Método para Processamento e Análise Computacional de Imagens Histopatológicas Visando Apoiar o Diagnóstico de Câncer de Colo do Útero

Gisele Helena Barboni Miranda

DISSERTAÇÃO APRESENTADA
AO
PROGRAMA INTERUNIDADES EM BIOINFORMÁTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO GRAU DE MESTRE
EM
CIÊNCIAS

Área de Concentração: Bioinformática

Orientador: Prof. Dr. Joaquim Cezar Felipe

Durante a elaboração deste trabalho a autora recebeu apoio financeiro da FAPESP

- Ribeirão Preto, novembro de 2011 -

Método para Processamento e Análise Computacional de Imagens Histopatológicas Visando Apoiar o Diagnóstico de Câncer de Colo do Útero

Este exemplar corresponde à redação
final da dissertação de Mestrado
devidamente corrigida e defendida por
Gisele Helena Barboni Miranda
e aprovada pela comissão julgadora.

Ribeirão Preto, 24 de novembro de 2011.

Banca Examinadora:

- Prof. Dr. Joaquim Cezar Felipe (orientador) - FFCLRP-USP
- Prof. Dr. Edson Garcia Soares - FMRP-USP
- Prof. Dr. Odemir Martinez Bruno - IFSC-USP

Aos meus familiares e amigos,

Pelo apoio, preocupação e incentivo.

AGRADECIMENTOS

À Deus, por mais uma conquista.

Aos meus pais, pelo apoio constante e pelo maior incentivo que eu poderia ter recebido. Agradeço também por terem sido os primeiros patrocinadores deste projeto durante o período em que estive sem bolsa, do contrário, seria muito mais difícil começar este trabalho.

Ao Joaquim, pelas reuniões e discussões essenciais ao desenvolvimento deste trabalho, sempre me auxiliando nas dificuldades que encontramos ao longo do projeto. Pela confiança e amizade ao longo dos últimos anos e pelo apoio nos primeiros passos da pesquisa científica.

À equipe do Laboratório de Citopatologia da FMRP-USP, em especial ao Edson, pelas reuniões efetuadas ao longo de mais de dois anos, pela colaboração nas diferentes etapas do projeto, e, também, por ter se mostrado um entusiasta da proposta desenvolvida, principalmente do ponto de vista de um usuário.

Ao Junior Barrera, pelos ensinamentos, críticas e sugestões ao longo deste projeto. Pela disponibilidade com a qual sempre atendeu às minhas solicitações, abrindo mão de sábados e domingos na USP.

Ao Renato e à equipe do Laboratório de Neurologia Aplicada e Experimental da FMRP-USP pela disponibilização do equipamento necessário para aquisição das imagens microscópicas.

Aos colegas e amigos do grupo de Computação Aplicada à Informação Biomédica, Marlon, Newton, Rômulo e Danilo, com os quais pude trocar experiências sobre a utilização das ferramentas necessárias ao desenvolvimento deste projeto.

Ao Programa de Bioinfo por todo o apoio recebido no período do Mestrado. Aos amigos da primeira turma de pós-graduandos do programa em Ribeirão, cuja convivência tornou mais agradável a passagem por esta etapa através da troca de experiências e de muitos cafés na Cantina do Valter. E, claro, à Patrícia por sempre atender com muita eficiência, mesmo à distância, os nossos inúmeros pedidos.

À Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, como sede do desenvolvimento deste trabalho fornecendo a infra-estrutura necessária.

Aos membros da banca examinadora pela disponibilidade na avaliação e discussão deste projeto.

Às amigas e pós-graduandas Luana, Carla e Mariana, cuja convivência possibilitou muitas conversas e reflexões a respeito das alegrias e dificuldades da Pós-Graduação.

A CAPES pelo auxílio financeiro no período de setembro de 2009 a fevereiro de 2010.

À FAPESP pelo auxílio financeiro no período de março de 2010 a agosto de 2011.

RESUMO

A histopatologia é considerada um dos recursos diagnósticos mais importantes na prática médica e caracteriza-se pelo estudo das alterações estruturais e morfológicas das células e dos tecidos causadas por doenças. Atualmente, o principal método utilizado no diagnóstico histopatológico de imagens microscópicas, obtidas por meio de amostras em exames convencionais, é a avaliação visual do patologista, a qual se baseia na experiência do mesmo. O uso de técnicas de processamento computacional de imagens possibilita a identificação de elementos estruturais e a determinação de características inerentes, subsidiando o estudo da organização estrutural das células e de suas variações patológicas. A utilização de métodos computacionais no auxílio ao diagnóstico visa diminuir a subjetividade do processo de avaliação e classificação realizado pelo médico. Diferentes características dos tecidos podem ser mapeadas por meio de métricas específicas que poderão ser utilizadas em sistemas de reconhecimento de padrões. Dentro desta perspectiva, o objetivo geral deste trabalho inclui a proposta, a implementação e a avaliação de um método para a identificação e a análise de estruturas histológicas, a ser utilizado para a análise de lesões neoplásicas do colo do útero (NICs) a partir de amostras histopatológicas. Este trabalho foi desenvolvido em colaboração com uma equipe de patologistas, especialistas do domínio. As imagens microscópicas digitalizadas foram adquiridas a partir de lâminas previamente fixadas, contendo amostras de biópsias. Para segmentação dos núcleos celulares, foi implementado um *pipeline* de operadores morfológicos. Métodos de segmentação baseados em cor também foram testados e comparados à abordagem morfológica. Foi proposta e implementada uma abordagem baseada em camadas para representação do tecido, adotando-se a Triangulação de *Delaunay* (TD) como modelo de grafo de vizinhança. A TD apresenta algumas propriedades particulares que permitem a extração de métricas específicas. Foram utilizados algoritmos de agrupamento e morfologia de grafos, adotando-se critérios de semelhança e relações de adjacência entre os triângulos da rede, a fim de se obter a fronteira entre as camadas histológicas do tecido epitelial de forma automática. As seguintes métricas foram extraídas dos agrupamentos resultantes: grau médio, entropia e taxa de ocupação dos triângulos da rede. Finalmente, foi projetado um classificador estatístico levando-se em consideração os diferentes agrupamentos que poderiam ser obtidos a partir das imagens de treinamento. Valores de acurácia, sensitividade e especificidade foram utilizadas para avaliação dos resultados obtidos. Foi implementada validação cruzada em todos os experimentos realizados e foi utilizado um total de 116 imagens. Primeiro, foi avaliado a acurácia da metodologia proposta na determinação correta da presença de anomalia no tecido, para isto, todas as imagens que apresentavam NICs foram agrupadas em uma mesma classe. A maior taxa de acurácia obtida neste experimento foi de 88%. Em uma segunda etapa, foram realizadas avaliações entre as seguintes classes: Normal e NIC-I; NIC-I e NIC-II, e, NIC-II e NIC-III, obtendo-se taxas de acurácia máximas de 73%, 77% e 86%, respectivamente. Além disso, foi verificada também, a acurácia na discriminação entre os três tipos de NICs e regiões normais, obtendo-se acurácia de 64%. As taxas de ocupação relativas aos agrupamentos representativos das camadas basais e superficiais, foram os atributos que levaram às maiores taxas de acurácia. Os resultados obtidos permitem verificar a adequação do método proposto na representação e análise do processo de evolução das NICs no tecido epitelial do colo uterino.

Palavras-chave: Processamento de Imagens Médicas, Diagnóstico Auxiliado por Computador, Grafos de Vizinhança, Neoplasia Intraepitelial Cervical (NIC).

ABSTRACT

Histopathology is considered one of the most important diagnostic tools in medical routine and is characterized by the study of structural and morphological changes of the cells in biological tissues caused by diseases. Currently, the visual assessment of the pathologist is the main method used in the histopathological diagnosis of microscopic images obtained from biopsy samples. This diagnosis is usually based on the experience of the pathologist. The use of computational techniques in the processing of these images allows the identification of structural elements and the determination of inherent characteristics, supporting the study of the structural organization of tissues and their pathological changes. Also, the use of computational methods to improve diagnosis aims to reduce the subjectivity of the evaluation made by the physician. Besides, different tissue characteristics can be mapped through specific metrics that can be used in pattern recognition systems. Within this perspective, the overall objective of this work includes the proposal, the implementation and the evaluation of a methodology for the identification and analysis of histological structures. This methodology includes the specification of a method for the analysis of cervical intraepithelial neoplasias (CINs) from histopathological samples. This work was developed in collaboration with a team of pathologists. Microscopic images were acquired from blades previously stained, containing samples of biopsy examinations. For the segmentation of cell nuclei, a pipeline of morphological operators were implemented. Segmentation techniques based on color were also tested and compared to the morphological approach. For the representation of the tissue architecture an approach based on the tissue layers was proposed and implemented adopting the Delaunay Triangulation (DT) as neighborhood graph. The DT has some special properties that allow the extraction of specific metrics. Clustering algorithms and graph morphology were used in order to automatically obtain the boundary between the histological layers of the epithelial tissue. For this purpose, similarity criteria and adjacency relations between the triangles of the network were explored. The following metrics were extracted from the resulting clusters: mean degree, entropy and the occupation rate of the clusters. Finally, a statistical classifier was designed taking into account the different combinations of clusters that could be obtained from the training process. Values of accuracy, sensitivity and specificity were used to evaluate the results. All the experiments were taken in a cross-validation process (5-fold) and a total of 116 images were used. First, it was evaluated the accuracy in determining the correct presence of abnormalities in the tissue. For this, all images presenting CINs were grouped in the same class. The highest accuracy rate obtained for this evaluation was 88%. In a second step, the discrimination between the following classes were analyzed: Normal/CIN 1; CIN 1/CIN 2, and, CIN 2/CIN 3, which represents the histological grading of the CINs. In a similar way, the highest accuracy rates obtained were 73%, 77% and 86%, respectively. In addition, it was also calculated the accuracy rate in discriminating between the four classes analyzed in this work: the three types of CINs and the normal region. In this last case, it was obtained a rate of 64%.The occupation rate for the basal and superficial layers were the attributes that led to the highest accuracy rates. The results obtained shows the adequacy of the proposed method in the representation and classification of the CINs evolution in the cervical epithelial tissue.

Key-words: *Medical Image Processing, Computer-Aided Diagnosis, Neighborhood Graphs, Cervical Intraepithelial Neoplasia (CIN).*

SUMÁRIO

LISTA DE FIGURAS **V**

LISTA DE TABELAS **VIII**

LISTA DE QUADROS **IX**

LISTA DE SIGLAS **X**

1. INTRODUÇÃO **1**

1.1. CONSIDERAÇÕES INICIAIS **1**
1.2. CONTEXTO DE APLICAÇÃO DA PROPOSTA **2**
1.3. MOTIVAÇÃO E OBJETIVOS **6**
1.4. ORGANIZAÇÃO DA DISSERTAÇÃO **9**

2. FUNDAMENTOS TEÓRICOS **10**

2.1. RECONHECIMENTO DE PADRÕES E O PROCESSAMENTO DIGITAL DE IMAGENS **10**
2.1.1. PRÉ-PROCESSAMENTO 11
2.1.2. SEGMENTAÇÃO 12
2.1.3. EXTRAÇÃO E SELEÇÃO DE ATRIBUTOS 14
2.1.4. CLASSIFICAÇÃO 14
2.1.5. INDEXAÇÃO E RECUPERAÇÃO 15
2.2. MORFOLOGIA MATEMÁTICA **15**
2.2.1. OPERADORES ELEMENTARES 16
2.2.2. OPERADORES CONEXOS 17
2.2.3. RECONSTRUÇÃO 17
2.2.4. SEGMENTAÇÃO POR *WATERSHEDS* 18
2.3. GRAFOS DE VIZINHANÇA **19**
2.3.1. TRIANGULAÇÃO DE *DELAUNAY* (DT) 21
2.3.2. GABRIEL GRAPH (GG) 21
2.3.3. REDE DE VIZINHANÇA RELATIVA 22
2.4. ANÁLISE BASEADA NA TEORIA DE REDES COMPLEXAS **24**

3. TRABALHOS CORRELATOS **26**

3.1. CONSIDERAÇÕES INICIAIS **26**
3.2. DESCRIÇÃO DOS TRABALHOS CORRELATOS **26**
3.3. CONSIDERAÇÕES FINAIS **33**

4. MÉTODO PARA ANÁLISE ESTRUTURAL DE TECIDOS **35**

4.1. CONSIDERAÇÕES INICIAIS **35**
4.2. MAPEAMENTO DAS RELAÇÕES DE VIZINHANÇA ENTRE NÚCLEOS CELULARES **35**

4.3. ANÁLISE BASEADA EM REGIÕES 36
4.3.1. GERAÇÃO DE AGRUPAMENTOS 36
4.3.2. ROTULAÇÃO DE AGRUPAMENTOS 40
4.4. REPRESENTAÇÃO DE AGRUPAMENTOS: GERAÇÃO DO VETOR DE CARACTERÍSTICAS 42
4.5. CLASSIFICAÇÃO DA IMAGEM 42
4.6. CONSIDERAÇÕES FINAIS 43

5. MÉTODO PARA ANÁLISE AUTOMÁTICA DE IMAGENS DE COLO UTERINO 44

5.1. CONSIDERAÇÕES INICIAIS 44
5.2. LEVANTAMENTO DE PARÂMETROS PERCEPTUAIS 45
5.3. SEGMENTAÇÃO 45
5.4. ANÁLISE ESTRUTURAL 47
5.4.1. IDENTIFICAÇÃO DAS CAMADAS DO TECIDO EPITELIAL 47
5.4.2. GERAÇÃO DO VETOR DE CARACTERÍSTICAS 49
5.4.3. PROJETO DE UM CLASSIFICADOR 49
5.5. AVALIAÇÃO DOS RESULTADOS 51
5.6. CONSIDERAÇÕES FINAIS 53

6. RESULTADOS E DISCUSSÃO 54

6.1. CONSIDERAÇÕES INICIAIS 54
6.2. AQUISIÇÃO DAS IMAGENS HISTOPATOLÓGICAS 54
6.3. BASE DE IMAGENS 55
6.4. AMBIENTE DE DESENVOLVIMENTO 56
6.5. SEGMENTAÇÃO DOS NÚCLEOS 56
6.5.1. APLICAÇÃO DO *PIPELINE* DE OPERADORES MORFOLÓGICOS 57
6.5.2. TÉCNICAS BASEADAS EM COR 57
6.5.3. AVALIAÇÃO DA SEGMENTAÇÃO AUTOMÁTICA 64
6.6. AVALIAÇÃO DA REPRESENTAÇÃO DOS TECIDOS POR MEIO DE MÉTRICAS GLOBAIS 64
6.6.1. AVALIAÇÃO DOS DIFERENTES MODELOS DE GRAFOS DE VIZINHANÇA 65
6.6.2. AVALIAÇÃO DA CLASSIFICAÇÃO ENTRE NICS 66
6.7. AVALIAÇÃO DA ABORDAGEM BASEADA EM REGIÕES 67
6.7.1. IDENTIFICAÇÃO AUTOMÁTICA DAS CAMADAS DO TECIDO EPITELIAL 67
6.7.2. ANÁLISE DO PARÂMETRO δS POR MEIO DA TO 73
6.7.3. CLASSIFICAÇÃO DAS NICS 75
6.8. AVALIAÇÃO DA COMPARAÇÃO ENTRE NICS 79
6.9. CONSIDERAÇÕES FINAIS 82

7. CONCLUSÃO 83

REFERÊNCIAS BIBLIOGRÁFICAS 87

APÊNDICE A - MATRIZES DE CONFUSÃO 93

ANEXO A - APROVAÇÃO DO COMITÊ DE ÉTICA EM PESQUISA 98

LISTA DE FIGURAS

Figura 1.1. Tecido epitelial de colo uterino obtido a partir de um exame de biópsia, com destaque para os tipos celulares encontrados em cada camada desse tecido [Araújo-1999] 4

Figura 1.2. Camadas do tecido epitelial e suas variações estruturais e morfológicas de acordo com o tipo de lesão encontrada [Stevens-1997] 4

Figura 1.3. Evolução das lesões do tecido epitelial do colo uteriono. (a) Tecido epitelial normal. (b) Lesão Intraepitelial de baixo grau (LSIL). (c) Lesão intraepitelial cervical de alto grau (HSIL) 6

Figura 2.1. Simulação do transformada watershed. (a) Sinal de entrada. (b) Regiões de mínimo da imagem e pontos iniciais do algoritmo. (c) Simulação da "inundação". (d) Processamento final da *watershed*: três linhas divisórias e quatro "bacias" [Dougherty-2003] 19

Figura 2.2. (a) Critério para um triângulo pertencer a Triangulação de Delaunay (*TD*). (b) Triangulação de *Delaunay* (linhas cheias) *vs* Diagrama de *Voronoi* (linhas pontilhadas) 21

Figura 2.3. Critério para definição de arestas no modelo *Gabriel Graph* 22

Figura 2.4. Critério para definição de arestas no modelo Rede de Vizinhança Relativa 22

Figura 2.5. Exemplos de modelos de geração de redes. (a) Triangulação de *Delaunay*. (b) *Gabriel Graph*. (c) Rede de Vizinhança Relativa. [Rosenberg-2011] 23

Figura 4.1. Representação do tecido histológico pela Triangulação de *Delaunay* 36

Figura 4.2. Relações de adjacência. O triângulo de referência é representado pela cor amarela e os seus vizinhos pela cor verde. (a) Triângulos com arestas em comum. (b) Triângulos com pelo menos um vértice em comum 37

Figura 4.3. O triângulo exemplificado, tal que $\boldsymbol{x1} \leq \boldsymbol{x2} \leq \boldsymbol{x3}$, passará a ser representado pelo vetor $\boldsymbol{X} = [\boldsymbol{x1}, \boldsymbol{x2}, \boldsymbol{x3}]$ mantendo a relação de ordem entre as medidas dos lados . 38

Figura 4.4. Processo de geração de agrupamentos. 42

Figura 5.1. Processo de análise automática das imagens de tecido epitelial do colo uterino. Os processos em branco (sem preenchimento) não foram executados neste projeto 44

Figura 5.2. *Pipeline* de operadores morfológicos utilizado na segmentação dos núcleos celulares. 46

Figura 5.3. As linhas pretas representam a separação manual em camadas do tecido realizada pelo patologista 48

Figura 5.4. Classificador implementado e suas respectivas partições 51

Figura 6.1. Diagrama Entidade Relacionamento da base de imagens criada. 56

Figura 6.2. Etapas do processo de segmentação para uma imagem sem lesão: (a) Imagem de referência. (b) Imagem de referência em níveis de cinza. (c) Resultado da aplicação do operador *close-by-reconstruction top-hat*. (d) Imagem limiarizada após a eliminação de ruídos e da aplicação do operador *area open*. (e) União dos marcadores internos e externos. (f) Aplicação da *watershed*: contornos dos núcleos celulares 59

Figura 6.3. Etapas do processo de segmentação para uma imagem que apresenta NIC-I: (a) Imagem de referência. (b) Imagem de referência em níveis de cinza. (c) Resultado da aplicação do operador *close-by-reconstruction top-hat*. (d) Imagem limiarizada após a eliminação de ruídos e da aplicação do operador *area open*. (e) União dos marcadores internos e externos. (f) Aplicação da *watershed*:contornos dos núcleos celulares..... 60

Figura 6.4. Etapas do processo de segmentação para uma imagem que apresenta NIC-II: (a) Imagem de referência. (b) Imagem de referência em níveis de cinza. (c) Resultado da aplicação do operador *close-by-reconstruction top-hat*. (d) Imagem limiarizada após a eliminação de ruídos e da aplicação do operador *area open*. (e) União dos marcadores internos e externos. (f) Aplicação da *watershed*: contornos dos núcleos celulares 61

Figura 6.5. Etapas do processo de segmentação para uma imagem que apresenta NIC-III: (a) Imagem de referência. (b) Imagem de referência em níveis de cinza. (c) Resultado da aplicação do operador *close-by-reconstruction top-hat*. (d) Imagem limiarizada após a eliminação de ruídos e da aplicação do operador *area open*. (e) União dos marcadores internos e externos. (f) Aplicação da *watershed*: contornos dos núcleos celulares 62

Figura 6.6. Resultados do processo de segmentação utilizando-se técnicas baseadas em cor. (a) Imagem original. (b) Imagem após a aplicação do algoritmo de subtração de fundo (c) e (d) Imagens relativas à hematoxilina e à eosina geradas pela técnica de *Color Deconvolution*. (e) e (f) Imagens relativas à hematoxilina e à eosina geradas pela técnica *k-means* 63

Figura 6.7. Comparação da segmentação manual (contornos pretos) com a segmentação automática (contornos vermelhos)..... 64

Figura 6.8. Redes geradas para diferentes modelos de grafos de vizinhança a partir dos núcleos segmentados. (a1), (a2) e (a3): Região Normal. (b1), (b2) e (b3): NIC-I. (c1), (c2) e (c3): NIC-II. (d1), (d2) e (d3): NIC-III. Essas redes foram obtidas a partir das seguintes imagens de referência: 6.2a, 6.3a, 6.4a e 6.5a, respectivamente 66

Figura 6.9. Comparação entre diferentes relações de adjacência. (a) Agrupamentos gerados a partir da RA definida na Figura 4.2 (a), levando-se em consideração arestas em comum. Agrupamentos gerados a partir da RA definida da Figura 4.2 (b), levando-se em consideração vértices em comum 68

Figura 6.10. Resultado da aplicação do algoritmo de agrupamento descrito no Quadro 1 utilizando a proporção relativa entre áreas como critério de agrupamento. (a) $\boldsymbol{\delta A}$ = 0,15. (b) $\boldsymbol{\delta A}$ = 0,30. (c) $\boldsymbol{\delta A}$ = 0,50. (d) $\boldsymbol{\delta A}$ = 0,60 69

Figura 6.11. Resultado da aplicação do algoritmo de agrupamento descrito no Quadro 1 utilizando a semelhança entre triângulos como critério de agrupamento. (a) $\boldsymbol{\delta S}$ = 0,05. (b) $\boldsymbol{\delta S}$ = 0,10. (c) $\boldsymbol{\delta S}$ = 0,20. (d) $\boldsymbol{\delta S}$ = 0,30 70

Figura 6.12. Análise dos agrupamentos resultantes obtidos sem a relação de adjacência e considerando-se o critério de semelhança entre triângulos ($\boldsymbol{\delta S}$) igual a 0,15. 71

Figura 6.13. Identificação das camadas do tecido epitelial. (a) Agrupamentos obtidos com a abordagem não supervisionada. (b) Agrupamentos obtidos com a abordagem supervisionada 72

Figura 6.14. Agrupamentos obtidos para uma imagem (a) Normal, (b) NIC-I, (c) NIC-II e (d) NIC-III. Os agrupamentos amarelos foram classificados como basais, os verdes como intermediários e os azuis como superficiais 73

Figura 6.15. Análise das taxas de sensitividade e especificidade obtidas comparando-se imagens normais e imagens que apresentam lesão 80

Figura 6.16. Análise das taxas de sensitividade e especificidade obtidas comparando-se imagens normais e imagens que apresentam NIC-I 81

Figura 6.17. Análise das taxas de sensitividade e especificidade obtidas comparando-se imagens que apresentam NIC-I e imagens que apresentam NIC-II 81

Figura 6.18. Análise das taxas de sensitividade e especificidade obtidas comparando-se imagens que apresentam NIC-II e imagens que apresentam NIC-III 81

LISTA DE TABELAS

Tabela 1. Comparação entre os diferentes sistemas de classificação para lesões do colo do útero 5
Tabela 2. Comparação entre diferentes modelos de redes..... 65
Tabela 3.Taxas de acurácia obtidas na classificação entre NICs 67
Tabela 4. Avaliação da acurácia na diferenciação entre NICs e imagens normais para diferentes valores de $\boldsymbol{\delta S}$ utilizando-se a abordagem não supervisionada na rotulação das camadas 74
Tabela 5. Avaliação da acurácia na diferenciação entre NICs e imagens normais para diferentes valores de $\boldsymbol{\delta S}$ utilizando-se a abordagem supervisionada na rotulação das camadas 74
Tabela 6. Avaliação da acurácia na detecção da presença de anomalias 76
Tabela 7. Avaliação da acurácia na diferenciação entre as classes Normal e NIC-I..... 77
Tabela 8. Avaliação da acurácia na diferenciação entre as classes NIC-I e NIC-II 78
Tabela 9. Avaliação da acurácia na diferenciação entre as classes NIC-II e NIC-III 78
Tabela 10. Valores de MCC referentes às comparações entre diferentes classes realizadas na seção anterior para diferentes vetores de atributos. 80

LISTA DE QUADROS

Quadro 1. Detecção de agrupamentos 39
Quadro 2. Algoritmo para filtragem de agrupamentos por área de ocupação 41

LISTA DE SIGLAS

CAD	*Computer Aided Diagnosis* (Diagnóstico Auxiliado por Computador)
CA	Critério de Agrupamento
DER	Diagrama Entidade Relacionamento
DV	Diagrama de *Voronoi*
GG	Gabriel Graph
HC-FMRP	Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto
H&E	Hematoxilina & Eosina
HPV	*Human Papilloma Virus* (Vírus do Papiloma Humano)
HSIL	*High-grade Squamous Intraepithelial Lesion* (Lesão Intraepitelial Escamosa de Alto grau)
INCA	Instituto Nacional de Câncer
k-NN	*k-Nearest Neighbors* (*k*-vizinhos mais próximos)
LSIL	*Low-grade Squamous Intraepithelial Lesion* (Lesão Intraepitelial Escamosa de Baixo grau)
MCC	*Matthews Correlation Coefficient* (Coeficiente de Correlação de Matthews)
MM	Morfologia Matemática
NIC	Neoplasia Intraepitelial Cervical
OPAS	Organização Pan-americana de Saúde
RA	Relação de Adjacência
RP	Reconhecimento de Padrões
ROC	*Receiver Operating Characteristic* (curvas ROC)
RVR	Rede de Vizinhança Relativa
SERPAT	Serviço de Patologia do HC-FMRP
SIL	*Squamous Intraepithelial Lesion* (Lesão Intraepitelial Cervical)
SKIZ	*Skeleton of Influence Zone* (Esqueleto por Zona de Influência)
TD	Triangulação de *Delaunay*

1.INTRODUÇÃO

1.1. Considerações Iniciais

Nas últimas décadas, inúmeras técnicas de processamento digital de imagens têm sido desenvolvidas para auxiliar o estudo de organismos em escalas que variam desde as dimensões milimétricas até as dimensões de estruturas moleculares [Tsien-2003]. O aperfeiçoamento e o surgimento de diferentes modalidades de microscopia estão relacionados a esse avanço. Novas tecnologias de aquisição de imagens aliadas a experimentos que utilizam técnicas de aquisição de dados em larga escala possibilitam a criação de grandes bases que contêm desde imagens bi e tridimensionais até sinais biológicos representativos dos níveis de expressão de RNA de diferentes genes [Lein-2007]. De acordo com Peng, uma nova área denominada Informática de Bioimagens (*Bioimage Informatics*) pode ser destacada neste contexto. Essa área consiste no desenvolvimento e na aplicação de métodos de processamento de imagens, mineração de dados, bancos de dados e visualização, voltados à extração, comparação, recuperação e gerenciamento do conhecimento biomédico. A Informática de Bioimagens pode ser caracterizada como uma subárea da Bioinformática e suas aplicações podem levar à melhor compreensão do processo dinâmico das células e organismos vivos, pois agregam às áreas do conhecimento computacional técnicas utilizadas em outras áreas [Peng-2008].

Exemplos de aplicações da Informática de Bioimagens incluem: o rastreio de fenótipos celulares para identificação de padrões resultantes da presença de anomalias, e também para a determinação de funções de genes; a construção de atlas para organismos-modelo com o objetivo de integrar o conhecimento anatômico e ontológico em bases de dados [Dorr-2008]; o estudo e a visualização de processos dinâmicos das células em organismos vivos [Altinok-2007]; a reconstrução tridimensional de estruturas neuronais e a junção das técnicas da Informática de Bioimagens com as técnicas tradicionais de Bioinformática. Adicionalmente, trabalhos relacionados mostram que técnicas de processamento de imagens e sinais, já amplamente utilizadas, podem ser adequadas a esta área de aplicação, assim como novas abordagens têm sido desenvolvidas para a resolução de diferentes problemas biológicos que envolvem a detecção de padrões em imagens.

No auxílio ao diagnóstico de câncer e no mapeamento de sua evolução, diferentes metodologias têm sido utilizadas para a identificação de lesões precursoras, que são

encontradas nas fases iniciais da doença. Atualmente, sabe-se que a cura de muitos tipos de câncer está relacionada à sua identificação precoce e ao tratamento adequado, de acordo com o nível de malignidade da lesão. Patologistas realizam a avaliação dessas lesões por meio da observação de lâminas contendo material cito/histopatológico. Geralmente, essa avaliação é baseada em classificações internacionais responsáveis pela padronização do diagnóstico e das condições de realização dos exames. No entanto, esse processo ainda é bastante subjetivo e apresenta grande variabilidade, dado que o diagnóstico final é proveniente da experiência pessoal do patologista [Ismail-1989] [McCluggage-1998].

Por outro lado, o diagnóstico automático é baseado em medidas quantitativas extraídas de atributos intrínsecos a imagens representativas das lâminas. De acordo com Demir & Yener, o avanço das pesquisas relacionadas a esse processo, ao longo das últimas décadas, deve-se às perspectivas de utilização dos sistemas de suporte à decisão em larga escala como parte de tratamentos avançados de câncer. Além disso, trata-se de uma área com muitos desafios a serem superados [Demir-2006]. Os chamados sistemas CAD (*Computer-Aided Diagnosis*) podem tornar a rotina de trabalho do especialista mais dinâmica, por meio da filtragem de imagens obviamente benignas, permitindo que ele se concentre na análise dos casos mais difíceis ou que possam gerar mais dúvidas quanto ao diagnóstico.

Além da definição do diagnóstico ou estágio de evolução da lesão, a análise quantitativa fornecida pela abordagem computacional das imagens também pode auxiliar a compreensão dos processos subjacentes ao diagnóstico gerado. Por exemplo, a textura específica da cromatina, apresentada por núcleos de células cancerosas, pode indicar certas anomalias genéticas [Gurcan-2009]. Dessa forma, além da importância clínica, a análise quantitativa automatizada da imagem fornece subsídios para o estudo dos mecanismos biológicos relacionados ao processo de geração/formação de anomalias.

1.2. Contexto de aplicação da proposta

A histopatologia é considerada um dos recursos diagnósticos mais importantes na prática médica e caracteriza-se pelo estudo das alterações estruturais e funcionais das células e dos tecidos causadas por doenças. As amostras histopatológicas são obtidas por meio de exames de biópsias e fornecem uma visão global da arquitetura do tecido assim como os efeitos das alterações celulares sobre ela. O diagnóstico obtido por meio dessas amostras é considerado o padrão-ouro para quase todos os tipos de câncer. Já as amostras citopatológicas

são provenientes de exames mais simples e fornecem a análise isolada de células ou de agrupamentos de células. Por este motivo, as imagens citopatólogicas apresentam estruturas menos complexas de serem analisadas. A segmentação de células individuais, por exemplo, é um processo relativamente mais fácil [Gurcan-2009]. No entanto, esse tipo de exame é muito importante na prevenção e acompanhamento de muitas doenças.

O processamento de imagens citopatológicas foi bastante explorado ao longo das últimas décadas e as ferramentas utilizadas na análise de imagens histopatológicas são provenientes desses estudos. O rastreamento das características nucleares associadas à presença de câncer foi realizado *a priori* em núcleos isolados de imagens e, posteriormente, adaptado para o estudo das amostras histológicas e caracterização de seus componentes [Gurcan-2009].

De maneira geral, a manipulação automática dessas imagens passa por três etapas: pré-processamento, extração de atributos e geração do diagnóstico. A primeira consiste na filtragem e eliminação de ruídos provenientes do processo de captura da imagem e/ou do processo de fixação das lâminas, no realce da imagem e na segmentação dos objetos de interesse. A segunda etapa consiste na extração de atributos que representem as alterações observadas nas imagens histológicas. Essas alterações podem ser divididas em dois grupos: as alterações em nível celular e as alterações na distribuição das células ao longo do tecido. No primeiro grupo, os atributos a serem caracterizados são representados pelas propriedades individuais das células, tais como, morfologia, textura, intensidade de brilho, entre outros. No segundo grupo, estão os atributos que representam a dependência espacial entre as células, e, entre as mesmas e outros componentes do tecido, fornecendo informações sobre a topologia do mesmo. Por fim, a etapa de geração do diagnóstico provê informações acerca do grau de malignidade de uma lesão a partir dos atributos extraídos e também a classificação da lesão em benigna ou maligna.

Neste trabalho serão utilizadas imagens histopatológicas de tecido epitelial do colo uterino como fonte de dados para o estudo da organização estrutural celular. A Figura 1.1 apresenta um corte desse tecido, obtido a partir de exame de biópsia do segmento inferior do útero, assim como os tipos de células encontradas nesse epitélio. A camada profunda apresenta células basais e parabasais, cujas características são núcleos grandes e área citoplasmática pequena, apresentando formato arredondado. As células da camada intermediária apresentam formato poligonal com vacúolos e glicogênio. E, por fim, a camada

superficial, que contém células descamativas, normalmente achatadas e ausentes de vacúolos [Araújo-1999].

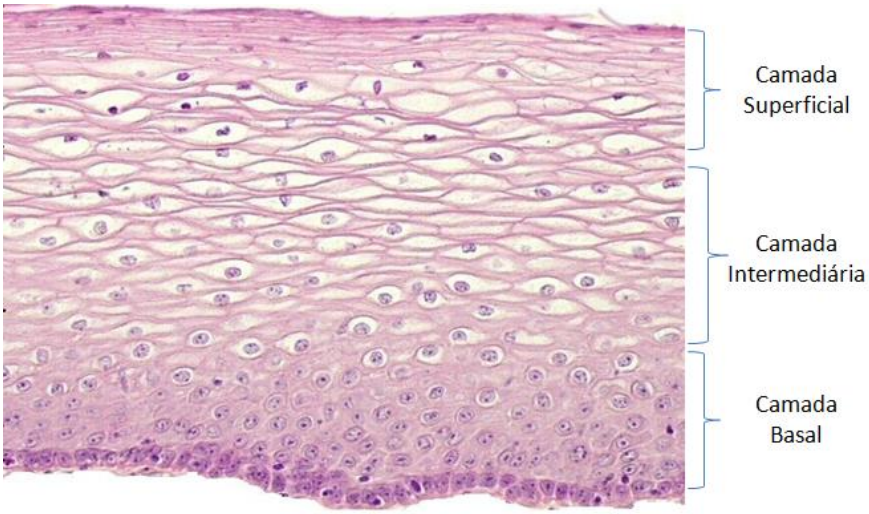


Figura 1.1. Tecido epitelial de colo uterino obtido a partir de um exame de biópsia, com destaque para os tipos celulares encontrados em cada camada desse tecido [IARC-2011]

As chamadas neoplasias intra-epiteliais cervicais (NICs) precedem o carcinoma das células escamosas do colo uterino, e, quando não tratadas podem evoluir para um carcinoma invasor. As NICs são lesões proliferativas que levam à maturação celular irregular nos tecidos. Essas lesões podem provocar a substituição parcial ou total do epitélio escamoso cervical e representam o principal fator na gênese do câncer de colo uterino. As NICs podem ser divididas em: displasia leve (NIC I), displasia moderada (NIC II) e displasia acentuada (NIC III) [Araújo-1999]. A Figura 1.2 apresenta uma representação esquemática das alterações que ocorrem a nível celular e estrutural no tecido epitelial do colo uterino.

