UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DETECÇÃO DE CÂNCER DE PRÓSTATA EM IMAGENS DE MICROSCOPIA UTILIZANDO GRAFOS DE CONTEXTO GLANDULAR

RODRIGO DE PAULA MENDES

ORIENTADOR: PROF. DR. CESAR HENRIQUE COMIN

São Carlos – SP

Abril/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DETECÇÃO DE CÂNCER DE PRÓSTATA EM IMAGENS DE MICROSCOPIA UTILIZANDO GRAFOS DE CONTEXTO GLANDULAR

RODRIGO DE PAULA MENDES

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Processamento de Imagens e Sinais
Orientador: Prof. Dr. Cesar Henrique Comin

São Carlos – SP

Abril/2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Rodrigo de Paula Mendes, realizada em 19/04/2021.

Comissão Julgadora:

Prof. Dr. Cesar Henrique Comin (UFSCar)

Prof. Dr. Alexandre Luis Magalhães Levada (UFSCar)

Prof. Dr. Francisco Aparecido Rodrigues (USP)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

A todos que de alguma maneira fazem esta vida valer a pena verdadeiramente.

AGRADECIMENTOS

Dizer obrigado, às vezes, não é suficiente para agradecer a todas as pessoas que nos momentos de nossas vidas, aqueles mais difíceis, nos estendem a mão e nos oferecem ajuda. Agradeço também aqueles que nos oferecem sermos melhor do que éramos, seja lá de que maneira.

A gratidão é imensa ao chegar neste momento, lembrar de todas as pessoas, que desde o início participaram. Algumas já se foram, algumas permanecem, algumas fazem questão de aparecer e estarem presentes, algumas não fazem nem questão de participarem, algumas nem sabem que de alguma maneira contribuíram para que este momento chegasse, algumas já se foram antes mesmo de quem estiver lendo este agradecimento nascer. Mas que, permanecem vivas em nossos pensamentos, algumas são meras lembranças, algumas nos deixam a saudade e vontade de estarem juntas e principalmente aqueles que gostam de permanecer ao nosso lado. Agradeço a todas essas pessoas, essas que fizeram e fazem parte disto tudo que somos. Agradeço a toda estrutura disponibilizada pela UFSCar e a todo o programa de pós-graduação em ciência da computação que me deu esta oportunidade de ser melhor em todos os sentidos e principalmente ao meu orientador Professor Doutor Cesar Henrique Comin que teve paciência e muita sabedoria durante toda esta dissertação, por todos os ensinamentos e cada minuto de dedicação. Muito obrigado! Enfim, agradeço a Deus. Que um dia eu consiga também estender a mão para oferecer ajuda e evolução.

Um dia você descobre que se leva muito tempo para se tornar a pessoa que se deseja tornar, e que o tempo é curto. Aprende que não importa até o ponto onde já chegamos, mas para onde estamos, de fato, indo - mas, se você não sabe para onde está indo, qualquer lugar servirá. Aprende que heróis são pessoas que foram suficientemente corajosas para fazer o que era necessário fazer, enfrentando as consequências de seus atos. Aprende que paciência requer muita persistência e prática. Descobre que, algumas vezes, a pessoa que você espera que o chute quando você cair, poderá ser uma das poucas que o ajudarão a levantar-se.

Veronica Shoffstall

RESUMO

Todo ano uma grande parte da população de homens é afetada pelo Câncer de próstata *(Prostate Cancer - PCa)* com novos casos e milhares de mortes ocorrendo especialmente em homens acima dos 40 anos. O PCa é uma condição generalizada que se difunde e se manifesta em uma ampla gama de padrões histológicos, que podem ser visualizados em imagens histológicas obtidas por meio de biópsia ou prostatectomia. A detecção precoce do PCa pode melhorar o prognóstico e reduzir significativamente o risco de morte. Atualmente, a principal metodologia para diagnóstico de PCa consiste em uma análise qualitativa realizada por especialistas para definir o grau da doença no chamado Sistema de Escore de Gleason *(Gleason Grading System - GGS)*, definido originalmente por Donald Gleason e refinado pela Sociedade Internacional de Patologia Urológica. Dada a importância da identificação de tecido anormal da próstata (estadiamento) para melhorar o prognóstico, muitas metodologias computadorizadas destinadas a auxiliar os patologistas de forma sistemática no diagnóstico têm sido desenvolvidas. É frequentemente argumentado que um diagnóstico melhorado de uma região de tecido pode ser obtido considerando medidas que levem em consideração o contexto no qual o tecido se encontra, isto é, propriedades da região no entorno do tecido sendo caracterizado. Tal contexto é considerado um importante fator biológico na definição do estadiamento. Este trabalho propõe uma nova metodologia que pode ser usada para definir sistematicamente características contextuais relacionadas às glândulas presentes em tecidos da próstata. Para tal, é definida a chamada Rede de Contexto Glandular *(Gland Context Network - GCN)* que é uma representação da amostra da próstata contendo informações sobre a relação espacial entre as glândulas, bem como a semelhança entre suas aparências. É demonstrado que tal rede pode ser usada para estabelecer características contextuais em qualquer escala espacial. Portanto, fornecendo informações que não são facilmente obtidas a partir de propriedades tradicionalmente utilizadas. Além disso, é mostrado que mesmo características básicos derivadas de uma GCN podem levar ao estado-da-arte no desempenho de classificação em relação ao PCa. Assim, as GCNs podem auxiliar na definição de abordagens mais eficazes para a classificação de PCa.

Palavras-chave: Câncer de Próstata, Escore de Gleason, Redes Complexas

ABSTRACT

Every year a large part of the male population is affected by Prostate Cancer (PCa), with many cases of deaths occurring especially in men over 40 years old. PCa is a pervasive condition that diffuses and manifests itself in a wide range of histological patterns, which can be visualized with details in histological images acquire through biopsy or prostatectomy. Early detection of PCa can improve the prognosis and reduce the risk of death. Currently, the main methodology for the diagnosis of PCa consists of a qualitative analysis carried out by specialists to define the degree of the disease in the so-called Gleason Grading System (GGS), originally defined by Donald Gleason and refined by the International Society of Urological Pathology. Given the importance of identifying abnormal prostate tissue (staging) to improve the prognosis, many computerized methodologies have been developed to assist pathologists in a systematic way for the diagnosis. It is often argued that an improved diagnosis of a tissue region can be obtained by considering measures that take into account various properties of the tissue surroundings, henceforth referred as the context of the tissue. Such a context is considered an important biological factor in staging. This work proposes a new methodology that can be used to systematically define contextual features related to prostate glands. The Gland Context Network (GCN) structure is defined, which is a representation of the prostate sample containing information about the spatial relationship between the glands as well as the similarity between their appearance. It is shown that the GCN can be used to establish contextual features at any spatial scale. Therefore, information that is not easily defined from traditional features can be easily extracted using the proposed approach. In addition, it is identified that even basic properties derived from a GCN can lead to state-of-the-art classification performance in relation to PCa. All in all, GCNs can assist in defining the most effective approaches for PCa detection.

Keywords: Prostate Cancer, Gleason Score, Complex Networks

LISTA DE FIGURAS

LISTA DE TABELAS

GLOSSÁRIO

CC – *Componentes Conexos*

CIE – *International Commission on Illumination*

CT – *Computed tomography*

ER – *Erdős-Rényi*

FN – *False Negative*

FPR – *False Positive Rate*

FP – *False Positive*

GCN – *Gland Context Network*

GGS – *Gleason Grading System*

GGS – *Gleason Grading System*

J.I.E. – *Jonathan I. Epstein*

JI – *Jaccard Index*

KNN – k_n*–Nearest-Neighbor*

L*a*b* – *luminância e cores únicas da visão humana*

MRF – *Markov Random Field*

MRI – *Magnetic Resonance Imaging*

NLA – *Nuclei-Lumen Association*

PCA – *Principal Component Analysis*

PCa – *Prostate Cancer*

PPMM – *Probabilistic Pairwise Markov Models*

PSA – *Prostate-Specific Antigen*

SVM – *Support Vector Machine*

TC – *Tomografia Computadorizada*

TNR – *True Negative Rate*

TN – *True Negative*

TPA – *True Positive Accuracy*

TPR – *True Positive Rate*

TP – *True Positive*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Este capítulo apresenta o contexto ao qual a pesquisa se dedica, sua motivação para resolver o problema em questão e uma breve introdução ao conteúdo.

1.1 Contexto e Motivação

O Câncer de Próstata *(Prostate Cancer - PCa)* afeta uma grande porção da população de homens ao redor do mundo, em especial homens com idade acima dos 40 anos. Em muitos países ele representa a principal causa de morte de homens. No Brasil é o segundo tipo de câncer mais comum, estando atrás apenas do câncer de pele não-melanoma. Estimou-se em torno de 65.840 casos em 2020 e cerca de 15.983 mortes em 2019 [1]. Nos Estados Unidos da América (EUA) foram estimados em torno de 248.530 novos casos e cerca de 34.130 mortes em 2021 [2]. Em números globais, o PCa também é o segundo tipo mais comum de câncer (CULP et al., 2020; WONG et al., 2016). Muitas destas mortes poderiam ter sido evitadas pela detecção precoce da doença, o que aumentaria a chance de recuperação após tratamento e até mesmo um prolongamento de sobrevivência [3]. Um bom diagnóstico pode fazer a diferença e em alguns casos pode alcançar a cura total ou pelo menos diminuir o avanço da doença prolongando o tempo de vida do indivíduo com PCa.

O diagnóstico do PCa atualmente consiste em exames retais digitais *(digital rectal examination - DRE)*, testes laboratoriais de antígeno específico da próstata *(Prostate-Specific Antigen - PSA)*, biópsia que é a análise via microscópio de amostras de tecido extraídas aleatoriamente, técnicas *in vivo* de imagens como Tomografia Computadorizada *(Computed tomography - CT)*,

[1] https://www.inca.gov.br/tipos-de-cancer/cancer-de-prostata
[2] https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html
[3] https://gco.iarc.fr/survival/survmark/

Imagens de Ressonância Magnética *(Magnetic resonance imaging - MRI)* e Ultrassom (MOTTET et al., 2020; VIDAL et al., 2011). Além de consumirem muito tempo, os diagnósticos feitos por especialistas podem ser subjetivos. Atualmente, uma das principais formas de diagnóstico do PCa é realizada através da classificação visual da amostra da biópsia de acordo com o Sistema de Escore de Gleason *(Gleason Grading System - GGS)*. A Sociedade Internacional de Patologias Urológicas *(International Society of Urological Pathology - ISUP)* vem refinando a definição dos padrões do Escore de Gleason e propondo atualizações do GGS para uma prática mais contemporânea (EPSTEIN et al., 2005, 2016; KRYVENKO; EPSTEIN, 2016; LEENDERS et al., 2020).

Uma atualização do GGS foi proposta em 2014 por Jonathan I. Epstein (J.I.E.) utilizando dados do Hospital Johns Hopkins com base nas modificações de 2005 (EPSTEIN et al., 2016). Em 2019, novas modificações da prática clínica, novas diretrizes de classficação patológica e inovações no campo da inteligência artificial na classificação do PCa foram reunidas em um novo consenso (LEENDERS et al., 2020). Não nos concentraremos em discutir o que foi mudado ou na comparação das versões anteriores mas sim na versão mais atualizada do GGS e como é feita sua caracterização de maneira sistemática. O GGS consiste em associar uma pontuação de 1 a 5 ao padrão que ocorre com maior frequência na amostra (escore primário) e também ao padrão com segunda maior frequência (escore secundário). Esses dois escores são somados para definir o escore final. Os resultados dos dados capturados por J.I.E. foram agrupados em 5 distintos grupos. As graduações vão desde o grupo 1, com pontuação abaixo ou igual a 6, possuindo glândulas discretas e bem formadas, até o grupo 5, com pontuação 9 ou 10, caracterizado por grupos de glândulas mal formadas, fundidas e/ou cribriformes.

A partir da identificação das propriedades características de PCa, é possível implementar um sistema de visão computacional que se aproxime da interpretação do especialista para caracterizar as amostras de maneira automatizada para os diversos grupos definidos pelo GGS. Para tanto, é necessário primeiro realizar uma segmentação da imagem, de maneira que seja possível identificar as diversas glândulas em suas particularidades conforme sua graduação.

A partir de uma motivação biológica e das novas definições do ISUP, foi identificado que glândulas cancerígenas tendem a aparecer próximas a outras glândulas cancerígenas, e que elas tendem a possuir características parecidas (KUMAR et al., 2014). Desta forma, o especialista realiza a classificação do PCa dentro do GGS encontrando grupos de glândulas similares e as classificam não somente conforme suas características individuais mas principalmente de acordo com as características de contexto do grupo em questão. Esta definição contextual se torna importante no momento ao qual não é mais observado apenas uma glândula isoladamente,

mas sim, em qual contexto ela está e como ela é influenciada por outras glândulas. Com isso, é possível mensurar o nível de PCa dos grupos encontrados e até mesmo se um grupo possui glândulas que estão tendendo para o próximo nível do GGS. Em alguns trabalhos da literatura, os autores consideram a importância de caracterizar o contexto das glândulas, por exemplo, são implementadas medidas de contexto, como a lotação do grupo ou médias das áreas das glândulas em uma dada região (NGUYEN; SARKAR; JAIN, 2012; DOYLE et al., 2007; MONACO et al., 2008). As definições de contexto utilizadas tendem a se basear em propriedades específicas das glândulas, e em geral não podem ser utilizadas para outras propriedades.

A intenção do presente trabalho é definir uma metodologia que possibilite a geração de contexto para qualquer propriedade de glândulas da próstata. A metodologia consiste em definir o que chamaremos de *Gland Context Network (GCN)* e é construída a partir de medidas de forma e textura utilizando estruturas de redes. Mostraremos que a metodologia proposta possibilita diversas novas investigações sobre tecidos da próstata, e que ela é capaz de melhorar a detecção de PCa.

1.2 Objetivos

Este trabalho possui dois objetivos principais:

1. Definir uma nova metodologia para caracterização sistemática de contexto em regiões de tecidos da próstata;
2. Verificar se a metodologia proposta possibilita aprimorar a detecção de PCa em tecidos da próstata.

O objetivo 1 será investigado utilizando redes para representar as glândulas presentes no tecido. A conectividade entre as glândulas será realizada de acordo com suas propriedades de forma. A partir da rede gerada, novas propriedades poderão ser extraídas utilizando informações de contexto. Para o objetivo 2, serão utilizados métodos de classificação supervisionados para investigar se as características de contexto possibilitam aumentar a performance de detecção de tecidos doentes.

1.3 Organização do Texto

A divisão do texto desta dissertação dá-se da seguinte maneira, no capítulo 2 é feita uma revisão da literatura discorrendo sobre a base do problema em questão e os principais desafios

encontrados atualmente seja na área computacional quanto na área clínica. No capítulo 3 são apresentadas as ferramentas e metodologias básicas de processamento de imagens necessárias para o desenvolvimento da metodologia proposta. No capítulo 4 é apresentada a metodologia proposta. No capítulo 5 são sintetizados os resultados obtidos a partir das GCNs criadas no estudo de caso e as contribuições para o estado-da-arte com exemplos de identificação de PCa. No capítulo 6 é apresentado a conclusão da pesquisa.

Capítulo 2

REVISÃO DA LITERATURA

Este capítulo discorre sobre trabalhos propostos na literatura que exploram o PCa e o GGS na área computacional e clínica. São descritos os desafios, lacunas e contribuições para a comunidade científica no diagnóstico de PCa.

2.1 Visão Geral

Com o intuito de pesquisar sobre a caracterização, segmentação e classificação do PCa em imagens de microscopia da próstata é importante fazer uma revisão sobre propostas da literatura, identificando quais são as melhores práticas dos últimos anos. A classificação sistemática que permanece como a mais utilizada desde os anos 70 é o GGS que foi proposto por Donald Gleason em 1974 (GLEASON; MELLINGER, 1974). Desde então, o GGS tem sido refinado conforme o conhecimento da área evolui, principalmente pela ISUP (LEENDERS et al., 2020; KRYVENKO; EPSTEIN, 2016; EPSTEIN et al., 2016, 2005). Partindo das definições descritas no GGS, técnicas de processamento de imagens e reconhecimento de padrões foram propostas na literatura com o objetivo de estimar e classificar o PCa de maneira robusta e eficiente. Neste capítulo serão discutidos algumas propostas com estas premissas.

2.2 Patologia do Câncer de Próstata

A base da caracterização celular surge de modelos com coloração a partir de hematoxilina e eosina (H&E) dos tecidos da próstata e tem a intenção de realizar prognóstico de estruturas doentes e sadias. As principais estruturas a serem analisadas são as diversas glândulas presentes no tecido da próstata. Estas são compostas por núcleos, citoplasma e lúmen (NGUYEN; SARKAR; JAIN, 2012). A figura 2.1 ilustra um exemplo dessa estrutura. São mostradas regiões contendo

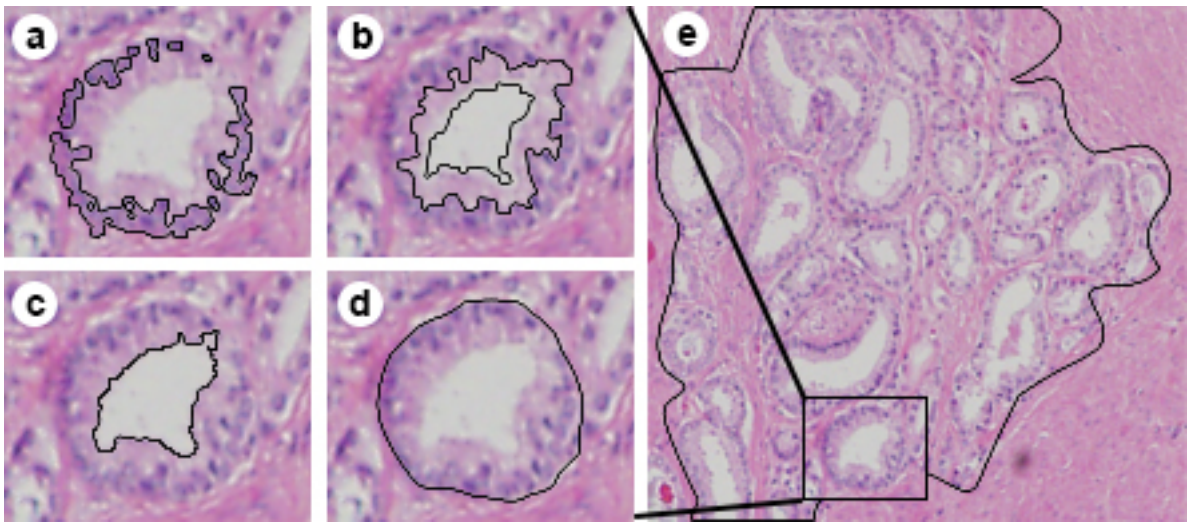

Figura 2.1: Estrutura Glandular. Estão indicados (a) os núcleos, (b) o citoplasma, (c) o lúmen e (d) uma glândula. Em (e) é mostrada uma região do tecido da próstata.

os núcleos da glândula (figura 2.1(a)), o citoplasma (figura 2.1(b)), o lúmen (figura 2.1(c)) e a glândula formada pela união das três estruturas (figura 2.1(d)). Na figura 2.1(e) é mostrada uma região de interesse com a glândula de exemplo em destaque. A aparência dessas glândulas é a principal característica utilizada no prognóstico do PCa.

Donald Gleason propôs em 1974 uma metodologia para a classificação sistemática dos graus (níveis) de câncer em pacientes com PCa. Para a definição dos níveis, os pacientes foram acompanhados desde o ano de início da doença detectada clinicamente até a morte deles utilizando-se dos possíveis tratamentos da época. Sabe-se, desde aquela época, que o estadiamento clínico do câncer diagnosticado pela primeira vez se correlaciona bem com o a extensão do tempo de vida do paciente (previsão de morte ou cura). Na tentativa de estabelecer padrões bem definidos entre o grau do câncer e o tempo de vida (restante em casos graves, cura em casos iniciais ou impedir o avanço da doença em casos médios[1]) do paciente, foram mapeados os primeiros padrões relacionados ao avanço da doença. Foi estabelecido um sistema de pontuação por escore de uma maneira sistemática relacionado com as formas das glândulas (estudo histológico) observadas em determinados graus estadiados clinicamente. Foi apresentado um sistema de escore histológico em 5 padrões de crescimento que foram enumerados em ordem crescente aparente com a malignância histológica. Dentro do escore 1 observou-se o padrão de glândulas individuais uniformes e bem diferenciadas, compactadas em massas com bordas relativamente bem delimitadas. Dentro do escore 2 observou-se glândulas individuais uniformes e bem diferenciadas mas com mais variações, ligeiramente espaçadas com bordas do tumor bem menos delimitadas. Dentro do escore 3 observou-se glândulas moderadamente diferenciadas, podendo

[1] `https://gco.iarc.fr/survival/survmark`

variar de pequenas a grandes, crescendo em padrões espaçados e infiltrativos, podendo ser papilares ou cribriformes. Dentro do escore 4 observou-se infiltração irregular, tumor glandular fundido, frequentemente com células pálidas, podendo assemelhar-se ao hipernefroma do rim. Dentro do escore 5 foi observado carcinoma anaplásico com mínima diferenciação glandular e estroma prostático difusamente infiltrado. Desta maneira, pode-se combinar o estadiamento clínico com o escore histológico sistematizado para prover um aumento na predição da malignância do carcinoma prostático estadiada pelo médico especialista (GLEASON; MELLINGER, 1974). A partir do diagnóstico feito pela definição do GGS é feito o auxílio no direcionamento do tratamento do paciente. A chance de cura aumenta quando a detecção do PCa é descoberta no início da doença de maneira assertiva.

Tendo um sistema de escore bem definido, o diagnóstico depende de análises qualitativas nas observações inter, intra glandulares (NAIK et al., 2007) e a observação da área de concentração de glândulas parecidas ou iguais formando os grupos com seus respectivos graus. Desde então, pesquisadores utilizam-se da computação para tentar mitigar a variabilidade nos diagnósticos, desenvolvendo sistemas para automatizar a estimação do escore de Gleason através da extração de características em diversos domínios como tamanho, texturas, formas, estruturas e estatísticas associadas.

2.3 Trabalhos Relacionados

2.3.1 Segmentação de Glândulas

O desafio de implementar algoritmos para auxiliar na análise realizada por patologistas vem sendo estudado por diversos autores ao longo da história. A primeira etapa é conseguir identificar as glândulas de PCa e então segmentá-las de maneira robusta. A segmentação possibilita extrair características das glândulas com o intuito de classificá-las ou até mesmo identificar características antes nunca notadas. Muitos trabalhos foram propostos na literatura para realizar a segmentação de glândulas. As metodologias propostas possuem procedimentos similares entre si, ilustrados na Figura 2.2. Tais metodologias tendem a diferir de acordo com os tipos de características extraídas para criar o vetor de características dos tecidos.

