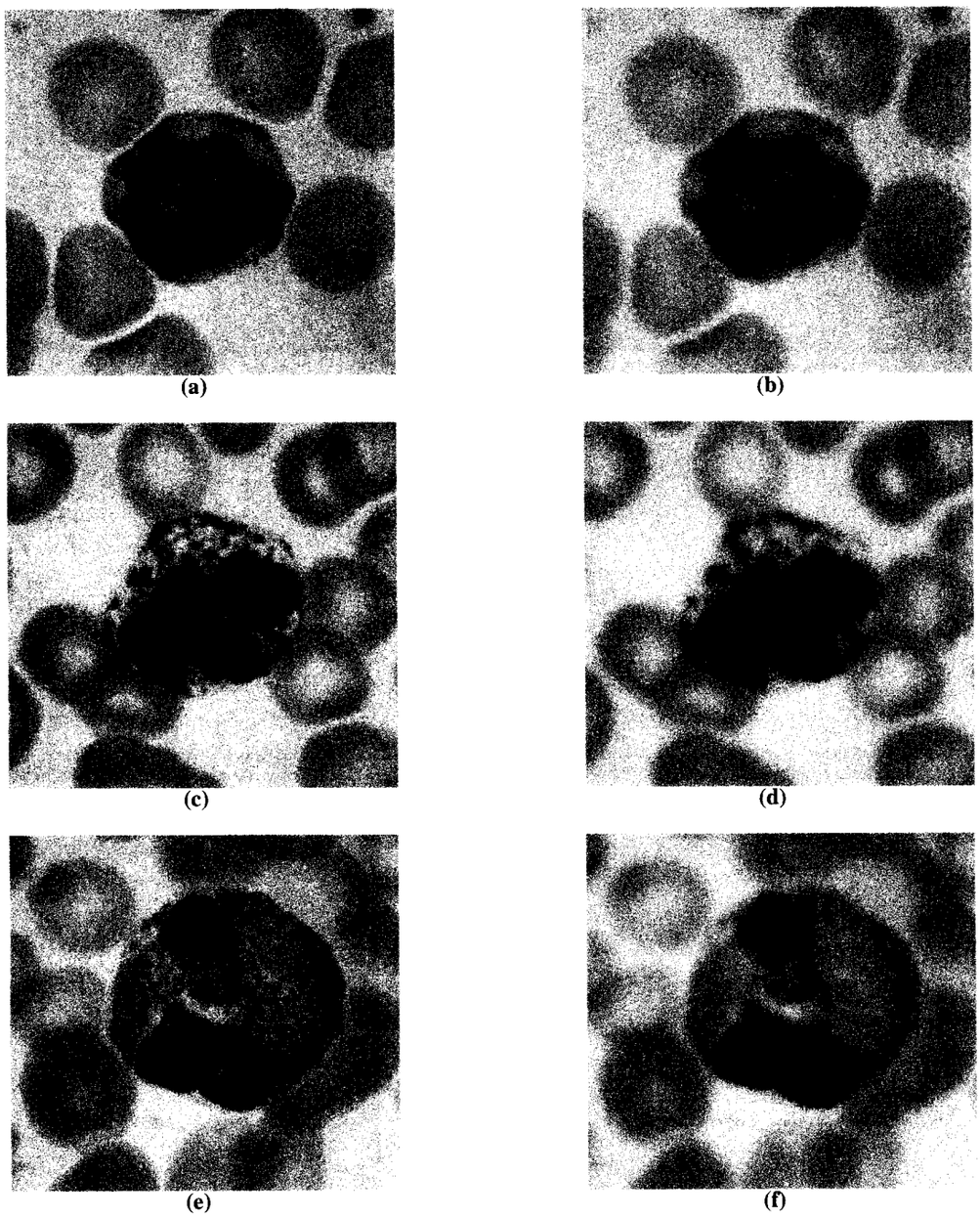


Figura C.4: **(a,c,e)** Micrografias. **(b,d,f)** Resultado após filtragem usando o método de difusão anisotrópica.

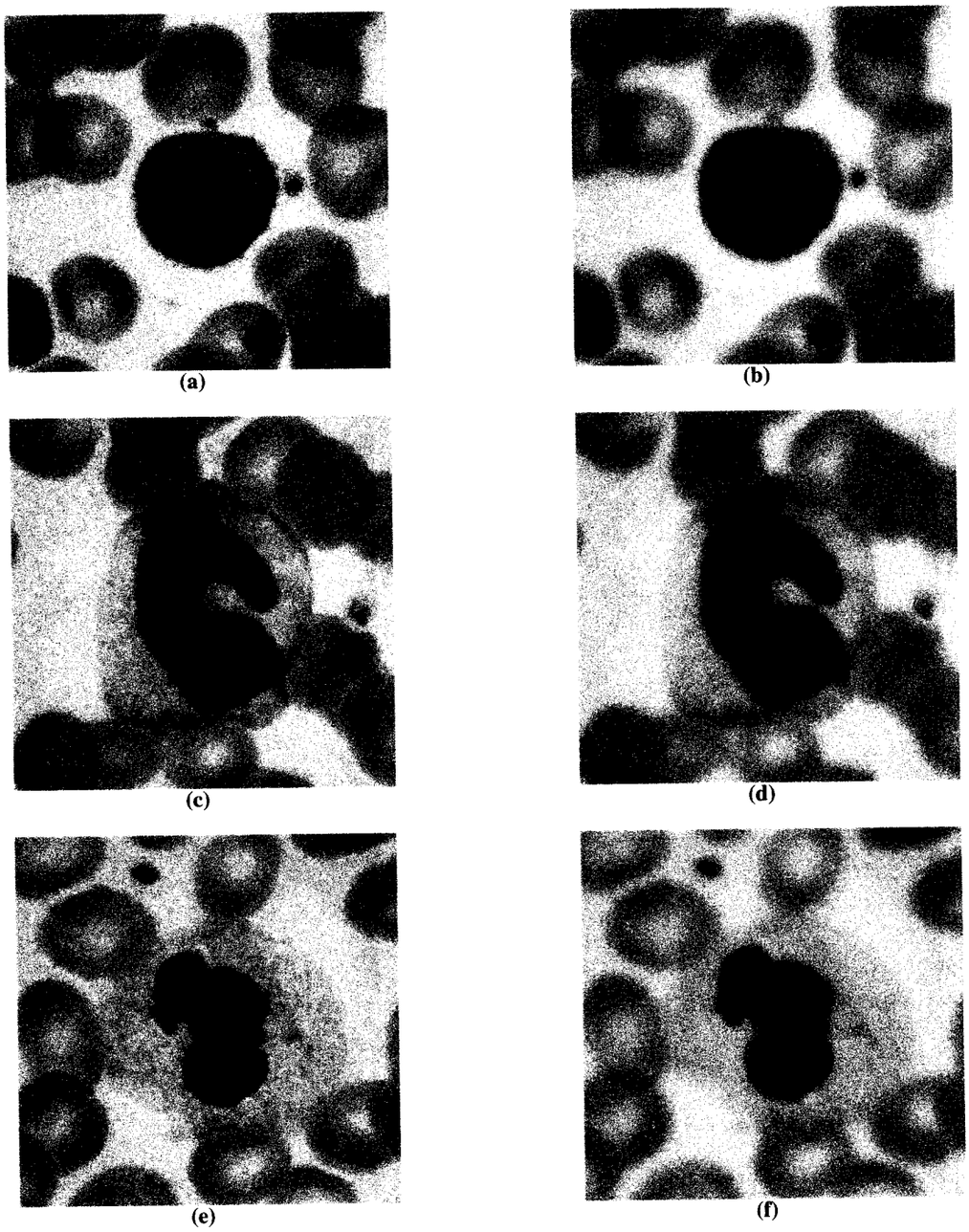


Figura C.5: **(a,c,e)** Micrografias. **(b,d,f)** Resultado após filtragem usando o método de difusão anisotrópica.

C.3 Segmentação baseada em evolução de curvas

Esta seção ilustra as segmentações utilizando o algoritmo proposto por Baris et al [68], conforme discutido na Seção 5.3.3.

