

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Elias Paulino Medeiros

**Caracterização multiclasse do tecido colorretal usando Matriz de
Coocorrência Estrutural**

Fortaleza – CE

14 de abril de 2021

Elias Paulino Medeiros

Caracterização multiclasse do tecido colorretal usando Matriz de Coocorrência Estrutural

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Linha de Pesquisa: Inteligência Artificial

Orientador: Geraldo Luis Bezerra Ramalho

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PRPI)

Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Fortaleza – CE

14 de abril de 2021

Dados Internacionais de Catalogação na Publicação
Instituto Federal do Ceará - IFCE
Sistema de Bibliotecas - SIBI
Ficha catalográfica elaborada pelo SIBI/IFCE, com os dados fornecidos pelo(a) autor(a)

- M488c Medeiros, Elias Paulino.
 Caracterização multiclasse do tecido colorretal usando Matriz de Coocorrência Estrutural / Elias Paulino Medeiros. - 2021.
 66 f. : il. color.
- Dissertação (Mestrado) - Instituto Federal do Ceará, Mestrado em Ciência da Computação, Campus Fortaleza, 2021.
 Orientação: Prof. Dr. Geraldo Luis Bezerra Ramalho.
1. Câncer colorretal. 2. Matriz de Coocorrência Estrutural. 3. Classificação de imagem. 4. Aprendizagem de máquina. I. Título.

CDD 004



Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PRPI)
Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Elias Paulino Medeiros

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, sendo aprovada pela Coordenação do Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará e pela banca examinadora:

Orientador: Geraldo Luis Bezerra Ramalho

Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

Daniel Silva Ferreira

Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

Elias Teodoro da Silva Júnior

Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

Fortaleza – CE
14 de abril de 2021

Dedico este trabalho aos meus pais e avós.

Agradecimentos

Agradeço primeiramente a Deus, o ser supremo criador de todo o universo, que me concedeu saúde e discernimento para concluir este trabalho.

A minha mãe Cidronia, por todo o suporte dado em todas as minhas escolhas, por seus incentivos diários, preocupações e acreditar no meu potencial.

A meu pai Edivan, pelos seus conselhos e experiências de vida compartilhadas.

A minha namorada Letícia, pelo seu companheirismo diário, mesmo a distância, sempre me escultou e tranquilizou nas situações adversas.

Aos colegas do Laboratório de Processamento de Imagens, Sinais e Computação Aplicada (LAPISCO) pela valorosa colaboração.

Ao colega Prof. Daniel Ferreira, por sua estimada contribuição, com críticas e sugestões ao trabalho.

Ao Prof. Dr. Geraldo L. B. Ramalho, por sua orientação, contribuição, paciência e acreditar no trabalho.

Ao Instituto Federal do Ceará (IFCE), por permitir o afastamento das minhas funções no trabalho, na cidade de Cedro, para poder desenvolver o projeto de mestrado na cidade de Fortaleza.

*“Sometimes it is the people no one can imagine anything of who do the things no one can
imagine.”
(Alan Turing)*

Resumo

O câncer colorretal causa a morte de milhares de pessoas em todo o mundo, de acordo com a Organização Mundial da Saúde. Nesse contexto, imagens histológicas são usadas por especialistas para avaliar e identificar o nível de avanço da doença. A análise manual dessas imagens é um processo que apresenta algumas desvantagens: (i) pode demandar tempo excessivo à rotina clínica, devido à quantidade de características a serem observadas; e (ii) pode apresentar natureza tendenciosa, pois o trabalho do especialista é guiado por meio de experiências passadas e nem sempre existe uma segunda opinião. O reconhecimento automático do tecido da região afetada, através de um exame de imagem é essencial para o diagnóstico precoce da doença. A maioria das pesquisas consiste em empregar descritores de textura para capturar características do tumor e classificá-lo em benigno e maligno. Uma classificação multiclasse acurada é desafiadora devido à complexidade e diversidade dos tecidos colorretais. Recentemente, pesquisas mostraram que a análise de padrões estruturais de textura degradada por uma filtragem de imagens fornece recursos valiosos para pré-diagnóstico em várias aplicações médicas. Portanto, esse trabalho propõe uma abordagem para classificação automática de oito tipos de tecidos colorretal usando Matriz de Coocorrência Estrutural (do inglês, *Structural Co-occurrence Matrix*). Foram realizados experimentos em 5000 amostras de tecido de um conjunto de imagens público para avaliar a abordagem, considerando dois cenários: caracterização multiclasse do tecido colorretal usando características da Matriz de Coocorrência Estrutural de forma isolada; e combinada com características de outros descritores bem conhecidos na literatura. Os resultados mostram que a estratégia melhora o estado da arte, alcançando, acurácia: 91.30%; precisão: 91.41%; revocação: 91.31%; especificidade: 98.76% e F_1 : 91.31%.

Palavras-chave: Câncer colorretal, Matriz de Coocorrência Estrutural, Classificação de imagem, Aprendizagem de máquina.

Abstract

Colorectal cancer causes the deaths of thousands of people worldwide according to the World Health Organization. In this context, histology images are used by specialists to assess and identify the level of progress of the tumors. The manual analysis of these images is a process that has some disadvantages: (i) it may require excessive time for the clinical routine due to the number of characteristics to be observed; and (ii) may be biased, as the specialist's work is guided by past experiences and there is not always a second opinion. Automatic tissue recognition of histopathological images is essential for early disease diagnosis. Most research consists of employing texture descriptors to capture features that identify tumor samples as benign and malignant. An accurate multi-class classification is a challenge due to the complexity of colorectal tissue images. Recently, researchers have shown that the analysis of texture structural patterns degraded by image filtering provides valuable features for pre-diagnosis in several medical applications. Therefore, this work proposes an approach to automatically classify eight types of colorectal tissues using Structural Co-occurrence Matrix. Experiments were carried on 5000 tissue patches from a public image dataset to evaluate the algorithm, considering two scenarios: multiclass characterization of colorectal tissue using Structural Cooccurrence Matrix features in isolation; and combined with features of other descriptors well known in the literature. The results show that the strategy improves the state-of-the-art, achieving, accuracy: 91.30%, precision: 91.41%, recall: 91.31%, specificity: 98.76% e F_1 : 91.31%.