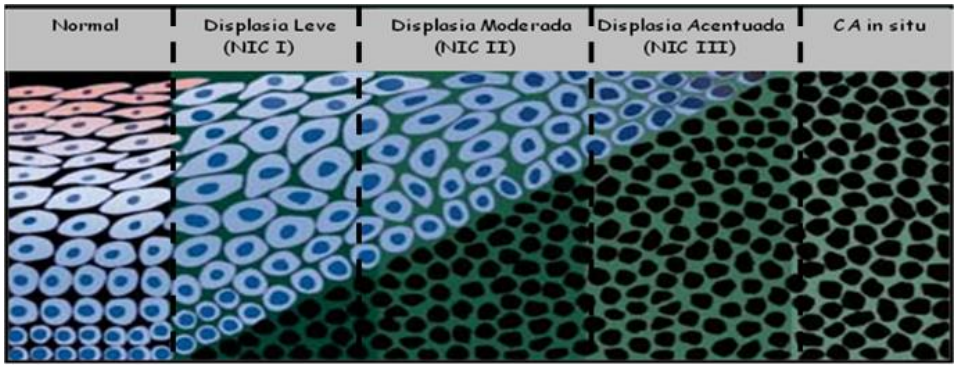


Figura 1.2. Camadas do tecido epitelial e suas variações estruturais de acordo com o tipo de lesão encontrada [Stevens-1997]

O carcinoma *in situ* é caracterizado pela maturação celular anormal em todo o tecido, porém ainda não é uma lesão invasiva. A partir da criação do Sistema de Bethesda, em 1988, predominantemente utilizado na América do Norte, foi criado o termo lesão intraepitelial escamosa (*Squamous Intraepithelial Lesion* - SIL), o qual pode ser dividido em baixo grau (*Low-grade Squamous Intraepithelial Lesion* - LSIL) e alto grau (*High-grade Squamous Intraepithelial Lesion* - HSIL) [Solomon-2002]. O Bethesda agrupa o NIC II, o NIC III e o carcinoma *in situ* na mesma classe (HSIL) devido ao fato de estes levarem a condutas diagnósticas semelhantes. A Tabela 1 correlaciona alguns sistemas de classificação, com destaque para as displasias.

Tabela 1. Comparação entre dois diferentes sistemas de classificação para lesões do colo do útero

Classes Papanicolaou	***Descrição***	***Gradação NIC***	***Sistema Bethesda***
I	Normal	Normal	Normal
II	Atipia Reativa/Inflamatória	Atipia	Normal
II / III	Atipia Suspeita	Atipia	ASCUS
II / III	Atipia com HPV	Atp.condiloma	SIL de baixo grau (LSIL)
III	**Displasiva Leve**	**NIC-I**	**SIL de baixo grau (LSIL)**
III	**Displasia Moderada**	**NIC-II**	**SIL de alto grau (HSIL)**
III	**Displasia Acentuada**	**NIC-III**	**SIL de alto grau (HSIL)**
IV	Carcinoma *in situ*	NIC-III	SIL de alto grau (HSIL)
V	Câncer Invasivo	Câncer Invasivo	Câncer Invasivo

É interessante notar que, do ponto de vista patológico, as lesões cervicais precursoras apresentam-se em graus evolutivos que variam desde displasias leves até o câncer invasivo, por meio de um processo gradual. Há também as classificações de lesões atípicas inflamatórias que não estão envolvidas em processos cancerígenos e não estão no escopo deste projeto.

A proporção entre o volume nuclear e o volume do citoplasma (representados pelas respectivas áreas nas imagens bidimensionais) é um dos parâmetros perceptuais utilizados pelo patologista na definição do diagnóstico. Porém, a porção do tecido epitelial atingida pela lesão destaca-se com alta relevância nessa análise: quando o terço inferior do tecido foi afetado o diagnóstico é caracterizado como lesão de baixo grau (LSIL); ou se de dois a três terços foram atingidos, configura-se uma lesão de alto grau (HSIL). A Figura 1.3 apresenta a evolução dessas lesões em amostras obtidas em exames de biópsias.

O presente trabalho propõe métodos para avaliação automática da porção do tecido afetada, utilizando esta avaliação como principal parâmetro na determinação da presença e do grau das NICs.

1.3. Motivação e Objetivos

De acordo com estimativas do Instituto Nacional de Câncer (INCA), o câncer do colo do útero é o segundo tipo mais comum entre as mulheres, sendo responsável pelo óbito de aproximadamente 230 mil mulheres por ano no mundo. Sua incidência é cerca de duas vezes maior em países em desenvolvimento quando comparada à incidência nos países desenvolvidos. A sobrevida média mundial estimada é de 49% dos casos após cinco anos de prevalência da doença. O câncer do colo do útero é o câncer que apresenta maior potencial de cura e prevenção quando diagnosticado precocemente. Uma redução de 80% na mortalidade pode ser alcançada por meio do rastreamento de mulheres na faixa etária de 25 a 65 anos [INCA-2008].

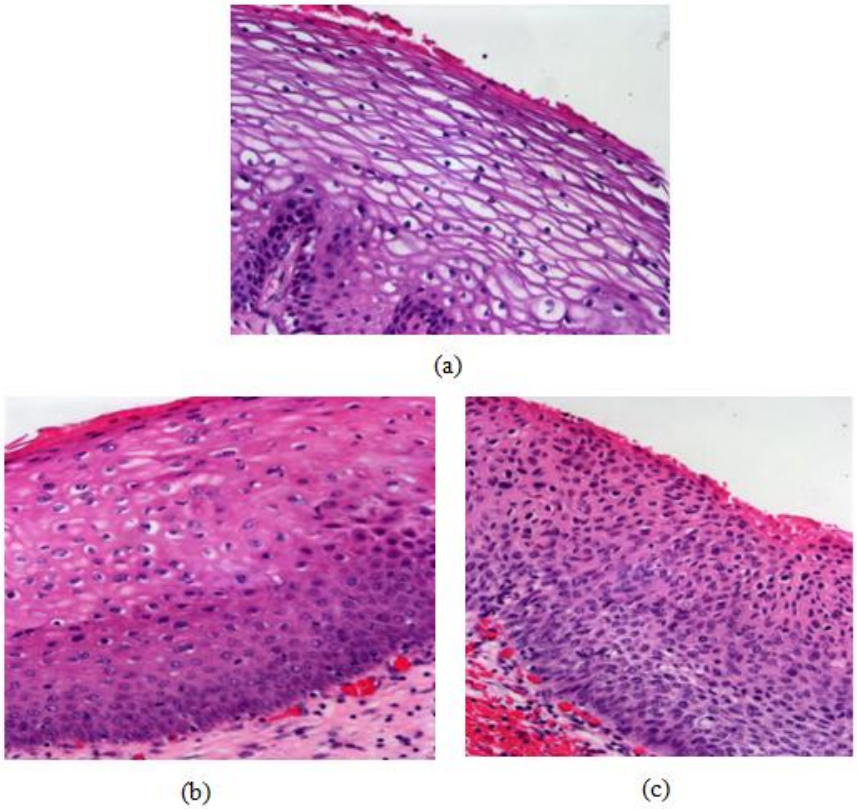


Figura 1.3. Evolução das lesões do tecido epitelial do colo uteriono. (a) Tecido epitelial normal. (b) Lesão Intraepitelial de baixo grau (LSIL). (c) Lesão intraepitelial cervical de alto grau (HSIL)

O exame de Papanicolaou é considerado pela Organização Pan-Americana da Saúde (OPAS) o mais adequado para o diagnóstico precoce de lesões precursoras do câncer cervical. Porém, verifica-se que a incidência desse tipo de câncer em mulheres jovens, triadas adequadamente, pela citologia oncológica é elevada, devido ao grande número de resultados falso-negativos daquele método [Lonky-1999] [Nanda-2000]. A biópsia do colo uterino

também é usada frequentemente em casos nos quais a presença de anomalias foi identificada em exames anteriores. Desta forma, a análise histológica tem por objetivo confirmar o diagnóstico obtido na análise citológica, assim como auxiliar o rastreamento de mulheres nas faixas etárias de risco.

Atualmente, a principal ferramenta utilizada no diagnóstico de imagens microscópicas, obtidas através de amostras em exames convencionais, é a avaliação visual do patologista, a qual está sujeita à sua interpretação pessoal. A utilização de métodos computacionais no auxílio ao diagnóstico visa diminuir a subjetividade e aumentar a precisão dos processos de avaliação e classificação realizados pelo médico. Diversos estudos realizados na área de radiologia apresentaram diferentes ferramentas computacionais de auxílio ao diagnóstico para a geração de segunda opinião. Esses estudos demonstraram que o uso dessas ferramentas promove um aumento na precisão dos diagnósticos, a partir da diminuição tanto de resultados falso-positivos quanto de falso-negativos [Doi-2007a] [Doi-2007b].

Para as imagens histopatológicas, a manipulação e a análise computacional permitem a identificação de características inerentes, subsidiando o estudo da organização estrutural dos tecidos e suas variações patológicas. Sistemas computacionais que forneçam a identificação de padrões anormais podem auxiliar o especialista no processo de tomada de decisão. Esses sistemas podem abarcar funcionalidades que permitam a realização automática de tarefas a partir da varredura da imagem. A eficácia dos métodos de visão computacional tornaram-se fundamentais para os sistemas de auxílio computadorizado. Em decorrência, verifica-se a importância do aperfeiçoamento de técnicas de Aprendizado de Máquina e Reconhecimento de Padrões.

Dentro deste contexto, o processo de segmentação de imagens é importante para a detecção e caracterização de estruturas de interesse. O formato dos núcleos celulares, por exemplo, é um dos principais atributos de análise do patologista.Técnicas de segmentação baseadas em Morfologia Matemática (MM) podem ser utilizadas para separação dos mesmos. A MM constitui-se como uma subárea do processamento de imagens focada no estudo da geometria da imagem e nos últimos anos tem sido amplamente utilizada em tarefas de segmentação, detecção de bordas, análise de forma, entre outras [Dougherty-2003]. A MM fornece ferramentas que possibilitam a descrição quantitativa de estruturas de interesse na imagem referentes a sua morfologia e estrutura topológica. Exemplos da utilização de operadores morfológicos no processamento de bioimagens podem ser encontrados em [Hirata-2001] e [Angulo-2003].

Além da importância da caracterização morfológica dos componentes histológicos, abordagens relacionais têm se destacado neste contexto. A utilização de modelos que descrevam as conexões estabelecidas entre componentes histológicos permite a exploração de um conjunto adicional de atributos, fornecendo suporte à análise estrutural do tecido. Desta forma, além das características morfológicas das células, a disposição espacial e a interação entre elas também poderá ser descrita, por exemplo, por meio da extração de medidas. Essas medidas podem diferenciar estruturas de interesse e serem usadas em uma abordagem de reconhecimento de padrões [Costa-2007]. Além disso, representações mais sofisticadas também permitem um estudo da dinâmica da rede por meio da análise de suas variações estruturais ao longo do tempo.

Por meio da Figura 1.2, que representa o modelo teórico das alterações que ocorrem a nível celular e histológico na presença de NICs, verifica-se a adequação das abordagens descritas acima na caracterização dos principais componentes de interesse. Dentro desta perspectiva, o objetivo geral deste trabalho inclui a proposta, a implementação e a avaliação de um método para a identificação automática de estruturas histológicas. Este método deverá ser utilizado para a análise de lesões neoplásicas do colo do útero (NICs), auxiliado por um estudo dos processos envolvidos na organização estrutural celular a partir de amostras histopatológicas. Especificamente, os seguintes objetivos podem ser citados:

- Aplicação de técnicas de processamento de imagens, obtidas a partir de amostras de biópsias do colo uterino, visando à identificação de seus principais componentes por meio da aplicação de técnicas baseadas em Morfologia Matemática;
- Proposta de um método para análise estrutural de tecidos histopatológicos visando o estudo de sua organização estrutural por meio de técnicas baseadas na teoria de Grafos de Vizinhança e Redes Complexas;
- Proposta de um método para a análise de neoplasias intraepiteliais cervicais, as NICs, a fim de se identificar de maneira automática as alterações estruturais características dessas lesões.
- Implementação de um classificador para identificação dos padrões celulares das lesões precursoras do tipo de câncer em análise neste trabalho, assim como sua evolução ao longo dos diferentes estágios.

Por fim, a proposta apresentada deverá contribuir para um aumento do grau de precisão do processo de geração de laudos e diagnósticos em exames histopatológicos, por meio da identificação automática da presença de lesões intraepiteliais no tecido e, também, da identificação de seu nível de malignidade. O estudo e a aplicação de técnicas de processamento de imagens e reconhecimento de padrões adequadas a este contexto podem auxiliar na identificação precoce dos processos cancerígenos associados a essas lesões, podendo atuar como uma segunda opinião para o patologista.

1.4. Organização da Dissertação

Esta dissertação está organizada da seguinte forma: a Seção 2 apresenta os fundamentos teóricos relacionados às áreas de Reconhecimento de Padrões, Morfologia Matemática, Grafos de Vizinhança e análise baseada em Redes Complexas. A Seção 3 apresenta os trabalho relacionados ao contexto deste projeto. A Seção 4 apresenta o método proposto para análise estrutural de imagens histopatológicas. A Seção 5 descreve o método proposto para o estudo das imagens histopatológicas provenientes de amostras do colo uterino. A Seção 6 apresenta e discute os resultados obtidos. E, finalmente, a Seção 7 apresenta as conclusões e as possibilidades de trabalhos futuros.

2. FUNDAMENTOS TEÓRICOS

2.1. Reconhecimento de Padrões e o Processamento Digital de Imagens

Ao longo das últimas décadas, diversos sistemas computacionais vêm sendo propostos visando à automatização de tarefas normalmente executadas por seres humanos, trazendo benefícios para diferentes áreas do conhecimento [Duda-2001]. A criação de sistemas que possam realizar tarefas com habilidades comparáveis à performance humana é uma das grandes motivações para a realização de estudos na área de Reconhecimento de Padrões (RP) [Reis-2001].

Aplicações de técnicas de RP podem ser encontradas em áreas como processamento de sinais e voz, visão computacional e biometria [Duda-2001]. Na área médica destacam-se os sistemas de diagnóstico clínico, análise de radiografias e tomografias, estudos genéticos, entre outros. Nos últimos anos, a utilização de diferentes ferramentas computacionais na área de bioinformática, por exemplo, tornou-se indispensável devido aos dados armazenados em grandes bases de dados geradas a partir do Projeto Genoma Humano [Liew-2005]. Devido a essa demanda computacional, diferentes técnicas de reconhecimento de padrões foram utilizadas visando auxiliar a análise de seqüências de DNA e de dados provenientes de experimentos de *microarrays*.

A construção de sistemas de reconhecimento de padrões envolve três aspectos: a aquisição de dados e o pré-processamento, a representação dos dados e a tomada de decisão. De acordo com o tipo dos objetos a serem classificados, a implementação de um sistema de RP poderá adotar diferentes abordagens, dentre as quais, as mais conhecidas são [Jain-2000]:

- Estatística: permite a criação de fronteiras de decisão *n*-dimensionais no espaço de atributos, que separam padrões pertencentes a classes diferentes. As fronteiras de decisão são determinadas utilizando-se distribuições de probabilidade referentes aos padrões de cada classe.
- *Template Matching*: consiste na comparação entre um padrão de entrada e um modelo previamente reconhecido baseada em medidas de similaridade.
- Sintática: cria uma analogia entre a estrutura de um padrão e a sintaxe de uma linguagem. Sentenças são geradas de acordo com uma gramática facilitando a

representação de padrões complexos por meio de um pequeno número de primitivas e regras gramaticais.

- Redes Neurais: compreendem modelos capazes de adquirir conhecimento através da experiência, ou seja, a partir do próprio conjunto de dados utilizado para treinamento. A capacidade de generalização, tolerância a falhas e adaptatividade são características das redes neurais que permitem que regras básicas sejam extraídas de dados reais sem a necessidade de um conjunto de regras pré-fixadas.

Dentro do contexto desse projeto, diferentes tipos de microscopia e outros dispositivos de aquisição de imagens têm auxiliado a descoberta e o aperfeiçoamento do conhecimento biológico nas últimas décadas. A aquisição automática de imagens através de sistemas integrados à rotina dos laboratórios possibilita a criação de grandes bases de dados, inviabilizando o seu processamento manual. Como consequência, o processamento automático dessas imagens tem se destacado em muitos experimentos biológicos. No caso das imagens microscópicas, por exemplo, fatores tais como a elevada densidade de células interligadas, as variações de morfologia e intensidade de um mesmo objeto em diferentes imagens, ou ainda, a presença de muitos canais de cores evidenciam a dificuldade de aplicação de modelos convencionas e a necessidade do desenvolvimento de novas técnicas de análise [Peng-2008] [Shamir-2010]. As próximas seções apresentam, de maneira geral, as principais etapas do processo automático de imagens.

2.1.1. Pré-Processamento

A aplicação de técnicas de pré-processamento tem por objetivo corrigir eventuais artefatos do processo de aquisição da imagem, ou até mesmo, destacar estruturas de interesse na mesma a fim de melhorar os resultados das etapas seguintes. O pré-processamento envolve a aplicação de técnicas de realce de contraste e remoção de ruídos. No caso das imagens histológicas devem ser considerados os ruídos provenientes dos equipamentos de aquisição e da distribuição não uniforme dos corantes sobre a amostra durante o processo de fixação. A correção de fundo, por exemplo, padroniza as imagens diminuindo os efeitos do processo de aquisição sob diferentes condições de iluminação. Esse processo é baseado na caracterização quantitativa de regiões da imagem que não contêm objetos de interesse ("vazias"). Algoritmos baseados em ajustes polinomiais, em características morfológicas de componentes da imagem

[Sternberg-1983] e em minimização de entropia [Likar-2000] podem ser utilizados para essa finalidade. A filtragem da imagem baseada em técnicas de limiarização e na aplicação de operadores mofológicos também podem ser utilizadas para eliminação de ruídos

2.1.2. Segmentação

A segmentação de imagens é utilizada para separar objetos ou regiões significativas da imagem, tais como formas, linhas e curvas, para uma identificação mais precisa dos objetos de interesse. A segmentação consiste, basicamente, na divisão da imagem em regiões distintas a partir de características comuns, como por exemplo, valores de pixels semelhantes. Dependendo do contexto de aplicação, pode ser necessária uma adequação da técnica utilizada, o que na maioria das vezes não é nada trivial. Dentre as técnicas de segmentação mais utilizadas encontram-se: métodos baseados em detecção de bordas (*boundary-based*) como a *Watershed* [Vincent-1991], utilizados para se identificar descontinuidades na imagem e, também, técnicas baseadas no crescimento de regiões por meio de contornos ativos, como a *Snake* [Kass-1987], que um é método interativo.

No contexto das imagens histológicas, os atributos morfológicas são muito importantes na classificação de componentes do tecido. Contornos nucleares irregulares, por exemplo, podem identificar alterações celulares com características de malignidade. A seção 2.2 introduz alguns conceitos da Morfologia Matemática que podem ser utilizadas na caracterização desse tipo de atributo. Além do formato, a cor é um outro atributo importante na análise desse tipo de imagem. A combinação dos corantes hematoxilina e eosina (H&E) é uma das mais utilizadas na coloração de tecidos biológicos. A hematoxilina cora estruturas ácidas, como é o caso do núcleo celular, e apresenta cor azul ou violeta. A eosina, por sua vez, cora estruturas básicas, como o citoplasma e apresenta coloração rósea ou avermelhada. Devido a este fato, muitas técnicas propostas para segmentação deste tipo de imagem utilizam as propriedades imuno-histoquímicas dos tecidos biológicos, sendo a cor o principal atributo de análise. Duas técnicas de segmentação baseadas em cor são descritas a seguir:

Color Deconvolution

A técnica desenvolvida por Ruifrok & Johnston, denominada *color deconvolution*, é um exemplo que utiliza as características imuno-histoquímicas dos tecidos [Ruifrok-2001]. Essa técnica pode ser generalizada a outros corantes e envolve transformações ortonormais entre o sistema RGB e um segundo sistema, cujas coordenadas são definidas de acordo com a

Densidade Óptica (DO) dos corantes utilizados. A DO é uma medida que quantifica a intensidade de luz incidente sob a amostra em análise e não absorvida, o que depende da concentração do corante utilizado. Ela é baseada na lei de Lambert-Beer que descreve a intensidade de luz transmitida através da amostra (1).

$$I = I_0 * 10^{-A*c*l} \qquad (1)$$

Sendo I a intensidade de luz observada, I_0 a intensidade da luz incidente, A a quantidade de corante utilizada, c o coeficiente de absorção do corante e l a distância percorrida através da amostra. A DO fornece uma relação linear entre a intensidade dos pixels da imagem e a densidade da coloração descrita na Equação 2. Para imagens RGB, a DO é calculada para cada canal.

$$DO = -\frac{1}{l} log_{10}\left(\frac{I}{I_0}\right) = A * c \qquad (2)$$

Em outras palavras, a imagem corada com H&E, passa a ser representada pela combinação linear das DOs de cada corante e da DO relativa ao fundo da imagem. Este sistema atua como um filtro, separando estruturas mais sensíveis a cada um dos corantes.

K-means clustering

A aplicação do algoritmo *k-means* tem por objetivo a representação da imagem por meio de agrupamentos (ou *clusters*) de referência, sendo amplamente utilizado como uma técnica de aprendizado não supervisionado. A partir de um conjunto de pontos em um espaço *n*-dimensional e um inteiro *k*, o algoritmo inicia-se com a escolha aleatória de *k clusters*, para os quais os respectivos centróides são calculados. Então, cada ponto é atribuído ao centróide mais próximo, e, posteriormente os centróides são recalculados para cada *cluster*. Esse processo é repetido interativamente até a estabilidade, ou seja, até que os valores dos centróides fiquem constantes [Theodoridis-2009].

Similar à técnica de *color deconvolution*, quando aplicado às imagens histológicas, o algoritmo *k-means* tem como objetivo representar a imagem por meio de cores de referência. No caso das imagens coradas com H&E, ignorando-se as variações de brilho, três cores podem ser identificadas: azul, rosa e branco, que passam a ser referência em cada um dos agrupamentos. É importante observar que a utilização do sistema de representação de cor *L*a*b** pode melhorar os resultados da segmentação por *k-means* devido a algumas vantagens desse sistema em relação a outros, como, por exemplo, a uniformidade perceptual

separando a informação de cor da luminosidade, que pode ser definida em função da componente *L**. Toda a informação de cor é representada pelas componentes *a** e *b**. Além disso, a distância euclidiana pode ser utilizada para medir a distância entre duas cores [MathWorks-2010].

2.1.3. Extração e Seleção de Atributos

Atributos de imagens podem corresponder a informações estatísticas, geométricas, de distribuição de pixels, entre outros. A extração e a seleção de atributos são requisitos para a maioria dos estudos envolvendo imagens. O extrator de características a ser utilizado depende do domínio de aplicação e normalmente está associado a descritores intrínsecos da imagem. Atributos morfológicos e relacionados a textura tem sido bastante explorados no diagnóstico automático de câncer [Gunduz-Demir-2007]. Diversas técnicas clássicas são utilizadas para a extração de características, tais como as matrizes de co-ocorrência e os descritores de *Haralick*, histogramas de cor, transformada de *Fourier*, transformada de *Wavelets* e momentos geométricos [Felipe-2005].

Técnicas de seleção de atributos objetivam remover do conjunto de dados atributos irrelevantes e/ou redundantes, diminuindo a dimensão do conjunto original. Esses atributos podem interferir negativamente na etapa de classificação. Em alguns casos, por exemplo, quanto maior o número de atributos utilizado, menor o significado (contribuição) de cada atributo separadamente, pois em muitos casos um grande número destes pode estar correlacionado, gerando informações redundantes e dificultando a extração de conhecimento [Egecioglu-2004].

2.1.4. Classificação

Um classificador tem por objetivo dividir o espaço de atributos em regiões que correspondam às classes de interesse. Muitas aplicações tais como a determinação de fenótipos celulares e a determinação da localização de proteínas estão envolvidas com técnicas de classificação e *clustering*. Através da utilização de técnicas de Aprendizado de Máquina e Inteligência Artificial é possível fornecer uma classificação automática a fim de auxiliar o especialista do domínio no processo de tomada de decisão.

2.1.5. Indexação e Recuperação

A utilização de descritores textuais para indexação em bases de dados relacionais pode ser uma escolha complicada devido à grande quantidade de padrões de nomenclatura na área, contudo vocabulários estruturados (ontologias) podem ajudar na organização do conhecimento biológico. Por outro lado, há o acesso baseado no conteúdo intrínseco das imagens, depois de realizado o processamento e a extração de atributos das mesmas [Peng-2008]. Nesse caso, faz-se uso de estruturas de indexação específicas para a manipulação de metadados, tais como os métodos de acesso espaciais e métricos [Traina-2007].

2.2. Morfologia Matemática

Caracterizada como uma subárea do Processamento de Imagens não-linear, a Morfologia Matemática (MM) tem como foco o estudo geométrico da imagem [Dougherty-2003]. A teoria de reticulados e a topologia digital formam a base teórica da MM, a qual, foi desenvolvida na década de 60 por George Matheron e Jean Serra com o objetivo inicial de experimentar uma abordagem baseada em transformações de forma para a extração de informação de imagens [Banon&Barrera-1994]. Essas transformações são caracterizadas pelas "transformações elementares", representadas pelos operadores de dilatação e erosão e dependentes do chamado elemento estruturante.

Um elemento estruturante é caracterizado por um subconjunto que define a transformação aplicada à imagem. Ele é usado para se definir a vizinhança de cada pixel, estabelecendo, dessa forma, a relação de conexidade entre eles. O tipo e as dimensões do elemento estruturante são dependentes da informação que se deseja obter da imagem [Dougherty -2003]. A translação e a reflexão de conjuntos são amplamente utilizadas em MM na implementação das transformações.

Seja B um conjunto definido em Z^2 (considerando Z o conjunto dos inteiros), A translação do conjunto B por um ponto z é descrita pela equação (3) e a transposição de B, pela equação (4):

$$B_z = \{b + z,\ para\ \ b \in B\} \qquad (3)$$

$$\hat{B} = \{-b,\ para\ \ b \in B\} \qquad (4)$$

2.2.1. Operadores Elementares

Operadores morfológicos são mapeamentos entre subconjuntos que apresentam propriedades particulares. O paradigma central da MM é a representação de qualquer conjunto de operadores por meio da concatenação dos operadores elementares através das operações básicas sobre conjuntos: composição, intersecção, união e complementação [Banon&Barrera-1993]. Os operadores elementares são a dilatação e a erosão e suas definições são apresentadas a seguir.

Sejam A e B dois conjuntos definidos em Z^2, representando, respectivamente, a imagem em análise e o elemento estruturante, as operações de erosão e dilatação apresentam as seguintes definições (5) e (6):

$$A \ominus B = \{z | B_z \subseteq A\} \quad (5)$$

$$A \oplus B = \{z | (\hat{B})_z \cap A \neq \emptyset\} \quad (6)$$

A erosão de *A* por *B* é o conjunto de todos os pontos *z*, tais que *B* transladados por *z* está contido em *A*. A dilatação de *A* por *B* é representado por todos os deslocamentos *z*, tais que, *B* transladado e refletido por *z*, sobrepõe-se em pelos menos um elemento com *A*. A dilatação é o operador dual da erosão com respeito a complementação de conjuntos. De maneira equivalente, a erosão e a dilatação podem ser definidas pelas equações (7) e (8).

$$A \ominus B = \cap_{b \in B}(A_{-b}) \quad (7)$$

$$A \oplus B = \cup_{b \in B}(A_b) \quad (8)$$

Ou seja, a erosão é representada por uma intersecção de translações efetuadas sobre a imagem, com o elemento estruturante refletido, e a dilatação, pela união de todos os pontos contidos no elemento estruturante definidos sobre a imagem transladada. Essa representação é baseada nas definições de soma e a subtração de *Minkowski* [Banon&Barrera-1994]. É interessante notar que, se a origem estiver contida no elemento estruturante, então a erosão terá um efeito de encolhimento sobre a imagem e a dilatação, de expansão [Dougherty -2003].

Como exemplo de composição, podem ser citados os operadores de abertura e fechamento. A abertura é caracterizada por uma erosão seguida de uma dilatação morfológica (9). Esse operador pode ser utilizado como um filtro para a eliminação de ruídos da imagem. Os ruídos filtrados são determinados pela dimensão do elemento estruturante. O fechamento é definido pela aplicação de uma dilatação seguida por uma erosão morfológica (10). E de

maneira similar a abertura, esse operador também pode ser utilizado para filtragem de ruídos. A abertura e o fechamento são operadores duais.

$$A \circ B = (A \ominus B) \oplus B \tag{9}$$

$$A \cdot B = (A \oplus B) \ominus B \tag{10}$$

2.2.2. Operadores Conexos

Um subconjunto de uma imagem é dito um componente conexo (C_i) se existe um caminho entre dois pontos quaisquer desse subconjunto, contido no próprio subconjunto. Para as imagens digitais, a conexidade é definida pela vizinhança do pixel e pelo elemento estruturante escolhido: dois pixels p e q são 4-conexo se q está contido no elemento estruturante E_4 (cruz), de maneira similar, p e q são 8-conexo se q está contido no elemento estruturante E_8 (quadrado) [Dougherty-2003].

Utilizando a definição sobre conjuntos, um operador ψ é dito conexo se para qualquer conjunto A, a diferença simétrica entre A e $\psi(A)$ é composta somente por componentes conexos de A ou de seu complemento A^c. O operador conexo também pode ser definido utilizando a definição de partição: uma partição P do espaço E, é definida como uma coleção de componentes conexos disjuntos $\{A_i\}$, cuja união resulta no espaço inteiro. Cada A_i é chamado classe de partição. Para cada pixel x, existe um conjunto único na partição que contém x, denotado $P(x)$. Uma partição P' é dita "mais fina" do que a partição P se, para todo x, $P'(x)$ está contido em $P(x)$, e, neste caso, P é considerada a partição "mais grossa". Desta forma, um operador ψ é conexo se para todo conjunto A, a partição associada à $\psi(A)$ é "mais grossa" do que a partição associada à A [Salembier-1995].

Um operador conexo preserva ou remove componentes conexas, não introduzindo novas componentes e, consequentemente, conservando as bordas dos objetos remanescentes. O filtro morfológico *area open* é um exemplo de um operador conexo. Esse operador é caracterizado pela filtragem de componentes cujas áreas são menores do que um limiar α.

2.2.3. Reconstrução

Transformações baseadas em reconstruções morfológicas podem ser utilizadas como filtros, eliminando regiões indesejadas. A reconstrução envolve a utilização de uma segunda imagem M, denominada "marcador", que contém os pontos iniciais da transformação. M é um subconjunto da imagem de entrada A. Essa última, é representada pelos seus componentes

conexos. A reconstrução de A por M é denominada $A\ \Delta_B\ M$ e definida pela equação (11), na qual B é o elemento estruturante e C_k é um componente conexo de A:

$$A\ \Delta_B\ M = \bigcup\{C_k : C_k \cap M \neq \emptyset\} \tag{11}$$

A reconstrução caracteriza-se como uma importante operação morfológica, apresentando muitas aplicações práticas, como por exemplo, a dilatação condicional. Como descrito na seção 2.2.1, quando uma imagem é dilatada por um elemento estruturante cuja origem está contida no mesmo, ela é expandida e essa expansão é condicionada pelo formato do elemento estruturante. A aplicação de dilatações sucessivas em uma mesma imagem leva à perda das fronteiras originais dos objetos. Essa situação pode ser contornada definindo-se condições para a aplicação da dilatação, ou seja, restringindo-se as translações [Dougherty-2003]. Considerando A como uma subimagem de C, e B como um elemento estruturante, a dilatação condicional pode ser definida como:

$$A \oplus_C B = \bigcup_{a \in A} B_a \cap C \tag{12}$$

Ou, de maneira equivalente:

$$A \oplus_C B = (A \oplus B) \cap C \tag{13}$$

A aplicação de n dilatações condicionais é denominada dilatação geodésica de tamanho n. Considerando-se M, o marcador, e A, a imagem de entrada, a reconstrução de A por M pode ser implementada aplicando-se uma sequência de dilatações geodésicas, até a convergência da transformação, de acordo com:

$$(M \oplus_A B)^n = (((M \oplus_A B) \oplus_A B) \oplus_A \ \dots \oplus_A B) \tag{14}$$

2.2.4. Segmentação por *Watersheds*

A transformada *watershed* é um dos métodos de segmentação mais conhecidos baseado no crescimento de regiões. Nesta abordagem a imagem em níveis de cinza é tratada como um relevo topográfico, no qual, os vales são caracterizados pelas regiões mais escuras da imagem e as montanhas, pelas regiões mais claras. O algoritmo "inunda os vales", agrupando pixels vizinhos com intensidades semelhantes. O algoritmo chega ao fim quando a água atinge os picos do relevo, o qual é particionado em bacias (regiões da imagem) por linhas divisoras [Vincent-1991] (Figura 2.1).

Geralmente, a *watershed* é aplicada sobre o gradiente morfológico da imagem, entretanto, por ser um operador sensível a ruídos, a imagem resultante desse operador pode apresentar muitos mínimos locais, gerando muitas "bacias" no processo final de segmentação. Esse resultado é conhecido como sobre-segmentação. A aplicação de filtros sobre a imagem do gradiente pode diminuir este problema.