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

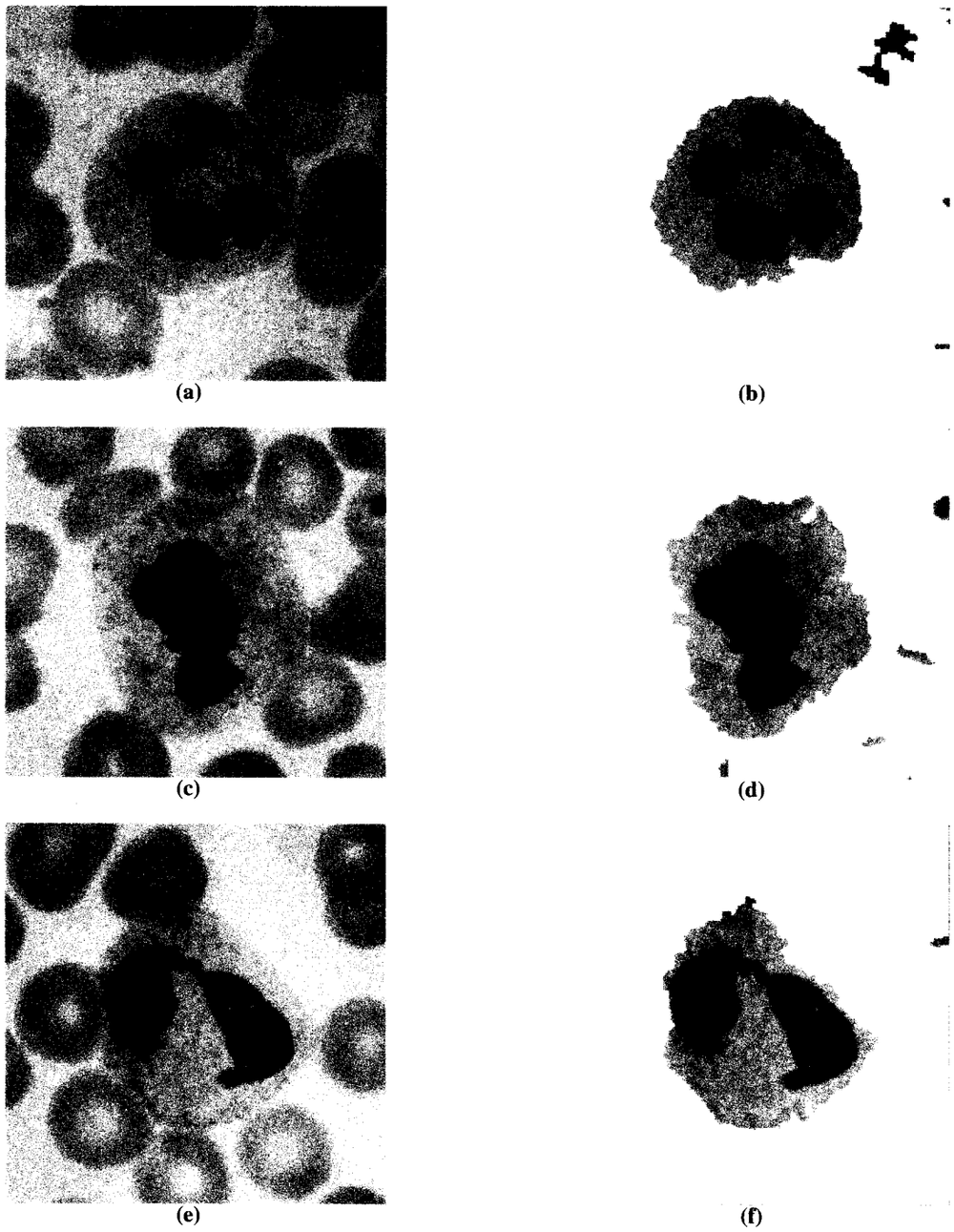


Figura C.6: **(a,c,e)** Micrografias de neutrófilos. **(b,d,f)** Segmentações por evolução de curva.

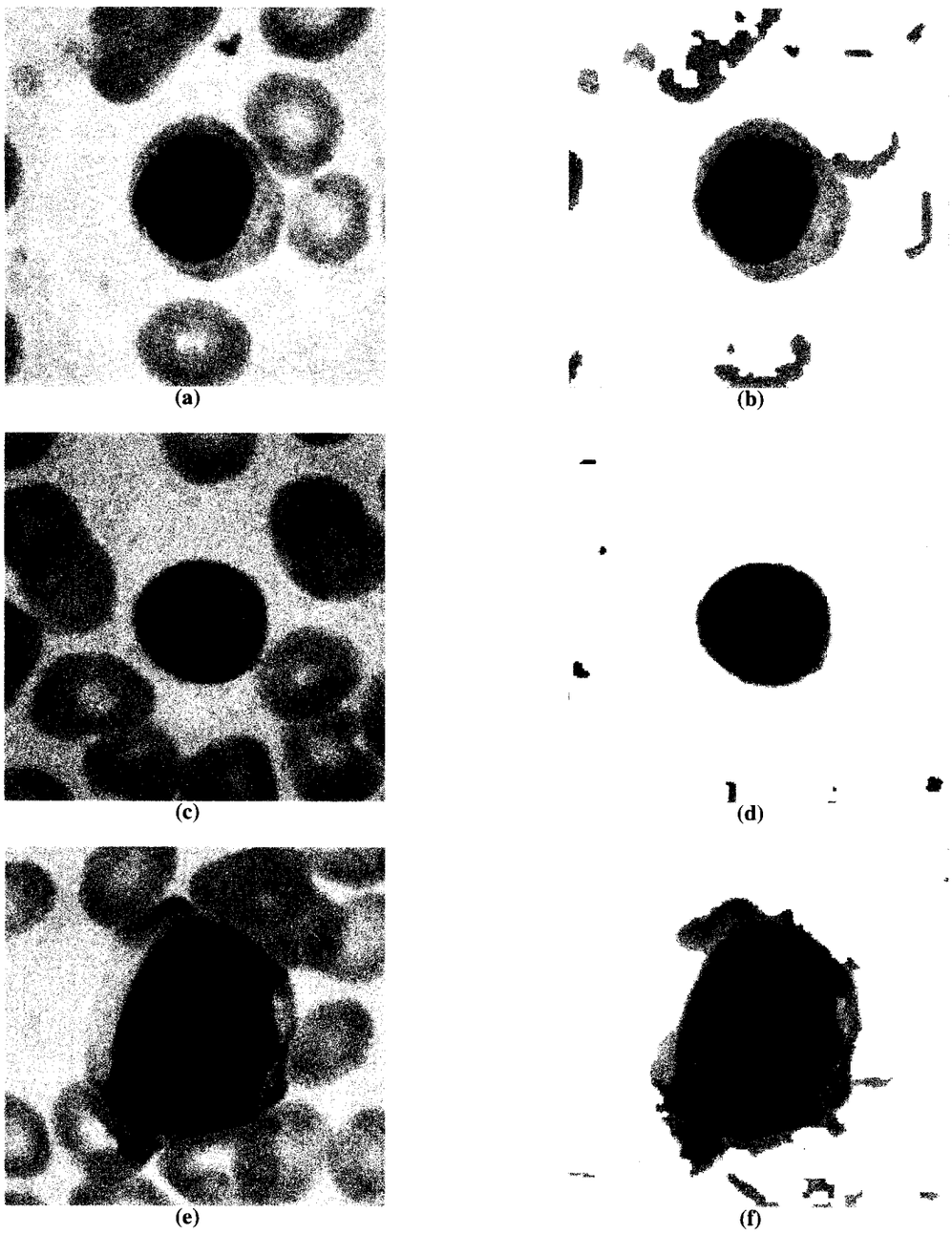


Figura C.7: **(a,c,e)** Micrografias de linfócitos. **(b,d,f)** Segmentações por evolução de curva.

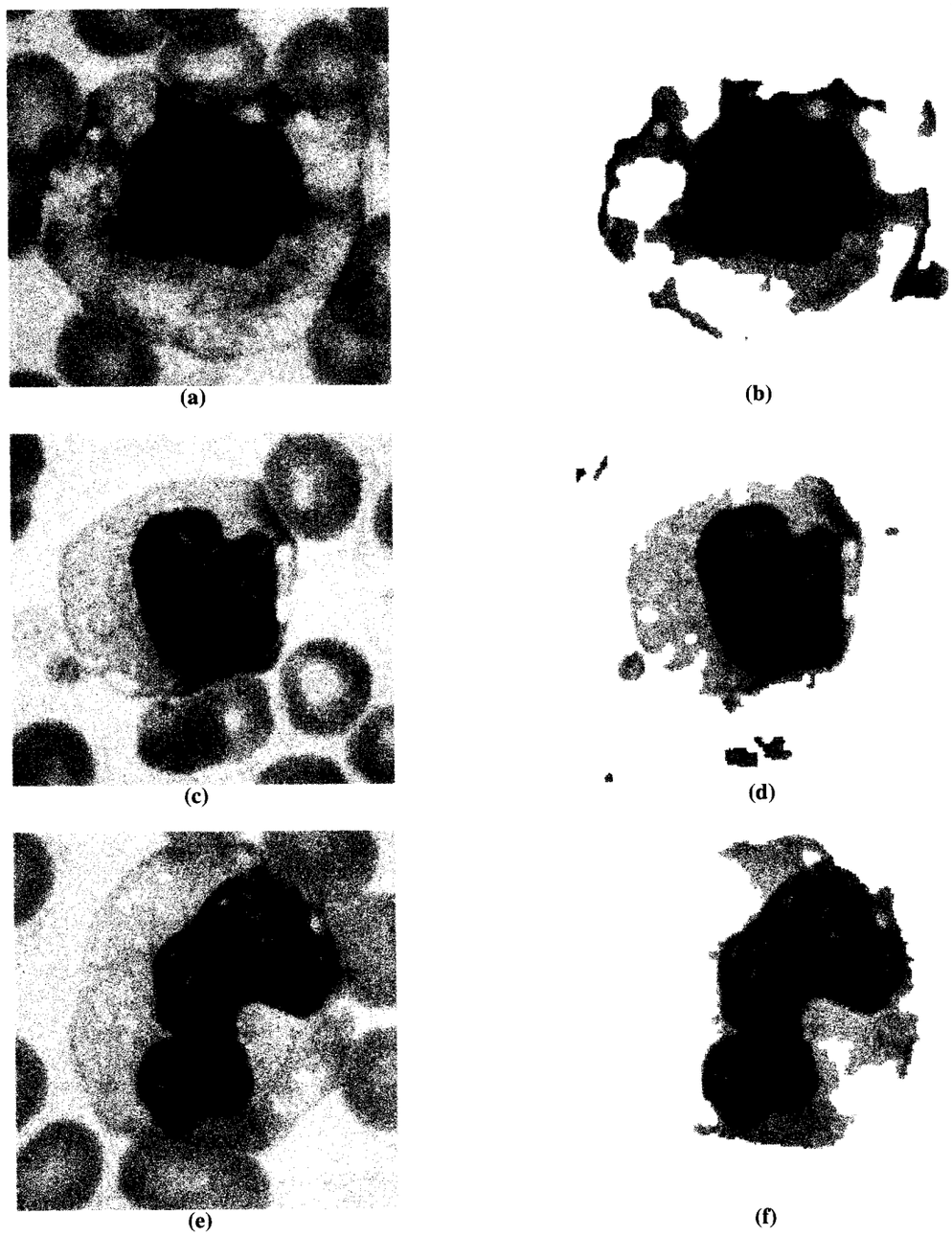


Figura C.8: **(a,c,e)** Micrografias de monócitos. **(b,d,f)** Segmentações por evolução de curva.

