Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

PathoSpotter: Um Sistema para Classificação de Glomerulopatias a partir de Imagens Histológicas Renais

George Oliveira Barros

Feira de Santana

Fevereiro, 2016

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

George Oliveira Barros

PathoSpotter: Um Sistema para Classificação de Glomerulopatias a partir de Imagens Histológicas Renais

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Angelo Amâncio Duarte

Coorientador: Washington Luís Conrado dos Santos

Feira de Santana

Fevereiro, 2016

Ficha Catalográfica – Biblioteca Central Julieta Carteado

Barros, George Oliveira
B274p PathoSpotter: um sistema para classificação de glomerulopatias a partir de imagens histológicas renais / George Oliveira Barros. - Feira de Santana, 2016.

108 f.: il.

Orientador: Angelo Amâncio Duarte
Coorientador: Washinton Luís Conrado dos Santos

Dissertação (Mestrado) – Universidade Estadual de Feira de Santana, Programa de Pós-graduação em Computação Aplicada, 2016.

1. Computação – PathoSpotter. 2. Histopatologia digital. 3. Imagens Histológicas Renais. I. Duarte, Angelo Amâncio, orient. II. Santos, Washington Luís Conrado dos, coorient. III. Universidade Estadual de Feira de Santana. IV. Título.

CDU: 681.3

George Oliveira Barros

PathoSpotter: Um Sistema para Classificação de Glomerulopatias a partir de Imagens Histológicas Renais

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Feira de Santana, 29 de Fevereiro de 2016.

BANCA EXAMINADORA

Dr. Angelo Amâncio Duarte (Orientador)
Universidade Estadual de Feira de Santana

Drª Michele Fúlvia Angelo
Universidade Estadual de Feira de Santana

Dr. Sérgio Marcos Arruda
Fundação Oswaldo Cruz

Abstract

The realization of an accurate diagnosis from histological images requires pathologists with practical experience because the characteristics of these images lead to a subjective analysis, which often hamper the accuracy of diagnosis. Systems that help to achieve better diagnoses can minimize doubts and improve the quality of diagnosis, influencing on increasing the effectiveness of medical treatments. This paper describes the research and development of PathoSpotter, a computer system to aid in the identification of diseases from histological images. The PathoSpotter proposes to reduce the lack of support work to histopathological diagnosis of renal diseases since much has been done in the area of cancer, but there is few published material in relation to the Digital Pathology applied to nephrology and hepatology. Our goal in this study was to apply the PathoSpotter the classification of proliferative glomerulopathy, which is a family of primary diseases affecting the kidneys. The work was based on a data set consisting of 811 histological pictures glomeruli and classical techniques of processing digital images and histopathology were used. The PathoSpotter presented a performance of 88.4% accuracy, which was similar to other Digital Pathology jobs that can be found in the literature.

Keywords: Glomerulopathy, Digital Image Processing, Machine Learning, Digital Histopathology.

Resumo

A realização do diagnóstico preciso a partir de imagens histológicas requer médicos patologistas com vasta experiência prática, pois as características dessas imagens conduzem a uma análise subjetiva que muitas vezes dificultam a exatidão do diagnóstico. Sistemas que auxiliam a obtenção de melhores diagnósticos podem minimizar dúvidas e melhorar a qualidade dos diagnósticos, influenciando no aumento da eficácia dos tratamentos médicos. Este trabalho descreve a pesquisa e o desenvolvimento do PathoSpotter, um sistema computacional para auxílio na identificação de patologias a partir de imagens histológicas. O PathoSpotter se propõe a reduzir a carência de trabalhos de apoio ao diagnóstico histopatológico das doenças renais, já que muito tem sido feito na área de neoplasias, mas há pouco material publicado em relação à Patologia Digital aplicada à nefrologia ou hepatologia. Nosso objetivo neste trabalho foi aplicar o PathoSpotter na classificação das glomerulopatias proliferativas, que é uma família de doenças primárias que afetam os rins. O trabalho se baseou em um conjunto de dados composto por 811 imagens histológicas de glomérulos, e foram utilizadas técnicas clássicas de processamento de imagens e histopatologia digital. O PathoSpotter apresentou um desempenho de 88,4% de acurácia, resultado similar ao de outros trabalhos de Patologia Digital que podem ser encontrados na literatura especializada.

Palavras-chave: Glomerulopatias, Processamento Digital de Imagens, Aprendizado de Máquina, Histopatologia Digital.

Prefácio

Esta dissertação de mestrado foi submetida à Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para a obtenção do grau de Mestre em Computação Aplicada.

A dissertação foi desenvolvida dentro do Programa de Pós-Graduação em Computação Aplicada (PGCA) tendo como orientador o Dr. Angelo Amâncio Duarte e coorientador o Dr. Washington Luís Conrado dos Santos.

Esta pesquisa foi financiada por uma bolsa de estudos fornecida pela CAPES.

Agradecimentos

Ao meu orientador Angelo Amâncio Duarte por todos os ensinamentos, apoio e colaboração.

Ao meu coorientador Washington Luis Conrado dos Santos.

Aos colegas do Programa de Pós-graduação em Computação Aplicada da UEFS.

A minha família, especialmente os meus pais, Higino Barros Meira e Jizonete Oliveira Silva Barros.

A Deus, por toda graça, bondade e misericórdia.

Sumário

Lista de Publicações

BARROS, G. O.; DUARTE, A. A.; DOS SANTOS, W. L. C. PathoSpotter: Um Sistema para Classificação de Glomerulopatias a partir de Imagens Histológicas Renais. In: CONFERENCE ON GRAPHICS, PATTERNS AND IMAGES, 28. (SIBGRAPI), 2015, Salvador. Proceedings... Porto Alegre: Sociedade Brasileira de Computação, 2015. On-line. Available from: <http://urlib.net/8JMKD3MGPBW34M/3JRK2U5>. Access in: 2015, Oct. 16.

Lista de Tabelas

Lista de Figuras

Lista de Abreviações

Abreviação	**Descrição**
H&E	Hematoxilyn and Eosin
IHC	Immunosthochemical
RGB	Red-Green-Blue
LoG	Laplacian of Gaussian
DAB	Diaminobenzidina
HED	Hematoxylin–Eosin–DAB
LUV	Luminescence, Saturation e Hue Angle
LAB	Luminescence, A=red/green e B=blue/yellow
HSV	Hue, Saturation e Value
kNN	*k* Nearest Neighbors
SVM	Suport Vector Machine

Capítulo 1

Introdução

"O começo é a parte mais difícil do trabalho."

-- Platão

As últimas décadas têm evidenciado a computação como uma ferramenta fundamental para o desenvolvimento da ciência e tecnologia, possibilitando avanços em diversas áreas do conhecimento. A visão computacional, especificamente, tem possibilitado avanços em diferentes áreas, como Agronomia, Biologia, Química e Geologia, além de aplicações militares e industriais [Bougouma *et al.* 2013; Meyer e Camargo Neto, 2008; Danuser, 2011; Młynarczuk *et al.* 2013; Karacor *et al.* 2011; Shahzad *et al.* 2015; Baravalle *et al.* 2015].

Entre as diferentes áreas de aplicação da visão computacional, as contribuições originadas através da parceria entre a computação e a medicina têm impactado a prática médica influenciando principalmente a qualidade dos diagnósticos. Segundo Ritter *et al.* [2011], a análise de dados de pacientes adquiridos por dispositivos de coleta de imagens médicas, como tomografia computadorizada (CT), tomografia de ressonância magnética (MRT), tomografia por emissão de pósitrons (PET), ou ultrassom, têm oferecido oportunidades nunca alcançadas antes, para os processos de diagnóstico e prognóstico.

Entre os trabalhos com imagens médicas, a área de histopatologia digital destaca-se como uma das maiores evoluções da medicina moderna [Irshard *et al.* 2014]. A histopatologia digital pode ser compreendida a partir dos conceitos de duas outras áreas, a histologia e a histopatologia. A histologia é o estudo da anatomia microscópica de tecidos de organismos. A histopatologia, por sua vez, é a análise microscópica de seções histológicas com o objetivo da diferenciação entre tecidos biológicos saudáveis e doentes, auxiliando assim no diagnóstico e prognóstico de patologias [Belsare e Mushirif, 2012; Lei He *et al.* 2012].

Para a realização de uma análise histológica de amostras de tecidos biológicos, os patologistas

extraem amostras de tecidos através de biópsia e examinam secções através de microscópio. O diagnóstico de patologias a partir de imagens contitui, até o presente momento, o "padrão ouro" para o diagnóstico de uma série de doenças, incluindo as neoplasias (tumores e cânceres) [Irshard *et al.* 2014].

O campo de estudo de histopatologia digital teve seu início simbólico com o surgimento da possibilidade de visualizar seções histológicas não apenas ao microscópio, mas através de monitores de computador. Na década de 90, Dirk G. Soenksen deu inicio ao que seria, naquele momento, um novo futuro para área de investigação de patologias por imagem. Soenksen criou a empresa Aperio que movia imagens de seções histológicas de microscópios para computadores [May, 2010]. Contudo, o número de propostas de trabalhos na área de histopatologia digital só aumentaria entre os anos 1990 e 2000, quando os investigadores das áreas de processamento de imagens e visão computacional de fato aceitariam o desafio de propor sistemas automáticos para análise histopatológica [Meijering *et al.* 2012]. Segundo Gurcan *et al.* [2009], o crescimento do número de trabalhos na área de histopatologia digital aumentou graças aos avanços da capacidade computacional de armazenamento de dados e da elaboração de bancos de amostras de tecidos histológicos digitalizados.

Atualmente é possível usar padrões histológicos com análise de imagem assistida por computador para facilitar a identificação de patologias [Belsare e Mushirif, 2012]. Os sistemas automáticos de diagnóstico por imagens histológicas podem apoiar as decisões dos patologistas sobre a presença ou ausência de uma patologia, sobretudo aumentando a eficiência e a precisão do diagnóstico médico e assim, reduzindo a subjetividade da análise. Adicionalmente, tais sistemas podem aperfeiçoar as tarefas de armazenamento e compartilhamento de informações patológicas, contribuindo para pesquisas científicas. Por fim, o sistema pode se tornar uma ferramenta com objetivos de apoio ao diagnóstico e auxílio didático para a formação de novos patologistas [Belsare e Mushirif, 2012; May, 2010; Irshad *et al.* 2014].

Gurcan *et al.* [2009] ilustram um exemplo da possível contribuição de sistemas automáticos de apoio ao diagnóstico, considerando que sendo cerca de 80% das biópsias de câncer de próstata examinadas nos Estados Unidos anualmente são benignas, se os patologistas tivessem o apoio de uma ferramenta de auxílio a sua tarefa, eles se concentrariam em casos de decisão mais difícil, poupando uma grande parcela de tempo e esforço.

Até o momento em que esse texto foi escrito, encontram-se na literatura científica diversos

trabalhos na área de histopatologia digital objetivando a automação da análise e interpretação de imagens médicas através da segmentação de células, para identificação e classificação de diferentes estruturas biológicas ou de patologias em diferentes órgãos do corpo humano [Gurcan *et al.* 2009; Irshad *et al.* 2014]. Contudo, nota-se uma grande concentração de trabalhos para o estudo de neoplasias, com pouca quantidade dedicada a outras patologias [Lei He *et al.* 2012; Belsare e Mushirif, 2012; Gurcan *et al.* 2009].

Apesar de diferentes doenças já serem diagnosticadas com o auxílio de sistemas de apoio ao diagnóstico [Kothari *et al.* 2013; Schöchlin *et al.* 2014; Sirinukunwattana *et al.* 2014], muitas ainda não contam com esse ganho oferecido pela histopatologia digital, como é o caso das glomerulopatias. Cohen e Glassok [1999, cap.3] definem as glomerulopatias, especificamente as primárias, como desordens que afetam a função e/ou estrutura dos glomérulos, que são as principais estruturas responsáveis pela filtragem do sangue nos rins. Segundo a OMS (Organização Mundial de Saúde), em dados apenas de 2004, condições descritas como nefrite ou nefrose, que incluem as glomerulopatias, estiveram associadas à morte de 739 mil pessoas [WHO, 2004, p.58]. No Brasil, estudos apontam que a incidência das glomerulopatias vem aumentando nas últimas décadas [Polito *et al*. 2010; Woo *et al*. 2010].

Atualmente, existe um sistema geral de classificação morfológica das doenças renais baseado nas características das lesões glomerulares. Apesar da sua importância para o tratamento e compreensão da fisiopatologia das glomerulopatias, essa classificação tem se mostrado insuficiente para definição das enfermidades renais [D'Agati, 2003; Weening *et al.* 2004]. Adicionalmente, estudos sistematicamente chamam a atenção para a necessidade de potenciais mudanças na ênfase conferida a determinados padrões de lesão glomerular, como definidores de diagnóstico e prognóstico de doenças renais. A despeito da relevância da classificação de glomerulopatias por imagens, até o momento em que este texto foi escrito, não foram encontrados na literatura trabalhos que proponham um sistema automático para a identificação e classificação das glomerulopatias.

Tendo em vista a carência de propostas de sistemas de classificação automática de glomerulopatias e os benefícios que trabalhos dessa natureza podem trazer à prática médica, o objetivo deste trabalho foi propor um sistema de apoio ao diagnóstico das glomerulopatias. Para alcançar este objetivo foi necessária a cooperação entre pesquisadores da área de computação e a área médica, compreender o estado da arte de trabalhos da área de histopatologia digital, e por fim, construir um sistema capaz de classificar imagens de

glomérulos renais quanto a presença das glomerulopatias proliferativas, o qual denominamos de PathoSpotter.

Pelo fato de não haverem trabalhos similares com quais fosse possível comparar o desempenho do PathoSpotter, foram utilizados como referência os trabalhos relacionados à classificação de neoplasias e estruturas biológicas, tanto para a investigação de métodos computacionais que poderiam ser utilizados na construção do PathoSpotter, quanto para a comparação de resultados finais obtidos por este na classificação das imagens histológicas.

Demonstraremos que, ao final da realização de todos os experimentos e implementação do sistema, os resultados obtidos através do PathoSpotter foram similares ou superiores aos resultados revelados em trabalhos de classificação de outras patologias, o que indica um futuro promissor para o desenvolvimento e aplicação do PathoSpotter em outras áreas da histopatologia digital.

1.1 Contextualização sobre as Glomerulopatias

As glomerulopatias são uma família de patologias renais. Essas patologias são caracterizadas por danos nos glomérulos, que por sua vez, são grupos de capilares pelos quais o sangue é filtrado. As glomerulopatias podem ter origem nos rins, sendo chamadas de primárias, ou podem ser secundárias a outras doenças, como diabetes, hepatites, doenças autoimunes, dentre outras [Barros *et al.* 2006; Guyton e Hall, 2006, p.309].

Diferentes estudos analisam a distribuição das glomerulopatias em suas diferentes formas, no Brasil e em um âmbito mundial. Polito *et al.* [2010] revelam que nas últimas décadas a incidência de doenças glomerulares vem aumentando no Brasil. Woo *et al.* [2010] realizaram uma comparação entre os casos de glomerulonefrite primária prevalentes em Cingapura e outros 28 países. Através de dados oriundos das últimas três décadas, o estudo revelou que em todo o mundo a prevalência de glomeruloesclerose segmentar e focal continua a crescer. Por fim, McGrogan *et al.* [2011] analisaram dados dos anos entre 1980 e 2010, incluindo 40 estudos de incidência das glomerulopatias primárias na Europa, América do Norte e do Sul, além da Austrália e Oriente Médio. Os autores concluíram que a taxa de incidência de glomerulonefrite primária varia entre 0.2/100 mil e 2.5/100 mil casos por ano.

A importância da função glomerular na fisiologia renal, e o fato de suas lesões afetarem outros segmentos do néfron, que é a menor unidade renal responsável pela filtração e

formação da urina, caracterizam as enfermidades glomerulares como um dos principais problemas da área de Nefrologia nos dias atuais [Alves Júnior *et al.* 2008; Guyton e Hall, 2006, p.310].

É crescente o surgimento de registros de glomerulopatias em diversos países [Castro *et al.* 2002]. Na Ásia e Oceania, as glomerulonefrites (tipo de glomerulopatia) são a causa de insuficiência renal entre 30% a 60% dos pacientes admitidos para tratamento dialítico. Na Europa e nos Estados Unidos esse valor está entre 10% a 15%. No Uruguai, cerca de 20% dos pacientes recebem o diagnóstico de glomerulonefrite na admissão para diálise [Queiroz *et al.* 2009; Bahiense-Oliveira e Malafronte, 2006]. Na Arábia Saudita, em um levantamento realizado entre os anos 1989 e 2007, dados de 568 casos de doença renal revelaram que 52,1% desses casos de doenças renais eram equivalentes a glomerulopatias primárias [Jalalah, 2009].

No Brasil, as glomerulopatias são a terceira causa de doença renal crônica. O que constitui uma etiologia frequente de insuficiência renal crônica dialítica, tendo a realização da biopsia renal um papel fundamental no correto diagnóstico e etiológico, e mesmo no prognóstico desses casos, apesar do entendimento de que a biópsia renal não deve ser analisada de forma isolada, pois o fator clínico de um paciente é o fator base para uma indicação [Alves Júnior *et al.* 2008; Ferrazi, 2010]. De tal modo, o diagnóstico das glomerulopatias é realizado com base em dados clínicos e laboratoriais (bioquímica do soro e exame de urina) e a análise histológica de biópsias renais. As biópsias renais, especialmente, constituem pequenos fragmentos do rim, obtidas por agulha ou cirurgicamente. Esses tecidos são fixados, cortados em secções de 2-3 μm de espessura e examinados no microscópico [Al Kofahi *et al.* 2010].

O investimento em novos estudos com o caráter epidemiológico sobre as glomerulopatias (no sentido de compreender as características fisiopatológicas da doença) e em pesquisas de novos paradigmas para a resolução dos problemas com glomerulopatias, reforçam a necessidade da implementação de novos estudos sobre doenças glomerulares no Brasil, ao se considerar a heterogeneidade da população brasileira em suas características étnicas, socioeconômicas e geográficas, além da necessidade de se obter métodos mais rápidos para a realização do diagnóstico [Alves Júnior *et al.* 2008; Lopes *et al.* 2001; Bahiense-Oliveira e Malafronte, 2006].

1.2 Contribuições deste Trabalho

As principais contribuições deste trabalho são:

- Auxílio na redução da subjetividade do diagnóstico de glomerulopatias através de um sistema automático de classificação;
- Produção de uma ferramenta computacional para o auxilio do processo de formação de novos patologistas;
- Contribuição científica para a área de histopatologia digital, ao aplicar métodos computacionais clássicos em um campo de estudo pouco explorado.

1.3 Organização da Dissertação

Este trabalho está organizado em 6 capítulos. No Capítulo 2, contextualizamos a histopatologia digital. No capítulo 3, falamos sobre os sistemas de classificação de imagens digitais. No capítulo 4 apresentamos o PathoSpotter, detalhando sua arquitetura e métodos computacionais utilizados. No capítulo 5, revelamos os experimentos realizados para construir e avaliar o sistema, além de apresentar os resultados parciais e finais do sistema. Por fim, no capítulo 6, apresentamos nossas conclusões, discutindo sobre os resultados e trabalhos futuros.

Capítulo 2

Histopatologia Digital

"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original."

-- Albert Einstein

Na área médica, a computação tem se mostrado uma ferramenta de grande valia, tornando os diagnósticos mais precisos e propiciando vários avanços para a medicina mundial [Ritter *et al.* 2011]. Estudos vêm sendo realizados com o objetivo de melhorar o desempenho dos atuais sistemas computacionais aplicados no apoio a diagnósticos médicos e classificação de diferentes patologias. No entanto, sabe-se que ainda há carências neste campo, o que gera a expectativa de que novas contribuições continuem a ser oferecidas [Chabat *et al.* 2000; Belsare e Mushrif, 2012].

De modo geral, os sistemas de análise de imagens médicas têm como propósito aumentar a percepção de determinadas características, extraindo e apresentando relevantes informações sobre imagens [Ritter *et al.* 2011]. Os avanços da visão computacional na medicina destacam-se principalmente em relação à instrumentação, diagnósticos e aplicações terapêuticas, sendo que a maioria desses feitos é baseada em análise de imagens digitais [Deserno *et al.* 2013].

Alguns exemplos de sistemas computacionais para apoio ao diagnóstico médico, mais recentes aparecem no trabalho de, Mas *et al.* [2015], que apresentam um algoritmo capaz de identificar malária a partir de imagens *in vitro*; Prabusankarlal *et al.* [2015], que realizam um estudo sobre o desempenho de sistemas de identificação de nódulos mamários a partir de imagens de ultrassom, combinando características de textura e aspectos morfológicos; e Cheng e Mandal [2015], que propõem um sistema de classificação de imagens, o qual segmenta determinadas partes de imagens histológicas e identifica células cancerosas, com o objetivo de apoiar o processo de diagnóstico de melanoma, que é o tipo mais agressivo de câncer de pele.

Dentre as subáreas de pesquisa computacional de imagens médicas, um dos segmentos que se destaca é a histopatologia digital. Irshard *et al.* [2014], evidenciam a área de histopatologia

digital como uma das maiores evoluções da medicina moderna.

A histopatologia digital pode ser compreendida a partir dos conceitos de histologia e histopatologia. A histologia é o estudo da anatomia microscópica de tecidos de organismos. Para a realização da análise histológica de amostras de tecidos biológicos, os patologistas extraem amostras de tecidos através de biópsia e examinam seções da amostra extraída através de microscópio. De modo similar, a histopatologia é o estudo microscópico de seções histológicas (pequeno corte transversal de tecido) com o objetivo da diferenciação entre tecidos normais e doentes, auxiliando assim no diagnóstico e prognóstico de patologias [Belsare e Mushirif, 2012; Lei He *et al.* 2012]. Por fim, através dos trabalhos de revisão de Meijering [2012], Gurcan *et al.* [2009] e Lei He *et al.* [2012], torna-se possível resumir a histopatologia digital como o estudo de métodos computacionais para a realização da análise e interpretação de patologias através de imagens histológicas.

2.1 Análise de Imagens Digitais Histológicas

As imagens histológicas são adquiridas a partir de biópsia[1], e os patologistas analisam as imagens através de um exame microscópico, no qual se observa uma série de detalhes na imagem: a estrutura do tecido, a distribuição das células, formas celulares e características citológicas de malignidade. Esse processo é demorado e propenso à subjetividade do observador [Belsare e Mushrif, 2012].

As técnicas de processamento de imagens histológicas podem ser aplicadas a diferentes patologias do sistema do corpo humano: identificação de tumores; câncer perirenal, mama, próstata, pulmão; boca, dentre outros [Belsare e Mushrif, 2012; Lei He *et al.* 2012].

Alguns estudos têm como objetivo diferenciar imagens de tecidos normais e anormais, enquanto outros classificam os tecidos em relação a determinado grau de uma doença, havendo a possibilidade de identificar uma imagem dentro de uma variedade de subcategorias patológicas. Cada patologia possui características únicas, decorrentes das diferenças entre as estruturas dos tecidos e células de cada órgão, além de existir diferentes técnicas de aquisição dessas imagens. Portanto, trabalhos com imagens histológicas se tornam mais complexos do que outros com diferentes tipos de imagens médicas, como é o caso das imagens radiológicas, por exemplo [Lei He *et al.* 2012].

[1] Retirada de material celular ou de um fragmento de tecido de um ser vivo para fins de diagnóstico.

Destacamos a seguir, exemplos de trabalhos com imagens histológicas, ilustrando o cenário geral dos objetivos de estudo neste campo. Algo que merece destaque é que os trabalhos com câncer (neoplasias) aparecem com grande frequência entre os estudos de segmentação e novas propostas de aperfeiçoamento da representação de imagens histológicas.

Com o objetivo de representar imagens histológicas de uma forma que as informações dos corantes utilizados na aquisição das imagens fossem extraídos em forma de intensidade de pixel, Ruifrok e Johnston [2001], propuseram o método deconvolução de cor (*color deconvolution),* comumente utilizado em trabalhos com imagens histológicas. Os trabalhos de Veillard *et al.* [2013] e Wang [2011], são exemplos de trabalhos que utilizam o método deconvolução de cor para representar imagens a partir das informações dos corantes utilizados. Devido ao seu uso frequente em histopatologia digital, Van der Walt *et al.* [2014], implementaram o método deconvolução de cor na biblioteca de processamento de imagens *scikit-image*, na forma de espaço de cor, o HED, que é constituído por três canais de cor, que armazenam as informações dos corantes *Hematoxylin*, *Eosin* e *Diaminobenzidina* (DAB).

Gavrilovic *et al.* [2013], propuseram um método para decomposição de cores em imagens histológicas para representar às imagens adquiridas com diferentes corantes (H&E, H&H e G&G), de forma que cada pixel da imagem possua uma intensidade adequada à estrutura biológica a qual faz parte. Entre as técnicas utilizadas estão: deconvolução de cor (de Ruifrok e Johnton [2001]) e a criação de modelos matemáticos para transformação de intensidade de pixels. Segundo os autores, o método proporciona uma relevante melhoria na representação das imagens histológicas.

Zarella *et al.* [2015], apresentaram um trabalho com objetivos parecidos aos de Gavrilovic *et al.* [2013]. Eles propuseram um método para melhorar a qualidade de representação das imagens coradas em H&E, padronizando a representação colorida de imagens histológicas. O método realiza um mapeamento e indica a probabilidade de cada pixel fazer parte de uma determinada estrutura biológica da imagem. Este trabalho utilizou um conjunto de dados com 44 imagens histológicas com câncer de mama. Os pixels foram classificados em quatro diferentes classes, equivalentes às estruturas biológicas da imagem, obtendo para cada classe, acurácias de 56%, 95%, 92% e 76%, na atribuição da cor adequada para cada pixel. Segundo os autores, os resultados são bons e esta técnica semiautomática pode auxiliar na identificação de estruturas biológicas e ajudar a calibrar a coloração de lâminas de tecido com problema.

Entre os trabalhos com propostas de segmentação, Wang [2011] propôs um método de

segmentação de estruturas biológicas em imagens histológicas de câncer no pulmão. O objetivo primordial do trabalho foi criar um novo marcador automático de segmentação de estruturas biológicas. A avaliação do sistema foi realizada através de dois conjuntos de dados, um formado por imagens coradas em H&E (9 imagens), atingindo 80% de acurácia, e outro formado por imagens coradas em IHC (também 9 imagens), atingindo 78% de acurácia. Entre os métodos computacionais utilizados neste trabalho destacam-se: deconvolução de cor, operadores morfológicos e manipulação de histogramas.

Mouelhi *et al.* [2013], apresentaram um método de segmentação de células de imagens de seções histológicas do tecido da mama, para auxílio ao diagnóstico de câncer de mama. Entre os métodos utilizados no trabalho estão a segmentação por divisor de águas (*watershed*), reconhecimento de borda, *Multilayer Neural Network* (MNN) e Fisher's *linear discriminant* (FDL). O conjunto de dados utilizado contou com 480 imagens histológicas da mama e o resultado obtido foi uma segmentação com 97,8%, em comparação a avaliação feita por patologistas.

Em relação aos trabalhos com o foco na classificação de patologias, Tabesh *et al.* [2007], apresenta um sistema de apoio ao diagnóstico de câncer de próstata. A avaliação do trabalho é realizada com dois conjuntos de dados distintos, ambos com imagens coradas em H&E. Um conjunto de dados foi formado por 367 imagens (com imagens onde a área de câncer equivalente entre 5% a 100% de toda a imagem) e outro conjunto de dados foi formado por 268 imagens (onde a área de câncer equivalia a no mínimo 80% de toda a área da imagem). Os resultados alcançados para cada um dos conjuntos de dados foram de 97% e 81% de acurácia, respectivamente. Entre os métodos utilizados destacam-se: filtros gaussianos, descritores de cor, morfologia e textura, além dos classificadores kNN (*k-Nearest Neighbors*) e SVM (*Support Vector Machines*).

Sharma *et al.* [2012], apresentaram um método de consulta de similaridade entre imagens histológicas. Tal método utiliza a teoria dos grafos para extrair características das imagens. Neste trabalho foram utilizados quatro conjuntos de dados. O primeiro conjunto de dados com 70869 imagens (64x64 pixels), o segundo conjunto de dados com 27596 imagens (128x128 pixels), o terceiro conjunto de dados com 9132 imagens (256x256 pixels) e o quarto conjunto de dados com 2485 (512x512 pixels), sendo todos os conjuntos de dados constituídos de imagens de câncer na mama. O método obteve resultado de 67% de acurácia, o que, segundo os autores, equivaleu a uma melhora de 81% em relação ao resultado obtido em métodos

anteriores, baseados em informações de histogramas.

Miranda *et al.* [2012] propuseram um método de classificação de imagens histológicas de câncer no colo do uterino. O objetivo do trabalho foi promover o apoio ao diagnóstico realizado pelos patologistas. O método visou classificar as imagens em três possíveis classes (C1, C2, e C3). As técnicas de segmentação e pré-processamento utilizadas foram: operadores morfológicos, teoria dos grafos (modelo de *Voronoi*) e limiarização automática. Entre as características extraídas estão: entropia, taxa de ocupação de pixels e média de grau. Por fim, a classificação foi realizada por método de aprendizado por agrupamento (*clustering),* apresentando o resultado de 73% acurácia em testes realizados com um conjunto de dados formado por 144 imagens, rotuladas em três diferentes classes.

No trabalho de Mathur *et al.* [2013], realizou-se a classificação de glóbulos brancos em relação cinco possíveis classes (*basophils, eosinophils, lymphocytes, monocytes, neutrophils*). A segmentação foi realizada através da conversão de espaço de cores RGB (espaço de cor *Red-Green-Blue*) para HSV (espaço de cor *Hue-Saturation-Value*), além de segmentação automática e operadores morfológicos como filtros. As características utilizadas basearam-se em informações de forma, textura, e descritores de núcleos. Por fim, a classificação foi realizada com o classificador *Naive Bayes*, obtendo um resultado de 92% de acurácia, em um teste realizado com um conjunto de dados de 267 amostras (80% como conjunto de treinamento e 20% como conjunto de teste).

Sirinukunwattana *et al.* [2014], realizaram um trabalho de detecção e classificação de células em imagens histológicas da mama, diferenciando células mitóticas de não mitóticas. Segundo os próprios autores, esta é uma tarefa difícil, dada a variabilidade das estruturas biológicas nas imagens, fazendo com que a automação proposta fosse importante para o processo posterior de identificação de câncer de mama. As principais técnicas utilizadas foram: Limiarização simples, crescimento de regiões, identificação de borda e filtros gaussianos. A etapa de classificação foi realizada com redes neurais, obtendo-se, como melhor resultado, 86% de acurácia. A avaliação foi realizada através da divisão de 10 subconjuntos de treinamento/teste (validação cruzada *10-fold*) sendo utilizado um conjunto de dados com 50 imagens oriundas de um banco de dados público, validado pela comunidade científica.