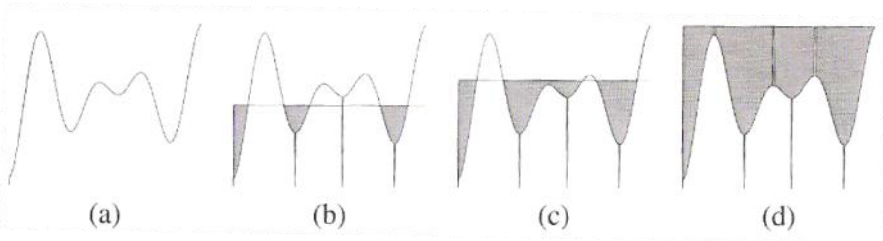


Figura 2.1. Simulação do transformada watershed. (a) Sinal de entrada. (b) Regiões de mínimo e pontos iniciais do algoritmo. (c) Simulação da "inundação". (d) Processamento final da *watershed*: três linhas divisórias e quatro "bacias" [Dougherty-2003]

Outra abordagem para se evitar a sobre-segmentação é a aplicação da *watershed* a partir de marcadores [Meyer-1990]. Os marcadores delimitam as "bacias" a serem segmentadas, ou seja, os pontos de mínimo a partir dos quais o algoritmo deverá iniciar. Ao final, cada região encontrada corresponderá a um único marcador. A implementação da *watershed* baseada em marcadores pode ser obtida por reconstrução, de maneira que os marcadores correspondam aos mínimos da imagem. Esse processamento é denominado imposição do mínimo.

2.3. Grafos de Vizinhança

Um grafo *G(V,E)* é um objeto formado por um conjunto de vértices ou nós (*V*) e um conjunto de arestas (*E*), que representam as conexões entre esses nós. A partir dessa estrutura, são derivadas algumas classificações:

- Nos **grafos direcionados ou dígrafos**, cada aresta apresenta uma origem $u \in V$ e um destino ($v \in V$), podendo ser representada pelo par ordenado (u,v). Nos **grafos não direcionados**, não há relação de ordem e a aresta é representada pelo conjunto $\{u,v\}$.
- Quando há informação de custo associado às arestas de um grafo, ele pode ser denominado **grafo ponderado** (*wighted graph*). O peso é representado por uma função w: $E \rightarrow \mathbb{R}$, que relaciona uma aresta $e \in E$ à um peso $w(e)$. Esse peso pode representar um valor de distância, capacidade, similaridade, entre outros, dependendo

do contexto de aplicação. Um grafo que não possui custo associados às suas arestas é equivalente a um grafo pesado com valor unitário para cada aresta.

- Em grafos não direcionados, o **grau** de um vértice *v* pode ser definido como o número de arestas conectadas a *v*. Em grafos direcionados, podemos determinar o **grau de saída** (*out-degree*) de um vértice (*v*), caracterizado pelo número de arestas que possuem *v* como origem, e, o **grau de entrada** (*in-degree*) de *v*, caracterizado pelo número de arestas que possuem *v* como destino.

A utilização de grafos no mapeamento das relações de vizinhança entre componentes de tecidos histológicos têm-se mostrado uma técnica bastante eficaz na caracterização de objetos de interesse em imagens [Landini-2004] [Gunduz-2004] [Zampirolli-2010]. Em [Vincent-1989] é apresentado um estudo aprofundado da utilização da modelagem por grafos aliada às ferramentas fornecidas pela Morfologia Matemática (MM). As operações clássicas de dilatação e erosão, além de diferentes filtros morfológicos, podem ser representados como operações sobre grafos. Nesse contexto, um grafo pode ser visto como um reticulado booleano completo, no qual, funções estruturantes específicas podem ser definidas.

A estrutura relacional fornecida pelos grafos, permite a extração de medidas estatísticas, coeficientes e outros parâmetros que os representem de maneira global. A geração do grafo pode levar em consideração diferentes critérios para definição das arestas. Assumindo-se uma imagem binária cujos componentes conexos são os objetos de interesse, temos um conjunto V de vértices, representado por esses elementos e um conjunto E de arestas representando as relações de vizinhança entre os mesmos. Um método bastante simples para a definição de nós vizinhos é o *k*-NN (*k-Nearest Neighbors*), no qual, cada $v \in V$ está ligado aos *k* elementos mais próximos. Outra abordagem similar, estabelece uma conexão entre vértices cuja distância entre eles não exceda um limiar *d*. No entanto, esses modelos não geram grafos necessariamente conectados e também não apresentam propriedades particulares que possam caracterizá-los sob certas configurações [Vincent-1989].

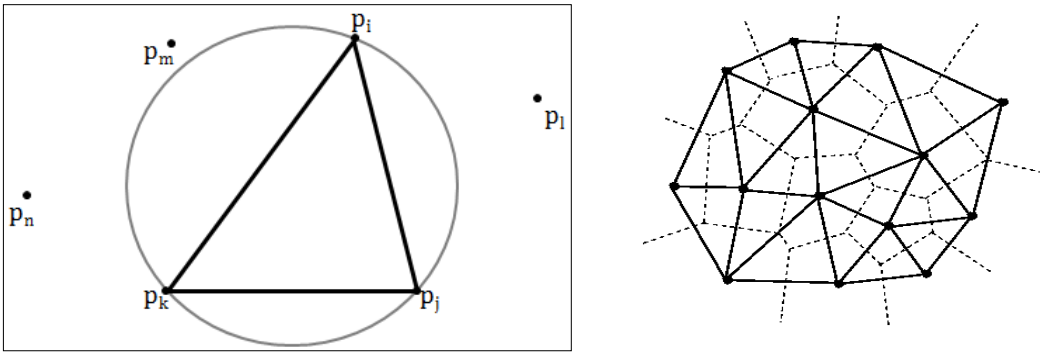
Outra classe de modelos de geração de grafos de vizinhança são representados pelos grafos obtidos a partir do Diagrama de *Voronoi* (DV). Este último, também denominado *Dirichlet Tesselation*, é uma partição do espaço formada por pontos equidistantes dos elementos de V. Para todo $v \in V$ pode ser definido um polígono $Z(v)$ formado pelos pontos mais próximos de v do que de qualquer outro elemento de V:

$$Z(v) = \{m \in \mathbb{R}^2, \forall\, q \in V \backslash \{v\}, distancia(m, v) < distancia(m, q)\} \quad (15)$$

$Z(v)$ também pode ser denominado zona de influência de *v*. A seguir serão descritos alguns modelos de grafos de vizinhança derivados do Diagrama de *Voronoi*:

2.3.1. Triangulação de *Delaunay* (DT)

Também conhecido como grafo dual do Diagrama de *Voronoi*, a triangulação de *Delaunay* (TD) estabelece conexões a cada tripla de pontos, sempre formando triângulos. Neste modelo, dado um conjunto de pontos $P = \{p_1, p_2, \dots, p_n\} \in \mathbb{R}^2$, o triângulo $p_i, p_j, p_k \in TD(P)$ se a circunferência circunscrita ao mesmo é vazia (Figura 2.2a). A dualidade entre a *TD* e o Diagrama de *Voronoi* poder ser visualizado na Figura 2.2b.



(a) (b)

Figura 2.2. (a) Critério para um triângulo pertencer a Triangulação de Delaunay (*TD*). (b) Triangulação de *Delaunay* (linhas cheias) *vs* Diagrama de *Voronoi* (linhas pontilhadas)

2.3.2. Gabriel Graph (GG)

É um modelo proposto da década de 60 por *Gabriel* e *Sokal* [Matula-1980], no qual, há conexão entre dois vértices da rede quando o seguinte critério é atendido: dado um conjunto de pontos $P = \{p_1, p_2, \dots, p_n\} \in \mathbb{R}^2$, p_i e p_j estão conectados se a circunferência que apresenta a distância entre *i* e *j* como diâmetro não contém nenhum ponto inscrito à mesma (Figura 2.3).

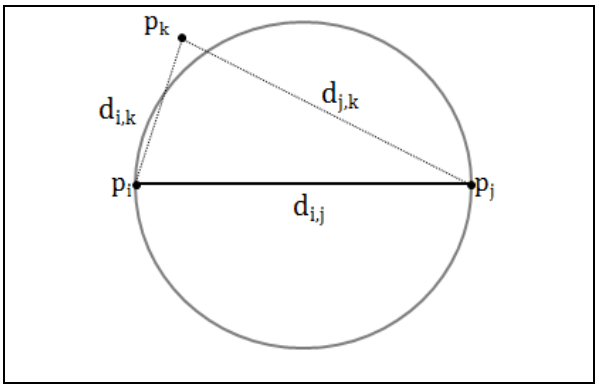


Figura 2.3. Critério para definição de arestas no modelo *Gabriel Graph*

Ou seja, p_i e p_j estão conectados se a distância entre eles for menor do que a soma das distâncias entre cada um destes pontos a um outro ponto qualquer k do espaço (Eq. 16).

$$dist(i,j)^2 < dist(i,k)^2 + dist(j,k)^2 \qquad (16)$$

Sendo *dist*, a distância euclidiana.

2.3.3. Rede de Vizinhança Relativa

Neste modelo, para dois pontos p_i e p_j estarem conectados, a intersecção entre as duas circunferências que apresentam distância entre p_i e p_j como raio comum, deverá estar vazia (Figura 2.4). Ou seja, p_i e p_j estarão conectados se a distância entre eles for menor do que a distância máxima entre p_i, p_k e p_j,p_k:

$$dist(i,j) < \ maximum(\ dist(i,k), dist(j,k)\ ) \qquad (17)$$

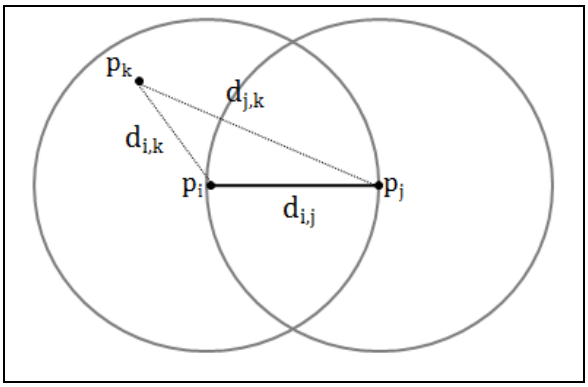


Figura 2.4. Critério para definição de arestas no modelo Rede de Vizinhança Relativa

Os modelos de geração de redes TD, GG e RVR apresentam algumas propriedades específicas [Vincent-1989]:

- São conexos;
- Podem ser obtidos a partir do Diagrama de *Voronoi*;
- Representam aproximações precisas de grafos planares;
- A cardinalidade do conjunto de arestas está limitado à: $3n - 6$, sendo n a ordem do grafo;
- Existe uma relação de inclusão entre eles dada por: $RVR \subset GG \subset TD$

A Figura 2.5 apresenta exemplos de redes geradas a partir dos três modelos descritos acima.

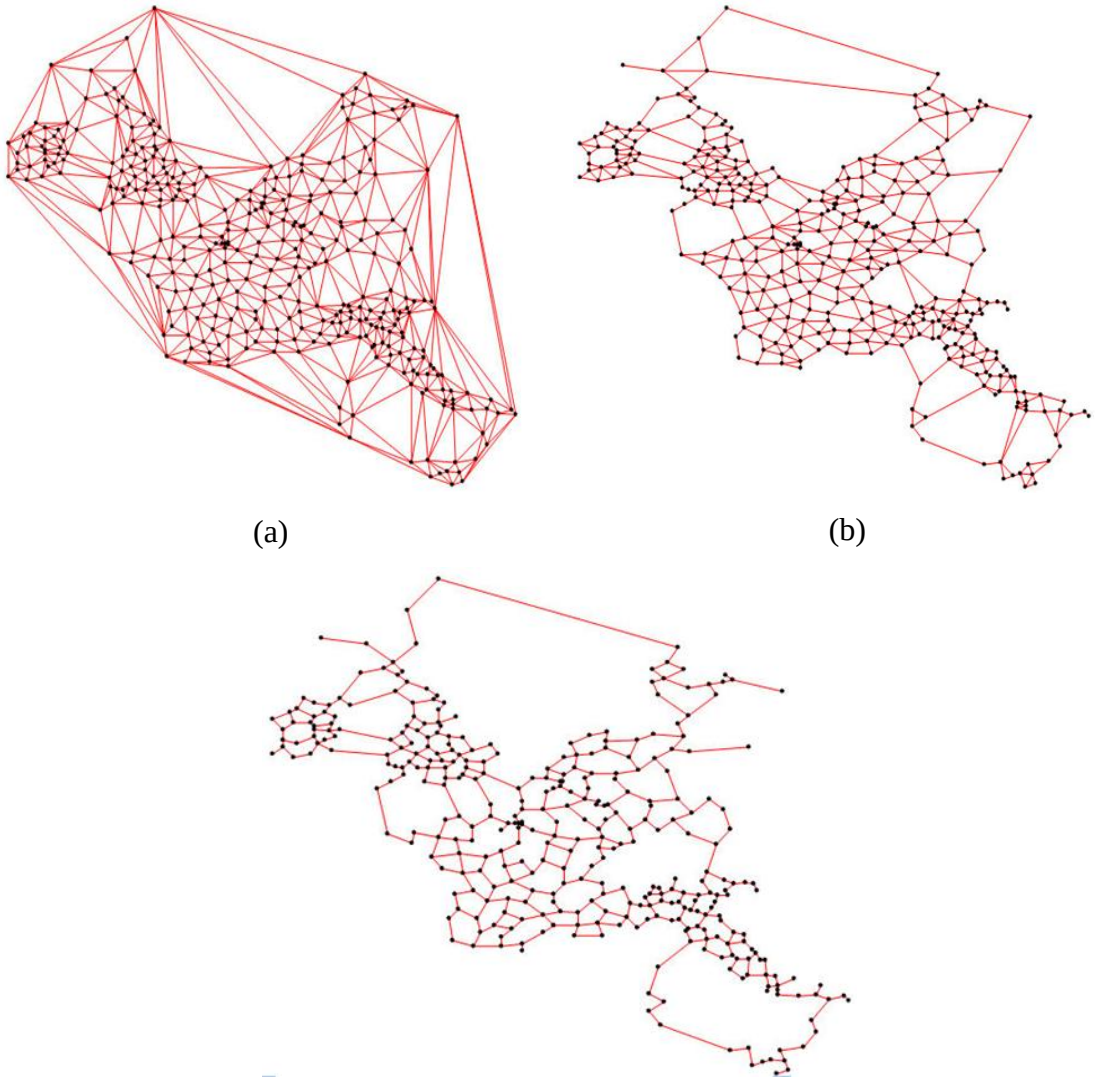


Figura 2.5. Exemplos de modelos de geração de redes. (a) Triangulação de *Delaunay*. (b) *Gabriel Graph*. (c) Rede de Vizinhança Relativa. [Rosenberg-2011]

2.4. Análise Baseada na Teoria de Redes Complexas

O conceito de "rede" é utilizado para caracterização de dados relacionais e a abrangência deste termo é tão grande que sua análise pode apresentar significados diferentes, dependendo do contexto de aplicação. A análise de redes tem sido utilizada em diferentes áreas, tais como, redes sociais, sistemas complexos, epidemiologia, bioinformática, sistemas hipertexto, redes de transporte, entre outras, e, por ser um tema inter-disciplinar, engloba o conhecimento teórico das áreas de ciência da computação, matemática, física, biologia e sociologia [Brandes-2005].

A representação de redes envolve o formalismo matemático da Teoria de Grafos na caracterização do conceito abstrato de "rede". Com o auxílio da área de mecânica estatística, novas ferramentas foram desenvolvidas ao longo dos últimos anos focadas na análise da estrutura de redes que apresentam grandes quantidades de vértices, possibilitando o estudo de sistemas complexos. Desta forma, ferramentas estatísticas aliadas a representação por grafos possibilitou um estudo mais robusto da estrutura topológica das redes, antes consideradas como estruturas aleatórias (em muitos casos, devido a falta de dados experimentais).

Brandes et *al*. define as ferramentas de análise da estrutura das redes em três níveis: a análise do elemento (vértice), do grupo, e, da rede como um todo. As medidas extraídas dessa estrutura abrangem esses três níveis. Essas medidas podem diferenciar estruturas de interesse e serem usadas em uma abordagem de reconhecimento de padrões. Além disso, também permitem um estudo dinâmico dos processos envolvidos na organização estrutural da rede [Brandes-2005].

Costa *et al.* [Costa-2007] descreve várias métricas relacionadas para identificação de características específicas às estruturas das redes que envolvem noções de distância (eqs. 18 e 19), agrupamento (eq. 20), disposição em comunidades, entre outras. Além disso, também são descritas métricas relacionadas à estrutura geral da rede, como a entropia (eq. 21). Algumas destas métricas são descritas a seguir:

- <u>Grau médio</u> (k_{med}):

$$k_{med} = \frac{1}{N}\sum_i k_i \qquad (18)$$

Na qual, k_i é o grau do nó i.

- Distância média (l):

$$l = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \qquad (19)$$

Sendo d_{ij}, é a distância geodésica entre i e j, ou seja, o menor número de arestas conectando i e j.

- Coeficiente de *cluster* (cc): utilizado para a caracterização da presença de *loops* na rede

$$cc = \frac{1}{N} \sum_{i} \frac{2E_i}{k_i(k_i-1)} \qquad (20)$$

Sendo E, o número de arestas entre os nós que estão conectados ao nó i.

- Entropia da distribuição do grau (H): relativa à heterogeneidade da rede

$$H = -\sum_{k} P(k) log P(k) \qquad (21)$$

Os trabalhos de [Gunduz-2004] e [Demir-2005] apresentam exemplos da utilização dessas métricas no processo de classificação em imagens histológicas.

3. TRABALHOS CORRELATOS

3.1. Considerações Iniciais

A extração de atributos a partir de imagens histológicas tem se apoiado, ao longo dos últimos anos, em técnicas computacionais baseadas em textura, morfologia e topologia. Nesta seção serão destacados alguns trabalhos relacionados ao contexto deste projeto. Esses trabalhos abordam técnicas de reconhecimento de padrões aplicados a imagens histológicas e serão descritos na próxima seção ordenados por grupos de pesquisa.

3.2. Descrição dos Trabalhos Correlatos

a) Landini et al.

Neste trabalho é apresentado um método para caracterização estrutural de tecidos histológicos baseado na representação da geometria das relações entre as células. O método proposto utiliza ferramentas da teoria de grafos para caracterização dos padrões encontrados em neoplasias das vias orais. Foi utilizada uma base de imagens contendo aproximadamente 400 imagens classificadas em normais, pré-malignas ou malignas [Landini-2004]. Na etapa de segmentação foi utilizada a técnica de *color deconvolution*, descrita em [Ruifrok-2001], com posterior aplicação da *watershed* para a obtenção das partições que caracterizam as fronteiras entre as células do tecido. Além disso, na abordagem proposta, para cada célula segmentada, cujo centróide é definido pelas coordenadas espaciais $C(x, y)$, é estabelecido o conceito de vizinhança local, caracterizada como uma região de influência de cada núcleo e definida por um raio de abrangência R. Assim, sobre cada vizinhança local, um grafo é construído e medidas são extraídas do mesmo e armazenadas para cada $C(x, y)$.

Foram testados e comparados diferentes modelos de grafos de vizinhança além de diferentes propriedades estatísticas, tais como, média, desvio padrão, medidas de assimetria e dispersão da distribuição das áreas celulares, número total de nós e de arestas, entre outros. As diferenças entre os parâmetros extraídos da rede para as três classes de interesse foram avaliados utilizando-se testes de significância estatística. Os resultados mostraram que para diferentes valores de R (30 e 60) as médias dos valores da maioria dos parâmetros eram

estatisticamente diferentes para as três classes, com melhores taxas de acurácia para R = 60 na classificação das imagens. Na comparação entre imagens normais e malignas, também foi verificado que quando utilizado apenas a área das células no processo de classificação, sem atributos relacionados a características topológicas ou de arquitetura das redes geradas, a taxa de acurácia diminuía (62%), quando comparada a taxa obtida com a utilização desses atributos (75%). Isso evidencia que os parâmetros obtidos por meio das redes forneceram informações relevantes na separação entre as classes de interesse.

Quanto às lesões pré-malignas, foi realizada uma investigação das diferenças estruturais entre três tipos de displasias que podem ser identificadas nesse tecido. Apesar das diferenças estatísticas significativas entre alguns parâmetros observados, as taxas de acurácia foram de 42% para R = 30 e 51% para R = 60. Como discutido pelo autor, verifica-se que a caracterização das transformações graduais nas lesões pré-malignas é uma tarefa mais complexa quando comparada à análise das diferenças topológicas entre imagens normais e malignas, e também, quando comparada às diferenças estruturais entre imagens normais e pré-malignas, e, entre imagens pré-malignas e malignas.

A avaliação das lesões pré-malignas realizada por Landini & Othman, utilizou um método desenvolvido pelos próprios autores e descrito em [Landini-2003] para a identificação automática de camadas em imagens histopatológicas. Este método é baseado em transformações de distância e reconstrução morfológica, através de um processo iterativo que pode ser resumido em três etapas: 1) identificação de um elemento ou de uma camada de referência; 2) identificação dos elementos, neste caso as células, em contato direto com a camada de referência, e, 3) rotulação da camada identificada como a nova camada de referência. O processo é repetido até que todas as células estejam agrupadas em camadas. Uma operação essencial para se definir a próxima camada é a reconstrução binária. Sendo que a imagem utilizada como marcador é sempre definida em função da camada atual. Como camada de referência inicial, poderia ser escolhida, por exemplo, a camada formada pelas células presentes na borda da imagem (superficiais), em contato com o fundo.

b) Geusebrock et al.

O trabalho descrito em [Geusebrock-1999] apresenta um método para segmentação de tecidos a partir das relações de vizinhança entre seus componentes. A construção da rede que identifica os relacionamentos entre as estruturas de interesse é obtida pela técnica de k-vizinhos mais próximos. Para cada v representando um marcador celular, é definido o

subconjunto de vértices mais próximos de v: $N_v^k = \{d_1, d_2, \dots, d_k | \, d_i = dist(v, v_i), \, d_i < d_{i+1}\}$. A partir dessa representação, podem ser extraídos padrões de vizinhança entre as células baseados nas distâncias entre seus componentes. Esses padrões são pontuais, pois traduzem o relacionamento estabelecido entre uma célula de referência e outras células do tecido dado o valor de k. Dessa forma, partindo-se de uma célula de referência, um padrão ou *template*, baseado na distribuição espacial dos elementos do tecido, pode ser definido e utilizado como protótipo na classificação das outras células quanto a similaridade em relação à célula de referência. Dentro deste contexto, células semelhantes são agrupadas gerando como resultado a segmentação do tecido em componentes de interesse.

A determinação da pertinência de um objeto a um agrupamento específico é definida pela comparação do vetor de distâncias observadas para este objeto com o vetor de distâncias do objeto de referência. Para a comparação entre esses dois vetores, são consideradas as inserções, deleções ou substituições que deverão ser realizadas no vetor do objeto para ser equivalente ao vetor de referência. Foram definidos custos para cada uma dessas operações e o custo total da comparação entre os vetores foi utilizado como limiar para o agrupamento. A acurácia obtida na geração dos agrupamentos foi analisada em função da modificações necessárias entre os vetores comparados. O método proposto foi comparado a outros métodos também baseados na geração das relações de vizinhança pelo técnica de *k*-vizinhos mais próximos, assim como a métodos que utilizam o Diagrama de *Voronoi* para definição das relações de vizinhança. Verificou-se que este último, por exemplo, é dependente da identificação correta das posições celulares no tecido, uma vez que a remoção de um determinado ponto leva a uma mudança das partições definidas pelo diagrama. No entanto, quanto à operação de substituição, métodos baseados na utilização do mesmo mostram-se mais robustos. Já o método proposto apresentou melhor desempenho em relação às operações de adição e deleção e, quando comparado a outros métodos baseados na técnica de *k*-vizinhos, mostrou-se mais robusto em relação à detecção de erros. O autor ainda apresenta uma aplicação do método proposto à segmentação da região do hipocampo em imagens histológicas de ratos.

c) Gunduz et al. e *Demir et al.*

No trabalho de Gunduz *et al*. é apresentada uma abordagem baseada na utilização de atributos topológicos para caracterização e diagnóstico de um tipo de câncer cerebral, o glioma. [Gunduz-2004]. O método proposto utiliza as diferenças estruturais dos grafos gerados sobre o

tecido de interesse para diferenciação de regiões normais, regiões que apresentam apenas inflamação e regiões com células cancerosas. A base de imagens utilizada pelo autor é composta de imagens de baixa resolução (384 *x* 384 *pixels*) representando as regiões de interesse. O objetivo na utilização dessas imagens é mostrar a adequação da abordagem estrutural na classificação desse tipo de lesão sem a necessidade de se obter os contornos exatos dos núcleos celulares, tarefa que fica mais difícil em imagens com essas características.

A primeira etapa do método proposto é a aplicação do algoritmo *k-means* para separação das células em relação ao fundo da imagem. O resultado é uma imagem binária, na qual o valor 1 corresponde aos *pixels* pertencentes às células e o valor 0 aos *pixels* pertencentes ao fundo. Essa separação levou em consideração a informação de cor, assim como a participação de um especialista do domínio para a classificação dos dois padrões de interesse na imagem, utilizados então pelo algoritmo *k-means*. A segunda etapa consistiu na determinação do conjunto de nós do grafo. Para isto, foi definida uma grade sobre a imagem resultante da etapa anterior. Essa grade é composta por quadrados de *n* x *n pixels* e para cada entrada da grade é calculada a probabilidade de seus pixels representarem uma célula do tecido. Esse cálculo é baseado na quantidade de valores 1 e 0 encontrados em cada entrada da grade. Desta forma, dois parâmetros devem ser controlados nessa etapa: o tamanho da grade (*n*) e um limiar de probabilidade a partir do qual os *pixels* presentes em cada quadrado da grade serão considerados células e consequentemente nós do grafo. O autor caracteriza essa etapa como uma amostragem dos nós da rede a ser gerada. Posteriormente, as arestas são definidas em função da distância euclidiana entre os nós a partir do modelo de *Waxman* (22), no qual a probabilidade de ligação entre dois nós *u* e *v*, decai exponencialmente com o aumento da distância euclidiana entre eles:

$$P(u,v) = \alpha . e^{-d(u,v)/(\beta . L)} \qquad (22)$$

Os parâmetros α e β desse modelo variam entre 0 e 1, e, controlam a conectividade entre os elementos do grafo. Quanto mais próximos de 1, mais denso será o grafo resultante. Neste trabalho, foram adotados os valores $\alpha = 1$ e $\beta = 0.01$ para evitar grafos densamente conectados, assim como, grafos muito esparsos. Por fim, as seguintes métricas foram extraídas partir de cada nó da rede: grau (maior para células cancerosas), coeficiente de *cluster*, excentricidade e *closeness*. As duas últimas estão relacionadas a distância entre dois nós da rede. Os resultados mostram uma porcentagem de acerto mínima de 85% na diferenciação entre as três classes de interesse. É importante notar que a classificação foi gerada inicialmente para cada célula, e, posteriormente, o tecido foi classificado com base no

número de células de cada categoria presente no mesmo. Redes neurais foram utilizadas na etapa de classificação deste trabalho.

Na abordagem apresentada por Gunduz *et al* observa-se que apesar de oferecer uma solução local para geração das arestas, a utilização de valores de probabilidade para definição das mesmas pode levar a não identificação de arestas importantes na caracterização de regiões de interesse na imagem, já que somente a distância é considerada no cálculo da probabilidade. Regiões normais tendem a apresentar células mais distantes entre si. Porém a ligação entre essas células é um atributo a mais que pode ser explorado, devendo ser representada.

De maneira similar ao trabalho de Gunduz, o método proposto por Demir *et al* também foi aplicado na diferenciação entre regiões normais, regiões que apresentam inflamação e regiões com glioma [Demir-2005a]. No entanto, neste trabalho, são utilizadas métricas globais (calculadas levando-se em consideração todos os nós da rede) para representação dos tecidos em análise. Os resultados são comparados ao resultados obtidos por Gunduz, que utilizou métricas locais (extraídas de cada nó da rede) na diferenciação entre as três classes de interesse. A abordagem de extração de atributos globais aumentou a acurácia na classificação das regiões de interesse, pois permitiu a representação da informação a nível do tecido.

Além disso, o método proposto foi comparado a duas outras abordagens: a primeira utilizou apenas atributos espaciais para representar os componentes de interesse no tecido, sem considerar qualquer ligação ou relação de vizinhança entre os mesmos e a segunda utilizou atributos relacionados à textura, refletindo as relações espaciais entre os níveis de cinza da imagem. Os resultados experimentais mostraram que a utilização de relações de vizinhança fornece um aumento da acurácia de classificação. Foi verificado que a distribuição espacial das células permite a diferenciação entre tecidos com câncer e tecidos normais com acurácia comparável as abordagens relacionais. No entanto, só a distribuição espacial não foi suficiente para a diferenciação entre tecidos com câncer e tecidos que apresentavam inflamação devido à semelhança na densidade de células nestes dois tecidos. Desta forma, o autor conclui que a utilização de atributos obtidos a partir das relações de vizinhança definidas entre as células, por meio de grafos, fornece informações relevantes na comparação entre diferentes tecidos. Os resultados obtidos utilizando-se atributos relacionados a textura foram equivalentes aos resultados obtidos utilizando-se apenas a densidade espacial das células e a acurácia também diminui na diferenciação entre regiões com inflamação e regiões com células cancerosas. As taxas de acurácia obtidas da diferenciação entre as três classes foram 95,45% a partir dos atributos extraídos dos grafos gerados sobre os tecidos; 78,66% a

partir dos atributos relacionados apenas à distribuição espacial das células, e, 89,03% utilizando atributos relacionados à textura.

Em outro trabalho, Demir apresenta uma metodologia baseada nos chamados *augmented cell graphs* (ACG), mostrando como essa abordagem pode aumentar a acurácia do diagnóstico automático de câncer [Demir-2005b]. ACGs são grafos não direcionados ponderados (as arestas apresentam custo associado) e completos (que apresentam uma aresta entre cada par de vértices). Neste trabalho, os nós dos grafos são representados por agrupamentos celulares, que substituem a vizinhança local. O objetivo de se usar ACGs é deixar o método independente não só da morfologia exata da célula, mas também da localização exata do núcleo / célula. Os parâmetros utilizados para se definir os nós dos grafos (tamanho do *grid* e limiar de probabilidade) foram otimizados por validação cruzada. O autor conclui que o método de definição dos nós e de representação dos mesmos por meio de agrupamentos é menos sensível às variações do processo de aquisição das imagens, sendo adequado, inclusive, para o tratamento de imagens de baixa resolução.

Em [Gunduz-Demir-2007], é apresentado um método que vai além da classificação de padrões em imagens histológicas e fornece análise do mapeamento da evolução do glioma baseada em uma análise de conectividade entre os nós da rede gerada sobre o tecido. As arestas definidas entre os nós da rede são determinadas em função da distância euclidiana entre os nós, de acordo com a seguinte probabilidade: $P(u,v) = d(u,v)^{\alpha}$. Essa abordagem é semelhante ao método utilizado por Gunduz *et al* para a definição das arestas, porém, neste último a probabilidade é definida em função do modelo de *Waxman*. Dessa forma, a conectividade pode ser definida em função de dois parâmetros: a distribuição espacial dos nós (células) e o valor do expoente α. Como o primeiro parâmetro é específico do tecido, o segundo parâmetro foi utilizado para se controlar a conectividade do grafo.

Nos experimentos realizados, o valor de α foi aumentado gradualmente e os grafos obtidos para cada valor do parâmetro foram avaliados quanto aos componentes conexos gerados. Essa avaliação levou em consideração os tipos de componentes encontrados, tais como, nós isolados, árvores, estruturas complexas, entre outros e a quantidade dos mesmos para cada classe de interesse (normal, glioma de baixo-grau e glioma de alto grau). Foram analisados valores de α contidos no intervalo [-2, 0]. É interessante notar que o número de componentes conexos tende a diminuir com o aumento do valor de α, já que a probabilidade também aumenta e, dessa forma, mais ligações entre os nós da rede serão geradas. Em outro experimento, quando fixado o valor de α, verificou-se que o número de componentes conexos

para cada classe de interesse era estatisticamente diferente. O autor conclui que os grafos gerados para diferentes tipos de tecido evoluem diferentemente apesar de apresentarem quantidade de nós semelhante. Assim, a fase de um grafo é determinada pelo número de arestas do mesmo, assim como, pelas características de seus componentes conexos.

f) Rizzio et al. e Zampirolli et al.

A caracterização de tecidos histológicos por meio de grafos também se mostra adequada ao processo de diferenciação de tipos de tecidos. Rizzio *et al*. apresentam uma metodologia para classificação da arquitetura de tecidos utilizando modelagem em grafos [Rizzio-2011]. Um vetor de atributos é gerado para cada grafo contendo medidas extraídas dos mesmos previamente calculadas. A metodologia proposta foi testada em dois tipos de tecidos, epitelial e adiposo, sendo que ambos apresentam arquiteturas semelhantes. Os resultados obtidos indicam que a distância média (DM) pode ser utilizada como um classificador linear na distinção desses tecidos, além de possuir uma interpretação biológica: como a DM é maior nas amostras de tecido adiposo em relação às amostras de tecido epitelial, conclui-se que as células deste último apresentam maior capacidade de comunicação entre si. Uma abordagem semelhante é descrita por Zampirolli *et al*. que utiliza atributos morfológicos extraídos de tecidos, além de atributos topológicos, como fonte de dados [Zampirolli-2010] para a classificação de tipos diferentes de tecidos histológicos. Cada tecido é representado por um grafo obtido a partir da geometria de suas células, considerando-se a vizinhança entre as mesmas. Na etapa de classificação foi utilizado SVM (*Support Vector Machines*).

g) Keenan et al. Em um contexto mais próximo ao deste trabalho, pode ser destacado o trabalho de Keenan *et al.*, cujo objetivo é a análise automática de lesões do colo uterino [Keenan-2000]. O processo de segmentação e detecção de núcleos das imagens utilizadas é baseado em uma técnica de *thresholding* iterativo, na qual, em cada etapa, tenta-se selecionar os pixels pertencentes aos núcleos de acordo com o novo valor de limiar adotado. A classificação dos pixels como pertencentes ou não ao núcleo é auxiliada por operadores morfológicos baseados em detecção de bordas e extração de contorno. O processamento remanescente tem como objetivo a geração de uma rede sobre o tecido, definida pela triangulação de *Delaunay* e extração de medidas relacionadas à área dos triângulos formados pelas arestas, assim como o comprimento médio das mesmas. Além disso, há uma proposta de divisão do tecido em camadas de tamanhos regulares, de forma que a espessura é sempre

igual para as três camadas e atributos são extraídos de cada uma. Características discriminantes para esses atributos foram verificadas entre as classes normal e NIC III (98,7% de acurácia de classificação); entre as classes NIC I e amostras que continham apenas coilócitos (76,5%); e entre as classes NIC-I, NIC II e NIC III (62,3%). Esses resultados indicam melhor acurácia na separação entre as classes normal e lesão de alto nível (NIC III), porém esse valor decresce na diferenciação entre lesões intermediárias.