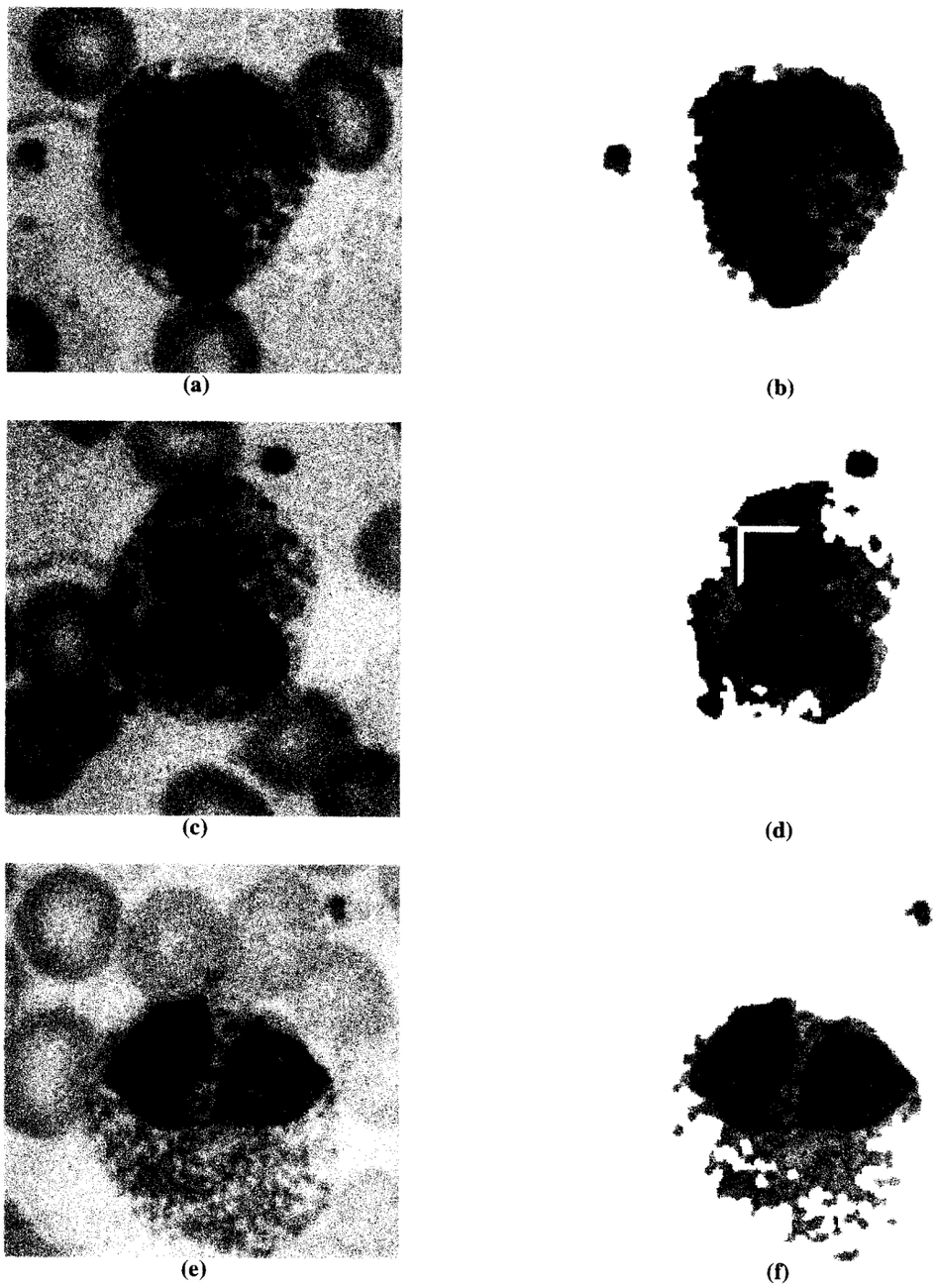


Figura C.9: **(a,c,e)** Micrografias de eosinófilos. **(b,d,f)** Segmentações por evolução de curva.

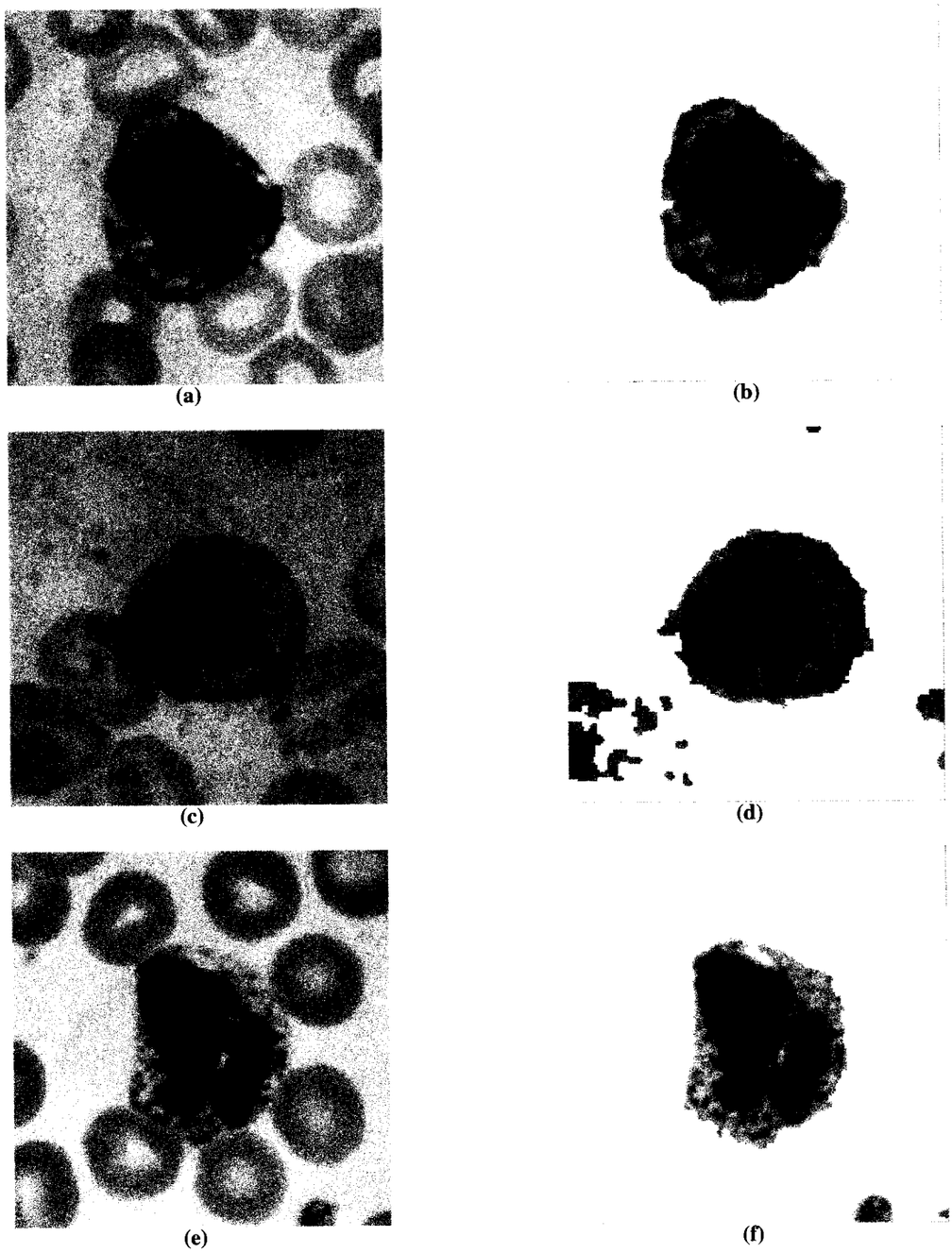


Figura C.10: **(a,c,e)** Micrografias de basófilos. **(b,d,f)** Segmentações por evolução de curva.