Por fim, ainda citamos o trabalho de Schöchlin *et al.* [2014], que propôs um classificador chamado de *nuclear circularity-based classifier*, desenvolvido especificamente para o auxilio da tarefa de identificação de melanoma (câncer de pele) através da classificação de duas

classes de células (*cell melanoma* e *desmoplastic melanoma*). Este trabalho apresenta 88,9% de acurácia, utilizando um conjunto de dados com 18 imagens. Entre os métodos que compõem o sistema destacamos a limiarização automática, deconvolução de cor na etapa de segmentação dos núcleos e a utilização das características: circularidade e raio.

2.2 Análise de Imagens Digitais Histológicas Renais

No que diz respeito à quantidade de trabalhos com imagens histológicas renais, o cenário não é muito diferente dos trabalhos com imagens histológicas de outros órgãos. As propostas mais comuns neste campo estão voltadas ao estudo de métodos para resolução de problemas de identificação de neoplasias. Os trabalhos apresentados a seguir comprovam essa afirmação.

Isitor e Thorne [2007] utilizaram técnicas de segmentação por pixel e textura, com o objetivo de estabelecer um ponto de referência para a identificação rápida de multiplicação de células em imagens histológicas renais de mamíferos, estudando assim, a evolução celular de tecidos suíno, bovino, de ratos e humanos.

Kothari *et al.* [2011], apresentaram um sistema de extração de características de diferentes imagens de câncer renal para auxiliar no diagnóstico. As características possibilitaram que os patologistas classificassem as imagens entre seis tipos de câncer renal, utilizando um conjunto de dados com 58 imagens. Em um trabalho similar, Kothari *et al.* [2013] propuseram um método automático de identificação e classificação de imagens histológicas de tumores renais, superando e complementando os modelos propostos anteriormente, conseguindo exatidão máxima de 77% em um teste realizado com 151 imagens de quatro diferentes tipos de câncer renal. Entre os principais métodos utilizados estão a segmentação por máscara de cor, descritores de forma e textura e o classificador SVM.

Stewart *et al.* [2014], propuseram um método computacional para diferenciar dois tipos de tumores renais, o *chromophobe renal cell carcinoma* e *renal oncocytoma*. Estes tumores, segundo os autores, apresentam características confusas utilizando métodos histopatológicos convencionais. O sistema proposto atingiu resultados finais de 86% de sensibilidade e 81% especificidade, utilizando um conjunto de dados formado por amostras extraídas a partir de 88 pacientes.

Tae-Yun Kim *et al*. [2014] aplicaram métodos de análise de textura tridimensionais para a extração de características contidas em imagens de células do tecido renal com o propósito de

facilitar o processo de identificação de carcinomas pela observação visual dos patologistas.

Por fim, em relação aos trabalhos com imagens histológicas renais que não tratam de neoplasias (câncer e tumores), pode-se citar Cui *et al.* [2012], que realizaram uma comparação de imagens renais histológicas e tomográficas com o objetivo de melhorar a compreensão e reconhecimento de novas categorias de tumores renais, e Herrmann *et al.* [2012], que propuseram um método semiautomático de quantificação de células epiteliais de glomérulos, as quais representam um importante papel na função glomerular. Esta análise semiautomática de quantificação produziu um marcador válido e sensível para o estudo de dano glomerular.

2.3 Propostas Similares ao Trabalho Atual

O trabalho de revisão de Gurcan *et al*. [2009] ratifica a importância de pesquisas que proponham sistemas de análise automática de imagens histológicas. Tanto os pesquisadores de análise de imagens quanto patologistas reconhecem a importância de serem desenvolvidos métodos automáticos de análise de imagens cada vez mais eficientes e precisos, não apenas com foco em aplicações clínicas, mas também objetivando pesquisas científicas. Há uma vasta quantidade de trabalhos de análise histológica para cada etapa de sistemas de análise de imagens (aquisição, pré-processamento, segmentação, extração de características e classificação).

Como se pôde perceber nos trabalhos citados nas seções 2.1 e 2.2, a maioria dos trabalhos com imagens histológicas, realizados até o momento e disponíveis na literatura, estão voltados para a detecção de neoplasias malignas, tendo como principais objetivos a melhoria no processo de representação, segmentação e (ou) classificação de imagens.

Os trabalhos de aquisição e pré-processamento tem como principal objetivo melhorar a representação digital das imagens histológicas e proporcionar métodos eficazes de eliminação de ruídos. Entre os trabalhos propostos destacam-se os métodos de representação em cores (deconvolução de cor), modelagem 3D, operações entre espaços de cor (RGB, HSV e outros), suavização, redução de ruído e aperfeiçoamento de bordas [Gurcan *et al.* 2009; Belsare e Mushirif, 2012].

Os trabalhos de segmentação visam extrair determinadas estruturas biológicas dos tecidos e são motivados pela necessidade de se realizar a contagem de núcleos ou observação particular

das estruturas. Os trabalhos de segmentação de células, em especial, são bastante numerosos. Uma justificativa para tal quantidade é o fato do processo de segmentação ser uma etapa prévia e de essencial importância para as etapas de extração de características e classificação.

Entre os principais métodos computacionais utilizados para segmentar as células destacam-se: Limiarização de Otsu, operações morfológicas de filtragem e reconstrução, laplaciano de gaussiano, métodos de aprendizado por agrupamento (*clustering*), teoria dos grafos, segmentação por divisor de águas e vários outros. Apesar de serem utilizadas várias técnicas nos processos de segmentação, a principal motivação dos trabalhos encontrados foi oferecer suporte a sistemas de apoio ao diagnóstico de câncer [Gurcan *et al.* 2009; Meijering, 2012; Belsare e Mushirif, 2012; Irshad *et al.* 2014].

Em relação à etapa de extração de características, os recursos úteis para classificação de uma determinada patologia são inspirados, geralmente, pelos dados visuais definidos pelos próprios patologistas. No entanto, como a capacidade humana de interpretação e compreensão de imagens é diferente da capacidade de um computador, assim como o modo como as imagens são representadas na mente humana são diferentes da representação digital (matrizes e pixels), em vários casos, se faz necessária a procura por diferentes características (marcadores) que elevassem a capacidade computacional de interpretação [Gurcan *et al.* 2009]. Segundo Belsare e Mushirif [2012], em um âmbito mais amplo, as técnicas presentes na literatura utilizadas na extração características concentram-se nos aspectos de cor, morfologia, textura, intensidade de brilho e outras diversas informações, específicas ao domínio do problema.

Para a etapa de classificação existem trabalhos com classificadores clássicos (redes neurais, SVM, kNN e sistemas *fuzzy*) ou exclusivamente desenvolvidos para a resolução de problemas específicos. A escolha de um classificador pode variar de acordo com a quantidade de amostras, características, ou especificações do problema. Os trabalhos de classificação de imagens histológicas estão centrados desde a identificação de um tipo de célula ou estrutura, até a identificação e classificação de uma patologia, sendo encontrados casos de classificação binária e multiclasse. Além disso, pode-se destacar que é comum tanto abordagens de aprendizado supervisionado, quanto aprendizado não supervisionado [Gurcan *et al.* 2009; Belsare e Mushirif, 2012; Irshad *et al.* 2014]. Adicionalmente, Lei He *et al.* [2012] ratificam a grande incidência de trabalhos com foco na identificação e classificação de carcinomas, apresentando uma revisão de trabalhos com esse caráter. Os principais órgãos (ou regiões)

estudados nos trabalhos de classificação de carcinomas são pulmão, próstata, mama e colo uterino.

Por fim, com base em todos os trabalhos citados nas seções 2.1 e 2.2, além dos levantamentos de Gurcan *et al.* [2009], Lei He *et al.* [2012], Belsare e Mushirif [2012], e Meijering [2012], que apresentam revisões completas sobre os trabalhos com imagens histológicas com diferentes órgãos, métodos e objetivos, bem como de Irshad *et al.* [2014], que são mais específicos e realizam o levantamento de trabalhos que utilizam especificamente imagens coradas com *Hematoxylin* e *Eosin* (H&E), constatamos a grande carência e até inexistência de trabalhos que se assemelhem ao que será apresentado neste texto no que diz respeito ao objetivo de classificar imagens histológicas renais de glomérulos quanto à presença de glomerulopatias.

A Tabela 2.1 resume detalhes de alguns dos principais trabalhos publicados entre os anos de 2012 e 2015, e tem como objetivo a classificação de imagens histológicas.

Tabela 2.1: Resumo de trabalhos similares ao PathoSpotter

Autores	**Objetivo**	**Tamanho do conjunto de dados**	**Métodos**	**Acurácia**
Miranda *et al.* [2012]	Classificação de Câncer no colo uterino	144 imagens	Limiarização por divisor de águas, morfológia, grafos e agrupamento.	73%
Sirinukunwattana *et al.* [2014]	Detecção e classificação de células em imagens histológicas da mama	50 imagens	Redes neurais, limiarização simples, crescimento de regiões, detector de borda e filtros gaussianos.	86%
Schöchlin *et al.* [2014]	Identificação de câncer de pele (melanoma)	18 imagens	Limiarização automática, deconvolução de cor e classificador criado para esta tarefa.	88,9%
Kothari *et al.* [2013]	Classificação de câncer Renal	151 imagens	SVM, segmentação por máscara de cor, descritores de forma e textura.	77%
Mathur *et al.* [2013]	Classificação de glóbulos brancos	287 amostras de 237 imagens	*Naive Bayes*, descritores de forma e textura, morfologia, conversão de espaço de cor.	92%

Além dos trabalhos citados na Tabela 2.1, as revisões de Gurcan *et al.* [2009], Lei He *et al.* [2012], Belsare e Mushirif [2012], Meijering [2012] e Irshad *et al.* [2014] tiveram papel fundamental no processo de investigação e experimentos de métodos para a construção do PathoSpotter.

Capítulo 3

Classificação de Imagens

"O que sabemos é uma gota, o que não sabemos é um oceano."
-- Isaac Newton

O sistema de visão humano é extremamente complexo e nos permite distinguir com grande precisão diferentes objetos, estruturas e pessoas em nossa volta. A visão computacional, por sua vez, é um campo de pesquisa que se propõe a simular o sistema de visão humano, e tem como principal objetivo extrair dados através de fotos ou vídeos e transformá-los em informação útil, melhorando processos, reconhecendo pessoas ou objetos e classificando elementos em problemas de diferentes domínios [Nixon e Aguado, 2008, p.1].

Como exemplos das diferentes aplicações da visão computacional, podem-se destacar os trabalhos de Danuser [2011] com a extração de informações para estudo da vida celular; Meyer e Camargo Neto [2008], com uma ferramenta de auxílio no controle de ervas daninhas a partir de partes da imagem da planta; Bougouma *et al*. [2012] que simulam o comportamento e composição de cristais simples de $MoSe_2$ através de microscopia eletrônica, microscopia óptica e análise de imagens; ou Młynarczuk *et al*. [2013] que aplicam técnicas de analise de imagens na área de geologia, automatizando o processo de classificação de rochas. As aplicações se dão em uma ampla área da ciência e tecnologia e está fora do escopo deste texto fazer um apanhado geral sobre as aplicações da visão computacional.

Os sistemas de visão computacional, em casos mais específicos, são utilizados como sistemas de classificação de imagens, como é o caso do PathoSpotter. A seguir, descreveremos como funcionam estes sistemas destacando quais as técnicas que serviram de base para a construção do PathoSpotter.

3.1 Sistemas de Classificação de Imagens

A arquitetura típica dos sistemas de classificação de imagens possui etapas básicas, as quais podem ser visualizadas mais especificamente por Gonzalez e Woods [2006, p.48] e de

maneira mais resumida por Pedrini e Schwartz [2008, p.4] em:

1. Aquisição;
2. Pré-processamento;
3. Segmentação;
4. Extração de características;
5. Classificação.

Além das etapas acima, existe ainda a etapa de avaliação da qualidade do classificador, a qual não é parte do classificador em si, mas serve para validar sua operação dentro do domínio em que está sendo usado.

A seguir, detalhamos as etapas que compõem um sistema de classificação de imagens, além de apresentar os respectivos métodos utilizados em cada etapa do sistema PathoSpotter. Adicionalmente, apresentamos as métricas e métodos de avaliação utilizados no processo de avaliação do classificador usado no sistema PathoSpotter.

3.1.1 Aquisição

Uma imagem é compreendida pelo computador como um conjunto de dados discretos representados através de informações de espaço e intensidade [Solomon e Breckon, 2011]. Outra forma de compreender a representação computacional de uma imagem é como uma função de intensidade luminosa, denotada por $f(x,y)$, cujo valor nas coordenadas espaciais (x,y) fornece a intensidade (ou o brilho) da imagem naquele ponto. Uma imagem digital é representada por meio de uma matriz bidimensional, na qual cada elemento da matriz é um pixel. A intensidade de brilho de cada pixel varia em uma um intervalo de valores inteiros, por exemplo, de 0 a 255 que é faixa usada para representar 256 níveis de cinza com 1 byte [Pedrini e Schwartz, 2008, p.3].

Em um sistema de classificação de imagens, o processo de aquisição de imagens trata da captura de uma imagem por meio de um dispositivo ou sensor e a conversão dessa imagem para uma representação digital [Nixon e Aguado, 2008, p.1]. A representação de uma imagem em nível de cinza, por exemplo, é realizada através de uma matriz única. Já para representar uma imagem colorida, na qual o espectro completo de cor possa ser representado, é necessário tratar a imagem como um conjunto de dados formado pela combinação de

diferentes matrizes de intensidade de cor [Solomon e Breckon, 2011, p.2].

3.1.2 Pré-processamento

A etapa de pré-processamento, também chamada de aperfeiçoamento (*enhancement*) tem o objetivo de reduzir ou eliminar imperfeições e ruídos presentes nas imagens, geralmente oriundos da etapa de aquisição [Pedrini e Schwartz, 2008, p.4]. Os principais métodos utilizados com o intuito de aperfeiçoar as imagens envolvem testes com espaços de cores, técnicas de manipulação de histogramas, aplicação de filtros e operações morfológicas, operações estas que podem ser utilizadas em diferentes etapas de um sistema de classificação de imagens [Petrou e Petrou, 2010, p.293].

Os espaços de cores possibilitam diferentes representações das imagens. Gonzalez e Woods [2006, p. 423] definem que espaço de cor é uma especificação de um dado sistema de coordenadas, em que cada cor é representada por um simples ponto. Existem diferentes espaços de cor, e talvez um dos mais populares seja o RGB, no qual a cor de cada pixel é representada pela combinação linear das cores de base (*Red*, *Green* e *Blue*). Outro exemplo de espaço de cor é o HSV, que representa uma imagem a partir das informações de tonalidade, saturação e intensidade (*Hue, Saturation* e *Value*) [Gonzalez e Woods, 2006, p. 423].

A realização de conversão entre espaços de cores é uma tarefa de pré-processamento comum em trabalhos da área de histopatologia digital, sendo realizada com o objetivo de auxiliar as etapas de segmentação ou extração de características [Gurcan *et al.* 2009]. Mathur *et al*. [2013] realizaram a conversão de imagens histológicas de RGB para HSV com o foco na etapa de segmentação e Lei He *et al.* [2012] destacam a conversão entre espaços de cor como a operação que antecede a extração de características de cor (em forma de intensidade de brilho em uma determinada região).

Dentre os mecanismos utilizados para observar dados em imagens, o histograma se destaca como uma ferramenta útil. O histograma de uma imagem é uma função discreta, formada através da contagem do número de pixels, da imagem, que possuem um determinado valor de intensidade de cinza. O gráfico gerado a partir dessa função, que descreve a quantidade de pixels por nível de cinza, é o histograma [Petrou e Petrou, 2010, p. 367; Pedrini e Schwartz, 2008, p.104]. A Figura 3.1 ilustra um histograma de uma imagem considerada ruim (a) e uma imagem boa (b) [Petrou e Petrou, 2010, p.367]:

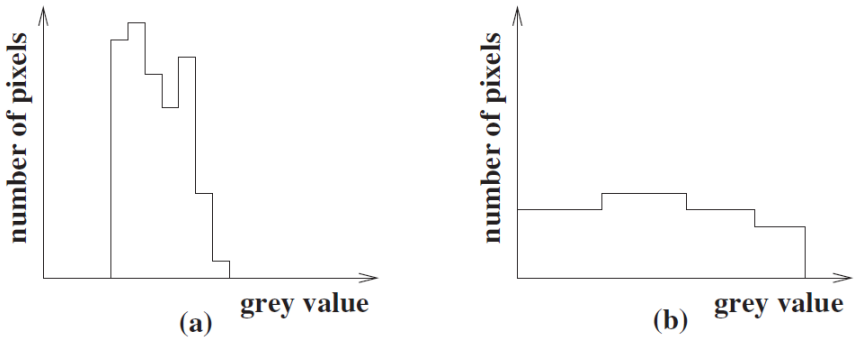

Figura 3.1: Histograma de uma imagem com distribuição de intensidade de pixels ruím (a) e histograma de uma imagem com boa distribuição de intensidade de pixels. Fonte: Petrou e Petrou [2010, p. 367].

As operações de manipulação de histogramas servem para melhorar ou enfatizar uma determinada região de uma imagem. Dois exemplos comuns de operações de manipulação de histogramas são a equalização e o realce logarítmico. A equalização também é conhecida como expansão de histograma e é uma operação que modifica a imagem original em uma nova imagem com distribuição de nível de cinza mais uniforme. A equalização pode aperfeiçoar uma imagem evidenciando regiões com baixa intensidade de pixels [Pedrini e Schwartz, 2008, p. 109]. A figura 3.2 ilustra o processo de equalização do histograma de uma imagem, antes (a) e após (c) a realização da equalização.

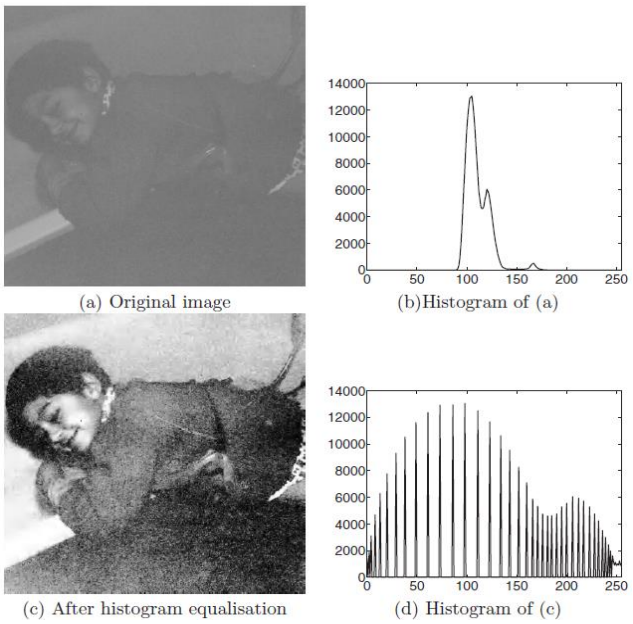

Figura 3.2: Equalização de histogramas. Imagem ruím (a,b) image melhorada (c,d). Fonte: Petrou e Petrou [2010, p. 371].

O realce logarítmico substitui o valor de cada pixel da imagem pelo seu logaritmo, o que resulta em um realce maior nos pixels de menor intensidade de brilho (regiões mais escuras) [Pedrini e Schwartz, 2008, p. 110]. A Figura 3.3 mostra o resultado da aplicação do realce logarítmico em uma imagem.

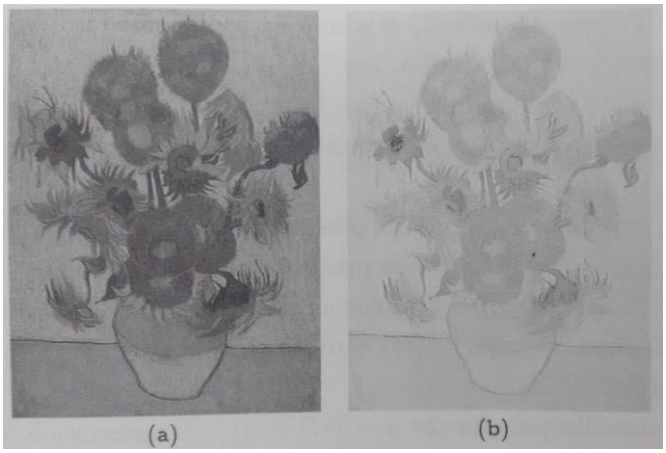

Figura 3.3: Imagem com o realce logarítmico (a) e imagem original (b) Fonte: Pedrini e Schwartz [2008, p.110].

Além das operações de manipulação de histogramas, que de modo geral, operam pixel por pixel, destacamos também as operações de filtragem, mais especificamente, os filtros de suavização, que são utilizados para eliminar ruídos.

O processo de filtragem, na prática, pode ser realizado através da operação de convolução, que se trata de uma operação onde uma matriz H, conhecida comumente como máscara de convolução, é operada com outra matriz M, que se trata da imagem sobre a qual se deseja realizar a filtragem [Davies, 2012, p.32; Burger e Burge, 2009, p.107].

Os dois tipos mais comuns de filtros de suavização são os filtros de média e mediana. No filtro de média a mascara de convolução é aplicada a imagem original percorrendo toda a imagem. A cada iteração, calcula-se a média dos valores dos pixels referentes à região da máscara de convolução e atribui-se o valor de média encontrada ao pixel referente à posição central da imagem operada. O filtro de mediana, por sua vez, reduz o efeito de *bluring* (embaçamento de informações de curvas e linhas) e preserva as informações de bordas porventura presentes na imagem, o que não ocorre no filtro de média. No filtro de mediana os valores da máscara de convolução são ordenados e o valor médio encontrado é atribuído ao pixel referente à posição central da imagem operada [Petrou e Petrou, 2010, p.326; Davies, 2012, p.43].

3.1.3 Segmentação

A etapa de segmentação é a etapa que tem como objetivo a extração de uma região específica da imagem, como a separação de um ou vários objetos em relação ao seu fundo (*background)*. De modo geral, esta etapa resulta em uma imagem binarizada, que possui apenas dois possíveis valores de pixels, branco na região equivalente do objeto e preto na região do fundo. A Figura 3.4 ilustra o histograma ideal para a realização da segmentação de uma imagem, na qual se percebe que há duas classes claramente definidas [Davies, 2012, p.83].

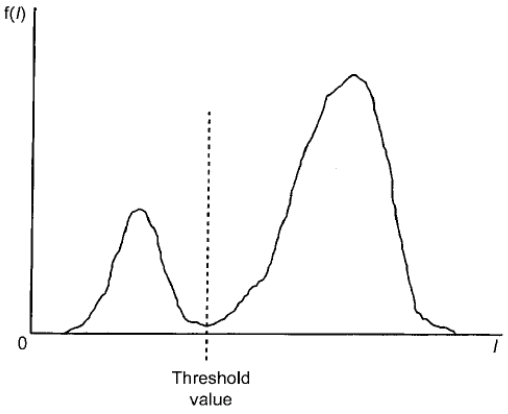

Figura 3.4: Escolha do limiar de segmentação. Fonte: Davies [2012, p.86.].

Devido à diversidade de imagens, nem sempre o limiar de segmentação (*threshold*) utilizado para separar as classes de pixels de uma imagem é facilmente encontrado. Destacamos a seguir alguns dos métodos utilizados no PathoSpotter para realizar a segmentação das áreas de interesse através da definição de um limiar adequado:

- Limiarização simples
- Limiarização automática de Otsu (*Otsu thresholding*)
- Segmentação por divisor de águas (*watershed*)
- Crescimento de regiões (*region growing*)

Na limiarização simples escolhe-se um limiar de segmentação manualmente, com a observação do histograma e testes com diferentes valores de limiar.

No caso da limiarização automática, um dos métodos mais comuns e eficientes são os métodos de Otsu, global e local. Na limiarização automática de Otsu global admite-se que uma imagem possua duas classes de pixels, então se encontra um limiar que minimize a

variância dentro de cada uma dessas classes. Por fim, utiliza-se como limiar, a soma ponderada dos limiares de cada classe. O limiar utilizado para segmentar a imagem pode ser descrito por: $t = (\mu 1 + \mu 2)/2$, sendo t igual ao limiar. A Limiarização de Otsu local é realizada de maneira similar, com a diferença que se aplica o método de Otsu em regiões da imagem (localmente), calculando assim um limiar de segmentação para cada região específica da imagem [Pedrini e Schwartz, 2008, p.187; Nixon e Aguado, 2008, p.78; Davies, 2012, p.95].

O método de segmentação por divisor por águas (*watershed)* propõe uma abordagem morfológica para o problema de segmentação de imagens, interpretando estas como superfícies em que cada pixel corresponde a uma posição, e os níveis de cinza determinam as altitudes. A partir desta noção, deseja-se então identificar bacias hidrográficas, definidas por mínimos regionais e suas regiões de domínio. Um dos problemas encontrados ao utilizar a segmentação por divisor de águas é sua suscetibilidade a ruídos. O ruído pode fazer com que regiões indesejadas sejam segmentadas, algo que também é conhecido como segmentação excessiva (*over segmentation*) [Preim e Botha, 2014, p.129]. A Figura 3.5 ilustra o processo de segmentação por divisor de águas, mostrando os pontos mínimos encontrados na superfície.

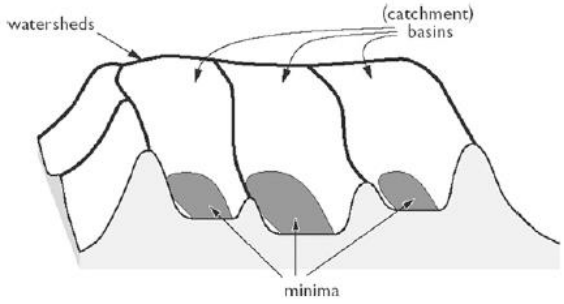

Figura 3.5: Método de segmentação por divisor de águas. Fonte: Hahn [2005] apud Preim e Botha [2014, p.129].

O método de segmentação baseada em crescimento de regiões é um método pelo qual pontos apresentando propriedades similares (valor de intensidade de pixel, por exemplo) são agrupados para formar uma região. O caso mais simples para se aplicar esse método de segmentação é em imagens binárias (com pixels de valor *true* ou *false*). Neste procedimento, pixels que possuem um vizinho com valor de intensidade de pixel diferente, são considerados como pixels que compõe a borda da região, caso contrário, onde sua vizinhança possua o mesmo valor de intensidade, esse conjunto de pixels é selecionado como parte integrante da

região [Pedrini e Schwartz, 2008, p.196; Preim e Botha, 2014, p.141].

Por fim, um conjunto de operações que podem ser utilizadas tanto na etapa de segmentação quanto no pré-processamento e extração de características são as operações morfológicas [Miranda *et al.* 2012]. Em geral, as operações morfológicas são realizadas com uma imagem proveniente da operação de limiarização (imagens binarizadas) [Sonka *et al.* 2006, p.682]. As operações morfológicas fundamentais são a dilatação e a erosão. A dilatação é a expansão de um conjunto de pixels ligados pela vizinhança e a erosão por sua vez é o contrário, a redução dos pixels [Dougherty, 2009, p.275]. As Figuras 3.6 e 3.7 exemplificam as operações morfológicas de dilatação e erosão, respectivamente.

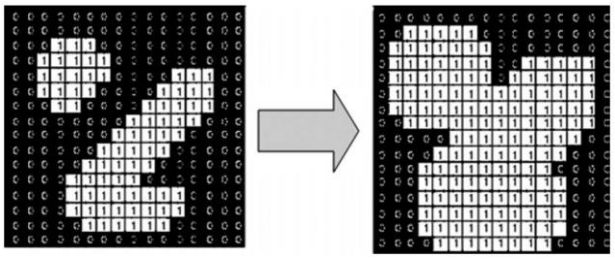

Figura 3.6: Exemplo de dilatação. Fonte: Dougherty [2009, p. 276.]

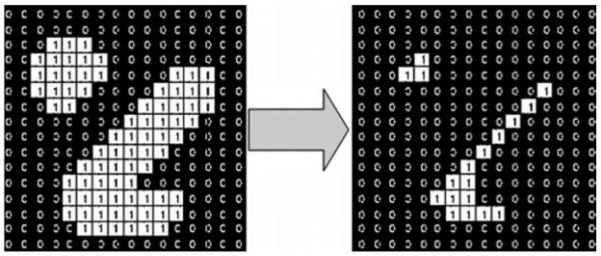

Figura 3.7: Exemplo de erosão. Fonte: Dougherty [2009, p. 278.]

As operações morfológicas de dilatação e erosão são as operações morfológicas fundamentais, a partir das quais as demais operações morfológicas são implementadas. A operação de fechamento, por exemplo, é definida pela dilatação seguida da erosão [Dougherty, 2009, p.281]. Em uma operação de reconstrução morfológica, se assumimos que uma forma deve ser reconstruída deve-se utilizar uma imagem referência, a partir da qual se realiza a reconstrução. Na reconstrução morfológica por fechamento, a imagem utilizada como referência para reconstruir a imagem original é a imagem resultante da operação morfológica de fechamento [Sonka *et al.* 2006, p. 682; Miranda *et al.* 2012].

3.1.4 Extração de Características

A etapa de extração de características, também chamada de etapa de descrição, tem como objetivo a extração de propriedades ou dados quantitativos e mensuráveis de uma imagem, pelos quais um sistema computacional se torna capaz de interpretar, identificar ou classificar cada imagem que lhe for apresentada [Pedrini e Schwartz, 2008, p.4].

Na etapa de extração de característica, a utilização de um determinado método descritor ou extrator está diretamente associada ao domínio do problema e ao objetivo do sistema de classificação de imagens. No entanto, apresentamos aqui algumas informações relevantes quando o objetivo da etapa de extração de características é extrair dados que possam identificar um objeto.

Através de descritores de bordas podem-se extrair informações (características) como: diâmetro, perímetro, curvatura, etc. Com descritores de região pode-se extrair: área, circularidade, propriedades topológicas entre outras [Pedrini e Schwartz, 2008]. Em imagens histológicas, por exemplo, as características morfológicas, de cor e textura aparecem como as mais utilizadas [Belsare e Mushirif, 2012].

Um exemplo de método de extração de características que foi usado neste trabalho é o método implementado por Van der Walt *et al.* [2014] chamado de detecção de bolhas (*blob detection*), que se baseia no método *Laplacian of Gaussian* [Lindeberg, 1993] para a detecção de bolhas em uma imagem.

3.1.5 Classificação

A classificação é a etapa que agrupa ou rotula adequadamente uma amostra dentro de um conjunto de possíveis classes pré-definidas. Os dados de entrada da etapa de classificação são dados quantitativos, oriundos da etapa de extração de características, comumente representados na forma de vetores de características. Problemas de classificação, geralmente, são aqueles cujas possíveis classes existentes possuem caráter qualitativo, diferente do que ocorre em problemas de regressão linear, cujos possíveis resultados de classificação são quantitativos [Pedrini e Schwartz, 2008, p.397; James *et al.* 2013, p.28].

O propósito de um sistema de classificação de imagens típico para diagnóstico de câncer pode ser classificar cada amostra (imagem) como saudável (0) ou neoplásica (1). Um exemplo de regressão linear pode ser um sistema de predição de preços de casas, levando em consideração

um conjunto de dados que contem informações sobre localização, tamanho do terreno e preço de casas vizinhas. [Conway e White, 2012, p.62; James *et al.* 2013, p.129].