Keywords: Colorectal cancer, Structural Co-occurrence Matrix, Image classification, Machine learning.

Lista de ilustrações

Figura 1	– (a) Imagem histológica do tecido colorretal com coloração H&E. (b) - (i) Recortes de diferentes regiões de (a). (a) Tumor epitelial, (b) Estroma simples, (c) Estroma complexo, (d) Células imunes, (e) Detritos, (f) Glândulas mucosas, (g) Tecido adiposo e (h) Fundo.	18
Figura 2	– Resumo gráfico da Tabela 1. Cada valor adotada na coluna Bases de imagens representa uma base de imagem diferente.	27
Figura 3	– Segmentos do Cólon.	29
Figura 4	– Modelo genérico do SCM.	30
Figura 5	– Aspectos da disposição das probabilidades na matriz M de acordo com as diferenças estruturais de f e $k(g)$	31
Figura 6	– Matriz SCM ($N=16$) e suas marginais X e Y resultante da entrada de uma imagem f e função de degradação $k(g)$	32
Figura 7	– Ilustração do processo de classificação do KNN.	40
Figura 8	– O hiperplano ótimo, construído pelo SVM, separando os dados com a máxima margem ρ . Os pontos apontados pelas setas são os vetores de suporte.	41
Figura 9	– Ilustração do processo de classificação do DT.	42
Figura 10	– Amostras H&E de diferentes estruturas do tecido colorretal. (a) Tumor epitelial, (b) Estroma simples, (c) Estroma complexo, (d) Células imunes, (e) Detritos, (f) Glândulas mucosas, (g) Tecido adiposo e (h) Fundo. . .	45
Figura 11	– Abordagem proposta para classificação multiclasse do CRC. n representa as diferentes métricas de avaliação empregadas. $S1$ e $S2$ corresponde aos dois cenários explorados.	46
Figura 12	– Abordagem desenvolvida para buscar a melhor configuração do SCM na caracterização multiclasse do CRC.	47
Figura 13	– Criação dos conjuntos de características combinados com base no SCM.	48
Figura 14	– Avaliação quantitativa e individual para cada descritor pesquisado. Todos os valores apresentados estão em (%). Os resultados estão ordenados da menor para a maior ACC . Resultados obtidos usando o classificador SVM RBF.	53
Figura 15	– Comparação quantitativa dos melhores resultados alcançados com os conjuntos de características combinados. A linha horizontal (pontilhada e na cor vermelha) representa a melhor média de cada métrica apresentada na Tabela 3, alcançadas com o uso da SCM de forma isolada.	54

Figura 16 – Resultado da correlação de Pearson (BENESTY; CHEN; HUANG, 2008) entre as características de diferentes métodos. Observa-se alta correlação apenas entre as características dos métodos GLCM e HH.	55
Figura 17 – Comparações entre os conjuntos combinados usando o teste post-hoc Turkey. As interseções em cinza representam pares com diferença significativa em $\alpha = 0,05$	57
Figura 18 – Matriz de confusão da classificação multiclasse do CRC. a) Resultados obtidos do <i>best5</i> e b) <i>set_01</i> . As classes verdadeiras e preditas são descritas como: 1 - Tumor epitelial, 2 - Estroma simples, 3 - Estroma complexo, 4 - Células imunes, 5 - Detritos, 6 - Glândulas mucosas, 7 - Tecido adiposo e 8 - Fundo.	57
Figura 19 – (a) Aplicação do método SCM, com parâmetro $N = 8$, para extração de características de uma imagem da classe de tecido colorretal estroma complexo. (b) Conjuntos de características construídos a partir da matriz SCM.	66

Lista de tabelas

Tabela 1	– Um resumo do estado da arte na classificação do CRC em imagens histológicas. A numeração adotada na coluna Bases de imagens representa as bases de imagens usadas nas respectivas abordagens. 1 - imagens adquiridas em <i>Atrium Medical Centrum</i> (privada); 2 - base de imagens privada; 3 - imagens adquiridas em <i>Rawalpindi Medical College</i> (privada); 4 - imagens adquiridas em <i>Pathology Department archives of Rawalpindi Medical College</i> (privada); 5 - base de imagens privada; 6 - base de imagens descrita em Kather et al. (2016) (pública); 7 - base de imagens descrita em Sirinukunwattana et al. (2017) (pública).	26
Tabela 2	– Formato experimental para os cenários S1 e S2.	49
Tabela 3	– Resultado da classificação, em oito classes do CRC, usando diferentes configurações do SCM.	52
Tabela 4	– Tempo médio de execução para cada método de extração de características. Os valores apresentados estão ordenados do menor para o maior tempo médio.	53
Tabela 5	– Resultado da classificação, em oito classes do CRC, usando conjunto de características agrupadas. Os valores apresentados são médias resultantes de 1000 repetições do experimento.	56
Tabela 6	– Comparação da abordagem proposta com trabalhos que utilizaram a mesma base de imagens para classificação multiclasse do CRC. Foram apresentados apenas os desvios padrões fornecidos pelos autores.	56

Lista de abreviaturas e siglas

ACC	Acurácia
BRL	<i>Bayesian Logistic Regression</i>
CAD	<i>Computer Aided Diagnosis</i>
CAD	<i>Complementary Absolute Difference</i>
COR	Correlação
CNN	<i>Convolutional Neural Network</i>
CRC	Câncer Colorretal
CSD	<i>Chi-Square Distance</i>
CSR	<i>Chi-Square Distance Ratio</i>
D_{kl}	<i>Divergence of Kullback and Leibler</i>
DT	<i>Decision Tree</i>
ENT	Entropia
FN	Falso Negativo
FP	Falso Positivo
GLCM	<i>Gray Level Co-occurrence Matriz</i>
H&E	<i>Hematoxilina-Eosina</i>
HL	<i>Histogram-Lower</i>
HH	<i>Histogram-Higher</i>
IDM	<i>Inverse Difference Moment</i>
KNN	<i>K-Nearest Neighbor</i>
LBP	<i>Local Binary Patterns</i>
MLP	<i>Multilayer Perception</i>
PREC	Precisão