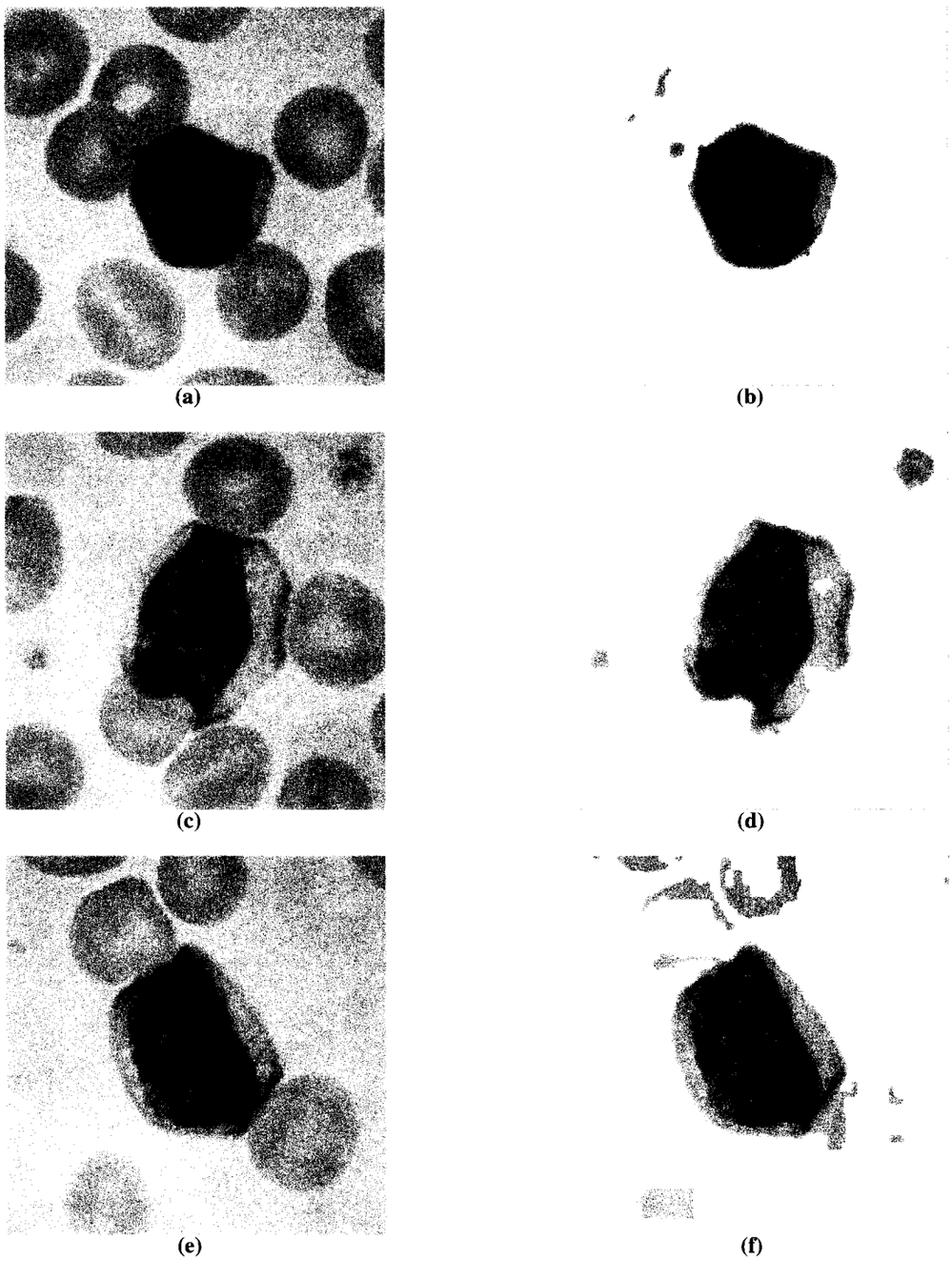


Figura C.11: **(a,c,e)** Micrografias de LLC. **(b,d,f)** Segmentações por evolução de curva.

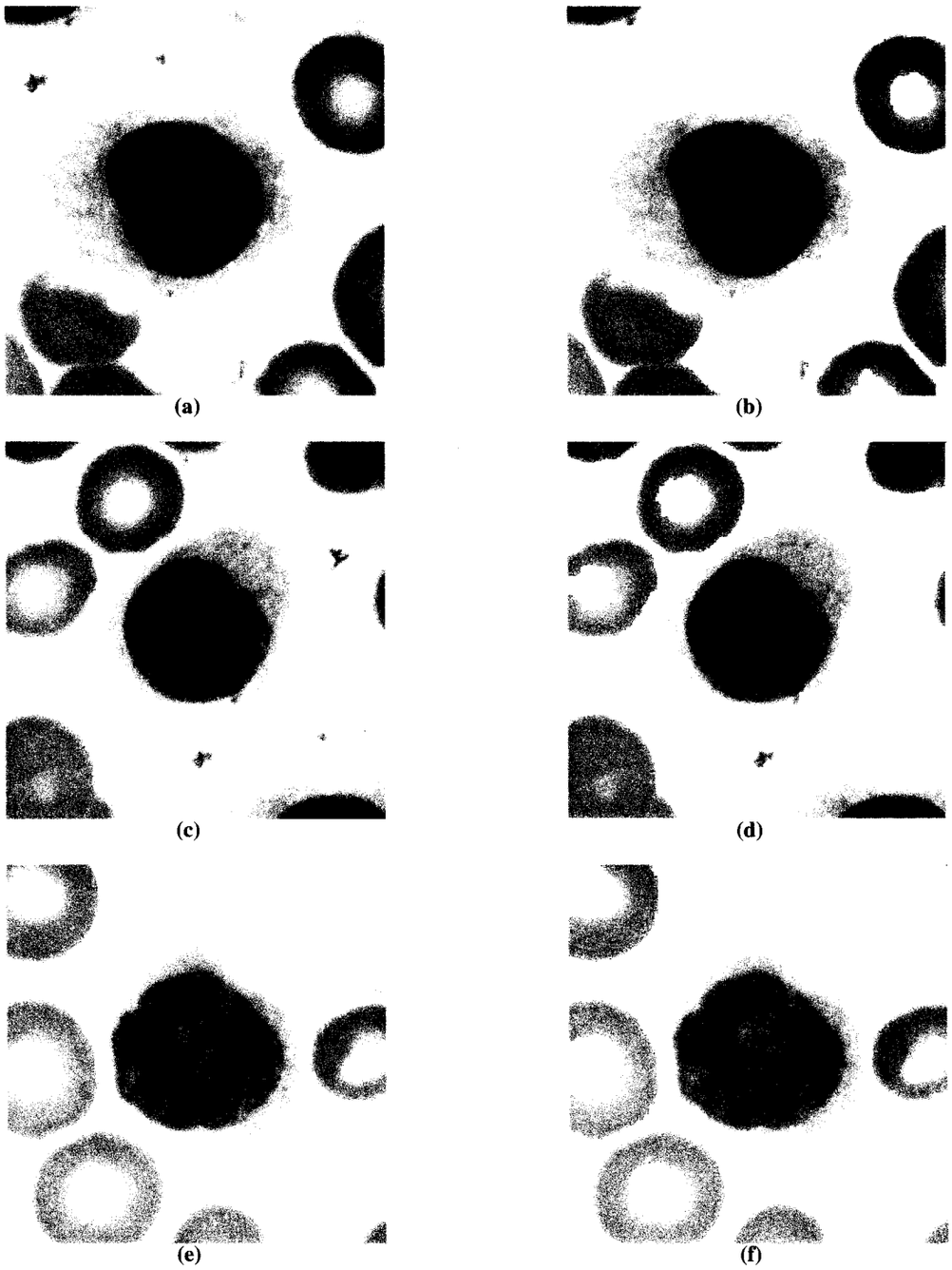


Figura C.12: **(a,c,e)** Micrografias de Hairy cells. **(b,d,f)** Segmentações por evolução de curva.

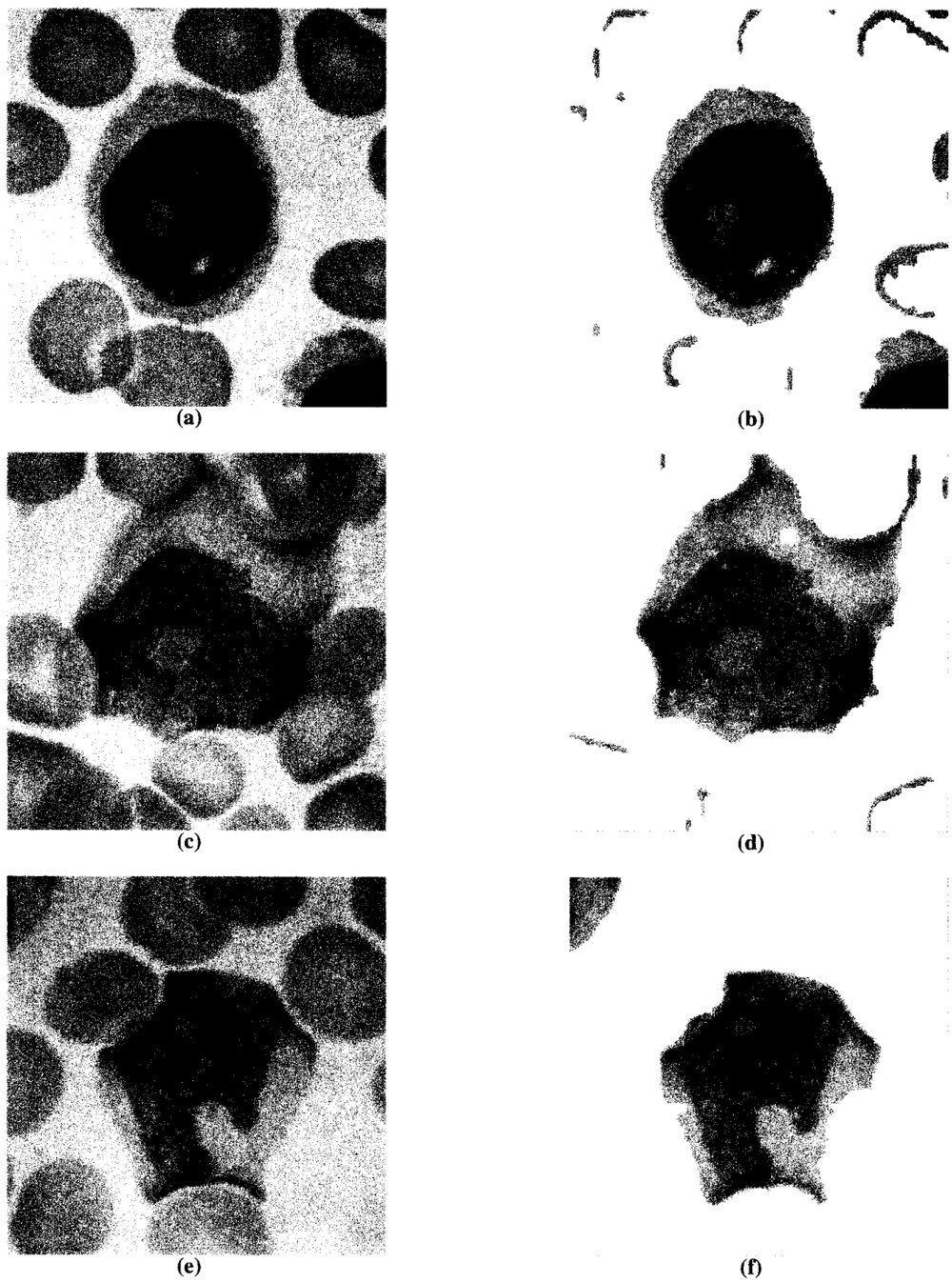


Figura C.13: **(a,c,e)** Micrografias de prolinfócitos. **(b,d,f)** Segmentações por evolução de curva.