O modelo mais simples de classificação é por meio de uma única característica com decisão através de uma estrutura condicional simples: isto ou aquilo. Porém, boa parte dos problemas do mundo real requer a extração de mais de uma característica. Logo, se faz necessária a utilização de modelos de classificação mais sofisticados. O campo de estudo do Aprendizado de Máquina é uma área que possui uma grande interseção com a visão computacional, propondo diversos modelos de classificadores tanto para a resolução de problemas de classificação binária (duas classes) quanto classificação multiclasse (com mais de duas classes) [James *et al.* 2013; Dean, 2014, p.64].

Como exemplo de classificador, citamos um dos algoritmos de classificação mais simples, o algoritmo de *k* vizinhos mais próximos ou *k-Nearest Neighbors* (kNN). O kNN é um algoritmo de classificação baseado em medidas de distância. Para tomar uma decisão, o kNN calcula os *k* vizinhos mais próximos de uma amostra desconhecida, dessa forma utiliza um conjunto de dados previamente rotulados (classificados) para classificar uma amostra desconhecida [James *et al.* 2013, p.151; Kirk, 2015, p.17]. A Figura 3.8 ilustra o processo de classificação do kNN em um espaço de características simples contendo 6 amostras, três rotuladas como *Good* e 3 rotulados como *Bad*, e uma amostra desconhecida, a qual se deseja classificar. Neste caso, se o valor de *k* fosse igual a 3 (quantidade de vizinhos a ser analisados), a amostra desconhecida seria classificada como *Bad*, pois das três amostras mais próximas da amostra desconhecida, duas são *Bad* e uma é *Good*. Dependendo da escolha do valor de *k*, o resultado da classificação da amostra desconhecida poderia ser alterado. Ainda no exemplo da Figura 3.8, caso o valor de *k* fosse 1, por exemplo, a amostra desconhecida seria classificada como *Good*, pois o vizinho mais próximo da amostra desconhecida é uma amostra rotulada como *Good*.

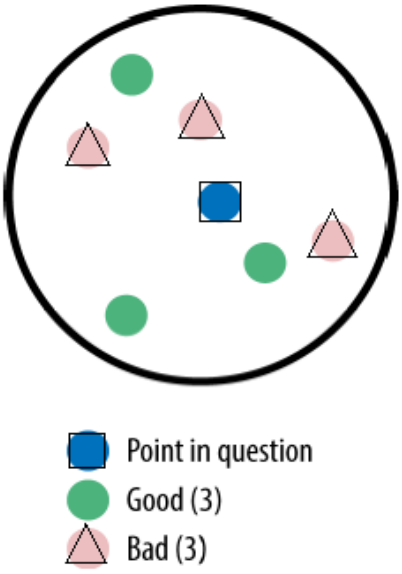

Figura 3.8: Algoritmo de classificação kNN. O círculo maior, em negrito, diz respeito a um dado espaço de características. Os círculos em verde correspondem as amostras *Good*, os triângulos são as amostras *Bad* e na forma de quadrado (em azul) encontra-se uma amostra desconhecida, a qual deseja-se classificar. Fonte: Kirk [2015, p.26].

Podemos afirmar que este classificador é um algoritmo baseado em aprendizado supervisionado, pois utiliza uma base de dados já rotulados para classificar dados de classes desconhecidas. É um método que funciona bem para dados com sensibilidade de distância, por outro lado, sofre com problemas de dados com alta dimensionalidade [Kirk, 2015, p.22]. Em relação a aplicação do kNN em trabalhos específicos da área de histopatologia digital, ele é apontado por Belsare e Mushirif [2012] como uma estratégia comum entre os trabalhos com imagens histológicas.

Sistemas de aprendizado automático também podem utilizar como estratégia de classificação o método de regressão logística. O objetivo geral de uma regressão logística é encontrar uma função que discrimine duas ou mais classes de amostras. Na regressão logística, a etapa de treinamento dos dados é o processo de ajuste da curva (função) que melhor discrimine os dados. As vantagens da utilização de uma regressão logística são sua fácil implementação, o fato de utilizar pouco recurso computacional e ser de simples compreensão. Como pontos negativos estão a baixa acurácia, quando comparada a algoritmos de classificação (como o kNN) [Harrington, 2012, p.83].

No caso de problemas de classificação binária (duas classes), o resultado da função de classificação deve ser 0 ou 1. Na regressão logística, para se obter resultados é utilizada a

função *sigmoid*, dada pela Equação 3.1. Independente dos valores de entrada da função *sigmoid*, os valores de saída estarão entre 0 e 1. Na regressão logística classifica-se os valores acima de 0.5 como 1, e valores abaixo de 0.5 como 0 [Hackeling, 2014, p.72; Harrington, 2012, p.84].

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.1}$$

A variável de entrada da função *sigmoid* é z, a qual é obtida pela Equação 3.2. O vetor *x* são as características utilizadas na classificação e o vetor *w*, são os melhores coeficientes que tornará a classificação o mais bem sucedida quanto possível [Harrington, 2012, p.86].

$$z = w_0x_0 + w_1x_1 + w_2x_2 ... + w_nx_n \tag{3.2}$$

Para calcular os melhores coeficientes que constituem o vetor *w*, para isso, se faz necessária a utilização de algum método de otimização. O método mais comum de otimização é o *Gradient Ascent*, dado pela Equação 3.3. O método *Gradient Ascent* parte da ideia de que quando se deseja encontrar o ponto máximo em uma função, a melhor maneira de se mover é na direção de gradiente [Harrington, 2012, p.86].

$$\delta f(x,y) = \begin{pmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{pmatrix} \tag{3.3}$$

A função *f(x,y)* se move em direção a *x* por $\frac{\partial f(x,y)}{\partial x}$ e em direção a *y* por $\frac{\partial f(x,y)}{\partial y}$. Isto se explica pelo fato de que *f(x,y)* precisa ser diferenciável em torno dos pontos em que está sendo avaliada. O algoritmo de otimização é aplicado aos dados até alcançar alguma condição de parada ou atender a uma determinada margem de tolerância [Harrington, 2012, p.86].

A Figura 3.9 ilustra um exemplo em que, através da regressão logística se encontrou uma função discriminante para um determinado conjunto de dados (*dataset*).

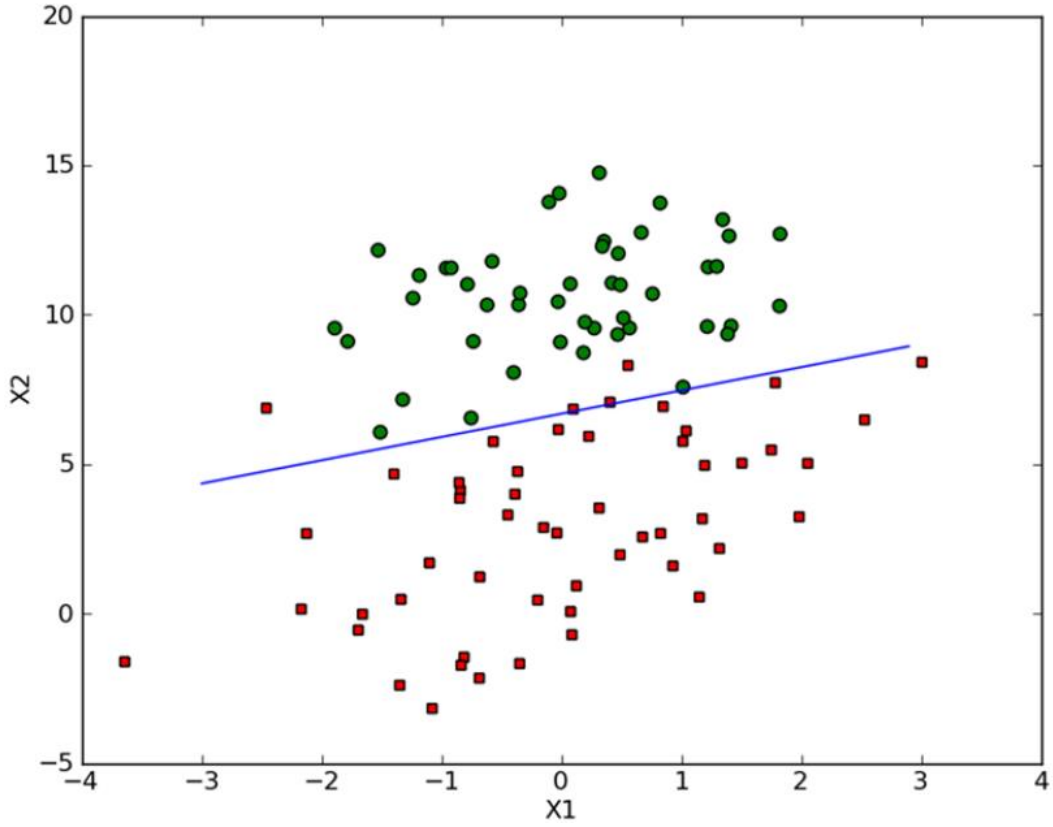

Figura 3.9: Exemplo de regressão logística. A reta em azul é a função ajustada através da regressão para discriminar as duas classes de amostras, que são ilustradas pelos círculos verdes (Classe 0) e os quadrados vermelhos (Classe 1). Fonte: Harrington [2012, p.91].

3.1.6 Avaliação

Há diferentes métodos utilizados para avaliar um sistema de classificação ou ajustar um determinado parâmetro. Estes métodos de avaliação se adequam ao domínio e especificações do problema, e são utilizados na escolha de um método em detrimento a outros ou para tornar um modelo de classificação mais confiável [Japkowicz e Shah, 2011, p.18]. Os métodos de avaliação utilizados no PathoSpotter foram a validação cruzada e matriz de confusão. A partir desses métodos foram obtidas as métricas de sensibilidade, especificidade, precisão, acurácia, *recall*, taxa de erro e desvio padrão.

Uma matriz de confusão é uma ferramenta útil para analisar o quão bem um classificador pode reconhecer amostras de diferentes classes. Dada *m* classes, a matriz de confusão é uma matriz de tamanho *m x m*. A Figura 3.10 mostra uma matriz de confusão típica, um caso de classificação binária. Nesse caso, a primeira coluna (C_1) equivaleria às amostras positivas e a segunda coluna (C_2), às amostras negativas. As linhas da matriz armazenam os resultados da

classificação, falso ou verdadeiro. No melhor dos casos, a matriz possuiria todas as amostras classificadas, distribuídas na diagonal principal (C_1-C_1 e C_2-C_2) [Ham e Kamber, 2006, p.361; Hackeling, 2014, p.77].

		Predicted class	
		C_1	C_2
Actual class	C_1	true positives	false negatives
	C_2	false positives	true negatives

Figura 3.10: Matriz de confusão. Fonte: Ham e Kamber [2006, p. 361].

A tupla *true positives* (t_{pos}) diz respeito a quantas amostras positivas foram classificadas corretamente por um classificador, enquanto *true negatives* (t_{neg}) contem quantas amostras negativas foram classificadas corretamente por um classificador. A tupla *false positive* (f_{pos}) contem quantas amostras negativas foram classificadas incorretamente como se fossem positivas, enquanto a posição *false negatives* (f_{neg}) contem quantas amostras positivas foram classificadas incorretamente como se fossem negativas [Ham e Kamber, 2006, p.361].

Algumas importantes métricas de avaliação são calculadas através da matriz de confusão: sensibilidade, especificidade, precisão, *recall* e acurácia. As fórmulas dessas métricas são apresentadas por [Ham e Kamber, 2006, p.361; Harrington, 2012, p.144]:

$$recall = \frac{t_{pos}}{f_{neg} + t_{pos}} \tag{3.4}$$

$$\text{precisão} = \frac{t_{pos}}{t_{pos} + f_{pos}} \tag{3.5}$$

$$\text{acurácia} = \frac{t_{pos} + t_{neg}}{t_{pos} + t_{neg} + f_{pos} + f_{neg}} \tag{3.6}$$

O *recall* (apresentado na Equação 3.4) mensura a fração de amostras positivas classificadas corretamente por um classificador. A precisão (apresentada na Equação 3.5) mensura a fração de amostras negativas classificadas corretamente por um classificador. Por fim, a acurácia (apresentada na Equação 3.6) é a fração de número de amostras classificadas corretamente pelo número de amostras classificadas de maneira equivocada.

As métricas de sensibilidade (Equação 3.7) e especificidade (Equação 3.8) são comumente utilizadas em diagnósticos médicos e também são utilizadas para mensurar detalhes específicos de uma classificação. A sensibilidade é igual a precisão e diz respeito a fração de pacientes com uma dada patologia foram diagnosticados corretamente, por outro lado a especificidade diz respeito a fração de pacientes sem essa determinada patologia foram diagnosticados corretamente [Ham e Kamber, 2006, p.361; Harrington, 2012, p.144].

$$\text{sensibilidade} = \frac{t\_pos}{t_{pos} + f_{pos}} \tag{3.7}$$

$$\text{especificidade} = \frac{t\_neg}{f_{neg} + t_{neg}} \tag{3.8}$$

Outra importante métrica de avaliação de classificação é a taxa de erro (Equação 3.9), descrita por [James *et al.* 2013; p.37] como a proporção de erros de classificação em um conjunto de teste qualquer:

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n} I(y_i \neq y'_i). \tag{3.9}$$

Na Equação 3.9, *n* representa o número de amostras classificadas, *y* indica o resultado obtido em uma dada classificação, *y'* indica o verdadeiro resultado das respectivas amostras classificadas (*y*). Caso $I(y_i \neq y'_i)$ seja verdade o resultado é 1, caso seja falso o resultado é 0. Harrington [2012, p.24] aponta a taxa de erro como métrica muito comum no processo de avaliação de um modelo de classificação.

Outra importante ferramenta de avaliação é a validação cruzada (*cross validation*), utilizada para avaliar um modelo de classificação de maneira mais robusta e confiável, além de também ser utilizada para comparar diferentes modelos de classificação.

De um modo geral, para avaliar um modelo de classificação, é necessário dividir o conjunto de dados em um conjunto de treinamento, que é o conjunto que serve como base de informação para a tomada de decisão de um método, e o conjunto de teste, que é o conjunto responsável por verificar as taxas de acerto do modelo de classificação [Ham e Kamber, 2006, p.364].

A grande contribuição do método de validação cruzada é a possibilidade de avaliar a

capacidade de generalização de um modelo de classificação, pois o mesmo é testado com diferentes combinações de conjuntos de treinamento/teste. Ao avaliar um modelo de classificação com conjuntos treinamento/teste fixos, os resultados finais são limitados a estes conjuntos de amostras, não podendo validar a capacidade geral de um modelo classificar um conjunto qualquer de amostras [James *et al.* 2013, p.181].

O método de validação cruzada *k-fold* estratificado, divide um conjunto de dados em *k* conjuntos idênticos de treinamento e teste, treinando e avaliando o modelo de classificação *k* vezes. Por fim, a acurácia final da avaliação é a média das acurácias dos *k* conjuntos treinamento/teste [James *et al.* 2013, p.181]. A Figura 3.11 ilustra a divisão de um conjunto de dados por validação cruzada *k-fold* igual a 5. Neste caso, as amostras são divididas em conjunto de teste (equivalente a 1/5 das amostras) e treinamento (4/5 das amostras). A divisão é realizada 5 vezes, de modo que todas as amostras sejam utilizadas como conjunto de treinamento e teste.

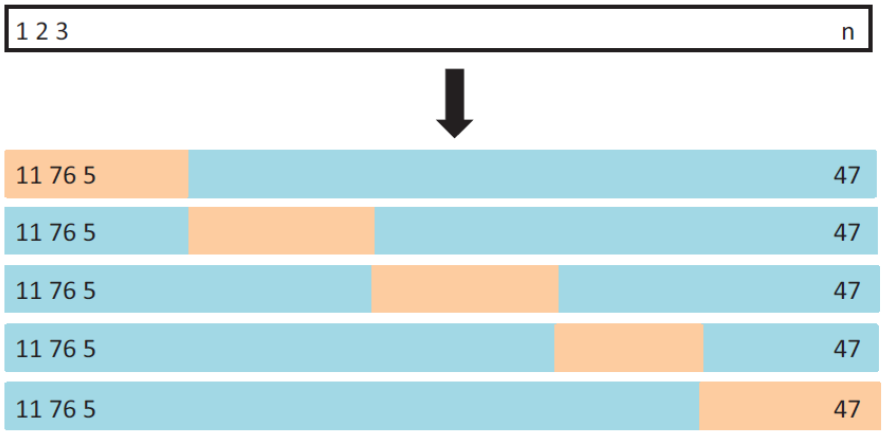

Figura 3.11: Exemplo de validação cruzada *5-fold*. Fonte: James *et al.* [2013, p.181].

O valor de *k* apontado como uma boa escolha para *k-fold* é o valor 10, que consegue validar e testar a exatidão de um modelo de classificação além de evitar altos custos computacionais [Japkowicz e Shah, 2011, p.7; James *et al.* 2013, p.181].

Adicionalmente ao método de validação cruzada *k-fold* estratificado, ainda destaca outra importante prática de avaliação, realizada previamente à validação cruzada, que é a divisão do conjunto de dados em um conjunto de validação e generalização. Inicialmente divide-se o conjunto de dados em *k* conjuntos e separa-se um dos conjuntos treinamento/teste como conjunto de validação. As demais amostras restantes do processo são utilizadas em uma nova aplicação da validação cruzada, como subconjuntos que formam o conjunto de generalização.

As métricas (acurácia, sensibilidade, etc.), obtidas com o conjunto de generalização, são comparadas com os resultados obtidos com o conjunto de validação. A ideia é que os resultados obtidos com o conjunto de validação estejam no mesmo intervalo dos resultados de generalização, geralmente através da alguma medida de dispersão, tal como o desvio padrão [James *et al.* 2013, p.181; Ham e Kamber, 2006, p.55].

Capítulo 4

O sistema PathoSpotter

"Está morto: podemos elogiá-lo à vontade."

-- Machado de Assis

Neste capítulo nós apresentamos o PathoSpotter, especificamos a sua arquitetura e os métodos que compõem cada etapa do sistema.

4.1 Visão Geral

PathoSpotter é um acrônimo (*Patho*= Patologia, *Spotter*= Indicador) criado como nome fantasia para um sistema computacional para identificação e classificação de patologias através de imagens histológicas. O PathoSpotter é um sistema gerado a partir da investigação de métodos computacionais com o objetivo de classificar imagens histológicas de glomérulos renais, indicando se as imagens são de glomérulos sem glomerulopatia (controle normal) ou com glomerulopatia proliferativa.

Para construir o sistema, propusemos uma abordagem baseada em métodos utilizados em outros trabalhos com imagens histológicas de diferentes órgãos, os quais, por sua vez, tratavam de outras patologias, adaptando os métodos às particularidades das imagens de histopatologia renal. A Figura 4.1 mostra um esquema que ilustra a arquitetura geral do PathoSpotter.

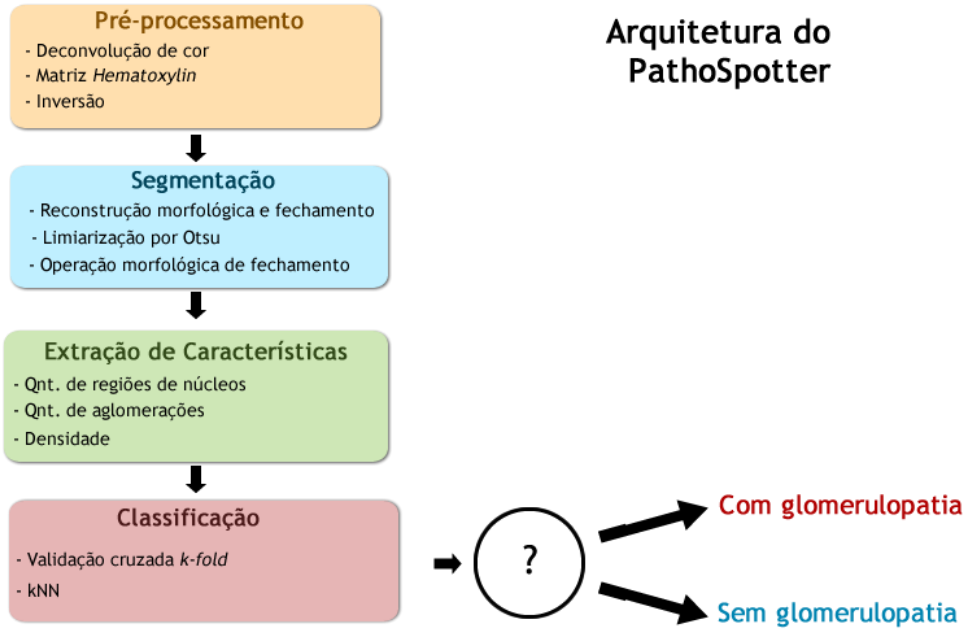

Figura 4.1: Arquitetura do PathoSpotter.

4.2 Etapas do PathoSpotter

4.2.1 Pré-processamento

O objetivo dos métodos de pré-processamento foi destacar as regiões dos núcleos e representar a imagem original (adquirida em RGB) por uma única matriz de intensidade de brilho (níveis de cinza), facilitando as operações nas etapas posteriores.

A primeira operação realizada na etapa de pré-processamento foi a aplicação do método deconvolução de cor, proposto por Ruifrok e Johnston [2001] e implementado na biblioteca *scikit-image* por Van der Walt *et al.* [2014]. Através da aplicação desse método foi possível extrair matrizes em nível de cinza, referentes aos corantes utilizados no processo de obtenção das amostras histológicas.

A matriz obtida no método deconvolução de cor, utilizada no PathoSpotter foi a matriz referente ao corante *Hematoxylin*, que foi escolhido por ser este o corante utilizado para marcar os núcleos presentes nas amostras histológicas. Nomeamos a matriz com informações de *Hematoxylin* (na forma de intensidade de brilho) como matriz H. A Figura 4.2 exemplifica os resultados das operações de pré-processamento em uma imagem.

Figura 4.2: Exemplo de pré-processamento. Imagem original (a); imagem do canal *Hematoxylin*, resultante da operação de deconvolução de cor (b).

4.2.2 Segmentação

A etapa de segmentação se baseou nos trabalhos de Miranda *et al*. [2012], Mathur *et al*. [2013] e Schöchlin *et al*. [2014], que segmentaram núcleos de imagens de células do colo uterino, glóbulos brancos e imagens de células de câncer de pele, respectivamente. Utilizamos as operações morfológicas de fechamento e reconstrução por fechamento, além da limiarização automática de Otsu, ajustados ao problema de detecção das glomerulopatias.

O primeiro método aplicado foi à reconstrução morfológica *top-hat* por fechamento, que reconstrói uma dada imagem H a partir de uma imagem referência F através da operação condicional de interseção ($H \cap F$), sendo H a matriz *Hematoxylin* e F o resultado da operação de fechamento de H. A reconstrução morfológica gerou uma imagem em que os núcleos se encontraram com maior intensidade de brilho do que as demais regiões.

Após a reconstrução morfológica de fechamento, aplicamos o método de limiarização automática de Otsu, com o objetivo de binarizar a imagem. Por fim, aplicamos novamente a operação morfológica de fechamento, dessa vez com o objetivo de corrigir a representação de núcleos e reduzir ruídos oriundos da binarização. A Figura 4.3 mostra o passo a passo das operações realizadas na etapa de segmentação através de uma imagem exemplo.

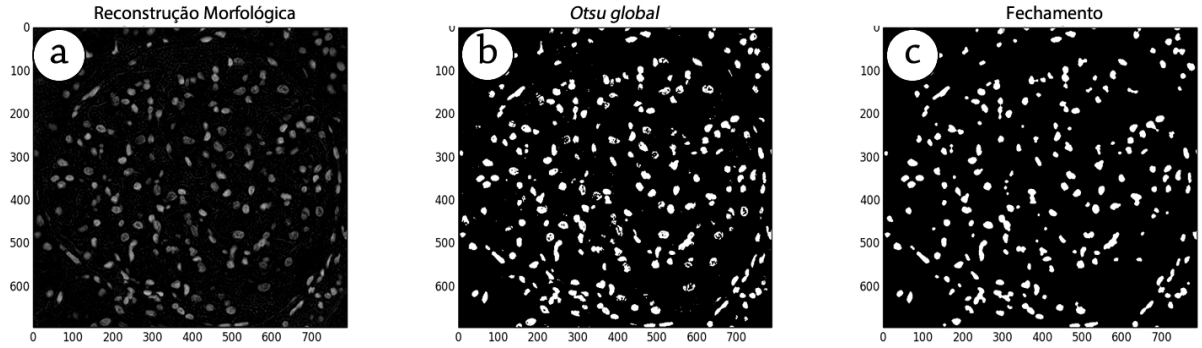

Figura 4.3: Exemplo de segmentação. Resultado da operação de reconstrução morfológica (a); resultado da limiarização automática por Otsu (b); resultado final da segmentação através da realização da operação de fechamento morfológico (c).

4.2.3 Extração de Características

A etapa de extração de características teve como finalidade obter dados quantitativos para viabilizar a diferenciação entre as imagens sem glomerulopatia das imagens com glomerulopatias. As características extraídas foram: a quantidade de regiões de núcleos, a quantidade de aglomerações e a densidade. A Figura 4.4 mostra os métodos utilizados e características extraídas nesta etapa.

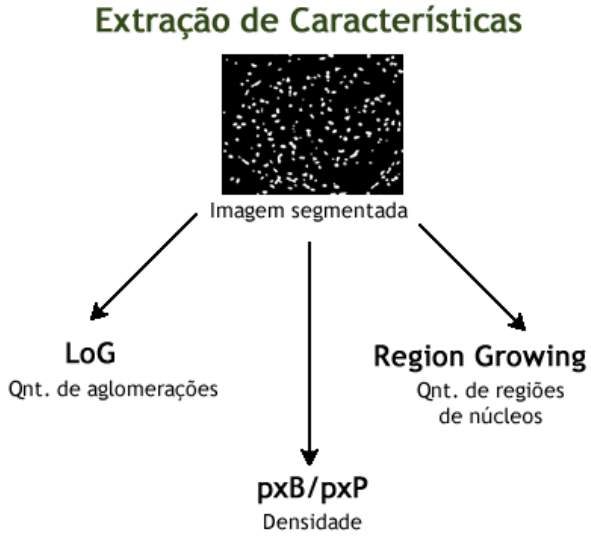

Figura 4.4: Etapa de extração de características. Características extraídas e os respectivos métodos utilizados.

A quantidade de regiões de núcleos foi extraída através da propriedade *regioprops* da biblioteca de processamento de imagens *scikit-image*, que utiliza análise de borda e crescimento de regiões para identificar e contar cada região de núcleo na imagem. O objetivo

dessa característica foi mensurar a proliferação de núcleos (aumento do número de núcleos em uma imagem com glomerulopatia proliferativa) a partir da contagem dos núcleos de uma imagem que se encontram unidos ou isolados. A Figura 4.5 ilustra contagem de regiões de núcleos em uma região de uma imagem.

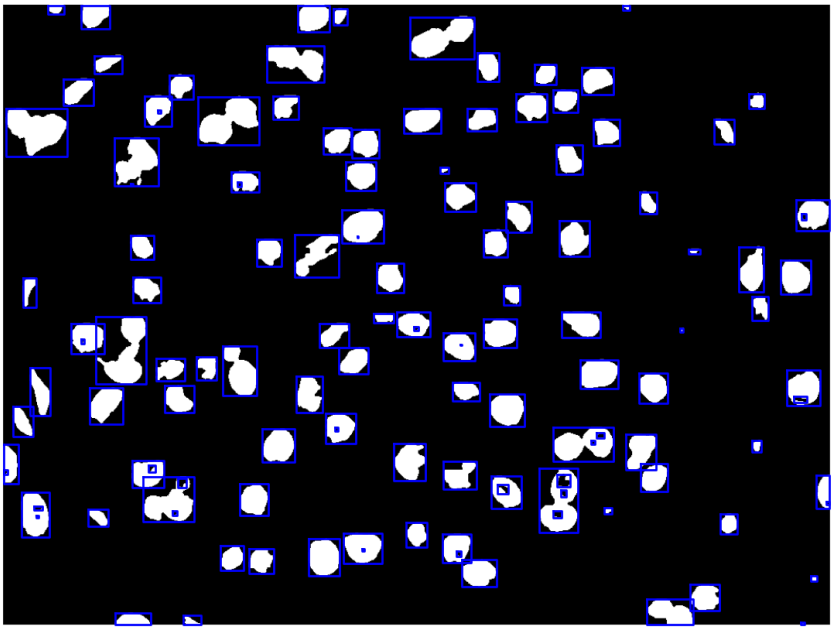

Figura 4.5: Regiões de núcleos representadas dentro dos quadrados em azul.

A quantidade de aglomerações foi extraída pelo método *Laplacian of Gaussian* (LoG), implementado na biblioteca *scikit-image* como um método de detecção de bolhas. O método LoG foi utilizado como um extrator da quantidade de núcleos que se encontravam próximos, porém, não necessariamente unidos, ligados entre pixels da mesma intensidade de brilho, como é o caso da quantidade de regiões de núcleos. O objetivo da contagem da quantidade de aglomerações em uma imagem foi mensurar a proliferação de núcleos de modo a complementar a simples contagem de regiões de núcleos, tendo em vista que a contagem de regiões de núcleos não descreve os núcleos que estão próximos, mas não necessariamente colados, que é a informação obtida através da quantidade de aglomerações. A figura 4.6 mostra um exemplo da identificação de aglomerações em uma imagem.

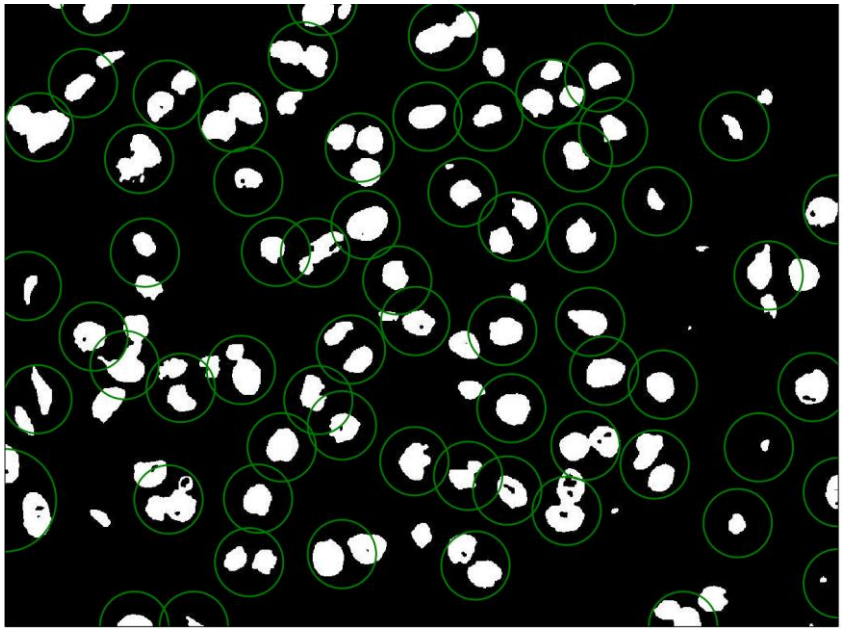

Figura 4.6: Aglomerados representados dentro dos círculos em verde.