RBF	<i>Radial Basis Function</i>
SENS	Sensitividade
SPEC	Especificidade
SCM	<i>Structural Co-occurrence Matrix</i>
SMO	<i>Sequential Minimal Optimization</i>
SRM	<i>Structural Risk Minimization</i>
SVM	<i>Support Vector Machine</i>
TAE	Teoria de Aprendizado Estatístico
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo

Lista de símbolos

f	Sinal de entrada do SCM.
g	Sinal de entrada do SCM.
k	Fitro do SCM.
Q	Função de quantização.
M	Matriz de coocorrência.
P	Propriedade
μ_i	Média da i -ésima linha.
μ_j	Média da j -ésima coluna.
σ_i	Desvio padrão da i -ésima linha.
σ_j	Desvio padrão da j -ésima linha.
m_{ij}	Valor da matriz M referente ao ponto (i,j).
$\log$	Logarítmo.
Σ	Somatório.
L	Número de tons de cinza da imagem quantizada.
p	Probabilidades marginais.
$\mathbf{Q}$	Coefficiente de correlação máxima
$P(b)$	Distribuição de probabilidade de um intervalo b .
$\sqrt{}$	Raiz quadrada.
$I(x, y)$	Ponto (x,y) em uma imagem I .
P	Amostragem.
R	Raio.
T	Distribuição conjunta.
s	Função limiar.

g	Valor de tons de cinza de um pixel.
γ	Razão de aspecto espacial.
ϕ	Deslocamento de fase.
θ	Orientação espacial.
G	Gradiente.
ϕ	Valor de pi.
$\mathfrak{R}_n$	Espaço n-dimensional.

Sumário

1	INTRODUÇÃO	17
1.1	Justificativa	19
1.2	Objetivo	20
1.2.1	Objetivos Específicos	20
1.3	Organização do texto	20
2	ESTADO DA ARTE	22
2.1	Diagnóstico auxiliado por computador de imagens histológicas do CRC	22
2.2	Conclusão	25
3	FUNDAMENTAÇÃO TEÓRICA	28
3.1	Câncer carcinoma e adenocarcinoma	28
3.1.1	Câncer colorretal	28
3.2	Técnicas de extração de características	29
3.2.1	Matriz de Coocorrência Estrutural	29
3.2.1.1	Atributos do SCM	31
3.2.2	Matriz de Coocorrência de Níveis de Cinza	33
3.2.3	Estatística de baixa e alta ordem	34
3.2.4	Padrões Binários Locais	35
3.2.5	Filtro de Gabor	36
3.2.6	Características perceptuais do sistema visual humano	37
3.3	Técnicas de aprendizagem de máquina	39
3.3.1	K-vizinhos mais próximos	39
3.3.2	Máquina de Vetores de Suporte	40
3.3.3	Árvores de Decisão	42
3.4	Conclusão	43
4	METODOLOGIA	44
4.1	Base de imagens	44
4.2	Pré-processamento	44
4.3	Experimento de classificação	46
4.4	Avaliação dos resultados	49
5	RESULTADOS E DISCUSSÕES	51
5.1	Cenário S1	51

5.2	Cenário S2	53
6	CONCLUSÕES E TRABALHOS FUTUROS	58
	REFERÊNCIAS	59
	APÊNDICE A – DETALHES DA CONSTRUÇÃO DOS CONJUN- TOS DE CARACTERÍSTICAS OBTIDOS A PAR- TIR DO SCM	65