Nguyen et al. definem uma metodologia para a detecção e classificação de PCa partindo da segmentação das glândulas utilizando critérios similares aos utilizados por patologistas (NGUYEN; SABATA; JAIN, 2012; NGUYEN; SARKAR; JAIN, 2012). Os autores propõem extrair características com informações estruturais sobre a distribuição espacial das glândulas de forma a detectar glândulas em regiões normais e de câncer. São extraídas também informações con-

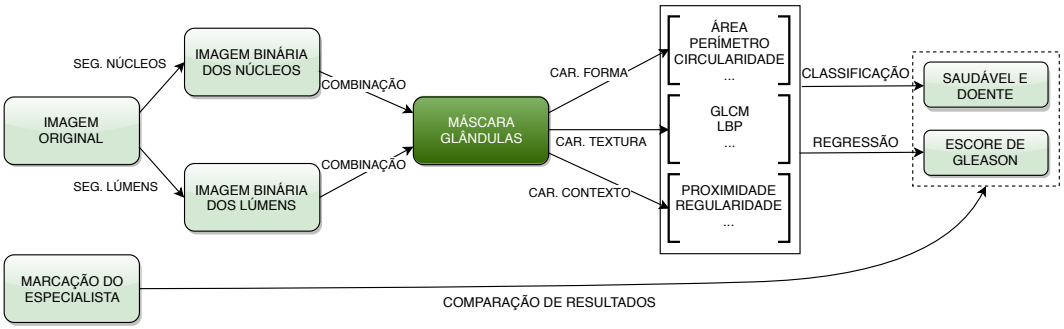

Figura 2.2: Diagrama de fluxo comum em análise de imagens de PCa. É realizada a segmentação dos núcleos e dos lúmens. As imagens segmentadas são unidas para gerar a segmentação das glândulas. Características de forma, textura e de contexto são extraídas e utilizadas para classificação no GGS.

textuais como a similaridade de tamanho e forma entre glândulas próximas. Para a detecção de glândulas, Nguyen et al. definiram um algoritmo chamado *segmentação de associação lúmen e núcleos (nuclei-lumen association - NLA)*. O primeiro passo do algoritmo é a detecção do lúmen de cada glândula. Na sequência, é feita uma busca por núcleos ao redor da borda de cada lúmen em uma região definida por um cone. Uma máscara circular é utilizada para limitar as regiões de núcleo a serem unidas à glândula. Durante a busca, *outliers* são detectados e eliminados de forma a se obter uma segmentação mais suave. Os pontos encontrados de lúmens e núcleos são utilizados na extração de características e posterior classificação das glândulas (NGUYEN; SARKAR; JAIN, 2012). A Figura 2.3 mostra uma ilustração do método de busca de núcleos relacionados ao lúmen com o objetivo de encontrar a glândula.

A avaliação dos resultados obtidos pelo algoritmo de segmentação foi feita utilizando o Índice de Jaccard *(Jaccard Index - JI)*. Cada glândula segmentada manualmente, G_i^0, é comparada com a i-ésima glândula segmentada pelo algoritmo m, G_i^m. O cálculo é feito como $J(G_i^0, G_i^m) = |G_i^0 \bigcap G_i^m| / |G_i^0 \bigcup G_i^m|$. Os resultados são valores entre 0 e 1, sendo que 1 indica uma segmentação perfeita. O método superou os outros dois métodos utilizados na comparação, alcançando o índice Jaccard de 0.66 contra 0.31 e 0.43. O algoritmo proposto visa detectar os núcleos em volta do lúmen enquanto os outros algoritmos focam na detecção do lúmen e citoplasma, o que pode ter levado a uma melhor segmentação. Também foi levado em consideração a complexidade computacional do algoritmo em relação aos outros métodos. O método NLA executou 10x mais rápido que os outros métodos implementados. Essa diferença ocorreu porque os métodos que foram comparados realizam operações complicadas utilizando as intensidades dos pixels e também necessitam de um algoritmo de crescimento de região, enquanto que o algoritmo NLA faz o processamento considerando apenas os rótulos dos pixels, o que o torna melhor também no custo computacional (NGUYEN; SARKAR; JAIN, 2012). A figura 2.4 (a) representa o *Ground Truth* criado manualmente, (b) representa o resultado do método proposto

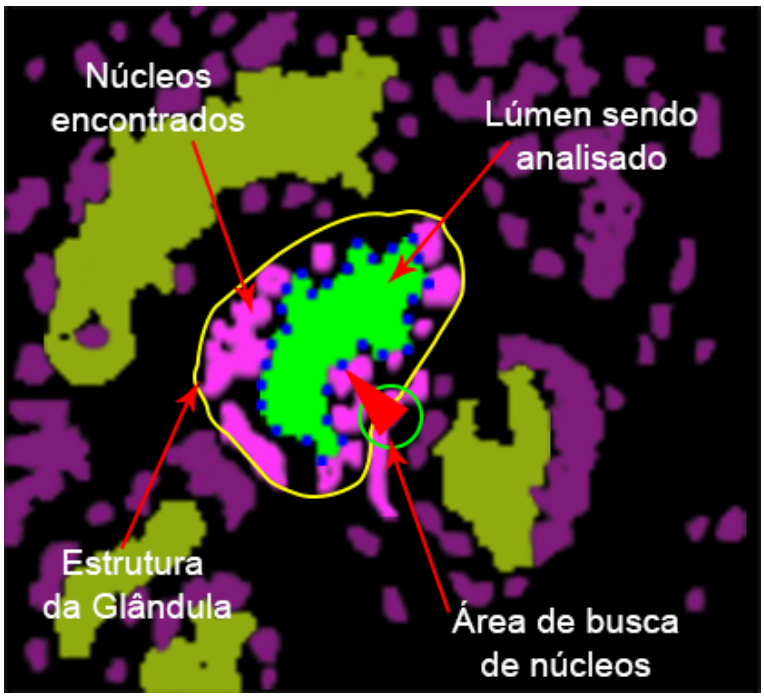

Figura 2.3: Variação da busca de núcleos a partir do lúmen. Algoritmo proposto por Nguyen et al.

por Nguyen et. al.

Naik et al. em 2007 propuseram análises de textura para seleção de características de baixo nível (valores dos pixels) para encontrar os lúmens (NAIK et al., 2008). Os autores mencionam que a técnica não é apropriada para segmentar o citoplasma e propuseram a utilização de características de baixo e alto nível (relação dos pixels com os objetos a serem detectados) para criar o algoritmo de segmentação para estruturas de interesse na histopatologia de câncer de próstata e mama. Amostras com informações de textura e intensidade das imagens foram selecionadas e utilizadas numa combinação de conhecimento de domínios (baixo nível, alto nível e características de domínio específicas das estruturas histológicas). Limitações de estruturas foram impostas e um algoritmo de seleção de nível (derivada de segunda ordem) e correspondência de modelo foram utilizados, respectivamente, para a identificação das bordas das glândulas e para a segmentação nuclear. Os autores mencionam a dificuldade de realizar a detecção de borda nas imagens de câncer de próstata e mama de forma a capturar todos os núcleos na borda da glândula (NAIK et al., 2008).

Alguns estudos utilizam-se de imagens *whole-mount*, representando o órgão glandular inteiro, para fazer a detecção de regiões de carcinoma. Monaco et al. propõe uma metologia que tem uma ótima performance computacional de detecção e classificação de câncer mesmo em grandes imagens como a *whole-mount* (MONACO et al., 2010). O algoritmo proposto faz a segmentação das glândulas usando o canal de luminância no espaço de cores CIE (International

Commission on Illumination) L*a*b*, pois os autores argumentam que a informação de cor não é necessária para identificar os lúmens nas imagens de seções histológicas (*whole-mount*). No canal de luminância, as glândulas aparecem como regiões de pixels contíguos e de alta intensidade circunscritos por bordas nítidas e pronunciadas. A imagem é analisada em múltiplas escalas de forma a levar em conta os tamanhos variáveis das glândulas. Os picos são considerados candidatos para os centros das glândulas, que são sementes para o procedimento de crescimento de região que opera na imagem original. Na segmentação final, as regiões detectadas podem se sobrepor. Isto é resolvido descartando as regiões com bordas menos nítidas.

Estudos recentes utilizam-se de estratégias como a classificação de *microarrays* via *deep learning* ou métodos de regressão para a seleção de características e associação com parâmetros clínicos (anotações feitas por especialistas) (ARVANITI et al., 2018).

De forma resumida e comparativa, alguns autores e suas estratégias de segmentação estão descritos na tabela 2.1. Onde a coluna "Estudo" é a referência do estudo comparado, a coluna "Carac. da Glândula" representa as características utilizadas durante o estudo proposto, a coluna "Objetivo" é a finalidade do estudo e a coluna "Base de Dados" referencia as imagens utilizadas durante o estudo.

Tabela 2.1: Comparação de estratégias para segmentação e classificação de glândulas no contexto de PCa.

Estudo	Carac. da Glândula	Objetivo	Base de Dados
(Nguyen K and A, 2012)	Características estruturais	Classificação de Glândulas	48 imagens em 5x
(Monaco et al., 2010)	Tamanho da glândula	Detectar regiões de câncer	40 imagens em 1.23x
(Naik et al., 2007)	Características de forma do lúmen e glândula	Classificar imagens de tecido	44 imagens em 40x
(Naik et al., 2008)	Características morfológicas	Detectar e segmentar estruturas de interesse	24 imagens de tecido de biópsia

2.3.2 Detecção de Câncer de Próstata

Com estratégias bem definidas para segmentar as glândulas, é possível realizar a estimação do Escore de Gleason. Trataremos a seguir trabalhos relacionados com este intuito. Nguyen et. al propuseram agrupar as características estruturais das glândulas em quatro grupos que são percebidos de forma distinta pela visão humana (NGUYEN; SARKAR; JAIN, 2012). São eles: núcleos (8 características), citoplasma (6 características), forma do lúmen (3 características)

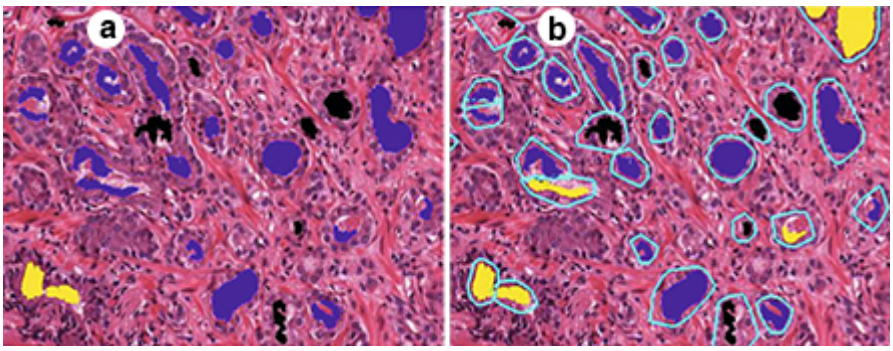

Figura 2.4: Resultado do método obtido por Nguyen et al. a) Marcação do especialista. b) Resultado do método, contornos azul claro denotam a segmentação obtida, preto são glândulas não rotuladas, vermelho são artefatos, amarelo são glândulas normais e azuis são glândulas cancerígenas. Fonte: Figura adaptada de (NGUYEN; SARKAR; JAIN, 2012)

e propriedades globais (2 características). Tendo os quatro grupos formados, foram definidas também três características contextuais aplicadas a agrupamentos de glândulas. São elas: lotação da vizinhança, similaridade de forma e similaridade de tamanho. Finalmente, foi criado um vetor de dimensionalidade 22 (19 + 3) para caracterizar cada glândula, e o classificador *SVM (Support Vector Machine)* foi utilizado para realizar a classificação das glândulas. Um *kernel* linear com $C = 1$ foi utilizado.

Durante a realização dos experimentos, é necessário ter uma base de comparação seja ela manual ou de outras propostas na literatura. Para isto, a base de dados foi manualmente classificada (NGUYEN; SARKAR; JAIN, 2012) por um patologista para a criação de máscaras *ground truth* das glândulas e devidamente rotuladas como "artefatos", "glândulas normais"e "glândulas cancerígenas". De forma a encontrar a acurácia foi implementado uma validação cruzada com 10 partições do conjunto de dados. A classificação se concentra em resolver dois problemas: (i) artefato vs. glândulas e (ii) glândulas normais vs. glândulas cancerígenas. As soluções dos problemas anteriores foram utilizadas para criar uma classificação de forma hierárquica em três classes. A figura 2.4 ilustra o resultado obtido (NGUYEN; SARKAR; JAIN, 2012). Glândulas pretas não foram rotuladas pelo especialista mas detectadas, vermelho são artefatos, amarelo são glândulas normais e azuis são glândulas cancerígenas.

Alguns autores utilizaram-se de secções histológicas provenientes de prostatectomias radicais, que dependendo do tipo de algoritmo utilizado pode demorar algumas horas para fazer a estimação do Escore de Gleason na imagem toda. Para este tipo de abordagem, Monaco et. al propuseram um algoritmo separado em duas etapas, a primeira etapa é encontrar as regiões de câncer e a segunda etapa é realizar a estimação do Escore de Gleason nas regiões encontradas (MONACO et al., 2010). Ajustando o sistema para analisar as seções histológicas em baixa resolução foi possível atingir um alto rendimento em cerca de 3 minutos para percorrer a seção

toda. Para a estimação do Escore de Gleason foi proposto outro algoritmo com três etapas, a primeira etapa é segmentar as glândulas, a segunda etapa é classificar as glândulas em benigna ou maligna e a terceira e última etapa é consolidar as glândulas malignas em regiões contínuas. A classificação individual das glândulas aproveita as características biológicas de tamanho e a tendência de glândulas próximas compartilharem a mesma classe, uma última característica utilizada descreve a dependência espacial que é modelada utilizando prioridades de Markov em termos de funções de densidade de probabilidade PPMM (*Probabilistic Pairwise Markov Models*). As Glândulas são modeladas utilizando um Campo Aleatório de Markov (*Markov Random Field MRF*) que é estabelecido através da construção de condições de funções de probabilidade de densidade descrevendo as inter-dependências entre regiões próximas. Os autores argumentam que a introdução de um Campo Aleatório de Markov para a modelagem da imagem representa uma importante contribuição para a estimação do Escore de Gleason, pois proporciona uma caracterização mais intuitiva dos dados, excedendo resultados obtidos com o modelo de Potts (MONACO et al., 2010).

Naik et al. propôs um método para estimar o escore de Gleason nível 3, 4 ou epitélio benigno via SVM (NAIK et al., 2007). O método faz uso de objetos incorporados em um espaço dimensional reduzido obtido através de uma técnica de *manifold learning*, abordagem para redução de dimensionalidade não-linear utilizando grafos. O método reduz o espaço de características de alta dimensão M para um espaço de características de baixa dimensão N, preservando da melhor forma possível as relações inter e intra dos dados. A proposta do algoritmo de incorporação de grafos utiliza-se de cortes normalizados. O grafo incorporado produz uma matriz de confusão descrevendo a similaridade entre duas imagens C_p e C_q, calculada como:

$$\gamma(p,q) = e^{-||F_p - Fq||}, \tag{2.1}$$

onde F_p e F_q são os vetores de características das imagens. $p,q \in \{1,2,...,k\}$ e k é o número total de imagens no conjunto de dados. Então, o vetor X é obtido a partir da maximização da função:

$$\varepsilon_y(X) = 2\alpha \frac{X^T(\mathcal{D}-\gamma)X}{X^T\mathcal{D}X}, \tag{2.2}$$

onde $\mathcal{D}(p,p) = \Sigma_q \gamma(p,q)$ e $\alpha = |I| - 1$. O espaço incorporado N-dimensional é definido pelos autovetores correspondentes aos N menores autovalores de $(\mathcal{D}-\gamma)X = \lambda \mathcal{D}X$. O valor de N foi otimizado pela obtenção da precisão da classificação para $N \in \{1,2,...,10\}$ e foi selecionando o N que forneceu a precisão mais alta em cada tarefa da classificação. Os dados são classificados utilizando SVM, sendo que a função de kernel K(·,·) é calculada utilizando os autovetores

gerados. A estimação então é realizada para todas as amostras utilizando ambos espaços de características reduzido e não reduzido.

A tarefa de detecção e segmentação automática de estruturas nucleares e glandulares é crítica para a estimação da histopatologia do câncer de próstata e de mama. Desta maneira, Naik et. al propôs uma metodologia que utiliza-se da combinação de três níveis de informação da imagem para realizar a segmentação dos glândulas e núcleos (NAIK et al., 2008). Utilizou-se de informação de baixo nível com base no valor dos pixels da imagem, informação de alto nível com base na relação entre os pixels da imagem e informação específica de domínio com base na relação histológica das estruturas. Para a extração de características foram calculadas 8 características de bordas do interior dos núcleos e das bordas dos lúmens num total de 16 características morfológicas que quantificam o tamanho e a forma das glândulas. Também foram calculadas 51 características com base em grafos gerados pelo diagrama de Voronoi, árvore de expansão mínima e a triangulação de Delaunay utilizando as centroides dos núcleos para quantificar a relação espacial entre os núcleos. Todas as características foram escolhidas com base em características tipicamente utilizadas por patologistas para a detecção e estimação de câncer na prática. A qualidade da segmentação automática foi avaliada comparando a estimação do câncer de próstata e mama. As precisões de discriminação benigna versus câncer com os correspondentes dados obtidos por meio da detecção manual de glândulas e núcleos foram usados na SVM, para então, realizar a comparação da mesma maneira proposta por Naik et al. em 2007 para as características do câncer de próstata e mama. Foram utilizadas metodologias de redução de dimensionalidade utilizando o PCA (*Principal Component Analysis*). Os resultados obtidos foram favoráveis comparados com as segmentações manuais (NAIK et al., 2008).

2.4 Considerações Finais

Os trabalhos apresentados demonstram interessantes metodologias em relação a automatização da segmentação das glândulas e a respectiva caracterização das mesmas para a detecção de PCa. A possibilidade de extensão das metodologias propostas pode trazer benefícios não somente para o diagnóstico do câncer de próstata como também para outras histopatologias como o câncer de mama e o de pele. A grande maioria os estudos se preocupam com a automatização da segmentação e na separação do tecido benigno do tecido com câncer. Vários estudos propõem formas diferentes de extração de características e até mesmo metodologias diferentes de transformação de dados na tentativa de reduzir a dimensionalidade ou até mesmo mudar de domínio espacial para obter um melhor resultado. Outros estudos buscam prever o escore de Gleason, sendo que, muitas vezes são propostos métodos para caracterizar apenas alguns níveis

do escore.

As fontes dos dados na sua maioria são feitas com apenas uma parte do tecido capturado pela biópsia realizada no paciente com suspeita de câncer. Alguns trabalhos propõem a análise com imagens *whole-mount* da próstata realizada pela captura de imagens após feita a prostatectomia radical e algumas imagens obtidas a partir de pacientes *ex-vivo*. Imagens *whole-mount* possibilitam o estudo da eficiência de se analisar apenas partes da próstata (biópsia).

Uma classe de propriedades que se mostrou muito relevante nos trabalhos analisados são as chamadas medidas de contexto. Tais medidas levam em conta características de conjuntos de glândulas na análise dos tecidos. Por exemplo, uma glândula com características típicas de um tecido doente pode não ser relevante se ela estiver rodeada de glândulas claramente saudáveis. Os trabalhos analisados definem metodologias e critérios distintos para criação de medidas de contexto. O principal objetivo deste trabalho é propor uma metodologia que possibilite a criação sistemática de medidas de contexto a partir de medidas de forma obtidas a partir das glândulas. Com isso, a importante informação sobre o contexto da glândula pode ser obtida facilmente e utilizada como característica adicional na classificação ou na determinação do escore de Gleason do tecido. A metodologia proposta GCN, será apresentada no Capítulo 4.

Capítulo 3

CONCEITOS BÁSICOS

O capítulo atual apresenta as ferramentas e técnicas básicas de processamento de imagens necessárias para o desenvolvimento da metodologia proposta.

3.1 Segmentação não supervisionada

Quando tratamos do caso de segmentação não supervisionada trabalhamos com critérios abertos, ou seja, são dados conjuntos de objetos e então solicita-se a busca por classes adequadas, sem o uso de protótipos específicos (JR; COSTA, 2009). O interesse em segmentação não supervisionada possui diversas motivações. Coletar e rotular um grande conjuntos de dados pode ser muito custoso. Em um possível cenário, a fala de algum ator gravado pode ter custo zero, mas rotular as palavras e fonemas em cada instante com exatidão pode ser muito caro e demorado. Se for possível desenvolver um classificador com um pequeno conjunto de rótulos de exemplo e então melhorá-lo para que possa realizar suas tarefas sem supervisão em um conjunto de dados grande e sem o conjunto de rótulos, muitos problemas poderiam ser resolvidos e um bom tempo poderia ser ganho. Pode-se querer seguir numa direção inversa onde é feito o treinamento com grandes quantidades de dados não rotulados (sem custo) e somente depois usar a supervisão para rotular os grupos encontrados. Este tipo de abordagem é apropriada para grandes aplicações de mineração de dados que possuem uma grande base de dados desconhecida. Em alguns casos, as características dos padrões podem mudar lentamente com o tempo. Por exemplo, a classificação do estilo de comida consumida, de modo não supervisionado, poderia ser detectada conforme as mudanças de estações ocorrem. Em estágios iniciais de investigação em uma pesquisa pode ser valioso obter alguma compreensão sobre a natureza ou estrutura dos dados. Então, a descoberta de subclasses ou similaridade entre os dados podem sugerir alterações significantes na estruturação do método utilizado (DUDA; HART; STORK,

2012).

Para algumas classes de aplicações em processamento de imagens digitais o principal objetivo é utilizar os operadores adequados para encontrar formas ou descritores que sejam capazes de caracterizar o objeto a ser encontrado. Por exemplo, objetos como bolhas, aerossóis, gotas, partículas de pigmento, núcleos de células e até mesmo glândulas cancerígenas e glândulas saudáveis que precisarão ser analisadas a posteriori. Dentro do espaço de características, podemos ter dois tipos principais de procedimentos que podem ser distinguidos em: classificação baseada em pixels (baixo nível) e classificação baseada em objetos (alto nível). Em casos complexos de segmentação não é possível utilizar apenas uma única característica. Então, é necessário utilizar-se de múltiplas características e talvez alguns processos de classificações para decidir qual pixel pertence a qual tipo de objeto e quais pixels pertencem ao fundo da imagem. Por exemplo, a posteriori pode-se realizar a classificação baseada nos objetos utilizando características geométricas (JÄHNE, 1997).

Para imagens complexas com características dependentes de suas cores e formas, uma ótima abordagem é realizar a segmentação no espaço de cores RGB (*Red, Green, Blue*). Dado um conjunto inicial de pontos representativos das cores de interesse, precisamos obter uma estimativa das cores típicas as quais desejamos segmentar. O objetivo da segmentação é classificar cada pixel em RGB da imagem dada como pertencendo a cor dentro de um grupo especificado ou não.