3.3. Considerações Finais

A partir dos trabalhos relacionados acima, quanto ao contexto de aplicação, diferente dos trabalhos de Landini, Gunduz e Demir, o presente trabalho visa a caracterização e a classificação automática de lesões que antecedem o diagnóstico de câncer e que se instalam de maneira gradual no tecido epitelial do colo uterino. Naqueles trabalhos os autores apresentaram métodos para diferenciação de tecidos normais, com inflamação e cancerosos, não levando em consideração o grau da anomalia identificada, apenas a detecção da mesma. Essa diferença de contexto de aplicação também pode ser verificada em relação aos trabalhos de Rizzio e Zampirolli, cujo objetivo era a diferenciação de diferentes tipos de tecidos (epitelial e adiposo).

A caracterização da evolução temporal proposta por Gunduz-Demir está relacionada a identificação de lesões de alto e baixo grau relativas ao glioma. No caso das lesões de colo uterino foi utilizado o sistema NIC de classificação composto por três classes: NIC-I, NIC-II e NIC-III. Porém, a principal diferença do método proposto em relação ao trabalho de Gunduz-Demir e em relação a outros trabalhos descritos na seção anterior é a geração do grafo sobre o tecido. Naqueles trabalhos foram utilizados métodos probabilísticos para se definir as arestas entre os vértices dos grafos. O presente trabalho, assim como os de Rizzio e Zampirolli, utilizam relações de vizinhança entre as células para se estabelecer as ligações entre as mesmas.

Como relação ao trabalho de Keenan, para extração de atributos por camadas, o tecido epitelial é dividido indistintamente em três partes iguais, sendo a área média dos triângulos da rede utilizada para classificação entre as NICs. No presente trabalho, um novo método para identificação automática das camadas deste tecido é proposto, baseado na organização estrutural de seus componentes, utilizando, para isto, um algoritmo de rotulação em grafos baseado nas propriedades da Triangulação de *Delaunay* (TD). Este método é independente da escala da imagem, assim como da posição angular de suas estruturas. Além disso, a área dos triângulos da TD foi utilizada para identificação de agrupamentos de interesse na imagem. Para classificação das NICs, foi proposto um classificador baseado no

número de agrupamentos que podem ser identificados por imagem e em métricas extraídas da rede gerada. Valores de acurácia, sensitividade e especificidade foram utilizadas na avaliação da metodologia proposta.

4. MÉTODO PARA ANÁLISE ESTRUTURAL DE TECIDOS

4.1. Considerações Iniciais

A proposta deste projeto baseia-se na exploração de atributos topológicos a fim de auxiliar a caracterização de diferentes estruturas em tecidos, identificando alterações nas mesmas, quando for o caso. Neste capítulo são apresentadas as etapas que constituem um método genérico proposto para análise de imagens histológicas, visando a identificação de componentes e regiões de interesse.

4.2. Mapeamento das relações de vizinhança entre núcleos celulares

As características dos núcleos celulares são parâmetros importantes na análise de imagens histológicas e podem descrever alterações funcionais específicas. Em decorrência, grande parte dos métodos de segmentação aplicados a essas imagens tem por objetivo a separação dos núcleos. O método de segmentação empregado é determinado pelos atributos que se deseja extrair. Quando há interesse na morfologia dos núcleos, métodos baseados em detecção de bordas são mais indicados, já que fornecem contornos bem precisos. No caso dos atributos topológicos, a identificação aproximada da localização dos núcleos pode ser suficiente para a representação da dependência espacial entre os mesmos. Desta forma, a imagem resultante do processo de segmentação é uma imagem binária, na qual cada núcleo identificado é um componente conexo. Assim, os núcleos segmentados poderão ser representados pelo conjunto de vértices de um grafo, usando a localização dos seus centróides.

Partindo-se desse conjunto, é possível modelar o relacionamento entre seus elementos a partir de diferentes critérios. Os grafos de vizinhança citados na seção 2.3, por exemplo, consideram alguns critérios de adjacência entre os elementos do conjunto para o estabelecimento de conexões entre os mesmos. Por considerarem a posição geográfica dos vértices, são adequados para a representação da distribuição dos núcleos no tecido.

Devido às suas propriedades particulares, a Triangulação de *Delaunay* (TD) foi o modelo adotado neste trabalho como parte do método proposto. A uniformidade quanto ao tipo de polígono gerado por este modelo (sempre triângulos) permite que atributos adicionais

sejam explorados. Além disso, a TD pode ser facilmente obtida pelo diagrama de *Voronoi*, definido sobre os núcleos segmentados.

Assim, um grafo $G(V, E)$, pode ser obtido por meio da $TD(V)$, sendo V composto pelos centróides dos núcleos celulares e E pelas conexões definidas entre os mesmos.

4.3. Análise Baseada em Regiões

Estudos relacionados [Rizzio-2011][Zampirolli-2010] mostram que uma abordagem baseada na extração de atributos globais (obtidos levando-se em consideração toda a estrutura do grafo) fornece suporte à análise da organização estrutural de tecidos distintos. Contudo, para amostras provenientes de um mesmo tecido, normalmente, as diferenças entre elas podem ser verificadas por alterações graduais, em apenas algumas regiões. Por isso, mostra-se mais adequada a extração local de atributos, caracterizada pela análise de regiões ou agrupamentos característicos dessas alterações. O método proposto neste capítulo foi definido de maneira que pudesse fornecer esses agrupamentos.

Dentro desse contexto, propriedades específicas da TD e algoritmos de rotulação em grafos foram utilizados na implementação do método proposto. Nota-se que, para este modelo de grafo de vizinhança, regiões do tecido que apresentam aglomerados nucleares tendem a gerar TDs cuja área média dos triângulos é pequena. Por outro lado, regiões com distribuição nuclear mais esparsa tendem a gerar TDs com área média relativamente maior (Figura 4.1). Análise semelhante pode ser feita quanto ao comprimento médio dos lados dos triângulos.

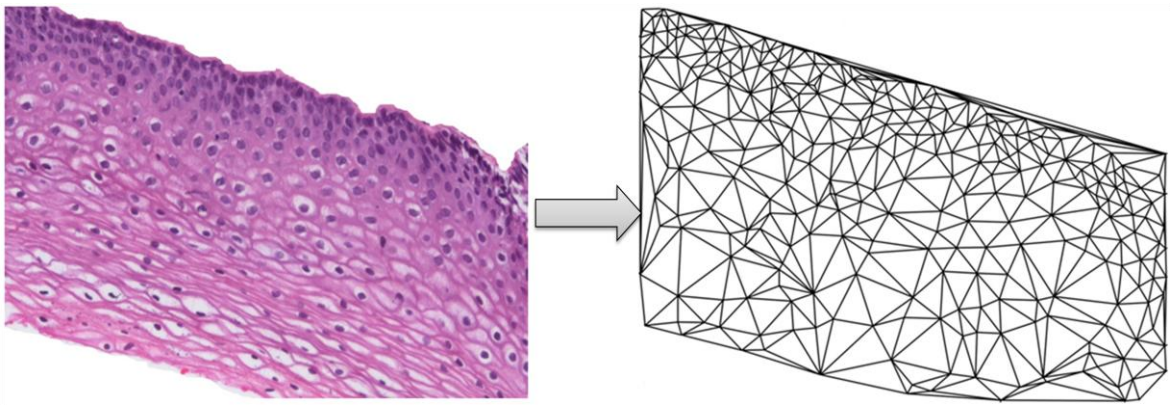


Figura 4.1. Representação do tecido histológico pela Triangulação de *Delaunay*

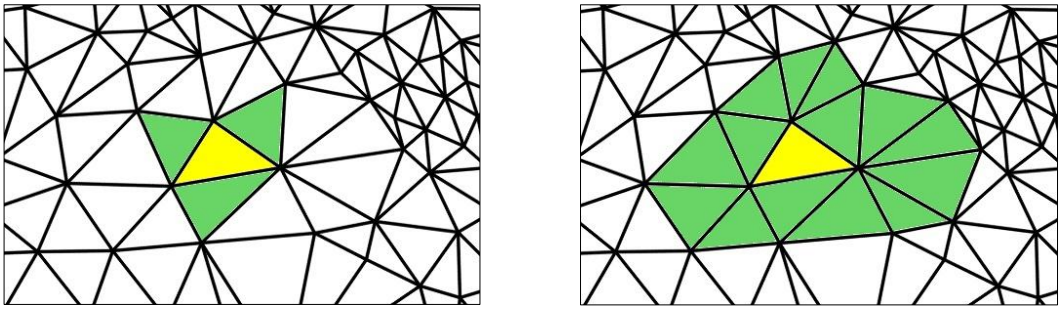
4.3.1. Geração de Agrupamentos

A fim de utilizar as propriedades específicas da TD, um novo grafo $G'(V', E')$ foi definido sobre a $TD(V)$, no qual V', o novo conjunto de vértices, passa a ser representado

pelo conjunto de triângulos pertencentes à $TD(V)$ e o conjunto E', definido por critérios de semelhança e adjacência entre os pares de triângulos. Esses critérios serão apresentados nas seções subsequentes.

4.3.1.1. Relações de Adjacência

As relações de adjacência definem a vizinhança entre os elementos do grafo, neste caso, entre os triângulos da *TD*. Dois tipos de adjacências foram definidas e utilizadas: na primeira, dois triângulos são considerados adjacentes se apresentarem pelo menos uma aresta em comum (Figura 4.2a). Na segunda, dois triângulos são adjacentes se apresentarem pelo menos um vértice em comum (Figura 4.2b). É possível verificar que a segunda RA é bem menos restritiva, considerando um número maior de vizinhos em relação à primeira.



(a) (b)

Figura 4.2. Relações de adjacência. O triângulo de referência é representado pela cor amarela e os seus vizinhos pela cor verde. (a) Triângulos com arestas em comum. (b) Triângulos com pelo menos um vértice em comum

4.3.1.2. Critérios de Agrupamento

Um agrupamento é definido como um subgrafo de G', ou seja, seu conjunto de vértices V'' é um subconjunto da $TD(V)$. Um critério de agrupamento define um limiar δ a partir do qual um triângulo t_i pertencerá a um subgrafo G''. Foram definidos dois critérios de agrupamento:

- *Proporção relativa entre áreas* (δ_A)

  Dois triângulos t_i e t_j, serão agrupados se:

  $$\frac{A_{t_j}}{A_{t_i}} \leq 1 + \delta_A \qquad e \qquad \frac{A_{t_i}}{A_{t_j}} \leq 1 + \delta_A$$

  Sendo A_{t_i} a área do triângulo t_i, A_{t_j} a área do triângulo t_j e $\delta_A \in [0,1]$

- *Semelhança entre triângulos* (δ_S)

  Dois triângulos t_i e t_j serão agrupados se: $dist(t_i, t_j) < \delta_S$,

Sendo $dist$ a distância euclidiana calculada entre os vetores formados pelos valores das medidas dos lados de t_i e t_j e $\delta_S \in [0,1]$. O vetor gerado para cada triângulo é ordenado como descrito na Figura 4.3.

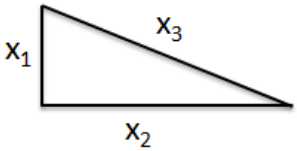


Figura 4.3. O triângulo exemplificado, tal que $\boldsymbol{x_1} \leq \boldsymbol{x_2} \leq \boldsymbol{x_3}$, passará a ser representado pelo vetor $\vec{\boldsymbol{X}} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}]$ mantendo a relação de ordem entre as medidas dos lados

4.3.1.3. Algoritmo de Agrupamento

O algoritmo apresentado no Quadro 1 agrupa triângulos da TD utilizando as relações de adjacência e os critérios de agrupamento definidos anteriormente. Esse algoritmo permite o mapeamento da TD em agrupamentos ou subgrafos que possam fornecer melhor representação de objetos e separação de regiões de interesse na imagem sob análise.

A escolha de uma RA estabelece o critério de vizinhança entre núcleos e pode ser implementada por uma matriz de adjacência (M). O critério de agrupamento, quando definido em função da proporção relativa entre áreas (δ_A), é representado por um valor percentual que traduz o limiar superior pelo qual dois triângulos ainda poderão ser agrupados. Quando definido em função da semelhança entre triângulos (δ_S), é representado por um valor percentual da semelhança máxima entre dois triângulos quaisquer da rede, ou seja, é um limiar de semelhança abaixo do qual dois triângulos poderão ser agrupados.

O algoritmo inicia-se a partir de um triângulo t_i, denominado triângulo de referência, para o qual são verificados os triângulos que respeitam os critérios de adjacência e semelhança através dos métodos *booleanos: adjacente*(t_i, t_j) e *similar*(t_i, t_j, δ). Esses triângulos são empilhados em $\mathcal{P}$ e juntamente com t_i são adicionados a um agrupamento C_i. Enquanto $\mathcal{P}$ não estiver vazio, o mesmo processo é repetido analisando-se os triângulos vizinhos aos elementos de $\mathcal{P}$. Ou seja, enquanto houver elementos em $\mathcal{P}$, mais triângulos poderão ser adicionados ao agrupamento C_i. O método *visitado*(t_j) verifica se o triângulo analisado já foi adicionado a algum agrupamento. Quando a pilha estiver vazia o agrupamento

C_i não receberá mais triângulos e um novo agrupamento é criado. Então, o processo descrito acima é repetido para os triângulos remanescentes até que todos sejam agrupados. Cada triângulo é visitado apenas uma vez. É importante notar que as verificações dos critérios de adjacência e semelhança são sempre efetuadas em relação ao triângulo de referência t_i.

Quadro 1. Detecção de agrupamentos

O algoritmo de Agrupamento

Entrada. Imagem Binária (I) contendo a posição dos centróides dos núcleos celulares.
Saída. Conjunto de n agrupamentos: $\mathcal{C}$.

Passo 1. Definir RA: matriz (M)

Passo 2. Definir CA: δ

Passo 3. Agrupar*:*
$\mathcal{P} \leftarrow \emptyset$
$\mathcal{C} \leftarrow \emptyset$
$C_i \leftarrow \emptyset$

Para cada $t_i \in TD(V, E)$
  Se !*visitado*(t_i)
    $agrupar(t_i, C_i)$

  Para cada $t_j \in TD(V, E)$
    Se !*visitado*(t_j) **e** *adjacente*(t_i, t_j) **e** *similar*(t_i, t_j, δ)
      $agrupar(t_j, C_i)$
      $inserir(t_j, \mathcal{P})$

  Enquanto $(\mathcal{P} \neq \emptyset)$
    $remover(t_p, \mathcal{P})$
    Para cada $t_k \in TD(V, E)$
    Se !*visitado*(t_k) **e** *adjacente*(t_k, t_p) **e** *similar*(t_i, t_k, δ)
      $agrupar(t_k, C_i)$
      $inserir(t_k, \mathcal{P})$

  $\mathcal{C} \leftarrow C_i$
  $C_i \leftarrow \emptyset$

Retorna$(\mathcal{C})$

Quanto maior o valor de δ, menor o número de agrupamentos gerados, porém, mais heterogêneos poderão ficar os agrupamentos. Portanto, a escolha desse valor é um compromisso entre a cardinalidade de $\mathcal{C}$ e a tolerância quanto ao critério de agrupamento.

4.3.2. Rotulação de Agrupamentos

A obtenção de um grande número de agrupamentos nem sempre é desejável. O algoritmo de agrupamento poderá ser aplicado novamente sobre $\mathcal{C}$ em um processo hierárquico, no qual os valores médios de cada grupo passam a ser referência na verificação dos critérios de agrupamento. No entanto, deve-se considerar que os próximos níveis de aplicação deste algoritmo exigirão valores de δ maiores do que para as iterações anteriores, podendo levar a agrupamentos com elementos bastante distintos. As próximas seções apresentam duas abordagens baseadas na área média dos triângulos da rede que auxiliam o controle do número de agrupamentos. Esse controle é efetuado de acordo com as regiões de interesse que se deseja caracterizar na imagem. A primeira é uma abordagem não supervisionada que utiliza filtragem de agrupamentos por taxa de ocupação e a segunda é uma abordagem supervisionada na qual valores de área média são estimados e utilizados como referência para classificação dos agrupamentos.

4.3.2.1. Abordagem não-supervisionada: filtragem de agrupamentos por área de ocupação

O método apresentado nesta seção pode ser utilizado para se reduzir o número de agrupamentos gerados a N'. Para isso, é utilizada a taxa de ocupação (TO) de cada agrupamento, definida como:

$$TO = A_{C_i}/A_{\mathcal{C}}, \qquad (23)$$

Na qual A_{C_i} é a área do agrupamento C_i, que corresponde à soma das áreas dos triângulos pertencentes a este agrupamento, e, $A_{\mathcal{C}}$ é a área total da rede, que corresponde a soma das áreas de todos os triângulos. Com base no valor de TO um limiar α pode ser utilizado para filtragem de agrupamentos considerados muito pequenos em relação à área ocupada. O algoritmo de eliminação desses agrupamentos pela incorporação dos mesmos a outros com maiores TOs é descrito no Quadro 2.

Seja um agrupamento permanente definido por: $C_p \in \mathcal{C},\ TO_{C_p} \geq \alpha$ e um agrupamento a ser filtrado definido por: $C_f \in \mathcal{C},\ TO_{C_f} < \alpha$, a primeira etapa do algoritmo tem como objetivo classificar os agrupamentos de $\mathcal{C}$ em C_p ou C_f. Posteriormente, para cada C_f, calcula-se a diferença absoluta (Δ) entre sua área média e a área média de cada C_p. O menor valor de Δ_a define a qual C_p, C_f deverá ser incorporado. Ao final do processo será

retornado o conjunto $\mathcal{C}_P$, contendo todos os agrupamentos permanentes após a etapa de filtragem.

Quadro 2. Algoritmo para filtragem de agrupamentos por área de ocupação

Filtragem por área de ocupação

Entrada. Conjunto de n_1 agrupamentos: $\mathcal{C}$.
Saída. Conjunto de n_2 agrupamentos: $\mathcal{C}'$, sendo $n_2 \leq n_1$

$\mathcal{C}_P \leftarrow \emptyset$
$\mathcal{C}_F \leftarrow \emptyset$

Para cada $C_i \in \mathcal{C}$
    Se $(TO_{C_i} \geq \alpha)$
        $inserir(C_i, \mathcal{C}_P)$
    Se não
        $inserir(C_i, \mathcal{C}_F)$

Para cada $C_j \in \mathcal{C}_F$
    Para cada $C_k \in \mathcal{C}_P$
        Se $\Delta_{C_j C_k} \geq \Delta_{C_j C_{k+1}}$
            $C_m \leftarrow C_{k+1}$
    $agrupar(C_j, C_m)$

Retorna $\mathcal{C}_P$

4.3.2.2. Abordagem supervisionada: estimação da área média

Este método utiliza-se de uma referência externa na definição de limiares que forneçam uma classificação dos agrupamentos da rede. Essa classificação é determinada pelo contexto de aplicação e engloba a participação de um especialista do domínio de análise da imagem na definição *a priori* dos componentes que se deseja obter da imagem.

Mais especificamente, o especialista contribui segmentando manualmente objetos ou regiões de interesse a partir de imagens de treinamento. Essas imagens atuam como *templates*, a partir dos quais podem ser obtidas medidas ou limiares que caracterizem os seus componentes. Para cada região segmentada, é gerada uma TD e a área média de seus triângulos é calculada. Os valores estimados são utilizados então utilizados na classificação de outros agrupamentos. Dessa forma, o número máximo de agrupamentos possível torna-se um valor fixo.

Esse processo pode ser aplicado, por exemplo, na identificação das camadas do tecido epitelial do colo uterino, as quais seriam segmentadas manualmente pelo patologista em um conjunto de imagens utilizadas para treinamento. A partir desta segmentação, cada camada identificada pelo especialista passa a ser caracterizada como um agrupamento, de acordo com o modelo proposto. Esse agrupamento poderá ser, então, representado quantitativamente e passará a representar uma das três camadas do tecido epitelial.

4.4. Representação de agrupamentos: geração do vetor de características

Os algoritmos apresentados na seção anterior permitem a representação do tecido por meio de agrupamentos característicos, como descrito da Figura 4.4.

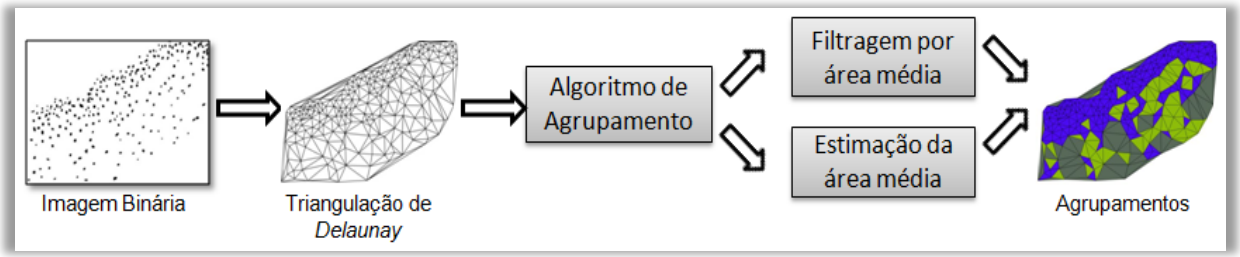


Figura 4.4. Processo de geração de agrupamentos.

Define-se então um conjunto de atributos representando medidas a serem extraídas de cada agrupamento por meio do seguinte vetor:

$$\vec{X} = [X_{11}, X_{12}, \dots, X_{1n}, X_{21}, X_{22}, \dots, X_{2n}, \dots, X_{An}] \qquad (24)$$

Sendo n o número de agrupamentos, A o número de atributos e X_{ij} o valor do atributo i obtido sobre o agrupamento j.

4.5. Classificação da Imagem

Como parte do modelo proposto para avaliação de imagens histopatológicas, a última etapa tem como objetivo a modelagem de um classificador que deverá fornecer a saída desejada de acordo com as características do problema em análise. Os parâmetros necessários para definição do classificador são obtidos por meio do processo de treinamento, efetuado sobre um conjunto de dados utilizado para o aprendizado do conceito de interesse, como descrito a seguir:

Seja Y a variável que descreve as classes de interesse e $\hat{E}[X_i]$ a média amostral para o atributo X_i, um vetor representativo $\vec{X}_r$ pode ser estimado para cada $y \in \{1, 2 \cdots N_c\}$:

$$\vec{X}_{r_y} = \left[\hat{E}[X_{11}], \hat{E}[X_{12}], \dots, \hat{E}[X_{AN_y}]\right] \qquad (25)$$

Sendo A o número de atributos; N_y, o número de instâncias pertencentes à classe y e N_C, o número de classes. Ou seja, cada classe passa a ser representada por $\vec{X}_{r_y}$. Para a rotulação de uma nova instância ($\vec{X}_i$), são calculadas as distâncias euclidianas entre $\vec{X}_i$ e cada $\vec{X}_{r_y}$. A instância i receberá o rótulo da classe que fornecer a menor distância.

O cálculo da distância euclidiana entre o vetor de características da instância i e o vetor representativo de uma determinada classe é adequada quando estes dois vetores possuem a mesma dimensão. Quando uma nova instância não apresenta o número máximo de agrupamentos possível, um classificador específico deverá ser projetado para cada combinação de agrupamentos, uma vez que a ausência de um agrupamento específico pode ser uma informação importante na classificação final da imagem.

4.6. Considerações Finais

Neste capítulo foi apresentado o método proposto para análise estrutural de imagens histopatológicas, a partir do qual, podem ser caracterizados os componentes de interesse nos tecidos sob análise, assim como, suas alterações estruturais.

5. MÉTODO PARA ANÁLISE AUTOMÁTICA DE IMAGENS DE COLO UTERINO

5.1. Considerações Iniciais

O método descrito neste capítulo tem por objetivo a análise automática de imagens histopatológicas do colo uterino visando a identificação de lesões precursoras do câncer que aí se instala. A Figura 5.1 apresenta as etapas que serão descritas nas seções subsequentes. Os processos em branco (sem preenchimento) não foram executados neste projeto. A aquisição de imagens e o levantamento de parâmetros perceptuais de análise das mesmas foi realizado com o auxílio de especialistas da área. Operadores morfológicos e a transformada *watershed* foram utilizados na etapa de segmentação. O método apresentado do capítulo 4 foi utilizado para análise estrutural dessas imagens e envolve as etapas de geração do grafo, identificação das camadas do tecido, extração de métricas e classificação das NICs.

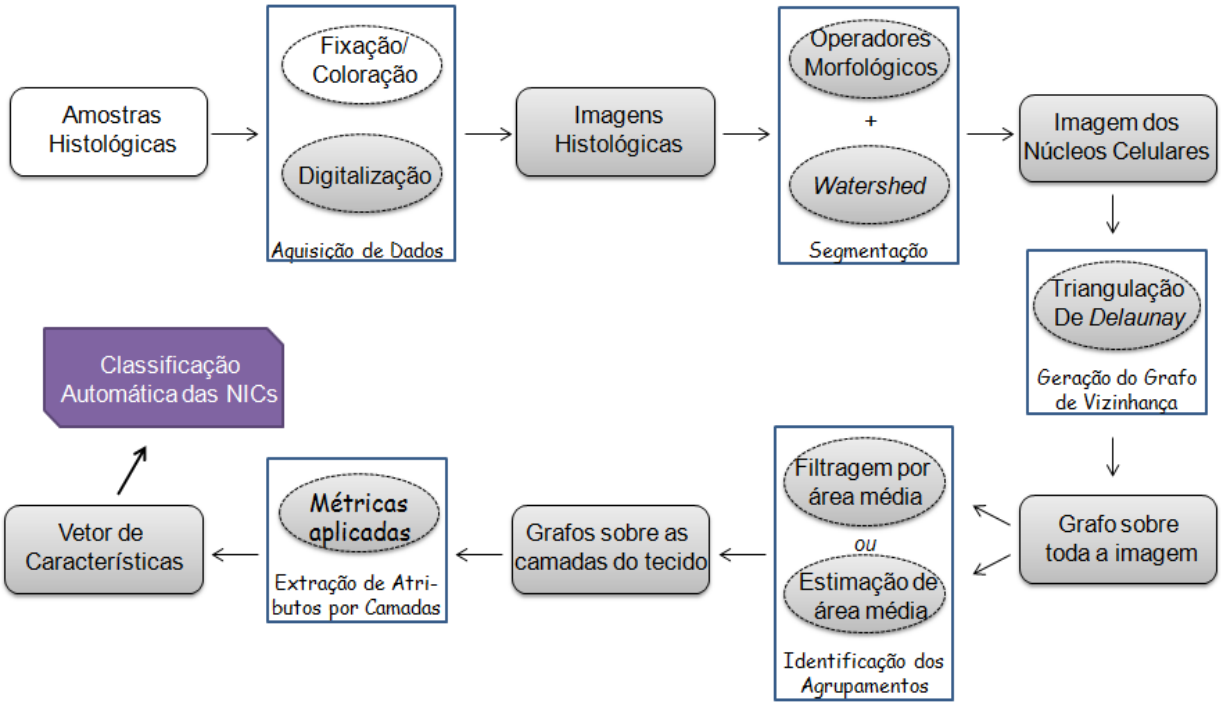


Figura 5.1. Processo de análise automática das imagens de tecido epitelial do colo uterino. Os processos em branco (sem preenchimento) não foram executados neste projeto

5.2. Levantamento de Parâmetros Perceptuais

Os parâmetros perceptuais são caracterizados pelos atributos de análise utilizados pelos patologistas no processo de determinação do diagnóstico. Especificamente, no caso das lesões neoplásicas do colo uterino, alguns parâmetros de interesse para o patologista são: a perda da estratificação do tecido, ou seja, a despolarização das três diferentes camadas do tecido epitelial; a proporção relativa do volume nuclear em relação ao volume citoplasmático (na imagem, representados pelas respectivas áreas); a presença de HPV (*Human Papilloma Virus*), identificado por estruturas denominadas coilócitos, entre outros. Apesar de subjetiva, a definição desses parâmetros é auxiliada por padrões internacionais, como o Bethesda e o NIC, apresentados na seção 1.2.

No presente projeto, esses parâmetros foram definidos em reuniões regulares com patologistas, assim como, por meio do estudo da literatura de referência. O parâmetro escolhido como objeto de estudo no presente trabalho foi a alteração estrutural causada pela despolarização das camadas do tecido epitelial.

5.3. Segmentação

Para a segmentação das imagens histopatológicas do colo uterino foi implementado um *pipeline* de operadores morfológicos com a adoção da transformada *watershed* baseada em marcadores. A parte específica desse algoritmo proposto por [Meyer-1990], é a definição dos marcadores, pois estes são dependentes do contexto de análise da imagem, o que nem sempre é uma tarefa fácil em uma abordagem totalmente automática. Para isto, podem ser utilizadas a marcação manual, a semi-automática e, quando possível, a automática. No método aqui proposto são utilizados operadores morfológicos para a determinação automática de marcadores para os núcleos celulares.

Para a abordagem apresentada nesta seção, as imagens de interesse são tratadas em níveis de cinza e um *pipeline* de operadores morfológicos é definido e aplicado em um processamento inicial para se determinar os marcadores internos ao núcleos. Esses marcadores são representados por componentes conexos, que definem as regiões de interesse a serem segmentadas. O *pipeline* de operadores é detalhado a seguir:

- *close-by-reconstruction top-hat*: destaca a diferença entre pequenos agrupamentos de pixels mais escuros (que apresentam valores de pixels mais baixos) circundados por pixels mais claros (valores de pixels mais altos). A transformada

close top-hat é caracterizada pela subtração da imagem original pelo fechamento morfológico (eq 8) e definida pela seguinte equação:

$$f \hat{\cdot} b = (f \cdot b) - f \qquad (26)$$

- *abertura / fechamento*: esses operadores efetuam uma filtragem, eliminando ruídos sobre a imagem resultante da aplicação do operador *close-by-reconstruction top-hat*. Antes da aplicação desses operadores, a imagem de referência deve ser limiarizada, gerando-se uma imagem binária.
- *area open*: este operador é um filtro conexo e sua aplicação tem como objetivo garantir que cada marcador encontrado seja uma componente conexa da imagem.

Em uma segunda etapa, são obtidos os marcadores externos aos núcleos celulares, cuja função é limitar a área máxima de crescimento da região segmentada quando da aplicação da *watershed*. Inicialmente, a partir da imagem obtida na etapa anterior, é calculado o diagrama de *Voronoi*, que consiste na determinação de fronteiras equidistantes entre os objetos de interesse na imagem. Para as imagens em análise, os centróides das regiões conexas, definidas como marcadores internos, são utilizados para representar os núcleos celulares. Cada região obtida a partir do diagrama de *Voronoi* é chamada "zona de influência". A coleção do todas as zonas de influência é denominada SKIZ (*Skeleton of Influence Zone*). A imagem resultante desse processo contém os marcadores internos e os externos. Finalmente, é aplicada a *watershed* sobre o gradiente da imagem original, utilizando-se os marcadores obtidos nas etapas anteriores. A sequência de operadores aplicada é resumida na Figura 5.2.

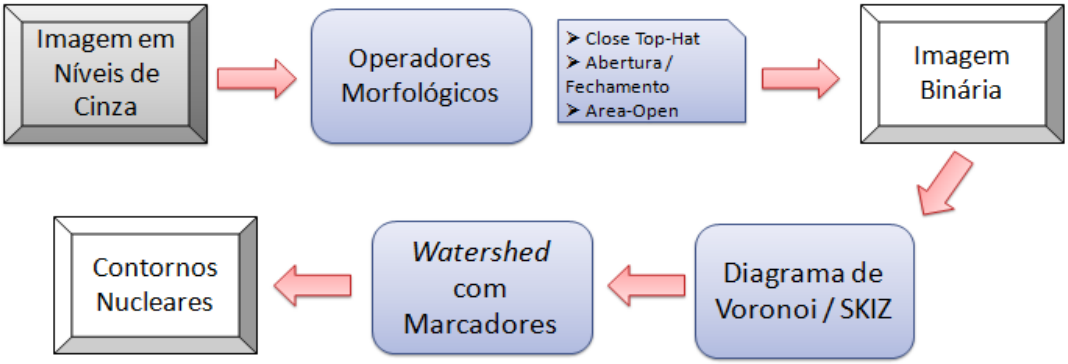


Figura 5.2. *Pipeline* de operadores morfológicos utilizado na segmentação dos núcleos celulares.

5.4. Análise Estrutural

A identificação dos núcleos celulares em imagens de tecido epitelial do colo uterino por meio de técnicas de segmentação configura-se como a primeira etapa da análise estrutural do mesmo. À imagem binária resultante desse processo, contendo a posição relativa dos núcleos, aplica-se o método de análise estrutural de tecidos histológicos proposto e apresentado no capítulo 4. As diferenças estruturais presentes nas lesões neoplásicas desse epitélio podem ser quantificadas por meio de métricas extraídas das redes geradas e combinadas em um vetor de características a fim de serem avaliadas. As subseções seguintes apresentam o método proposto para representação e análise automática das lesões intraepiteliais cervicais.

5.4.1. Identificação das camadas do tecido epitelial

A organização estrutural do tecido epitelial é avaliada em cada camada pelo patologista. A camada atingida pela lesão define a malignidade da mesma. Devido a esse fato e à análise da Figura 1.2, verifica-se a adequação de uma abordagem local de extração de atributos, como descrito na seção 4.3. Essa abordagem é caracterizada pela obtenção de atributos a partir de diferentes agrupamentos da rede, em contraste com a abordagem global. Neste caso, os agrupamentos são representados pelas camadas de interesse: basal, intermediária e superficial.

A identificação automática dessas camadas foi efetuada em duas etapas. Na primeira, foram definidas as fronteiras entre elas aplicando-se o algoritmo descrito no Quadro 1 para geração de agrupamentos. Foram avaliadas as relações de adjacência e os critérios de agrupamento que fornecessem uma representação ótima. Na segunda etapa, os agrupamentos identificados foram rotulados. Para isto, foram utilizadas e testadas duas abordagens: não supervisionada e supervisionada.