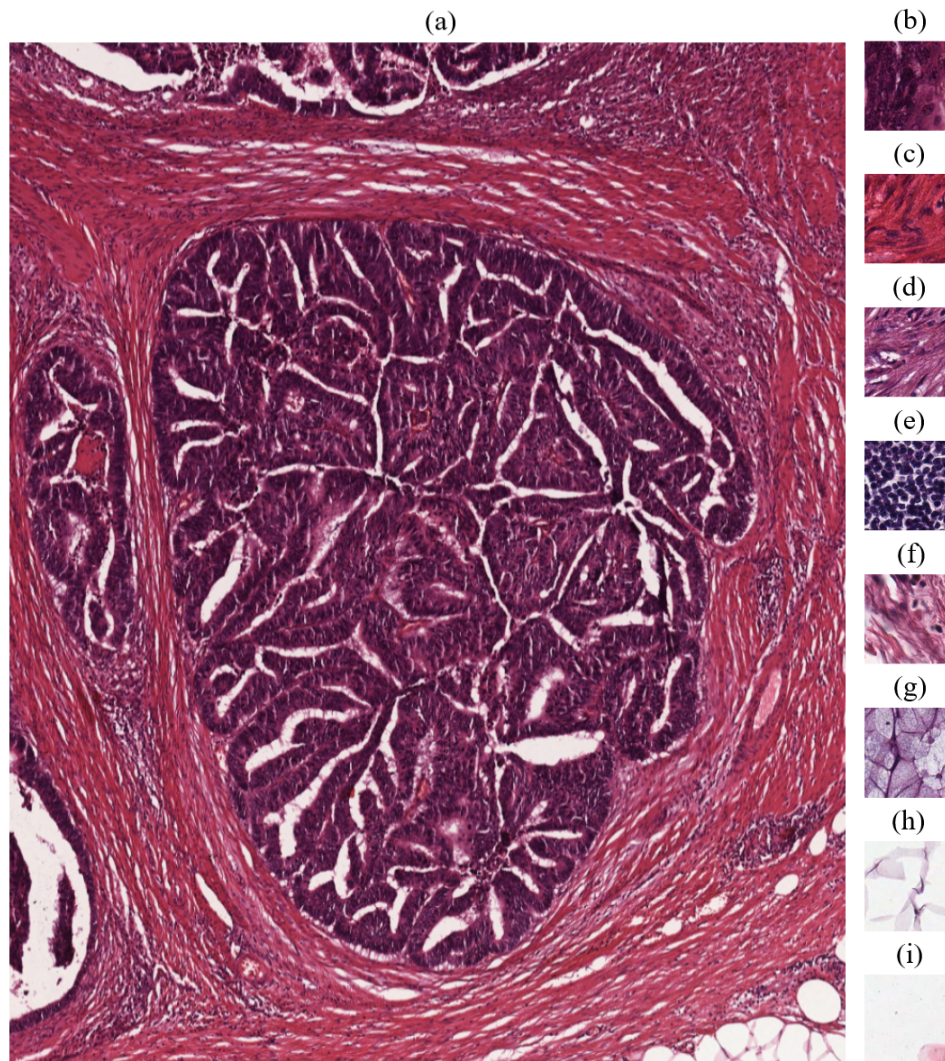
1 Introdução

O câncer é uma doença caracterizada pela mutação e consequente reprodução desordenada das células no corpo. Esse fenômeno acarreta a formação de agrupamentos chamados de tumores, que podem ser caracterizados como malignos devido ao seu potencial de lesionar tecidos adjacentes (SOCIETY, 2013). O colorretal é o terceiro tipo de câncer que mais tira a vida de pessoas no mundo. Em 2018, a Agência Internacional para Pesquisas sobre o Câncer (do inglês, *International Agency for Research on Cancer*) estimou que foram detectados 1,8 milhões de novos casos e 881,000 mortes (BRAY et al., 2018). Segundo o Instituto Nacional do Câncer (INCA), no Brasil, estima-se 40.990 novos casos em 2020, sendo 20,520 homens e 20,470 mulheres.

Câncer Colorretal (CRC) é o termo atribuído a uma patologia quando um tumor é encontrado em regiões do cólon, mais especificamente entre o cólon sigmóide e o reto. Também pode ser chamado de câncer de cólon ou reto, dependendo de onde os sintomas iniciaram (SOCIETY, 2017). A detecção precoce da doença é uma importante estratégia para encontrar o tumor na fase inicial e, assim, aumentar as chances de sucesso do tratamento. A presença de anormalidades causadas por esse câncer pode ser detectada principalmente pelos exames de sigmoidoscopia e colonoscopia. Ao encontrar anormalidades no exame, a confirmação ocorre através da biópsia do tecido da região suspeita, usando a técnica de coloração de tecidos em histologia Hematoxilina-Eosina (H&E) (SANTOS et al., 2018). Os patologistas são profissionais responsáveis por realizar o diagnóstico e o estudo de propriedades do tecido, tais como, estrutura, quantidade e forma das células (GURCAN et al., 2009). Exemplos de imagens de histologia H&E de diferentes partes do tecido colorretal são apresentados na Figura 1. Elas possuem uma grande variação de iluminação, intensidade de manchas e texturas de tecidos. Por conta dessa diversidade de características e classes, o processo de análise dos tecidos pode apresentar alguns problemas como: tempo gasto; e uma natureza tendenciosa, pois, a análise manual é orientada pelo conhecimento preexistente do patologista (MITTAL; SARASWAT, 2019). Portanto, a fim de reduzir o tempo de análise e interpretações subjetivas, sistemas computacionais vêm sendo desenvolvidos e utilizados por profissionais da saúde como ferramentas de auxílio durante o diagnóstico de imagens médicas.

Os sistemas CAD (do inglês, *Computer Aided Diagnosis*) são recursos computacionais que processam imagens médicas digitais com objetivo de destacar regiões e fornecer informações importantes que possam ajudar na avaliação prognóstica e apoiar o processo de tomada de decisão (SANTOS et al., 2019). Esses sistemas atualmente são usados para auxiliar no diagnóstico de várias doenças do corpo humano, tais como: pele, mama, esperma, coração e sistema colorretal (TSCHANDL; SINZ; KITTLER, 2019; RIORDON;

Figura 1 – (a) Imagem histológica do tecido colorretal com coloração H&E. (b) - (i) Recortes de diferentes regiões de (a). (a) Tumor epitelial, (b) Estroma simples, (c) Estroma complexo, (d) Células imunes, (e) Detritos, (f) Glândulas mucosas, (g) Tecido adiposo e (h) Fundo.



Fonte: Adaptado de Kather et al. (2016).

MCCALLUM; SINTON, 2019; SAN-SEGUNDO et al., 2019; TSOCHATZIDIS; COSTA-RIDOU; PRATIKAKIS, 2019; DEBELEE et al., 2017; TAMAI et al., 2017). Todos esses trabalhos investigam técnicas de processamento de sinais que envolvem na sua metodologia os passos de extração de características e classificação.

Durante o processo de construção dos sistemas CAD, a extração de características é uma etapa importante para o desempenho dessas ferramentas. Portanto, vários trabalhos concentram-se em buscar um método de extração de características ou um conjunto deles que melhor represente importantes regiões dos sinais analisados. Por exemplo, características extraídas pelo método Matriz de Coocorrência de Níveis de Cinza, GLCM (do inglês, *Gray Level Co-occurrence Matriz*) foram aplicadas para investigar doenças do cérebro (PS; DHARUN, 2016) e do pulmão (FILHO et al., 2017). Já Rouhi e Jafari (2016) experi-