Um classificador comumente utilizado para classificação de tipos (grupos) é o K-médias onde cada observação pertence ao grupo possuindo a média mais próxima. Com a finalidade de simplificar a computação e acelerar a convergência das observações existem várias técnicas que podem ser utilizadas. Duda et al. brevemente consideraram um método elementar de aproximação onde a probabilidade $\widehat{P}(\omega_i|x_k,\widehat{\theta})$ é grande quando a distância de Mahalanobis (invariante a escala) $(x_k-\widehat{\mu}_i)^t\widehat{\sum}_i^{-1}(x^k-\widehat{\mu}_i)$ é pequena (DUDA; HART; STORK, 2012). De mesma maneira, ao computarmos a distância Euclidiana $\|x_k-\hat{\mu}_i\|^2$ encontramos a média $\hat{\mu}_m$ mais próxima de x_k e aproximando a probabilidade $\widehat{P}(\omega_i|x_k,\widehat{\theta})$, teremos a equação 3.1.

$$\widehat{P}(\omega_i|x_k,\widehat{\theta}) \approx \begin{cases} 1 & \text{se } i=m \\ 0 & \text{caso contrário.} \end{cases} \tag{3.1}$$

A partir da equação 3.1 é possível realizar o procedimento de busca entre todos os elementos de determinado conjunto de dados. O algoritmo de iteração entre os elementos é historicamente referenciado como K-médias que pode ser descrito como um método para se obter as estimativas de máxima verossimilhança a partir de diferentes pontos de partida (sementes).

3.2 Caracterização de Formas

Georges Matheron e Jean Serra em 1982, definiram a morfologia matemática concentrando seus esforços no estudo da estrutura geométrica das entidades presentes em uma imagem. A base da morfologia matemática é a teoria de conjuntos na qual a principal aplicação é extrair características da imagem que sejam úteis em sua representação e descrição de formatos (FILHO; NETO, 1999). Por exemplo, o conjunto de pixels de determinada cor em uma imagem colorida que descreve uma forma.

A evolução produziu sistemas visuais que em sua própria natureza dedicam sua atenção para detectar variações abruptas ao longo dos objetos, assim como, suas delimitações nas bordas criando fronteiras em suas regiões que levaram ao conceito humano de *forma*. A evolução das habilidades de processamento visual dedicadas, ou pelo menos relacionadas à captura, processamento, análise e classificação de formas deram origem à morfologia computacional. Durante a etapa de caracterização de formas, comumente adotam-se medidas (ou características) para utilizar durante o processo de classificação e reconhecimento de padrões. Dependendo da complexidade da forma é necessário abordar descritores mais complexos, como por exemplo, as dimensões fractais e descritores de *Fourier* para a extração de informações importantes sobre suas características (JR; COSTA, 2009).

No contexto deste trabalho, análises morfológicas têm como objetivo descrever a aquisição, processamento e suas representações com base nos modelos do GGS e compará-los aos modelos de padrões analisados por um patologista. A caracterização e reconhecimento de padrões em glândulas partem de alternativas simples como área, perímetro e elongação, que são obtidas individualmente de cada glândula, e vai até descritores mais sofisticados e complexos para classificar e analisar grupos de glândulas, como por exemplo através do uso de grafos.

A caracterização de formas não envolve apenas descobrir e mensurar características mas sim compreender quais conjuntos de características possuem bom poder discriminativo. Cada medida, pode então, ser normalizada utilizando transformações estatísticas para então treinar um algoritmo de classificação. Não existe uma solução genérica definitiva para escolher características ótimas e obter um algoritmo de classificação ideal. Ao longo das análises são utilizadas diferentes medidas de forma propostas por outros trabalhos da literatura que analisaram imagens de PCa. Tais medidas costumam ser baseadas em critérios utilizados por patologistas na quantificação do escore de Gleason. O sistema de graduação descrito por Gleason(GLEASON, 1977) (GGS) aceito e refinado pela ISUP considera, entre outras características, a forma e o tamanho das glândulas para identificar os padrões de crescimento arquitetural (LEENDERS et al.,

2020; EPSTEIN et al., 2016, 2005). Nas revisões do GGS feitas pela ISUP em 2014 e 2019, a forma é mencionada como um preditor para várias pontuações. Por exemplo, os padrões benignos tendem a ter grandes glândulas circulares e ovais, enquanto os padrões de Gleason grau 3 geralmente têm glândulas menores e mais circulares quando comparados aos padrões benignos. Além disso, estruturas glandulares irregulares podem ser observadas em padrões de Gleason grau 4, uma vez que algumas glândulas são indiferenciáveis, fundidas com outras glândulas ou contêm muitos pequenos orifícios (padrões cribriformes) (NGUYEN; SABATA; JAIN, 2012; EPSTEIN et al., 2016).

O GGS fornece uma referência abrangente para os muitos padrões diferentes que podem ser observados nos tecidos da próstata. Mesmo assim, as pontuações ainda são avaliadas qualitativamente. Desta maneira, é relevante definir valores quantitativos para as características de forma consideradas pelos especialistas no prognóstico do PCa. Muitas características de forma foram usados em estudos de prognóstico automatizados de PCa (NGUYEN; SABATA; JAIN, 2012; NGUYEN; SARKAR; JAIN, 2012; NAIK et al., 2008, 2007). Com a finalidade de mostrar o potencial das *Gland Context Networks (GCNs)*, foram selecionadas características de forma identificadas como relevantes para o prognóstico do PCa nos estudos citados. As características consideradas e as respectivas definições utilizadas neste trabalho são:

Área: quantifica o tamanho da glândula. É calculado como o número de pixels que representa a glândula. A área é considerada uma característica importante no prognóstico automatizado do PCa, uma vez que tende a mudar conforme o câncer evolui (NGUYEN; SARKAR; JAIN, 2012; EPSTEIN et al., 2005, 2016).

Perímetro: está associado ao tamanho e, portanto, à área de uma glândula, mas fornece informações complementares, já que duas formas com áreas semelhantes podem ter perímetros muito distintos. É calculado como o comprimento do arco do contorno paramétrico da forma da glândula (JR; COSTA, 2009).

Solidez: quantifica o quão próxima a forma da glândula está de um polígono convexo. É calculado como a razão entre a área da glândula e a área do casco convexo da glândula (JR; COSTA, 2009). Glândulas com contornos irregulares tendem a ter baixa solidez.

Excentricidade: representa o alongamento de uma glândula. Neste trabalho ele é calculado como a razão entre os autovalores da matriz de covariância calculada a partir das posições dos pixels que representam a forma (JR; COSTA, 2009). Curiosamente, os padrões com uma pontuação de 5 no GGS tendem a conter glândulas com formas cilíndricas (EPSTEIN et al., 2016). A seção transversal dessas glândulas pode aparecer como formas alongadas em amostras de tecido.

Diâmetro: quantifica a distância máxima entre quaisquer dois pontos no contorno de uma glândula.

3.3 Redes Complexas

Nos últimos 20 anos, redes complexas se tornaram uma ferramenta importante para analisar sistemas interconectados, como por exemplo interações proteína-proteína (HAN et al., 2004; VELLA et al., 2018), redes sociais (CASTELLANO; FORTUNATO; LORETO, 2009) e cidades (DOMINGUES et al., 2018). Uma rede ou grafo pode ser entendido como uma abstração topológica de um sistema na qual os elementos são representados por nós (COSTA, 2004, 2006), que são interligados por arestas se estiverem relacionados de acordo com algum critério. Tal abstração fornece uma representação comum para diversos sistemas distintos, permitindo a respectiva aplicação de metodologias de teoria de redes para caracterizar e modelar sua dinâmica e também são potencialmente úteis em análise de formas e visão computacional (COSTA, 2004, 2006). Para o propósito de definir redes de contexto, dois aspectos principais precisam ser considerados. A metodologia utilizada para construir a rede e as propriedades utilizadas para caracterizar seus nós. Esses aspectos são apresentados e discutidos a seguir.

3.3.1 Definindo conectividade

Uma rede pode ser representada matematicamente como $G = (V, E)$, onde V é o conjunto de nós e E o conjunto de arestas, refletindo a conectividade entre pares de nós. Portanto, dois nós a, b de V são ditos adjacentes, ou conectados, se $(a, b) \in E$. Uma rede também pode ser ponderada, caso em que um valor é associado a cada uma de suas arestas. Normalmente, o peso está associado ao grau de relacionamento entre os nós, quantificado por alguma métrica de similaridade ou dissimilaridade. As redes também podem ser categorizadas como espaciais ou não espaciais. Em redes espaciais, cada nó tem uma posição específica em algum espaço métrico, geralmente o espaço euclidiano, e a distância entre pares de nós influencia sua conectividade. No caso de redes não espaciais, outras metodologias de conexão dos nós são consideradas. As abordagens mais comuns incluem conectar os nós aleatoriamente (ERDOS; RéNYI, 1959), com preferência para nós de alto grau (BARABÁSI; ALBERT, 1999) ou de acordo com uma sequência fixa do número de conexões feitas por cada nó (NEWMAN, 2018). Muitas abordagens diferentes podem ser usadas para levar em consideração a distância entre os nós ao decidir se eles devem ser conectados ou não. Critérios relevantes são:

1. Conectar pares de nós possuindo distância menor que um valor fixo R. Este critério é

utilizado para gerar os chamados Grafos Geométricos (DALL; CHRISTENSEN, 2002);

2. Conectar os nós de acordo com uma probabilidade que decai exponencialmente com a distância entre os nós, conhecido como o critério de Waxman (WAXMAN, 1988);

3. Conectar nós com células adjacentes do Voronoi (BARTHÉLEMY, 2011);

4. Conectar pares de nós apenas se um disco que está centrado no ponto médio dos dois nós e possui um diâmetro igual à distância entre os nós não contiver nenhum outro nó (GABRIEL; SOKAL, 1969).

Tendo em vista as diversas possibilidades de conectividade dos nós da rede, torna-se necessário elaborar alguma estratégia a fim de criar uma rede que auxilie na investigação do problema em estudo. A partir da rede com suas adjacências construídas, se torna possível explorar características que são relevantes para o problema.

Uma construção que foi utilizada em alguns trabalhos da literatura sobre câncer de próstata é baseada na Tesselação de Voronoi. Esse método consiste em recobrir uma superfície bidimensional, tendo como unidades básicas polígonos congruentes ou não, sem que existam espaços entre eles e de modo que a superfície total seja igual ao espaço particionado. A tesselação é gerada a partir de um conjunto de pontos semente isolados P_i, $i = 1,2,...,N$ em $\mathbb{R}^2$. O espaço $\mathbb{R}^2$ é então particionado em N células, representadas como $\mathbb{R}_i$, que possuem associação com cada ponto semente P_i, ou seja, qualquer ponto dentro da célula $\mathbb{R}_i$ está mais próximo ao ponto semente P_i do que a qualquer outro ponto semente. Por exemplo, a figura 3.1(a) mostra uma imagem com glândulas da próstata a serem representadas pela Tesselação de Voronoi. Na figura 3.1(b) é mostrado as glândulas segmentadas. A respectiva tesselação de Voronoi utiliando os centroides das glândulas é mostrada na figura 3.1(c), na forma de linhas azuis tracejadas. Células de Voronoi adjacentes são então conectadas, definindo a representação em rede.

Apesar de redes geradas por Tesselação de Voronoi serem uma importante ferramenta na análise de glândulas da próstata, elas possuem uma característica que dificulta a interpretação dos resultados. A distância entre glândulas conectadas pode variar muito ao longo da rede, pois mesmo glândulas muito distantes podem acabar sendo conectadas dependendo da distribuição das posições da glândulas. Para a análise dos resultados, é interessante que seja possível controlar de forma sistemática a escala da análise, ou seja, especificar as distâncias típicas entre glândulas conectadas. Um dos modelos mais simples de redes espaciais que possuem essa característica são os chamados grafos geométricos aleatórios (DALL; CHRISTENSEN, 2002). Nós são distribuídos uniformemente em um espaço bidimensional e dois nós são conectados por uma aresta se eles estiverem separados por uma distância menor que R. Portanto, um disco de

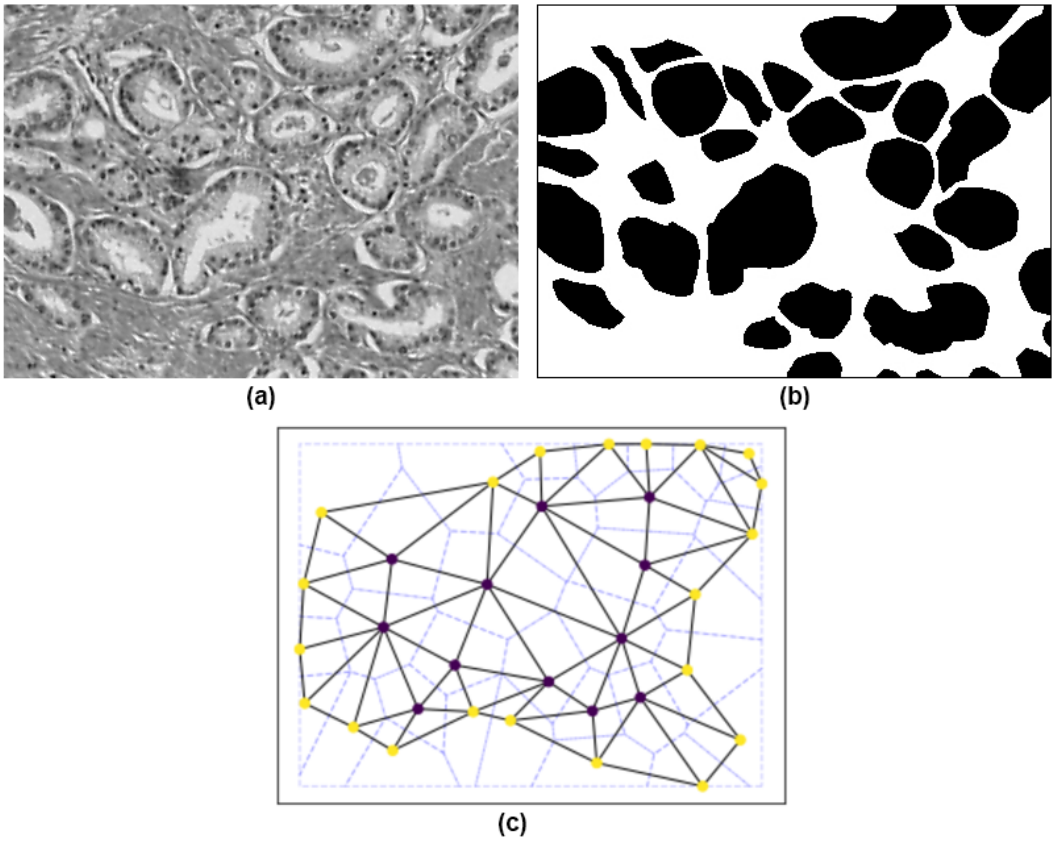

Figura 3.1: Exemplo de Tesselação de Voronoi para formar a rede com as glândulas de uma parte do tecido da próstata.

alcance pode ser associado a cada nó, e vértices são conectados se os discos correspondentes se intersectam. Os grafos geométricos randômicos também têm sido utilizados nos modelos de percolação contínua (BALBERG, 1985; QUINTANILLA; TORQUATO; ZIFF, 2000). Segundo Marc Barthélemy (BARTHÉLEMY, 2014), em d dimensões a probabilidade p que dois vértices escolhidos randomicamente estejam conectados é igual ao volume da hiperesfera de raio $R = 2r$, ou seja:

$$p = V(R) = \frac{\pi^{d/2} R^d}{\Gamma(1 + d/2)}. \tag{3.2}$$

De maneira geral, os grafos geométricos aleatórios tendem a possuir alto coeficiente de clusterização (BARTHÉLEMY, 2014). Isso porque o modelo leva em consideração a sobreposição dos discos e conexões mais longas do que R são proibidas. Este fato implica que se ambos vizinhos i e j estão conectados ao vértice k, eles devem estar na mesma vizinhança espacial de k.

3.3.2 Caracterização de Redes

Dada uma rede, é relevante caracterizar os nós de acordo com a respetiva topologia, ou seja, o padrão de ligações efetuadas pelos nós. Para fazer isso, primeiro definimos dois conceitos-chave. O *comprimento do caminho mais curto* entre dois nós V_i e V_j é calculado como o número de arestas no menor caminho entre os dois nós, onde um caminho é definido como uma sequência de arestas na rede.

Ao caracterizar a conectividade dos nós, não apenas as conexões imediatas do nó precisam ser consideradas. Por exemplo, pode ser interessante medir o número de conexões entre os vizinhos de um nó, ou entre os vizinhos dos vizinhos desse nó. Se forem considerados apenas os nós da primeira vizinhança de um nó, a respectiva caracterização será muito local, ou seja, se referirá apenas aos nós diretamente associados ao nó de referência. Ao contrário, se uma grande vizinhança for considerada, os nós indiretamente associados ao nó de referência também serão levados em consideração. Essa escolha de localidade da medida também se relaciona com o contexto no qual as glândulas serão analisadas.

Inúmeras propriedades de redes foram definidas na literatura (COSTA et al., 2007). Com a finalidade de apresentar possíveis aplicações da metodologia proposta neste trabalho, estudaremos apenas algumas características simples de redes, mas outras características poderiam ser consideradas. As características estudadas serão:

Grau: O número de conexões feitas por um nó. Geralmente está associado à importância, ou centralidade, de um nó, uma vez que nós com grandes graus estão associados a muitos outros nós da rede. A distribuição de probabilidade dos graus dos nós, conhecida como distribuição de graus, pode revelar muitas propriedades importantes sobre a rede (COSTA et al., 2007). Uma ilustração dessa propriedade é mostrada na Figura 3.2 (a).

Força: Definido para redes ponderadas, é calculada como a soma dos pesos das arestas conectadas a um nó. Pode ser considerada a contraparte de grau para redes ponderadas. Assim, também pode ser interpretada como uma medida de importância de um nó (COSTA et al., 2007). A Figura 3.2 (b) mostra uma rede ponderada e os pesos das arestas correspondentes e as forças dos nós.

Centralidade de intermediação: O comprimento do caminho mais curto entre dois nós é de grande importância, uma vez que fundamenta o conceito de proximidade para nós em uma rede. A centralidade de intermediação de um nó V_i está associada ao número de caminhos mais curtos entre quaisquer dois nós na rede que passam pelo nó V_i (COSTA et al., 2007) normalizado pelo número de caminhos entre os nós. Nós com grande intermediação geralmente podem ser vistos como "pontes" entre diferentes grupos de nós na rede. A Figura 3.2 (c) mostra em vermelho exemplos de nós com alta intermediação.

3.4 Classificação Supervisionada

A resolução de muitos problemas consiste em reconhecer e classificar padrões em classes ou categorias que possuam características distintas. Para tal tarefa, é interessante utilizar um método que se utiliza de protótipos ou exemplos entre os envolvidos e que são pré-determinados por alguma supervisão (JR; COSTA, 2009). Em outras palavras, é necessário ter os dados organizados em conjuntos estruturados a partir das características, medidas e propriedades encontradas nos protótipos e nos dados a serem reconhecidos ou classificados. Alguns problemas são praticamente impossíveis de serem processados manualmente devido à alta quantidade de dados. Nestes casos, é necessário implementar algum método que seja capaz de reconhecer tais padrões de forma sistemática e automatizada. Tais métodos utilizam a chamada classificação supervisionada. Existem diversos tipos de classificadores que podem ser utilizados para solucionar os mais diversos problemas. Desta maneira, é importante a escolha de um classificador versátil que possibilite encontrar os padrões com bons resultados conforme a estrutura dos dados e suas particularidades (AMANCIO et al., 2014).

Neste trabalho é utilizado o clássico classificador k-vizinhos (k_n*–Nearest-Neighbor - KNN*)

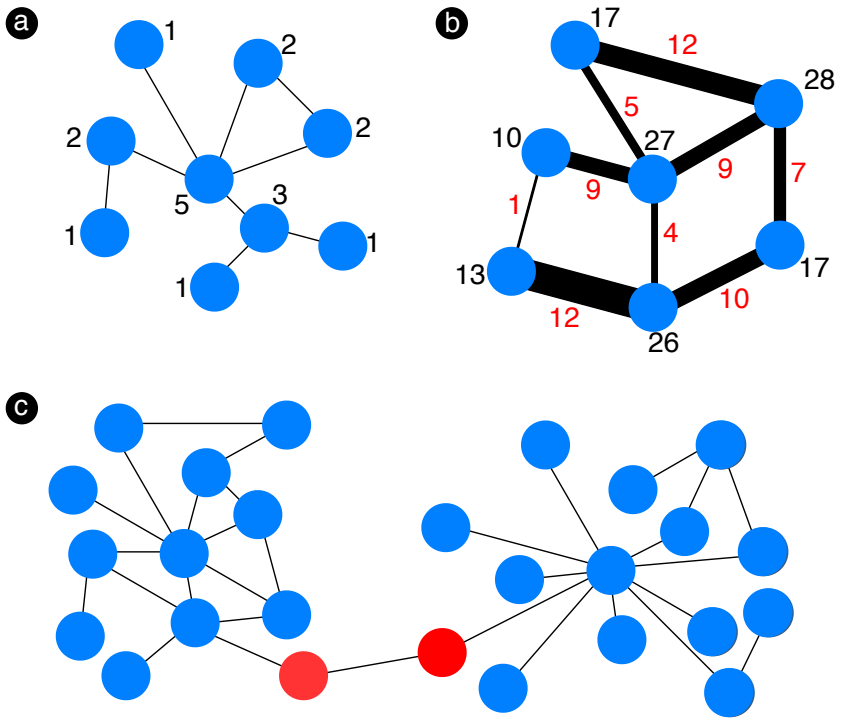

Figura 3.2: Exemplo de propriedades de rede: (a) graus dos nós, indicados por números ao lado de cada nó; (b) pesos das arestas (em vermelho) e respectivas forças dos nós; (c) centralidade de intermediação, nós vermelhos possuem alto valor desta propriedade.

pela sua simplicidade e eficácia. O classificador recebe como entrada um conjunto de dados de referência. O termo *treinamento* na verdade é de certa forma incorreto, pois na verdade não há treinamento do classificador. Os dados de entrada são apenas usados como referência para classificar novos dados. O principal parâmetro do método é o número k de vizinhos analisados durante o processo de inferência. Para cada novo objeto a ser classificado, os k vizinhos mais próximos são analisados, e a classe do objeto é associada à classe mais frequente dentre os vizinhos. Um exemplo de classificação é mostrado na Figura 3.3. Para que o método funcione corretamente, é importante que as propriedades estejam em escalas compatíveis. Normalmente, as propriedades são normalizadas antes de serem utilizadas no KNN.

A distância mais comumente utilizada no KNN é a Euclidiana. Porém existem outros tipos de distâncias, como por exemplo, Bray Curtis, Canberra, correlação de Pearson, cosseno, Manhattan e Hamming que podem ser utilizadas.

Com a finalidade de validar os resultados obtidos pelo KNN também será utilizado o classificador *Support Vector Machine* (SVM) que é um método poderoso, popular e muito utilizado na área de processamento de imagens (BISHOP, 2006; DUDA; HART; STORK, 2012; AMANCIO et al., 2014).