Abordagem Não Supervisionada

Nesta abordagem foi utilizada a filtragem de agrupamentos por área de ocupação (Seção 4.3.2.1). Assim, os três maiores agrupamentos foram selecionados e ordenados pelos respectivos valores de área média dos triângulos. O agrupamento com menor valor de área média foi rotulado como basal (B), o segundo menor, como intermediário (I) e, o de maior área, como superficial (S). Nota-se que, com essa abordagem, todas as imagens apresentarão as três camadas de interesse.

Abordagem Supervisionada

É esperado que para uma imagem normal, sem lesão, triângulos pequenos sejam encontrados em maior quantidade na camada basal do tecido epitelial do colo uterino, triângulos médios na camada intermediária e, por fim, triângulos grandes na camada superficial. A presença de lesão no tecido tende a gerar redes mais homogêneas quanto ao tipo de triângulo encontrado, como, por exemplo, em lesões de alto grau (NIC-III), nas quais espera-se encontrar uma grande quantidade de triângulos pequenos. Deste modo, foram definidas três classes de triângulos, considerando-se sua área: basal (B), intermediário (I) e superficial (S). Para a obtenção dos intervalos que representassem cada uma dessas classes foi utilizada a abordagem supervisionada (Seção 4.3.2.2). Foi solicitado a um patologista a identificação manual das fronteiras entre as camadas histológicas em um conjunto de imagens normais. Exemplos da identificação realizada por ele podem ser visualizadas na Figura 5.3.

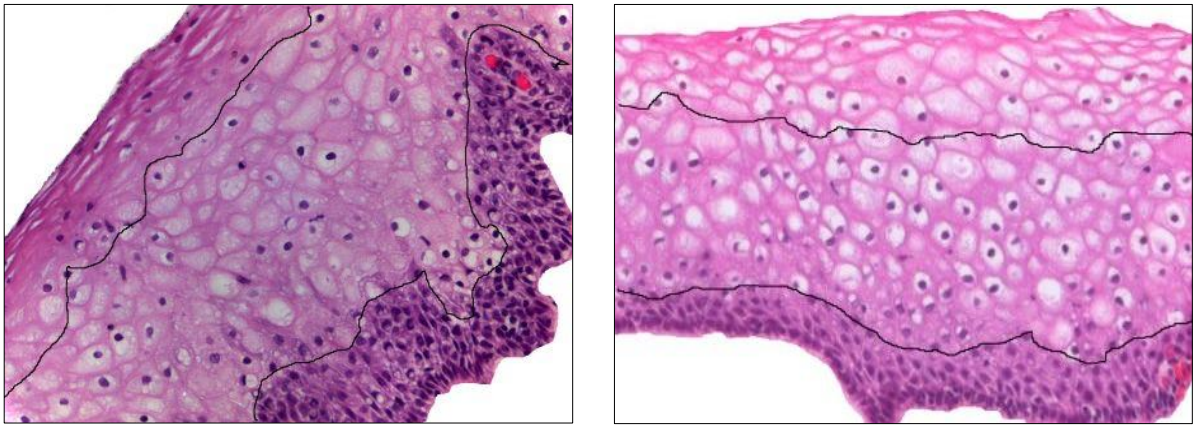


Figura 5.3. As linhas pretas representam a separação manual em camadas do tecido realizada pelo patologista

Após a identificação manual, essas imagens foram separadas em três agrupamentos definidos de acordo com as fronteiras traçadas pelo patologista e uma *TD* foi gerada sobre cada um. Foram calculados os valores de área média dos triângulos para cada agrupamento, representando as camadas basal, intermediária e superficial, para cada imagem do conjunto de treinamento. Esse processo resultou na estimação de 3 valores: área média dos triângulos da camada basal ($\hat{A}_B$), área média dos triângulos da camada intermediária ($\hat{A}_I$) e área média dos triângulos da camada superficial ($\hat{A}_S$). A partir desses valores, o critério para classificação dos triângulos é descrita a seguir:

- $basal, \quad se \quad A_t \leq \frac{(\hat{A}_B+\hat{A}_I)}{2}$
- $intermediário, \quad se \quad \frac{(\hat{A}_B+\hat{A}_I)}{2} \leq A_t \leq \frac{(\hat{A}_I+\hat{A}_S)}{2}$
- $superficial, \quad se \quad A_t \geq \frac{(\hat{A}_I+\hat{A}_S)}{2}$

Sendo A_t a área de um triângulo pertencente à *TD* gerada. Diferente da abordagem não supervisionada, a rotulação dos triângulos pode gerar de um a três agrupamentos.

5.4.2. Geração do vetor de características

A modelagem realizada na etapa anterior permitiu a representação do tecido em camadas por meio da aplicação do algoritmo de agrupamento e do processo de rotulação dos mesmos. A etapa descrita nesta seção caracteriza-se pela extração de medidas das camadas obtidas. As métricas adotadas nessa fase foram escolhidas com base nas diferenças estruturais que os agrupamentos de interesse fornecem, como por exemplo, a taxa de ocupação (TO), caracterizada pela soma das áreas dos triângulos de um determinado agrupamento em relação à soma das áreas dos triângulos de toda a rede; o grau médio (k_{med}) e a entropia com relação ao grau (H). Essas duas últimas métricas foram descritas na seção 2.4.

Experimentos foram realizados considerando-se as métricas isoladamente, representadas pelo seguinte vetor $\vec{X} = [X_B X_I X_S]$, no qual, X representa uma determinada métrica e B, I e S, representam os agrupamentos previamente definidos (basal, intermediário e superficial). Se uma instância apresentar somente agrupamentos basal e intermediário, o vetor passa a ser representado por $\vec{X} = [X_B X_I]$. Além disso, foram consideradas também combinações entre as diferentes métricas, assim como, a exclusão de valores extraídos de um determinado agrupamento com o objetivo de se avaliar a acurácia de classificação das NICs para diferentes conjuntos de atributos e, também, entre agrupamentos (*B*, *I* e *S*).

5.4.3. Projeto de um classificador

A partir dos vetores de características obtidos na modelagem baseada em camadas, foi projetado e implementado um classificador para as NICs. Como descrito anteriormente, de acordo com a abordagem de geração de camadas utilizada, no máximo três agrupamentos podem resultar desse processo. Desta forma, partições foram criadas para atender os diferentes vetores de características.

Baseado no modelo teórico adotado neste trabalho (Figura 1.2), foram identificadas algumas combinações de agrupamentos possíveis de serem geradas dentro do método

proposto. Por exemplo, verificou-se que a existência de vetores que contenham apenas métricas do agrupamento basal seria uma situação aceitável, uma vez que, lesões de alto grau tendem a gerar redes mais homogêneas, com triângulos pequenos. Já em redes obtidas a partir de imagens normais, é esperado que os três agrupamentos sejam identificados, e, em redes obtidas a partir de lesões NIC-I ou NIC-II, é esperado que apresentem um número intermediário de agrupamentos. No entanto, nas redes geradas para esses dois últimos conjuntos, na fase de treinamento, verificou-se que ambas tendem a gerar três agrupamentos, porém com variações bem distintas da taxa de ocupação da rede para cada classe.

Devido a essa variação do número de agrupamentos obtidos por imagem, foram definidas quatro partições do classificador, as quais permitiram que apenas vetores de mesma dimensão fossem comparados entre si. Essa modelagem evitou que o valor zero fosse usado inadequadamente na representação de um atributo ausente. Por exemplo, no cálculo do grau médio, a adoção do valor zero para um agrupamento ausente, indicaria que ele apresenta não apresenta vizinhos, quando, na realidade era impossível calcular esse valor.

Esses conflitos na representação de valores para agrupamentos ausentes foram resolvidos com a utilização das seguintes partições, definidas de acordo com a quantidade e o tipo de agrupamentos obtidos Figura 5.4:

1) identificação somente do *cluster* basal: $\vec{X} = [X_b]$

2) identificação dos *clusters* basal e intermediário: $\vec{X} = [X_b X_i]$

3) identificação dos três *clusters*: $\vec{X} = [X_b X_i X_s]$

4) qualquer vetor $\vec{X}$, distinto dos vetores especificados nos três itens acima

Nesta última partição estão incluídos, por exemplo, vetores que apresentem agrupamentos basais e superficiais $\vec{X} = [X_B X_S]$ ou intermediários e superficiais $\vec{X} = [X_I X_S]$. Esses vetores não são representativos do modelo teórico de evolução das lesões de interesse, pois são consideradas combinações impossíveis de serem representadas em situações reais, e, portanto, são considerados ruídos e não são classificados.

As partições *1* e *2* levam a classificação NIC-III devido aos exemplos encontrados durante a etapa de treinamento do classificador e que representaram apenas 1,72% do conjunto de exemplos. A grande maioria encontra-se na partição *3*. Para esta partição foi estimado um vetor representativo para cada classe de interesse: Normal, NIC-I, NIC-II e NIC-III, como descrito na seção 4.5. Esse vetor representativo é caracterizado pelo valor esperado

de cada atributo considerado no vetor. Na etapa de testes, cada imagem gerou um vetor de características, o qual foi comparado a cada um dos quatro vetores representativos estimados no treinamento por meio da cálculo da distância entre eles. A classe adotada para o novo exemplo foi a classe que apresentou menor distância entre o seu vetor representativo e o vetor da imagem de referência, representado na Figura 5.4 por $d_{min} = d(\vec{X}_{r_y}, \vec{X})$.

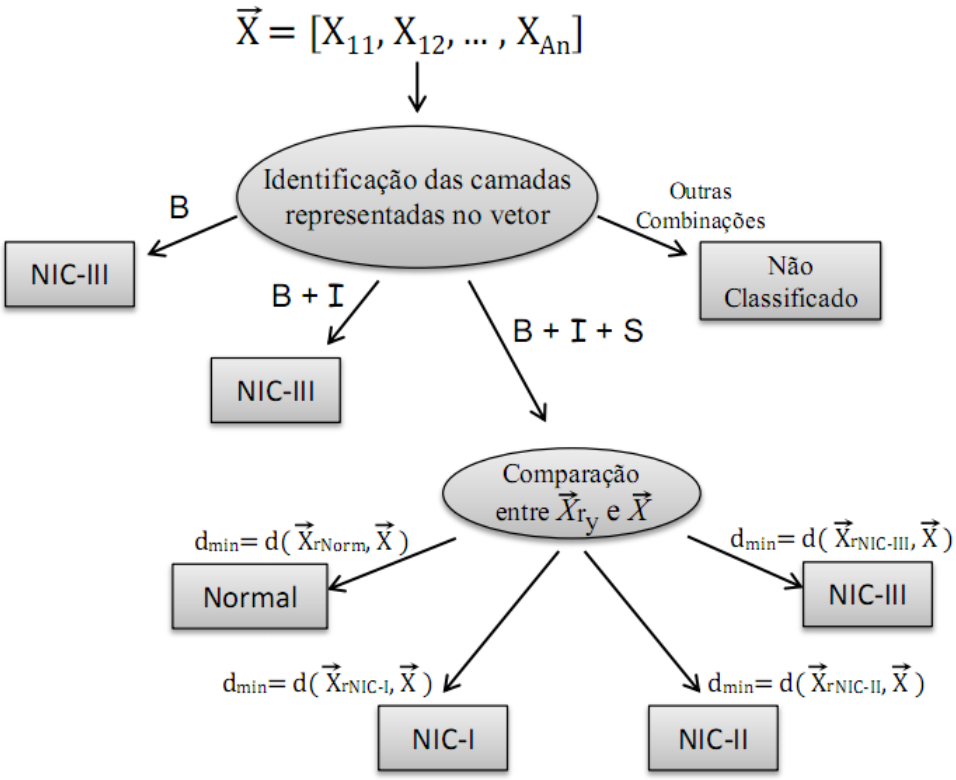


Figura 5.4. Classificador implementado e suas respectivas partições

5.5. Avaliação dos Resultados

Para analisar a eficácia dos resultados obtidos na etapa de segmentação, foram comparadas as imagens resultantes do processamento automático com as imagens obtidas a partir da segmentação manual realizada por um especialista da área. Essa comparação foi obtida calculando-se a sobreposição entre imagens dos núcleos segmentados pelos processos automático e manual. Para este cálculo, são contabilizados os pixels identificados como pertencentes ao núcleo pelos dois processos de segmentação para uma mesma imagem. A razão entre esses números fornece a taxa de sobreposição. Uma taxa de sobreposição média foi obtida após o processamento de imagens a partir de um conjunto de treinamento.

Para avaliar os atributos utilizados e a classificação das NICs foram utilizadas medidas de acurácia (27), sensitividade (28), especificidade (29) e curvas ROC (*Receiver*

Operating Characteristic) [Metz-2000]. A acurácia representa a fração de imagens classificadas corretamente. A sensitividade é caracterizada pela frequência com que o algoritmo identifica corretamente a existência de uma anomalia e a especificidade, pela frequência com que o algoritmo identifica corretamente a ausência de anomalia.

$$acurácia = \frac{VP+VN}{VP+FN+VN+FP} \qquad (27)$$

$$sensitividade = \frac{VP}{VP+FN} \qquad (28)$$

$$especificidade = \frac{VN}{VN+FP} \qquad (29)$$

Nas quais,

- VP (fração de verdadeiros positivos): número de exemplos que apresentam anomalias e que algoritmo classificou corretamente;
- FN (fração de falsos negativos): número de exemplos que apresentam anomalias e que o algoritmo não classificou corretamente;
- VN (fração de verdadeiros negativos): número de exemplos que não apresentam anomalias e que o algoritmo classificou corretamente;
- FP (fração de falsos positivos): número de exemplos que não apresentam anomalias e que o algoritmo não classificou corretamente.

Para avaliar o compromisso entre sensitividade e especificidade foi utilizado o Coeficiente de Correlação de Matthews (MCC - *Matthews Correlation Coefficient*), calculado através da fórmula:

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}} \qquad (30)$$

O MCC é um coeficiente de correlação entre as classificações observadas e as preditas, quer varia no intervalo [-1,1]. Quanto mais próximo de 1, maior o número de predições corretas. O valor 0 indica que o número de predições corretas é praticamente aleatório e o valor -1 indica que as classificações preditas são totalmente diferentes das observadas. Similar à acurácia, o MCC considera também considera os valores de VP e VN no cálculo de predições corretas, porém este último não está sujeito ao viés do tamanho

amostral, que pode ocorrer quando as dimensões dos conjuntos de treinamento são diferentes para diferentes classes [Baldi-2000].

Diferentes etapas do método proposto foram acompanhadas por especialistas da área possibilitando a validação de várias etapas do projeto. Como, por exemplo, durante o levantamento de parâmetros perceptuais, na distinção das anomalias de interesse, permitindo a caracterização morfológica e estrutural das lesões. A etapa de aquisição de imagens também contou com a participação de especialistas. As imagens capturadas foram avaliadas por um patologista quanto à qualidade e representatividade das lesões de interesse. Esse processo auxiliou a filtragem de ruídos presentes na base de imagens. A definição dos limiares de área para caracterização dos agrupamentos basal, intermediário e superficial, na etapa de extração de atributos também foi auxiliada pelo patologista, por meio da delimitação manual das camadas do tecido epitelial nas imagens de interesse. Esses limiares foram utilizados na fase de treinamento para estimação de parâmetros do modelo.

Em todos os experimentos realizados foi utilizado o processo de validação cruzada que consiste no particionamento do conjunto de dados em *n* subconjuntos, de mesma dimensão e sem intersecção de seus elementos. Alguns desses subconjuntos são utilizados na fase de treinamento para estimação dos parâmetros de interesse e os outros na fase de teste, para avaliar o desempenho do método utilizado. Esse processo é repetido alternando-se os subconjuntos de treinamento e de teste e valores médios de desempenho podem ser obtidos.

5.6. Considerações Finais

Neste capítulo foi apresentado o método proposto para caracterização e avaliação automática de imagens provenientes de biópsias do colo uterino a fim de se identificar as alterações graduais causadas pelas NICs no tecido epitelial.

6. RESULTADOS E DISCUSSÃO

6.1. Considerações Iniciais

O capítulo atual apresenta os experimentos realizados e os resultados obtidos ao longo do processo de geração automática das NICs, empregando-se o método descrito no capítulo 5. Primeiro, são descritos os resultados referentes aos processos de aquisição das imagens de interesse e segmentação. Neste último, foram comparados os resultados obtidos por meio de técnicas que utilizam cor e por meio de técnicas baseadas na morfologia dos componentes da imagem. Posteriormente, são descritos os resultados obtidos a partir da caracterização topológica do tecido epitelial do colo uterino.

6.2. Aquisição das Imagens Histopatológicas

De acordo com Junqueira, a preparação de tecidos biológicos para análise microscópica envolve três etapas [Junqueira-2004]: a) Fixação, que consiste no tratamento das amostras com substâncias químicas ou físicas, as quais evitam que os tecidos sejam digeridos por enzimas intracelulares ou por bácterias; b) Inclusão, etapa na qual os tecidos são infiltrados com substâncias que proporcionem rigidez às amostras a fim de serem seccionadas pelo micrótomo; c) Coloração, caracterizada pela utilização de corantes, a fim de facilitar a sua visualização, uma vez que a grande maioria dos tecidos biológicos é incolor. Posteriormente, a amostra é seccionada pelo micrótomo e disposta em lâminas. Com uma câmera fotográfica acoplada a um microscópio é possível digitalizar regiões de interesse em diferentes ampliações. No presente trabalho, a aquisição de imagens histopatológicas pode ser descrita em três etapas:

a) Seleção de amostras de interesse: Junto ao Serviço de Patologia (SERPAT) do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto - USP foi realizada uma pesquisa no arquivo de biópsias a fim de selecionar exames de colo de útero que apresentassem diagnósticos de interesse. Posteriormente, foram adquiridas lâminas previamente fixadas com H&E contendo amostras de tecido referentes aos exames selecionados, para posterior identificação de regiões de tecido epitelial.

b) Digitalização: A partir dessas amostras, foram obtidas as imagens digitalizadas. O sistema de captura utilizado foi disponibilizado pelo Laboratório de Neurologia Aplicada e Experimental do Departamento de Neurologia da FMRP-USP e consiste de um microscópio Óptico *Zeiss Axiophot*, equipado com uma câmera CCD *AxioCam-MRC*, controlada pelo software *AxioVision* versão 4.8. As imagens capturadas foram padronizadas com um aumento de 320x e apresentam resolução de 1388 *x* 1040 pixels.

c) Validação e Classificação: Posteriormente, as imagens obtidas foram avaliadas junto a um patologista. Essa avaliação consistiu na eliminação de exemplos que não apresentassem qualidade suficiente para serem classificadas ou que não permitissem uma visualização de todas as camadas do tecido epitelial. Deste modo, foi realizada uma filtragem inicial nos dados para eliminação de ruídos. O especialista também determinou a classificação NIC de cada imagem selecionada.

Além da aquisição das imagens, a pesquisa aos arquivos do SERPAT permitiu a coleta de informações adicionais para cada exame de interesse, tais como, a data de ocorrência, a classificação NIC correspondente ao exame, um registro que caracteriza a paciente que realizou o exame, entre outras informações[1].

6.3. Base de Imagens

Com o objetivo de facilitar a recuperação das informações obtidas, assim como o acesso às imagens, um banco de dados foi modelado e implementado. Por meio deste, podem ser recuperadas, por exemplo, as imagens provenientes de exames de uma mesma paciente. Essa base também armazena as estruturas de dados resultantes do processamento realizado sobre as imagens, tais como as informações extraídas sobre as redes geradas: quantidade de nós, a relação de nós vizinhos, pesos das arestas, etc. O Diagrama Entidade Relacionamento (DER) do banco pode ser visualizado na Figura 6.1. Atualmente, a base de imagens contém um total de 166 imagens sendo 45 representativas de NIC-I, 39 de NIC-II, 32 de NIC-III e 50 normais.

[1] O material obtido no SERPAT não contém dados de identificação de pacientes, sendo resultado de um estudo retrospectivo. A metodologia para aquisição de dados neste trabalho foi aprovada pelo Comitê de Ética em Pesquisa HCRP / FMRP-USP sob o processo nº 11761/2010 (Anexo A).

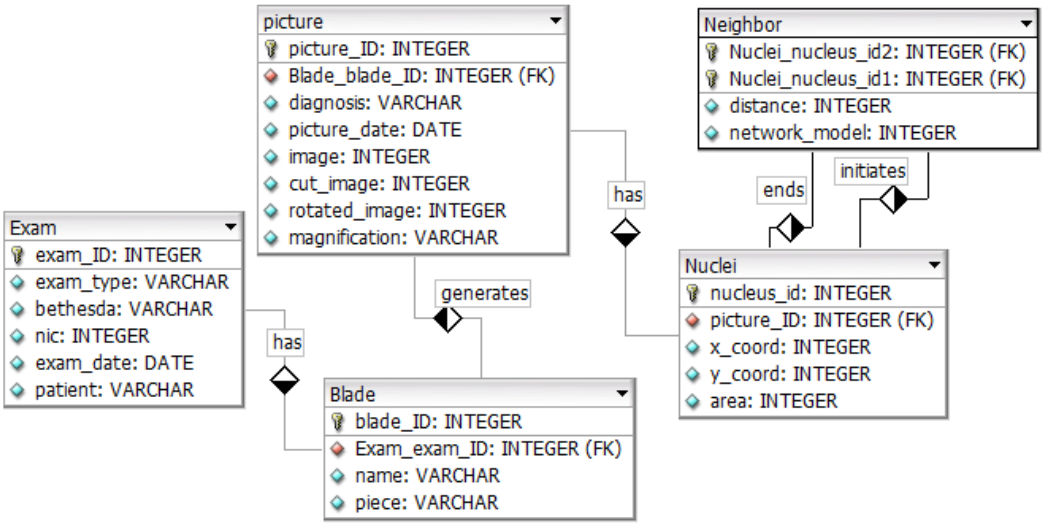


Figura 6.1. Diagrama Entidade Relacionamento da base de imagens criada.

6.4. Ambiente de Desenvolvimento

Para implementação das tarefas propostas e realização de testes foi utilizada a linguagem *C++* e como ambiente integrado de desenvolvimento foi escolhido o *Microsoft Visual C++*[2]. O *PostgreSQL*[3] foi utilizado como Sistema Gerenciador de Banco de Dados.

Além dessas ferramentas, duas bibliotecas estão sendo utilizadas:

- **OpenCV** (*Open Computer Vision Library*)[4] - É uma biblioteca escrita em C/C++ que possui uma coleção de algoritmos para o desenvolvimento de aplicações de processamento de imagens e visão computacional. Foi criada em 2000 por um grupo de desenvolvimento da Intel® e atualmente vem sendo bastante utilizada nesta área.
- **SDC Morphology Toolbox**[5]: Esta biblioteca apresenta uma coleção de algoritmos amplamente utilizados na área de Morfologia Matemática aplicados a imagens em níveis de cinza. Esses algoritmos podem ser utilizados na segmentação, filtragem e análise de imagens.

6.5. Segmentação dos Núcleos

Como descrito anteriormente, os núcleos celulares nas imagens histopatológicas representam um dos principais atributos de análise do patologista. Sendo assim, o objetivo da

[2] http://www.microsoft.com/express/Downloads/
[3] http://www.postgresql.org.br/
[4] http://opencv.willowgarage.com/wiki/
[5] http://www.mmorph.com/

etapa de segmentação deste trabalho é a identificação dos mesmos nas imagens adquiridas. Para isto, foi aplicado o *pipeline* de operadores morfológicos descrito na seção 5.3. Além disso, foram aplicadas e comparadas duas técnicas de segmentação baseadas em cor. As sub-seções seguintes apresentam os resultados.

6.5.1. Aplicação do *pipeline* de operadores morfológicos

A Figura 6.2 apresenta os resultados obtidos a partir da aplicação do *pipeline* de operadores morfológicos, detalhado na seção 5.3, a uma imagem que apresenta um exemplo de tecido epitelial normal (Figura 6.2a). A Figura 6.2b apresenta a mesma imagem em níveis de cinza e a Figura 6.2c apresenta o efeito da aplicação do operador *close top-hat*, por reconstrução, realçando as diferenças entre os núcleos (agrupamentos de pixels mais escuros) e o citoplasma e a matriz extracelular (agrupamentos de pixels mais claros). Na Figura 6.2d é possível visualizar os marcadores internos de cada núcleo resultantes da aplicação dos operadores de abertura, fechamento e *area open*. Cada núcleo possui apenas um marcador, sendo este um componente conexo. Essa imagem foi limiarizada antes da aplicação desses operadores. A Figura 6.2e apresenta o Diagrama de *Voronoi* gerado a partir dos marcadores internos. As fronteiras definidas nessa etapa caracterizam os marcadores externos, que também foram utilizados na aplicação da *Watershed*. Esses últimos limitam a área de crescimento de cada região conexa. Por fim, a Figura 6.2f apresenta o resultado final da aplicação da *Watershed*, utilizando-se a Figura 6.2e como marcador.

O processo descrito acima pode ser visualizado também nas Figuras 6.3, 6.4 e 6.5 que apresentam exemplos de imagens NIC-I, NIC-II e NIC-III, respectivamente.

6.5.2. Técnicas baseadas em cor

Para imagens coradas com H&E, quando não consideradas as variações de brilho, basicamente, podemos identificar três cores: violeta, rosa e branco. De maneira geral, as técnicas baseadas em cor levam em consideração a proximidade de cada pixel da imagem a essas três cores de referência.

Em trabalhos relacionados [Landini-2004] [Landini-2009], a técnica de *color deconvolution* mostrou-se adequada ao processamento automático de imagens histopatológicas. Devido a esse fato, essa técnica foi aplicada às imagens de interesse a fim de se avaliar sua adequação à segmentação dos núcleos celulares. Foi utilizado um *plugin* para

ImageJ[6], desenvolvido e disponibilizado por Landini[7]. Para resultados mais precisos, é necessário a aplicação de algoritmos de subtração de fundo e correção de iluminação em imagens microscópicas.

A Figura 6.6 apresenta as imagens resultantes da aplicação dessa técnica sobre uma imagem que apresenta uma lesão NIC-II. A imagem original pode ser visualizada na Figura 6.6a e a imagem corrigida, na Figura 6.6b. As Figuras 6.6c e 6.6d correspondem às imagens geradas pela técnica, separando hematoxilina e eosina, respectivamente. As Figuras 6.6e e 6.6f correspondem às imagens geradas para hematoxilina e para eosina, por meio da técnica *k-means clustering*[8], para k = 3. De maneira similar à técnica anterior, as características imuno-histoquímicas foram utilizadas a fim de se obter a separação entre as estruturas coradas predominantemente por cada um desses corantes. A técnica de *color deconvolution* apresentou melhores resultados quanto à separação dos componentes corados com hematoxilina, facilitando a obtenção de marcadores para cada núcleo da imagem.

Apesar da aplicabilidade das técnicas baseadas em cor, sob análise do sistema RGB, as imagens coradas por H&E apresentam variações de intensidade para a mesma cor, com predominância dos canais R e B na mesma proporção. Devido a esse fato, a utilização dessas imagens em níveis de cinza também mostra-se adequada a este contexto. Além disso, as abordagens baseadas em cor fornecem apenas uma separação inicial dos elementos corados predominantemente por cada um dos corantes utilizados. No caso da imagem resultante relativa à hematoxilina, uma filtragem adicional é necessária para a separação dos núcleos. Essa filtragem poderia ser efetuada, por exemplo, por meio de operadores morfológicos para identificação de marcadores internos e posterior aplicação da *watershed*. Desta forma, verifica-se que os resultados obtidos com as técnicas baseadas em cor são equivalentes à aplicação do operador *top-hat* por reconstrução nas imagens em níveis de cinza.

Por fim, o uso da cor torna o método de análise extremamente dependente do processo de aquisição das lâminas, uma vez que se observam grandes variações entre as lâminas na concentração dos fixadores, fazendo com que a informação de cor nas imagens resultantes se torne de difícil controle, além de estar desvinculada do problema em estudo. Portanto, optou-se pelo processamento em níveis de cinza das imagens de interesse adotando-se a abordagem descrita na seção 5.3.

---

[6] http://rsbweb.nih.gov/ij/
[7] http://www.dentistry.bham.ac.uk/landinig/software/software.html
[8] http://www.mathworks.com/products/image/demos.html?file=/products/demos/shipping/images/ipexhistology.html

Região normal do tecido epitelial do colo uterino

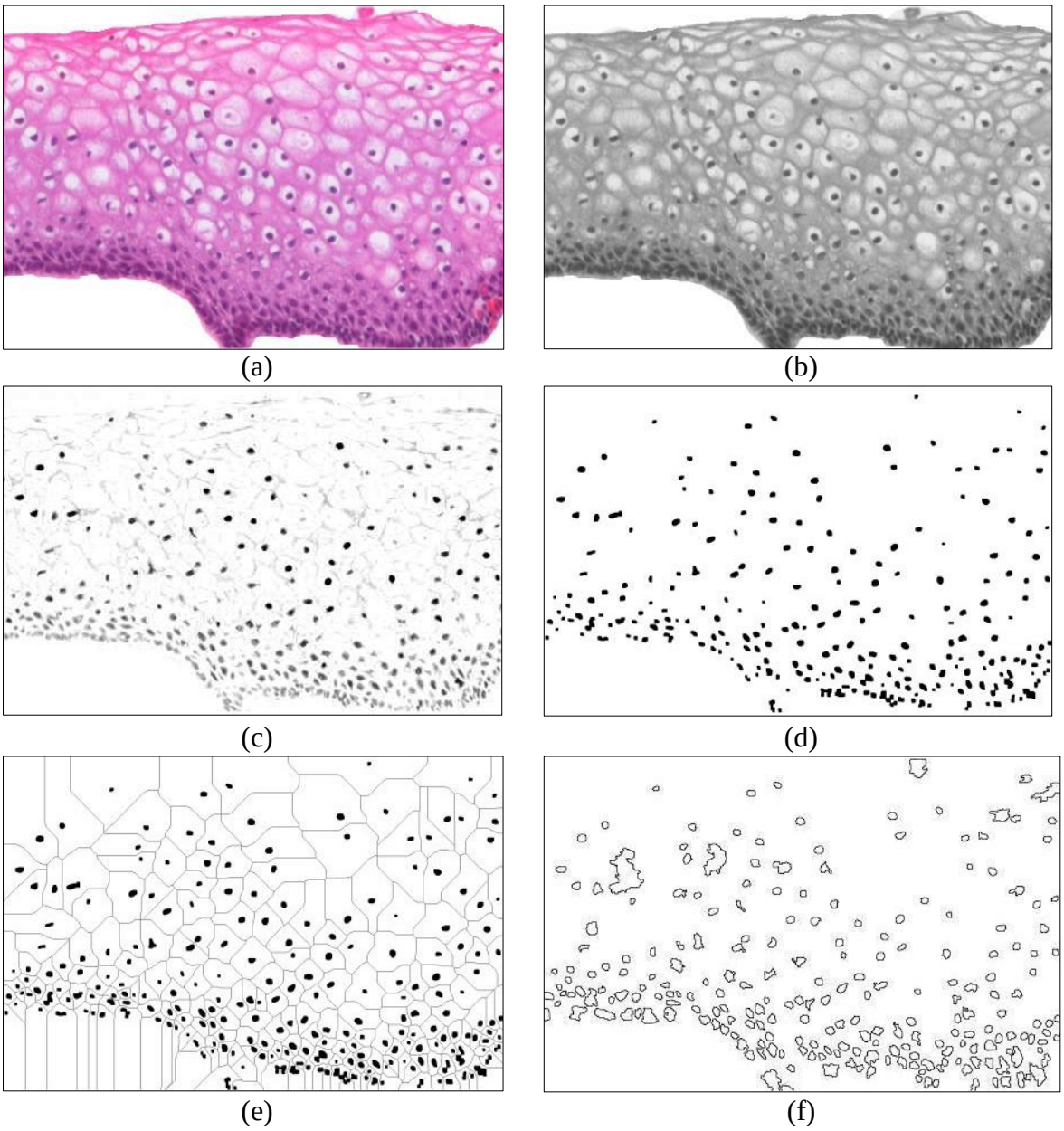


Figura 6.2. Etapas do processo de segmentação para uma imagem sem lesão: (a) Imagem de referência. (b) Imagem de referência em níveis de cinza. (c) Resultado da aplicação do operador *close-by-reconstruction top-hat*. (d) Imagem limiarizada após a eliminação de ruídos e da aplicação do operador *area open*. (e) União dos marcadores internos e externos. (f) Aplicação da *watershed*: contornos dos núcleos celulares

Resultado do processo de segmentação para uma imagem com NIC-I

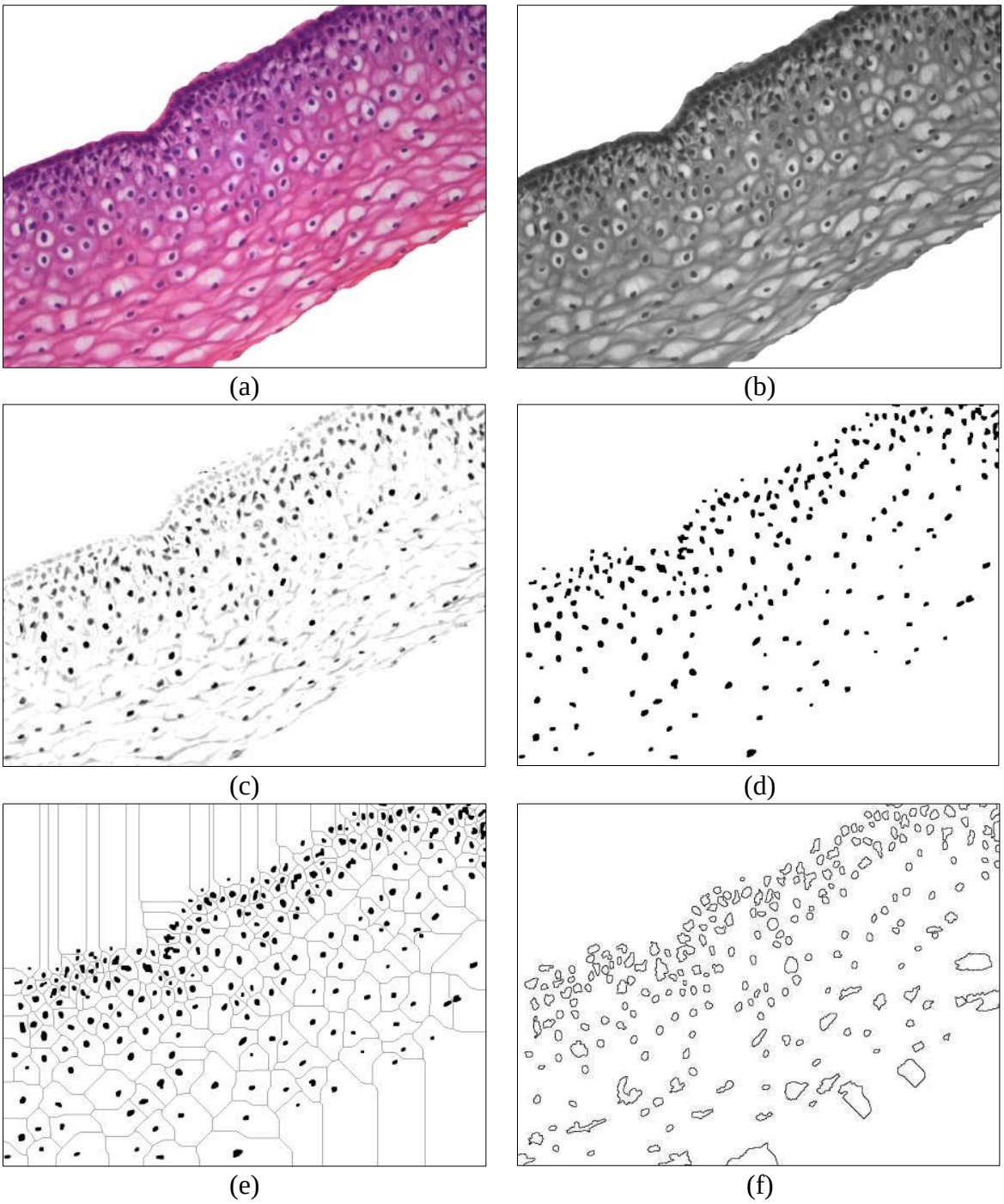


Figura 6.3. Etapas do processo de segmentação para uma imagem que apresenta NIC-I: (a) Imagem de referência. (b) Imagem de referência em níveis de cinza. (c) Resultado da aplicação do operador *close-by-reconstruction top-hat.* (d) Imagem limiarizada após a eliminação de ruídos e da aplicação do operador *area open*. (e) União dos marcadores internos e externos. (f) Aplicação da *watershed*:contornos dos núcleos celulares