mentaram usar apenas quatro atributos (energia, homogeneidade, correlação e contraste) desse extrator para segmentar nódulos da mama. Por outro lado, características extraídas pelo método de *Gabor* foram usadas para descrever recortes de mamogramas nas classes normal-anormal e benigno-maligno (BUCIU; GACSADI, 2011; KHAN et al., 2016). Han et al. (2015) propuseram uma investigação sistemática com objetivo de diferenciar nódulos do pulmão usando apenas informações de textura extraídas com o método Padrões binários locais, LBP (do inglês, *Local Binary Patterns*). O LBP obteve resultados significativamente superiores quando comparado com o GLCM e *Gabor*. Investigados por Tamura, Mori e Yamawaki (1978), seis atributos tentaram aproximar na forma computacional a maneira com que os humanos diferenciam as tonalidades de textura em imagens. Chamados de características de percepção visual, elas foram aplicadas por Bianconi, Álvarez-Larrán e Fernández (2015) para discriminar tecidos colorretal nas formas epitelial e estroma. Estatísticas de histograma também podem ser usadas para descrever textura de imagens médicas. Elas foram usadas por SV e Gunasundari (2018) para extrair informações de imagens dos olhos e posteriormente classificá-los em 3 classes (normal e mais duas variações de doenças oculares).

Outro descritor de textura, o método Matriz de Coocorrência Estrutural, SCM (do inglês, *Structural Co-occurrence Matrix*) (RAMALHO et al., 2016), analisa o relacionamento entre estruturas de baixo nível das imagens. Peixoto e Filho (2018), Filho et al. (2017), Filho et al. (2018) aplicaram essa técnica para caracterizar imagens de acidente vascular cerebral, câncer de pele e câncer de pulmão, respectivamente. Em todos os trabalhos o método apresentou resultados superiores ou comparáveis a literatura como descritor de textura e mostrou ser computacionalmente eficiente em relação a outras técnicas. Portanto, percebe-se um esforço científico contínuo no sentido de melhorar cada vez mais o desempenho dos sistemas CAD.

1.1 Justificativa

Os tumores do câncer colorretal (CRC) são estruturas complexas que normalmente contêm vários tipos de tecidos distintos (Figura 1). Esses diferentes tecidos podem ser distinguidos pela avaliação manual histopatológica em imagens com coloração Hematoxilina e Eosina (H&E). A arquitetura do tumor muda durante sua progressão e é de extrema importância verificar seu estado durante o prognóstico do paciente. Quantificar a composição do tecido no CRC é, portanto, uma tarefa relevante na histopatologia. Contudo, a maioria dos trabalhos publicados apresentam duas limitações: (i) eles consideram apenas duas ou três categorias de tecido, o que torna essas abordagens inadequadas para caracterização de uma quantidade maior de classes de tecidos do CRC; ou (ii) os estudos usam dados privados, o que dificulta a comparação quantitativa do desempenho da classificação (KALKAN et al., 2012, 2012; AKBAR; GOPI; BABU, 2015; WANG et al., 2017) .

Sempre que um patologista avalia uma imagem histológica, ele classifica manualmente as regiões do tecido nas categorias existentes. Automatizar essa tarefa é importante, pois, as imagens podem ser avaliadas de forma reproduzível e de forma rápida. A análise automática é particularmente útil quando se trata de quantificar regiões do tecido. Por exemplo, regiões com porção tumor-estroma provou ser um importante fator prognóstico em uma série de doenças neoplásicas. Com isso, Kather et al. (2016) foi o primeiro trabalho a propor uma classificação multiclasse, em oito classes de tecido, do câncer colorretal com validação em um banco de imagens público. Alguns desafios foram deixados por ele a comunidade científica: (i) melhorar o resultado de acurácia; (ii) diminuir o tempo de extração de características da abordagem e; (iii) identificar quais classes que causam mais confusão durante a avaliação do problema. O mesmo trabalho ainda deixa evidente que características de textura são informações valiosas para caracterização de imagens histológicas. Então, há um direcionamento nas pesquisas que buscam encontrar um descritor de características que possa contribuir para aprimorar sistemas CAD aplicados ao CRC. Portanto, esse trabalho propõe uma metodologia de classificação multiclasse do CRC, a fim de contribuir com soluções para os problemas encontrados na literatura.

1.2 Objetivo

O objetivo deste trabalho é propor uma abordagem com o SCM para descrever imagens de diferentes classes de tecidos do CRC.

1.2.1 Objetivos Específicos

Os seguintes objetivos específicos foram traçados:

- Investigar um conjunto métodos de extração de características para caracterização multiclasse do CRC, utilizando esses descritores de forma isolada e em grupos.
- Analisar qual combinação [extrator(es) - classificador] apresenta o melhor resultado no processo de classificação multiclasse do CRC.

1.3 Organização do texto

O texto está organizado da seguinte maneira: no Capítulo 2 são discutidos os trabalhos mais recentes no desenvolvimento de sistemas CAD para análise de imagens biomédicas. O Capítulo 3 descreve a teoria necessária para o entendimento e construção desse trabalho. O Capítulo 4 detalha os passos e materiais necessários para o desenvolvimento dessa pesquisa, tais como, os cenários analisados, descrição da base de imagens e métricas de avaliação. No Capítulo 5 são expostos e discutidos os resultados dos cenários abordados

nessa pesquisa. Por fim, no Capítulo 6 estão as conclusões finais desta dissertação e sugestões de investigações futuras.

2 Estado da Arte

Este capítulo apresenta uma revisão das abordagens direcionadas ao diagnóstico auxiliado por computador de imagens histológicas do CRC. Por meio deste, pode-se identificar as técnicas que vêm sendo utilizadas e os principais desafios da área.