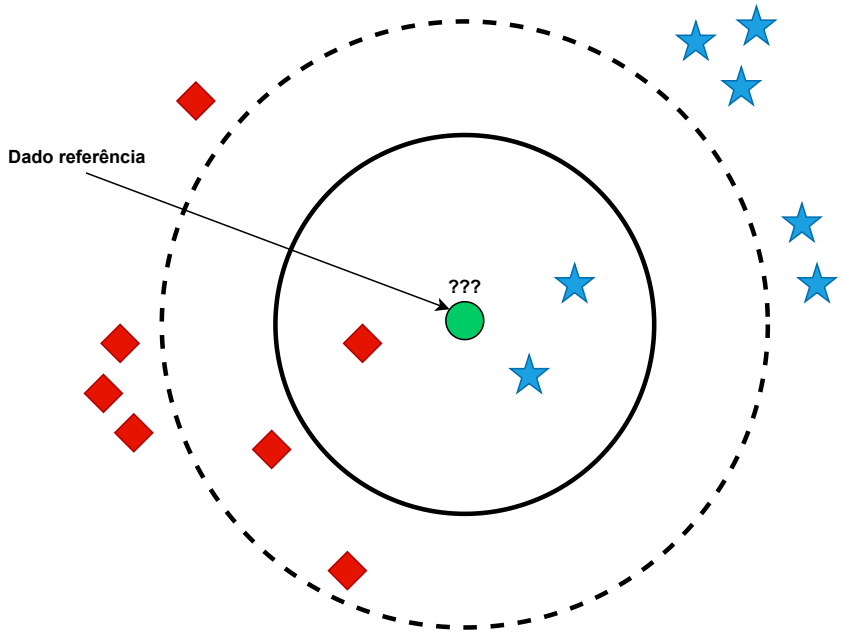

Figura 3.3: Exemplo de classificação feita pelo KNN. O dado referência a ser classificado é indicado pelo círculo verde. Ele deve ser classificado entre losango vermelho ou estrela azul. Se $k = 3$ (círculo com linha sólida) o dado referência é associado à estrela azul, pois, há duas estrelas azuis e apenas um losango vermelho dentro do círculo. Se $k = 5$ (círculo com linha pontilhada) o dado referência é associado ao losango vermelho, pois, há três losangos vermelhos e apenas duas estrelas azuis dentro do círculo.

3.4.1 Quantificação da performance de classificação

Com o intuito de averiguar a performance da classificação realizada, é interessante utilizar a chamada matriz de confusão ou matriz de erro, que em algumas área também é chamada de matriz de correspondência (STEHMAN, 1997). Esta é uma matriz quadrada contendo linhas que representam as classes conhecidas (padrão ouro) e as colunas que representam as predições do classificador. A cada previsão feita pelo classificador o respectivo valor da matriz é incrementado. Assim, o desempenho pode ser facilmente visualizado conforme a matriz se aproxima de uma matriz diagonal (JR; COSTA, 2009). No caso de um problema de classificação binário, isto é, envolvendo duas classes, os elementos da matriz de confusão são definidos da seguinte forma:

- TP: Quanto um objeto Positivo é classificado verdadeiramente como Positivo;
- FP: Quando um objeto Negativo é classificado falsamente como Positivo;
- FN: Quando um objeto Positivo é classificado falsamente como Negativo;
- TN: Quando um objeto Negativo é classificado verdadeiramente como Negativo.

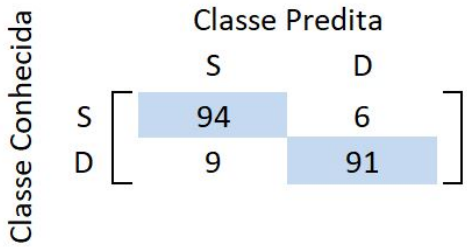

Figura 3.4: Ilustração da matriz de confusão com dados de glândulas de PCa. As classes são descritas como S - Saudável e D - Doente. As células azuis enfatizam a diagonal da matriz.

A Figura 3.4 ilustra um exemplo de matriz de confusão para o caso de classificação de glândulas da próstata em Saudável (S) e Doente (D). No caso, e ao longo do trabalho, a classe Doente estará associada com o resultado Positivo e a classe saudável ao resultado Negativo. No exemplo, um total de 200 elementos são preditos pelo classificador. As células em azul na figura indicam a matriz diagonal.

É possível realizar análises mais detalhadas do que apenas a acurácia que mede a proporção de predições corretas. Dependendo da quantidade de observações e o equilíbrio entre a variação das classes a acurácia pode produzir resultados enganosos. Por exemplo, se houver 95 glândulas doentes e apenas 5 saudáveis nos dados, algum determinado classificador pode predizer todas as observações como doentes. Com esta sensibilidade deste classificador temos uma precisão de 100% para glândulas doentes e 0% para glândulas saudáveis sendo que a precisão geral seria de 95%. Desta maneira, algumas métricas podem ser analisadas com as combinações das 4 medidas da matriz de confusão (TP, FP, TN, FN). Algumas delas são:

Precision: também conhecido como confiança no contexto de mineração de dados, é a proporção de casos TP comparada com o número de positivos retornados pelo classificador. Matematicamente, a medida é definida como

$$Precision = \frac{TP}{PP} = \frac{TP}{TP+FP}, \tag{3.3}$$

onde PP é o número de positivos retornados pelo classificador. Portanto, para que essa medida possua valores altos, é necessário que o classificador não retorne muitos falsos positivos.

Recall: é a proporção dos casos positivos que foram corretamente preditos como positivos. A métrica tem o intuito de refletir a quantidade de casos relevantes que o classificador conseguiu identificar. Ela é definida como

$$Recall = \frac{TP}{P} = \frac{TP}{TP+FN}, \tag{3.4}$$

onde P é o número de elementos positivos. Portanto, elementos positivos não identificados pelo classificador diminuem o valor desta propriedade.

Specificity: mede a proporção de negativos que foram corretamente classificados. É calculada por TN dividido pelo somatório da condição negativa N, ou seja,

$$Specificity = \frac{TN}{N} = \frac{TN}{TN+FP}. \tag{3.5}$$

Accuracy: é o acerto geral do classificador, que pode ser definido como

$$Accuracy = \frac{TP+TN}{Total} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.6}$$

As combinações das medidas *Precision* e *Recall* são focadas apenas nas previsões positivas, mas também capturam algumas informações sobre taxas e tipos de erros gerados. Métodos capazes de obter simultaneamente altos valores dessas duas propriedades são considerados de alta qualidade. Nenhuma destas duas medidas lidam com qualquer informação sobre casos negativos. Por isso também é interessante utilizar a medida *Specificity* .

Durante a etapa de classificação, é interessante evitar o ajuste enviesado do modelo ao conjunto de dados utilizado no treinamento, que é chamado de *over-fitting*. Desta maneira, é importante ter algum método que evite o *over-fitting* no modelo estatístico utilizado e que seja eficaz para prever novos resultados.

Neste trabalho é utilizado o método de validação cruzada *cross-validation* que permite uma proporção $(S-1)/S$ do conjunto de dados a ser utilizado para treinamento enquanto faz o uso de todos os dados para avaliar a performance. A validação cruzada *S-fold* utiliza-se da técnica de particionar os dados em S grupos. Então $S-1$ grupos são utilizados para treinar o classificador, que é então avaliado no grupo restante. Este procedimento é repetido por todas as S possibilidades restantes (BISHOP, 2006). A figura 3.5 exemplifica a validação cruzada utilizando $S = 5$.

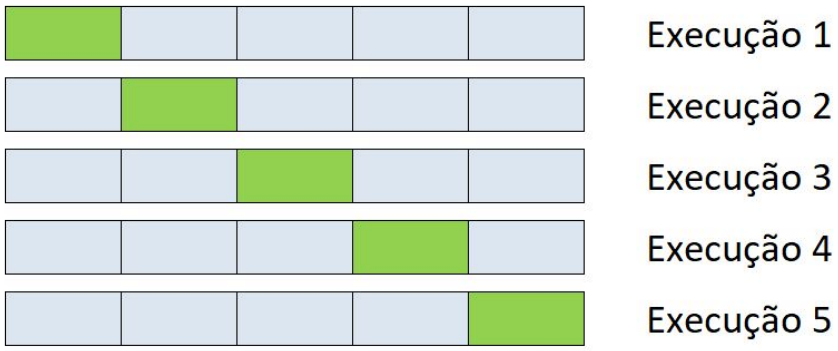

Figura 3.5: Ilustração da validação cruzada com $S = 5$, isto é, dados particionados em S grupos. O treinamento é feito com $S - 1$ grupos, representados pelos retângulos em azul. A classificação é realizada com o grupo em verde. A soma ou média da performance nas 5 execuções é então calculada.

Capítulo 4

METODOLOGIA

Este capítulo tem como objetivo motivar e descrever a metodologia de criação de redes de contexto com o intuito de caracterizar glândulas para a detecção de PCa.

4.1 Importância do Contexto na Detecção do Câncer de Próstata

Conforme apresentado no Capítulo 2, as glândulas de PCa podem ser caracterizadas usando medidas de forma e textura. No entanto, o Gleason Grading System (GGS) afirma que a aparência da glândula não deve ser considerada isoladamente (KUMAR et al., 2014; NGUYEN; SARKAR; JAIN, 2012; EPSTEIN et al., 2005, 2005), ou seja, o contexto da glândula também é importante. Por exemplo, uma única glândula com uma aparência que pode estar associada a glândulas saudáveis, se cercada por glândulas indiferenciadas, na verdade estará associada a um alto índice de Gleason. Por tais razões, algumas metodologias para definição de contexto foram desenvolvidas (NGUYEN; SARKAR; JAIN, 2012; DOYLE et al., 2007; MONACO et al., 2008). Esta definição contextual se torna importante no momento o qual não é mais observado apenas uma glândula isoladamente, mas sim, em qual contexto ela se enquadra e como ela é influenciada por outras glândulas. Por exemplo, os autores Nguyen et. al (NGUYEN; SARKAR; JAIN, 2012) exploram as seguintes informações contextuais: aglomeração de vizinhança, similaridade de forma e similaridade de tamanho. Para calcular os atributos contextuais, eles consideram, para cada glândula, uma região de raio $R < 65$ em torno da glândula. Ainda assim, o contexto geralmente não é levado em consideração de maneira sistemática, ou é considerado apenas para propriedades específicas.

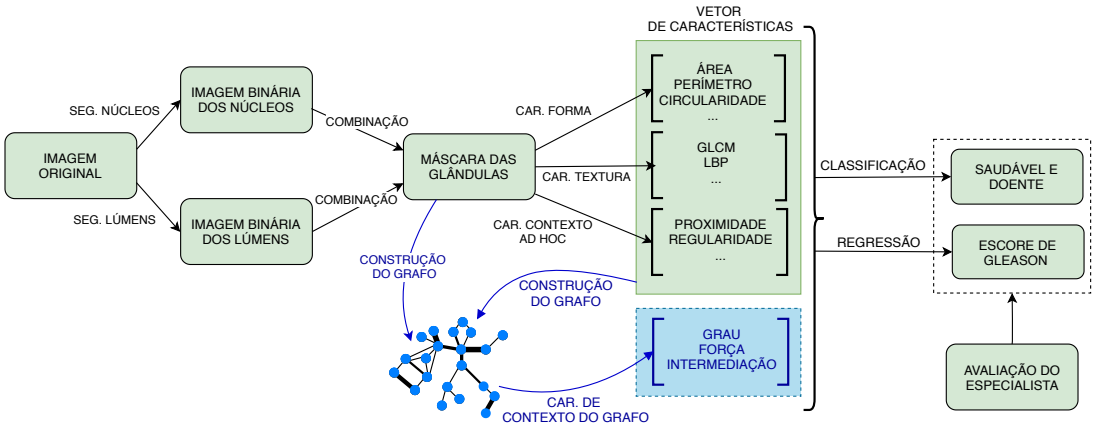

Figura 4.1: Fluxograma da figura 2.2 modificado com a metodologia proposta, de forma a sistematizar a criação de medidas de contexto.

4.2 Rede de Contexto Glandular

Baseando-se na necessidade mencionada por Nguyen et. al e também por outros autores de se definir características de contexto para caracterizar imagens PCa, este trabalho explora uma metodologia para a construção de um vetor de características que agregue informações contextuais em imagens de PCa. Desta maneira, foi definida uma abordagem para criação de uma rede complexa, aqui chamada de *Gland Context Network (GCN)* onde cada glândula é representada por um vértice e as adjacências entre os vértices indicam algum relacionamento entre as glândulas. Esta relação pode ser espacial, no sentido ao qual as distâncias entre as glândulas são utilizadas durante a definição de conectividade, ou pode ser também baseada na forma ou textura das glândulas, o que significa que glândulas similares tendem a estar conectadas. É possível também utilizar uma mistura dos dois aspectos. A rede obtida pode então ser caracterizada utilizando propriedades de rede (COSTA et al., 2007), as quais, são medidas adicionais para utilização durante o diagnóstico de PCa. O fluxograma da metologia é apresentado na figura 4.1.

Para a criação da GCN, é interessante que glândulas mais próximas possuam maior chance de serem conectadas e que glândulas mais parecidas possuam maior relacionamento no grafo. Por exemplo, a figura 4.2 ilustra um possível grafo de contexto, nota-se que cada adjacência criada possui um peso diferente que é obtido a partir do vetor de características de cada glândula. Nesse caso, a similaridade da área entre as glândulas foi utilizada na ponderação das arestas, criando o que podemos chamar de GCN de área.

A caracterização da topologia resultante, considerando a ponderação das arestas, leva à imediata contextualização da área de cada glândula. Por exemplo, a força de cada nó, dada pela

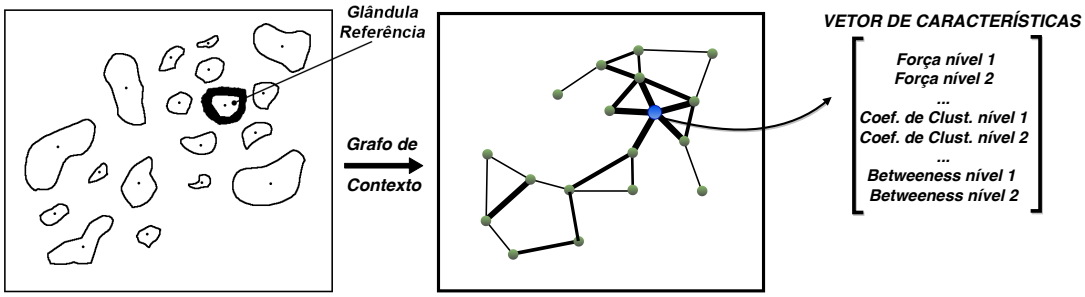

Figura 4.2: Ilustração de uma GCN de área. Glândulas próximas são conectadas, e as arestas são ponderadas pela similaridade entre as áreas das glândulas. Medidas de rede podem então ser calculadas para cada glândula.

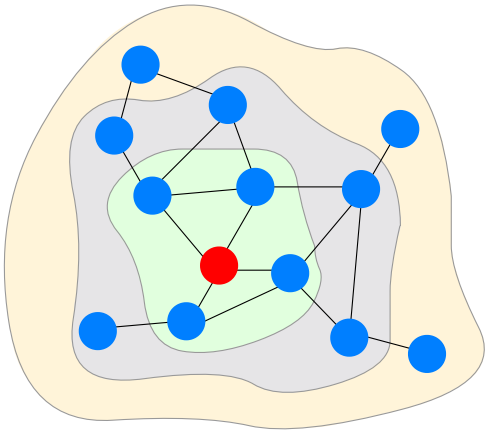

Figura 4.3: Ilustração dos níveis hierárquicos de uma rede a partir de um nó referência (vermelho).

soma dos pesos das arestas conectadas ao nó, indica a similaridade entre a área da glândula de referência e as áreas dos vizinhos imediatos desta glândula. O coeficiente de clusterização ponderado indica a similaridade entre as áreas dos vizinhos. Adicionalmente, não é preciso se limitar aos vizinhos imediatos à glândula de referência, qualquer vizinhança (ou nível) pode ser utilizada, levando à caracterização sistemática do contexto em diferentes escalas de análise, as quais, podem ser consideradas para caracterizar o nó da rede. Estas diferentes escalas podem ser observadas na Figura 4.3. A região verde indica os vizinhos imediatos ao nó de referência (nó vermelho), as regiões roxa e laranja contém nós que são respectivamente, segundos e terceiros vizinhos na escala do nó de referência.

Para a criação da GCN, é possível utilizar apenas uma medida específica, criando um grafo de contexto daquela propriedade, ou até mesmo conjuntos de medidas, criando contextos mais gerais. Por exemplo, o uso de um conjunto de medidas de forma pode levar a criação de um grafo de contexto de forma.

A criação de uma GCN é feita da seguinte forma. Seja S o conjunto de glândulas na amostra,

onde cada glândula i é referenciada como s_i. Suponha que m características de forma e textura $f_1^i, f_2^i, \ldots, f_m^i$ foram obtidas para cada glândula s_i. Essas características podem ser representadas por um vetor $\vec{f}^i$. Cada glândula também tem uma posição $\vec{p}^i = (p_1^i, p_2^i)$. Consideraremos que a posição é dada pelo centroide da glândula.

A distância espacial entre duas glândulas i e j é calculada por:

$$d_{ij}^s = \sqrt{\sum_{k=1}^{2} (\tilde{p}_k^i - \tilde{p}_k^j)^2} \tag{4.1}$$

onde $\tilde{p}_k^i$ é o escore padrão da coordenada k da glândula i, dado por

$$\tilde{p}_k^i = \frac{p_k^i - \mu_k}{\sigma_k} \tag{4.2}$$

onde μ_k e σ_k são, respectivamente, a média e o desvio padrão da coordenada k de todas as glândulas. O intuito da normalização é permitir uma definição de distância geral envolvendo ambas coordenadas espaciais das glândulas e suas características de forma e textura.

A diferença entre as características de duas glândulas i e j é calculada por

$$d_{ij}^f = \sqrt{\sum_{k=1}^{m} (\tilde{f}_k^i - \tilde{f}_k^j)^2} \tag{4.3}$$

onde $\tilde{f}_k^i$ representa o escore padrão de f_k^i, definido utilizando a mesma abordagem descrita anteriormente.

A combinação de distância entre as glândulas e suas similaridades de forma e textura é calculada como

$$h_{ij} = \sqrt{\alpha \left(d_{ij}^s\right)^2 + (1-\alpha)\left(d_{ij}^f\right)^2}, \tag{4.4}$$

O peso da conexão entre as glândulas i e j, é então, dado por

$$w_{ij} = e^{-h_{ij}}. \tag{4.5}$$

O parâmetro α ajusta a importância relativa da distância espacial e a similaridade das características entre as glândulas. Quando $\alpha = 1$, apenas a distância espacial é levada em conta na ponderação, enquanto que $\alpha = 0$ dá origem a pesos definidos apenas pela aparência das glândulas. É importante salientar que o contexto da rede pode ser criado para todas as características

ou utilizando apenas um sub-conjunto delas.

Uma GCN pode ser criada utilizando diferentes combinações de distâncias d_{ij}^{s} e d_{ij}^{f}. Por exemplo, é possível usar um valor de α para definição de conectividade entre os nós e outro valor para a definição do peso das arestas. No restante deste trabalho, será considerada uma abordagem para gerar uma rede que seja fácil de visualizar e interpretar. As conexões entre os nós serão definidas utilizando a abordagem do grafo geométrico. Desta maneira, a conectividade é definida utilizando $\alpha = 1$ e um dado raio R que é um parâmetro do método. Na sequência, os pesos das arestas são definidos utilizando $\alpha = 0$, ou seja, considerando apenas as características de forma e textura das glândulas. Isso gera a rede onde apenas as distâncias entre as glândulas definem a existência ou ausência de arestas e o peso das arestas depende apenas da similaridade entre a aparência da glândula. É importante ter em mente que GCNs permitem abordagens alternativas para a construção da conectividade.

Capítulo 5

RESULTADOS OBTIDOS

Este capítulo tem como principal objetivo sintetizar os resultados obtidos a partir das GCNs criadas no estudo de caso e as contribuições para o estado-da-arte com exemplos de identificação de PCa.

5.1 Base de imagens *whole-mount*

Foi elaborado um estudo de caso como forma de validação para a metologia elaborada durante os experimentos da pesquisa em conjunto com especialistas da área médica com amostras *whole-mount* da próstata demarcadas e fornecidas por eles. O conjunto de dados utilizado no estudo de caso consiste em seções de duas amostras *whole-mount* da próstata a partir de modelos de coloração com hematoxilina e eosina (H&E). Os slides são gerados por um microscópio automático de campo claro, que é configurado para percorrer toda a amostra e formar um grande mosaico de imagens composto de 40 a 50 imagens individuais. A figura 5.1(a) mostra parte de uma das imagens do conjunto de dados. Especialistas analisaram a amostra e identificaram duas regiões avaliadas com pontuação 3+3 no GGS para PCa que estão demarcadas em vermelho na imagem. A figura 5.2(a) mostra uma segunda imagem do conjunto de dados que também foi identificada por especialistas. Esta, contém apenas uma região demarcada em vermelho e avaliada com pontuação 3+3 no GGS para PCa.

5.2 Segmentação automatizada de glândulas

Foi implementada uma variação da proposta de Nguyen et al. (NGUYEN; SARKAR; JAIN, 2012), que utilizou segmentação com K-médias para a identificação das cores das imagens associadas com cada tipo de estrutura. A figura 5.3 mostra uma parte da imagem *whole-mount*

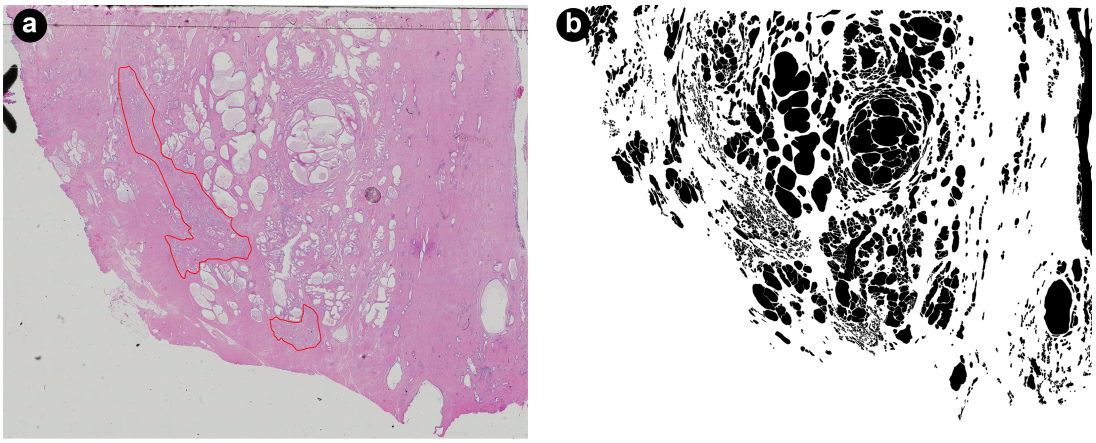

Figura 5.1: Primeira imagem utilizada no estudo de caso. (a) Imagem *whole-mount* da próstata com regiões de pontuação 3+3, conforme o GGS, demarcadas em vermelho. (b) Imagem binária que mostra a segmentação manual das glândulas.