Apêndice D

Manual dos programas

As seções abaixo visam familiarizar futuros usuários dos programas desenvolvidos através de uma documentação bastante simplificada. Maiores detalhes devem ser disponibilizados em futura página na internet.

D.1 Leuko

A seguir, segue descrição sucinta sobre as funcionalidades de *Leuko*, sendo que o próprio programa possui balões explicativos a respeito das funções dos botões. A Fig.D.1 mostra as janelas usadas durante o processo de segmentação da micrografia nas regiões de interesse para classificação. Os números nos círculos indicam as funções dos botões no painel, cujas funções são indicadas abaixo:

1. Abre uma janela contendo uma micrografia colorida
2. Uma vez aberta a imagem, pode-se realizar o treinamento das ROI para classificação supervisionada baseada em cor;
3. Apresenta a segmentação da imagem aberta no passo 1 levando em consideração o treinamento do passo 2;
4. Extrai as características da imagem corrente;
5. Treina um classificador bayesiano baseando-se em vetores de características provenientes de outros leucócitos pré-classificados (classificação supervisionada do leucócito);

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

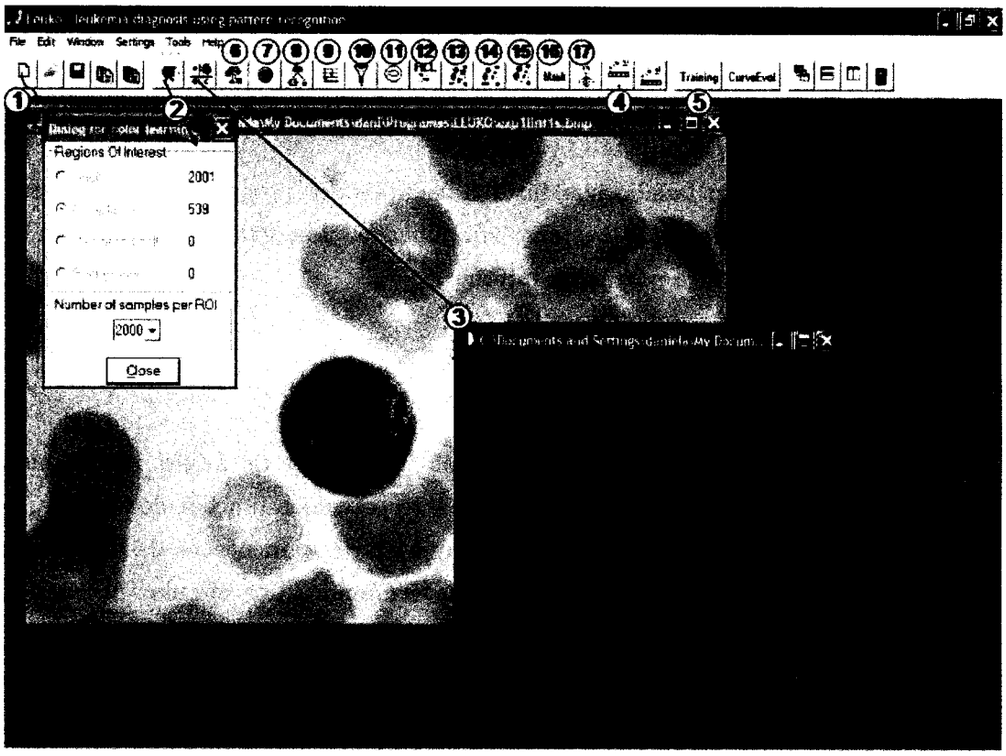


Figura D.1: Interface de Leuko e suas funcionalidades

6. Segmenta várias imagens, dado um conjunto de arquivos de treinamento (*nuc,cyt,rbc,bkg.txt*)
7. Segmenta o núcleo do leucócito usando o histograma do canal G, dado que o passo 1 tenha sido executado anteriormente;
8. Divide a imagem segmentada, dada pelo passo 3, em duas janelas com imagens binárias, uma contendo o núcleo e outra o citoplasma;
9. Mostra o histograma do canal G dado que o passo anterior tenha sido o passo 1;
10. Aplica filtros morfológicos a uma imagem binária;
11. Extrai o contorno de uma imagem binária;
12. Preenche os buracos de um objeto, representado por pixels branco, em imagem binária;
13. Rotula objetos a partir de uma imagem binária, marcando-os com diferentes cores dado um limiar de área, que pode ser mudado no menu *Settings*;
14. Faz contagem do número de objetos em uma imagem binária, colocando a área correspondente sobre os objetos;
15. Seleciona objetos de uma imagem binária dado a área e proximidade ao maior objeto;
16. Dado duas imagens em janelas dentro do programa, uma contendo a imagem colorida e outra contendo a segmentação, mascara a imagem colorida mostrando somente a região dada pela segmentação;
17. Extrai os esqueletos do objeto em imagem binária;

D.2 Seleção de características

A interface do programa de seleção de características é flexível pois é montada em tempo de execução a partir de um arquivo (*features.txt*) que deve estar no mesmo diretório que o programa; esse arquivo deve conter todas as características utilizadas, uma em cada linha do arquivo. O botão *Load samples and Run* permite carregar

uma matriz (de arquivos *.txt*) de amostras rotuladas de acordo com a classe a que pertencem seguindo o padrão mais aceito onde cada linha representa uma amostra, a última coluna de cada linha, representada por um número inteiro, dá a classe a que pertence a amostra; cada uma das outras colunas representa uma determinada característica. Diferentes amostras relativas a uma mesma classe devem ter o mesmo código de classe. As amostras não precisam estar agrupadas por classe e podem estar em um ou mais arquivos, que podem ser utilizados em uma execução concomitantemente. Para carregar mais de um vetor ou matriz de amostras simultaneamente usa-se $CTRL + mousedown$.

O programa permite a execução de 14 diferentes buscas nas amostras, dadas todas as combinações possíveis entre abordagens de busca, algoritmo de busca, critério de separabilidade e distância entre pontos. É possível executar todas as buscas em uma única requisição ao software, que após executada, retorna dois arquivos, *allMethDist.txt* e *allMethOrder.txt*, contendo a evolução das distâncias entre as classes e o *ranking* das características, respectivamente, para as diferentes buscas. Apenas a distância euclidiana e de Mahalanobis foram implementadas até o momento. O diagrama de blocos da Fig.5.3 mostra as funções atualmente implementadas, com as possibilidades de exploração dos dados para estimar qual o grupo de características mais promissor.

Entretanto, melhorias devem ser adotadas para estimação da distância entre as classes. A avaliação adotada utiliza a média das distâncias entre as classes, que não é uma boa estimativa da separação entre as classes quando o número de classes é grande. Uma possibilidade seria testar a separabilidade entre uma classe e todas as outras, realizando um estudo das características mais relevantes para todas as classes.

D.3 Weka

Acrônimo de **W**aikato **E**nvironment for **K**nowledge **A**nalysis, esse programa foi desenvolvido na Universidade de Waikato na Nova Zelândia. Weka é uma coleção de algoritmos de aprendizagem de máquina para resolver problemas de mineração de dados do mundo real. Ele é escrito em Java e executa em quase todas as plataformas. Os algoritmos podem ser tanto aplicados diretamente ao conjunto de dados ou

chamados de seu próprio código Java. Weka é também adequado para desenvolvimento e teste de novos esquemas de aprendizado de máquina. Weka é um software de código fonte livre sob licença pública geral GNU.

Dentre os vários esquemas de aprendizado disponibilizados pelo Weka, escolhemos 4 métodos:

1. Naive Bayes (comando weka.classifiers.NaiveBayes): implementa o classificador probabilístico bayesiano ingênuo, discutido na seção 30. Por default, usa a distribuição normal para modelar atributos numéricos, embora exista opção de estimativa de *kernel*;

2. árvore de decisão (comando weka.classifiers.j48.j48);

3. rede neural (comando weka.classifiers.neural.NeuralNetwork): é uma rede neural que usa *backpropagation* para treinar. O número de camadas escondidas utilizada é $a = (atributos + classes)/2$;

4. k-médias (comando weka.clusteres.SimpleKMeans): algoritmo de clusterização.

A questão de desempenho na predição de classes baseada em dados limitados é interessante, mas ainda controversa. Dentre as diferentes técnicas, a validação cruzada (*crossvalidation*) tem predominado e é provavelmente o método de avaliação de escolha na maioria das situações práticas onde os dados são limitados.

O ambiente Weka conta com alguns parâmetros para medir custo de erros, ou seja, avaliar e quantificar o desempenho dos classificadores [41], explicados a seguir (maiores detalhes[1]):

1. TP rate (TP = true positives): dá a taxa de classificações corretas da classe em questão.

2. FP rate (FP = false positives): dá a taxa de classificações incorretas da classe em questão.

3. Precision (P): é definida como o número de amostras corretamente classificadas como positivas dividido pelo número total de amostras classificadas como positivas.

[1] http://www.gsi.dit.upm.es/ lssii/soft/papr/weka/doc/weka.classifiers.Evaluation.html

4. Recall (R): é definida como o número de amostras corretamente classificadas como positivas dividido pelo número total de amostras positivas.

5. F-measure = $\frac{2 \times P \times R}{P \times R}$

Exemplificando os conceitos acima, suponha a matriz de confusão (Tabela D.1):

Tabela D.1: Exemplo de uma matriz de confusão relacionando o número de amostras com o resultado de classificação, dadas as classes C1 e C2

Classe	C1	C2	$\sum_{linha}$
C1	533	79	612
C2	276	99	375
$\sum_{coluna}$	809	178	987

Dado um classificador, o ambiente Weka produz os seguintes parâmetros de saída (Tabela D.2):

Tabela D.2: Precisão detalhada por classe considerando as 10 características mais relevantes

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.871	0.736	0.659	0.871	0.75	C1
0.264	0.129	0.556	0.264	0.358	C2

Assim, TP=0.871 é resultado de 533 / (533 + 79) a partir dos valores da primeira linha da matriz de confusão. Já FP=0.736 é resultado de 276 / (276 + 99) a partir dos valores da segunda linha da matriz de confusão. O resultado ideal é que FP seja zero. O valor de P = 0.659 para C1 é resultado de 533 / (533 + 276). O valor de R = 0.871 para C1 é resultado de 533 / (533 + 79).