A diferença entre as regiões de núcleos e os aglomerados de núcleos pode ser visualizada na Figura 4.7, que mostra a aplicação do método *regionprops* (que identifica regiões de núcleos) e *LoG* (que identifica aglomerados) em uma mesma região de uma imagem.

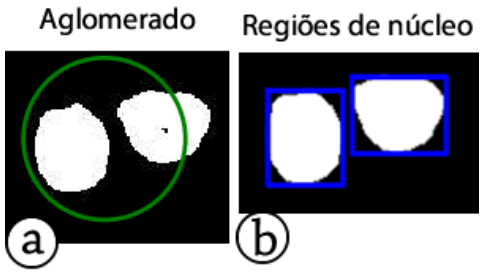

Figura 4.7: Comparação entre um exemplo de aglomerado (a) e regiões de núcleos (b).

Por fim, a última característica extraída foi a densidade, calculada através da razão entre os pixels brancos (núcleos) e pixels pretos (fundo da imagem). O cálculo da densidade foi dado pela Equação 4.1:

$$densidade = \frac{px.brancos}{px.pretos} \tag{4.1}$$

Após a extração das características, organizamos os dados obtidos em uma matriz de características (Figura 4.8), que foi utilizada posteriormente na etapa de classificação de imagens. Cada linha da matriz diz respeito a uma amostra específica (n) e cada coluna possui

informações especificas sobre cada amostra. Na primeira coluna armazenamos o rótulo da imagem (valor 1 para imagens com glomerulopatias e valor 0 para imagens sem glomerulopatias), informação previamente oferecida pelos patologistas. Na segunda coluna, armazenamos as informações de densidade, na terceira a quantidade de regiões e na quarta coluna a quantidade de aglomerações. A Figura 4.8 ilustra a matriz de características obtida na etapa de extração de características.

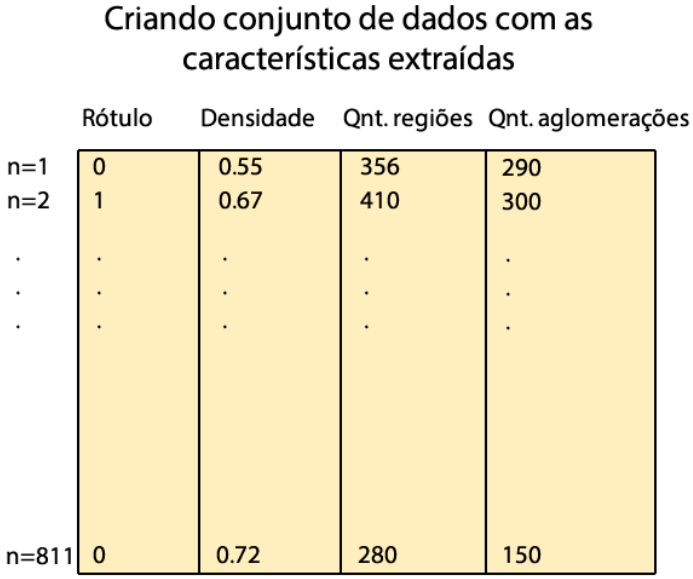

Figura 4.8: Matriz de características.

4.2.4 Classificação

O PathoSpotter usou um classificador binário (imagens sem glomerulopatia e imagens com glomerulopatia), também chamado de dicotomizador. Para a realização da classificação, utilizamos o algoritmo kNN e validamos os resultados através do método de validação cruzada *k*-fold. A Figura 4.9 ilustra a etapa de classificação do sistema PathoSpotter.

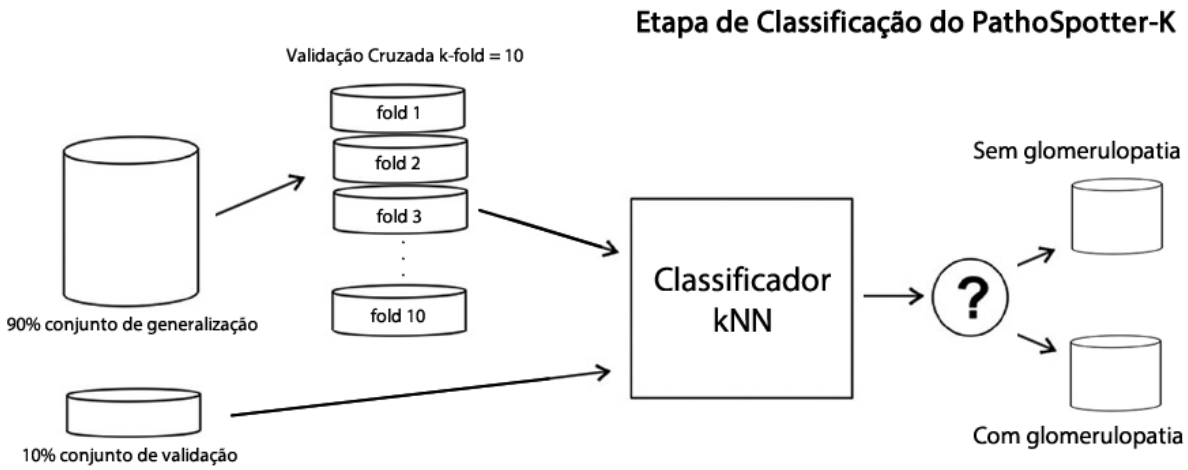

Figura 4.9: Etapa de classificação. Aplicação da validação cruzada e o classificador kNN.

A validação dos resultados obtidos na etapa de classificação do PathoSpotter foi realizada através de dois testes básicos: generalização e validação final. A generalização serviu para avaliar os resultados em meio a diferentes conjuntos de teste/treinamento. A validação final teve como objetivo conferir a qualidade dos resultados obtidos no teste de generalização.

Os resultados da etapa de classificação foram calculados da seguinte forma:

1. Dividimos a matriz de características em duas novas matrizes, de generalização e validação final;
2. Aplicamos a validação cruzada *k-fold* no conjunto de generalização;
3. Calculamos os resultados finais do conjunto de generalização através da média dos resultados parciais de cada *fold* (subconjuntos de teste/treinamento oriundos da validação cruzada);
4. Validamos o sistema utilizando o conjunto de validação final composto por 10% das amostras do conjunto de dados (que não foram utilizadas no teste de generalização).

O resultado do teste de classificação realizado com o conjunto de validação final serviu para, de fato, validar o resultado final obtido com o conjunto de generalização e validação cruzada.

Todos os resultados e experimentos realizados no processo de construção do sistema PathoSpotter são apresentados a seguir no capítulo 5, no qual detalhamos as operações de ajustes de parâmetros, testes de desempenho de métodos e avaliação de resultados.

Capítulo 5

Experimentos e Resultados

"Muitas vezes as coisas que me pareceram verdadeiras quando comecei a concebê-las, tornaram-se falsas quando quis colocá-las sobre o papel."

-- René Descartes

Neste capítulo descrevemos o processo de construção do PathoSpotter, especificamos detalhes sobre seu código, o processo de aquisição das amostras e os experimentos que possibilitaram a seleção dos métodos e parâmetros dos algoritmos que compõem o sistema.

Os experimentos realizados no processo de desenvolvimento do PathoSpotter serviram para selecionar os métodos computacionais mais eficazes no problema de classificação das glomerulopatias proliferativas. Os experimentos foram realizados de forma cíclica e o principio estabelecido no processo foi o de iniciar os testes com os métodos e técnicas mais simples e, caso não fossem atingidos os critérios de precisão requeridos, métodos mais sofisticados seriam agregados para aumentar a qualidade do sistema até que atingíssemos o resultado esperado.

Essa decisão de projeto nos levou ao primeiro parâmetro do sistema que foi a precisão mínima aceitável. Baseando-nos no que trazia a literatura sobre histopatologia digital (ver capítulo 2), e considerando que o trabalho atual do PathoSpotter é pioneiro na área de patologia renal, assumimos que a precisão de classificação mínima para o sistema seria de 80%.

O conjunto de dados (*dataset*) utilizado para a realização dos experimentos foi composto por 811 imagens histológicas de glomérulos renais. Dessas, 511 eram de imagens com glomerulopatia e 300 de imagens sem glomerulopatias (controle normal). No caso específico de avaliação da etapa de segmentação utilizou-se um subconjunto deste conjunto de dados, de 50 imagens (25 imagens sem glomerulopatias e 25 imagens com glomerulopatia). A seleção

de um subconjunto para avaliar a segmentação foi uma estratégia concebida a partir da necessidade da realização de contagem manual das regiões de núcleos das amostras para o estabelecimento de um padrão ouro.

A primeira etapa dos experimentos foi a criação de uma abordagem preliminar, na qual testamos uma abordagem para simular o modo como os patologistas classificam as imagens. Após a construção da abordagem preliminar, que não atingiu a precisão esperada, iniciamos testes para criar uma abordagem de classificação que utilizasse outras características, diferentes das utilizadas pelos patologistas, porém, baseadas no mesmo principio, discriminar as amostras através da identificação de proliferação de núcleos.

Os ciclos de testes se concentraram em cada etapa do sistema (pré-processamento, segmentação, extração de características e classificação). As etapas de pré-processamento e segmentação foram implementadas e avaliadas em um único ciclo de experimentos, logo após realizamos mais dois ciclos, com o foco nas etapas de extração de características e classificação. Por fim, após os ciclos de experimentos de cada etapa do sistema pôde-se obter a versão atual do PathoSpotter. A Figura 5.1 ilustra os ciclos de experimentos realizados na construção do PathoSpotter.

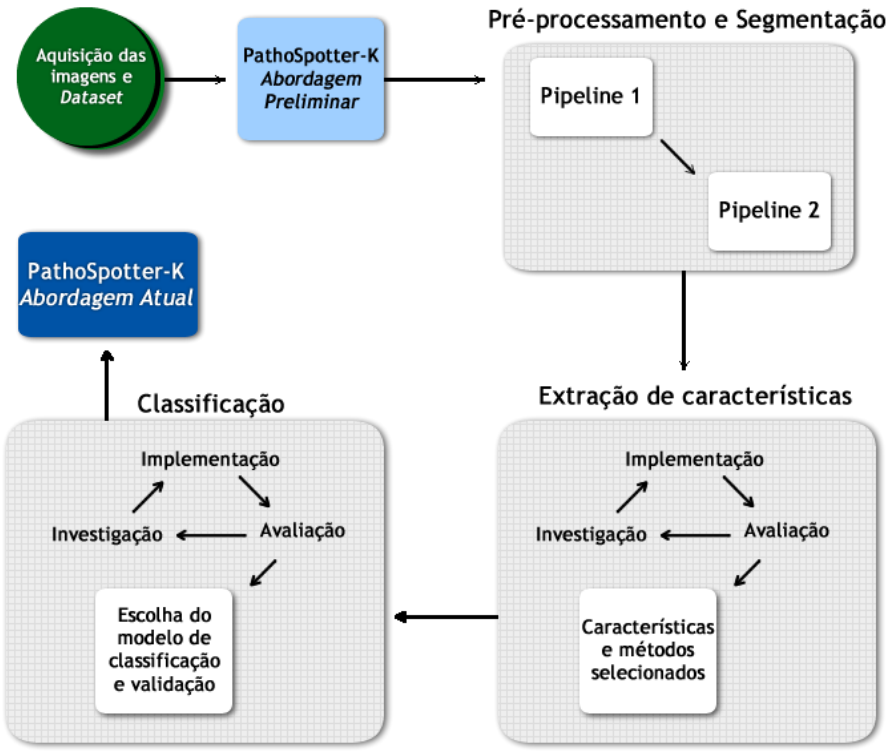

Figura 5.1: Etapas de experimentos.

As etapas de pré-processamento e segmentação foram construídas a partir de duas propostas ou *pipelines* (sequência de operações) distintas. Na extração de características analisamos quais seriam as melhores características discriminantes. Por fim, na classificação definimos e ajustamos o modelo do classificador.

Os testes realizados também tiveram como propósito o estudo da integração de uma etapa do sistema com a etapa posterior, como foi o caso da extração de características e classificação, em que analisamos a capacidade discriminatória das características extraídas antes de utiliza-las em um classificador. Nesse sentido, a preocupação nos testes da etapa de extração de características foi obter informações discriminatórias eficazes para a classificação correta das imagens.

Em relação às métricas de avaliação, por se tratar de um sistema de apoio ao diagnóstico médico, a análise do PathoSpotter considerou como métricas fundamentais os critérios de sensibilidade e especificidade. Essa escolha levou em consideração que, em sistemas dessa natureza, se um paciente for diagnosticado com uma patologia sem tê-la de fato (falso positivo), o diagnóstico errado pode provocar transtornos emocionais para ele e seus familiares, além de leva-lo a submeter-se a um tratamento médico desnecessário. Por outro lado, caso um paciente seja diagnosticado como saudável de maneira equivocada (falso negativo), as consequências podem ser ainda piores, já que isso pode fazer que o mesmo não seja tratado, degradando seu quadro de saúde, com eventual risco de óbito.

5.1 Código do PathoSpotter

O código do PathoSpotter foi implementado em linguagem Python 2.7, podendo ser executado nos sistemas operacionais Windows®, Linux e MacOS X®. A IDE (*Integrated Development Environment*) utilizada foi a Spyder IDE 2.3. A implementação do sistema teve como base as seguintes bibliotecas:

- *Scikit-image* (versão 0.11.0 04/03/2015), para a aplicação de métodos clássicos da área de processamento de imagens;
- *PyMorph* (versão 0.96.0) para a realização de reconstrução morfológica.
- *Scikit-learn* (versão 0.14) para a utilização de algoritmos de aprendizado de máquina.
- *Numpy* (versão 1.10) para a manipulação de matrizes.

- *Matplotlib* (versão 1.5.1) para a visualização de gráficos, histogramas e imagens.
- *PIL* (versão 1.1.6) para a leitura de imagens.

5.2 Aquisição das Imagens e Conjunto de Dados

As imagens processadas no PathoSpotter fazem parte do banco de imagens obtidas pelos patologistas do Centro de Pesquisas Gonçalo Mouniz (CPqGM) da Fundação Osvaldo Cruz (FIOCRUZ).

De maneira simplificada, o processo de aquisição de imagens histológicas é iniciado através da extração de um pequeno pedaço de tecido renal (biópsia). As biópsias renais constituem pequenos fragmentos de tecido biológico, obtidos por agulha ou cirurgicamente. Após o procedimento cirúrgico as biópsias são desidratadas, emblocados em parafusa e cortados em secções de 2 a 3 µm de espessura, quando enfim são coradas, montadas em lâminas de vidro e examinadas ao microscópico.

Em seguida, os patologistas realizam uma ampliação óptica de 2x, 4x, 10x, 40x e 200x (vezes). As imagens do PathoSpotter são imagens renais de néfrons (que incluem os glomérulos) e e receberam grau de ampliação de 200 vezes. O microscópio utilizado na aquisição das imagens do PathoSpotter foi o microscópio óptico Nikon E600, apresentado na Figura 5.2, que mostra o microscópio em um dos ambientes de trabalho dos patologistas.

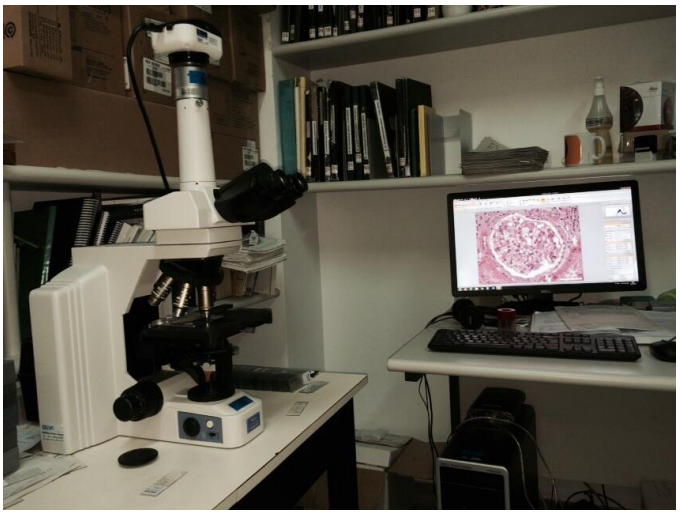

Figura 5.2: Microscópio óptico Nikon E600.

Após o processo de ampliação das imagens através do microscópio, a imagem ampliada foi capturada por uma câmera Olympus Qcolor 3, que se encontrou acoplada ao microscópio

através de um tubo trinocular. A Figura 5.3 mostra com mais detalhes a câmera acoplada ao microscópio e o tubo pelo qual a imagem é capturada pela câmera.

Figura 5.3: Câmera Olympus Qcolor 3 acoplada ao tubo trinocular.

As amostras utilizadas no estudo e execução dos testes do sistema foram disponibilizadas pela equipe do Dr. Washington Luís Conrado dos Santos, médico patologista e pesquisador da FIOCRUZ, o qual possui um conjunto de amostras composto por aproximadamente 10.000 imagens histológicas de diversas patologias que acometem os rins. Todas as amostras utilizadas no PathoSpotter foram classificadas previamente por pelo menos dois patologistas, permanecendo no conjunto de dados do sistema apenas as amostras que foram classificadas por consenso.

O conjunto de dados usado no desenvolvimento do PathoSpotter contou com 811 amostras, sendo 300 de imagens de glomérulos sem glomerulopatias e 511 de imagens que apresentam glomerulopatias.

As imagens histológicas renais que compõem o conjunto de dados do PathoSpotter são imagens específicas de glomérulos. Essas imagens apresentam estruturas comuns entre si, como tecido (Figura 5.4a), e núcleos (na Figura 5.4b) que são as estruturas utilizadas no PathoSpotter para a identificação das glomerulopatias proliferativas.

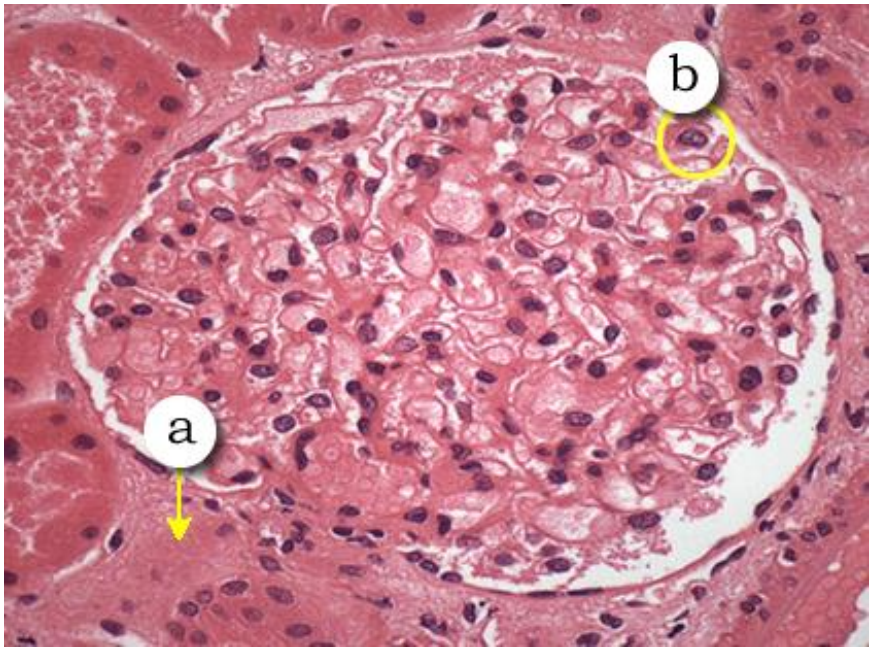

Figura 5.4: Estruturas presentes nas imagens histológicas renais de glomérulos. Tecido (a) e núcleo (b).

Outra característica das imagens que compõem o conjunto de dados do sistema é a presença de cores similares para a representação das estruturas presentes nas amostras. No entanto, essas cores se apresentam com intensidades de brilho diferentes. A Figura 5.5 mostra exemplos de amostras com glomérulos de diferentes formas e diferentes intensidades de brilho nas cores das estruturas, as quais sempre aparecem com um tom avermelhado no tecido (*background* da imagem), tom azul escuro ou preto nos núcleos, e em branco em torno dos glomérulos (Figura 5.5 a, b, c), algo que nem sempre ocorre (Figura 5.5d).

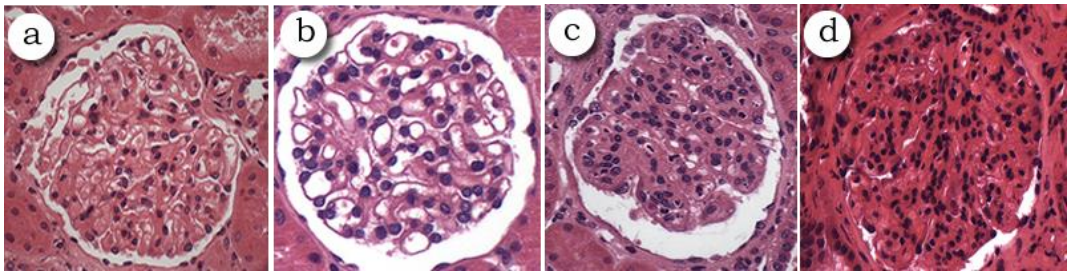

Figura 5.5: Exemplos de imagens que compõem o conjunto de dados do PathoSpotter. Imagens com borda do glomérulo evidente (a, b e c), imagem com glomérulo pouco evidente (d).

A distribuição das amostras do sistema não foi igual para as duas classes de imagens, pois, para os patologistas, as imagens sem glomerulopatias (controle normal) não apresentam informações patológicas importantes para os seus estudos. Dessa forma, não há o hábito de registrar imagens sem glomerulopatias com a mesma frequência que imagens com glomerulopatias.

Além da diferenciação de classe (com glomerulopatia e sem glomerulopatias), as amostras do conjunto de dados também se diferenciam em relação aos corantes utilizados no processo de aquisição, os quais foram o PAS e o H&E, em relação ao tamanho das imagens (que varia entre 362x362 pixels a 1024x768 pixels) e formato (JPEG ou TIFF). A Figura 5.6 mostra um quadro com diferentes imagens que compõem o conjunto de dados do PathoSpotter.

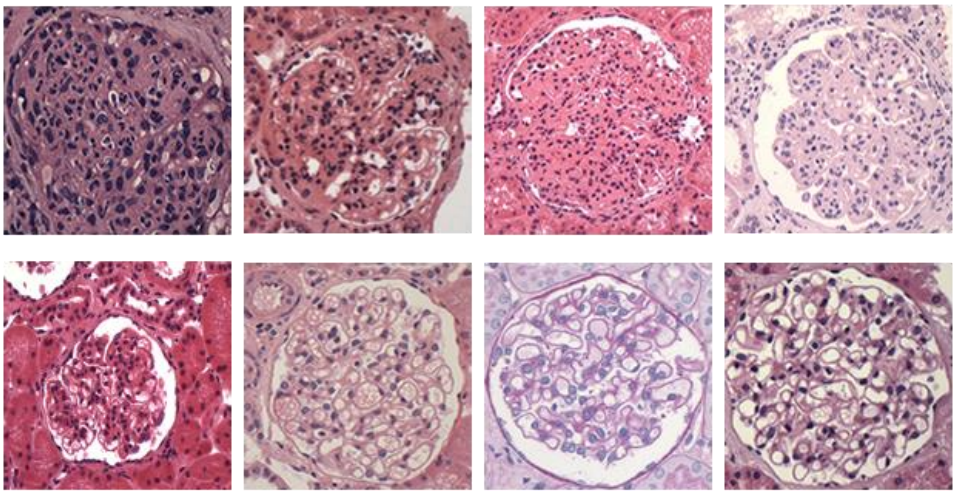

Figura 5.6: Diversidade das imagens que compõem o conjunto de dados do PathoSpotter.

5.3 Abordagem preliminar

A primeira abordagem proposta como arquitetura do PathoSpotter, chamada aqui de abordagem preliminar, utilizou como princípio de classificação das imagens uma das características discriminantes utilizadas pelos patologistas na classificação manual, que é a observação de formação de aglomerações de núcleos (*clusters*).

Nessa abordagem, as imagens nas quais é possível constatar a presença de mais de três núcleos unidos, formando um *cluster*, são classificadas como "com glomerulopatia", caso contrário às imagens são classificadas como "sem glomerulopatia". Essa proposta representa, de um modo simplificado, a forma como os patologistas classificam as imagens quanto as glomerulopatias. A Figura 5.7 mostra a arquitetura da Abordagem Preliminar do PathoSpotter e todos os métodos utilizados em cada etapa do sistema.

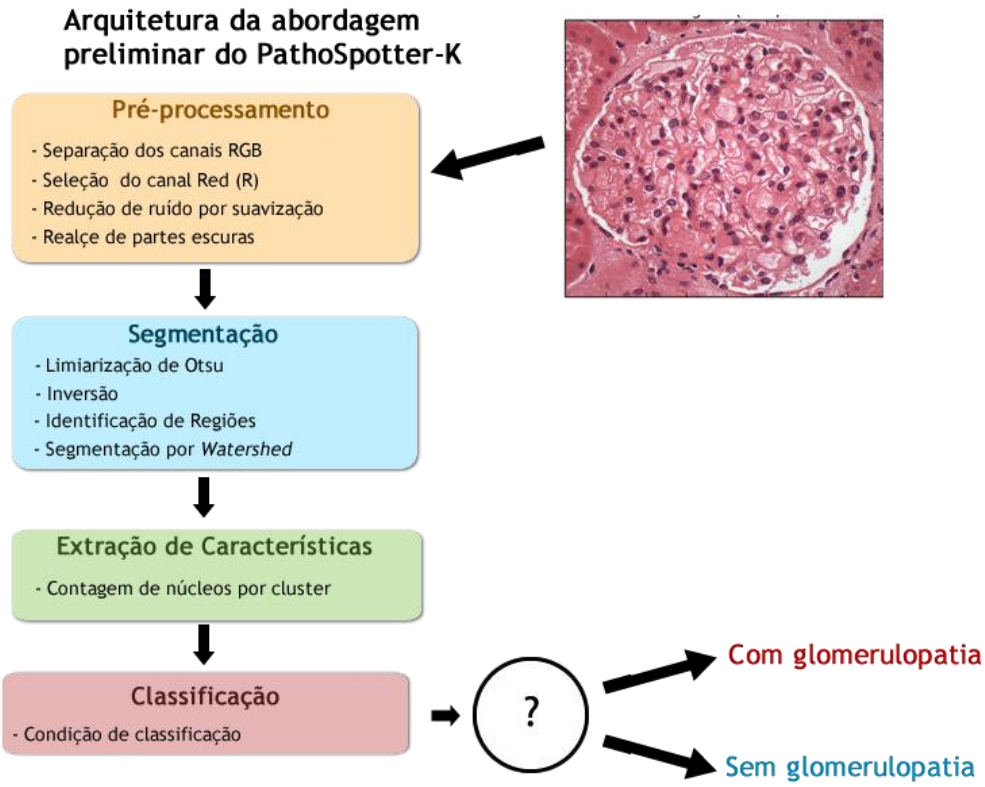

Figura 5.7: Arquitetura da abordagem preliminar.

A etapa de pré-processamento serviu para reduzir ruídos e enfatizar as regiões dos núcleos. Para isso, inicialmente separou-se o canal *Red* (R) das imagens, pelo fato desse canal apresentar maior contraste entre os núcleos e o fundo das imagens. O canal R representa as informações de cor vermelha das imagens, logo, neste canal, o fundo (que possui tonalidade vermelha) é mais claro que os núcleos. Logo depois da separação do canal R aplicou-se um filtro de média (com mascara de convolução de tamanho 3x3) para a realização da suavização, o que proporcionou a redução de ruídos. Por fim, após a suavização, realizou-se um realce logarítmico (ver seção 3.1.2), com o objetivo de enfatizar as partes escuras da imagem, que são os núcleos. O realce logarítmico calcula um novo valor de intensidade para cada pixel da imagem através da função *log*.

Na segmentação, o método aplicado foi o Otsu global (apresentado na sessão 3.1.3), que calcula, de maneira automática, um parâmetro de limiarização para binarizar para cada imagem. Depois, realizamos a identificação dos *clusters* através da análise de borda e vizinhança entre pixels, usando o método de crescimento de regiões (apresentado na sessão 3.1.3). Neste método, os pixels que possuem um vizinho com valor de intensidade de pixel diferente do seu, é considerado como um dos pixels que compõe a borda de uma região, caso contrário, onde sua vizinhança possua o mesmo valor de intensidade de pixels, o conjunto de

pixels é selecionado como parte integrante de uma mesma região, a qual pode ser um núcleo isolado ou um cluster.

A última técnica de segmentação foi o método de divisor de águas, *watershed* (apresentado na sessão 3.1.3) para realizar a contagem de núcleos presentes em cada região. No método de segmentação por divisor de águas a imagem a ser segmentada é interpretada como uma superfície topográfica, em que as intensidades dos pixels correspondem a valores de altitude dos pontos. Na prática, esse procedimento separa objetos que estão "unidos", atribuindo uma cor diferente para cada núcleo identificado. Ao converter as imagens oriundas da segmentação por divisor de águas em nível de cinza e realizar a contagem de tons, obtivemos o número de núcleos identificados em cada *cluster* presente na imagem.

Na etapa de extração de características, a tarefa realizada foi a contagem de núcleos em cada *cluster* (ou região de núcleos), tarefa realizada a partir do resultado da segmentação de cada *cluster*.

Na etapa de classificação, utilizamos um classificador binário baseado em uma simples estrutura condicional. Caso fossem detectados três ou mais núcleos unidos (formando um *cluster*), a imagem seria classificada como uma amostra "com glomerulopatia", caso contrário (menor do que três) classificaria como "sem glomerulopatia".

A Figura 5.8 mostra os resultados de cada etapa da Abordagem Preliminar (pré-processamento e segmentação, extração de características e classificação). Pelo fato da imagem apresentada ser uma imagem com glomerulopatia e ter sido identificado um aglomerado de núcleos composto por 4 núcleos, o exemplo apresentado na Figura 5.8 revela um caso em que a classificação foi bem sucedida.