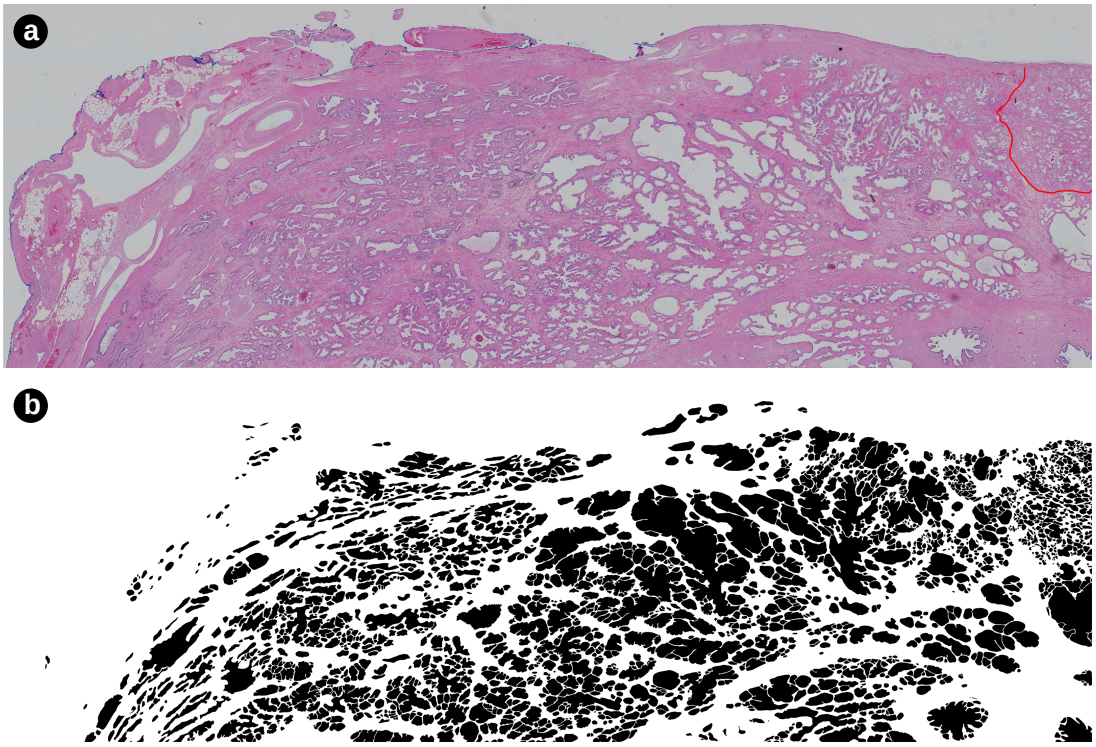

Figura 5.2: Segunda imagem utilizada no estudo de caso. (a) Imagem *whole-mount* da próstata com região de pontuação 3+3, conforme o GGS, demarcada em vermelho. (b) Imagem binária que mostra a segmentação manual das glândulas.

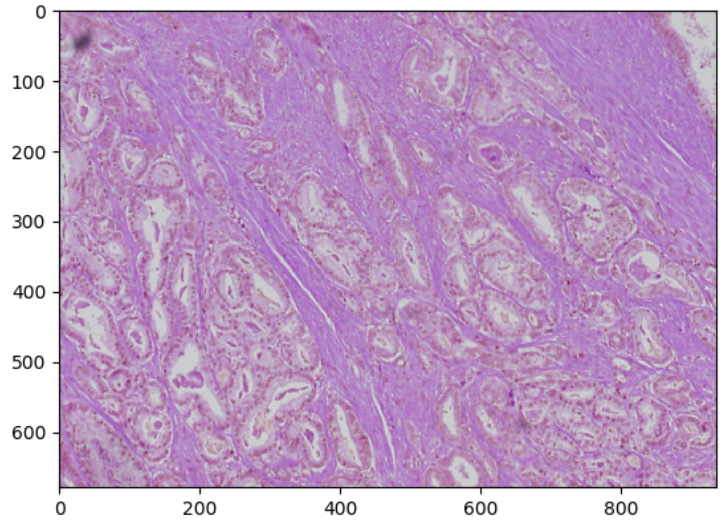

Figura 5.3: Imagem analisada para a definição de um método de segmentação automatizado.

que foi inicialmente explorada durante a identificação de núcleos e lúmens. Para a detecção dos lúmens, o método K-médias com $K = 4$ foi aplicado no espaço de cores CIE L*a*b*. Dentre os grupos identificados, o grupo possuindo o centroide com a maior intensidade foi associada ao lúmen das glândulas. O resultado pode ser observado na figura 5.4.

Para a identificação dos núcleos, cada glândula foi analisada isoladamente em uma janela com tamanho proporcional à área da glândula. Então, iniciou-se a busca pelos núcleos ao seu redor. A fim de se obter o melhor resultado foram considerados 8 níveis de cores para utilização do K-médias neste caso. A segmentação dos núcleos resulta em alguns núcleos sobrepostos e não tão bem segmentados. Desta maneira, com o intuito de identificar melhor esses núcleos a segmentação obtida foi aprimorada utilizando o método Divisor de Águas (*Watershed*) (WÄHLBY et al., 2004).

Os lúmens e núcleos segmentados podem então ser combinados para gerar uma única imagem. Um exemplo de resultado pode ser observado na figura 5.5. Na figura 5.5(a) é possível visualizar o lúmen detectado em uma janela proporcional ao seu tamanho. Os núcleos detectados na mesma região são mostrados na figura 5.5(b). Na figura 5.5(c) é mostrada a junção duas imagens anteriores, formando a glândula. Para efeito de comparação, a glândula original é mostrada na figura 5.5(d). Adicionalmente, alguns artefatos identificados na segmentação dos lúmens podem ser removidos através de uma heurística simples. Lúmens em geral possuem núcleos ao seu redor. Portanto, regiões detectadas como lúmen mas que não possuem um número significativo de núcleos podem ser consideradas artefatos e removidas da imagem.

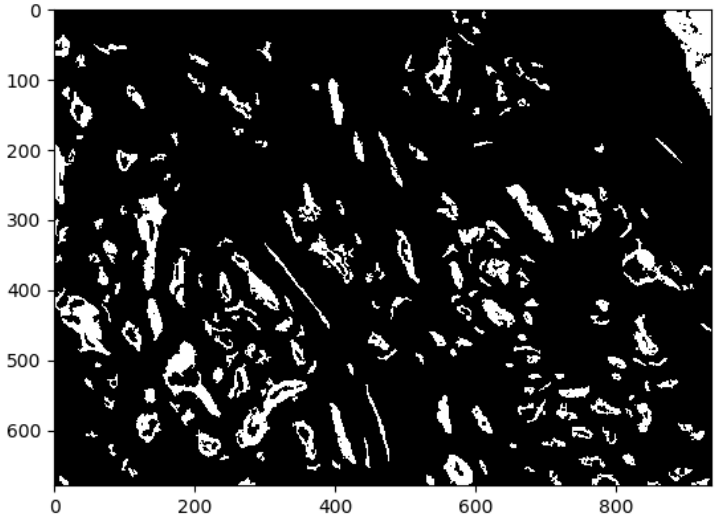

Figura 5.4: Lúmens detectados na imagem mostrada na figura 5.3.

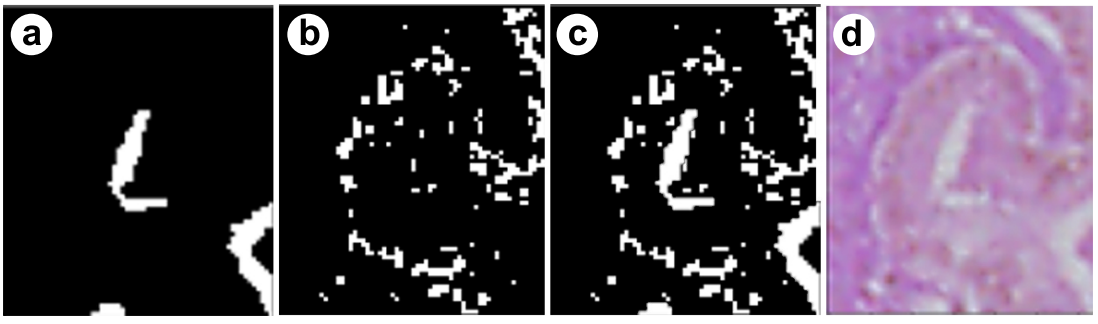

Figura 5.5: Ilustração dos resultados obtidos para a segmentação automática. (a) lúmen segmentado. (b) núcleos detectados. (c) junção do lúmen e do núcleo. (d) glândula original.

Após diversos experimentos e tentativas de otimização de parâmetros, foram obtidos resultados razoáveis com a segmentação automática. Ainda assim, foi observado que muitas glândulas não foram segmentadas corretamente. Tais problemas de segmentação teriam influência no método proposto. Como o foco deste trabalho não é a segmentação, e sim a aplicação e validação do método proposto, decidimos realizar a segmentação das glândulas de forma manual, com o auxílio e supervisão de especialistas.

5.3 Segmentação manual de glândulas

Como mencionado, o foco deste trabalho é a investigação do potencial das redes complexas para fornecer suporte aos patologistas durante a avaliação de tecidos anormais da próstata. Para isso, é interessante remover qualquer influência causada por problemas de segmentação das glândulas. Como a segmentação automatizada ainda é um problema em aberto, a segmentação ideal das glândulas pode ser realizada somente de forma manual. Desta forma, foi possível definir critérios e validações da marcação das glândulas juntamente com especialistas, evitando qualquer viés durante o processo.

As glândulas marcadas manualmente estão mostradas nas figuras 5.1(b) e 5.2(b). No total foram marcadas 5479 glândulas, o que foi julgado suficiente para avaliar o potencial das GCNs em representar e caracterizar as imagens da próstata.

5.4 Criação da Rede de Contexto Glandular

A criação da GCN é iniciada pela identificação de cada glândula demarcada na máscara que se torna uma região de interesse. Cada região possui um centro de massa que é identificado numa posição espacial (x, y) da imagem. Os centros de massa definem as coordenadas dos nós da GCN. Para referência, a figura 5.6(a) mostra as glândulas identificadas juntamente com os respectivos centros de massa marcados em vermelho. Uma imagem mostrando apenas os centros de massa é mostrada na figura 5.6(b). Desta maneira, é possível criar a CGN mantendo a relação espacial das glândulas nas imagens utilizadas. Também é possível identificar as características de formas, redes e suas relações espaciais que sejam relevantes durante a identificação do PCa. Durante a visualização geral dos centros de massa demonstrado pelos pontos vermelhos em cima da máscara na figura 5.6(a), fica claro que a hipótese criada no capítulo 4 tem um potencial muito grande. A alta densidade dos centros de massa é nítida próxima a demarcação feita pelo especialista (mostrada na figura 5.1(a)).

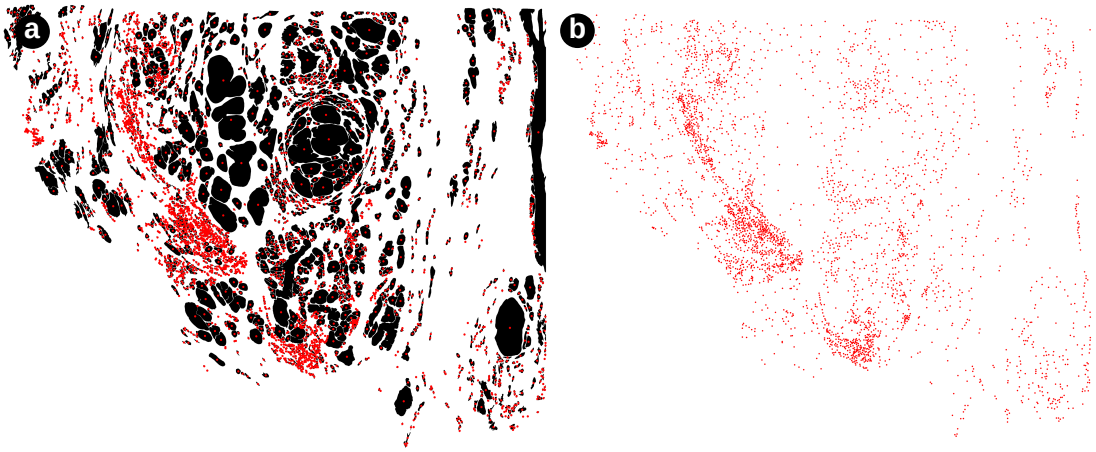

Figura 5.6: Representação dos centros de massa de cada glândula (em vermelho).

Após a visualização dos nós sobre a máscara das glândulas e a observação das densidades próximas à região demarcada pelos especialistas é necessário a identificação de algum método que crie da melhor forma possível as arestas. Também é importante levar em consideração a motivação biológica encontrada na literatura, que foi a base da hipótese, somado ao *ground truth* obtido a partir da demarcação feita pelos especialistas. Assim sendo, foram criadas algumas redes geométricas com o intuito de observar a inluência do raio de alcance na criação das conexões. A figura 5.7 demonstra essas visualizações produzidas. É possível observar em (a) a rede geométrica gerada com raio de 100 pixels, em (b) raio de 200 pixels, em (c) raio de 300 pixels e em (d) raio de 350 pixels.

O aumento do raio provoca um crescimento significativo do número de arestas na rede, o que faz com que seja mais custoso calcular diversas propriedades de rede. Por outro lado, um raio muito pequeno faz com que a rede fique muito desconexa. Portanto, o ideal é que valores intermediários de raio sejam utilizados para a criação da rede. Outro possível critério é utilizar o raio que leve à melhor detecção de PCa na amostra. Esse critério será aplicado na próxima seção. Para os resultados seguintes apresentados nesta seção utilizaremos um raio de 350 pixels.

Na figura 5.8 são mostradas regiões específicas do grafo criado. A principal diferença a ser notada nas regiões mostradas está relacionada com o número de arestas criadas, que tendem a acompanhar a densidade de glândulas nas regiões. Dado o número de arestas envolvidas, é difícil notar visualmente outras diferenças. Por outro lado, é importante notar que a rede mostrada possui apenas informação sobre a conectividade entre as glândulas, por isso a chamamos de rede geométrica. A GCN é criada a partir da ponderação das arestas desta rede.

Para gerar a GCN, diversas características de forma foram obtidas para as glândulas estudadas. A lista de propriedades obtidas está apresentada e discutida na seção 3.2. É impor-

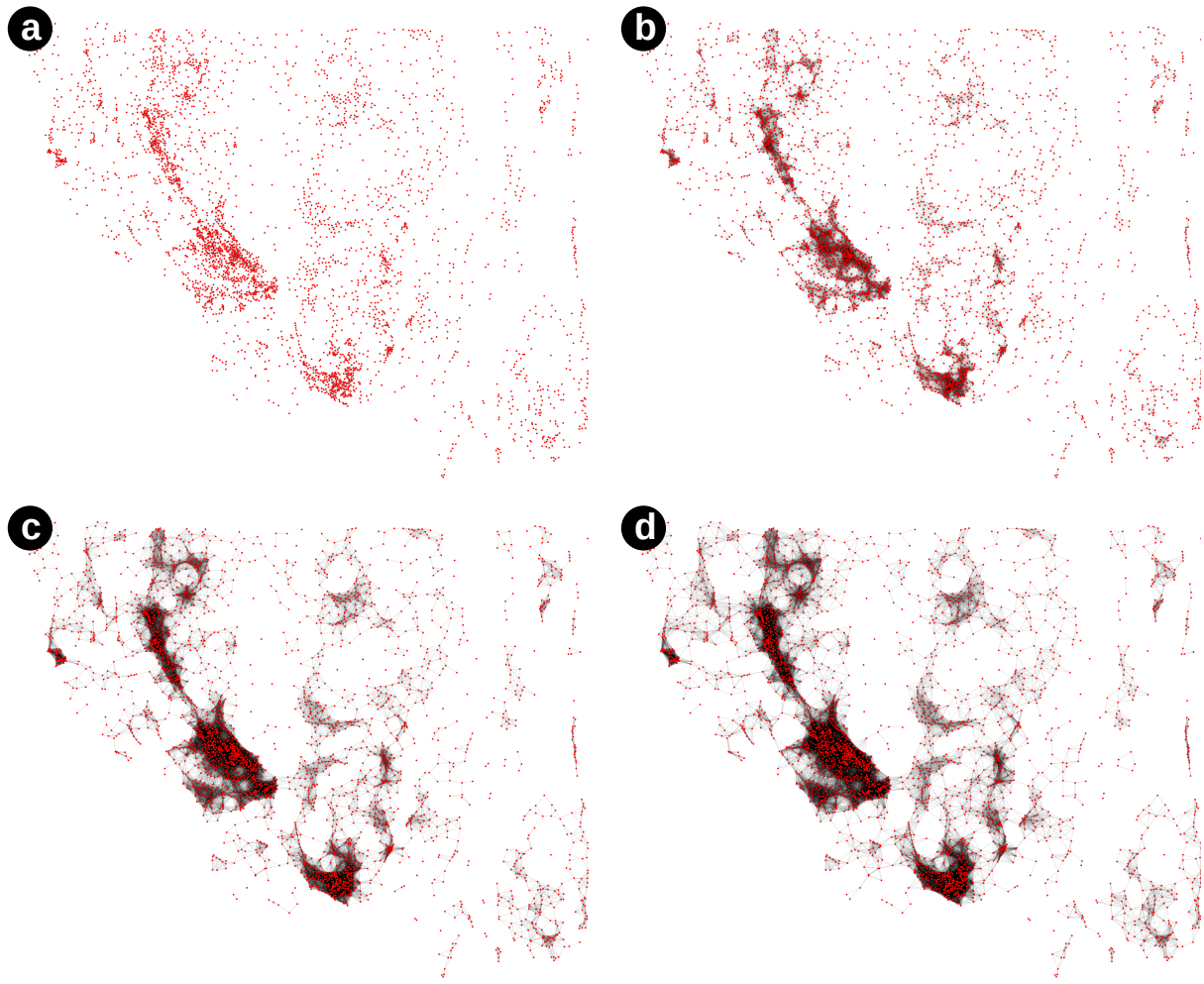

Figura 5.7: Observação da rede geométrica com diferentes raios. A figura (a) possui raio de 100 pixels, a figura (b) possui raio de 200, a figura (c) possui raio de 300 pixels e a figura (d) possui raio de 350 pixels.

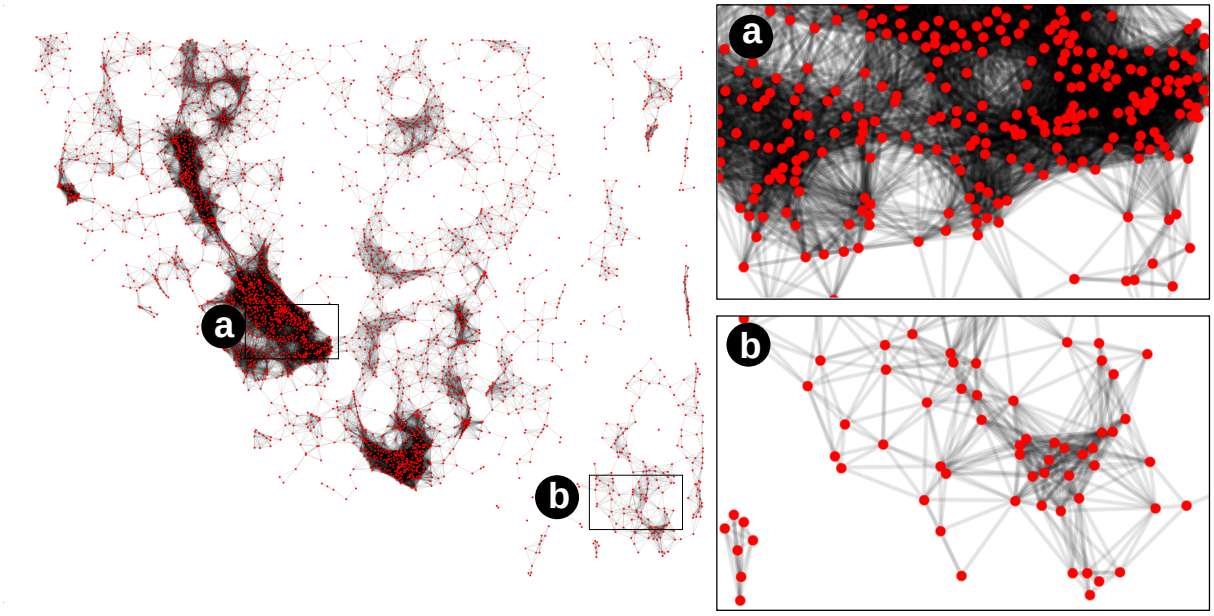

Figura 5.8: (a) Rede geométrica gerada a partir da imagem mostrada na figura 5.1. Visualizações ampliadas da rede são mostradas em (a) e (b). Um raio de $r = 350$ pixels foi usado para gerar a rede.

tante salientar que outras propriedades poderiam ser utilizadas para gerar a ponderação. Em particular, dependendo do tipo de imagem e processamento utilizado, propriedades de textura possuem interessante potencial em identificar glândulas atingidas por PCa (MOSQUERA-LOPEZ et al., 2015).

Tendo obtido as propriedades, é possível gerar a ponderação das arestas utilizando as equações 4.4 e 4.5. Como já mencionado, utilizamos $\alpha = 0$ nos experimentos para poder separar os dois aspectos da criação da rede: a definição da presença ou não das arestas é dada somente pela posição espacial, enquanto que a ponderação é indicada pela forma das glândulas. Outras variações podem ser utilizadas.

Inicialmente, foi gerada uma GCN para cada propriedade analisada. As redes geradas estão mostradas nas figuras 5.9 (área), 5.11 (perímetro), 5.10 (diâmetro), 5.12 (solidez) e 5.13 (excentricidade). As figuras mostram as redes geradas para as duas regiões indicadas na figura 5.8, pois a inclusão da rede inteira neste texto ficaria com uma visualização comprometida. A largura das arestas está relacionada com a similaridade entre as glândulas. Arestas mais grossas indicam glândulas mais similares. Verifica-se que glândulas próximas possuindo propriedades similares tendem a possuir arestas mais grossas. É interessante observar a criação de grupos de glândulas similares e próximas. Por exemplo, na figura 5.9 vemos um grupo de glândulas bem conectadas na região direita inferior da figura. Esse mesmo grupo possui conectividade bem mais heterogênea nos casos das propriedades *Solidity* (figura 5.12) e *Eccentricity* (figura 5.13).