Resultado do processo de segmentação para uma imagem com NIC-II

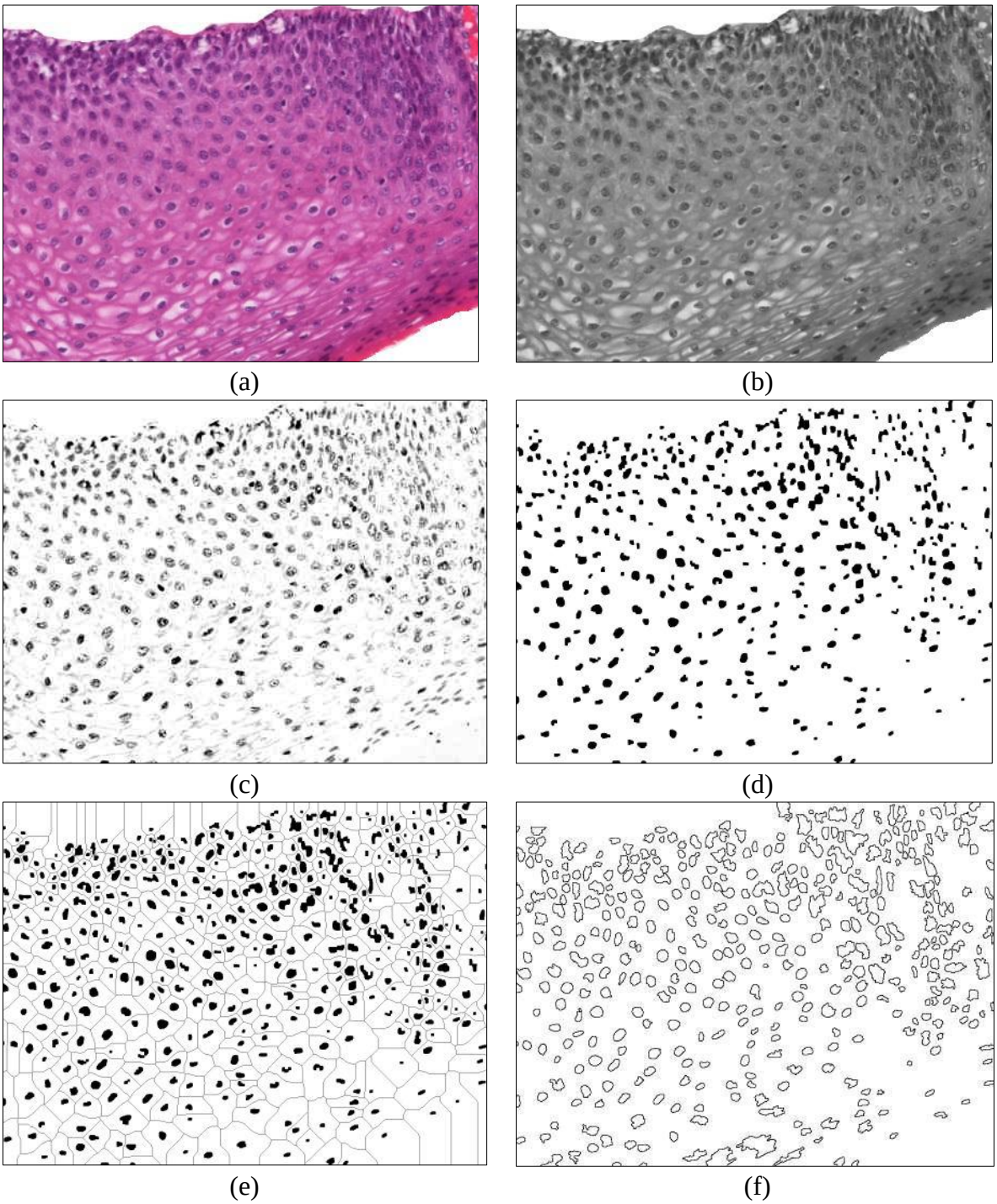


Figura 6.4. Etapas do processo de segmentação para uma imagem que apresenta NIC-II: (a) Imagem de referência. (b) Imagem de referência em níveis de cinza. (c) Resultado da aplicação do operador *close-by-reconstruction top-hat*. (d) Imagem limiarizada após a eliminação de ruídos e da aplicação do operador *area open*. (e) União dos marcadores internos e externos. (f) Aplicação da *watershed*: contornos dos núcleos celulares

Resultado do processo de segmentação para uma imagem com NIC-III

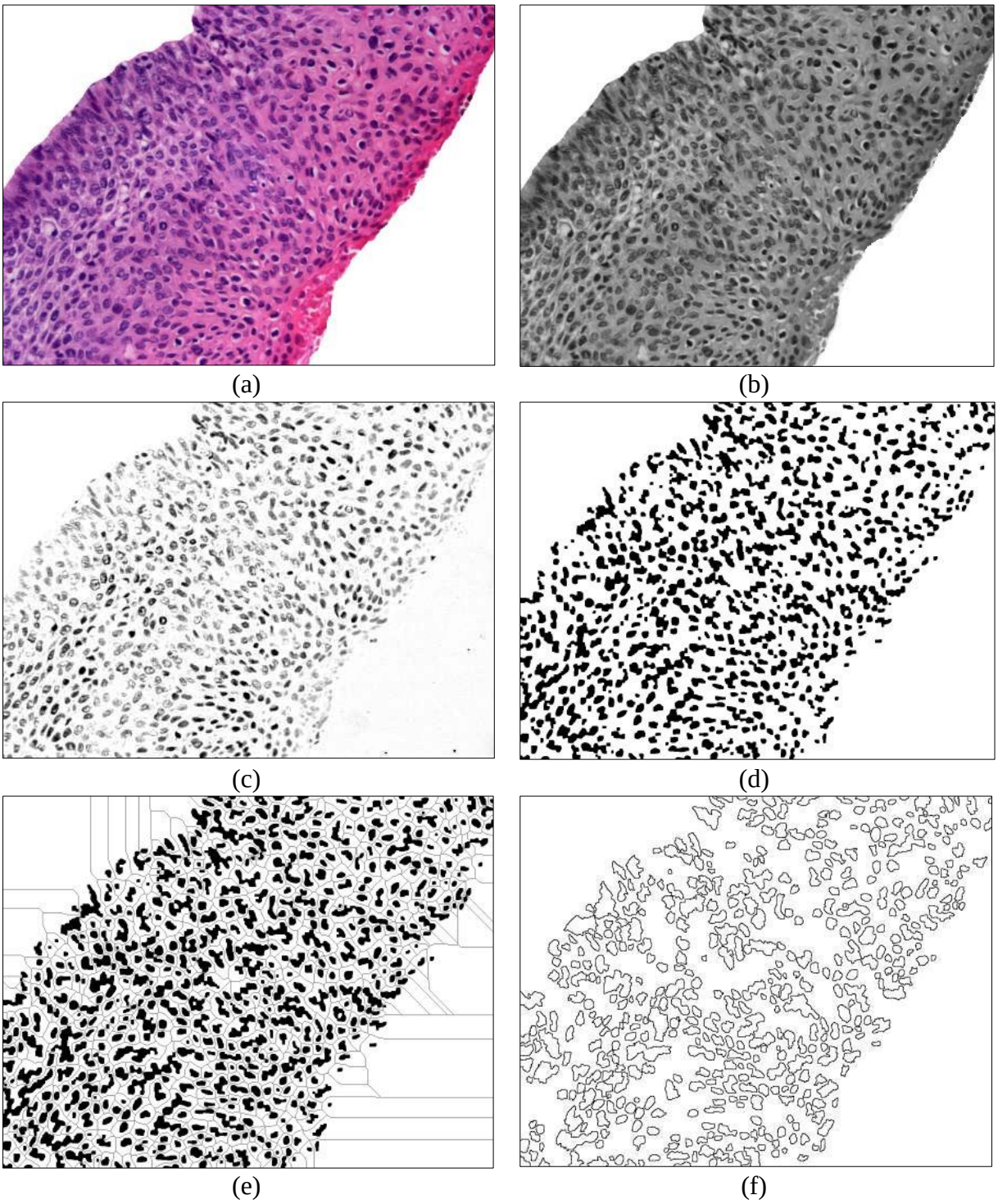


Figura 6.5. Etapas do processo de segmentação para uma imagem que apresenta NIC-III: (a) Imagem de referência. (b) Imagem de referência em níveis de cinza. (c) Resultado da aplicação do operador *close-by-reconstruction top-hat.* (d) Imagem limiarizada após a eliminação de ruídos e da aplicação do operador *area open*. (e) União dos marcadores internos e externos. (f) Aplicação da *watershed*: contornos dos núcleos celulares

Comparação entre as técnicas *color deconvolution* e *k-means clustering*

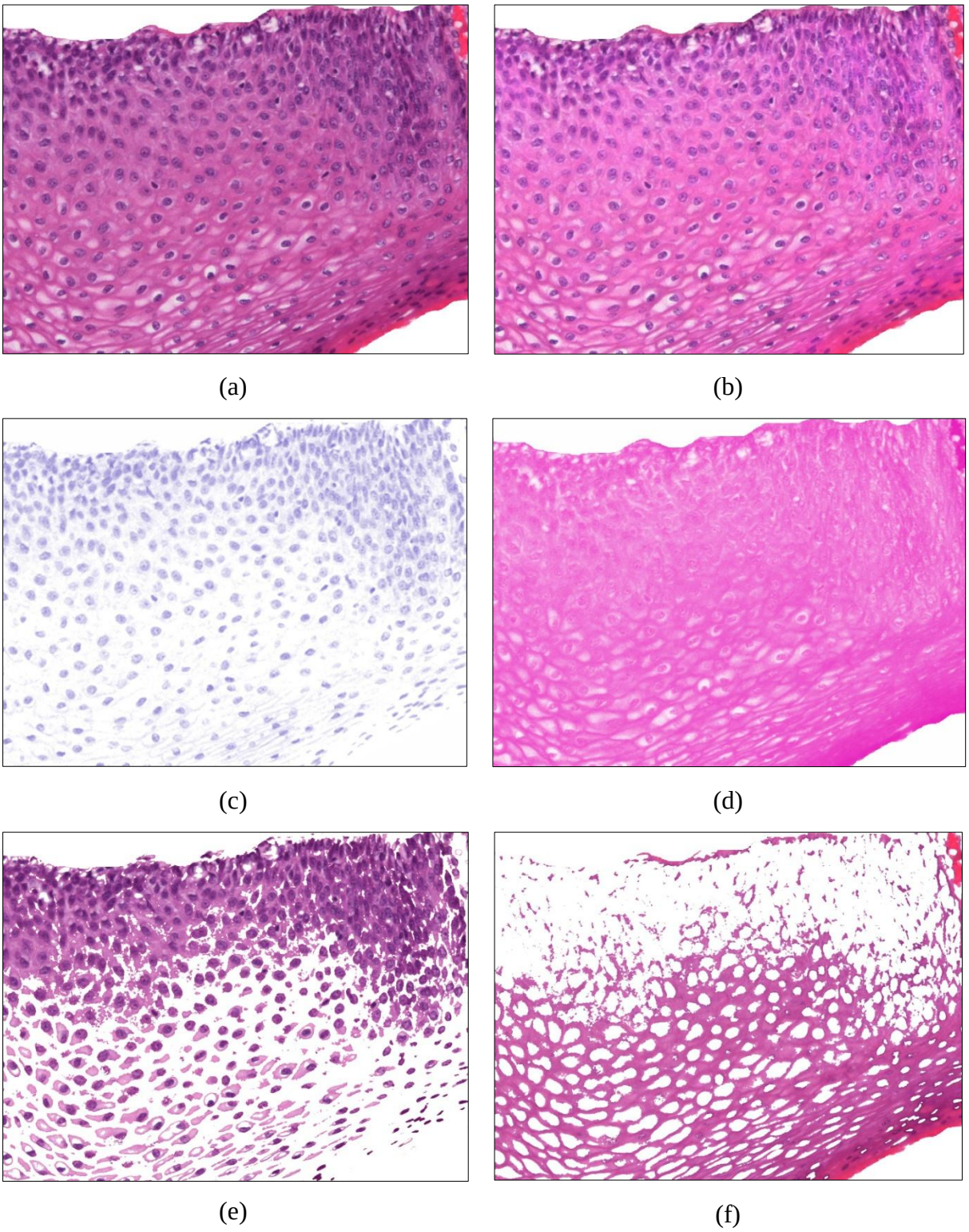


Figura 6.6. Resultados do processo de segmentação utilizando-se técnicas baseadas em cor. (a) Imagem original. (b) Imagem após a aplicação do algoritmo de subtração de fundo (c) e (d) Imagens relativas à hematoxilina e à eosina geradas pela técnica de *Color Deconvolution*. (e) e (f) Imagens relativas à hematoxilina e à eosina geradas pela técnica *k-means*

6.5.3. Avaliação da segmentação automática

A acurácia do processo de segmentação automática utilizando a abordagem morfológica foi avaliada por meio da comparação com a segmentação obtida manualmente. Foram utilizadas 3 imagens segmentadas por especialistas da área. A taxa de sobreposição média para essas imagens foi calculada de acordo com o procedimento descrito na seção 5.5, sendo que o valor médio obtido foi de 78%. Cada imagem analisada continha aproximadamente 300 núcleos. A Figura 6.7 apresenta uma comparação qualitativa da segmentação manual do especialista (contornos pretos) e da segmentação automática (contornos vermelhos).

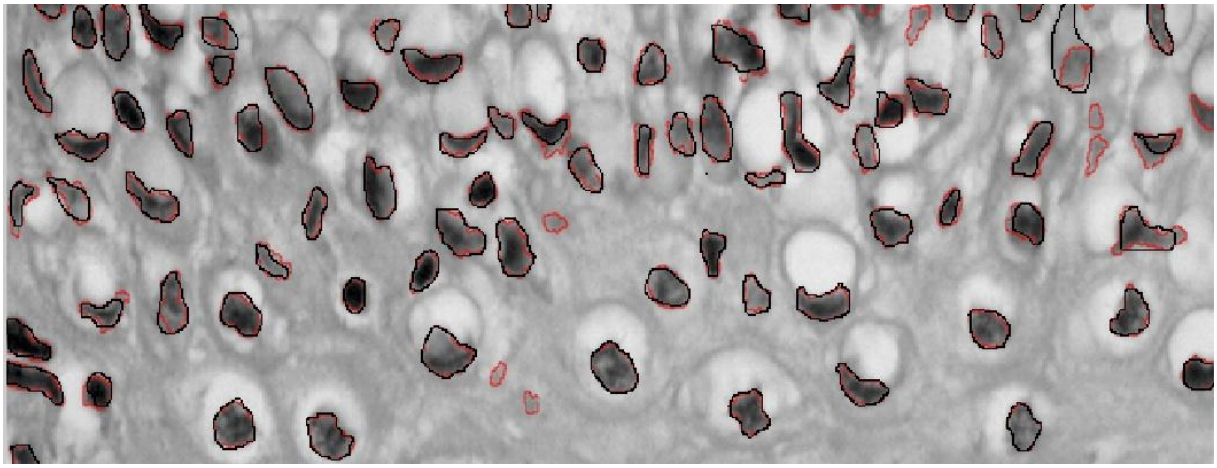


Figura 6.7. Comparação da segmentação manual (contornos pretos) com a segmentação automática (contornos vermelhos).

Apesar do método proposto não utilizar atributos morfológicos para geração do vetor de características, tais como contorno ou área nuclear, a comparação entre a segmentação automática e a manual permitiu uma avaliação quantitativa do desempenho da abordagem adotada na identificação correta dos núcleos celulares. Os centróides referentes aos núcleos segmentados foram utilizados no processo de geração da Triangulação de *Delaunay* para cada imagem da base.

6.6. Avaliação da representação dos tecidos por meio de métricas globais

Esta seção apresenta os resultados obtidos na caracterização estrutural dos tecidos de interesse a partir da extração de métricas globais, ou seja, considerando-se a totalidade da imagem (todo o conjunto de nós e arestas gerados), sem a divisão da mesma em agrupamentos representativos das camadas.

6.6.1. Avaliação dos diferentes modelos de grafos de vizinhança

A fim de se comparar a acurácia de classificação entre NICs, foi realizado um estudo preliminar com os três modelos de rede descritos na seção 2.3: TD, GG e RVR. A Figura 6.8 apresenta uma comparação visual dos diferentes modelos para cada uma das classes em análise.

Esses modelos foram implementados, gerando-se uma representação do tecido para cada um, e então, as seguintes métricas foram extraídas das redes obtidas: grau médio (k_{med}), distância média (l), coeficiente de *cluster* (cc), entropia (H) e correlação de *Pearson* (r). Desta forma, cada imagem passou a ser representada pelo seguinte vetor de características:

$$\vec{X} = [X_1\ X_2\ X_3\ X_4\ X_5]^t = [k_{med}\ l\ cc\ H\ r]^t$$

A Tabela 2 apresenta a acurácia (*ac*) na classificação das NICs para cada modelo de rede e para três classificadores distintos, assim como o erro quadrático médio (*eqm*). Para este experimento, foram utilizadas 92 imagens, contendo regiões de NIC-I, NIC-II, NIC-III e de imagens normais. A avaliação da acurácia foi implementada por meio de validação cruzada (10-*fold*) sobre o conjunto de imagens, portanto o valor exibido na tabela é um valor médio de acurácia. Os classificadores utilizados nesta etapa foram: uma árvore de decisão (J48), uma rede neural (*Multilayer Perceptron* - MLP) e um classificador bayesiano (*Naive Bayes* - NB) no ambiente *Weka*[9].

Tabela 2. Comparação entre diferentes modelos de redes.

Modelo de Rede	***J48***		***MLP***		***NB***	
	ac	***eqm***	***ac***	***eqm***	***ac***	***eqm***
TD	**0,39**	0,51	**0,40**	0,45	0,40	0,44
GG	**0,39**	0,49	0,33	0,45	0,30	0,45
RVR	0,33	0,50	0,37	0,43	**0,44**	0,41

Verificou-se que a *TD* apresentou melhor ou igual desempenho, em relação aos outros modelos, para os classificadores J48 e MLP. No entanto, todos os valores de acurácia apresentados foram insatisfatórios para a distinção das quatro classes de interesse (Normal, NIC-I, NIC-II e NIC-III).

[9] http://www.cs.waikato.ac.nz/ml/weka/

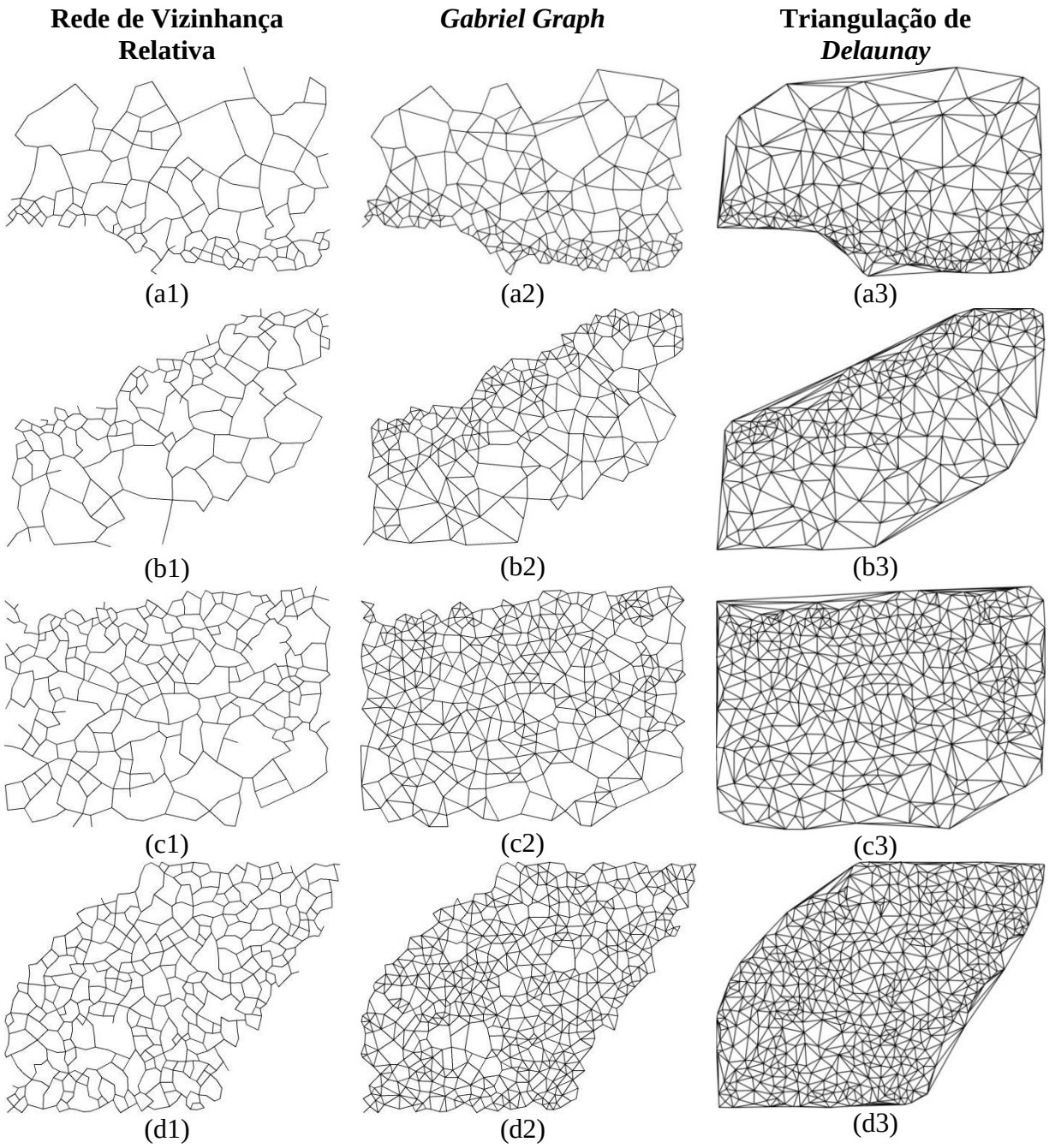


Figura 6.8. Redes geradas para diferentes modelos de grafos de vizinhança a partir dos núcleos segmentados. (a1), (a2) e (a3): Região Normal. (b1), (b2) e (b3): NIC-I. (c1), (c2) e (c3): NIC-II. (d1), (d2) e (d3): NIC-III. Essas redes foram obtidas a partir das seguintes imagens de referência: 6.2a, 6.3a, 6.4a e 6.5a, respectivamente

6.6.2. Avaliação da classificação entre NICs

Foram avaliadas também, a partir da extração de métricas globais, os valores de acurácia na diferenciação entre as classes Normal / NIC-I, NIC-I / NIC-II e NIC-II / NIC-III, que representam a transição gradual entre as lesões.

Inicialmente, foi avaliada a acurácia na detecção da presença de anomalias: Normal / Lesão, sendo a classe Lesão representada por todos os NICs. Os resultados podem ser visualizados na Tabela 3. A TD foi o modelo adotado neste experimento. Os classificadores que forneceram as melhores taxas de acurácia para cada conjunto em análise foram destacados em negrito. A detecção da presença de anomalias apresentou acurácia próxima a 80% com o J48. No entanto, para diferenciação entre NICs, representada pelos outros conjuntos descritos na Tabela 3, foram obtidas taxas menores ou iguais a 70%.

Tabela 3. Taxas de acurácia obtidas na classificação entre NICs

Comparação entre classes	**J48**		**MLP**		**NB**	
	ac	***eqm***	***ac***	***eqm***	***ac***	***eqm***
Normal / NIC-I	**0,64**	0,39	0,50	0,42	0,57	0,42
NIC-I / NIC-II	0,40	0,45	**0,67**	0,38	0,64	0,38
NIC-II / NIC-III	**0,70**	0,35	**0,70**	0,33	0,59	0,38
Normal / Lesão	**0,79**	0,41	0,70	0,47	0,70	0,47

A partir dos resultados apresentados nas Tabelas 2 e 3, verificou-se a possibilidade da utilização da abordagem baseada em regiões, ou seja, fragmentando a imagem em agrupamentos representativos das camadas epiteliais, para extração de atributos em cada agrupamento, a fim de verificar se essa abordagem permite um aumento das taxas de acurácia obtidas na classificação das NICs apresentada nesta seção.

6.7. Avaliação da abordagem baseada em regiões

Esta seção apresenta os resultados obtidos a partir do método proposto na capítulo 5 de análise estrutural das imagens de colo uterino. A primeira etapa consistiu na identificação e rotulação de agrupamentos característicos nas imagens, seguida da geração do vetor de características e, finalmente, da etapa de classificação. As seções seguintes descrevem esse processo.

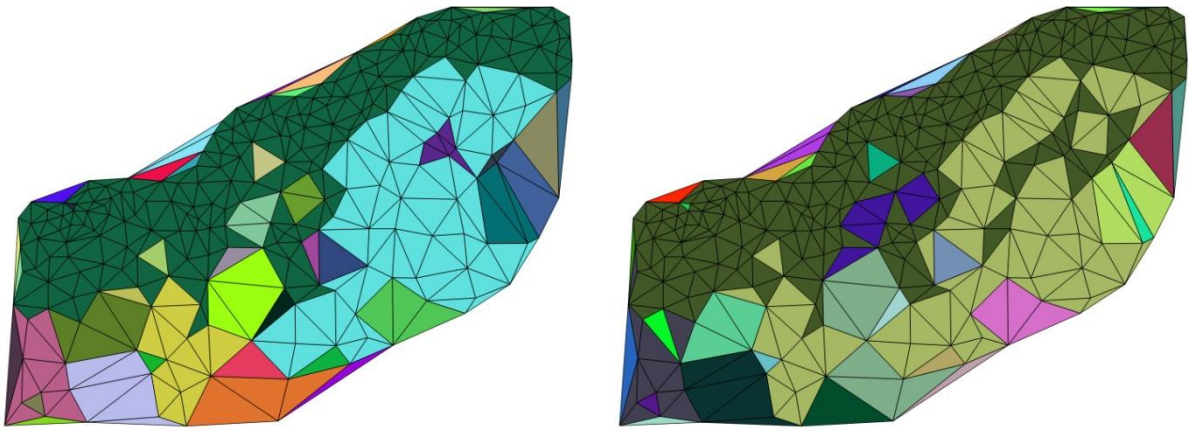
6.7.1. Identificação automática das camadas do tecido epitelial

O tecido epitelial é caracterizado por três camadas distintas que representam agrupamentos celulares com características semelhantes. A primeira etapa do processo de extração local de atributos é a definição desses agrupamentos que representam as fronteiras entre os objetos de interesse na imagem, nesse caso, entre as camadas do tecido epitelial. Como descrito na seção 5.4.1, a identificação automática dessas camadas foi efetuada em

duas etapas: a geração de agrupamentos característicos das regiões de interesse e a rotulação dos mesmos.

6.7.1.1. Geração dos Agrupamentos

Para a identificação dos agrupamentos representativos das camadas do tecido epitelial foi utilizado o algoritmo descrito no Quadro 1. A Figura 6.9 compara os agrupamentos gerados para as diferentes relações de adjacência, mantendo-se o mesmo limiar como critério de agrupamento: $\delta = 0.15$. As cores foram definidas aleatoriamente e cada uma representa um único agrupamento. A RA definida para a Figura 6.9a, por ser mais restritiva, fornece um número maior de agrupamentos. Foram gerados 40 agrupamentos para o exemplo da Figura 6.9a e 30 agrupamentos para o exemplo da Figura 6.9b.



(a) (b)

Figura 6.9. Comparação entre diferentes relações de adjacência. (a) Agrupamentos gerados a partir da RA definida na Figura 4.2 (a), levando-se em consideração arestas em comum. Agrupamentos gerados a partir da RA definida da Figura 4.2 (b), levando-se em consideração vértices em comum

Dependendo do formato do tecido, a TD pode gerar alguns triângulos de borda bastante irregulares. Para evitar esse tipo de artefato, antes da aplicação do algoritmo de agrupamento, os triângulos são filtrados em relação ao ângulo e ao comprimento da aresta. Desta forma, triângulos que possuem ângulos maiores do que θ e pelo menos uma das arestas maiores do que m, são eliminados da TD. Para que apenas triângulos das bordas fossem filtrados, os valores adotados para esses limiares foram: $\theta = 120°$ e $m = 350.0$. Essa filtragem auxiliou também a eliminação de agrupamentos formados nas bordas da rede contendo apenas um ou poucos elementos.

A Figura 6.10 apresenta o resultado da aplicação do algoritmo de agrupamento sobre o grafo gerado a partir de uma imagem normal utilizando-se o critério de proporção entre

áreas e a RA definida na Figura 6.9b para diferentes valores do parâmetro δ_A. Para estes exemplos, os valores 0,15 e 0.3 não forneceram uma boa aproximação das camadas de interesse. Valores de δ_A maiores ou iguais a 0,5, apesar de terem fornecido boas aproximações, são limiares muito altos para se considerar dois triângulos semelhantes levando à formação de agrupamentos menos homogêneos quanto à área dos triângulos.

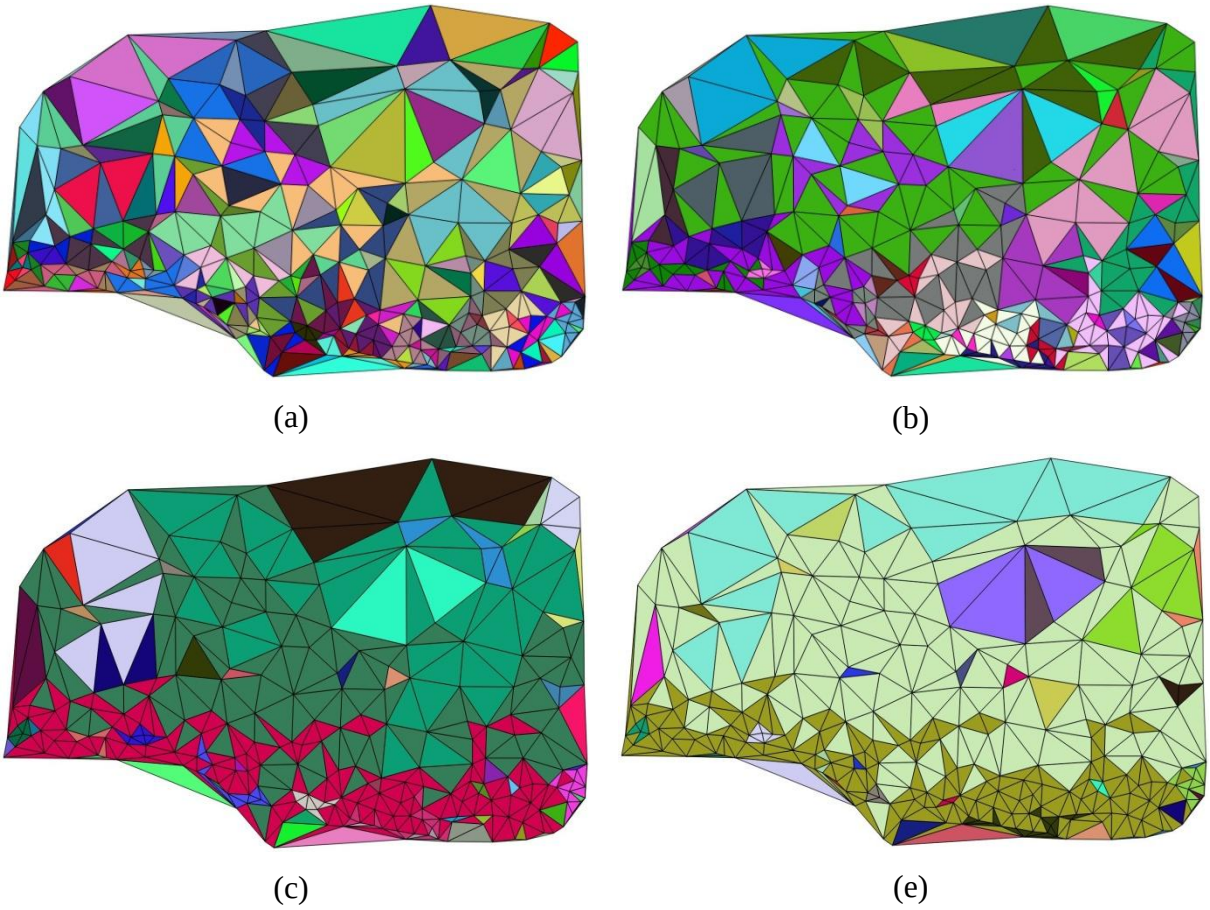


Figura 6.10. Resultado da aplicação do algoritmo de agrupamento descrito no Quadro 1 utilizando a proporção relativa entre áreas como critério de agrupamento. (a) $\boldsymbol{\delta_A}$ = 0,15. (b) $\boldsymbol{\delta_A}$ = 0,30. (c) $\boldsymbol{\delta_A}$ = 0,50. (d) $\boldsymbol{\delta_A}$ = 0,60

De maneira semelhante, a Figura 6.11 também apresenta o resultado da aplicação do algoritmo de agrupamento, porém, utilizando o critério de semelhança entre triângulos (δ_S). Para este exemplo, é possível verificar que os valores de δ_S iguais a 0,2 e 0,3 forneceram boas aproximações das camadas e valores de δ_S menores do que 0,15 praticamente não forneceram estruturas relevantes, gerando um grande número de agrupamentos. Comparando-se os agrupamentos gerados, nota-se ainda que, o critério de semelhança de triângulos fornece agrupamentos mais homogêneos. Já o critério de proporção entre áreas, por considerar medidas absolutas, pode levar ao agrupamento de triângulos bem distintos no seu formato,

desde que suas áreas sejam parecidas. Devido a esse fato o critério de semelhança entre triângulos foi o critério adotado nos outros experimentos descritos nesse capítulo.