Seguem alguns testes considerando a estimativa de erro 10-fold crossvalidation (métodos 1, 2 e 3) para um conjunto de 1055 imagens de 8 tipos de leucócitos diferentes, levando em consideração as 62 características discutidas na seção 5.7. As classes serão representadas de acordo com o código a seguir:

0. Basófilo;
1. Eosinófilo;
2. Linfócito;
3. Monócito;
4. Neutrófilo;
5. LLC;

6. Prolinfócito;
7. Hairy cell;

D.3.1 Naive Bayes

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 727 68.91 %

Incorrectly Classified Instances 328 31.09 %

Kappa statistic 0.6337

Mean absolute error 0.0775

Root mean squared error 0.2694

Relative absolute error 37.5842 %

Root relative squared error 83.9101 %

Total Number of Instances 1055

Tabela D.3: Precisão detalhada por classe considerando 62 características.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.909	0.018	0.69	0.909	0.784	0
0.868	0.063	0.605	0.868	0.713	1
0.63	0.07	0.768	0.63	0.692	2
0.685	0.074	0.514	0.685	0.587	3
0.58	0.006	0.966	0.58	0.724	4
0.715	0.038	0.761	0.715	0.737	5
0.809	0.048	0.442	0.809	0.571	6
0.771	0.04	0.581	0.771	0.663	7

Tabela D.4: Matriz de confusão.

a	b	c	d	e	f	g	h	↩	Classified as		Correct
40	1	3	0	0	0	0	0	a	=	0	90.91%
4	92	1	3	4	0	1	1	b	=	1	86.79%
2	0	179	27	1	23	23	29	c	=	2	63.03%
4	1	13	74	0	0	15	1	d	=	3	68.52%
4	56	6	34	142	1	0	2	e	=	4	57.96%
3	2	23	1	0	108	8	6	f	=	5	71.52%
1	0	2	3	0	3	38	0	g	=	6	80.85%
0	0	6	2	0	7	1	54	h	=	7	77.14%

D.3.2 Árvore de decisão

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 877 83.128 %

Incorrectly Classified Instances 178 16.872 %

Kappa statistic 0.7946

Mean absolute error 0.0479

Root mean squared error 0.1974

Relative absolute error 23.217 %

Root relative squared error 61.4774 %

Total Number of Instances 1055

Tabela D.5: Precisão detalhada por classe considerando 62 características.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.818	0.014	0.72	0.818	0.766	0
0.792	0.016	0.848	0.792	0.82	1
0.842	0.079	0.797	0.842	0.818	2
0.815	0.012	0.889	0.815	0.85	3
0.927	0.028	0.908	0.927	0.917	4
0.788	0.033	0.799	0.788	0.793	5
0.809	0.012	0.76	0.809	0.784	6
0.657	0.012	0.793	0.657	0.719	7

Tabela D.6: Matriz de confusão

a	b	c	d	e	f	g	h	↩P	Classified as		Correct
36	2	5	0	1	0	0	0	a	=	0	81.82%
3	84	2	0	14	3	0	0	b	=	1	79.25%
3	1	239	6	4	16	8	7	c	=	2	84.15%
3	0	11	88	1	2	2	1	d	=	3	81.48%
2	10	2	1	227	2	0	1	e	=	4	92.65%
2	1	23	1	2	119	0	3	f	=	5	78.81%
0	1	3	3	0	2	38	0	g	=	6	80.85%
1	0	15	0	1	5	2	46	h	=	7	65.71%

D.3.3 Rede neural

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 962 91.1848 %

Incorrectly Classified Instances 93 8.8152 %

Kappa statistic 0.8928

Mean absolute error 0.0272

Root mean squared error 0.1362

Relative absolute error 13.1671 %

Root relative squared error 42.4385 %

Total Number of Instances 1055

Tabela D.7: Precisão detalhada por classe considerando 62 características.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.932	0	1	0.932	0.965	0
0.962	0.003	0.971	0.962	0.967	1
0.891	0.047	0.875	0.891	0.883	2
0.907	0.006	0.942	0.907	0.925	3
0.976	0.016	0.948	0.976	0.962	4
0.894	0.017	0.9	0.894	0.897	5
0.851	0.011	0.784	0.851	0.816	6
0.771	0.009	0.857	0.771	0.812	7

Tabela D.8: Matriz de confusão

a	b	c	d	e	f	g	h	↩	Classified as		Correct
41	0	1	0	2	0	0	0	a	=	0	93.18%
0	102	0	0	4	0	0	0	b	=	1	96.23%
0	0	253	2	2	14	6	7	c	=	2	89.08%
0	0	6	98	2	0	2	0	d	=	3	90.74%
0	2	1	2	239	0	0	1	e	=	4	97.55%
0	1	9	0	2	135	3	1	f	=	5	89.40%
0	0	4	2	0	1	40	0	g	=	6	85.11%
0	0	15	0	1	0	0	54	h	=	7	77.14%

D.3.4 K-médias

Tabela D.9: Desempenho do k-médias para os diferentes agrupamentos

Classe	amostra	acerto
0	164	16%
1	144	14%
2	176	17%
3	34	3%
4	125	12%
5	141	13%
6	171	16%
7	100	9%

Tabela D.10: Classes para agrupamentos considerando 62 características.

0	1	2	3	4	5	6	7	↤ assigned to cluster
2	2	4	30	0	0	6	0	0
4	84	1	3	2	0	12	0	1
109	0	99	0	19	44	0	13	2
33	2	6	0	36	30	1	0	3
4	56	0	1	31	0	152	1	4
6	0	49	0	1	11	0	84	5
4	0	0	0	0	41	0	2	6
2	0	17	0	36	15	0	0	7

Tabela D.11: Número do cluster e correspondente conforme código do leucócito

Cluster	Código do leucócito
0	3
1	1
2	2
3	0
4	7
5	6
6	4
7	5

IFSC-USP SERVIÇO DE BIBLIOTECA INFORMAÇÃO

Seguem alguns testes considerando a estimativa de erro 10-fold crossvalidation (métodos 1, 2 e 3) para um conjunto de 1055 imagens de 8 tipos de leucócitos diferentes, levando em consideração as 10 características mais relevantes, segundo selector do Weka.

O método de seleção de características adotado considera os critérios seguintes:

Filtro: filtro de seleção de atributos;

Busca: *bestfirst* - busca o espaço de subconjunto de atributos através do aumento do número de características, ou seja, começa-se com um conjunto vazio de atributos, que é acrescido sempre do atributo que garantiu maior taxa de acerto;

Avaliador: avaliador de subconjunto por *wrapper* - avalia o conjunto de atributos através do uso de um esquema de aprendizado. A validação cruzada é usada para estimar a precisão do esquema de aprendizagem;

Aprendizagem: Naïve Bayes;

Estimativa de erro: 10-fold crossvalidation.

O código das 10 características selecionadas pelo Weka são 2, 3, 4, 14, 15, 35, 43, 53, 57, 61, que significam respectivamente: área do núcleo, área da maior bolota do núcleo, razão núcleo-citoplasma, correlação na textura do núcleo, inércia na textura com janela pequena do citoplasma, correlação na textura com janela grande do citoplasma, skewness do canal G do citoplasma, média do canal B do citoplasma e a energia do canal B do citoplasma.

D.3.5 Naive Bayes

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 884 83.7915 %

Incorrectly Classified Instances 171 16.2085 %

Kappa statistic 0.8031

Mean absolute error 0.0516

Root mean squared error 0.1827

Relative absolute error 25.0466 %

Root relative squared error 56.9151 %

Total Number of Instances 1055

Tabela D.12: Precisão detalhada por classe considerando as 10 características mais relevantes

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.864	0.005	0.884	0.864	0.874	0
0.887	0.017	0.855	0.887	0.87	1
0.82	0.079	0.793	0.82	0.806	2
0.898	0.023	0.815	0.898	0.855	3
0.914	0.012	0.957	0.914	0.935	4
0.801	0.042	0.761	0.801	0.781	5
0.638	0.011	0.732	0.638	0.682	6
0.671	0.008	0.855	0.671	0.752	7

Tabela D.13: Matriz de confusão

a	b	c	d	e	f	g	h	↩	Classified as		Correct
38	1	4	0	1	0	0	0	a	=	0	86.36%
0	94	3	0	9	0	0	0	b	=	1	88.68%
2	0	233	10	0	25	7	7	c	=	2	82.04%
2	0	7	97	0	0	2	0	d	=	3	89.81%
1	13	2	3	224	2	0	0	e	=	4	91.43%
0	2	26	1	0	121	1	0	f	=	5	80.13%
0	0	5	7	0	4	30	1	g	=	6	63.83%
0	0	14	1	0	7	1	47	h	=	7	67.14%

D.3.6 Árvore de decisão

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 852 80.7583 %

Incorrectly Classified Instances 203 19.2417 %

Kappa statistic 0.7666

Mean absolute error 0.0552

Root mean squared error 0.2081

Relative absolute error 26.7722 %

Root relative squared error 64.831 %

Total Number of Instances 1055

Tabela D.14: Precisão detalhada por classe considerando as 10 características mais relevantes.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.841	0.007	0.841	0.841	0.841	0
0.849	0.019	0.833	0.849	0.841	1
0.757	0.086	0.765	0.757	0.761	2
0.806	0.026	0.777	0.806	0.791	3
0.931	0.017	0.942	0.931	0.936	4
0.768	0.051	0.716	0.768	0.741	5
0.66	0.01	0.756	0.66	0.705	6
0.686	0.017	0.738	0.686	0.711	7