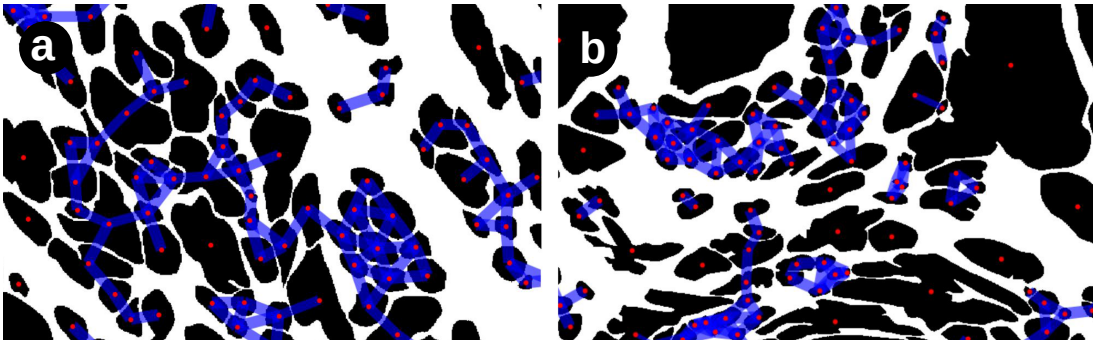

Figura 5.9: Visualização da ponderação da propriedade de Área na região (a) com distância 50 pixels e região (b) com distância 100 pixels.

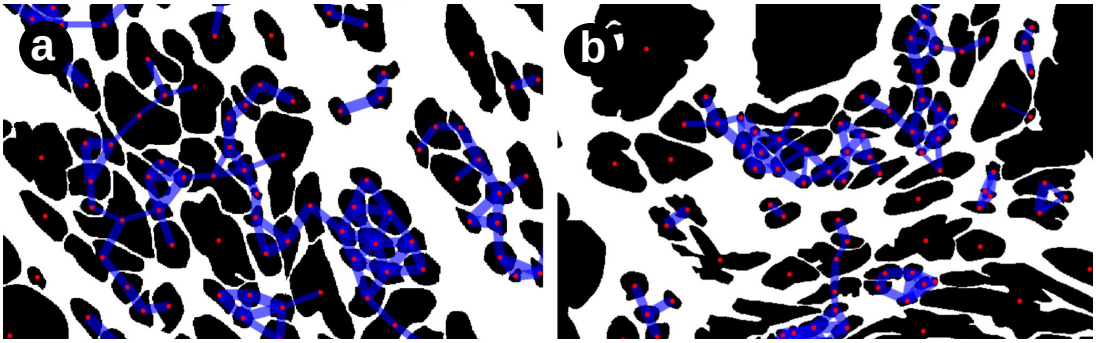

Figura 5.10: Visualização da ponderação da propriedade de Diâmetro na região (a) com distância 50 pixels e região (b) com distância 100 pixels.

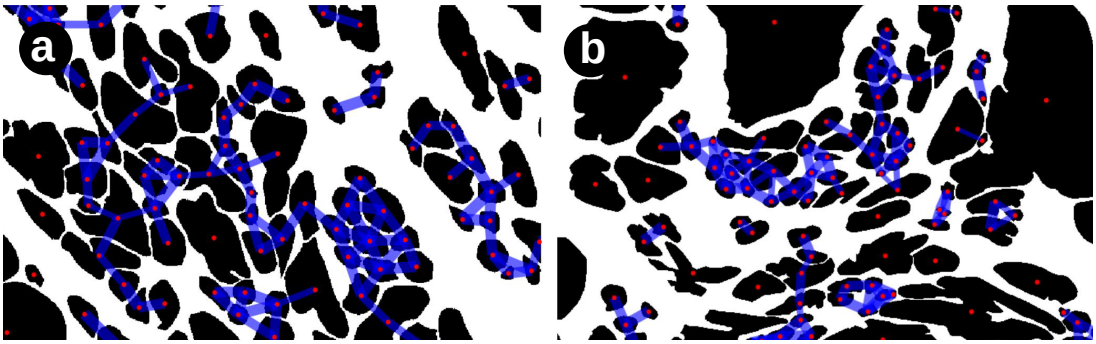

Figura 5.11: Visualização da ponderação da propriedade de Perímetro na região (a) com distância 50 pixels e região (b) com distância 100 pixels.

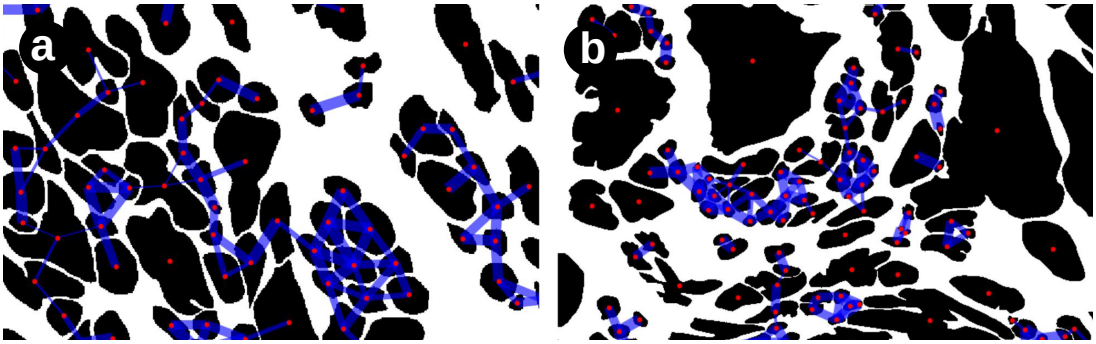

Figura 5.12: Visualização da ponderação da propriedade de *Solidity* na região (a) com distância 50 pixels e região (b) com distância 100 pixels.

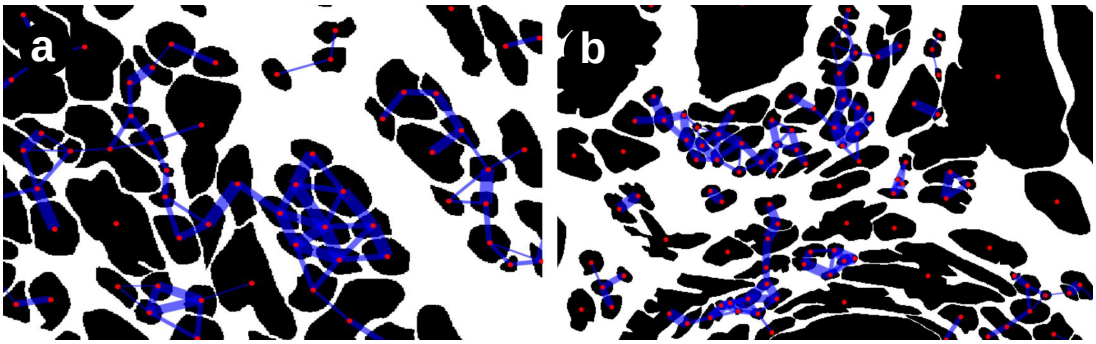

Figura 5.13: Visualização da ponderação da propriedade de *eccentricity* na região (a) com distância 50 pixels e região (b) com distância 100 pixels.

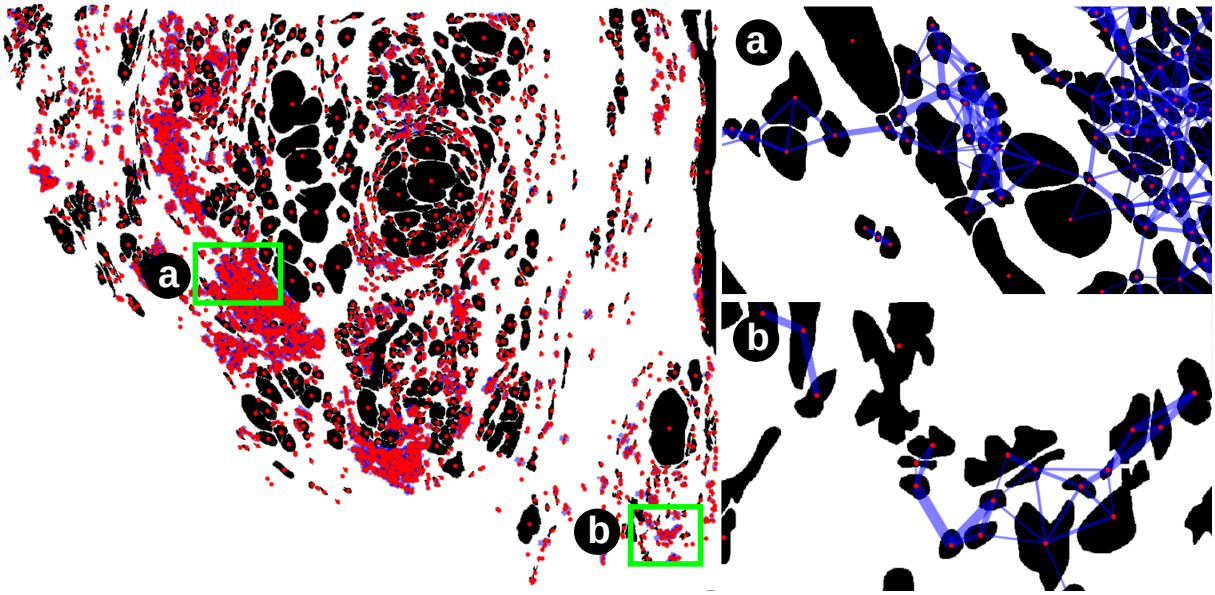

Figura 5.14: Visualização do grafo gerado utilizando todas as medidas de forma na ponderação das arestas com raio de 100 pixels. As regiões (a) e (b) possuem suas visualizações ampliadas. A espessura das arestas indica a similaridade de forma entre as glândulas.

Na figura 5.14 é mostrada a GCN gerada considerando simultaneamente todas as medidas de forma. É possível verificar uma heterogeneidade grande nas larguras as arestas, causada pela variação de forma entre glândulas vizinhas.

Tendo obtido a rede é possível então extrairmos propriedades topológicas da conectividade entre as glândulas, que estão relacionadas com o contexto no qual cada glândula está inserida. Como ilustração, calculamos o grau e o coeficiente de intermediação de cada glândula. Os valores calculados estão mostrados nas figuras 5.15 e 5.16. A rede utilizada para o cálculo envolve todas as medidas de forma. Enquanto que o grau proporciona um contexto local, o coeficiente de intermediação é influenciado por todas as glândulas da amostra, e portanto é mais difícil de ser interpretado.

Vemos que nós de grau maior tendem a estar localizados dentro ou próximos a região demarcada pelo especialista (mostrado na Figura 5.1), sugerindo uma relação entre o grau do nó e o PCa. No caso do coeficiente de intermediação, glândulas que estão entre regiões de alta densidade ou na borda de grupos de glândulas tendem a possuir valores maiores da medida.

5.5 Detecção de PCa

Para verificar o potencial da metodologia desenvolvida em aprimorar a performance da detecção de PCa, nesta seção utilizamos propriedades obtidas a partir da GCN para identificar

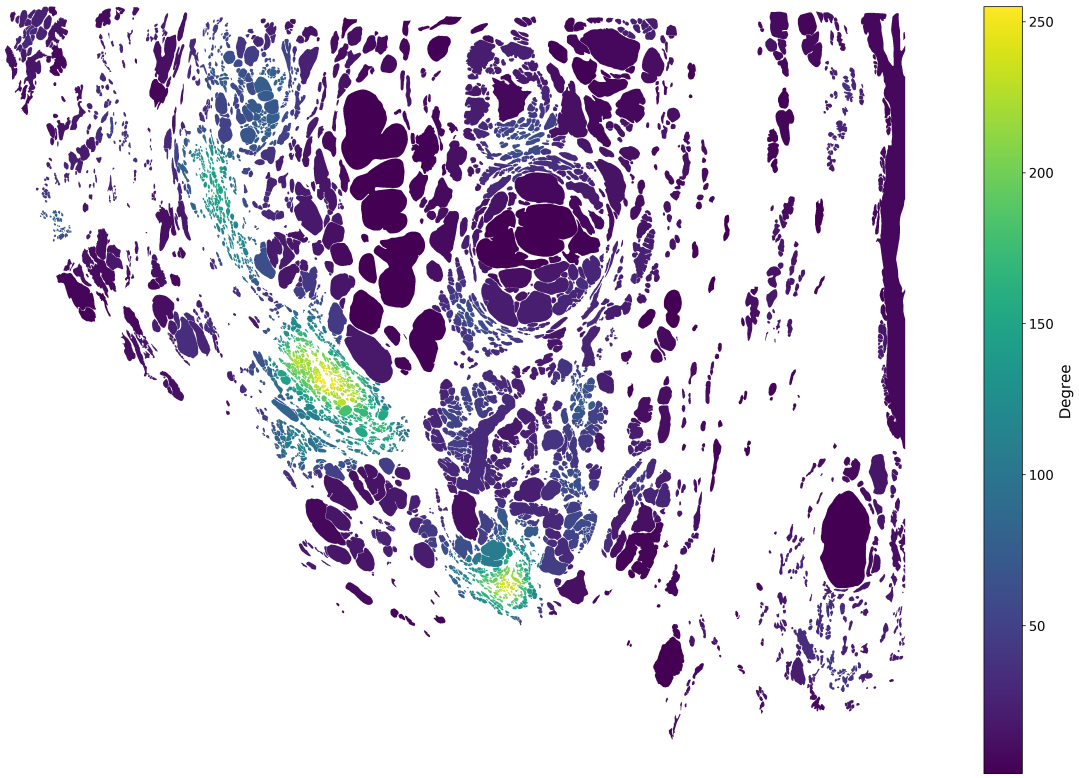

Figura 5.15: Visualização dos graus dos nós da GCN. Cada glândula é colorida de acordo com o grau de seu respectivo nó.

glândulas afetadas por PCa. Comparamos os resultados obtidos com o caso no qual apenas propriedades locais de forma são utilizadas.

O procedimento envolve duas amostras *whole mount* da próstata possuindo regiões de PCa demarcadas por um especialista. Dessa forma, temos o padrão ouro das glândulas nas categorias doente e saudável. O algoritmo KNN foi aplicado para classificar as glândulas nas duas categorias possíveis. Para verificarmos a performance do procedimento, foi utilizada validação-cruzada com 5 *folds*.

Como primeiro passo, é preciso selecionar um raio R adequado para criação da rede. Para isso, foram criadas redes possuindo diferentes valores de raio e a acurácia para cada raio foi calculada. A figura 5.17 apresenta um resultado típico obtido. No caso, foram criadas GCNs a partir de todas as medidas de forma e as propriedades utilizadas para a classificação incluem as medidas de forma e também as medidas de rede grau, força, e coeficiente de intermediação. Pelo resultado, é possível verificar que valores de raio entre 300 e 500 levam às melhores classificações. É interessante que seja utilizado um valor pequeno de raio para que a rede fique menos conectada, facilitando o cálculo das propriedades e a visualização da rede. Dessa forma, no restante dos experimentos foi utilizado um raio de 350 pixels.

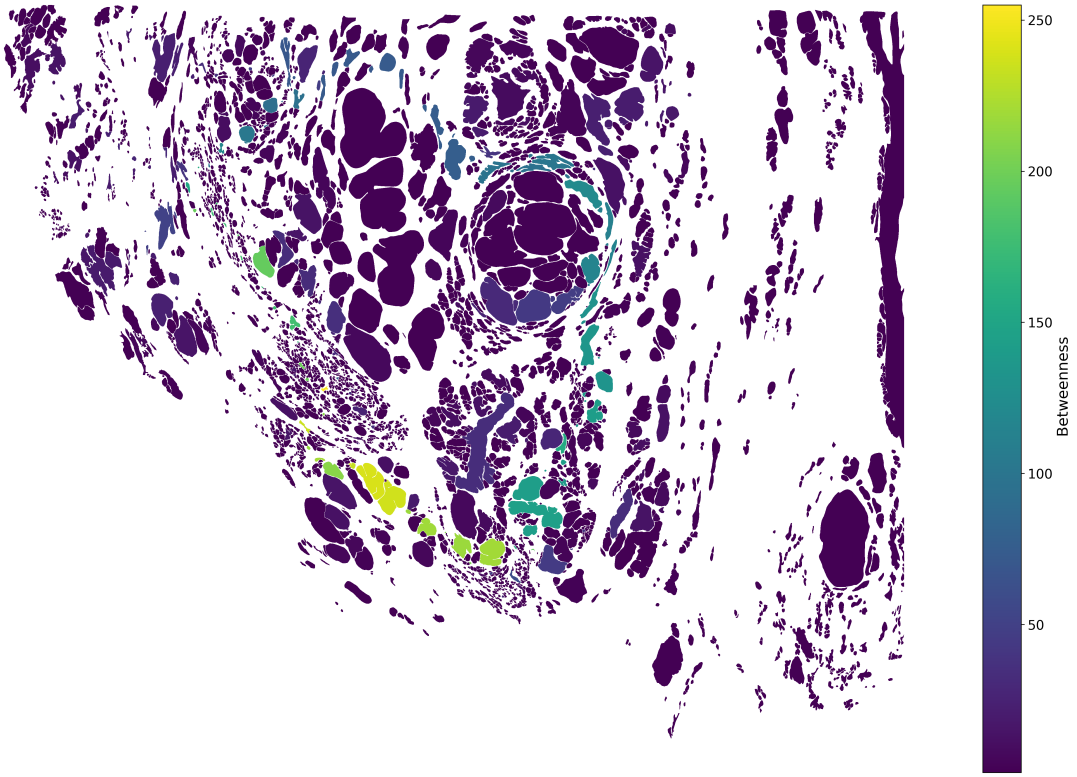

Figura 5.16: Visualização do coeficiente de intermediação intermediação dos nós na GCN. Cada glândula é colorida de acordo com o valor da medida de seu respectivo nó.

A partir da GCN criada e parametrizada pelo melhor raio, foi possível observar as características de rede considerando as ponderações das semelhanças das aparências. Em seguida, foram extraídos diferentes conjuntos de medidas das redes criadas, que foram utilizados para verificar a performance da classificação das glândulas das amostras em saudável (negativo para PCa) e doente (positivo para PCa). Os diferentes conjuntos de medidas considerados são: i) somente medidas de forma, ii) somente medidas de rede, iii) todas as medidas (forma+rede) e iv) somente a medida de grau das redes. Assim, foi possível fazer uma validação cruzada (conforme descrito na sessão 3.4.1) com cada tipo de propriedade na intenção de criar a matriz de confusão para verificar a acurácia obtida pelas propriedades. Os resultados da classificação das glândulas estão ilustrados nas figuras 5.18 (medidas de forma), 5.19 (medidas de rede), 5.20 (todas as medidas), 5.21 (medida de grau) para uma das amostras da próstata estudadas. As glândulas estão coloridas com base na classificação correta e incorreta da seguinte maneira: Azul é verdadeiro negativo, vermelho é verdadeiro positivo, amarelo é falso positivo e verde é falso negativo. Os respectivos valores de performance média obtidos para cada caso estão mostrados na tabela 5.1.

Comparando as classificações das glândulas com a região demarcada pelo especialista, mos-

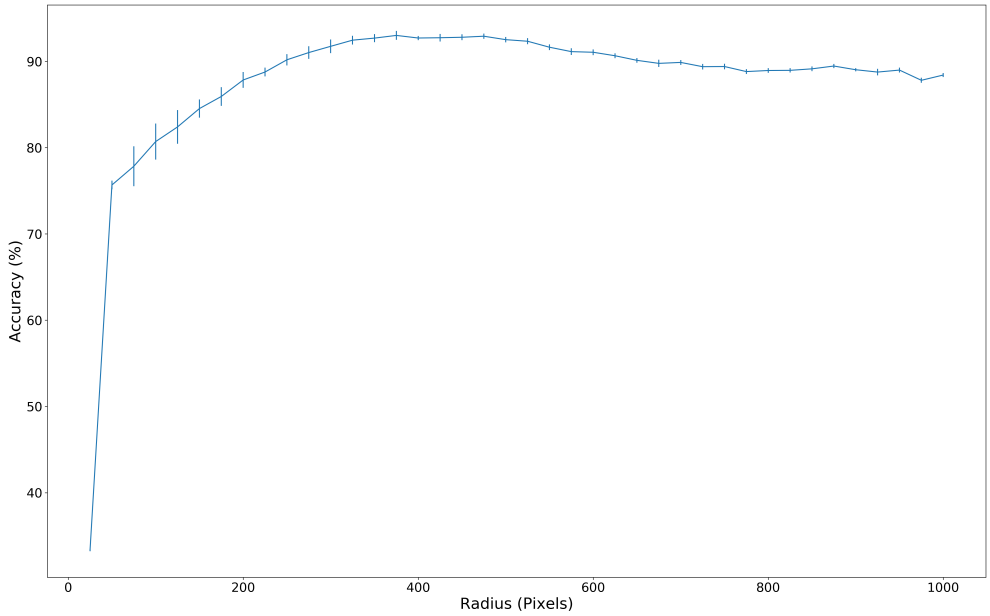

Figura 5.17: Visualização da acurácia da classificação de glândulas em doente e saudável em função do raio utilizado na criação da GCN. Barras verticais indicam o desvio padrão calculado para 10 realizações do procedimento.

trada na figura 5.1, vemos que as medidas de forma foram adequadas para identificar a maioria das glândulas saudáveis. Por outro lado, a qualidade de identificação de glândulas doentes foi muito baixa. Em contraste, o uso de propriedades de rede aprimorou muito o resultado. Em particular, utilizar apenas a medida de grau já aumentou muito a performance da classificação. Isso ocorreu porque as glândulas doentes tendem as ser menores e estar mais próximas entre si nas amostras estudadas, o que aumentou a densidade desse tipo de glândula. A propriedade de grau reflete a densidade de glândulas em regiões do tecido.

Tabela 5.1: Resultados obtidos nas duas imagens do estudo de caso utilizando o classificador *K-Nearest Neighbors* (K-NN). Os valores entre parênteses são o desvio padrão entre as duas imagens.