2.1 Diagnóstico auxiliado por computador de imagens histológicas do CRC

Segundo Altunbay et al. (2009), existem na literatura quatro abordagens diferentes para quantificar tecidos em imagens histológicas: morfológica, textural, baseada na intensidade e estrutural. Usando a morfológica, os tecidos são quantificados com as propriedades de tamanho e forma dos componentes celulares (STREET; WOLBERG; MANGASARIAN, 1993). Para a abordagem baseada em intensidade, os tecidos são quantificados de acordo com os tons de cinza e intensidade das cores dos pixels. Nessa abordagem, histogramas da imagem são calculados a fim de extrair características como média, desvio padrão e entropia (WILTGEN; GERGER; SMOLLE, 2003). Com o uso da abordagem textural, a textura de todo o tecido é caracterizada por um conjunto de características como as calculadas em matrizes de coocorrência (DOYLE et al., 2007; ESGIAR et al., 1998; SERTEL et al., 2009) e coeficientes *multiwavelet* (WEYN et al., 1999). Por fim, a abordagem estrutural caracteriza os tecidos de acordo com a distribuição espacial das estruturas celulares. Essa abordagem foi escolhida por Altunbay et al. (2009) para representar e quantificar matematicamente tecidos do CRC. Para isso, o método constrói um grafo de cores para uma imagem, considerando os centroides dos componentes do tecido como nós e atribuindo bordas de cores entre esses nós, utilizando *Delaunay triangulation* (DELAUNAY et al., 1934). O grafo de cores fornece as informações estruturais como grau médio, coeficiente médio de agrupamento e diâmetro. Essas características foram usadas para treinar e testar o método Máquina de Vetores de Suporte, SVM (do inglês, *Support Vector Machine*) com objetivo de separar as imagens nas classes normal e dois níveis de câncer (*low-grade* e *high-grade*). A abordagem obteve 82,65% de acurácia analisando 213 imagens.

Kalkan et al. (2012) propuseram uma classificação automática de imagens histológicas do cólon analisando quatro cenários: (normal-anormal), (câncer-não câncer) e (normal-câncer-adenoma-inflamação). As informações utilizadas pelo autor consistiam em forma e textura dos componentes estruturais do tecido do cólon: estroma, lúmen e núcleos. Além das informações de forma, os descritores de Haralick e Gabor foram aplicados para representar as características texturais. A classificação realizada das classes

(normal-câncer-adenoma-inflamação), utilizando regressão logística, o modelo apresentou apenas 75% de acurácia. No estudo um total de 8000 recortes de tecidos foram analisadas de uma base de imagens privada.

Akbar, Gopi e Babu (2015) compararam o desempenho dos classificadores Perceptron de Múltiplas Camadas, MLP (do inglês, *Multilayer Perception*); Otimização Sequencial Mínima, SMO (do inglês, *Sequential Minimal Optimization*); Regressão Logística Bayesiana, BLR (do inglês, *Bayesian Logistic Regression*) e o *K-star* durante a classificação de características extraídas com o método GLCM de imagens do cólon. A abordagem avaliou apenas 113 imagens (64 câncer e 49 normais). Eles mostraram que a técnica de aprendizagem de máquina MLP conseguiu separar com melhor acurácia as classes do problema, alcançando 83,33%. No entanto, os autores não apresentaram as taxas de verdadeiros positivos ou verdadeiros negativos da classificação, como também não aplicaram uma abordagem na separação multiclasse do tecido do cólon.

Naiyar, Asim e Shahid (2015) observaram que o componente lúmen¹ das imagens histológicas tendem a ser elípticas em imagens normais e irregulares quando o tecido é maligno. Com objetivo de separar as imagens em grupos normais e malignos, e considerando a disposição estrutural do lúmen, os autores experimentaram 3 extratores de características que capturam informações estruturais e morfológicas. O desempenho das características extraídas pelos métodos distância dos mínimos quadrados (do inglês, *Least square distance*), descritores elípticos de fourier (do inglês, *Elliptic Fourier descriptors*) e um grupo de características morfológicas foram investigadas de forma isolada e associadas. Apesar do modelo atingir 96,55% de acurácia, classificando características associadas com o método SVM, apenas duas classes de tecidos foram avaliadas.

No trabalho de Iftikhar, Hassan e Alquhayz (2016) foi proposto um sistema de avaliação do grau de agressividade do câncer do cólon. Três níveis (classes) foram definidos pelos autores: (i) O bem diferenciável é o nível inicial de câncer, onde os tecidos são quase semelhantes aos normais e a progressão é lenta. (ii) O moderadamente diferenciável, apresenta tecidos diferente dos normais e a progressão é mais rápida. (iii) No último nível, o pouco diferenciável, o tecido é totalmente deformado e a progressão do câncer é muito rápida. Para definir o nível dos tecidos, a abordagem baseia-se na análise de características de textura e intensidade das imagens, para isso foram usados os descritores de Haralick e momentos do histograma. A metodologia foi validada submetendo características de 92 imagens ao classificador SVM. A acurácia obtida foi de 95,4%.

Pesquisas desenvolvidas no contexto de imagens do CRC, geralmente dedicam esforços nas tarefas de segmentação de importantes estruturas do tecido, classificação e

¹ Lúmen é um componente da célula, delimitado pela parede celular, na qual, seu interior é preenchido pelo protoplasma, que consiste na: membrana plasmática, citoplasma e núcleo. Em células mortas, o lúmen não apresenta protoplasma.

detecção do grau de agressividade do câncer (RATHORE et al., 2013). Na pesquisa feita por Jørgensen et al. (2017), foi realizada a tarefa de segmentação das células e classificação dos tecidos em malignos e benignos. Os autores aplicaram a abordagem em 87 imagens. Na fase de segmentação, foram usados os métodos *color deconvolution* (RUIFROK; JOHNSTON et al., 2001), *k-means clustering* e limiarização adaptativa local. Na fase 2, foram extraídas informações das células segmentadas a partir dos métodos intensidade mínima; média e máxima de cada canal de cor; matrizes de coocorrência; matriz *run-length*; e estatísticas de primeira ordem. O poder discriminativo dessas informações foi avaliado através do classificador Floresta Aleatória, RF (do inglês, *Random Forest*), no qual alcançou acurácia de 91,0%.