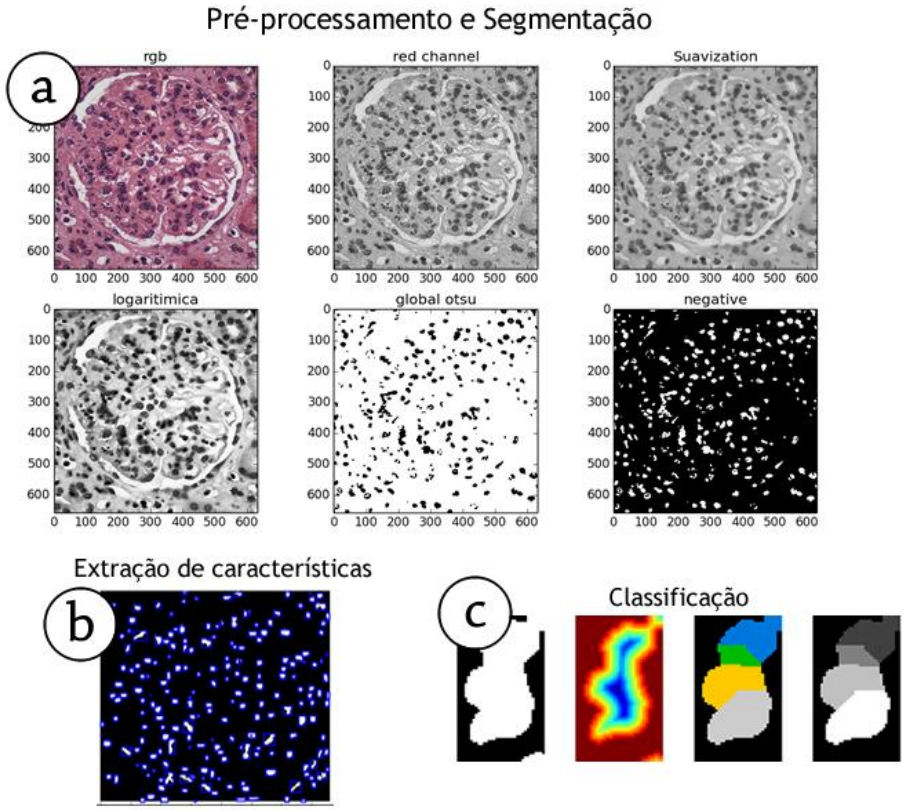

Figura 5.8: Aplicação da abordagem preliminar em uma imagem exemplo. Etapa de pré-processamento (a), etapa de extração de características (b), etapa de classificação (c).

Apesar do exemplo apresentado na Figura 5.8, de classificação bem sucedida, consideramos a abordagem preliminar frágil, por ser bastante subjetiva. Os patologistas contam com sua experiência prévia, e obviamente, a capacidade de compreensão de formas e estruturas complexas que, até o momento, só um humano tem. Além disto, esta abordagem é muito sensível à qualidade da segmentação, já que, com os métodos testados, a identificação dos *clusters* gerou muitos falsos positivos, por causa da dificuldade na escolha dos parâmetros corretos de segmentação em função da diversidade das imagens.

A Figura 5.9 mostra a etapa inicial e final da aplicação da abordagem preliminar em uma imagem onde a classificação não é realizada com sucesso, pelo fato do algoritmo de segmentação por divisor de águas segmentar excessivamente o *cluster* identificado. Nesse caso especifico, uma região onde há apenas dois núcleos (que não caracteriza uma lesão histológica) é reconhecida como tendo três núcleos. A esse acontecimento se dá o nome de segmentação excessiva (*over segmentation*). Na Figura 5.9 ainda é possível observar os núcleos (enumerados), na imagem original dos clusters e como a abordagem preliminar

contou os núcleos, em nível de cinza.

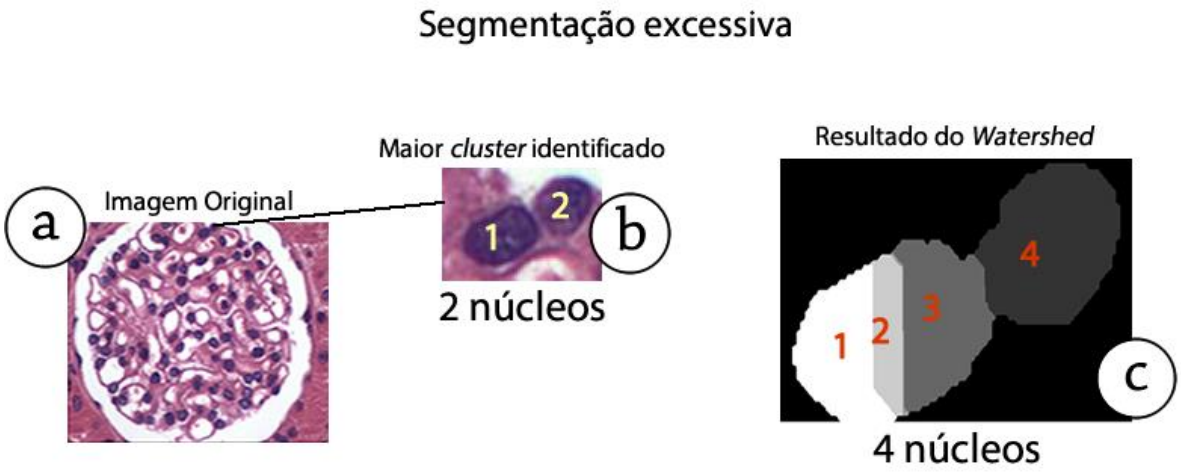

Figura 5.9: Segmentação excessiva. Imagem original (a), segmentação ideal (b) e segmentação realizada pela abordagem preliminar (c).

Para as 811 imagens que compõem o atual do conjunto de dados do sistema (300 sem glomerulopatia e 511 com glomerulopatia) obteve-se uma sensibilidade de 81%, especificidade de 19% e acurácia de 58%. Esse resultado nos fez concluir que a abordagem preliminar é inadequada e extremamente sensível à variação de imagens com diferentes aspectos. A matriz de confusão desta abordagem pode ser visualizada na Tabela 5.1

Tabela 5.1: Matriz de confusão da abordagem preliminar.

Resultado da Classificação	**Amostras com glomerulopatia (511)**	**Amostras sem glomerulopatia (300)**
Positivo	TP=418	FP=241
Negativo	FN=93	TN=59

Através dos resultados obtidos, pode-se constatar alguns problemas na abordagem preliminar, como a alta subjetividade do principio utilizado para realizar a classificação e o fato da segmentação por divisor de águas ser sensível a ruídos, influenciando assim na contagem dos núcleos. Sendo assim, descartamos a abordagem preliminar como uma arquitetura adequada, porém entendemos que a sua elaboração foi importante para amadurecermos o processo de construção do PathoSpotter, pois serviu para testarmos os primeiros métodos a serem experimentados, principalmente nas etapas de pré-processamento, segmentação e extração de

características.

5.4 Abordagem Atual

A abordagem atual do PathoSpotter foi obtida após a realização de experimentos com técnicas de processamento de imagem e de modelagem do classificador, visando aumentar a precisão final do sistema. Na etapa de pré-processamento e segmentação foram experientados duas propostas. Na etapa de extração de características, investigamos quatro diferentes características, até escolher as que apresentaram potencial para discriminar as amostras. Por fim, na etapa de classificação, testamos duas abordagens, baseadas em regressão logística e classificador kNN.

5.4.1 Pré-processamento e segmentação

As etapas de pré-processamento e segmentação tiveram como objetivo aperfeiçoar a imagem e selecionar as regiões onde houvesse núcleos. Os métodos de pré-processamento e segmentação foram avaliados de maneira conjunta, pelo fato das duas etapas estarem intimamente relacionadas. Portanto, apresentaremos o pré-processamento e a segmentação em conjunto.

Após investigar técnicas clássicas do campo de processamento de imagens e métodos utilizados comumente em trabalhos similares, criamos duas propostas de pré-processamento e segmentação. Para avaliar cada proposta, um subconjunto de 50 imagens foi selecionado do conjunto de dados do sistema e realizamos a contagem manual de regiões de núcleos presentes nas imagens desse subconjunto de amostras. O subconjunto de avaliação das propostas de pré-processamento e segmentação foi divido em 25 imagens sem glomerulopatia e 25 imagens com glomerulopatia, com diferentes características de resolução e corante.

Os resultados obtidos através da contagem manual foram considerados como o padrão ouro para a avaliação das respectivas propostas de pré-processamento e segmentação. A contagem considerou cada núcleo isolado, e cada cluster, como uma região, exatamente como as regiões foram identificadas através dos métodos automáticos. Logo, a contagem não ofereceu um resultado da quantidade de núcleos presentes em cada amostra, mas a quantidade de regiões de núcleos, que inclui tanto núcleos isolados quanto *clusters*.

A qualidade das propostas de pré-processamento e segmentação foi avaliada através da

medida de quão próxima à contagem automática ficou da contagem manual. Em alguns casos a contagem automática superou a contagem manual (*over segmentation*), contudo, isso não influenciou no resultado final, pois a avaliação foi realizada a partir da proximidade entre o resultado da contagem automática e o resultado da contagem manual. O cálculo da taxa de erro de cada proposta foi realizado pela Equação 5.1, onde *ca* é o resultado da contagem automática e *cm* o resultado da contagem manual.

$$taxa\ de\ erro = \frac{|cm - ca|}{cm} * 100 \tag{5.1}$$

Apresentamos a seguir as tarefas realizadas e também as razões que nos fizeram utilizar os métodos empregados, bem como os testes realizados na escolha desses métodos.

5.4.1.1 Proposta 1

As imagens a seguir mostram as atividades que compõem a proposta 1 (Figura 5.10) e ilustram a aplicação dos métodos em uma imagem exemplo (Figura 5.11).

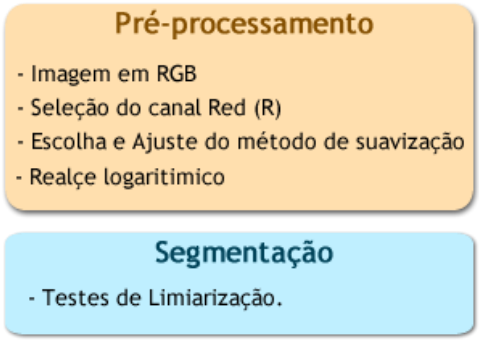

Figura 5.10: Pré-processamento e segmentação, proposta 1.

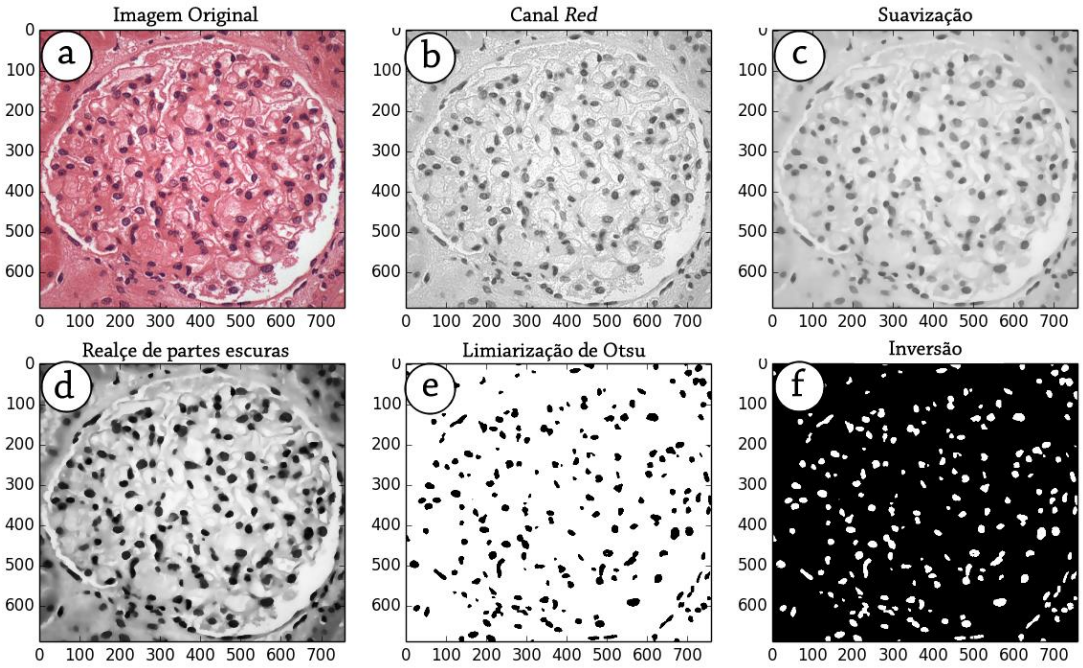

Figura 5.11: Exemplo de aplicação da proposta 1 em uma imagem. Imagem original (a), seleção do canal Red (b), aplicação do filtro de suavização (c), realização de realce de partes escuras (d), limiarização automática por Otsu (e), calculo da inversa da imagem (f).

Assim como na abordagem preliminar, a primeira tarefa realizada na proposta 1 foi a utilização do canal R da imagem, a qual estava representada no espaço de cor RGB. Entre os canais que compõem uma imagem RGB, o canal R se mostrou como o mais adequado para iniciar a abordagem de pré-processamento e segmentação, justamente pelo fato desse canal representar as informações de tonalidade vermelha, presentes nas imagens.

As imagens do conjunto de dados possuem como característica a tonalidade azulada ou preta nas regiões dos núcleos, branca na borda do glomérulo e tonalidade vermelha nas demais regiões da imagem.

Após a escolha do canal R como primeiro passo da etapa de pré-processamento, foi necessário reduzir ruídos relacionados à textura. Para isso realizou-se a suavização das imagens, tarefa que diminuiu a variação de intensidade entre os pixels.

Dois tipos de filtros de suavização foram testados. O primeiro filtro de suavização testado foi o filtro de média. A Figura 5.12 mostra o resultado da aplicação do filtro de média, na qual cada imagem diz respeito a um tamanho de janela de convolução (*kernel*) utilizado, 2x2, 3x3, 4x4, 5x5, 6x6 e 7x7 *pixels*, respectivamente.

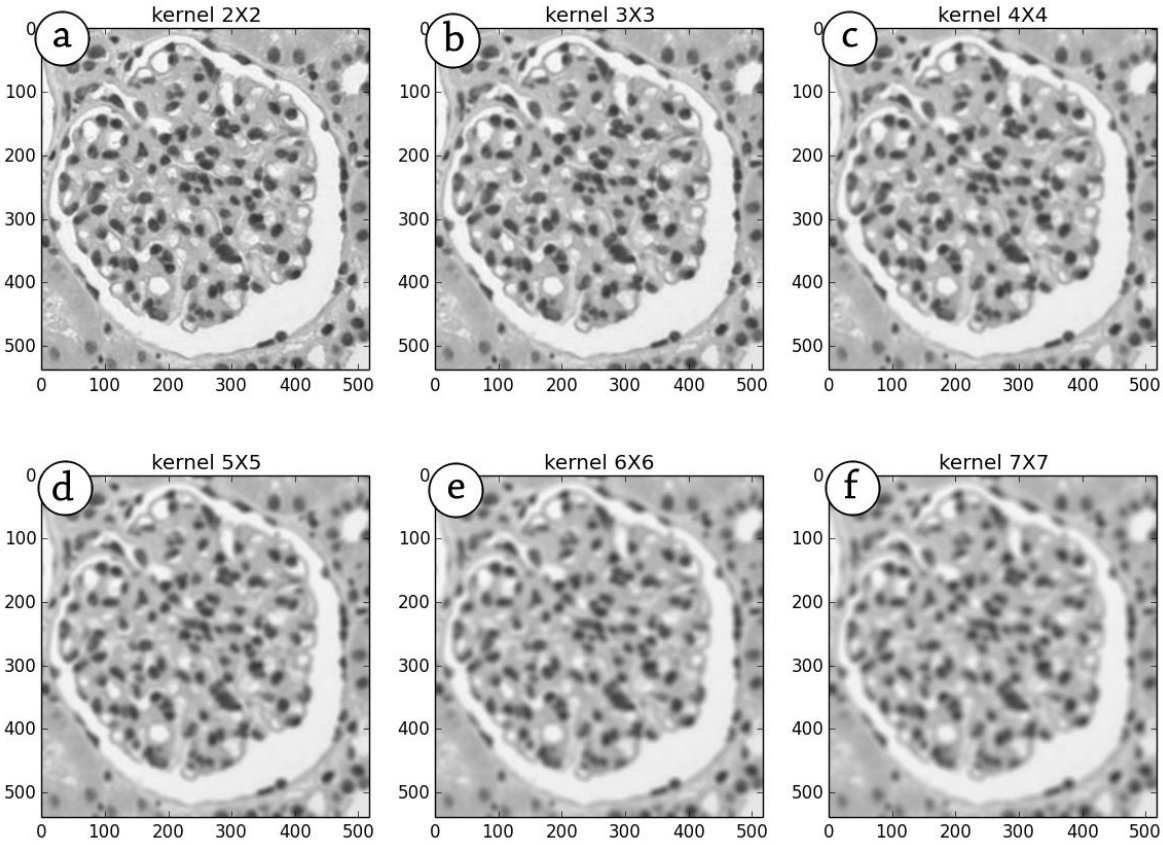

Figura 5.12: Filtro de média, aplicação de diferentes tamanhos de janela de convolução. Resultados da utilização dos filtros de tamanho 2x2 (a), 3x3 (b), 4x4 (c), 5x5 (d), 6x6 (e), 7x7 (f).

O segundo filtro de suavização testado nas imagens foi o filtro de mediana. A mesma variação de tamanho de janela de convolução realizada no teste com filtro de média foi realizada com o filtro de mediana. A Figura 5.13 mostra as imagens resultantes da aplicação do filtro de mediana.

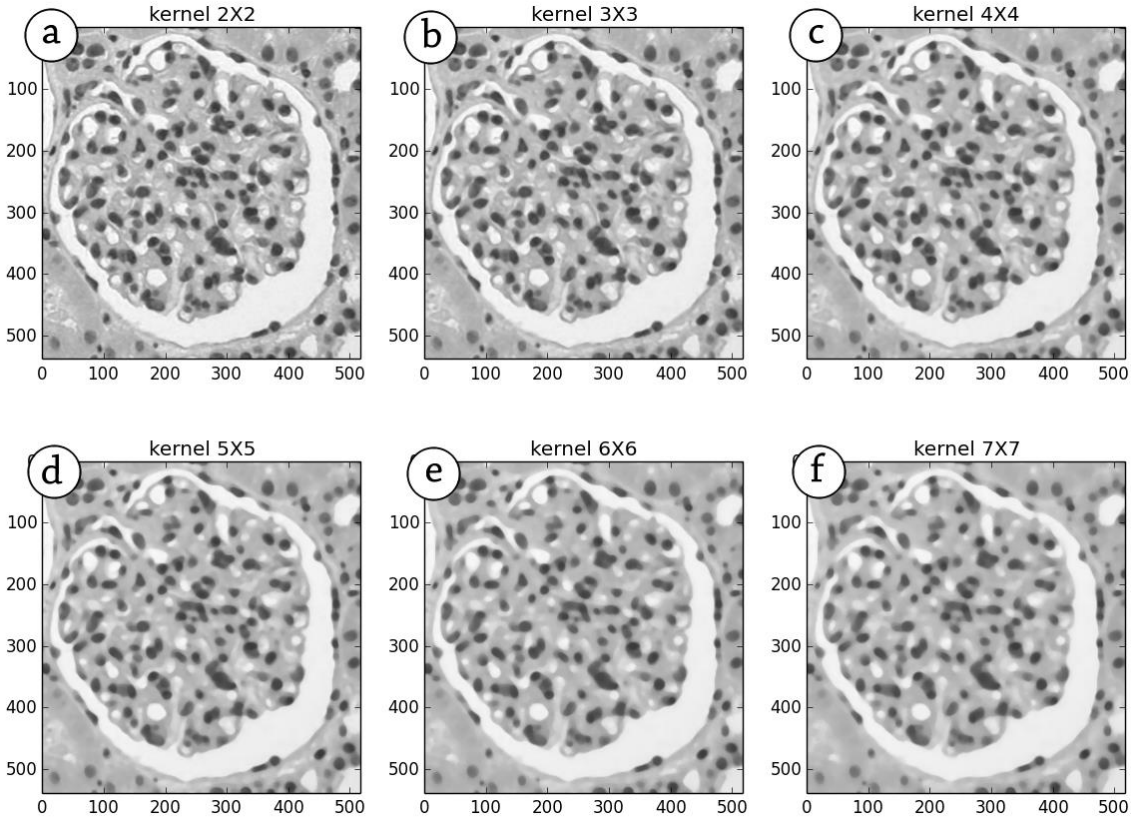

Figura 5.13: Filtro de mediana, aplicação de diferentes tamanhos de janela de convolução. Resultados da utilização dos filtros de tamanho 2x2 (a), 3x3 (b), 4x4 (c), 5x5 (d), 6x6 (e), 7x7 (f).

A escolha do filtro de suavização mais adequado se deu através do fato de que o filtro de mediana tem como característica preservar informações de bordas [Davies, 2012, p.44], algo importante para o PathoSpotter, pois o objetivo da suavização nessa abordagem é reduzir a variação de intensidade de pixels sem interferir drasticamente nos núcleos. Este fato pode ser observado nos testes de suavização através da Figura 5.14, que compara o filtro de média e mediana com um mesmo tamanho de janela de convolução.

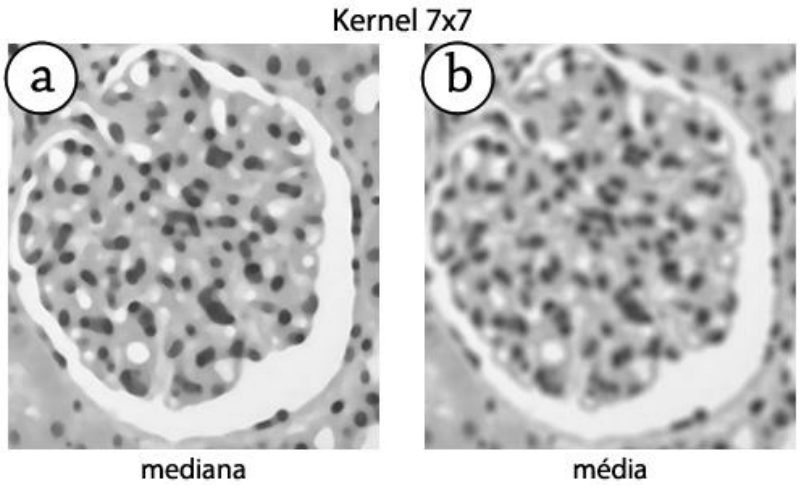

Figura 5.14: Diferença entre filtros de mediana (a) e média (b).

Pelo fato da literatura disponível apresentar o filtro de mediana superior ao filtro de média no que diz respeito a suscetibilidade ao efeito de *bluring* (ver seção 3.1.2), a comparação apresentada aqui objetivou simplesmente ilustrar as vantagens ao se utilizar o filtro de mediana.

Após a escolha do filtro de suavização, foi necessário escolher o tamanho da janela de convolução mais adequado para a realização da suavização. Os tamanhos de janela de convolução testados nas imagens foram 3x3, 4x4, 5x5, 6x6 e 7x7 *pixels*. A partir do fato de que o crescimento da janela de convolução está diretamente relacionado à perda de detalhes em uma imagem e uma menor janela de convolução executa menos operações matemáticas computacionais (menos recurso computacional) [Burger e Burge, 2009, p.112], o tamanho de janela de convolução escolhido para a suavização foi a menor janela testada, 3x3. A Figura 5.15 mostra o resultado da aplicação de um filtro de mediana com janela de convolução 3x3.

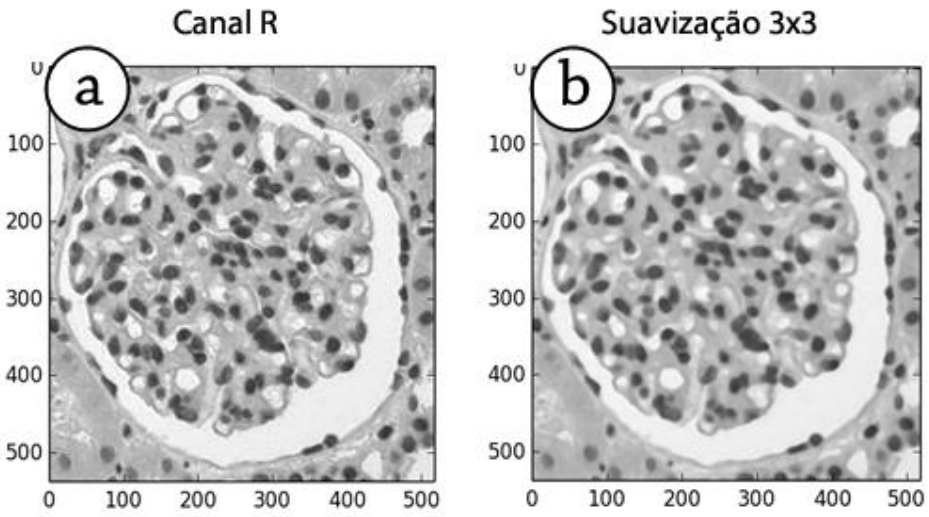

Figura 5.15: Imagem sem a suavização (a) e com a suavização (b).

A suavização das imagens é uma operação importante da etapa de pré-processamento, pois influencia diretamente o processo de segmentação. A Figura 5.16 revela o impacto da suavização das imagens na etapa de segmentação ao mostrar o resultado final de segmentação, com a suavização e sem a suavização. Como é possível observar nas imagens, especialmente nas áreas circuladas, a imagem cujo filtro de suavização não foi aplicado antes da segmentação, apresentou mais ruído, relacionado diretamente a informação de textura dos núcleos, caracterizada justamente pela alta variação de pixels nessas regiões.

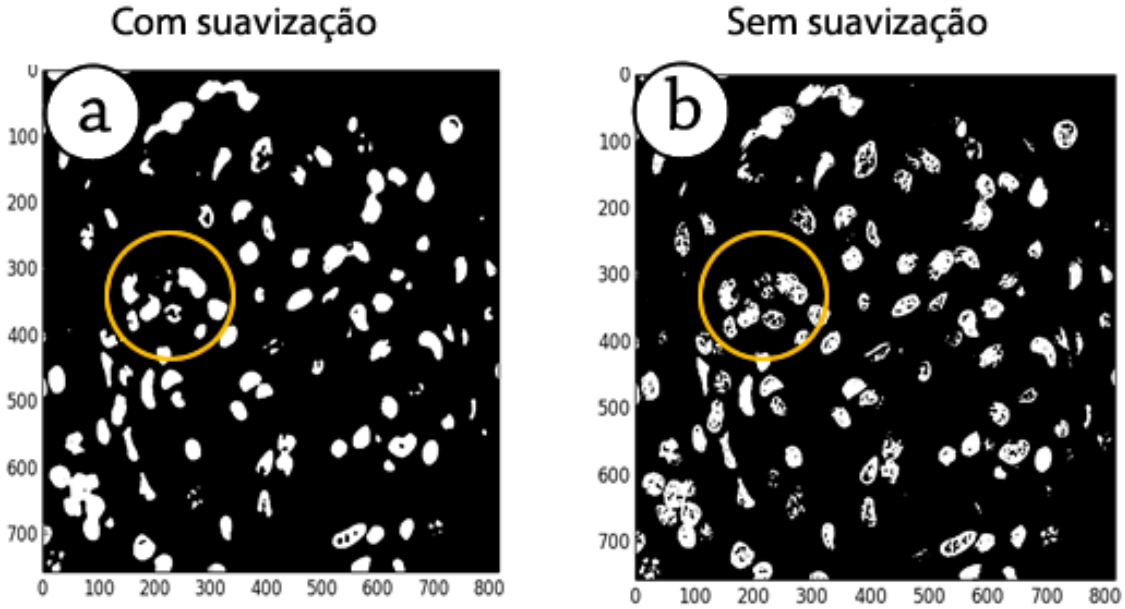

Figura 5.16: Imagens segmentadas com (a) e sem a suavização (b).

Como já foi mencionado, a suavização diminui o intervalo de intensidade de pixels, sendo assim, permite que os núcleos sejam segmentados com maior qualidade, como é o caso da imagem mostrada na Figura 5.16.

Antes de realizar a segmentação das imagens, foi necessário realçar os núcleos, facilitando assim o processo de escolha do limiar de segmentação das imagens (nesse caso, binarização), que é um valor de intensidade de pixels pelo qual os pixels da imagem são classificados em duas possíveis classes de pixels: região de núcleos e fundo.

Para realçar os núcleos foi realizada uma operação de realce logarítmico, também conhecida como realce de partes escuras, operação utilizada, pois a imagem resultante desse momento específico da abordagem possuíra nos núcleos um brilho menor, e um brilho maior no fundo. A Figura 5.17 ilustra a imagem antes da aplicação do realce de partes escuras e após o realce, além de apresentar seus respectivos histogramas.

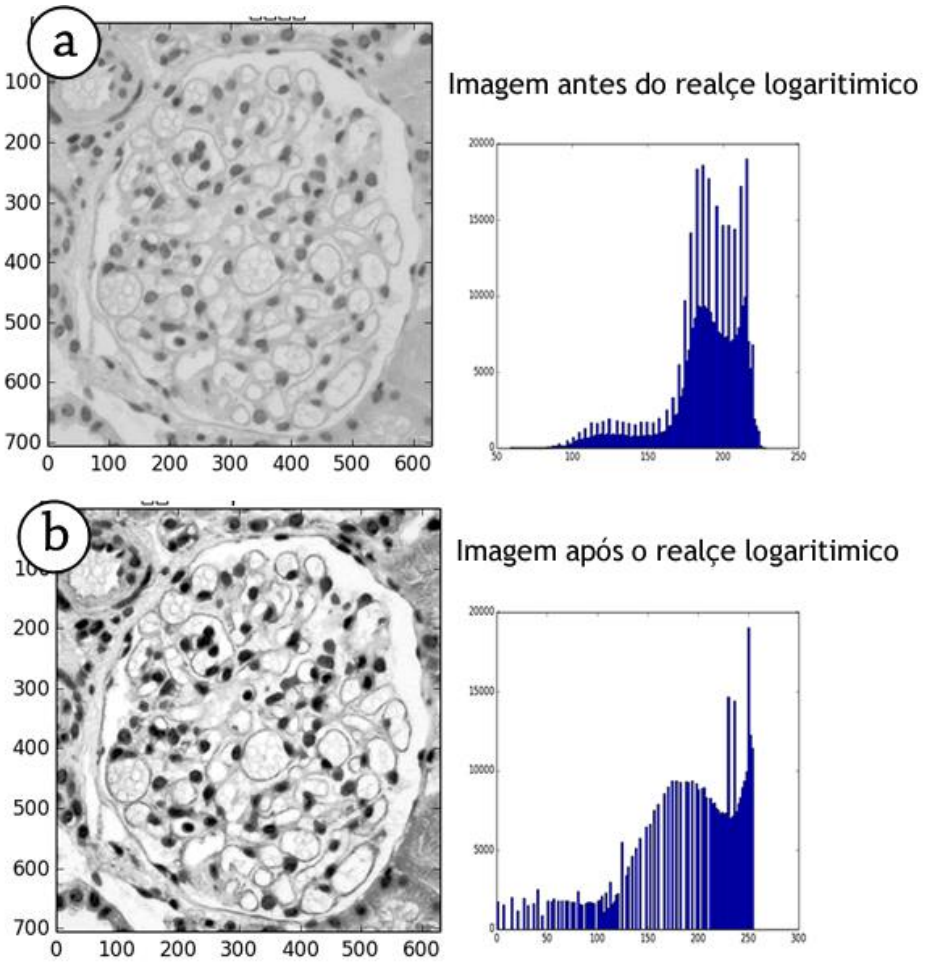

Figura 5.17: Realce de partes escuras. Imagem antes do realce (a) e após o realce (b).