	Formas	**Redes**	**Formas e Redes**	**Grau**
Precision	0.366 (0.103)	0.818 (0.053)	0.789 (0.085)	0.764 (0.075)
Recall	0.699 (0.012)	0.943 (0.017)	0.918 (0.012)	0.937 (0.014)
Specificity	0.606 (0.006)	0.939 (0.007)	0.932 (0.003)	0.916 (0.005)
Accuracy	0.627 (0.001)	0.939 (0.009)	0.927 (0.002)	0.921 (0.005)

Os valores mostrados na tabela 5.1 podem ser melhor visualizados na forma da um gráfico de barras, que é mostrado na figura 5.22. A figura indica que nitidamente as medidas de redes aprimoraram o resultado. Adicionalmente, utilizar todas as medidas (forma+rede) levou a um

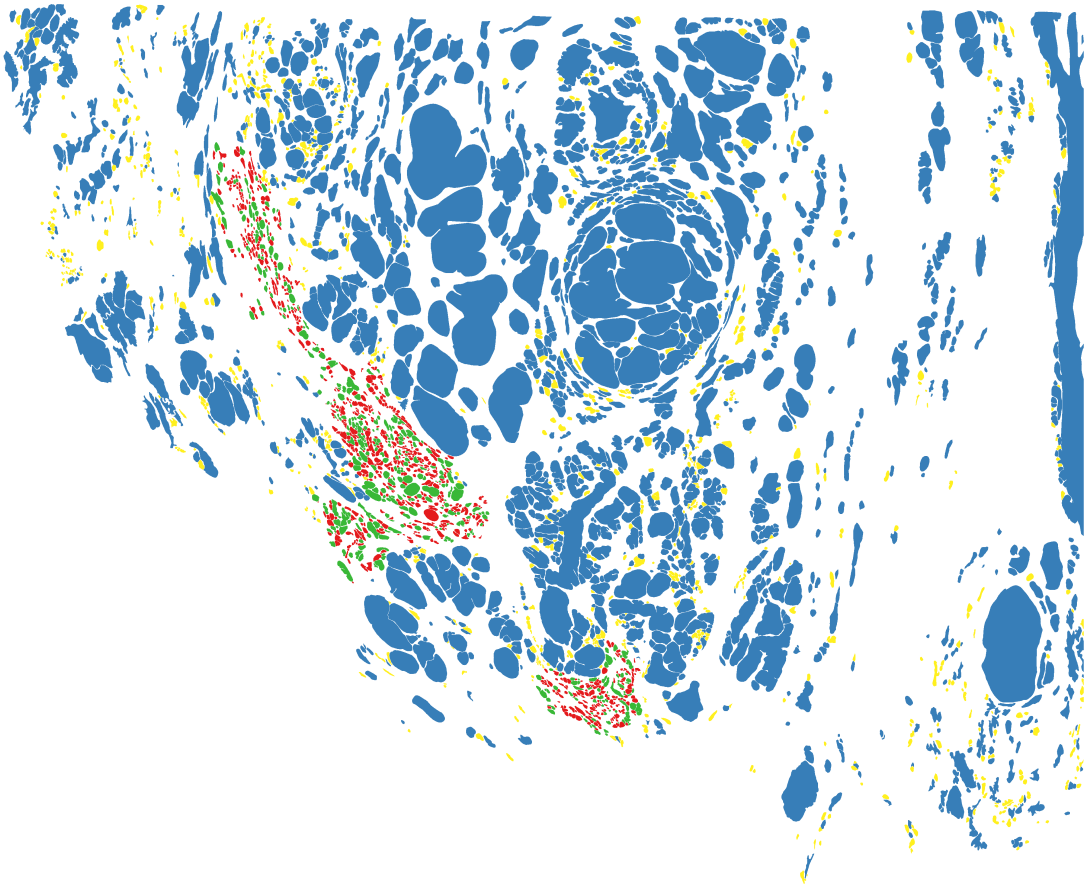

Figura 5.18: Visualização do resultado da classificação utilizando apenas as propriedades de forma e sua acurácia. As glândulas estão coloridas da seguinte forma: azul indica verdadeiro negativo, vermelho indica verdadeiro positivo, amarelo indica falso positivo e verde indica falso negativo.

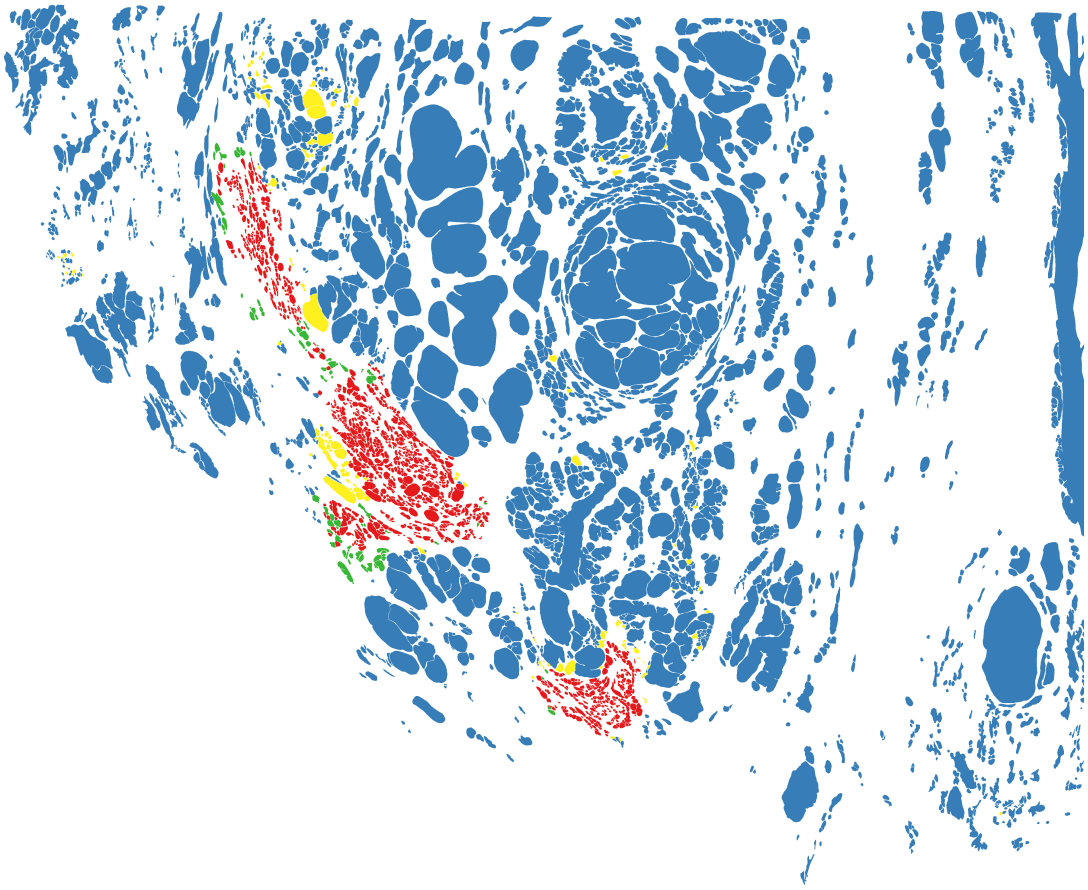

Figura 5.19: Visualização do resultado da classificação utilizando apenas as propriedades de rede e sua acurácia. As glândulas estão coloridas da seguinte forma: azul indica verdadeiro negativo, vermelho indica verdadeiro positivo, amarelo indica falso positivo e verde indica falso negativo.

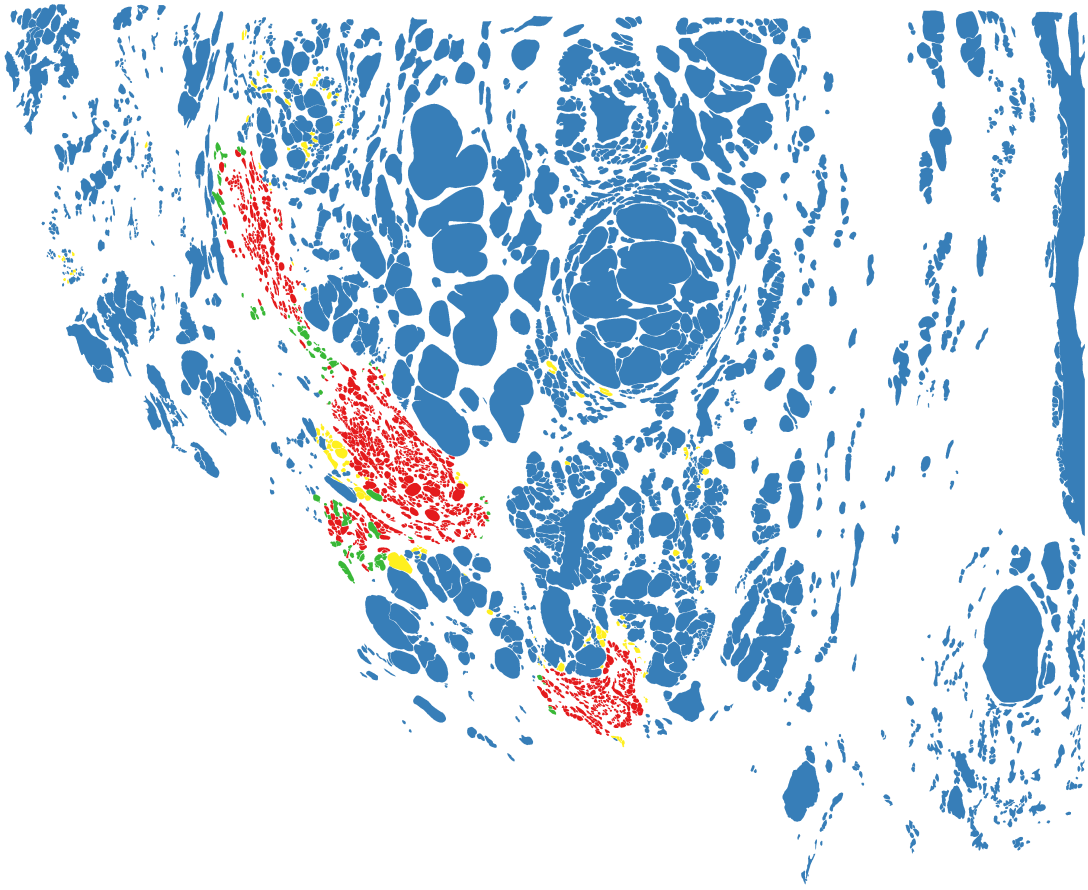

Figura 5.20: Visualização do resultado da classificação utilizando as propriedades de forma e rede. As glândulas estão coloridas da seguinte forma: azul indica verdadeiro negativo, vermelho indica verdadeiro positivo, amarelo indica falso positivo e verde indica falso negativo.

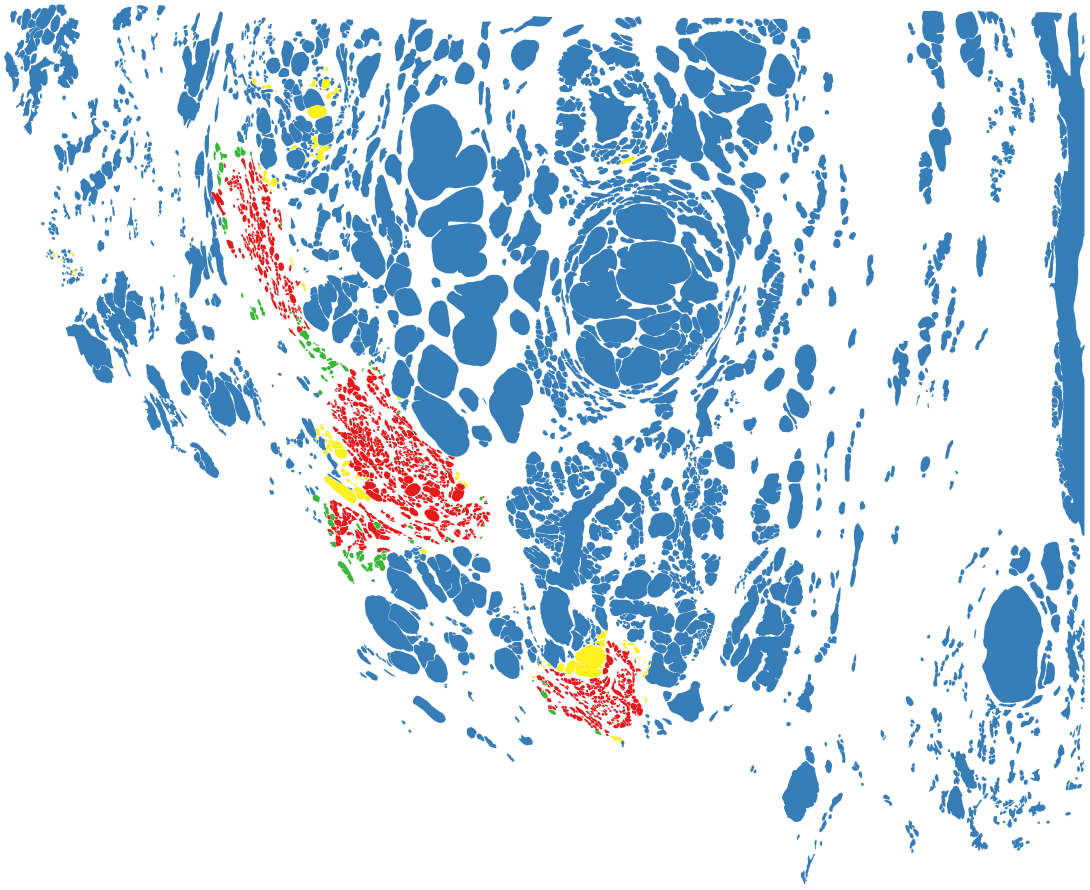

Figura 5.21: Visualização do resultado da classificação utilizando apenas a propriedade de grau. As glândulas estão coloridas da seguinte forma: azul indica verdadeiro negativo, vermelho indica verdadeiro positivo, amarelo indica falso positivo e verde indica falso negativo.

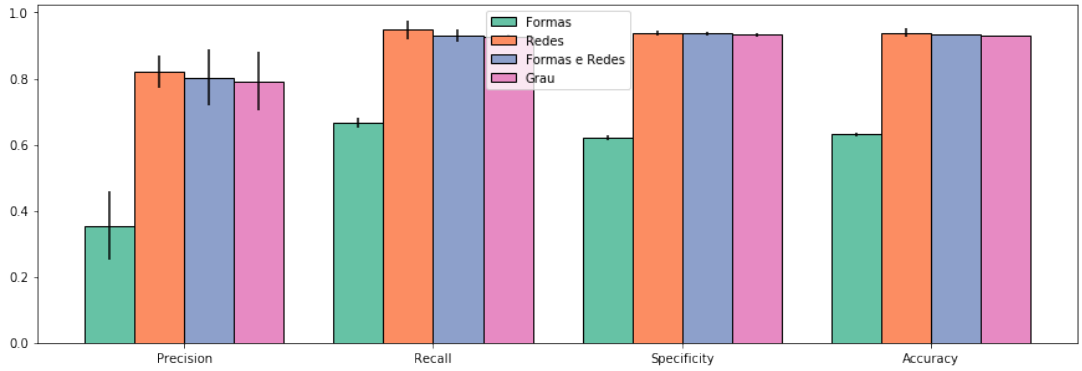

Figura 5.22: Resultado de classificação das glândulas quantificado utilizando as medidas de performance *precision, recall, specificity e accuracy* ao utilizar diferentes tipos de propriedades para a caracterização das glândulas. As linhas verticais indicam o desvio padrão da métrica de performance utilizando o classificador KNN.

resultado um pouco melhor do que o uso da medida de grau.

De forma a confirmar os resultados obtidos pelo KNN, foi aplicado também o classificador SVM nas propriedades das glândulas. O resultado é mostrado na tabela 5.2 e na figura 5.23. Os parâmetros do SVM foram otimizados para proporcionar a melhor performance possível. Vemos que os resultados são bem compatíveis com os obtidos pelo KNN.

Tabela 5.2: Resultados obtidos nas duas imagens do estudo de caso utilizando o classificador *Support vector machines* (SVMs). Os valores entre parênteses são o desvio padrão entre as duas imagens.

	Formas	**Redes**	**Formas e Redes**	**Grau**
Precision	0.372 (0.103)	0.813 (0.098)	0.785 (0.104)	0.800 (0.119)
Recall	0.824 (0.011)	0.948 (0.027)	0.941 (0.028)	0.942 (0.025)
Specificity	0.567 (0.029)	0.942 (0.013)	0.930 (0.019)	0.938 (0.021)
Accuracy	0.634 (0.004)	0.940 (0.004)	0.930 (0.004)	0.935 (0.010)

A acurácia média obtida no melhor caso para o KNN, respectivo ao uso do conjunto de medidas de rede, foi de $94.13\% \pm 0.01\%$, que é melhor ou equivalente aos melhores resultados que encontramos na literatura. Por exemplo, Nir et al. reportaram uma acurácia de 91.9% (NIR et al., 2018), Doyle et al. obteve 85.4% (DOYLE et al., 2007), Naik et al. alcançou um valor de 95.14% (NAIK et al., 2008) e Nguyen et al. obteve 79% (NGUYEN; SARKAR; JAIN, 2012). Por outro lado, é importante salientar que, com exceção de Naik et al., os autores utilizaram segmentação automática das glândulas.

Desta maneira, foi demostrado que mesmo as características mais básicas derivadas de uma GCN podem levar ao estado-da-arte em termos de performance de classificação em relação ao PCa. Considerando o todo, as GCNs podem auxiliar na definição de abordagens mais eficazes

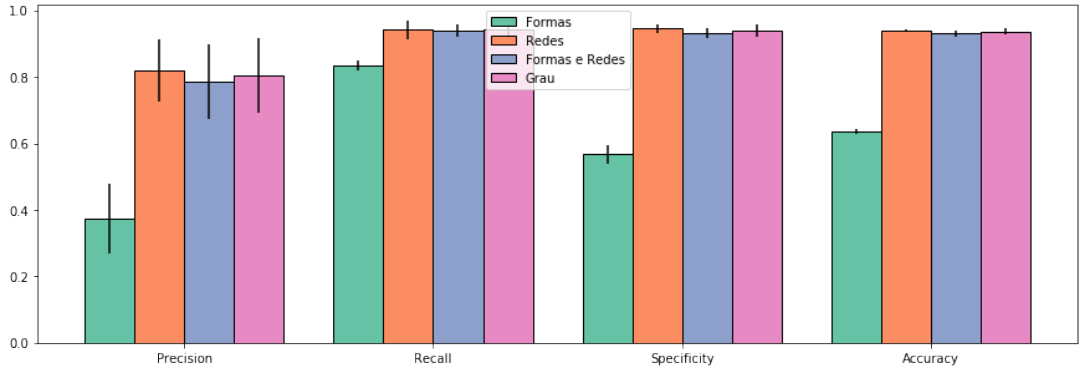

Figura 5.23: Resultado de classificação das glândulas quantificado utilizando as medidas de performance *precision, recall, specificity e accuracy* ao utilizar diferentes tipos de propriedades para a caracterização das glândulas. As linhas verticais indicam o desvio padrão da métrica de performance utilizando o classificador SVMs.

na classificação do PCa.

5.6 Robustez da GCN para detecção de PCa

5.6.1 Variação do parâmetro α

De forma a validar a robustez da GCN em relação à variação do parâmetro α, as glândulas foram classificadas utilizando valores de α no intervalo $[0,1]$. Para este experimento, foi utilizado o conjunto composto pelas medidas de rede e de forma. As Figuras 5.24(a) e 5.24(b) mostram a variação das medidas de performance em função do valor de α para, respectivamente, as amostras apresentadas nas Figuras 5.1(a) e 5.1(b). O parâmetro α ajusta a importância relativa entre a distância espacial e a similaridade das características entre as glândulas. Quando $\alpha = 1$, apenas a distância espacial é levada em conta na ponderação, enquanto que $\alpha = 0$ dá origem a pesos definidos apenas pela aparência das glândulas.

O método apresentou resultados similares mesmo com a variação de α. Isso porque o grau aparentemente possui um peso alto para a classificação, e essa medida não é alterada com os valores de α.

5.6.2 Perturbação da segmentação das glândulas

Segmentar glândulas é uma tarefa difícil. Alterações na segmentação podem influenciar a geração da GCN, que por sua vez influenciará o desempenho das tarefas de classificação. Para verificar o impacto que alterações na segmentação podem ter na classificação das glândulas,

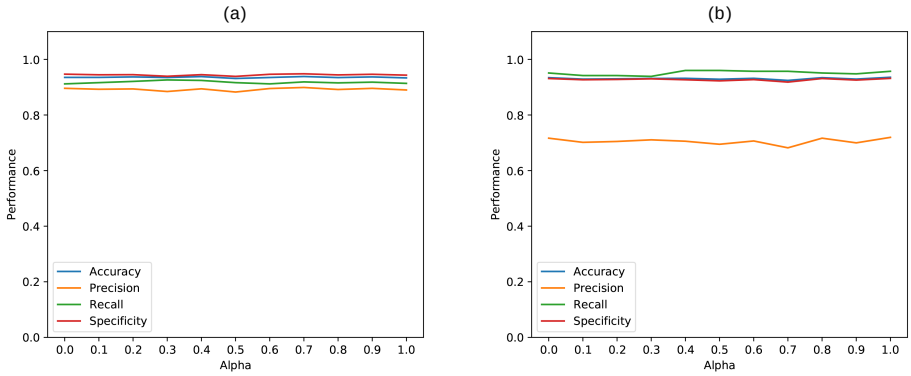

Figura 5.24: Performance da classificação das glândulas quando α é variado no intervalo $[0,1]$.

quatro tipos de perturbações foram aplicadas às anotações manuais das glândulas: i) erosão binária; ii) dilatação binária; iii) dilatação binária evitando fusões das glândulas e iv) remoção aleatória das glândulas. Os procedimentos de erosão e dilatação representam situações em que as glândulas segmentadas são menores ou maiores do que deveriam ser. O procedimento ii) tende a gerar componentes conectados muito grandes devido ao procedimento de dilatação, o que não representa uma situação real. Por outro lado, o procedimento iii) leva a glândulas dilatadas, mas o número de glândulas não muda. A remoção aleatória de glândulas representa situações em que glândulas não são identificadas por um anotador especialista ou por procedimentos de segmentação automatizados.