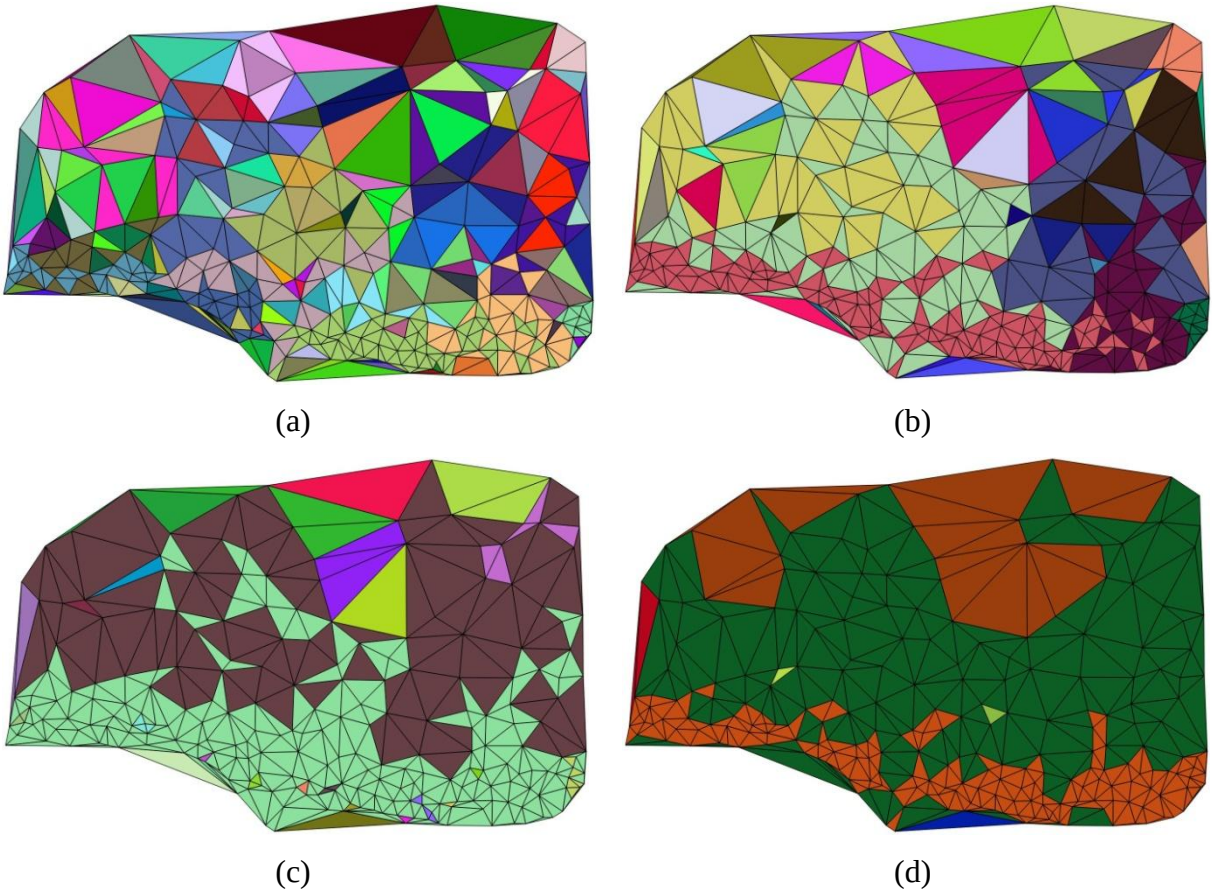


Figura 6.11. Resultado da aplicação do algoritmo de agrupamento descrito no Quadro 1 utilizando a semelhança entre triângulos como critério de agrupamento. (a) $\boldsymbol{\delta_S}$ = 0,05. (b) $\boldsymbol{\delta_S}$ = 0,10. (c) $\boldsymbol{\delta_S}$ = 0,20. (d) $\boldsymbol{\delta_S}$ = 0,30

É interessante notar que, apesar do grande número de agrupamentos que podem ser obtidos para valores de δ muito pequenos, o algoritmo de agrupamento pode ser aplicado novamente considerando-se a média dos agrupamentos obtidos nesta primeira iteração como referência para os agrupamentos seguintes. Verificou-se também a possibilidade de se usar somente o critério de agrupamento na geração das camadas, sem a relação de adjacência. Dessa forma, triângulos semelhantes são agrupados mesmo não sendo adjacentes. Por consequência, um número menor de agrupamentos é gerado na primeira iteração do algoritmo de agrupamento. A Figura 6.12 apresenta os agrupamentos gerados sem considerar a RA entre os triângulos da rede e para δ_S=0,15. O número de agrupamentos resultantes desse processo (dez), é bem menor quando comparado ao número gerado considerando-se as RAs.

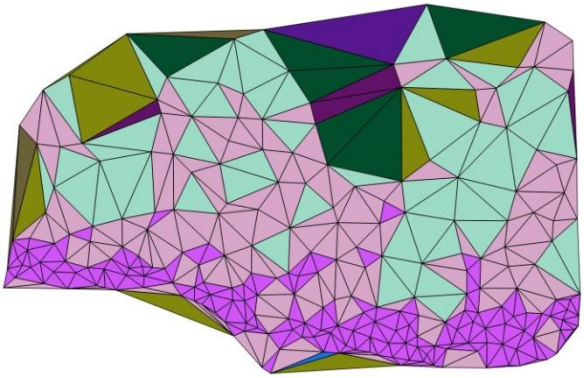


Figura 6.12. Análise dos agrupamentos resultantes obtidos sem a relação de adjacência e considerando-se o critério de semelhança entre triângulos ($\boldsymbol{\delta_S}$) igual a 0,15.

Desta forma, a não utilização da RA pode tornar o processo de identificação das camadas mais rápido e, de maneira equivalente, o limiar δ_S pode controlar a formação dos agrupamentos quanto a semelhança entre os triângulos dos mesmos. O próximo passo é a classificação de cada agrupamento e será descrito na próxima seção.

6.7.1.2. Rotulação dos Agrupamentos

O processo de rotulação dos agrupamentos consiste na classificação dos mesmos em basais, intermediários ou superficiais. Para isto, duas abordagens foram utilizadas e comparadas e são descritas a seguir.

a) Abordagem não supervisionada

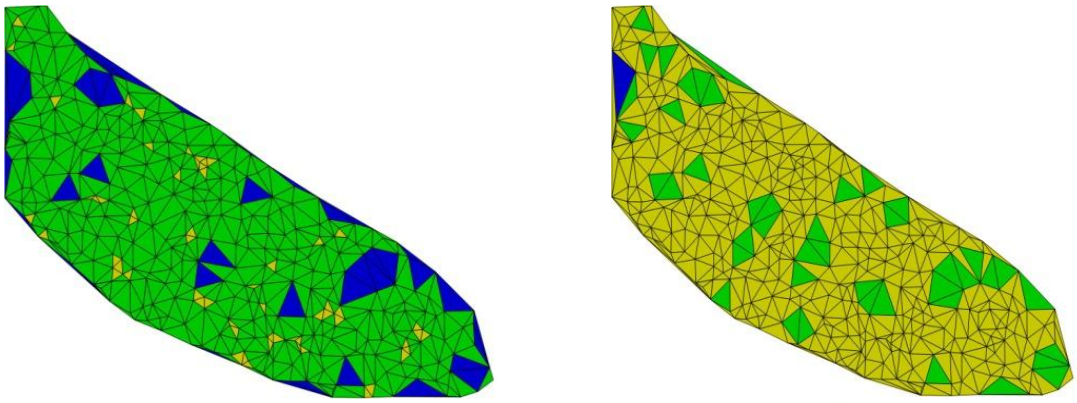
O algoritmo de filtragem por área de ocupação, descrito no Quadro 2, foi utilizado para se reduzir o número de agrupamentos a n, neste caso $n = 3$, devido ao número de camadas do tecido epitelial e, também, para rotulação desses agrupamentos. É importante destacar que, no processo de filtragem, os n agrupamentos de referência são os que apresentam maiores taxas de ocupação (TO). Os agrupamentos remanescentes serão filtrados de acordo com a distância de cada um em relação aos agrupamentos de referência. A rotulação dos agrupamentos resultantes da filtragem foi realizada utilizando-se os valores das áreas médias (A_m) dos mesmos. O agrupamento de menor área média foi classificado como basal, o agrupamento com a segunda menor área média foi classificado como intermediário e, por fim, o agrupamento com a maior área média como superficial.

b) Abordagem supervisionada

Em alguns casos, com a utilização da filtragem por área de ocupação, no entanto, podem ser obtidos agrupamentos de referência com valores de A_m muito próximos levando a

inconsistências nesse tipo de classificação. Por exemplo, para uma imagem que apresenta uma lesão NIC-III é possível que os três agrupamentos de referência, resultantes do processo de filtragem, apresentem valores de A_m próximos e pequenos devido a maior quantidade de células segmentadas, assim, os três devem ser classificados como basais, o que não ocorre. Para evitar esse tipo de inconsistência, foi utilizada a abordagem supervisionada no processo de classificação dos agrupamentos. A partir de uma referência externa, definida com a segmentação manual do patologista e posterior analise quantitativa, foram definidos intervalos de A_m, a partir dos quais, os agrupamentos podem ser classificados.

A Figura 6.13 ilustra a situação descrita no parágrafo anterior e compara os agrupamentos obtidos em cada abordagem para uma mesma imagem de referência com classificação NIC-III. As Figuras 6.13a e 6.13b apresentam os agrupamentos gerados com as abordagens não supervisionada e supervisionada, respectivamente. Os agrupamentos representados pela cor amarela correspondem à camada basal, os agrupamentos representados pela cor verde correspondem à camada intermediária e os agrupamentos representados pela cor azul, à camada superficial. Verifica-se que, na Figura 6.13a, o agrupamento com maior TO é aquele representativo da camada intermediária, já na Figura 6.13b, é aquele representativo da camada basal. Ou seja, essa última figura apresenta distribuição dos agrupamentos mais próxima da situação esperada para uma imagem NIC-III.



(a) (b)

Figura 6.13. Identificação das camadas do tecido epitelial. (a) Agrupamentos obtidos com a abordagem não supervisionada. (b) Agrupamentos obtidos com a abordagem supervisionada

A Figura 6.14 apresenta exemplos de agrupamentos gerados para diferentes imagens utilizando-se a abordagem supervisionada. As imagens 6.2a, 6.3a, 6.4a e 6.5a foram utilizadas como referência para a geração dos agrupamentos apresentados e correspondem a imagens

com as seguintes classificações: Normal, NIC-I, NIC-II e NIC-III, respectivamente. Ainda, por meio da Figura 6.14 é possível identificar visualmente as variações de TOs de cada agrupamento em relação às lesões encontradas no tecido.

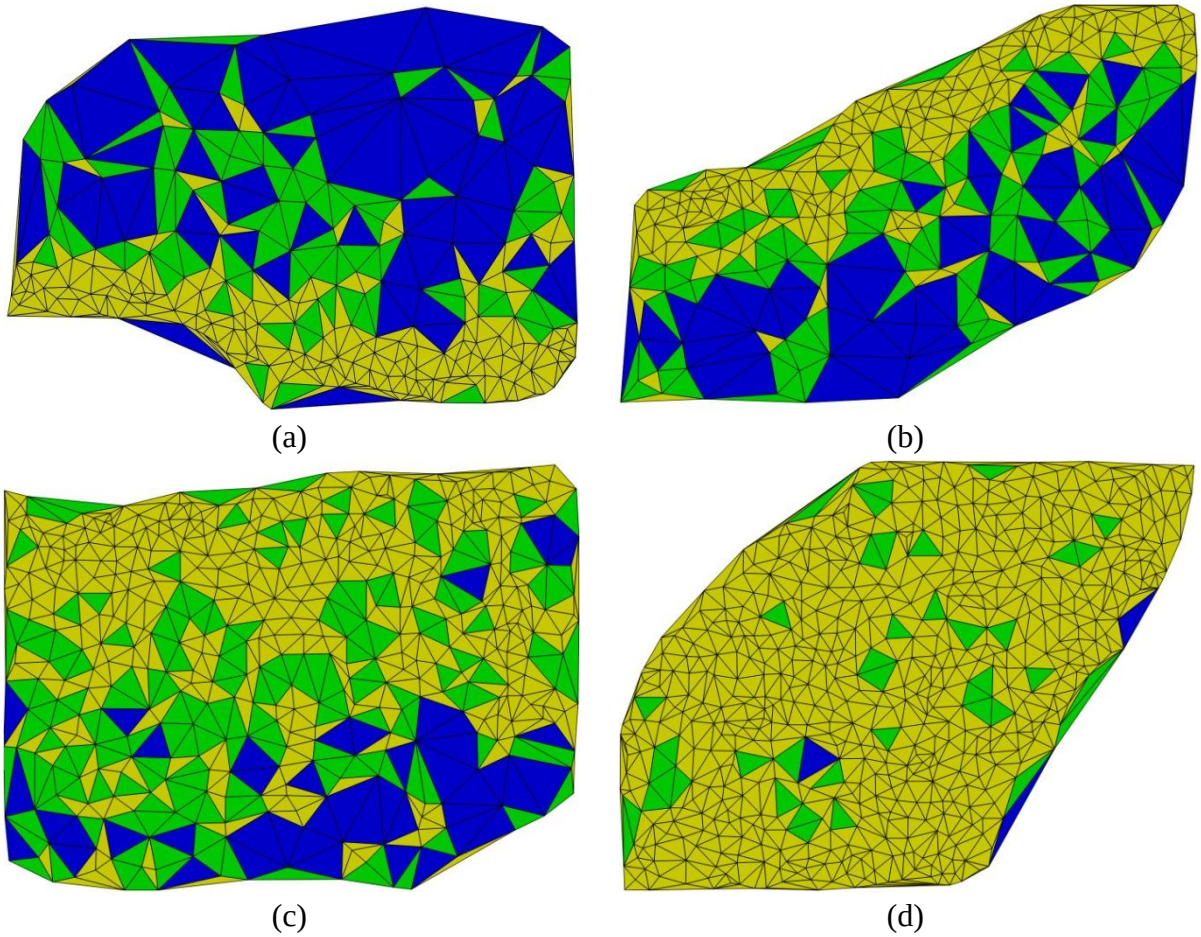


Figura 6.14. Agrupamentos obtidos para uma imagem (a) Normal, (b) NIC-I, (c) NIC-II e (d) NIC-III. Os agrupamentos amarelos foram classificados como basais, os verdes como intermediários e os azuis como superficiais

A seção seguinte apresenta uma comparação quantitativa entre as duas abordagens de rotulação das camadas.

6.7.2. Análise do parâmetro δ_S por meio da TO

Como apresentado nas figuras anteriores, o processo de identificação das camadas é extremamente dependente do valor do parâmetro δ_S. Para se determinar um valor ótimo para esse parâmetro, foi realizado um experimento no qual calculou-se a acurácia na identificação correta entre as quatro classes de interesse: imagens normais, imagens que apresentassem NIC-I, imagens que apresentassem NIC-II e imagens que apresentassem NIC-III, para diferentes valores de δ_S. Também foram comparadas as duas abordagens descritas na seção

anterior de rotulação das camadas do tecido. Para os experimentos apresentados a partir dessa seção foi utilizado um conjunto de 116 imagens e validação cruzada (5-*fold*). As Tabelas 4 e 5 apresentam os resultados obtidos a partir da abordagem não supervisionada e a da abordagem supervisionada, respectivamente. São apresentados os resultados de cada validação assim como a acurácia média para cada valor de δ_S, utilizando-se como atributos as TOs das camadas basal e superficial.

Tabela 4. Avaliação da acurácia na diferenciação entre NICs e imagens normais para diferentes valores de $\boldsymbol{\delta_S}$ utilizando-se a abordagem não supervisionada na rotulação das camadas

Validação	***Taxa de Acurácia - Abordagem Não Supervisionada***								
	δ_S=0,2	δ_S=0,15	δ_S=0,12	δ_S=0,10	δ_S=0,08	δ_S=0,06	δ_S=0,05	δ_S=0,03	δ_S=0,01
1	0,38	0,48	0,41	0,35	0,50	0,35	0,41	0,32	0,22
2	0,35	0,44	0,29	0,35	0,38	0,44	0,41	0,50	0,13
3	0,32	0,22	0,38	0,32	0,35	0,32	0,47	0,44	0,13
4	0,35	0,35	0,44	0,35	0,35	0,25	0,44	0,29	0,25
5	0,25	0,35	0,32	0,41	0,22	0,22	0,54	0,38	0,16
AC_m	**0,33**	**0,37**	**0,37**	**0,36**	**0,36**	**0,32**	**0,46**	**0,39**	**0,18**
DP	**0,05**	**0,10**	**0,06**	**0,03**	**0,10**	**0,09**	**0,05**	**0,08**	**0,05**

Tabela 5. Avaliação da acurácia na diferenciação entre NICs e imagens normais para diferentes valores de $\boldsymbol{\delta_S}$ utilizando-se a abordagem supervisionada na rotulação das camadas

Validação	***Taxa de Acurácia - Abordagem Supervisionada***								
	δ_S=0,2	δ_S=0,15	δ_S=0,12	δ_S=0,10	δ_S=0,08	δ_S=0,06	δ_S=0,05	δ_S=0,03	δ_S=0,01
1	0,53	0,53	0,50	0,50	0,47	0,67	0,56	0,66	0,63
2	0,50	0,34	0,41	0,41	0,56	0,56	0,59	0,66	0,66
3	0,53	0,53	0,47	0,50	0,56	0,69	0,59	0,63	0,66
4	0,63	0,66	0,56	0,53	0,50	0,53	0,50	0,63	0,63
5	0,53	0,50	0,47	0,47	0,50	0,59	0,59	0,56	0,59
AC_m	**0,54**	**0,51**	**0,48**	**0,48**	**0,52**	**0,61**	**0,57**	**0,63**	**0,64**
DP	**0,05**	**0,11**	**0,05**	**0,04**	**0,04**	**0,07**	**0,04**	**0,04**	**0,03**

Por meio desses resultados verifica-se que a abordagem não supervisionada de rotulação das camadas, caracterizada pela filtragem por área de ocupação, forneceu taxas de acurácia insatisfatórias. Apesar de alguns valores do parâmetro δ_S fornecerem valores de acurácia maiores quando comparados aos valores obtidos com a abordagem global (Seção 6.6.1), o processo de filtragem não forneceu uma boa diferenciação entre as classes de interesse. Já a abordagem supervisionada forneceu melhores taxas de acurácia média para todos os valores de δ_S, o que justifica sua utilização nos experimentos apresentados nas seções seguintes, nos quais foram comparadas as taxas de acurácia entre diferentes classes de interesse.

A partir dos resultados apresentados na Tabela 5, verifica-se que quanto menor o valor de δ_S, maior a acurácia obtida, o que foi observado também para outros conjuntos de atributos testados para a abordagem supervisionada. No entanto, é interessante notar que quando o valor de δ_S tende a zero, o agrupamento de dois triângulos da rede só é realizado se os mesmos forem idênticos e, consequentemente, o número de agrupamentos gerados pelo algoritmo de agrupamento tende ao número de triângulos da rede. Por meio desses resultados, pode-se concluir que para o problema abordado, utilizando esse conjunto de imagens, a rotulação direta de cada triângulo da rede, a partir dos intervalos de A_m definidos na abordagem supervisionada fornece melhor representação dos agrupamentos característicos de cada uma das classes em análise, e, portanto, maior acurácia na diferenciação entre NICs e imagens normais.

6.7.3. Classificação das NICs

A partir dos agrupamentos gerados, foram criados vetores de características extraindo-se medidas de cada agrupamento, como descrito na seção 5.4.2. Foram definidos nove conjuntos de atributos combinando-se as seguintes métricas: taxa de ocupação (TO), grau médio (k_{med}) e entropia (H) para os agrupamentos basal (B), intermediário (I) e superficial (S). Esses conjuntos de atributos são descritos abaixo.

$(a) = [TO_B, TO_S]$
$(b) = [k_{med}_B, k_{med}_I, k_{med}_S]$
$(c) = [k_{med}_B, k_{med}_S]$
$(d) = [H_B, H_I, H_S]$
$(e) = [H_B, H_S]$
$(f) = [TO_B, TO_S, k_{med}_B, k_{med}_S]$
$(g) = [TO_B, TO_S, H_B, H_S]$
$(h) = [k_{med}_B, k_{med}_S, H_B, H_S]$
$(i) = [TO_B, TO_S, k_{med}_B, k_{med}_S, H_B, H_S]$

Os resultados apresentados a seguir foram gerados utilizando-se a abordagem supervisionada e tiveram por objetivo a comparação entre as classes mais similares e que representam a transição gradual das lesões do colo uterino, de acordo com o modelo teórico apresentado na Figura 1.2, e, a identificação da presença de anomalias adotando-se a modelagem baseada em regiões.

6.7.3.1. Normal vs Lesão

Neste experimento, o método proposto foi avaliado quanto à detecção de anomalias. As NICs foram agrupadas em uma mesma classe, representando a presença de alterações estruturais no tecido. Os valores apresentados na Tabela 6 são valores médios de acurácia devido à utilização de validação cruzada. Além da acurácia média (AC_m), também são apresentados os valores de sensitividade (S_m) e especificidade (E_m) médios e os respectivos desvios padrão (dp) dessas medidas. Os valores destacados em negrito representam os maiores valores de acurácia, sensitividade e especificidade obtidas.

Na detecção de anomalias os melhores resultados foram obtidos utilizando-se o conjunto de atributos (a). Por terem sido obtidas acurácias semelhantes para os conjuntos (d) e (e), verifica-se que a utilização dos valores de entropia apenas dos agrupamentos basais e superficiais já são suficientes na distinção entre as classes analisadas. O mesmo ocorre com para os conjuntos (b) e (c) que avaliam o grau médio. Ou seja, a utilização de medidas da camada intermediária, para esse conjunto de dados, não alterou a acurácia obtida.

A combinação de atributos representadas pelos conjuntos (*f*), (*g*), (*h*) e (*i*) não aumentou os valores absolutos de acurácia obtidos em relação ao conjunto (*a*). No entanto, os valores de $dp(AC_m)$ fazem com que todos os conjuntos de atributos analisados neste experimento sejam bem próximos quanto a acurácia fornecida. Análise semelhante pode ser realizada para S_m e E_m, com destaque para a entropia (*d* e *e*) e a taxa de ocupação (a).

Neste experimento foram utilizadas 60 imagens na etapa de treinamento, sendo 30 imagens normais e 30 imagens que apresentam algum tipo de NIC, e, 15 imagens na etapa de teste.

Tabela 6. Avaliação da acurácia na detecção da presença de anomalias

Vetores de Atributos	$\boldsymbol{AC_m}$	$\boldsymbol{dp(AC_m)}$	$\boldsymbol{S_m}$	$\boldsymbol{dp(S_m)}$	$\boldsymbol{E_m}$	$\boldsymbol{dp(E_m)}$
$\vec{X}_a$	**0,88**	0,09	0,98	0,04	**0,68**	0,30
$\vec{X}_b$	0,85	0,09	0,96	0,09	0,64	0,33
$\vec{X}_c$	0,85	0,09	0,96	0,09	0,64	0,33
$\vec{X}_d$	0,83	0,11	**1,00**	0,00	0,48	0,33
$\vec{X}_e$	0,85	0,13	**1,00**	0,00	0,56	0,38
$\vec{X}_f$	0,85	0,09	0,96	0,09	0,64	0,33
$\vec{X}_g$	0,84	0,13	**1,00**	0,00	0,52	0,39
$\vec{X}_h$	0,87	0,09	0,98	0,04	0,64	0,33
$\vec{X}_i$	0,87	0,09	0,98	0,04	0,64	0,33

6.7.3.2. Normal vs NIC-I

As classes analisadas neste experimento apresentam padrões visuais bem semelhantes, já que o NIC-I é caracterizado por alteração estrutural apenas da camada basal do epitélio. O maior valor absoluto de acurácia na comparação entre essas duas classes foi obtido utilizando-se o vetor $\vec{X}_a$ (Tabela 7). As alterações estruturais apresentadas pelas lesões NIC-I aumentam o número de conexões entre as células e, consequentemente, o grau médio da camada basal do epitélio. Quando analisadas as taxas de sensitividade, verifica-se que a entropia fornece melhor taxa de identificação de verdadeiros positivos. No entanto, os conjuntos que forneceram taxas de sensitividade 1,0, também forneceram as menores taxas de especificidade.

De maneira similar ao experimento anterior, verificou-se valores de acurácia muito próximas para todos os conjuntos de atributos neste experimento devido ao elevado valor do desvio padrão. Neste experimento foram utilizadas 60 imagens na etapa de treinamento, sendo 30 de cada classe, e, 9 imagens na etapa de teste.

Tabela 7. Avaliação da acurácia na diferenciação entre as classes Normal e NIC-I

Vetores de Atributos	AC_m	$dp(AC_m)$	S_m	$dp(S_m)$	E_m	$dp(E_m)$
$\vec{X}_a$	**0,73**	0,23	0,90	0,14	**0,60**	0,47
$\vec{X}_b$	0,67	0,27	0,80	0,27	0,56	0,46
$\vec{X}_c$	0,67	0,27	0,80	0,27	0,56	0,46
$\vec{X}_d$	0,65	0,18	0,95	0,11	0,40	0,40
$\vec{X}_e$	0,67	0,22	**1,00**	0,00	0,40	0,40
$\vec{X}_f$	0,67	0,27	0,80	0,27	0,56	0,46
$\vec{X}_g$	0,67	0,22	**1,00**	0,00	0,40	0,40
$\vec{X}_h$	0,69	0,25	0,90	0,14	0,52	0,46
$\vec{X}_i$	0,69	0,25	0,90	0,14	0,52	0,46

6.7.3.3. NIC-I vs NIC-II

Na comparação realizada entre as classes NIC-II e NIC-III, os atributos relativos às TOs basal e superficial também forneceram os maiores valores absolutos de acurácia (Tabela 8), assim como para a sensitividade e a especificidade médias. Essas classes diferem-se pela camada do tecido atingida pela lesão, o que explica o fato do atributo TO ter se destacado neste experimento para todas as métricas utilizadas, apresentando o melhor compromisso entre sensitividade e especificidade. A análise da Tabela 8 também permite verificar que a

utilização da entropia relativa à camada intermediária melhora a acurácia absoluta obtida (comparação entre os vetores $\vec{X}_d$ e $\vec{X}_e$). As melhores taxas de sensitividade neste experimento também foram obtidas utilizando-se medidas relacionadas à entropia das três camadas. Foram utilizadas 44 imagens na etapa de treinamento, sendo 22 de cada classe analisada, e, 15 imagens na etapa de teste.

6.7.3.4. NIC-II vs NIC-III

A comparação entre as classes NIC-II e NIC-III é realizada pela identificação da camada do tecido atingida pela lesão. A TO também se destacou neste experimento na separação entre essas duas classes, como destacado na Tabela 9. Além disso, a TO também se destacou quanto aos valores de especificidade. Foram utilizadas 40 imagens na etapa de treinamento, sendo 20 de cada classe analisada, e, 7 imagens na etapa de teste.

Tabela 8. Avaliação da acurácia na diferenciação entre as classes NIC-I e NIC-II

Vetores de Atributos	$\boldsymbol{AC_m}$	$\boldsymbol{dp(AC_m)}$	$\boldsymbol{S_m}$	$\boldsymbol{dp(S_m)}$	$\boldsymbol{E_m}$	$\boldsymbol{dp(E_m)}$
$\vec{X}_a$	**0,77**	0,14	**0,80**	0,18	**0,77**	0,14
$\vec{X}_b$	0,65	0,06	0,60	0,28	0,67	0,06
$\vec{X}_c$	0,64	0,08	0,60	0,28	0,65	0,07
$\vec{X}_d$	0,65	0,08	0,73	0,13	0,63	0,10
$\vec{X}_e$	0,60	0,11	**0,80**	0,18	0,55	0,11
$\vec{X}_f$	0,65	0,09	0,60	0,28	0,67	0,08
$\vec{X}_g$	0,61	0,09	**0,80**	0,18	0,57	0,09
$\vec{X}_h$	0,67	0,08	0,60	0,28	0,68	0,07
$\vec{X}_i$	0,67	0,08	0,60	0,28	0,68	0,07

Tabela 9. Avaliação da acurácia na diferenciação entre as classes NIC-II e NIC-III

Vetores de Atributos	$\boldsymbol{AC_m}$	$\boldsymbol{dp(AC_m)}$	$\boldsymbol{S_m}$	$\boldsymbol{dp(S_m)}$	$\boldsymbol{E_m}$	$\boldsymbol{dp(E_m)}$
$\vec{X}_a$	**0,86**	0,20	0,70	0,45	**0,92**	0,20
$\vec{X}_b$	0,80	0,31	0,60	0,55	0,88	0,27
$\vec{X}_c$	0,83	0,26	0,70	0,45	0,88	0,27
$\vec{X}_d$	0,77	0,22	0,60	0,42	0,84	0,26
$\vec{X}_e$	0,83	0,23	**0,80**	0,27	0,84	0,26
$\vec{X}_f$	0,83	0,26	0,70	0,45	0,88	0,27
$\vec{X}_g$	0,80	0,24	0,70	0,45	0,84	0,26
$\vec{X}_h$	0,83	0,26	0,70	0,45	0,88	0,27
$\vec{X}_i$	0,83	0,26	0,70	0,45	0,88	0,27

Nos experimentos mostrados nas tabelas 6, 7, 8 e 9 a classe positiva é sempre representada pela anomalia mais grave ou simplesmente pela presença da anomalia. O Apêndice A apresenta as matrizes de confusão dos experimentos apresentados nesta seção para o conjunto de atributo (a).

6.8. Avaliação da comparação entre NICs

É possível verificar que o atributo TO permitiu uma boa separação entre os conjuntos e classes analisados, fornecendo valores absolutos de acurácia sempre maiores do que 73%, igualando-se aos demais conjuntos de atributos em alguns casos ou apresentando melhores resultados do que os mesmos. No entanto, de maneira geral, em todos os experimentos apresentados na seção anterior, verifica-se que os valores de desvios-padrão indicam que todos os conjuntos de atributos utilizados fornecem resultados muito próximos quanto aos valores de acurácia, sensitividade e especificidade.

Quanto aos valores de sensitividade, a entropia (H) apresentou melhores valores absolutos, porém, ela apresentou também uma especificidade muito baixa. Apesar de ser interessante a obtenção de altas taxas de sensitividade para o problema em análise, uma especificidade baixa pode levar a condutas diagnósticas mais agressivas em casos que poderiam receber tratamentos mais simples. Por exemplo, um falso positivo exemplificado na Tabela 8 poderia levar a uma intervenção cirúrgica em um caso que necessitaria apenas de um tratamento não intrusivo. Para analisar qual conjunto de atributos fornece o melhor compromisso entre as taxas de sensitividade e especificidade foi calculada a correlação de *Matthews* (MCC) para cada comparação realizada entre as diferentes classes de interesse. Esse resultado é apresentado na Tabela 10.

Verifica-se que o conjunto de atributos representado pelas TOs basal e superficial (a) também foi o conjunto que forneceu o melhor compromisso entre sensitividade e especificidade nas comparações analisadas exceto na comparação entre as classes Normal *vs* Lesão. Esse resultado também pode ser visualizado no gráfico da Figura 6.15 que apresenta os valores de sensitividade em função dos valores de especificidade para os diferentes conjuntos de atributos testados, comparando-se as classes Normal *vs* Lesão. Verifica-se que os conjuntos (d), (e) e (g) forneceram os maiores valores de sensitividade e o conjunto (a), o maior valor de especificidade.

Tabela 10. Valores de MCC referentes às comparações entre diferentes classes realizadas na seção anterior para diferentes vetores de atributos.

Vetores de Atributos	*Normal vs Lesão*		*Normal vs NIC-I*		*NIC-I vs NIC-II*		*NIC-II vsNIC-III*	
	MMC	dp	MMC	dp	MMC	dp	MMC	dp
$\vec{X}_a$	0,737	0,196	**0,664**	0,296	**0,492**	0,272	**0,773**	0,454
$\vec{X}_b$	0,684	0,197	0,478	0,560	0,217	0,207	0,613	0,774
$\vec{X}_c$	0,684	0,197	0,478	0,560	0,203	0,226	0,727	0,546
$\vec{X}_d$	0,704	0,172	0,514	0,237	0,302	0,134	0,571	0,467
$\vec{X}_e$	0,779	0,191	0,650	0,325	0,283	0,192	0,657	0,447
$\vec{X}_f$	0,684	0,197	0,478	0,560	0,219	0,235	0,727	0,546
$\vec{X}_g$	**0,741**	0,223	0,566	0,325	0,296	0,175	0,660	0,517
$\vec{X}_h$	0,706	0,213	0,552	0,461	0,234	0,236	0,727	0,546
$\vec{X}_i$	0,706	0,213	0,552	0,461	0,234	0,236	0,727	0,546

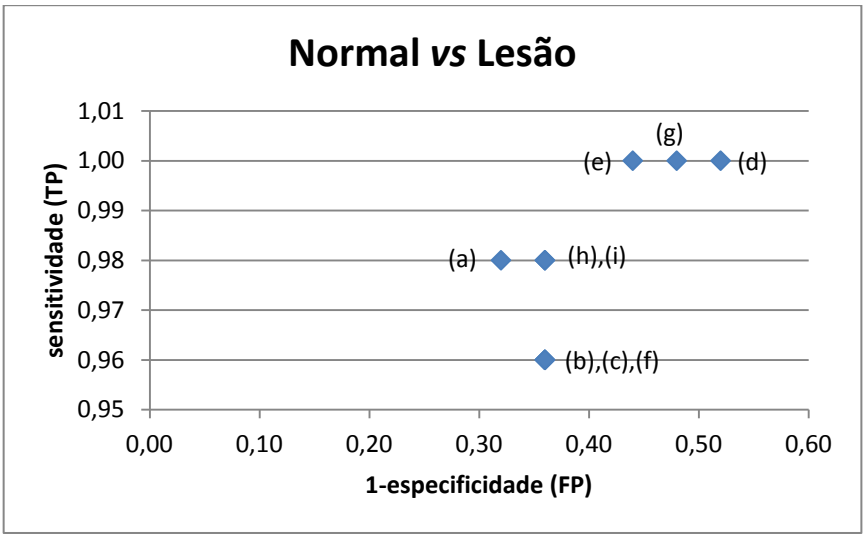


Figura 6.15. Análise das taxas de sensitividade e especificidade obtidas comparando-se imagens normais e imagens que apresentam lesão

De maneira similar, os gráficos das Figuras 6.16, 6.17 e 6.18 apresentam o compromisso entre as taxas de sensitividade e especificidade para as comparações entre as classes Normal *vs* NIC-I, NIC-I *vs* NIC-II e NIC-II *vs* NIC-III.