Como pode ser observado através das imagens e histogramas o realce de partes escuras

possibilitou o aumento de contraste na imagem, o que, por sua vez, facilitou a segmentação. Assim como foi testado na suavização, verificamos a influência do realce de partes escuras no resultado final da segmentação da região dos núcleos. Ao observar a Figura 5.18 é possível perceber ruídos na imagem final segmentada sem a realização prévia do realce. Por outro lado, na imagem com o realce há uma segmentação mais precisa dos núcleos.

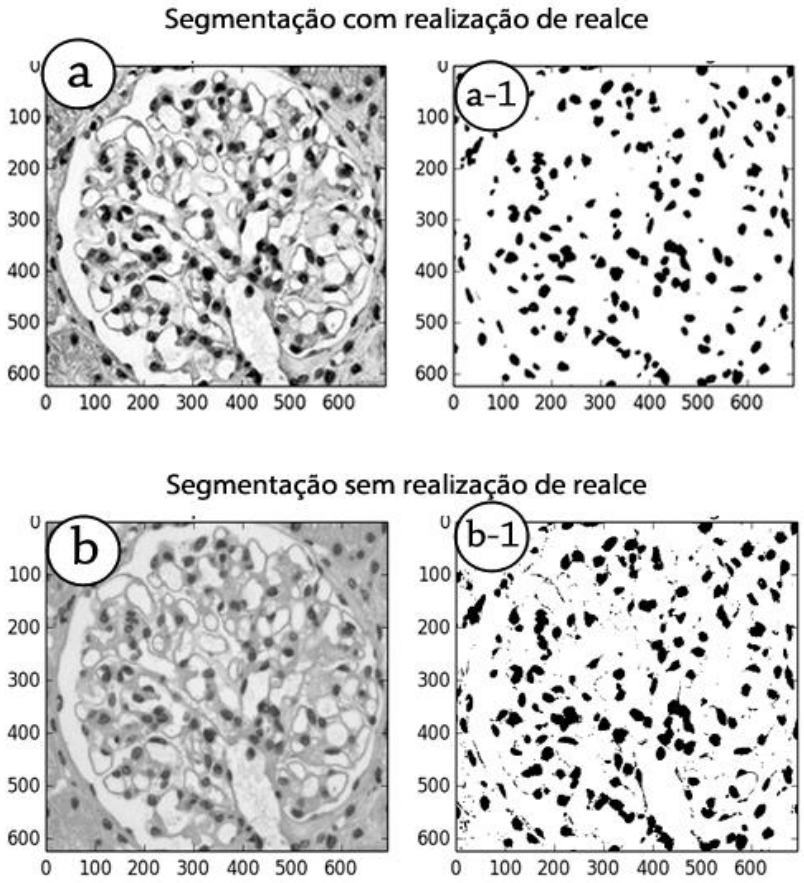

Figura 5.18: Influência do realce de partes escuras no final da segmentação. Imagem resultante do realce (a) e resultado da limiarização dessa imagem (a-1). Imagem sem realce (b) e o seu respectivo resultado de limiarização (b-1).

A última operação realizada no proposta 1 foi a binarização das imagens. Inicialmente, experimentamos a utilização de um limiar de binarização (*threshold*) fixo. A Figura 5.19 exemplifica os resultados obtidos ao variar um limiar fixo em duas imagens.

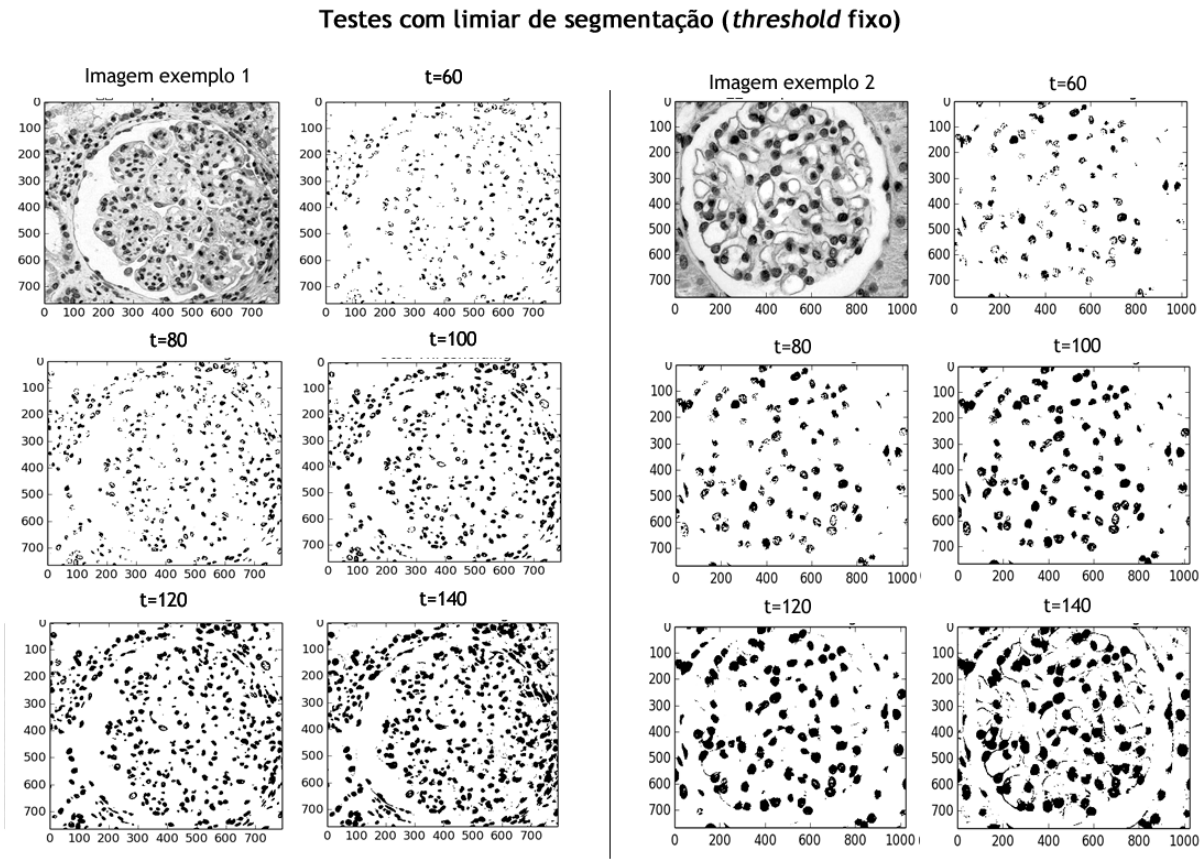

Figura 5.19: Testes com limiar de segmentação. Variando o valor do limiar (em duas imagens diferentes) de 60 a 140 (em um intervalo de 20 em 20).

No caso do PathoSpotter, onde a limiarização deve funcionar bem para diversas imagens, as quais possuem diferentes intervalos de intensidade de brilho, esta abordagem não se demonstrou eficiente. Nas imagens apresentadas na Figura 5.19, o valor de limiar igual a 140, por exemplo, proporciona um resultado relativamente aceitável na imagem 1 (pouco ruído), porém na imagem 2 o resultado é uma imagem com excesso de ruído. A estratégia mais propícia para o caso do PathoSpotter foi a utilização de um método automático de limiarização, no qual fosse calculado um limiar específico para cada imagem, tornando assim, o processo de segmentação dinâmico.

A forma como o processo de segmentação do PathoSpotter foi modelado e desenvolvido, fez com que os histogramas das imagens com realce apresentassem comumente duas classes básicas de pixels, referentes as regiões de núcleos e fundo. Portanto, o método apontado na literatura como a provável melhor solução foi o método automático de Otsu, o qual pode ser implementado de forma local ou global. A Figura 5.20 apresenta os resultados da aplicação do método de limiarização por Otsu na sua forma global e local. Como pode ser observado na

Figura 5.20, o método de limiarização global se apresentou como a melhor alternativa.

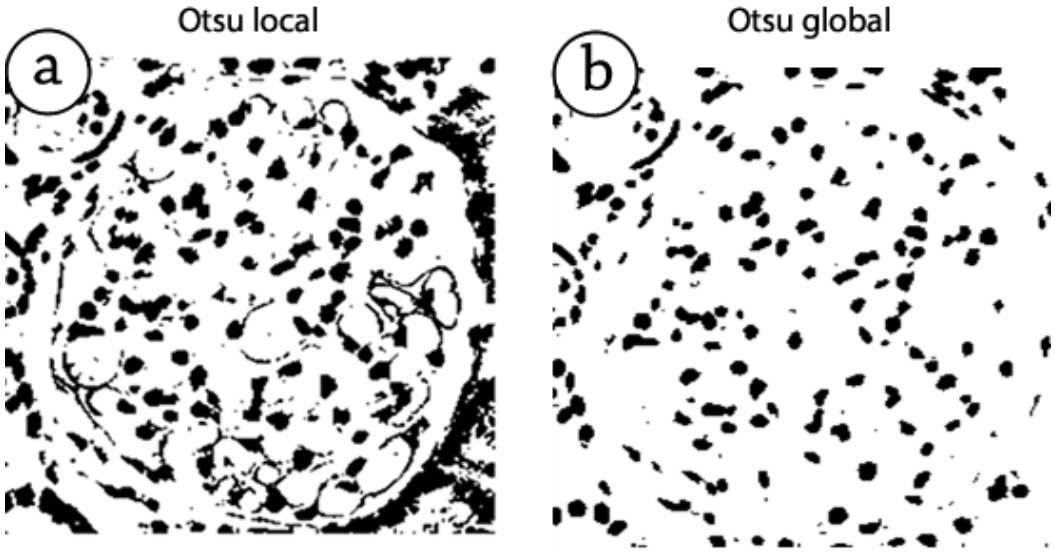

Figura 5.20: Limiarização por Otsu, na versão local (a) e global (b).

O método de Otsu local funciona melhor em casos onde os detalhes da imagem devem ser preservados. No caso do PathoSpotter, o objetivo foi extrair apenas as regiões de núcleos, regiões de intensidade de brilho mais baixa, logo o método global que segmenta a imagem a partir de um único calculo de limiar, apresentou resultados mais satisfatórios.

Tendo em vista as informações encontradas na literatura sobre as vantagens da utilização do método de Otsu em casos de histogramas com duas classes, além dos trabalhos com imagens histológicas que utilizaram esse método [Davies, 2012, p.108], resolvemos nos concentrar apenas no método de Otsu.

Após a segmentação final, utilizou-se a propriedade *regionprops* da biblioteca de processamento de imagens *scikit-image*, que utiliza o método de crescimento de regiões para identificar as regiões de núcleos presentes na imagem (ver seção 3.1.3). O método *regioprops* realiza a contagem das regiões detectadas, esta informação é o resultado final da contagem automática.

5.4.1.2 Proposta 2

A proposta 2 de pré-processamento e segmentação nos baseamos em diferentes trabalhos similares com imagens histológicas, utilizando técnicas e métodos distintos dos métodos utilizados no proposta 1. A Figura 5.21 mostra um quadro com as operações que compõem o proposta 2, e a Figura 5.22 mostra um exemplo da aplicação do proposta 2 em uma amostra.

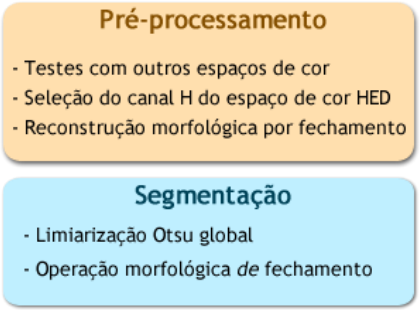

Figura 5.21: Segundo proposta de pré-processamento e segmentação.

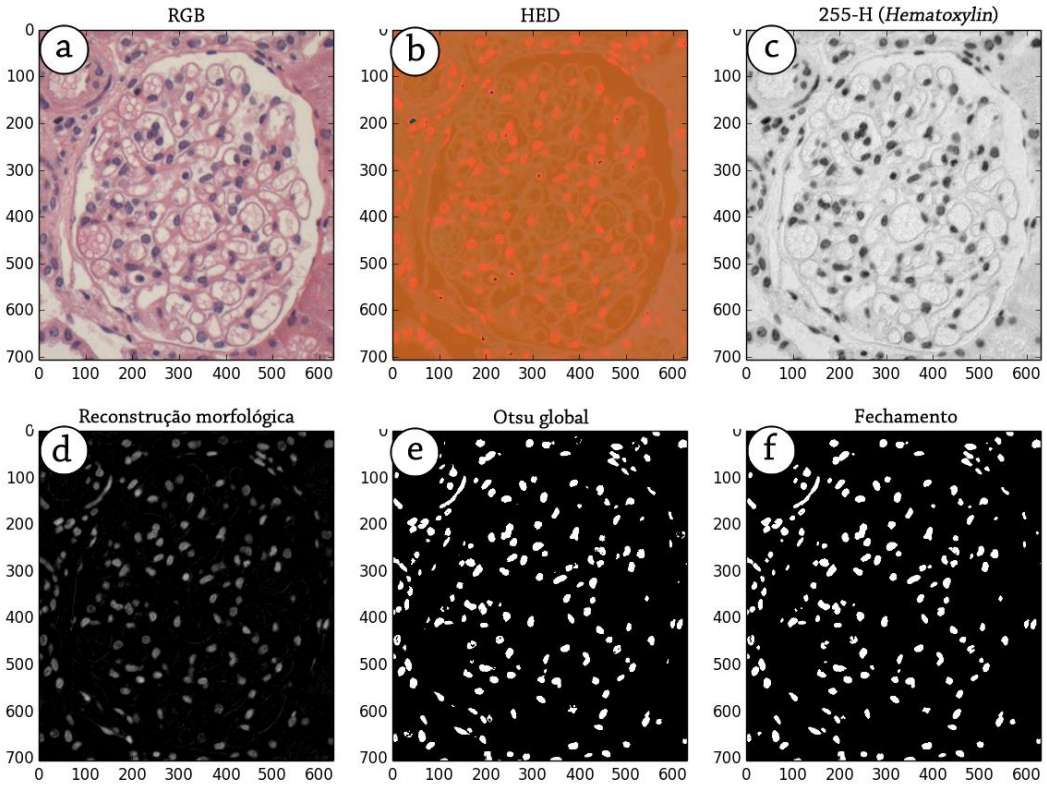

Figura 5.22: Aplicação do proposta 2 em uma imagem exemplo. Imagem original em rgb (a), conversão para o espaço de cor hed (b), seleção do canal *Hematoxylin* (c), resultado da reconstrução morfológica por fechamento (d), limiarização por Otsu global, e resultado final da segmentação com a operação morfológica de fechamento (e).

A primeira tarefa no proposta 2 foi ampliar as possibilidades de observação da imagem através de diferentes espaços de cor. Apesar do espaço de cor RGB apresentar bons resultados no pré-processamento, procuramos outros espaços de cor que pudessem representar melhor a região dos núcleos. A Figura 5.23 mostra um quadro com os espaços de cor testados e seus respectivos canais de cor.

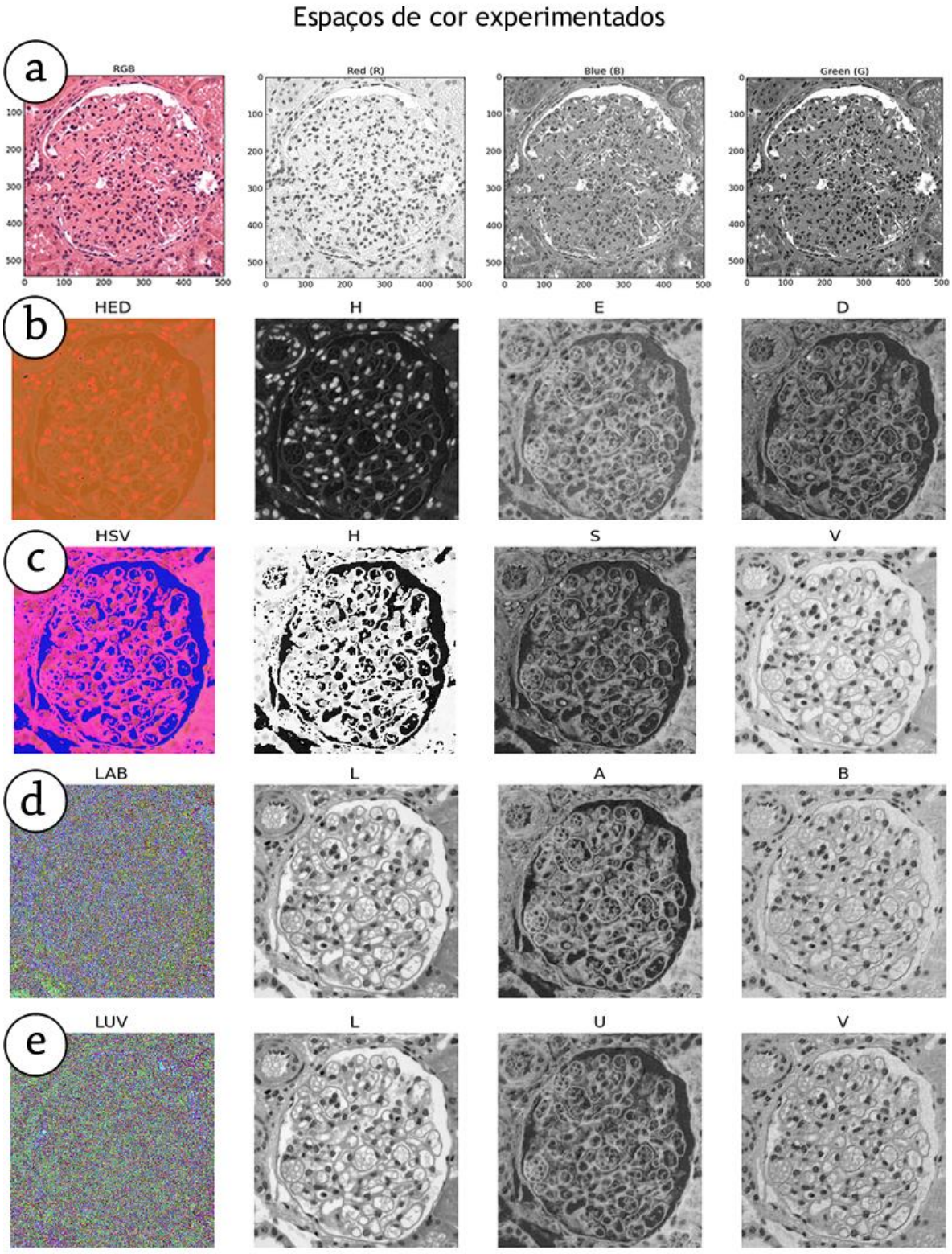

Figura 5.23: Espaços de cor testados e seus respectivos canais. RGB (a), HED (b), HSV (c), LAB (d), LUV (e).

Após separar os canais dos espaços de cor LUV (*luminescence*, *saturation* e *hue angle*), HSV (*hue*, *saturation* e *value*), LAB (*luminescence*, *a=red/green* e *b=blue/yellow*) e HED

(*hematoxylin*, *eosin* e *DAB*), observamos que alguns canais de cor se destacaram, ao apresentar bons resultados de contraste entre os núcleos e as demais regiões das imagens. A Figura 5.24 apresenta os melhores canais escolhidos a partir de observação visual, os quais foram os canais H (do HED), V (do HSV), L e B (do LAB). Adicionalmente a Figura 5.24 também apresenta os respectivos histogramas dos canais.

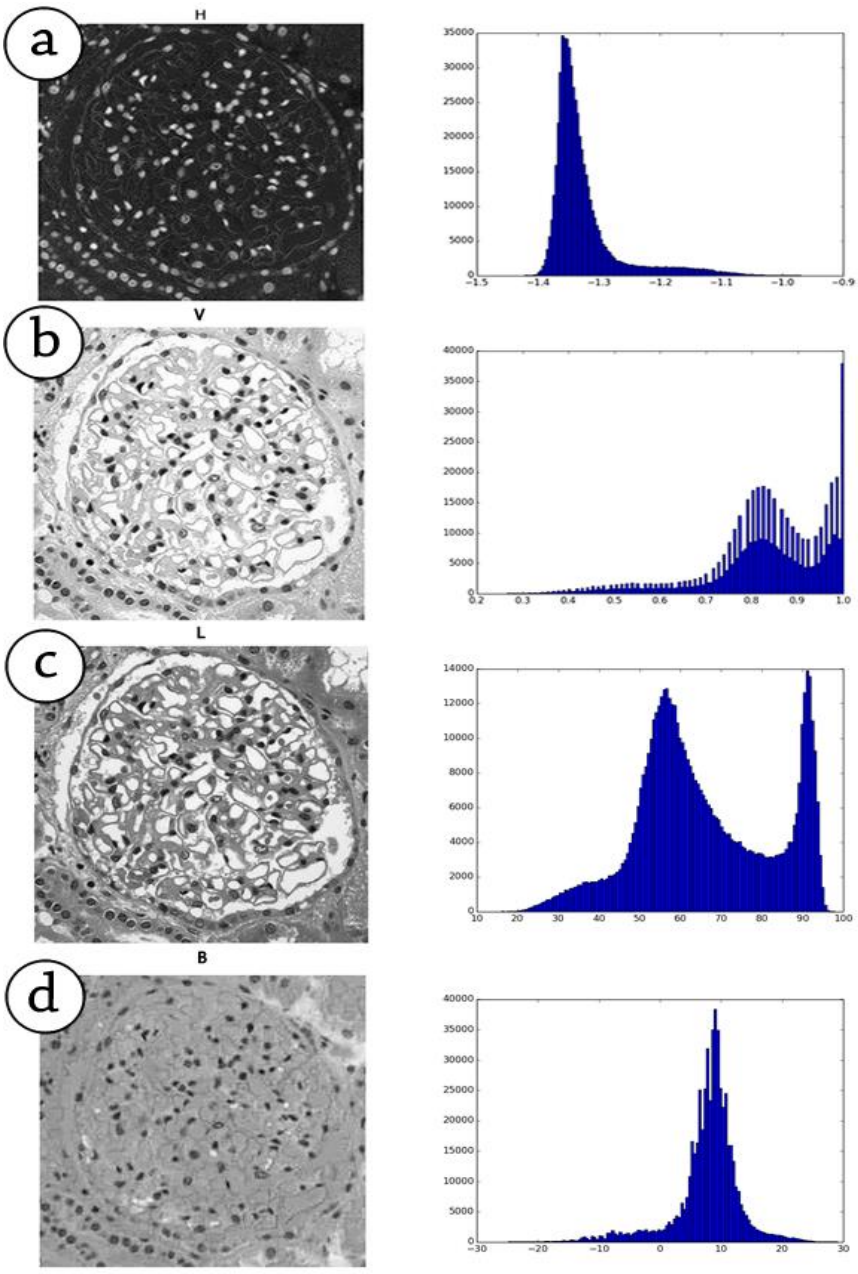

Figura 5.24: Canais com maior contraste entre regiões de núcleos e fundo. Canal H do espaço de cor HED (a), canal V do espaço de cor HSV (b), canais de cor L e B do espaço de cor LAB (c e d respectivamente).

Dos canais mostrados na Figura 5.24, o canal H (do espaço de cor HED) apresentou o maior contraste entre os núcleos e o fundo da imagem. O canal HED é resultado da aplicação do

método deconvolução de cor, implementado por Vander Walt *et al.* [2014], na biblioteca *scikit-image* e proposta originalmente por Ruifrok e Johnston [2001]. O histograma do canal H possui duas classes bem definidas e maior contraste do que os demais canais. Portanto, o canal de cor selecionado para compor o proposta 2 foi o H. Os trabalhos de Schöchlin *et al.* [2014], Wang [2011] e Veillard *et al.* [2013], também utilizam a operação deconvolução de cor para obter representações mais precisas das suas respectivas imagens histológicas. Os bons resultados obtidos com a utilização do canal H (*Hematoxylin*) podem ser explicados justamente pelo fato desse canal armazenar informações referentes às regiões coradas com *Hematoxylin*, que é o corante que interage com os núcleos presentes nas amostras que compõem o conjunto de dados do PathoSpotter.

Após selecionar o canal H**,** as tarefas realizadas no proposta 2 basearam-se no trabalho de Miranda *et al.* [2012], que realiza a segmentação de núcleos em imagens histológicas do colo uterino para posterior classificação de neoplasias. Sendo assim, invertemos a imagem H e aplicamos o método de reconstrução por fechamento, obtendo o resultado que pode ser observado na Figura 5.25.

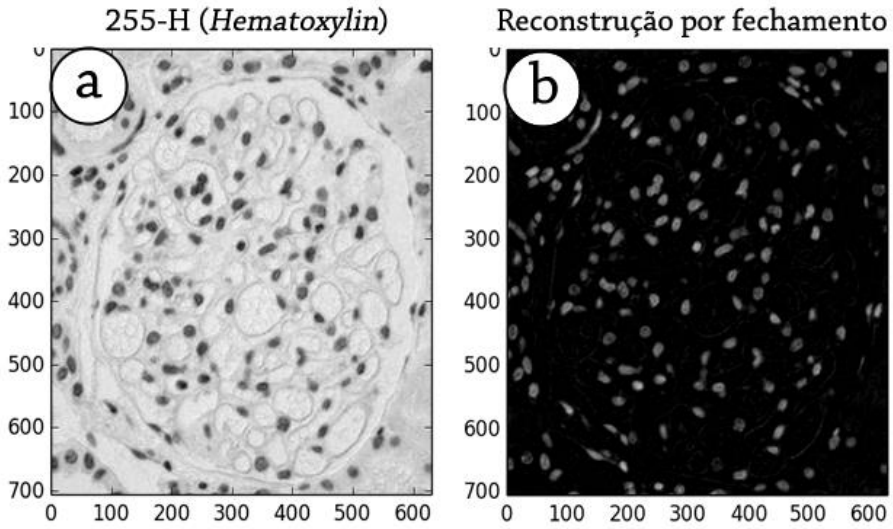

Figura 5.25: Reconstrução por fechamento. Matriz referente ao canal *Hematoxylin* (a) e resultado da reconstrução morfológica por fechamento (b).

A imagem obtida após a aplicação da técnica morfológica de reconstrução por fechamento possui alto contraste e duas classes de intensidade de pixels muito bem definidas. Isso pode ser facilmente observado no histograma da imagem resultante dessa reconstrução morfológica mostrado na Figura 5.26.

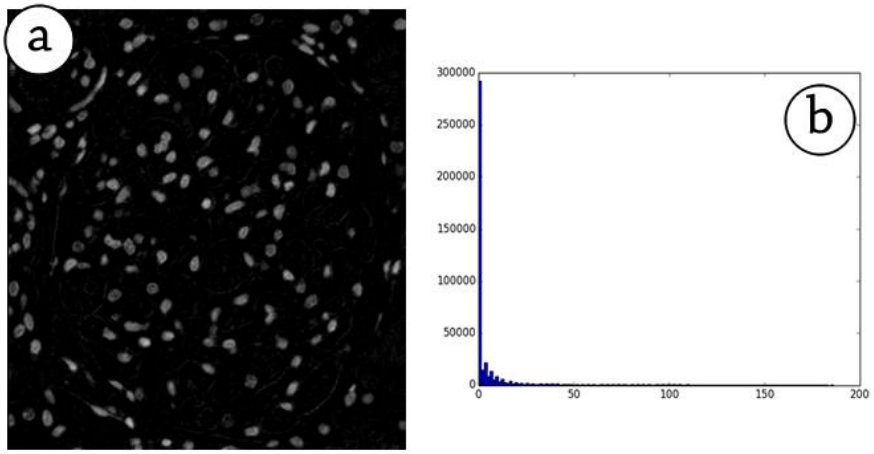

Figura 5.26: Imagem resultante da reconstrução morfológica (a) e o seu histograma (b).

Após a aplicação dessas novas técnicas, como utilização do espaço de cor HED e a reconstrução morfológica, o método de limiarização por Otsu foi mantido para a limiarização automática das imagens, tendo em vista a utilização com sucesso desse método nos trabalhos de Schöchlin *et al.* [2014] e Miranda *et al.* [2012]. Para aumentar ainda mais a qualidade das imagens, após a aplicação do método de Otsu, adicionamos uma nova técnica com propósito de eliminar ruídos oriundos da binarização das imagens.

A técnica utilizada foi a operação morfológica de fechamento (*closing*), inspirada em trabalhos similares com imagens histológicas [Miranda *et al.* 2012; Mathur *et al.* 2013 e Prabusankarlal *et al.* 2015]. A Figura 5.27 ilustra o resultado final do proposta 2 após a aplicação do método de Otsu e a operação morfológica de fechamento (*closing*).

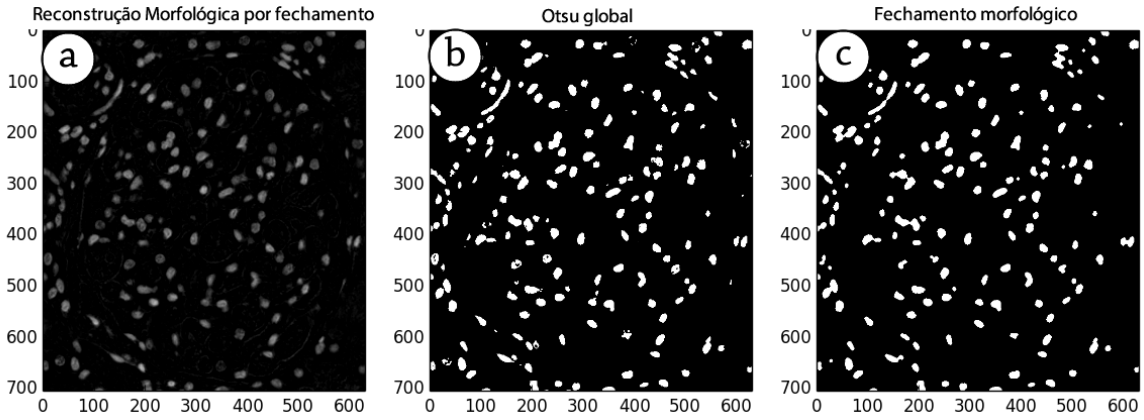

Figura 5.27: Reconstrução morfológica (a), limiarização por Otsu (b) e fechamento morfológico (c).

5.4.1.3 Escolha da proposta de pré-processamento e segmentação

Após a realização da contagem manual e automática do subconjunto de 50 imagens (25 sem glomerulopatia e 25 com glomerulopatia), foi possível calcular a taxa de erro geral na contagem, além das taxas de acerto em relação as amostras sem glomerulopatia e as amostras com glomerulopatia.

Os resultados finais obtidos com a proposta 1 (P1) foram de 14,4% de taxa de erro na contagem de regiões de núcleos presentes nas imagens sem glomerulopatia, 24,8% de taxa de erro na contagem de regiões de núcleos das imagens com glomerulopatia e uma taxa de erro geral de 19,6%.