O trabalho de Kather et al. (2016) foi o primeiro a propor uma abordagem de classificação multiclasse tecido colorretal com 8 classes. Para isso, eles avaliaram o poder de discriminação de 6 descritores (Gabor, GLCM, características de percepção visual, LBP e estatística de histograma de alta e baixa ordem), de forma isolada e associando características desses métodos. Eles perceberam que a associação de 5 dos métodos analisados obteve o melhor resultado na caracterização dos tecidos, alcançando 87,4% de acurácia. No entanto, a associação de métodos que obteve melhor acurácia, foi a que necessitou do segundo maior tempo médio de extração de características, o que motiva a comunidade científica a buscar modelos adequados a rotina clínica.

Ribeiro et al. (2019) apresentaram uma investigação que envolve métodos de extração de características e métodos de aprendizagem de máquina para o diagnóstico benigno e maligno do CRC. A abordagem propôs avaliar também, descritores de diferentes princípios de funcionamento de forma isolada e associados. As características foram extraídas por meio da transformação *curvelet* e descritores de Haralick. A etapa de classificação avaliou os métodos Árvores de Decisão, DT (do inglês, *Decision Tree*); RF; SVM; *Naive Bayes*; K-Star; e Polinomial, para a classificação de características extraídas das bases de imagens descritas em (SIRINUKUNWATTANA et al., 2017) e (KATHER et al., 2016). Os resultados obtidos com o classificador Polinomial foram 99,4% de AUC e 97,68% de acurácia para primeira e segunda base, respectivamente.

A generalização das Redes Neurais Convolucionais, CNN (do inglês, *Convolutional Neural Network*) tem revolucionado o campo da análise de imagens médicas, mostrando desempenhos similares ou superiores aos especialistas humanos (BEJNORDI et al., 2017; KOOHBANANI et al., 2018; SONG et al., 2018). Nesse contexto, Wang et al. (2017) propuseram um algoritmo chamado bilinear CNN (BCNN) para classificar imagens CRC divididas em oito tipos de tecidos. O algoritmo primeiro decompõe a imagem analisada nos componentes H (Hematoxilina) e E (Eosina) presentes em imagens de exames com coloração do tipo H&E. A BCNN é composta por duas CNNs que são responsáveis por extrair características das decomposições. As saídas são combinadas por um *pooling* bilinear

e classificadas com o SVM. Os resultados mostraram um desempenho comparável com a literatura, 92% de acurácia, porém os autores selecionaram para avaliação um conjunto de 1000 imagens de um universo de 5000.

Outra abordagem com CNN foi apresentada por Rachapudi e Devi (2020). Diferente das tradicionais CNNs, eles construíram uma nova arquitetura direcionada para o problema de classificação do CRC. O método consiste em cinco camadas de blocos subsequentes, onde cada bloco é representado por uma camada convolucional, um *drop-out* e um *max-pooling*. A abordagem foi avaliada com 5000 imagens, e alcançou apenas 77,3% de acurácia, desempenho inferior quando comparada com outras abordagens que avaliaram a mesma base de imagens. Embora alguns modelos de CNN mostrem melhor desempenho para vários problemas de classificação, eles podem apresentar algumas desvantagens em relação a tradicionais abordagens, por exemplo, o custo computacional e necessitam de grande quantidade de imagens marcadas para treinamento da rede.

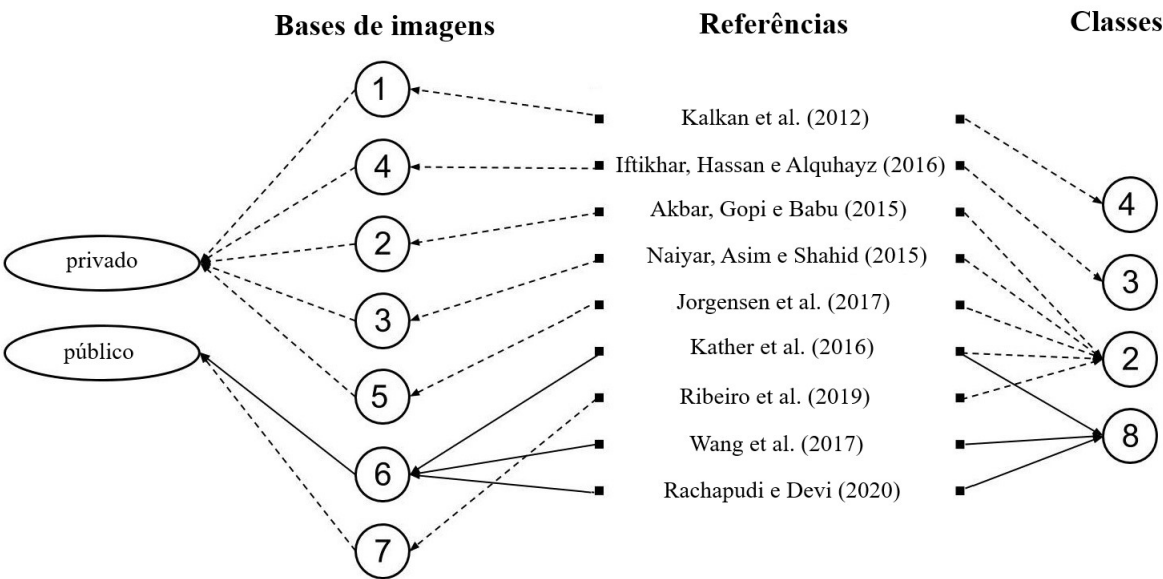
2.2 Conclusão

Neste capítulo foram discutidos os principais e mais recentes trabalhos vinculados ao desenvolvimento de abordagens computacionais para auxílio ao diagnóstico de imagens do CRC. Os trabalhos buscam melhorar a literatura com pesquisas dedicadas a encontrar descritores que aumentem o poder de discriminação de imagens para problemas de classificação. Nota-se que a informação de textura vem sendo bastante explorada e provando ser um importante recurso na representação de tecido em imagens histológicas. Um resumo dos trabalhos discutidos é apresentado na Tabela 1. Pode-se perceber que a classificação do tecido colorretal vem sendo pouco explorada, como também, não foi avaliado a capacidade das abordagens em identificar casos verdadeiros negativos ou verdadeiros positivos, excluindo métricas importantes como, revocação, especificidade, precisão e F_1 . Na Figura 2 tem-se um resumo gráfico da Tabela 1, onde mostra que a maioria dos trabalhos consideram classificar poucas categorias de tecidos, o que torna essas abordagens inadequadas para partes mais heterogêneas do tumor; e usam dados privados, o que dificulta a comparação quantitativa dos resultados. Apesar das CNNs apresentarem bons desempenhos na caracterização de diversos tipos de imagens médicas, no contexto do tecido colorretal, não mostraram serem superiores ao desempenho (classificação e tempo de extração) de abordagens que usaram tradicionais descritores de textura, o que desmotivou a inclusão dessas técnicas na abordagem de investigação desse trabalho.