Tabela D.15: Matriz de confusão

a	b	c	d	e	f	g	h	↩	Classified as		Correct
37	2	2	2	1	0	0	0	a	=	0	84.09%
2	90	1	0	10	3	0	0	b	=	1	84.91%
0	1	215	12	1	36	6	13	c	=	2	75.70%
1	0	13	87	2	0	3	2	d	=	3	80.56%
3	12	0	1	228	1	0	0	e	=	4	93.06%
1	3	28	1	0	116	1	1	f	=	5	76.82%
0	0	6	7	0	2	31	1	g	=	6	65.96%
0	0	16	2	0	4	0	48	h	=	7	68.57%

D.3.7 Rede neural

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 887 84.0758 %

Incorrectly Classified Instances 168 15.9242 %

Kappa statistic 0.8055

Mean absolute error 0.0463

Root mean squared error 0.1788

Relative absolute error 22.4307 %

Root relative squared error 55.6942 %

Total Number of Instances 1055

Tabela D.16: Precisão detalhada por classe considerando as 10 características mais relevantes.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.841	0.007	0.841	0.841	0.841	0
0.849	0.019	0.833	0.849	0.841	1
0.757	0.086	0.765	0.757	0.761	2
0.806	0.026	0.777	0.806	0.791	3
0.931	0.017	0.942	0.931	0.936	4
0.768	0.051	0.716	0.768	0.741	5
0.66	0.01	0.756	0.66	0.705	6
0.686	0.017	0.738	0.686	0.711	7

Tabela D.17: Matriz de confusão.

a	b	c	d	e	f	g	h	↩	Classified as		Correct
36	2	2	1	3	0	0	0	a	=	0	81.82%
3	93	2	0	8	0	0	0	b	=	1	87.74%
1	2	246	1	2	18	7	7	c	=	2	86.62%
1	0	4	99	2	0	2	0	d	=	3	91.67%
3	7	0	4	231	0	0	0	e	=	4	94.29%
1	3	33	1	1	108	1	3	f	=	5	71.52%
0	0	12	3	0	0	32	0	g	=	6	68.09%
0	0	19	0	2	3	4	42	h	=	7	60.00%

D.3.8 K-médias

Tabela D.18: Desempenho do k-médias para os diferentes agrupamentos

Classe	amostra	acerto
0	152	14%
1	121	11%
2	208	20%
3	41	4%
4	116	11%
5	150	14%
6	125	12%
7	151	14%

Tabela D.19: Classes para agrupamentos considerando as 10 características mais relevantes.

0	1	2	3	4	5	6	7	assigned to cluster
0	1	4	36	3	0	0	0	0
1	88	1	1	6	0	9	0	1
110	1	133	0	0	11	0	29	2
3	0	4	2	0	99	0	0	3
0	18	2	2	106	0	116	1	4
19	3	9	0	1	2	0	117	5
0	0	10	0	0	34	0	3	6
19	1	45	0	0	4	0	1	7

Tabela D.20: Número do cluster e correspondente conforme código do leucócito

Cluster	Código do leucócito
0	2
1	1
2	7
3	0
4	sem classe
5	3
6	4
7	5

Apêndice E

Artigos publicados sobre este projeto de doutorado

Listam-se abaixo, os artigos submetidos e aceitos para publicação, incluindo resumos e textos completos:

1. Resumo: Diferenciação de leucócitos por computador (Hemato 2001);
2. Resumo: Automatic leukemia diagnosis (CSBMM 2001);
3. **Artigo**: Automatic leukemia diagnosis (**Acta Microscopica 2003**);
4. **Artigo**: Chromatin texture characterization using multiscale fractal dimension (DSP 2002);
5. Resumo: Diagnóstico de leucemia por computador (Workshop LADI 2002);
6. Resumo: Clusterização de dados de leucócitos usando modelo superparamagnético (Escola de Computação de Alto Desempenho para Sistema Complexos);
7. Resumo: Feature selection methods applied to multiscale texture (SIBGRAPI 2003);
8. Resumo: Texture approach to leukocyte recognition (ICOBICOBI 2003);
9. **Artigo**: Texture approach to leukocyte recognition (**Real-Time Imaging 2004**);

10. **Artigo**: Toward leukocyte recognition using morphometry, texture and color (ISBI 2004) - **premiado**;

11. **Resumo**: Fast image segmentation: quantitative evaluation and image application (SIBGRAPI 2004);

Bibliografia

[1] Agosti S J, Cornbleet P J, Galagan K, Gewirtz A S, Glassy E F, Novak R, and Spier C. *Color Atlas of Hematology: an illustrated field guide based on proficiency testing.* College of American Pathologists Hematology and Clinical Microscopy Resource Commitee, 1st edition, 1998. xi, xv, 4, 7, 9, 11, 12, 13, 15, 16, 22, 23, 24, 26, 82

[2] O'Connor B H. *A Color Atlas and Instruction Manual of Peripheral Blood Cell Morphology.* Lippincott Williams and Wilkins, 1st edition, 1984. xi, 7, 9, 13, 14, 17, 20, 22, 23, 27

[3] Haralick R M, Shanmugan K, and Dinstein I. Textural features for image classification. *IEEE Transactions Systems, Man, and Cybernetics*, SMC(9):610–621, 1973. xi, 32, 53, 54, 101

[4] Jain A K. Advances in statistical pattern recognition. In P. A. Devijever and J. Kittler, editors, *Pattern Recognition Theory and Applications.* Springer-Verlag, Berlin, 1987. xii, 59

[5] Fayyad U, Piatetsky-Shapiro G, and Smyth P. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996. xii, 78

[6] Glymour C, Madigan D, Pregibon D, and Smyth P. Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery 1*, 1:11–28, 1997. xv, 77, 78

[7] Barbosa Filho S. Um quantificador da ploidia tumoral através da citofotometria. Engenharia elétrica e informática industrial, CEFET, Curitiba-PR, Brasil, Dezembro 1994. 2

[8] Hubel D H. *Eye, Brain, and Vision.* W. H. Freeman, New York, 2nd edition, 1995. 4

[9] Sabino D M U, Costa L da F, Rizzatti E G, and Zago M A. Toward leukocyte recognition using morphometry, texture and color. In *Proc. of International Symposium on Biomedical Imaging: from Nano to Macro*, Arlington, VA, USA, apr 2004. IEEE. 6, 85, 106, 114, 115

[10] Hoffbrand A V and Pettit J E. *Atlas Colorido de Hematologia Clínica.* Editora Manole, 3rd edition, 2000. 7, 9, 20

[11] D'onofrio G and Zini G. Morphology of the blood. *Butterworth Heinemann*, 1998. 8, 12, 14, 81

[12] Can J H and Rodak B F. *Atlas de hematologia.* Livraria Santos Editora, 2000. 9, 20

[13] Duda R O, Hart P E, and Stork D G. *Pattern classification.* Wiley Interscience, New York, 2001. 30, 59, 62, 63, 64, 65, 66, 114

[14] Sonka M, Hlavac V, and Boyle R. *Image, Processing, Analysis and Machine Vision.* Chapman & Hall, 1993. 30, 37, 38, 39, 46, 49, 98

[15] Pratt W K. *Digital Image Processing (2nd Edition).* John Wiley & Sons, Inc., New York, New York, 1991. 30, 33, 45, 47, 49, 52

[16] Costa L F and Marcondes Jr. R C. *Shape analysis and classification:Theory and Practice.* CRC Press, Boca Raton, 2001. 31, 33, 49, 50, 56, 60, 76

[17] Haralick R M. Statistical and structural approaches to texture. *Proc. IEEE*, 67(5):786–804, 1979. 32, 53, 54, 101

[18] Castleman K R. *Digital Image Processing.* Prentice Hall, 1st edition, 1996. 33

[19] Lucchese L and Mitra S. Advances in color image segmentation. In *Global Telecommunications Conference - Globecom*, Multimedia Services and Technology Issues, pages 2038–2044, 1999. 33, 47

[20] Serra J. *Image Analysis and Mathematical Morphology*, volume 1. Academic Press, 1982. 33

[21] Perona P and Malik J. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:629–639, 1990. 35

[22] Land E H and McCann J J. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971. 35, 36

[23] Toveé M J. *An introduction to the visual system.* Cambrigde University Press, 1996. 35

[24] Harms H and Aus H M. Tissue image segmentation with multicolor, multifocal algorithms. In P. A. Devijver and J. Kittler, editors, *Pattern Recognition Theory and Applications, (NATO Advanced Study Institute, Spa-Balmoral, Belgium, June 9–20, 1986)*, pages 519–528, Berlin, 1986. Springer. 36

[25] Katz A R J. Image analysis and supervised learning in the automated differentiation of white blood cells from microscopic images. Applied science in information technology, RMIT, Melbourne-VIC, Australia, Fevereiro 2000. Minor thesis submitted in partial fulfillment of the requirements for the degreee of Master. 39, 46, 56, 90, 97

[26] Stevens C F. *The six core theories of modern physics.* MIT Press, 1998. 40

[27] Consularo L A. *Qualificação por artigo: Two-dimensional entropic segmentatin.* Monografia de Qualificação para o Doutorado em Física Computacional apresentada ao FFI-IFSC-USP em Novembro de 1999, 1999. 40, 43

[28] Kapur J N, Sahoo P K, and Wong A K C. new method for gray level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics and Image Processing*, 29:273–285, 1985. 42

[29] Ushizima D M, Consularo L A, Costa L da F, and Manjunath B S. Fast image segmentation: quantitative evaluation and image applications. In *XV Brazilian Symposium on Computer Graphics and Image Processing*, 2004. 44, 92

[30] Aus H M, Harms H, ter Meulen V, and Gunzer U. Statistical evaluation of computer extracted blood cell features for screening populations to detect leukemias. *PRTA*, 87:509–518, 1987. 46