Os resultados obtidos na proposta 2 (P2) foram superiores ao da proposta 1, apresentando taxas de erro menores, 15,1% de taxa de erro na contagem de regiões de núcleos presentes nas imagens sem glomerulopatia, 16,4% na contagem de regiões de núcleos das imagens com glomerulopatias e uma taxa de erro geral de 15,7%.

Os resultados obtidos com as propostas revelaram-se satisfatórios, tendo em vista a diversidade de características das imagens que compõem o conjunto de dados do PathoSpotter, e a comparação com os resultados observados em diferentes casos de segmentação de núcleos em imagens histológicas, como 7,5% de taxa de erro na segmentação de células com câncer em imagens da mama, 5,1% na segmentação de imagens do intestino grosso ou 13,1% em imagens histológicas em geral [Belsare e Mushirif, 2012; Irshad *et al.* 2015], tendo como único objetivo a segmentação.

A Tabela 5.2 apresenta os resultados obtidos com as duas propostas de pré-processamento e segmentação.

Tabela 5.2: Avaliação das propostas de pré-processamento e segmentação.

Propostas	**Taxa de erro (imagens sem glomerulopatia)**	**Taxa de erro (imagens com glomerulopatia)**	**Taxa de erro geral**
P1	14,4%	24,8%	19,6%
P2	15,1%	16,4%	15,7%

Ao analisar os resultados apresentados nas duas propostas de pré-processamento e segmentação (*pipelines*) concluímos que o P2 foi a proposta que apresentou o melhor

resultado (15,7 de taxa de erro, equivalente a 84,3% de acurácia), portanto, assumimos essa proposta para as etapas de pré-processamento e segmentação do PathoSpotter.

5.4.2 Extração de Características

A etapa de extração de características foi a etapa do trabalho em que investigamos características que pudessem discriminar quantitativamente as imagens com glomerulopatia das imagens sem glomerulopatia. O padrão histológico de glomerulopatias estudado neste trabalho foi o de glomerulopatias proliferativas, as quais, de modo geral, podem ser identificadas através da proliferação de núcleos em um glomérulo.

As características testadas foram: densidade de núcleos presentes na imagem, quantidade de regiões de núcleos (que inclui núcleos isolados e clusters, contados como um único elemento), quantidade de aglomerações (que são os núcleos que estão próximos, mas não necessariamente colados) e, por fim, a distância entre as regiões de núcleos.

Para avaliar a capacidade discriminatória das características, realizamos a regressão logística (ver seção 3.1.5) das amostras em relação a cada característica extraída. De cada imagem que compõe o conjunto de imagens do PathoSpotter, extraímos a característica estudada e realizamos a regressão logística, classificando as imagens de acordo com cada característica em especial.

A estratégia de realizar a regressão logística para cada característica extraída teve o propósito de observar se as características estudadas poderiam ser individualmente úteis na etapa de classificação das imagens. A seguir, nós apresentaremos as análises de cada uma das quatro características avaliadas.

5.4.2.1 Densidade

A ideia inicial de utilização da densidade como característica partiu da hipótese de que se há uma proliferação de núcleos nas imagens com glomerulopatia, a taxa de ocupação de pixels que representem os núcleos deve aumentar. Adicionalmente, Irshad *et al.* [2014], aponta a densidade como uma informação relevante em trabalhos com imagens histológicas, como na detecção de núcleos, por exemplo. A densidade de cada imagem foi extraída pela razão de pixels brancos (regiões dos núcleos) por pixels (pretos) fundos da imagem. A Equação 5.3 descreve a densidade, onde *pxB* é a quantidade de pixels brancos em uma imagem e *pxP* é a quantidade de pixels pretos.

$$densidade = \frac{pxB}{pxP} \tag{5.3}$$

Para analisar se a densidade seria uma informação capaz de caracterizar quantitativamente a diferença entre imagens com glomerulopatias das imagens sem glomerulopatias, observamos a sua distribuição espacial em relação a ambas as amostras. Essa prática é citada por James *et al.* [2013, p.189], e a Figura 5.28 mostra um histograma que revela a distribuição das amostras com glomerulopatia e sem glomerulopatia através do número de amostras com uma determinada densidade.

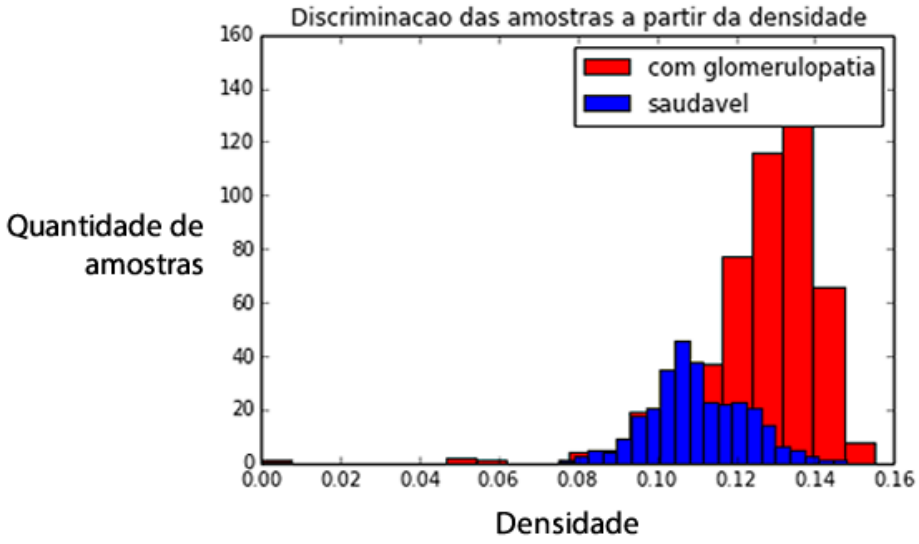

Figura 5.28: Histograma de densidade.

Ao realizar a matriz de confusão do classificador baseado em regressão, obtivemos 76% de precisão, 78% de *recall* e 77% de acurácia (Tabela 5.3).

Tabela 5.3: Matriz de confusão, regressão logística da densidade.

Resultado da Classificação	**Amostras com glomerulopatia (511)**	**Amostras sem glomerulopatia (300)**
Positivo	TP=389	FP=66
Negativo	FN=122	TN=234

5.4.2.2 Distância

Medidas de distância são utilizadas com diferentes finalidades em trabalhos com imagens histológicas, desde operações na etapa de segmentação até como característica discriminante para a classificação. Escolhemos neste trabalho a distância euclidiana, por sua grande aplicação em trabalhos com imagens histológicas, como pode ser constatado nos trabalhos de revisão de Irshad *et al.* [2014] e Lei He *et al.* [2012], ou ainda, no trabalho de Al-Kofahi *et al.* [2010].

Por hipótese, consideramos que imagens com glomerulopatia poderiam ter distância entre as regiões de núcleos menores do que em imagens sem glomerulopatias. No entanto, após os experimentos essa hipótese não se confirmou, pois se percebeu que a mesma não funcionou como um bom discriminante, como pode ser visto no histograma construído em relação à distância euclidiana entre as regiões de núcleos (Figura 5.29). O histograma mostra que amostras com glomerulopatia se apresentam bem sobrepostas às imagens sem glomerulopatia, o que inviabiliza a construção de um classificador a partir desta.

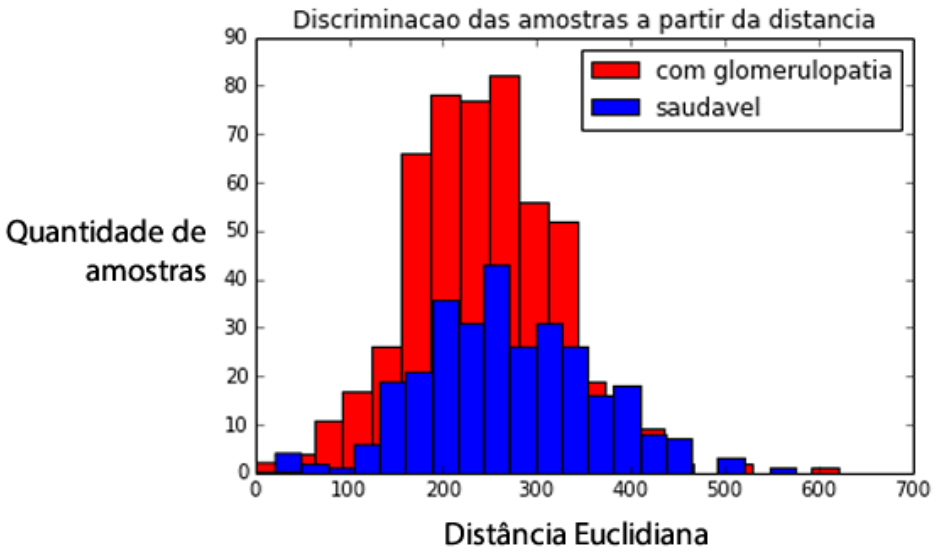

Figura 5.29: Histograma de distância.

Os resultados obtidos a partir da matriz de confusão para essa característica (Tabela 5.4) confirmam que ela não é adequada: 21% de precisão 64% de *recall* e 37% de acurácia.

Tabela 5.4: Matriz de confusão, regressão logística de distância.

Resultado da Classificação	**Amostras com glomerulopatia (511)**	**Amostras sem glomerulopatia (300)**
Positivo	TP=107	FP=108
Negativo	FN=404	TN=192

5.4.2.3 Quantidade de regiões de núcleos

A quantidade de *clusters* é uma característica extraída através da propriedade *regionprops* da biblioteca de processamento de imagens *scikit-image*. A propriedade *regionprops* utiliza o método de crescimento por regiões para identificar regiões da imagem que estejam conectadas por uma mesma intensidade de pixel. Além de identificar regiões, o *regionprops* gera uma lista de informações das áreas identificadas.

No caso do PathoSpotter, a única informação utilizada foi a quantidade de regiões identificadas, a qual nós nomeamos de quantidade de regiões de núcleos. A Figura 5.30 mostra o histograma dessa característica. Nota-se que, apesar de uma área de sobreposição, é possível observar uma boa separação entre as amostras.

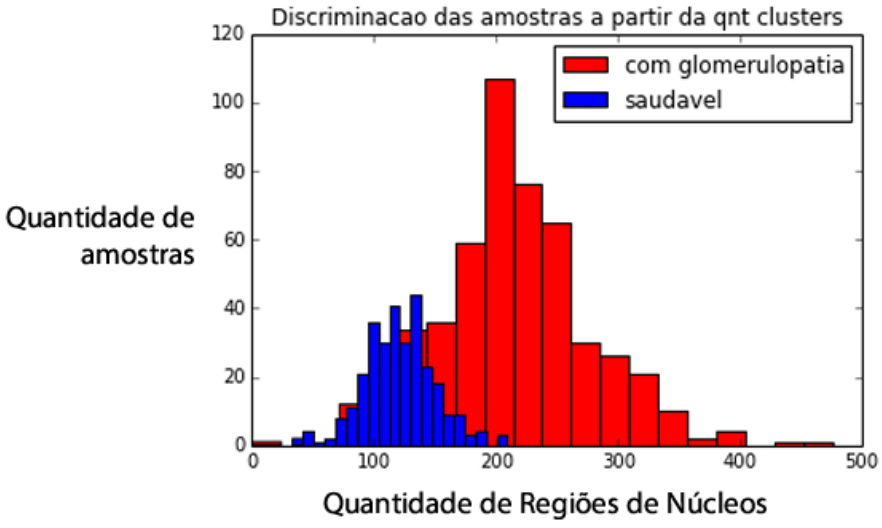

Figura 5.30: Histograma de quantidade de regiões de núcleos.

A matriz de confusão para essa característica (Tabela 5.5) apresentou 87% de precisão, 85% de *recall* e 83% de acurácia.

Tabela 5.5: Matriz de confusão, regressão logística de quantidade de regiões de núcleos

Resultado da Classificação	**Amostras com glomerulopatia (511)**	**Amostras sem glomerulopatia (300)**
Positivo	TP=239	FP=73
Negativo	FN=61	TN=438

5.4.2.4 Quantidade de aglomerações

A quantidade de aglomerações foi uma característica extraída através de um algoritmo de identificação de bolhas, o LoG (*laplacian of gaussian).* Nossa hipótese foi utilizar como discriminante as aglomerações de núcleos que estavam próximos, mas não necessariamente colados, como no caso da quantidade de *clusters*.

O método de LoG foi aplicado através da biblioteca de processamento de imagens *scikit-image* e a Figura 5.31 ilustra um exemplo de sua aplicação.

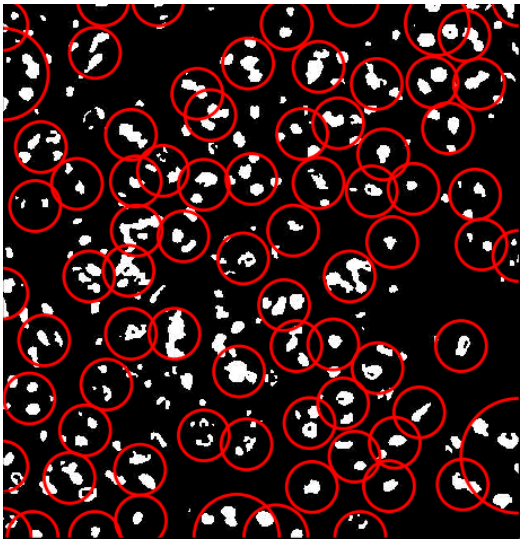

Figura 5.31: Identificação de aglomerações.

A Figura 5.32 mostra o histograma construído a partir da distribuição das amostras em relação a quantidade de aglomerações, no qual se pode observar que há uma razoável sobreposição entre as duas classes.

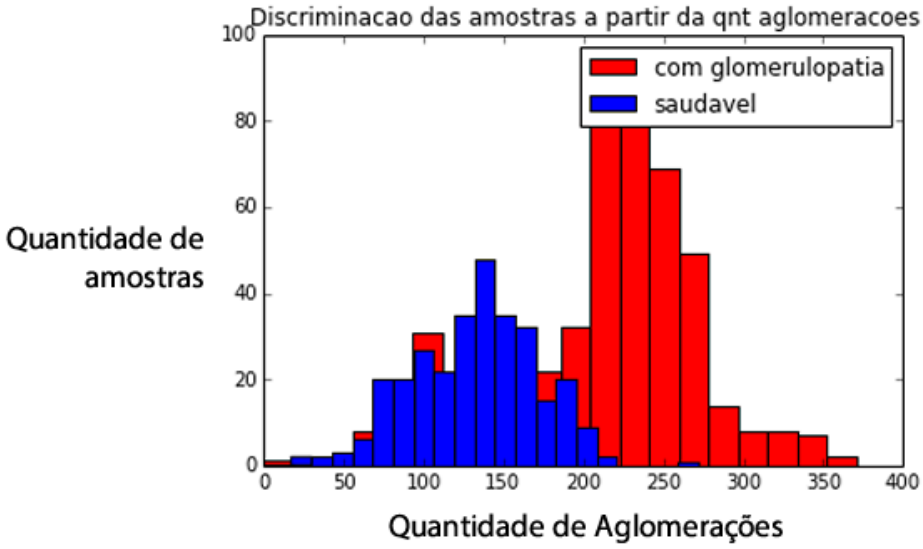

Figura 5.32: Histograma de quantidade de aglomerações.

A matriz de confusão para este caso (Tabela 5.6) apresentou precisão de 78%, *recall* de 81% e acurácia de 73%.

Tabela 5.6: Matriz de confusão, regressão logística de qnt. aglomerações

Resultado da Classificação	**Amostras com glomerulopatia (511)**	**Amostras sem glomerulopatia (300)**
Positivo	TP=414	FP=117
Negativo	FN=97	TN=183

5.4.2.5 Escolha das características

Os resultados obtidos no processo de investigação e avaliação das características indicaram as características densidade, quantidade de *clusters* e quantidade de aglomerações, como características com capacidade discriminatória relevante para utilização em um classificador. Contudo, essa análise também mostrou que a capacidade discriminatória da característica distância foi extremamente baixa, o que nos fez descartar a possibilidade de utilizar essa característica na etapa de classificação. A Tabela 5.7 os resultados obtidos com a regressão logística de todas as umas das quatro características avaliadas.

Tabela 5.7: Resultados obtidos com a regressão logística de cada caractística avaliada

Característica	Acurácia	*Recall*	Precisão
Densidade	77%	78%	76%
Distância	37%	64%	21%
Qnt. Aglomerados	73%	81%	78%
Qnt. Regiões de núcleos	83%	85%	87%

A Figura 5.33 mostra o espaço tridimensional composto pelas três características selecionadas (densidade, quantidade de regiões e quantidade de aglomerados) para compor a matriz de características do PathoSpotter.

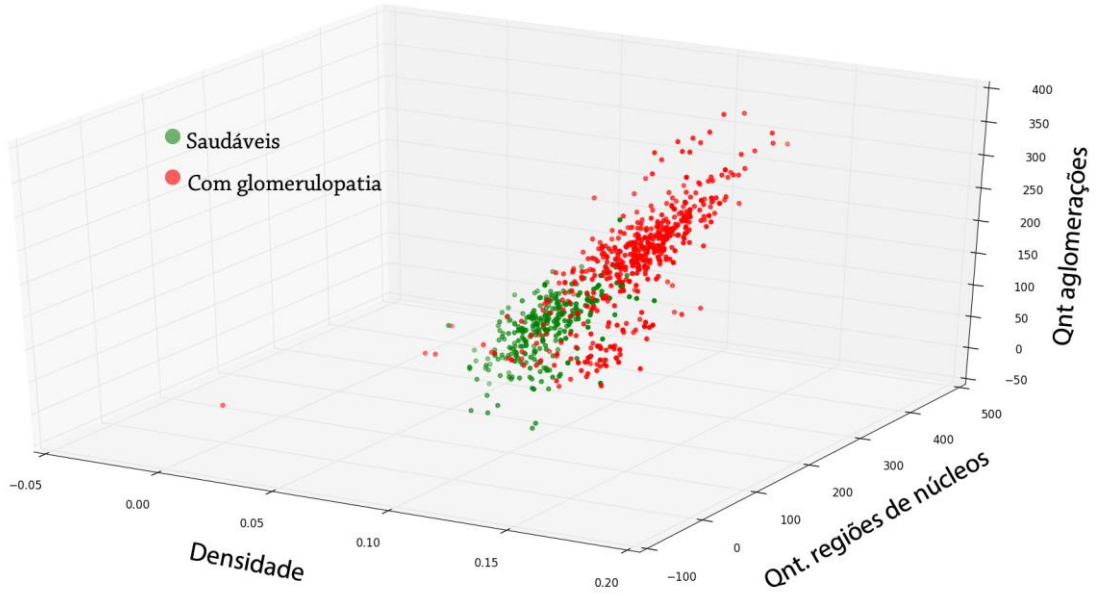

Figura 5.33: Espaço de características formado pelas informações de densidade, quantidade de regiões de núcleos e quantidade de aglomerações (3D).

5.4.3 Classificação e Avaliação

A etapa de classificação foi à última etapa do sistema a ser implementada. Nesta etapa, partimos do mesmo principio adotado em todo o sistema, trabalhando inicialmente com métodos mais simples de classificação e depois testando abordagens mais robustas. As duas abordagens testadas na etapa de classificação do PathoSpotter foram a regressão logística e a kNN.

5.4.3.1 Regressão logística

A classificação das imagens foi realizada através de uma regressão logística aplicada ao espaço tridimensional formado a partir das três características obtidas no processo de extração de características descrito anteriormente (densidade, quantidade de aglomerações e quantidade de regiões de núcleos).

Nessa abordagem todas as amostras foram utilizadas para a busca da função discriminante. Os resultados obtidos com a regressão foram de 89% de precisão, 88% de *recall* e 86% de acurácia. A Tabela 5.8 é a matriz de confusão da classificação realizada através da regressão logística das três características selecionadas.

Tabela 5.8: Matriz de confusão, regressão logística das 3 características.

Resultado da Classificação	**Amostras com glomerulopatia (511)**	**Amostras sem glomerulopatia (300)**
Positivo	TP=451	FP=52
Negativo	FN=60	TN=248

Também realizamos testes com combinação das características extraídas, observando o comportamento da regressão logística em relação aos diferentes espaços de características formados a partir das combinações. A Figura 5.34 (a) mostra a distribuição espacial das características em relação à quantidade de regiões de núcleos e quantidade de aglomerados, além do resultado da regressão logística. Do mesmo modo, a Figura 5.34 (b) realiza a regressão relacionando as amostras à quantidade de aglomerados e densidade e a Figura 5.34 (c) realiza a regressão em relação à quantidade de regiões de núcleos e densidade, realizando assim todas as combinações possíveis com duas características. As retas em cor branca é a função ótima que discrimina as amostras**.**

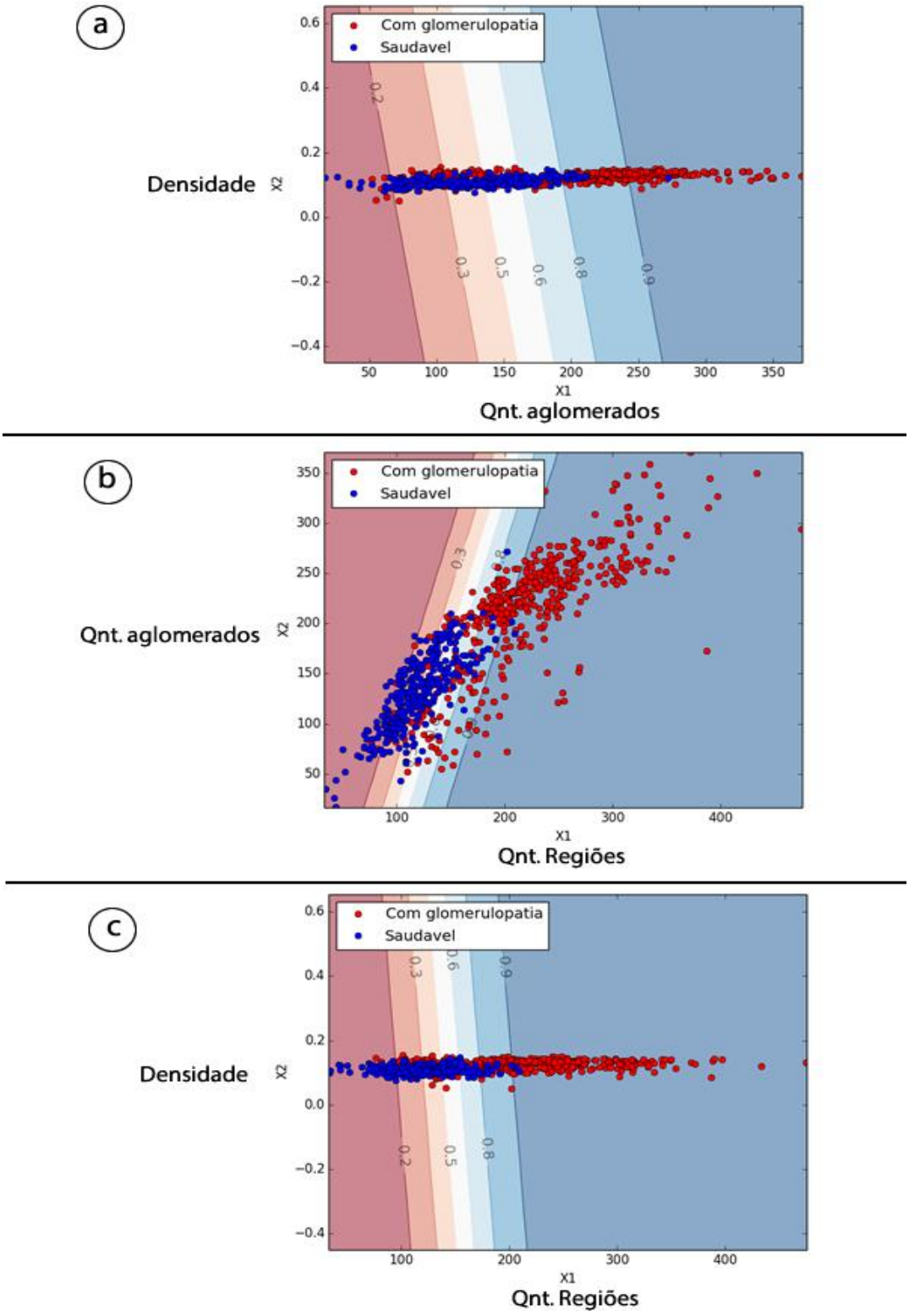

Figura 5.34: Realização da regressão logística em diferentes espaços de características formados pela combinação das características (densidade, quantidade de regiões de núcleos e quantidade de aglomerados).

A Tabela 5.9 apresenta os resultados obtidos com a regressão logística ao utilizar cada uma

das combinações de características apresentadas na Figura 5.34.

Tabela 5.9: Resultados da regressão em diferentes combinações de características.

Combinações de características	**Acurácia**	**Precisão**	***Recall***
Qnt. Regiões x Qnt. Aglomerados	86%	89%	88%
Qnt. Aglomerados X Densidade	73%	78%	80%
Qnt. Regiões X Densidade	83%	87%	88%

5.4.3.2 kNN

Optamos pela utilização do classificador kNN, que é citado na literatura disponível como um dos classificadores mais simples da área de aprendizado de máquina e que apresenta bons resultados em espaços com poucas características, como é o caso do PathoSpotter [Harrington, 2012, p.21]. A implementação do kNN utilizada neste trabalho foi a disponível na biblioteca de aprendizado de máquina *scikit-learn*. A Figura 5.35 mostra o esquema geral do que foi construído para testar essa técnica.

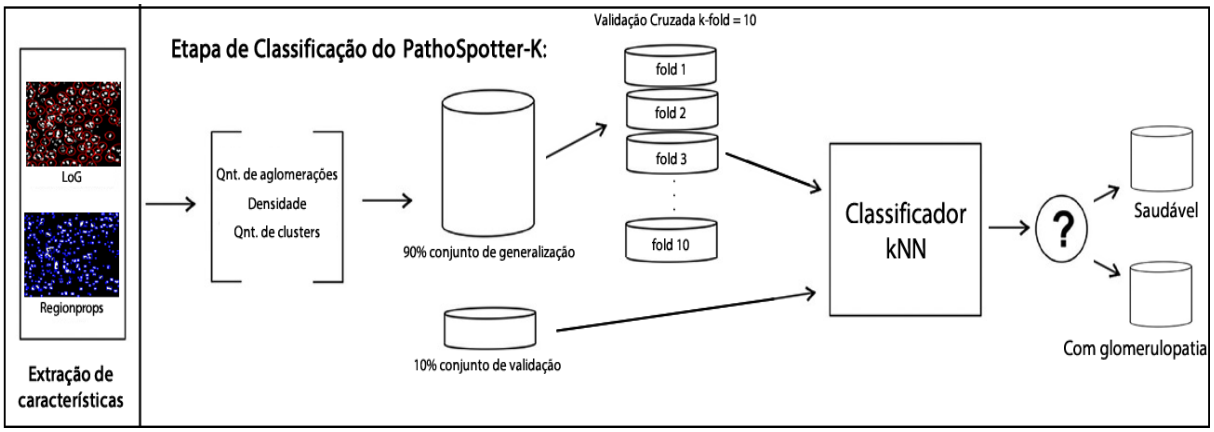

Figura 5.35: Etapa de classificação do PathoSpotter. Estratificação de características e validação dos resultados.

Dividimos o conjunto de dados em subconjuntos de amostras de teste e treinamento. Com o objetivo de realizar uma avaliação confiável e estratificação adequada das amostras do conjunto de dados, realizamos a estratificação das amostras através do método de validação cruzada k-fold estratificada, com *fold* igual a 10, valor sugerido por James *et al.* [2013, p.181] e Japkowicz e Shah [2011, p.18] e também utilizado por Sirinukunwattana *et al.* [2014].

A estratificação do conjunto de dados ocorreu em dois momentos. A primeira estratificação foi realizada apenas com o objetivo de se adquirir um conjunto equivalente a 10% das amostras do conjunto de dados. A esse conjunto demos o nome de conjunto de validação, o qual foi utilizado para realizar um teste final do desempenho do classificador [James *et al.* 2013, p.176]. O conjunto de validação herdou exatamente 81 imagens (10% das amostras do conjunto de dados, aproximadamente), dessas amostras 30 foram de imagens sem glomerulopatias (equivalente a 10% das imagens sem glomerulopatias) e 51 imagens com glomerulopatias (equivalente a 10% das imagens com glomerulopatia), o que tornou a divisão balanceada, evitando a formação de um conjunto amostral desproporcional.

A segunda estratificação ocorreu com o conjunto restante de 90% das amostras do conjunto de dados, o qual foi nomeado de conjunto de generalização, responsável por testar o classificador com diferentes conjuntos de teste e treinamento através da validação cruzada.

Após a criação do conjunto de validação e o conjunto de generalização realizamos a classificação por meio do kNN e a validação cruzada com k-fold igual a 10, no conjunto de generalização. A Figura 5.36 ilustra o processo de testes com a validação cruzada k-fold igual a 10 no conjunto de generalização [James *et al.* 2013, p. 181].

Além do *recall*, precisão e acurácia também calculamos o desvio padrão com o objetivo de

obter a informação da dispersão dos resultados obtidos na validação cruzada, tendo como fonte os resultados específicos obtidos a partir de cada subconjunto de teste e treinamento (fold 1, fold 2, ..., fold 10). Pelo fato de utilizarmos a estratégia de validação cruzada estratificada, os subconjuntos de testes e treinamento (*folds*) extraídos do conjunto de generalização foram divididos de forma balanceada, ou seja, a quantidade de amostras sem glomerulopatias e com glomerulopatias foi igual em todos os *folds*.

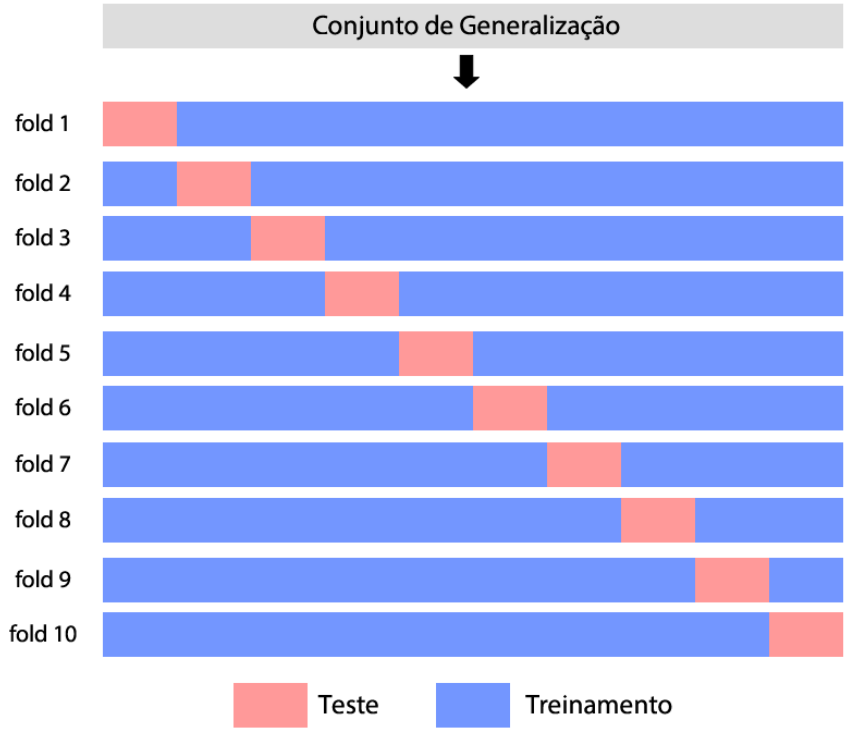

Figura 5.36: Conjunto de generalização e validação cruzada k-fold igual a10.