Na Figura 6.16, verifica-se que os atributos relacionados à entropia (H) tendem a fornecer maiores taxas de sensitividade e os atributos relacionados ao grau médio (k_{med}) tendem a fornecer melhores taxas de especificidade. Na Figura 6.17, o conjunto (a) também foi responsável pelo maior valor de especificidade, além da melhor acurácia. A comparação entre NIC-II e NIC-III (Figura 6.18) forneceu os maiores valores de especificidade dos experimentos realizados sendo que o melhor compromisso entre sensitividade e especificidade também foi fornecido pelo conjunto (a).

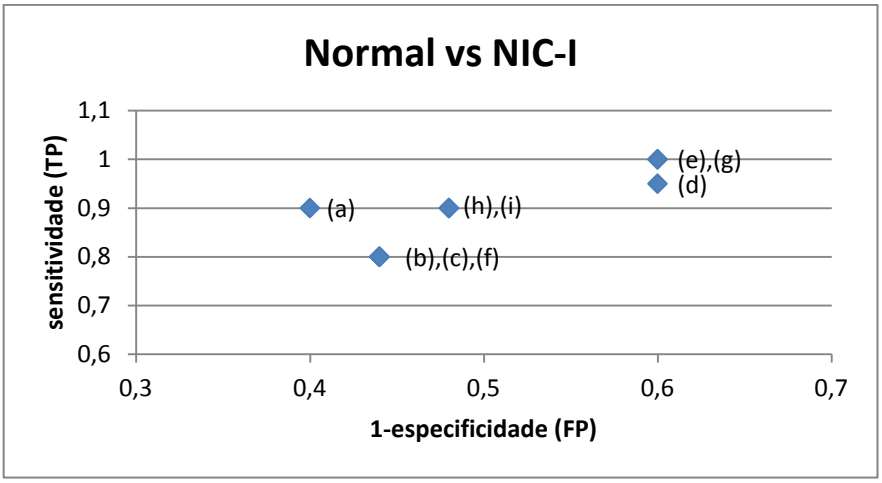


Figura 6.16. Análise das taxas de sensitividade e especificidade obtidas comparando-se imagens normais e imagens que apresentam NIC-I

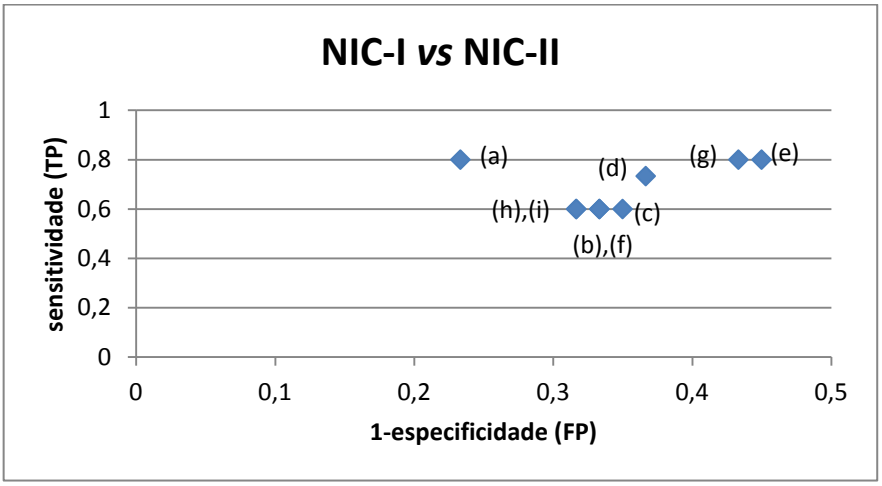


Figura 6.17. Análise das taxas de sensitividade e especificidade obtidas comparando-se imagens que apresentam NIC-I e imagens que apresentam NIC-II

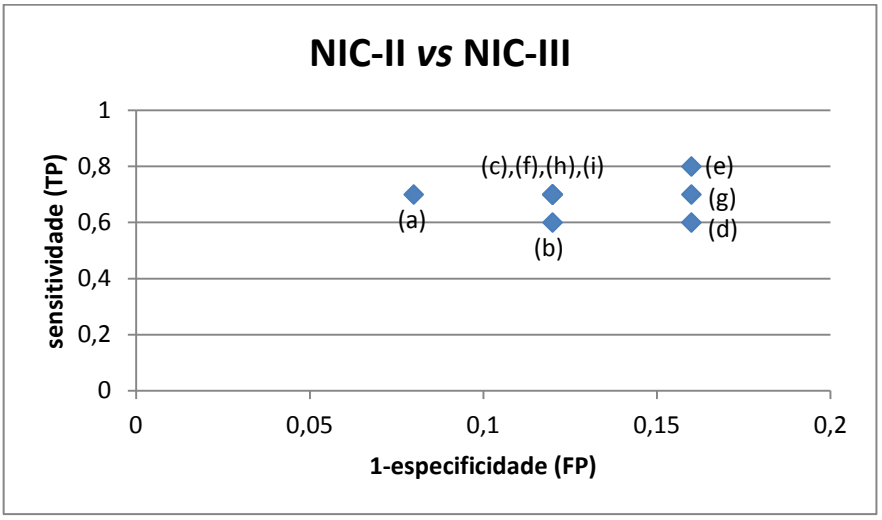


Figura 6.18. Análise das taxas de sensitividade e especificidade obtidas comparando-se imagens que apresentam NIC-II e imagens que apresentam NIC-III

Os gráficos apresentados corroboram os valores de MCC da Tabela 10, na qual verifica-se que o melhor compromisso entre sensitividade e especificidade é alcançado com o conjunto (a) para a maioria das comparações efetuadas, assim como os maiores valores de acurácia, como apresentado na seção 6.7. Portanto, conclui-se que a TO permite uma boa representação da porção do tecido atingida pela lesão, refletindo as alterações estruturais representadas na Figura 1.2.

No entanto, a escolha do conjunto de atributos deve levar em consideração a abordagem em questão. Se a sensitividade deve ser priorizada, atributos relacionados a entropia podem ser a melhor opção. Quanto à especificidade, atributos relacionados a TO e ao grau médio se destacaram para esse conjunto de dados.

6.9. Considerações Finais

Os experimentos apresentados ao longo desse capítulo permitem verificar a adequação do método proposto neste projeto ao objetivo do mesmo: caracterizar estruturas em imagens histológicas e, numa abordagem mais específica, identificar lesões neoplásicas de colo uterino em seus diferentes estágios de evolução. A utilização de métodos baseados em morfologia na etapa de segmentação permitiu a identificação dos núcleos de interesse nas imagens possibilitando uma representação topológica do tecido por meio de grafos de vizinhança. A Triangulação de *Delaunay* mostrou-se adequada e suas propriedades específicas permitiram a modelagem realizada em camadas. O método proposto para análise estrutural da arquitetura de tecidos histológicos fornece uma análise automática de agrupamentos característicos em imagens a partir de atributos intrínsecas às mesmas utilizando as propriedades da TD. A aplicabilidade deste método pode ser visualizada na caracterização das NICs, cujos resultados foram apresentados neste capítulo. É importante observar também que todo o processamento realizado sobre as imagens, inerente aos métodos propostos, é invariante a rotação e escala.

7. CONCLUSÃO

Este trabalho apresentou um método para análise de imagens histológicas a partir da exploração de atributos topológicos com o objetivo de se identificar de maneira automática a presença de alterações estruturais nos tecidos. Este método baseia-se na caracterização de regiões ou agrupamentos celulares com características semelhantes subsidiando a extração de atributos por regiões. Além disso, foi proposto e implementado um método para a análise automática de imagens histopatológicas do colo uterino, que utiliza o método anterior com o objetivo de se mapear as lesões precursoras do câncer que aí se instala. Para isto, foram executadas as seguintes etapas: levantamento de parâmetros perceptuais de análise do tecido epitelial do colo uterino juntamente com especialistas da área; digitalização de regiões de interesse a partir de lâminas provenientes de biópsias; modelagem e implementação de uma base de imagens representativas dessas regiões; segmentação dos núcleos celulares a partir de operadores morfológicos com posterior aplicação da *watershed*, e, análise estrutural a partir da representação do tecido por meio de grafos de vizinhança, o que resultou na identificação automática das camadas do tecido e posterior classificação das NICs.

A criação da base de dados descrita na seção 6.2, além de possibilitar fácil acesso às imagens de interesse, permitiu também o acesso às informações complementares aos exames referentes a essas imagens. Podem ser recuperadas todas as imagens provenientes de exames realizados por uma mesma paciente e as datas de ocorrência de cada um. Desta forma, a análise do histórico clínico de pacientes pode ser explorada em um trabalho futuro por meio da análise da evolução das NICs e empregando-se os métodos propostos neste trabalho. Além disso, o armazenamento dessas informações de maneira estruturada permite que a base seja facilmente expandida, a partir da inserção de novos exemplos.

O processamento das imagens de interesse em níveis de cinza amenizou os efeitos dos processos de aquisição das lâminas e das imagens sujeitos respectivamente às variações nas concentrações de fixadores e de intensidade de luz incidente sobre a amostra. Além disso, o resultado fornecido pelas técnicas baseadas em cor pode ser comparado ao resultado obtido com a aplicação do operador *top-hat* por reconstrução nas imagens em níveis de cinza no processo de identificação dos marcadores internos aos núcleos celulares. Quando comparadas as duas abordagens de segmentação baseadas em cor, a técnica de deconvolução apresentou

melhor separação das estruturas coradas por hematoxilina e eosina. A utilização de métodos de segmentação baseados na detecção de bordas, como a *watershed*, permitiu a caracterização morfológica dos núcleos das imagens. No entanto, para a geração dos grafos sobre os tecidos, foi necessário somente a identificação dos marcadores internos aos núcleos, os quais foram obtidos com a aplicação de um *pipeline* de operadores morfológicos. A extração de atributos morfológicos não está no escopo desse projeto.

No entanto, a avaliação da utilização de atributos morfológicos juntamente com os atributos topológicos, pode ser considerada em um trabalho futuro. No caso das NICs, a área nuclear, assim como a proporção núcleo/citoplasma são parâmetros importantes para a avaliação da lesão. Quando comparadas imagens normais do tecido epitelial do colo uterino com imagens que apresentam NIC-I verifica-se que a alteração estrutural ocorre apenas na camada basal. A inclusão da área nuclear como atributo poderia levar a um aumento da taxa de acurácia obtida com as TOs.

Com relação à avaliação dos métodos de análise estrutural, verificou-se que a extração de métricas globais da rede, levando-se em consideração todo o conjunto de nós e arestas, não fornece boas taxas de acurácia na diferenciação entre os diferentes NICs e imagens normais (Tabela 2). Com a utilização do algoritmo de agrupamento foi possível implementar uma modelagem baseada em regiões da imagem e aumentar as taxas de acurácia obtidas em relação à abordagem global. No caso das imagens de tecido epitelial do colo uterino, foi possível identificar de maneira automática as camadas do mesmo adotando-se critérios de semelhança e relações de adjacência.

É interessante notar que a adoção de valores de δ_S próximos a 0 tende a gerar uma grande quantidade de agrupamentos, com poucos ou apenas um triângulo em cada um. Além disso, os resultados apresentados na Tabela 5 mostram que quanto menor o valor de δ_S, maior a taxa de acurácia obtida. Assim, verifica-se que a adoção de valores de δ_S pequenos, com posterior classificação dos mesmos é equivalente à rotulação direta dos triângulos individualmente, para o conjunto de dados em análise. Também foi verificado que a abordagem supervisionada mostrou-se mais adequada ao processo de identificação das camadas do tecido epitelial. Por este motivo, nos experimentos da seção 6.8 foi utilizada a abordagem supervisionada.

Quanto à avaliação do método proposto na representação da transição gradual das lesões do colo uterino, foram obtidos valores de acurácia maiores do que 70% para as comparações entre as seguintes classes: Normal *x* NIC-I, NIC-I *x* NIC-II, e, NIC-II *x* NIC-III.

É importante destacar que a diferença entre as classes Normal e NIC-I encontra-se apenas na camada basal. A detecção da presença de anomalias também foi avaliada e, para isto, todas as NICs foram agrupadas em uma mesma classe: Lesão. Essa classe foi comparada a imagens normais e foi obtida acurácia igual a 88%. Além disso, foi verificada também, a acurácia na discriminação entre os três tipos de NICs e regiões normais, obtendo-se acurácia de 64%. O método descrito por Keenan *et al*, apresentado no Capítulo 3, forneceu uma taxa de acurácia de 62,3% comparado-se apenas os três NICs. Portanto, verifica-se que com o método proposto neste trabalho foi alcançada melhor acurácia na diferenciação entre NICs em relação ao trabalho de Keenan. O trabalho de Landini *et al* também avaliou a acurácia na classificação de lesões pré-malignas (neste caso, relacionadas ao carcinoma das vias orais) atingindo valor máximo de 51%, sendo comparadas apenas lesões de alto e baixo grau. Apesar do contexto de aplicação do trabalho de Landini ser diferente do presente trabalho, ambos são aplicados a tecidos epiteliais.

Não foram testadas abordagens probabilísticas na geração do grafo sobre o tecido, como nos trabalhos de Demir e Gunduz (Capítulo 3), sendo consideradas apenas relações de vizinhança para geração do grafo. Porém, de maneira semelhante a esses trabalhos, os resultados atingidos com o método proposto permitiram verificar a adequação da utilização de atributos topológicos na diferenciação de lesões de interesse. A Triangulação de *Delaunay* permitiu a representação do relacionamento entre as células e suas propriedades específicas puderam ser utilizadas da identificação das camadas do tecido. Também verifica-se que o método apresentado é invariante a rotação e a escala das imagens utilizadas.

Quanto ao atributos utilizados, em todas as comparações realizadas, o conjunto que forneceu melhores valores absolutos de acurácia, foi o conjunto composto pelas TOs das camadas basal e superficial. Como as TOs das três camadas são linearmente dependentes, foram utilizadas apenas os valores das camadas citadas. Verificou-se também que, para as outras métricas, quando avaliadas separadamente, o valor referente à camada intermediária não aumentava o valor absoluto da acurácia, assim como a combinação de diferentes métricas em um mesmo vetor. Quanto à avaliação da sensitividade, a entropia forneceu as melhores taxas. No entanto, para a maioria das comparações realizadas entre as classes de interesse, o melhor compromisso entre sensitividade e especificidade, foi obtido com a utilização do conjunto de atributos formado pelas TOs das camadas basal e superficial, verificado através do cálculo do coeficiente de *Matthews* (MCC). Além disso, os valores de desvios-padrão aproximaram bastante os resultados obtidos com os diferentes vetores de atributos testados.

O método apresentado ainda pode ser utilizado e avaliado na identificação de outros tipos de lesões de interesse e outros tipos de tecido, uma vez que, este método é genérico quanto ao número de agrupamentos que podem ser identificados e o classificador proposto pode ser facilmente adaptado a esses agrupamentos. Assim, em um trabalho futuro também pode ser considerada a análise de outros tipos de lesões e inflamações relacionadas a alterações estruturais em imagens histopatológicas.

No auxílio ao diagnóstico das NICs, e consequente prevenção de lesões malignas que possam levar ao câncer de colo uterino, também pode ser considerada a modelagem e a implementação de um sistema automatizado de auxílio ao diagnóstico. A construção e a validação desse sistema podem ser acompanhadas por especialistas da área de modo a oferecer fácil acesso a consultas de interesse. Por fim, espera-se que este trabalho possa representar uma contribuição para o aprimoramento dos métodos de análise automática de imagens para auxílio ao diagnóstico, aumentando o grau de precisão do processo de geração de laudos e diagnósticos em exames histopatológicos, e, podendo atuar como uma segunda opinião para o patologista.

REFERÊNCIAS BIBLIOGRÁFICAS

[Altinok-2007] ALTINOK, A. *et al*. Model based dynamics analysis in live cell microtubule image. *BMC Cell Biol.*, v.8 (Suppl. 1), S4, 2007.

[Angulo-2003] ANGULO, J.; SERRA, J. Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics*, v. 19, p. 553-562, 2003.

[Araújo-1999] ARAÚJO, S. R. *Citologia e Histopatologia Básicas do Colo Uterino para Ginecologistas*. Curitiba: VP, 1999. 112 p.

[Baldi-2000] BALDI, P.; BRUNAK, S.; CHAUVIN, Y.; ANDERSEN, C. A. F.; NIELSEN, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, v. 16, n. 5, p. 412-424, 2000.

[Banon&Barrera-1993] BANON, G. J. F.; BARRERA, J. Decomposition of mappings between complete lattices by mathematical morphology. *Part I. general lattices, Signal Processing*, v. 30, p. 299-322, 1993.

[Banon&Barrera-1994] BANON, G. J. F.; BARRERA, J. *Bases da Morfologia Matemática para Análise de Imagens Binárias*. IX Escola de Computacão. Pernambuco, Julho, 1994.

[Brandes-2005] BRANDES, U.; ERLEBACH, T. *Network Analysis: Methodological Foundations*. Berlin, NewYork: Springer, 2005.

[Costa-2007] COSTA, L. F.; RODRIGUES, F. A.; TRAVIESO, G.; VILLAS BOAS, P.R. Characterization of complex networks: A survey of measurements. *Advances in Physics*, v. 56, p. 167-242, 2007.

[Demir-2005a] DEMIR, C.; GULTEKIN, S. H; YENER, B. Learning the topological properties of brain tumors. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 2, n. 4, p. 1-9, 2005.

[Demir-2005b] DEMIR, C.; GULTEKIN, S. H; YENER, B. Augmented cell-graphs for automated cancer diagnosis. *Bioinformatics*, v. 21, Suppl.2, p. ii7-ii12, 2005.

[Demir-2006] Demir, C.; Yener, B. Automated cancer diagnosis based on histopathological images: A systematic survey. *Rensselaer Polytechnic Institute*, Troy, NY, 2006.

[Doi-2007a] DOI, K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *J Comp Med Imaging Graphics*, v. 31, p. 198-211, 2007.

[Doi-2007b] DOI, K. Computer-aided diagnosis moves from breast to other systems. *Diag Imaging*, May, p. 37-40, 2007.

[Dorr-2008] DORR, A. E. *et al*. High resolution three-dimensional brain atlas using an average magnetic resonance image of 40 adult C57Bl/6J mice. *Neuroimage*, v. 42, p. 60-69, 2008.

[Dougherty-2003] DOUGHERTY, E.; LOTUFO, R. A. *Hands-on Morphological Image Processing*. The Society of Photo Optical Instrumentation Engineers, 2003.

[Duda-2001] DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*, John Wiley Interscience, 2nd ed, 2001.

[Egecioglu-2004] EGECIOGLU, O.; FERHATOSMANOGLU, H.; OGRAS, U. Dimensionality reduction and similarity computation by inner-product approximations. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, v. 16, n. 6, p. 714-726, 2004.

[Felipe-2005] FELIPE, J. C. *Desenvolvimento de métodos para extração, comparação e análise de características intrínsecas de imagens médicas, visando à recuperação perceptual por conteúdo*, Tese de Doutorado, Departamento de Ciência da Computação - ICMC, Universidade de São Paulo, São Carlos, Brasil, 2005.

[Geusebroek-1999] GEUSEBROEK, J. M.; SMEULDERS, A. W. M.; CORNELISSEN, F.; GEERTS, H. Segmentation of Tissue Architecture by Distance Graph Matching. *Cytometry* v.35, p.11-22, 1999.

[Gunduz-2004] GUNDUZ, C.; YENER, B.; GULTEKIN, S. H. The cell graphs of cancer. *Bioinformatics*, v. 20, suppl. 1, p. i145-i151, 2004.

[Gunduz-Demir-2007] GUNDUZ-DEMIR, C. Mathematical modeling of the malignancy of cancer using graph evolution. *Mathematical Biosciences*, v. 209, p. 514-527, 2007.

[Gurcan-2009] GURCAN, M.; BOUCHERON, L.; CAN, A.; MADABHUSHI, A.; RAJPOOT, N.; YENER, B. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, v. 2, p. 147-171, 2009.

[Hirata-2001] HIRATA Jr, R.; BARRERA, J.; HASHIMOTO, R. F.; DANTAS, D. O. Microarray gridding by mathematical morphology. *In: Proceedings of the SIBGRAPI 2001*. Florianópolis: IEEE, v. 1, p.112-119, 2001.

[IARC-2011] INTERNATIONAL AGENCY FOR RESEARCH ON CANCER. Histopatologia do Colo Uterino - Atlas Digital. Disponível em: http://screening.iarc.fr/atlashisto.php?lang=4. Acessado em 23 de agosto de 2011.

[INCA-2008] INSTITUTO NACIONAL DE CÂNCER. Estimativas 2008: incidência de cancer no Brasil. Rio de Janeiro, 2008.

[Ismail-1989] ISMAIL, S. M.; COLCLOUGH, A. B.; DINNEN, J. S.; EAKINS, D. et al. Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *Br. Med. J.*, v. 298, p. 707-710, 1989.

[Jain-2000] JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1, p. 4-37, 2000.

[Junqueira-2004] JUNQUEIRA, L. C.; CARNEIRO, J. *Histologia Básica*. 10ª ed. Rio de Janeiro, RJ, Guanabara Koogan S.A., 2004.

[Kass-1987] KASS, M.; WITKIN, A.; TERZOPOULOS, D. Snakes: Active contour models. *Int. J. Comput. Vis.*, v. 1, p. 321-331, 1987.

[Keenan-2000] KEENAN, S. J.; DIAMOND, J.; McCLUGGAGE, W. G.; BHARUCHA, H.; THOMPSON, D.; BARTELS, P. H.; HAMILTON, P. W. An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN). *Journal of Pathology*, v. 192, n. 3, p. 351-62, 2000.

[Landini-2003] LANDINI, G.; OTHMAN, I.E. Estimation of tissue layer level by sequential morphological reconstruction. *Journal of Microscopy*, v. 209, n. 2, p. 118-125, 2003.

[Landini-2004] LANDINI, G.; OTHMAN, I. E. Architectural analysis of oral cancer, dysplastic, and normal epithelia. *Cytometry*, Part A v. 61, p. 45-55, 2004.

[Landini-2009] LANDINI, G; PERRYER, D. G. Digital enhancement of Haematoxylin and Eosin stained histological images for red-green colour blind observers. *Journal of Microscopy*, v. 234, n. 3, p. 293-301, 2009.

[Lein-2007] LEIN, E. *et al*. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, v. 445, p. 168-176, 2007.

[Liew-2005] LIEW, A. W.-C; YAN, H.; YANG, M. Pattern Recognition Techniques for the Emerging Field of Bioinformatics: A Review. *Pattern Recognition*, vol.38, n.11, p. 2055-2073, 2005.

[Likar-2000] LIKAR, B.; MAINTZ, J. B. A.; VIERGEVER, M. A.; PERNUS, F. Retrospective shading correction based on entropy minimization. *Journal of Microscopy*, v. 197, p. 285-295, 2000.

[Lonky-1999] LONKY, N.; SADEGUI, M.; TSADIK, G. W.; PETITTI, D. The clinical significance of the poor correlation of cervical displasia and cervical malignancy with referral cytology results. *Am J Obstet Gynecol*. v. 181, p. 560-566, 1999.

[MathWorks-2010] Color-Based Segmentation Using K-Means Clustering. Disponível em http://www.mathworks.com/products/image/demos.html?file=/products/demos/shipping/images/ipexhistology.html. Acessado em 23 de agosto de 2010.

[Matula-1980] MATULA, D. W.; SOKAL, R. R. Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane. *Geographical Analysis*, v. 12, p. 205-222, 198

[McCluggage-1998] MCCLUGGAGE, W. G.; WALSH, M.Y.; THORNTON, C. M. et al. Inter- and intra-observer variation in the histopathological reporting of cervical squamous intraepithelial lesions using a modified Bethesda grading system. *Br J Obstet Gynaecol*.; v. 105, p. 206-210, 1998.

[Metz-2000] METZ, C. E. “Fundamental ROC analysis,” in *Handbook of medical imaging*, v. 1, J. Beutel, H. L. Kundel,R. L. Van Metter, Eds. Bellingham, Washington: SPIE Press, *The International Society for Optical Engineering*, p. 751-770, 2000.

[Meyer-1990] MEYER, F.; BEUCHER, S. Morphological segmentation. *Journal of Visual Communications and Image Processing*, v. 1, n. 1, p. 21-46, 1990.

[Nanda-2000] NANDA, K.; MCCRORY, D. C.; MEYERS, E. R.; BASTIAN, L.A.; HASSELBLAD, V.; HICHEY, J. D.; MATCHAR, D. B. Accuracy of the Papanicolaou test in screening and follow-up of cervical cytologic abnormalities: a systematic review. *Ann Intern Med*; v. 132, p. 810 - 819, 2000.

[Peng-2008] PENG, H. Bioimage informatics: a new area of engineering biology, *Bioinformatics*, v **24**, p. 1827–1836, 2008.

[Reis-2001] REIS, C. F.; ALBUQUERQUE, M. P.; CASTRO, S. B. Introdução ao Reconhecimento de Padrões ultilizando Redes Neurais. Disponível em: http://biblioteca.cat.cbpf.br/pub/apub/nt/2001/nt00201.pdf. Acesso em 30 de ago. 2009.

[Rizzio-2011] RIZZIO, R.; STRANSKY, B.; ZAMPIROLLI, F.A.; BARRERA, J. Classifying biological tissue architectures with small samples. *Pattern Recognition Letters*. No prelo.

[Rosenberg-2011] ROSENBERG, M. S; ANDERSON, C.D. PASSaGE: Pattern Analysis, Spatial Statistics and Geographic Exegesis, 2011. http://www.passagesoftware.net.

[Ruifrok-2001] RUIFROK, A. C.; JOHNSTON, D. A. Quantification of histological staining by color deconvolution. *Anal Quant Cytol Histol*, v. 23, p. 291-299, 2001.

[Salembier-1995] SALEMBIER, P.; SERRA, J. Flat Zones Filtering, Connected Operators, and Filters by Reconstruction. *IEEE Transactions on Image Processing*, v. 4, n. 8, p. 1153-1160, 1995.

[Shamir-2010] SHAMIR, L.; DELANEY, J. D.; ORLOV, N.; ECKLEY, D. M.; GOLDBERG, I. G. Pattern Recognition Software and Techniques for Biological Image Analysis. *PLoS Computational Biology*, v. 6, n. 11, 2010.

[Solomon-2002] SOLOMON, D.; DAVEY, D.; KURMAN, R. et al. The 2001 Bethesda System: terminology for reporting results of cervical cytology. *JAMA*, v. 287, p. 2114-2119, 2002.

[Sternberg-1983] STERNBERG, S. Biomedical Image Processing, *IEEE Computer*, v. 16, n. 1, p. 22-34, 1983.

[Stevens-1997] STEVENS, A.; LOWE, J. *Human Histology*, 2nd Ed. Times Mirror International Publishers Ltd, UK, 1997.

[Theodoridis-2009] THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern recognition.* Academic Press, 4.ed, 2009.

[Traina-2007] TRAINA, C.; SANTOS-FILHO, R. F.; TRAINA, A. J. M.; VIEIRA, M. R.; FALOUTSOS, C. The OMNI-Family of All-Purpose Access Methods: A Simple and Effective Way to Make Similarity Search More Efficient. *The International Journal on Very Large Databases*, v. 16, p. 483-505, 2007.

[Tsien-2003] TSIEN, R. Y. Imagining imaging's future, *Nat. Rev. Mol. Cell Biol.*, v. 4, SS16–SS21, 2003.

[Vincent-1989] VINCENT, L. Graphs and Mathematical Morphology. *Signal Processing*, vol. 16, n. 4, p. 365-388, 1989.

[Vincent-1991] VINCENT, L.; SOILLE, P. Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 13, n. 6, p. 583-598, 1991.

[Zampirolli-2010] ZAMPIROLLI, F. A.; STRANSKY, B; LORENA, A. C.; Paulon, F. L. M. Segmentation and Classification of Histological Images - Application of Graph Analysis and Machine Learning Methods. In: *Brazilian Symposium on Computer Graphics and Image Processing* 2010. Gramado. Proceedings, 2010.

APÊNDICE A

Matrizes de Confusão dos experimentos realizados

A.1 Matrizes de confusão resultantes da validação cruzada (5-*fold*) descrita no experimento da seção 6.7.3.1: Normal *vs* Lesão

- Matrizes obtidas para o vetor de atributos $\vec{X}_a = [TO_B, TO_S]$

Validação 1		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	10	3
	Negativo	0	2

Sensitividade = 1,00 / Especificidade = 0,40

Validação 2		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	10	3
	Negativo	0	2

Sensitividade = 1,00 / Especificidade = 0,40

Validação 3		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	10	2
	Negativo	0	3

Sensitividade = 1,00 / Especificidade = 0,60

Validação 4		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	9	0
	Negativo	1	5

Sensitividade = 0,90 / Especificidade = 1,00

Validação 5		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	10	0
	Negativo	1	5

Sensitividade = 1,00 / Especificidade = 1,00

A.2 Matrizes de confusão resultantes da validação cruzada (5-*fold*) descrita no experimento da seção 6.7.3.2: Normal *vs* NIC-I

- Matrizes obtidas para o vetor de atributos $\vec{X}_a = [TO_B, TO_S]$

Validação 1		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	4	4
	Negativo	0	1

Sensitividade = 1,00 / Especificidade = 0,20

Validação 2		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	4	5
	Negativo	0	0

Sensitividade = 1,00 / Especificidade = 0,00

Validação 3		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	3	1
	Negativo	1	4

Sensitividade = 0,75 / Especificidade = 0,80

Validação 4		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	4	0
	Negativo	0	5

Sensitividade = 1,00 / Especificidade = 1,00

Validação 5		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	3	0
	Negativo	1	5

Sensitividade = 0,75 / Especificidade = 1,00

A.3 Matrizes de confusão resultantes da validação cruzada (5-*fold*) descrita no experimento da seção 6.7.3.3: NIC-I *vs* NIC-II

- Matrizes obtidas para o vetor de atributos $\vec{X}_a = [TO_B, TO_S]$

Validação 1		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	2	5
	Negativo	1	7

Sensitividade = 0,67 / Especificidade = 0,58

Validação 2		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	3	2
	Negativo	0	10

Sensitividade = 1,00 / Especificidade = 0,83

Validação 3		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	3	1
	Negativo	0	11

Sensitividade = 1,00 / Especificidade = 0,92

Validação 4		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	2	4
	Negativo	1	8

Sensitividade = 0,67 / Especificidade = 0,67

Validação 5		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	2	2
	Negativo	1	10

Sensitividade = 0,67 / Especificidade = 0,83

A.4 Matrizes de confusão resultantes da validação cruzada (5-*fold*) descrita no experimento da seção 6.7.3.4: NIC-II *vs* NIC-III

- Matrizes obtidas para o vetor de atributos $\vec{X}_a = [TO_B, TO_S]$

Validação 1		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	2	0
	Negativo	0	5

Sensitividade = 1,00 / Especificidade = 1,00

Validação 2		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	2	0
	Negativo	0	5

Sensitividade = 1,00 / Especificidade = 1,00

Validação 3		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	2	0
	Negativo	0	5

Sensitividade = 1,00 / Especificidade = 1,00

Validação 4		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	1	2
	Negativo	1	3

Sensitividade = 0,50 / Especificidade = 0,60

Validação 5		Classe Real	
		Positivo	Negativo
Classe Predita	Positivo	0	0
	Negativo	2	5

Sensitividade = 0,00 / Especificidade = 1,00

ANEXO A

Aprovação do Comitê de Ética em Pesquisa

HOSPITAL DAS CLÍNICAS DA FACULDADE DE MEDICINA
DE RIBEIRÃO PRETO DA UNIVERSIDADE DE SÃO PAULO

www.hcrp.usp.br

USP - RIBEIRÃO
assistência - ensino - pesquisa

Ribeirão Preto, 06 de outubro de 2010

Ofício nº 3392/2010
CEP/MGV

Prezada Senhora,

O trabalho intitulado **"SISTEMA PARA APOIO AO DIAGNÓSTICO DE CÂNCER DE COLO DO ÚTERO BASEADO EM TÉCNICAS COMPUTACIONAIS DE PROCESSAMENTO E ANÁLISE DE IMAGENS HISTOPATOLÓGICAS"** foi analisado pelo Comitê de Ética em Pesquisa, em sua 313ª Reunião Ordinária realizada em 04/10/2010 e enquadrado na categoria: **APROVADO,** de acordo com o Processo HCRP nº 11761/2010.

Este Comitê segue integralmente a Conferência Internacional de Harmonização de Boas Práticas Clínicas (IGH-GCP), bem como a Resolução nº 196/96 CNS/MS.

Lembramos que devem ser apresentados a este CEP, o Relatório Parcial e o Relatório Final da pesquisa.

Atenciosamente.

Marcia Villanova

DRª MARCIA GUIMARÃES VILLANOVA
Vice-Coordenadora do Comitê de Ética em Pesquisa do HCRP e da FMRP-USP

Ilustríssima Senhora
GISELE HELENA BARBONI MIRANDA
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto-USP – Depto. de Física e Matemática

Comitê de Ética em Pesquisa HCRP e FMRP-USP - Campus Universitário
FWA – 0000 2733; IRB – 0000 2186 e Registro SISNEP/CONEP nº 4
Fone (16) 3602-2228 - E-mail : cep@hcrp.usp.br
Monte Alegre 14048-900 Ribeirão Preto SP