A Figura 5.25 mostra as métricas de desempenho obtidas para diferentes quantidades de erosão, dilatação e remoção de glândulas. Para a erosão e dilatação, o desempenho é medido em função do raio do elemento estruturante utilizado na respectiva operação morfológica. Em relação à erosão (Figura 5.25(a)), o método é robusto a erosões com raio de até 13 pixels. Observamos que, para esse raio, as glândulas segmentadas são bem menores do que as glândulas reais da imagem. Isso pode ser verificado pelo desaparecimento de muitas glândulas para raios maiores que 13 pixels, conforme visto no gráfico. Por exemplo, para um raio de 15 pixels, cerca de 15% das glândulas desaparecem. Dado que o método apresenta bom desempenho mesmo quando a segmentação das glândulas tem baixa qualidade (por exemplo, para um raio de 13 pixels), consideramos que o método é robusto à perturbação da erosão. Quanto à robustez do método com relação à operação de dilatação (Figura 5.25(b)), fica claro que pequenas quantidades de dilatação podem diminuir significativamente o desempenho da classificação. Em particular, a precisão torna-se muito menor após apenas 5 dilatações. Por outro lado, como explicado acima, como as glândulas são muito próximas umas das outras, o procedimento de dilatação leva a grandes componentes conectados cobrindo quase toda a amostra, que não repre-

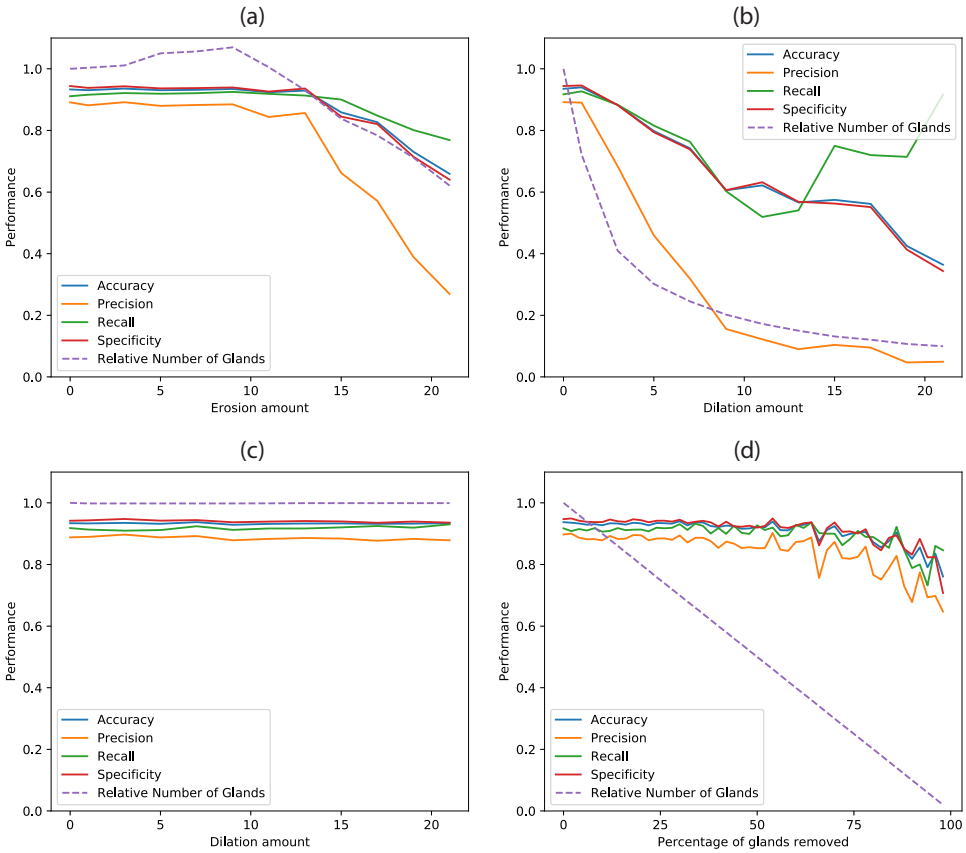

Figura 5.25: Performance da tarefa de classificação de glândulas em função da quantidade de perturbação aplicada às segmentações manuais. As perturbações foram: (a) erosão, (b) dilatação, (c) dilatação sem fusões de glândulas e (d) remoção aleatória de glândulas. A linha tracejada indica o número de glândulas após a perturbação dividido pelo número de glândulas quando nenhuma perturbação foi aplicada às imagens.

sentam de forma realista as glândulas ou possíveis erros nas anotações manuais. Isso pode ser observado pela diminuição do número de glândulas conforme a dilatação se torna maior. A Figura 5.25(c) mostra o resultado para o procedimento de dilatação evitando fusões de glândulas. Nesse caso, quase não há alteração no desempenho da classificação. Isso ocorre porque, como mencionado antes, as características relacionadas à distribuição espacial das glândulas tendem a ser mais relevantes para a classificação do que as associadas às suas formas. Como a dilatação sem fusão das glândulas não altera a distribuição espacial das glândulas, o desempenho permanece alto.

Os resultados obtidos para a remoção aleatória das glândulas são mostrados na Figura 5.25(d). Curiosamente, um bom desempenho é obtido mesmo quando cerca de 60% das glândulas são removidas. Isso ocorre porque quando as glândulas são removidas aleatoriamente com probabilidade uniforme, sua distribuição espacial geral não muda. Por exemplo, após a remoção, regiões com alta densidade de glândulas ainda terão densidade relativamente grande

quando comparadas a outras regiões da amostra. As informações sobre a distribuição espacial das glândulas remanescentes são suficientes para classificá-las corretamente. Os resultados da análise de perturbação indicam que, pelo menos para os dados considerados, as GCNs são robustas tanto no que diz respeito às mudanças consideradas na forma das glândulas quanto nas situações em que algumas glândulas não são segmentadas corretamente.

Um aspecto desafiador para quantificar o desempenho das metodologias de detecção de tumor é que não é simples definir os limites precisos de um tumor. Isso pode ser uma fonte de variação nas classes de glândulas *ground-truth* usadas para medir o desempenho do método. Para verificar de que forma a anotação manual das regiões tumorais influenciou nossos resultados, foi empregado um procedimento de avaliação intra-especialista. Os resultados da metodologia foram mostrados ao especialista que originalmente identificou as regiões tumorais em nossas amostras. As anotações foram reavaliadas pelo especialista considerando as glândulas TP, FN, TN e FP identificadas por nossa análise. A avaliação do especialista foi que as glândulas TP e TN foram de fato classificadas corretamente. Curiosamente, o especialista observou que a maioria das glândulas FN, ou seja, glândulas dentro da região tumoral delimitada, mas que foram classificadas por nosso método como normais (glândulas verdes na Figura 5.20), eram realmente glândulas normais. Essas glândulas devem ser consideradas normais porque, apesar de estarem dentro ou próximas da região tumoral, tinham distribuição espacial mais dispersa e ainda possuíam células basais. Além disso, muitas glândulas que estavam próximas, mas fora da região cancerígena e foram classificadas como anormais pelo método (glândulas amarelas na Figura 5.20), possuem suspeita de serem ácinos tumorais (ácinos atípicos), devido aos seus pequenos tamanhos e pequenas distâncias entre si. Observou-se também absência de células basais nessas glândulas. Assim, o método identificou com sucesso glândulas anormais que antes não estavam incluídas na região tumoral.

5.7 Considerações finais

A definição de características contextuais *ad hoc* pode negligenciar informações importantes para a classificação do tecido. Por exemplo, se as glândulas alongadas tendessem a aparecer juntas no tecido do PCa, mas se o alongamento da glândula for medido apenas individualmente, sua proximidade não poderia ser considerada para classificar as amostras de tecido. A definição de características contextuais específicas pode ser evitada usando, por exemplo, redes convolucionais (CNN) (ARVANITI et al., 2018; GOLDENBERG; NIR; SALCUDEAN, 2019). Entretanto, os resultados obtidos com CNNs não podem ser facilmente interpretados, o que dificulta o desdobramento clínico de soluções baseadas em CNNs. Assim, as GCNs podem ser usadas

não apenas para a identificação automática do PCa, mas também para identificar novas características associadas ao PCa que podem ser úteis em futuras revisões do GGS. Para mostrar o potencial da metodologia, criamos e caracterizamos GCNs para duas amostras de próstata de montagem inteira (*whole-mount*) e usamos as estruturas geradas para identificar regiões cancerígenas. Para o estudo de caso considerado, foi identificado que as glândulas alteradas podem ser detectadas com alta precisão, mesmo usando apenas uma propriedade de rede simples, o grau dos nós. A metodologia também pode ser aplicada a outros tipos de tecidos, desde que os recursos apropriados sejam usados para construir a GCN.

Capítulo 6

CONCLUSÕES

O PCa é um dos tipos mais comuns de câncer em homens, representando uma das principais causas de mortes relacionadas ao câncer em todo o mundo (BRAY et al., 2018). A detecção precoce do PCa pode melhorar o prognóstico e reduzir significativamente o risco de morte (WONG et al., 2016). Sendo uma doença que se difunde, o PCa se manifesta em uma ampla gama de padrões histológicos encontrados nas amostras de biópsia. Dada a importância da identificação de tecido anormal da próstata para melhorar o prognóstico, muitas metodologias computadorizadas destinadas a auxiliar os patologistas no diagnóstico foram desenvolvidas.

A caracterização automatizada ou semiautomatizada do tecido da próstata pode fornecer suporte importante para os patologistas identificarem tecidos anormais. Além disso, a análise de imagens da próstata em grande escala, usando amostras de montagem inteira (*whole-mount*), pode revelar fatores importantes para a classificação do PCa que não foram levados em consideração na revisão mais atualizada das diretrizes do ISUP (LEENDERS et al., 2020). Como consequência, alguns métodos automatizados foram desenvolvidos para segmentar, caracterizar e classificar o tecido da próstata (GOLDENBERG; NIR; SALCUDEAN, 2019; NIR et al., 2018; SINGH et al., 2017; ARVANITI et al., 2018; NGUYEN; SABATA; JAIN, 2012). Como os métodos desenvolvidos tendem a ter uma segmentação automatizada como primeira etapa, as demais análises tornam-se altamente dependentes da qualidade da segmentação. Por exemplo, pode ser difícil identificar se um mau desempenho na identificação de tecido anormal foi devido a características inadequadas ou se foi causado por tecido não estar corretamente segmentado.

Na literatura é frequentemente encontrado que um diagnóstico melhorado de uma região de tecido pode ser obtido considerando medidas que podem levar em consideração várias propriedades de seu entorno, ou seja, capaz de fornecer um contexto mais robusto para a análise. Neste trabalho partimos do pressuposto de que as glândulas já foram corretamente segmentadas, de forma manual ou automática, e focamos em uma metodologia de caracterização das glândulas

identificadas. Foi apresentado o conceito de *Gland Context Network (GCN)*, um grafo ponderado que descreve a relação espacial e visual entre as glândulas. Argumentamos que um dos principais benefícios do uso das GCNs é que a vasta literatura sobre caracterização de redes complexas torna-se imediatamente disponível para aumentar as medições individuais de glândulas.

A metodologia proposta pode ser usada para uma caracterização mais abrangente dos tecidos, integrando informações de escalas espaciais variáveis. De forma a ilustrar o procedimento, foram criadas GCNs para amostras de montagem inteira *whole-mount*. Foi mostrado que redes heterogêneas com muitas propriedades interessantes são criadas e podem ser investigadas posteriormente. Além disso, mostramos que mesmo em um procedimento simples de classificação de glândulas, usando classificadores básicos, as propriedades das GCNs levaram à identificação da grande maioria das glândulas anormais.

Para análises futuras, é interessante verificar se GCNs podem aprimorar não apenas a detecção de PCa, mas também proporcionar uma previsão do Escore de Gleason segundo as diretrizes do GGS. Adicionalmente, as GCNs podem ser usadas para melhorar ainda mais a classificação das glândulas usando um esquema de classificação em duas etapas. Os resultados indicam que muitas glândulas classificadas incorretamente tendem a estar próximas de glândulas classificadas corretamente. Um exemplo desse comportamento pode ser visto na figura 5.21. Assim, pode-se aplicar uma dinâmica de formação de consenso, como a difusão de opinião (SZNAJD-WERON; SZNAJD, 2000; CASTELLANO; FORTUNATO; LORETO, 2009), para propagar categorias de glândulas por toda a GCN, o que corrigiria a maioria das classificações errôneas.

REFERÊNCIAS

AMANCIO, D. R. et al. A systematic comparison of supervised classifiers. *PloS one*, Public Library of Science, v. 9, n. 4, p. e94137, 2014.

ARVANITI, E. et al. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports*, Nature Publishing Group, v. 8, 2018.

BALBERG, I. Universal percolation-threshold limits in the continuum. *Physical review B*, APS, v. 31, n. 6, p. 4053, 1985.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *Science*, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

BARTHÉLEMY, M. Spatial networks. *Physics Reports*, Elsevier, v. 499, n. 1-3, p. 1–101, 2011.

BARTHÉLEMY, M. *Spatial networks*. [S.l.]: Springer, 2014.

BISHOP, C. M. *Pattern recognition and machine learning*. [S.l.]: springer, 2006.

BRAY, F. et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 68, n. 6, p. 394–424, 2018.

CASTELLANO, C.; FORTUNATO, S.; LORETO, V. Statistical physics of social dynamics. *Reviews of Modern Physics*, APS, v. 81, n. 2, p. 591, 2009.

COSTA, L. d. F. Complex networks, simple vision. *arXiv preprint cond-mat/0403346*, 2004.

COSTA, L. d. F. et al. Characterization of complex networks: A survey of measurements. *Advances in physics*, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007.

COSTA, L. da F. Complex networks: New concepts and tools for real-time imaging and vision. *arXiv preprint cs.CV/0606060*, 2006.

CULP, M. B. et al. Recent global patterns in prostate cancer incidence and mortality rates. *European urology*, Elsevier, v. 77, n. 1, p. 38–52, 2020.

DALL, J.; CHRISTENSEN, M. Random geometric graphs. *Physical Review E*, APS, v. 66, n. 1, p. 016121, 2002.

DOMINGUES, G. S. et al. Topological characterization of world cities. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2018, n. 8, p. 083212, 2018.

DOYLE, S. et al. Automated grading of prostate cancer using architectural and textural image features. In: IEEE. *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*. [S.l.], 2007. p. 1284–1287.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley & Sons, 2012.

EPSTEIN, J. I. et al. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *The American journal of surgical pathology*, LWW, v. 40, n. 2, p. 244–252, 2016.

EPSTEIN, J. I. et al. The 2005 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, LWW, v. 29, n. 9, p. 1228–1242, 2005.

ERDOS, P.; RéNYI, A. On random graphs. *I". Publicationes Mathematicae (Debre*, 1959.

FILHO, O. M.; NETO, H. V. *Processamento digital de imagens*. [S.l.]: Brasport, 1999.

GABRIEL, K. R.; SOKAL, R. R. A new statistical approach to geographic variation analysis. *Systematic Zoology*, Society of Systematic Zoology, v. 18, n. 3, p. 259–278, 1969.

GLEASON, D. The Veteran's Administration Cooperative Urologic Research Group: histologic grading and clinical staging of prostatic carcinoma. In: TANNENBAUM, M. (Ed.). *Urologic Pathology: The Prostate*. [S.l.]: Philadelphia: Lea and Febiger, 1977. p. 171–198.

GLEASON, D. F.; MELLINGER, G. T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *The Journal of urology*, Wolters Kluwer Philadelphia, PA, v. 111, n. 1, p. 58–64, 1974.

GOLDENBERG, S. L.; NIR, G.; SALCUDEAN, S. E. A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology*, Nature Publishing Group, v. 16, n. 7, p. 391–403, 2019.

HAN, J.-D. J. et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, Nature Publishing Group, v. 430, n. 6995, p. 88–93, 2004.

JÄHNE, B. *Digital image processing: concepts, algorithms, and scientific applications*. [S.l.]: Springer Heidelberg, 1997. v. 6.

JR, R. M. C.; COSTA, L. da F. *Shape classification and analysis: theory and practice*. [S.l.]: Crc Press, 2009.

KRYVENKO, O. N.; EPSTEIN, J. I. Prostate cancer grading: a decade after the 2005 modified Gleason grading system. *Archives of Pathology & Laboratory Medicine*, the College of American Pathologists, v. 140, n. 10, p. 1140–1152, 2016.

KUMAR, V. et al. *Robbins and Cotran pathologic basis of disease, professional edition e-book*. [S.l.]: Elsevier health sciences, 2014.

LEENDERS, G. J. van et al. The 2019 international society of urological pathology (isup) consensus conference on grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, LWW, 2020.

MONACO, J. et al. Detection of prostate cancer from whole-mount histology images using markov random fields. In: CITESEER. *Workshop on Microscopic Image Analysis with Applications in Biology (in conjunction with MICCAI)*. [S.l.], 2008.

MONACO, J. P. et al. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models. *Medical image analysis*, Elsevier, v. 14, n. 4, p. 617–629, 2010.

MOSQUERA-LOPEZ, C. et al. Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. *IEEE reviews in biomedical engineering*, IEEE, v. 8, p. 98–113, 2015.

MOTTET, N. et al. Eau-eanm-estro-esur-siog guidelines on prostate cancer—2020 update. part 1: Screening, diagnosis, and local treatment with curative intent. *European Urology*, Elsevier, 2020.

NAIK, S. et al. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In: IEEE. *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. [S.l.], 2008. p. 284–287.

NAIK, S. et al. Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information. In: CITESEER. *MIAAB workshop*. [S.l.], 2007. p. 1–8.

NEWMAN, M. *Networks*. [S.l.]: Oxford University Press, 2018.

NGUYEN, K.; SABATA, B.; JAIN, A. K. Prostate cancer grading: Gland segmentation and structural features. *Pattern Recognition Letters*, Elsevier, v. 33, n. 7, p. 951–961, 2012.

NGUYEN, K.; SARKAR, A.; JAIN, A. K. Structure and context in prostatic gland segmentation and classification. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2012. p. 115–123.

NIR, G. et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical Image Analysis*, Elsevier, v. 50, p. 167–180, 2018.

QUINTANILLA, J.; TORQUATO, S.; ZIFF, R. M. Efficient measurement of the percolation threshold for fully penetrable discs. *Journal of Physics A: Mathematical and General*, IOP Publishing, v. 33, n. 42, p. L399, 2000.

SINGH, M. et al. Gland segmentation in prostate histopathological images. *Journal of Medical Imaging*, International Society for Optics and Photonics, v. 4, n. 2, p. 027501, 2017.

STEHMAN, S. V. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, Elsevier, v. 62, n. 1, p. 77–89, 1997.

SZNAJD-WERON, K.; SZNAJD, J. Opinion evolution in closed community. *International Journal of Modern Physics C*, World Scientific, v. 11, n. 06, p. 1157–1165, 2000.

VELLA, D. et al. Mtgo: Ppi network analysis via topological and functional module identification. *Scientific Reports*, Nature Publishing Group, v. 8, n. 1, p. 1–13, 2018.

VIDAL, J. et al. A fully automated approach to prostate biopsy segmentation based on level-set and mean filtering. *Journal of pathology informatics*, Medknow Publications, v. 2, 2011.

WÄHLBY, C. et al. Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections. *Journal of microscopy*, Wiley Online Library, v. 215, n. 1, p. 67–76, 2004.

WAXMAN, B. M. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, IEEE, v. 6, n. 9, p. 1617–1622, 1988.

WONG, M. C. et al. Global incidence and mortality for prostate cancer: analysis of temporal patterns and trends in 36 countries. *European urology*, Elsevier, v. 70, n. 5, p. 862–874, 2016.

Apêndices

APÊNDICE A - CÓDIGO DE CRIAÇÃO DA GCN

Os códigos desenvolvidos durante este trabalho e os resultados gerados estão disponibilizados de forma pública no repositório:

https://github.com/rodpmendes/GleasonScoreCharacterization

Introdução

Durantes os experimentos foram gerados códigos para validar as hipóteses criadas. A linguagem escolhida e utilizada foi Python 3.6.4 com a biblioteca Jupyter Notebook que é uma aplicação web *open-source* que permite criar e compartilhar documentos. Alguns scripts foram organizados em funções nos seguintes arquivos:

1.AppGleasonScoreCharacterization.ipynb

- •Este notebook produz um arquivo .gml contendo a GCN.

2.AppGleasonScoreDataAnalysis_UnbalancedCrossValidation.ipynb

- •Este notebook produz a análise de dados da GCN utilizando validação cruzada e KNN.

3.AppFindBestGeometricGraphRadius.ipynb

- •Este notebook encontra o melhor raio para a análise dos dados da GCN.

4.Algumas funções auxiliares foram reunidas em arquivos cujo os próprios nomes os descrevem

- •data_analysis_func.py
- •geometric_graph.py
- •misc.py

•prop.py

•unbalanced_cv.py

A seguir apresentamos os principais códigos para a reprodução das análises realizadas.

Criação da GCN

```
from igraph import Graph
from scipy.spatial import cKDTree as kdtree
import scipy.ndimage as ndi
import numpy as np

def geometric_graph(positions, radius):
    '''Generates a geometric graph
    positions: Positions of the points. A list where each item
    contains the position of a point.
    radius: Radius used for connecting the points'''

    tree = kdtree(positions)
    edges = list(tree.query_pairs(radius))
    g = Graph(n=len(positions), edges=edges)

    return g

def network_from_mask(img_mask, radius):
    '''Generate a geometric graph from a given mask image.
    img_mask: Binary image where 1 represents a pixel
    associated with an object.
    radius: Radius used for connecting the points'''

    lbl, nro = ndi.label(img_mask)
    idx = np.array(range(1, nro+1, 1))
    cm = ndi.measurements.center_of_mass(img_mask, lbl, idx)

    #ROTATE CENTER OF MASS
```

```
    cm = np.array(cm)
    cm = cm[:,::-1]
    cm[:,1] = img_mask.shape[0]-cm[:,1]

    g = geometric_graph(cm, radius)

    return g, cm
```

Ponderação das arestas

```
import numpy as np
from scipy.stats import zscore

def normalize_values(vals, means=None, stds=None):
    '''Normalize values in list ‘vals‘ using the equation

    vals_norm = (vals-means)/stds

    The list can be one-dimensional or bi-dimensional (N rows
    representing objects and M columns representing features).
    If ‘means‘ and ‘stds‘ are given, they are used for normalizing
    the values. Otherwise, the mean and standard deviation
    are calculated from the values. '''

    vals = np.array(vals)

    if (means is None) and (stds is None):
        vals = zscore(vals)
    else:
        vals = (vals - means)/stds

    return vals
```

```
def calculate_weight(pos_node1=None, pos_node2=None, att_node1=None,
att_node2=None, alpha=0.):
    '''Calculate edge weight for a pair of nodes with the given
    positions and attributes. ‘alpha‘ adjusts the relative
    importance between position and attribute. If alpha=0,
    only attributes are used. If alpha=1, only the
    position of the nodes are used.'''

    dist_pos2 = np.sum((pos_node1-pos_node2)**2)
    dist_att2 = np.sum((att_node1-att_node2)**2)

    dist2 = alpha*dist_pos2 + (1-alpha)*dist_att2
    weight = np.exp(-np.sqrt(dist2))
    return weight

def calculate_weight_all(nxgraph, pos_nodes, att_nodes,
alpha=0., att_idx=None, normalize_pos=True, normalize_att=True,
pos_means=None, pos_stds=None, att_means=None, att_stds=None):
    '''Calculate edge weights for all nodes in the graph.

    Parameters
    ----------
    nxgraph : networkx graph
      Graph to calculate the weights
    pos_nodes : list
      Positions of the nodes in the graph
    att_nodes : list
      List of nodes attributes (e.g.: shape properties)
    alpha : float
      Relative importante of positions and attributes when calculating
      the weight (see calculate_weight())
    att_idx : int
        Attribute index to use for weight calculation. If None, all
        attributes are used.
    normalize_pos : bool
```

```
    Whether the positions should be normalized
normalize_att : bool
    Whether the attributes should be normalized
pos_means : list
    Averages calculated for the positions, used in the
    normalization. Must have length two
pos_stds : list
    Standard deviations calculated for the positions, used in the
    normalization. Must have length two
att_means : list
    Averages calculated for the attributed, used in the
    normalization. Must have length equal to the
    number of attributes
att_stds : list
    Standard deviations calculated for the attributed, used in the
    normalization. Must have length equal to the
    number of attributes

Returns
-------
weight_dict : dict
    A dicitonary of the edge weights
'''

pos_nodes = np.array(pos_nodes)
att_nodes = np.array(att_nodes)
if normalize_pos or pos_means is not None:
    pos_nodes = normalize_values(pos_nodes, pos_means, pos_stds)
if normalize_att or att_means is not None:
    att_nodes = normalize_values(att_nodes, att_means, att_stds)
if att_idx is None:
    # Use all attributes
    att_idx = ...

weight_dict = {}
```

```
    for edge in nxgraph.edges:
        node1 = edge[0]
        node2 = edge[1]
        pos_node1 = pos_nodes[node1]
        pos_node2 = pos_nodes[node2]
        att_node1 = att_nodes[node1, att_idx]
        att_node2 = att_nodes[node2, att_idx]

        weight = calculate_weight(pos_node1, pos_node2, att_node1,
        att_node2, alpha)
        weight_dict[edge] = weight

    return weight_dict
```