Antes de obter um resultado final da classificação realizada com o kNN ainda foi necessário escolher o modelo de classificação a ser utilizado por definitivo na classificação. O processo de escolha de um modelo de classificação é uma abordagem realizada com diferentes classificadores ou até mesmo diferentes parâmetros de um único classificador [Japkowicz e Shah, 2011, p.18]. Analisamos três parâmetros com o objetivo de obter o melhor modelo de classificação com o kNN:

- Versão do algoritmo kNN.
- Medida de distância.
- Melhor valor de *k*.

As versões do algoritmo kNN que foram avaliadas na escolha do modelo de classificação

foram *Ball_tree*, *KD_tree* e *Brute Force*. As medidas de distância avaliadas foram, *Euclidean*, *Manhattan e Minkowski*. Por fim, o intervalo testado para escolha de *k* foi de 3 a 80 [Kirk, 2015, p.27]. As versões do algoritmo kNN e o calculo das medidas de distância foram métricas testadas através da biblioteca *scikit-learn*.

A métrica utilizada para o estudo dos parâmetros do classificador e a escolha do modelo de classificação foi a taxa de erro [Kirk, 2015, p.49]. O processo de escolha do modelo de classificação se iniciou através dos testes com as versões do kNN. Em cada uma das três versões do algoritmo kNN, nós variamos as três medidas de distância testadas e em cada uma das três medidas de distância testadas nós avaliamos o *k*, variando-o no intervalo de 3 a 80 (3,4,5,6...80), intervalo de número de características mais 1 até o total de amostras [Kirk, 2015, p.27]. A Figura 5.37 ilustra o processo de avaliação e escolha do modelo de classificação.

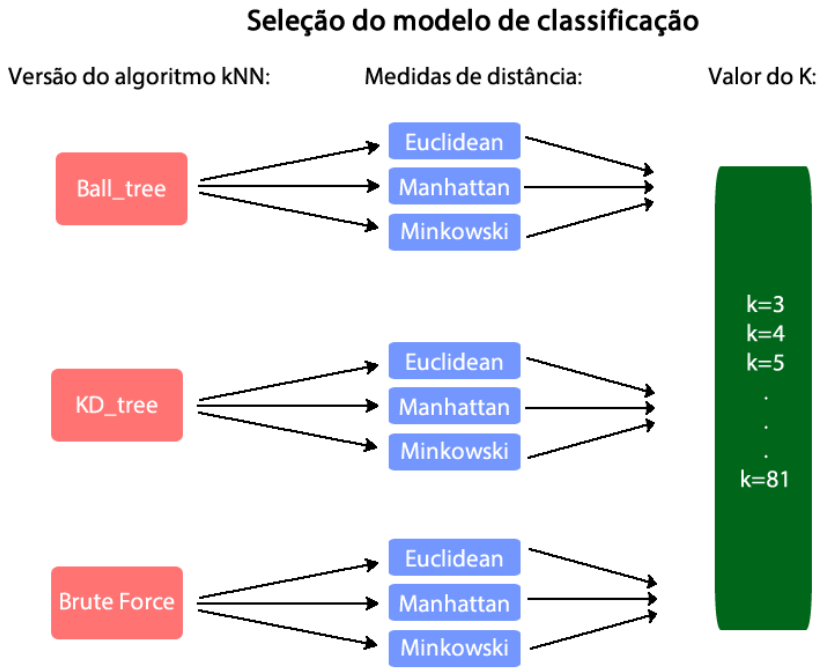

Figura 5.37: Processo de escolha do modelo de classificação através da combinação dos parâmetros.

Após realizar os testes de combinação de todos os parâmetros avaliadados foi possível construir um gráfico para cada versão do algoritmo kNN (*Ball_tree, KD_tree, Brute_force*), variando para cada versão as medidas de distância e o valor *k*.

Observamos que os gráficos obtidos para cada versão do algoritmo kNN foram idênticos. Portanto, não houve variações nas taxas de erro de cada modelo de classificação em relação às diferentes versões do kNN (*Ball_tree, KD_Tree, Brute Force*). A figura 5.38 apresenta um

dos três gráficos contruídos para avaliar as versões do kNN.

Como as diferentes versões do kNN não influenciaram na classificação das amostras, utilizamos o parâmetro *auto* (automático), disponível na biblioteca de aprendizado de máquina do *scikit-learn*, na versão final do modelo de classificação. Esse parâmetro faz com que o método do algoritmo kNN realize uma busca pela melhor versão do kNN a ser utilizada na classificação. Em relação aos demais parâmetros avaliados (medidas de distância e valor de *k*), ocorreram variações na taxa de erro da classificação.

O gráfico apresentado na Figura 5.38 mostra os resultados de taxa de erro da classificação obtidos de acordo com a variação do valor de *k* (3 a 81, variando de 1 em 1) em função das três medidas de distâncias avaliadas (*Manhattan*, *Euclidean* e *Minkowsk*). Cada curva apresentada no gráfico diz respeito a uma medida de distância avaliada em relação aos valores de *k* testados. A curva em azul representa a medida de distância *Manhatan*, a curva em verde representa a medida de distância *Euclidean* e a curva vermelha representa a medida de distância *Minkowsk*. O ponto em destaque (em amarelo e preto) indica a melhor combinação (menor taxa de erro) de medida de distância e valor de *k*.

Figura 5.38: Gráfico de resultados de taxa de erro para cada valor de *k*. Cada curva diz respeito as medidas de distância avaliadas.

A Tabela 5.10 mostra os menores resultados de taxa de erro encontrados para cada medida de distância avaliada e seu respectivo melhor valor de *k*.

Tabela 5.10: Menores taxas de erro obtidas a partir da combinação de medidas de distância e valor de *k*.

Medida de Distância	Valor de *k*	Taxa de Erro
Manhatan	11	11,7%
Euclidean	15	12,5%
Minkowsk	11	12,6%

A partir dos resultados apresentados na Tabela 5.10 é possível constatar que a melhor combinação de medida de distância e valor de *k* encontrados na avaliação e escolha do modelo do kNN são: Medida de distância *Manhatan* e valor de *k* igual a 11. Portanto, assumimos estes parâmetros no modelo de classificação com o kNN.

Após a escolha do melhor modelo de classificação tornou-se possível observar os resultados finais obtidos com o classificador kNN.

Os resultados finais obtidos com o conjunto de generalização foram de 88,3% de acurácia, 88,0% de *recall* e 92,3% de precisão, com desvio padrão de 3,6.

Após a realização dos testes com validação cruzada, aplicamos o kNN ao conjunto de validação final (que havia sido separado no início da etapa de classificação em uma primeira estratificação). Nesse teste o conjunto de validação final foi utilizado como conjunto de teste e o conjunto que havia sido separado como conjunto de generalização foi utilizado como conjunto de treinamento. Os resultados obtidos nesse teste final foram de 85% de acurácia, 88% de recall e 88% de precisão.

O valor do desvio padrão calculado a partir dos resultados de acurácia de cada subconjunto de generalização (3,6), revela que a dispersão do resultado de acurácia final (de 88,3%, obtido pela média dos 10 *folds*) pode variar entre 84,7% e 91,9% (88,3 ±3,6). Esse resultado é validado, justamente, através do conjunto de validação final, que foi previamente isolado e não participou em nenhum momento do treinamento das amostras, este conjunto foi criado exatamente com este propósito, validar o resultado final com um conjunto isolado.

O resultado de 85% de acurácia obtido na validação final confirma e valida o resultado final da classificação, pois se encontra dentro do intervalo de dispersão calculado através da acurácia final e desvio padrão.

Através do conjunto de validação final e generalização, tornou-se possível avaliar a classificação de uma maneira confiável, mostrando a capacidade de generalização do modelo

de classificação e por fim validando o resultado final com um conjunto neutro, o qual não havia sido utilizado anteriormente, em nenhum momento, como conjunto de treinamento do sistema.

Enfim, os resultados finais, apresentados na Tabela 5.11 mostra a capacidade do PathoSpotter de classificar imagens histológicas renais de glomérulos, como sem glomerulopatia ou com glomerulopatia (proliferativa).

Tabela 5.11: Resultado final do PathoSpotter.

Quantidade de amostras	**Precisão %**	***Recall* %**	**Acurácia %**
811	92,8	88,0	88,3

No próximo capítulo nós apresentamos as considerações finais do trabalho e discutimos o futuro do PathoSpotter.

Capítulo 6

Considerações Finais

"Depois de escalar uma montanha muito alta, descobrimos que há muitas outras montanhas por escalar."

-- Nelson Mandela

Neste capítulo nós discutimos e avaliamos os resultados obtidos com o PathoSpotter–K, comparando-os a outros trabalhos de natureza similar. Adicionalmente, apontamos os trabalhos futuros, relacionados diretamente e indiretamente ao PathoSpotter.

6.1 Conclusão

Este trabalho se propôs a desenvolver um sistema computacional para auxílio ao diagnóstico de glomerulopatias a partir da análise de imagens digitais. Como resultado, a partir de um conjunto de dados (*dataset*) construímos o sistema PathoSpotter, que classifica imagens histológicas de glomérulos renais como sem glomerulopatia ou com glomerulopatia. A versão do sistema apresentada neste trabalho analisou as glomerulopatias proliferativas (um entre os diferentes padrões histológicos de glomerulopatias).

Observou-se que há um grande número de trabalhos publicados sobre sistemas computacionais de apoio ao diagnóstico para diferentes patologias, surgindo inclusive um campo de estudo chamado de Patologia Digital. No entanto, também se constatou uma enorme carência de trabalhos com propostas de sistemas de apoio com o foco nas glomerulopatias. Isso faz com que o PathoSpotter, até o momento em que esse texto foi escrito, seja o primeiro sistema dedicado a este fim.

O PathoSpotter atingiu 88,4% de acurácia, resultado considerado bom pelos médicos patologistas que o validaram, além de ter sido compatível com os resultados encontrados em sistemas similares de Patologia Digital. Os trabalhos de Kothari *et al.* [2013], Miranda *et al.* [2012] e Sirinukunwattana *et al.* [2014], utilizando um número menor de amostras, mais características (aumenta a possibilidade de realizar a classificação de maneira eficaz) e em

alguns casos, métodos mais sofisticados, apresentaram acurácia menor do que no PathoSpotter, de 77%, 73% e 86%, respectivamente. Adicionalmente, o PathoSpotter ainda conseguiu resultados inferiores aos de Schöchlin *et al.* [2014] e Mathur *et al.* [2013] (88,9% e 92% de acurácia, respectivamente), que tratavam de problemas mais simples de caracterizar, pelo fato de partir de marcadores (características) pré-estabelecidas por especialistas.

Os resultados obtidos neste trabalho apontaram o PathoSpotter como uma ferramenta promissora para o apoio de diagnóstico das glomerulopatias proliferativas, podendo ser também uma ferramenta útil para o treinamento e auxílio de estudantes em medicina ou patologistas não familiarizados com essa patologia em especial. O sistema pode funcionar como uma verificação preliminar da presença de glomerulopatias em imagens de glomérulos renais, até a realização de um diagnóstico mais acurado de um especialista.

O PathoSpotter se valeu de métodos computacionais robustos e consagrados na área de reconhecimento de padrões e processamento de imagens, sendo, no momento em que este texto está sendo escrito, o único sistema especializado em identificação de glomerulopatias proliferativas. Todas as imagens usadas no desenvolvimento do sistema, bem como os códigos fontes estão disponíveis na página do grupo de pesquisa PathoSpotter, que pode ser acessada em: http://www.pathospotter.uefs.br.

6.2 Trabalhos Futuros

O PathoSpotter pode ser considerado uma inovação na Patologia Digital, já que os sistemas similares estão voltados principalmente ao estudo de neoplasias. Sendo assim, as possibilidades de trabalhos futuros são vastas.

Entre eles, destacam-se a melhoria da qualidade de predição do sistema e a implantação da capacidade de predizer outras patologias renais. Também serão trabalhadas as patologias hepáticas, fazendo com que o sistema evolua para a versão PathoSpotter-KL (*Kidney and Liver*).

Em relação ao aperfeiçoamento do sistema, serão investigadas novas características e testados novos métodos de extração e classificação das imagens. Faremos a ampliação do conjunto de dados, incluindo mais imagens de aspecto normal, além de incluir as imagens para as novas patologias que serão trabalhadas, como as glomerulopatias membranosas, membranoproliferativas e esclerótica. Também será criado o conjunto de

dados específico para o trabalho com patologias hepáticas.

Por fim, a página da internet do grupo PathoSpotter será melhorada para armazenar todas as informações sobre o desenvolvimento do sistema, para facilitar seu uso tanto por estudantes de computação quanto por estudantes de medicina. Estuda-se a possibilidade de viabilizar o uso do PathoSpotter como um *Web Service*, permitindo a exploração de seus recursos, e também auxiliando sua melhora através das críticas de um grande número de usuários.

Referências Bibliográficas

[Bougouma *et al.* 2013] Bougouma, M. *et al.*, 2012. Growth and characterization of large, high quality MoSe2 single crystals. *Journal of Crystal Growth. v. 363*, p.122–127, Elsevier, 2013.

[Meyer e Camargo Neto, 2008] Meyer G.E.; Camargo Neto, J. Verification of color vegetation indices for automated crop imaging applications. *Computers and Electronics in Agriculture, 63(2)*: 282293, October 2008.

[Danuser, 2011] Danuser, G. Computer Vision In Cell Biology. *Cell 147*, Elsevier Inc, November 23, 2011.

[Młynarczuk *et al.* 2013] Młynarczuk, M.; Górszczyk, A.; Ślipek, B. The application of pattern recognition in the automatic classification of microscopic rock images. *Computers & Geosciences. v. 60, 126–133*. Elsevier. 2013.

[Karacor *et al.* 2011] Karacor, A. G.; Torun, R.; Abay, S. Aircraft classification using image processing techniques and artificial neural networks. *International Journal of Pattern Recognition and Artificial Intelligence. vol. 25, no. 8*, 1321_1335, 2011.

[Shahzad *et al.* 2015] Shahzad, N.; *et al*. Evaluation of state and community/private forests in Punjab, Pakistan using geospatial data and related techniques, *Forest Ecosystems2:7*, 2015.

[Baravalle *et al.* 2015] Baravalle, R. G.; *et al*. Multifractal characterisation and classification of bread crumb digital images, *EURASIP Journal on Image and Video Processing, 2015:9*, 2015.

[Ritter *et al.* 2011] Ritter, F.; Boskamp, T.; Homeyer, A.; Laue, H.; Schwier, M.; Link, F.; Peitgen, H. Medical Image Analysis. *60 IEEE. PULSE*. November/December 2011.

[Irshad *et al.* 2014] Irshad, H. *et al.*, Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review—Current Status and Future Potential. *Ieee Reviews In Biomedical Engineering, Vol. 7*, 2014.

[Lei He *et al.* 2012] Lei He, L.; et al. Histology image analysis for carcinoma detection and grading. *Computer methods and programs in biomedicine. v 107* p. 538–556. 2012.

[Belasare and Mushirif, 2012]. Belsare, A. D.; Mushrif, M. M. Histopathological image analysis using image processing techniques: an overview. *Signal & image processing : an international journal (SIPIJ) vol.3*. 2012.

[May, 2010] M. May, "A better lens on disease: Computerized pathology slides may help doctors make faster and more accurate diagnoses," *IEEE. PULSE. November/December 2011. Sci. Amer., vol. 302*, pp. 74–77, 2010.

[Meijering, 2012]. Meijering, E. Cell Segmentation: 50 Years Down the Road. *IEEE Signal Processing Magazine, vol. 29*, no. 5, September 2012, pp. 140–145

[Gurcan *et al.* 2009] Gurcan, M. N. Histopathological Image Analysis: A Review. *IEEE Rev Biomed Eng. 2009* ; 2: 147–171. doi:10.1109/RBME.2009.2034865.

[Cohen e Glassok, 1999] Cohen, A. H. and Glassock, R.J. The Schrier Atlas of Diseases of the Kidney. Disponível em: http://www.cybernephrology.ualberta.ca/cn/Schrier/Volume2/chapt2/ADK2_02_1-3.pdf. Acessado 18 de DEZ. de 2015.

[WHO, 2004] World Health Organization. The global burden of disease. Disponível em: http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf. Acesso em: 15 de DEZ. de 2015.

[Polito *et al.* 2010] Polito MG, De Moura LAR, Kirsztajn GM. An overview on frequency of renal biopsy diagnosis in Brazil: clinical and pathological patterns based on 9,617 native kidney biopsies. *Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association - European Renal Association.* 2010 Feb;25(2):490–6.

[Woo *et al.* 2010] Woo K-T, Chan C-M, Mooi CY, -L-Choong H, Tan H-K, Foo M, et al. The changing pattern of primary glomerulonephritis in Singapore and other countries over the past 3 decades. *Clinical nephrology. 2010 Nov;74*(5):372–83.

[D'agati, 2004] D'agati, V. D. Pathologic Classification of Focal Segmental Glomerulosclerosis: A Working Proposal. *American Journal of Kidney Diseases, Vol 43*, No 2 (February): p. 368-382, 2004.

[Weening *et al.* 2004] Weening, J. J.; et al. The classification of glomerulonephritis in systemic lupus erythematosus revisited. *Kidney int 65,* 521-530. 2004.

[Chabat *et al.* 2000]Chabat, F.; Hansell, D. M.; Yang, G. Visual Information Processing. *IEEE Engineering in Medicine and Biology. September/October*, 2000.

[Deserno *et al.* 2013]. Deserno T. M.; et al. Viewpoints on Medical Image Processing: *From Science to Application. Current Medical Imaging Reviews, v. 9*, p. 79-88, 2013.

[Mas *et al.* 2015]. Mas, D.; *et al*. Novel image processing approach to detect malaria. *Optics Communications 350(2015)*13–18. 2015.

[Prabusankarlal *et al.* 2015]. Prabusankarlal, K. M.; *et al*. Assessment of combined textural and morphological features for diagnosis of breast masses in ultrasound. *Human-centric Computing and Information Sciences 5:12*, 2015.

[Cheng e Mandal, 2015] Cheng, L. E Mandal, M. Automated analysis and diagnosis of skin melanoma on whole slide histopathological images. *Pattern Recognition 48(2015)*, 2738–2750, 2015.

[Tolles, 1955] W. E. Tolles, "The Cytoanalyzer — An example of physics in medical research", *Transactions of the New York Academy of Sciences, vol. 17*, no. 3, pp. 250–256, January 1955.

[Ruifrok and Johnston] Ruifrok AC, Johnston DA (2001). Quantification of histochemical staining by color deconvolution. Analytical and Quantitative Cytology and Histology 23:291–299.

[Veillard *et al.* 2013] Veillard et al. Cell nuclei extraction from breast cancer histopathologyimages using colour, texture, scale and shape information. *Diagnostic Pathology 2013, 8 (Supp1)*: S5. http://www.diagnosticpathology.org/content/8/S1/S5.

[Wang, 2011]. Wang, C. W. Robust Automated Tumour Segmentation on Histological and Immunohistochemical Tissue Images. *PLoS ONE, 6(2):* e15818. 2011.

[Van der Walt *et al.* 2014]. Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu and the scikit-image contributors. scikit-image: Image processing in Python. *PeerJ 2:e453 (2014)* http://dx.doi.org/10.7717/peerj.453.

[Gavrilovic *et al.* 2013]. Gavrilovic et al., 2013. Blind Color Decomposition of Histological Images. *Ieee Transactions On Medical Imaging, Vol. 32, No. 6*, June 2013.

[Zarella *et al.* 2015] Zarella MD, Breen DE, Plagov A, Garcia FU. An optimized color transformation for the analysis of digital images of hematoxylin & eosin stained slides. *J Pathol Inform 2015*;6:33.

[Mouelhi *et al.* 2013] Mouelhi A, *et al.* A new automatic image analysis method for assessing estrogen receptors' status in breast tissue specimens. *Computers in Biology and Medicine 43 (2013)* 2263–2277.

[Tabesh *et al.* 2007] Tabesh A, *et al*. Multifeature Prostate Cancer Diagnosis and Gleason Grading of Histological Images. *Ieee Transactions On Medical Imaging, Vol. 26*, No. 10, October 2007.

[Sharma *et al.* 2012]. Sharma, H.; *et al*. Determining similarity in histological images using graph-theoretic description and matching methods for content-based image retrieval in medical diagnostics. *Diagnostic Pathology, 7:134*. 2012.

[Miranda *et al.* 2012] Miranda, G. H. B. *et al.* Structural Analysis of Histological Images to Aid Diagnosis of Cervical Cancer. In: *Graphics, Patterns and Images(SIBGRAPI), Conference on - 2012 – IEEE Conference, Ouro Preto, Minas Gerais, Brasil. V. 25*, P. 316 – 323, 2012.

[Mathur *et al.* 2013] Mathur A, Tripathi AS, Kuse M. Scalable system for classification of white blood cells from Leishman stained blood stain images. *J Pathol Inform* 2013;4:15.

[Sirinukunwattana *et al.* 2014] Cell words: Modelling the visual appearance of cellsin histopathology images. *Computerized Medical Imaging and Graphics 42 (2015) 16–24.*

[Schöchlin *et al.* 2014]. Schöchlin M, Weissinger SE, Brandes AR, Herrmann M, Möller P, Lennerz JK. A nuclear circularity-based classifier for diagnostic distinction of desmoplastic from spindle cell melanoma in digitized histological images. *J Pathol Inform* 2014;5:40.

[Isitor e Thorne, 2007]. Isitor G. N.; Thorne R. Comparison between nuclear chromatin patterns of digitalized images of cells of the mammalian testicular and renal tissues: an imaging segmentation study. *Comput Med Imaging Graph. Mar;31(2):63-70*, 2007.

[Kothari *et al.* 2011]. Kothari, S.; *et al*. Histological image classification using biologically interpretable shape-based features. *BMC Medical Imaging. v. 13*, p.13:9, 2013.

[Stewart *et al.* 2014]. Stewart, S; *et al*. Distinguishing between renal oncocytoma and chromophobe renal cell carcinoma using Raman molecular imaging. *J. Raman Spectrosc. 45, 274–280*, 2014.

[Tae-Yun Kim *et al.* 2014] Kim, T. Y.; *et al*. 3D Texture Analysis in Renal Cell Carcinoma

Tissue Image Grading. *Computational and Mathematical Methods in Medicine. Article ID 536217*, 12 pages, 2014.

[Cui *et al.* 2012] Cui, L *et al*. CT imaging and histopathological features of renal epithelioid angiomyolipomas. *Clinical Radiology 67 (2012)* e77ee82

[Herrmann *et al.* 2012]. Herrmann, A., Tozzo, E., Funk J. Semi-automated quantitative image analysis of podocyte desmin immunoreactivity as a sensitive marker for acute glomerular damage in the rat puromycin aminonucleoside nephrosis (PAN) model. *Experimental and Toxicologic Pathology 64* (2012) 45– 49.

[Guyton e Hall, 2006] Guyton A.C. and Hall J.E. *Tratado de fisiologia médica. 11ª edição.* Sanders Elsevier. Rio de Janeiro, 2006. P. 309.

[Barros *et al.* 2006]Barros RT, Alves MAR, Kirztajn GM, Sens YAS, Dantas M. *Glomerulopatias: patogenia, clínica e tratamento.* 2ed. São Paulo: Editora Sarvier, 2006.

[McGrogan *et al.* 2011] McGrogan A, Franssen CF, de Vries CS. The incidence of primary glomerulonephritis worldwide: *a systematic review of the literature. Nephrol Dial Transplant.* 2011 Feb;26(2):414-30.

[Castro *et al.* 2002] Castro YAM, Núñez LG, Barry HG, Guerrero MA, González LMG. Estudio clinicopatológico de las glomerulopatías primarias. *Rev Cubana .Med.* 2002; 41(6).

[Queiroz *et al.* 2009] Queiroz MM, Júnior GBS, Lopes MSR, Nogueira JOL, Correia JW, Jerônimo ALC, et al. Estudo das Doenças Glomerulares em Pacientes Internados no Hospital Geral César Cals - Fortaleza, Ceará, Brasil. *J. Bras. Nefrol. 2009*;31(1):6-9

[Bahiense-Oliveira e Malafronte, 2006] Bahiense-Oliveira M, Malafronte P. *Epidemiologia das glomerulopatias*. In: Barros RT, Alves MAR, Dantas M, Kirsztajn GM, Sens YAS, editores. *Glomerulopatias: patogenia, clínica e tratamento. 2. ed*. São Paulo: Sarvier; 2006, p. 55-63.

[Ferrazi *et al.* 2010]Ferrazi FHRP, Martinsii CGB, Cavalcantiii JC, Oliveiraii FL, Quirinoi RM, Chiconi R, et al. Perfil das doenças glomerulares em um hospital público do Distrito Federal. *J. Bras. Nefrol. 2010*;32(3):249-256

[Al Kofahi *et al.* 2010]Y. Al Kofahi; et al. Improved Automatic Detection And Segmentation Of Cell Nuclei In Histopathology Images. *Ieee Transactions On Biomedical Engineering, Vol. 57*, No. 4, April 2010.

[Lopes *et al.* 2001]Lopes AA, Silveira MA, Martinelli RP, Rocha H. Associação entre raça e .incidência de doença renal terminalsecundária a glomerulonefrite: .influência do tipo histológico e da presença de hipertensão arterial. *Rev .Ass Med Brasil. 2001; 47(1):*78-84.

[Nixon e Aguado, 2008] Nixon, M.; Aguado, A. *Feature Extraction e Image Processing. 2. ed*. Elsevier. 2008.

[Alves Júnior *et al*. 2008] Alves Júnior JM; Pantoja RKS; Barros CV. Estudo Clínicopatológico Das Glomerulopatias No Hospital De Clínicas Gaspar Vianna. *Revista Paraense de Medicina V.22 (1)* janeiro a março 2008.

[Ganzalez e Woods, 2006] Gonzalez, R.C. e Woods, R.E. e Eddins, S.L., *Digital Image Processing using MATLAB*, Pearson, 2006.

[Pedrini e Schwartz, 2008]Pedrini, H.; Schwartz, W. R. *Análise de Imagens Digitais: Princípios, algoritmos e aplicações*. São Paulo: Thomson Learning, 2008.

[Burger e Burge, 2009] W. Burger, M.J. Burge, Principles of Digital Image Processing, *Undergraduate Topics in Computer Science,* DOI 10.1007/978-1-84800-191-6_5, Springer-Verlag London © Limited, 2009.

[Burger e Burge, 2009] W. Burger, M.J. Burge, Principles of Digital Image Processing, *Fundamental Techniques,* DOI 10.1007/978-1-84800-191-6_5, Springer Verlag London © Limited, 2009.

[Peixoto *et al.* 2015]. Peixoto, Fm; Reboucas, Es; Xavier, Fgl And Reboucas Filho, P P. Desenvolvimento de um Software para cálculo da densidade de nódulos de grafita em ferro fundido nodular através de Processamento Digital de Imagens. *Matéria (Rio J.) [online]. 2015, vol.20, n.1*, pp. 262-272. ISSN 1517-7076.

[Lindeberg, 1993]. Lindeberg (1993)``Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention", *International Journal of Computer Vision, vol. 11,* pp. 283--318, Dec. 1993. (PostScript 1.6Mb) (PDF 787kb).

[James *et al.* 2013] G. James et al., An Introduction to Statistical Learning: with Applications in R, *Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7* 5, © Springer Science+Business Media New York 2013.

[Solomon e Breckon, 2011] Solomon, C and Breckon T.; *Fundamentals of Digital Image Processing, A Pratical Approach With Examples in Matlab*. John Wiley & Sons Ltd. ISBN 978 0 470 84472 4. 2011

[Petrou e Petrou, 2010] M. Petrou and C. Petrou; *Image Processing: The Fundamentals*, Second Edition © 2010 John Wiley & Sons, Ltd. ISBN: 978-0-470-74586-1. 2010.

[Dean, 2014] J. Dean; *Big Data, Data Mining and Machine Learning: Value Creation for Business Leaders and Practitioners*. Published by John Wiley & Sons, Inc., Hoboken, New Jersey, 2014.

[Kirk, 2015] M. Kirk; *Thoughtful Machine Learning*. O'Reilly Media. p. 236. Ebook ISBN: 978-1-4493-7405-1. 2014.

[Ham e Kamber, 2006] J. Ham, M Kamber. *Data Mining: Concepts and Techniques.* Second Edition. Morgan Kaufmann Publishers is an imprint of Elsevier. 500 Sansome Street, Suite 400, San Francisco, CA 94111.

[Japkowicz e Shah, 2011] N. Japkowicz and M. Shah. Evaluating Learning Algorithms. A Classification Perspective. ISBN 978-0-521-19600-0 Cambridge University Press 2011.

[Harrington, 2012] P. Harrington; *Machine Learning in Action*. ISBN 9781617290183. ©2012 by Manning Publications Co. All rights reserved. 2012.

[Jalalah, 2009] Jalalah SM; Patterns of Primary Glomerular diseases among Adults in the Western Region of Saudi Arabia. *Saudi Journal of Kidney Diseases and Transplantation*. 2009; 20 (2):295-299

[Sonka *et al.* 2006] M. Sonka, V. Hlavac, R. Boyle; *Image Processing, Analysis, and Machine Vision*. Third edition. ISBN-10: 1133593607. p.912. CL Engineering; 2007.

[Conway e White, 2012] D. Conway and JM White, *Machine Learning for Hackers*. ISBN:

978-1-449-30371-6. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. 2012.

[Dougherty, 2009] G. Dougherty; *Digital Image Processing for Medical Applications*. Cambridge University Press The Edinburgh Building, Cambridge CB2 8RU, UK. ISBN-13 Ebook: 978-0-511-53343-3.(2009).

[Davies, 2012] ER Davies; *Computer and Machine Vision: Theory, Algorithms,Practicalities*, Fourth edition . Academic Press is an imprint of Elsevier 225 Wyman Street, Waltham, 02451, USA. The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK. ISBN: 978-0-12-386908-1 (2012).

[Preim e Botha, 2014] B. Preim and C. Botha; *Visual Computing For Medicine Theory, Algorithms, And Applications Second Edition*. Morgan Kaufmann is an imprint of Elsevier 225 Wyman Street, Waltham, MA, 02451, USA © 2014 Elsevier Inc. All rights reserved. ISBN: 978-0-12-415873-3 (2014)