[31] Rodenacker K and Bengtsson E. A feature set for cytometry on digitized microscopic images. *Analytical Cellular Pathology*, 25:1–36, 2003. 46, 53, 89, 100

[32] Young I T, Verbeeck P W, and Mayall B H. Characterization of chromatin distribution in cell nuclei. *Cytometry*, 7:467–474, 1986. 46

[33] Young I T and Roos R. Acuity: Image analysis for the personal computer. *Pattern Recognition and Artificial Intelligence*, pages 5–16, 1988. 46

[34] Moreira J. *Uma proposta de estruturacao e integração de processamento de cores em sistemas artificiais de visão.* IFSC-USP, 1st edition, 1999. 49

[35] Nilsson B and Heyden A. Segmentation of dense leukocyte clusters. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'01)*, pages 221–227, Kauai, Hawaii, December 2001. 50, 117

[36] Zhong Y and Jain A K. Object localization using color, texture and shape. *Pattern Recognition*, 33:671–684, jan 2000. 51

[37] Song X, Abu mostafa Y, Sill J, and Kasdan H. Incorporating contextual information in white blood cell image identification. In *Neural Information Processing Systems*, volume 10 of *NIPS97*. Mit Press, 1997. 52, 85

[38] Danielsen H E, Albregtsen F, and Baheerathan S. New texture features based on the complexity curve. *Pattern Recognition*, 32:605–618, 1999. 53

[39] Braga A P, Carvalho A C P L F, and Ludermir T B. *Redes neurais artificiais: teoria e aplicações.* LTC Editora S.A., 2000. 56

[40] Bowman A and Robinson D. *Introduction to statistics: a computer illustrated text.* Iop Publishing Ltd, 1987. 56, 62

[41] Witten I H and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, 2000. 56, 57, 64, 65, 71, 73, 86, 106, 114, 149

[42] Michalski R S and Carbonell J G Mitchell T M. *Machine learning: an artificial intelligence approach.* Morgan Kaufmann Publishers, Inc., 1983. 56

[43] Jain A K, Duin R P W, and Mao J. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, jan 2000. 57, 58, 59, 64, 67, 69, 115

[44] Kohn A F. *Reconhecimento de Padrões: uma abordagem estatística.* Edusp - Dept. Eng Eletrônica, Escola Politécnica da USP, 1st edition, 1998. 60, 75

[45] Consularo L A. *Técnicas de mineração de dados para análise de imagens.* Tese de Doutorado apresentada ao FFI-IFSC-USP em 26.09.2000, 1st edition, 2000. 63, 64, 76, 78

[46] Fukunaga K. *Introduction to statistical pattern recognition.* Academic Press, 2nd edition, 1990. 64, 67, 75

[47] Campos T E. *Técnicas de seleção de atributos e de classificação para reconhecimento de faces.* Tese de Mestrado apresentada ao IME-USP, 1st edition, 2000. 64, 66, 108

[48] Langley P and Sabe S. Induction of selective bayesian classifiers. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 399–406, 1994. 64, 76, 114

[49] Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38. Society for Industrial and Applied Mathematics (SIAM); SERIES:CBMS-NSF Regional Conference Series in Appli, 1982. 69

[50] Dash M and Liu H. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997. 72

[51] Gagliardi I, Schettini R, and Ciocca G. Retrieving color images by content. *Proc. Image and Video Content based Retrieval*, pages 57–64, 1998. 74, 108

[52] John G H, Kohavi R, and Pfleger K. Irrelevant features and the subset selection problem. In Cohen W W; Hirsh H, editor, *Machine learning: Proceedings of the Eleventh International Conference*, San Francisco, CA, 1994. Morgan Kaufmann Publishers. 76, 77

[53] Kohavi R and John G H. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2):273–323, 1997. 76

[54] Two Crows Corporation. Introduction to data mining and knowledge discovery. *http://www.twocrows.com*, pages 1–36, 1999. 78

[55] Siqueira M, Scharcanski J, and Navaux P O A. Echocardiographic image sequence segmentation and analysis using self-organizing maps. *The Journal of VLSI Signal Processing*, 32:135–145, 2002. 85

[56] Camapum J, Silva A, Freitas A, and Bassani H. Segmentation of clinical structures from images of the human pelvic area. In *XV Brazilian Symposium on Computer Graphics and Image Processing*, 2004. 85

[57] Sabino D M U, Costa L da F, Martins S L R, and Zago M A. Diferenciação de leucócitos por computador. In *XVIII Congresso Nacional da Colégio Brasileiro de Hematologia*, 2001. 87

[58] Sabino D M U, Costa L da F, Calado R T, and Zago M A. Automatic leukemia diagnosis. In *XVIII Congress of the Brazilian Society For Microscopy and Microanalysis*, Águas de São Pedro, SP, Brasil, oct 2001. SBMM. 87

[59] Sabino D M U, Costa L da F, Calado R T, and Zago M A. Automatic leukemia diagnosis. *Acta Microscopica*, 12(1):1–6, dec 2003. 87

[60] Sabino D M U and Costa L da F. Diagnóstico de leucemia por computador. In *I Workshop Regional de Engenharia Biomédica*, São Carlos, SP, Brasil, oct 2002. EESC. 87, 100

[61] Sabino D M U, Mendes T, and Costa L da F. Clusterização de dados de leucócitos usando o modelo superparamagnético. In *Escola de Computação de Alto Desempenho para Sistemas Complexos*, São Carlos, SP, Brasil, jul 2003. IFSC-USP, Tereza Mendes. 87, 101

[62] Sabino D M U, Nakamura E K, Costa L da F, Calado R T, and Zago M A. Chromatin texture characterization using multiscale fractal dimension. In *14th International Conference on Digital Signal Processing Proceedings*, volume 2, pages 529–533, Santorini, Greece, jul 2002. IEEE. 87

[63] Sabino D M U, Costa L da F, Rizzatti E G, and Zago M A. A texture approach to leukocyte recognition. In *Proc. International Conference on Bioinformatics*

and Computational Biology, Ribeirão Preto, SP, Brasil, may 2003. IME-USP. 87

[64] Sabino D M U, Costa L da F, Rizzatti E G, and Zago M A. A texture approach to leukocyte recognition. *Real-Time Imaging*, 10(4):205–216, 2004. 87, 97, 115

[65] Sabino D M U and Costa L da F. Feature selection methods applied to multiscale curvature. In *XIV Brazilian Symposium on Computer Graphics and Image Processing*, São Carlos, SP, Brasil, oct 2003. ICMC-USP, SIBGRAPI. 87, 102, 115

[66] Lucchese L and Mitra S. Color segmentation based on separate anisotropic diffusion of chromatic and achromatic channels. In *IEEE Proceedings: Vision, Image, and Signal Processing*, pages 2038–2044, Sept. 89

[67] Rosin P L and Ioannidis E. Evaluation of global image thresholding for change detection. *PRL*, 24:2345–2356, 2003. 92

[68] Sumengen B, Manjunath B S, and Kenney C. Image segmentation using curve evolution and flow fields. In *IEEE International Conference on Image Processing (ICIP)*, Sep 2002. 99, 116, 135

[69] Blatt M, Wiseman S, and Domany E. Data clustering using a model granular magnet. *Neural Computing*, pages 1805–1842, 1997. 100

[70] Jain A K and Vailaya A. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, aug 1996. 108

[71] Sethian J A. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry.* Cambridge University Press, 1999. 116

[72] Catovsky D. Small b and t leukemias and the leukaemic phase of non-hodgkin lymphomas (nhl). In *Eigth European Tutorial On Haematopathology - Leukaemias and Lymphomas*, London, England, march 1999. European Haematology Association. 117

Índice

abertura, 33
aprendizado, 56
atributos, 49

basófilos, 8
Bayes, 62
borda, 38

célula cabeluda, 23
características densitométricas, 52
ceguinho, 50
circularidade, 50
citômetro de fluxo, 3
classificação, 58
coeficiente de assimetria, 52
concavidade, 51
contorno, 50
convexidade, 51
cor, 38, 51, 102
curse of dimensionality, 57
curtose, 52
curvatura, 50

Delphi, 82
desvio padrão, 52
difusão anisotrópica, 34, 131
dilatação, 33
distribuição normal, 61

energia, 52
energia de dobramento, 51
entropia, 40, 52
eosinófilos, 8
equalização, 31
erosão, 33
esfregaços de sangue, 26
espaço-escala, 53
estatística primeira ordem, 54
estatística segunda ordem, 54
evolução por curva, 100
experimentos, 87
extração de características, 49

fechamento, 33
filtro, 53, 131
filtros, 32
forma, 49, 100

gaussiana, 60
gaussiana multivariada, 61
GLCM, 53, 101, 114
GLCM normalizada, 53
granulócitos, 8

hemácia, 8
histograma, 38, 90

imagens utilizadas, 89
imunofenotipificação, 117
inferência estatística, 56

isotropia, 53

Lei da Probabilidade Total, 62
Leuko, 82
limiarização, 37–39, 44, 92
linfócitos, 8
linfoma, 15
LLC, 19
LLP, 21

média, 52
matriz de confusão, 100
matrizes de co-ocorrência, 53
mediana, 53
micrografias, 88
mineração de dados, 107
monócitos, 8
morfologia matemática, 33

não-supervisionado, 57
Naive bayesian classifier, 64
neutrófilos, 8
normalização, 56, 85

overfitting, 57
overtraining, 57

pós-processamento, 92
PCC, 131
peaking phenomena, 58
plaqueta, 8
plataforma, 82
pré-processamento, 89
probabilidade a priori,posteriori, 63
probabilidade condicional, 62
processamento, 30
programas desenvolvidos, 82

quantização, 31, 53

reconhecimento de padrões, 29
retinex, 35

segmentação, 36, 46, 89
segmentação por cor, 98
seleção de características, 71, 148
seleção de escala, 106
semântica, 96
superparamagnética, 87, 100
supervisionado, 56

textura, 38, 52, 101

vetor de característica, 82

Weka, 87