Tabela 1 – Um resumo do estado da arte na classificação do CRC em imagens histológicas. A numeração adotada na coluna **Bases de imagens** representa as bases de imagens usadas nas respectivas abordagens. 1 - imagens adquiridas em *Atrium Medical Centrum* (privada); 2 - base de imagens privada; 3 - imagens adquiridas em *Rawalpindi Medical College* (privada); 4 - imagens adquiridas em *Pathology Department archives of Rawalpindi Medical College* (privada); 5 - base de imagens privada; 6 - base de imagens descrita em Kather et al. (2016) (pública); 7 - base de imagens descrita em Sirinukunwattana et al. (2017) (pública).

Referências	Bases de imagens		Imagens usadas	Classes	Extrator		Classificador	ACC(%)
Kalkan et al. (2012)	1	8000	4	2	GLCM, Gabor e informações de forma	Regressão logística	MLP	75,15
Akbar, Gopi e Babu (2015)	2	113	2	2	GLCM			83,33
Naiyar, Asim e Shahid (2015)	3	174	2	2	Least square lumen distance, Elliptic Fourier descriptorse e características morfológicas	SVM		96,55
Iftikhar, Hassan e Alquhayz (2016)	4	92	3	3	Haralick e momentos do histograma	SVM		95,40
Jørgensen et al. (2017)	5	87	2	2	GLCM, matriz run-length e estatísticas de primeira ordem	Random Forest		91,00
Kather et al. (2016)	6	5000	2	2	GLCM, PERC, LBP, HH e HL.	SVM		98,60
			8	8	GLCM, PERC, LBP, HH e HL.	SVM		87,40
Ribeiro et al. (2019)	7	500	2	2	transformação curvelet e Haralick	Polinomial		97,68
Wang et al.(2017)	6	1000	8	8	CNN	SVM		92,60
Rachapudi e Devi (2020)	6	5000	8	8	CNN	CNN		77,30

Figura 2 – Resumo gráfico da Tabela 1. Cada valor adotada na coluna **Bases de imagens** representa uma base de imagem diferente.



3 Fundamentação Teórica

Este capítulo apresenta conceitos associados ao CRC, uma revisão de técnicas de extração de características e classificação comumente empregadas na construção de sistemas de diagnóstico auxiliado por computador em imagens biomédicas. O objetivo é constituir um embasamento teórico para a compreensão da metodologia proposta e para avaliação dos resultados obtidos.

3.1 Câncer carcinoma e adenocarcinoma

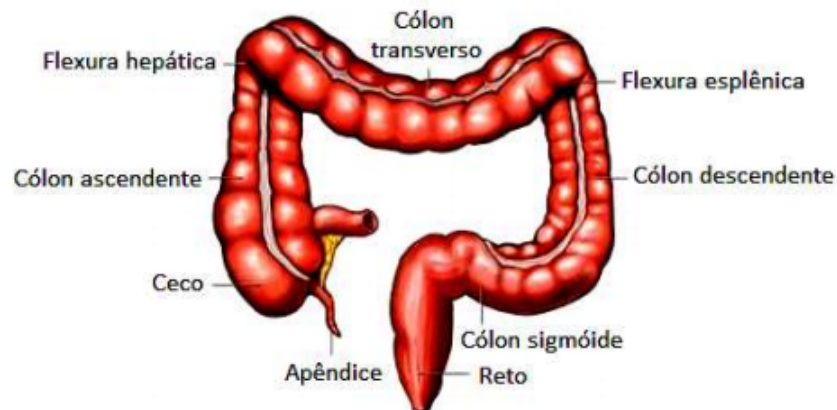
O corpo humano é composto por basicamente quatro tipos de tecidos: o tecido conjuntivo, tecido nervoso, tecido muscular e tecido epitelial. O carcinoma é um tipo de câncer que surge a partir de uma célula epitelial (BRENER et al., 2007). Como o tecido epitelial recobre a pele e a maioria dos órgãos do corpo humano, esse câncer pode aparecer em diversas partes do corpo. Por exemplo, se a célula que sofreu mutação é uma célula epitelial do rim, o câncer que surge dela é o carcinoma de células renais; se a origem do câncer for a célula epitelial do fígado, conhecida como hepatócito, o câncer se chamará hepatocarcinoma. Por outro lado, se o câncer surge a partir de um tecido epitelial que contém glândulas, o carcinoma passa a ser chamado de adenocarcinoma. Portanto o adenocarcinoma pode aparecer no estômago, mama, pâncreas, próstata, cólon ou reto (NEILAN, 1983).

3.1.1 Câncer colorretal

O colorretal é o terceiro tipo de câncer que mais tira a vida de pessoas no mundo. Em 2018, a Agência Internacional para Pesquisas sobre o Câncer (do inglês, *International Agency for Research on Cancer*) estimou que foram detectados 1,8 milhões de novos casos e 881,000 mortes (BRAY et al., 2018). Segundo o Instituto Nacional do Câncer (INCA), no Brasil, estima-se 40.990 novos casos em 2020, sendo 20.520 homens e 20.470 mulheres.

A grande maioria dos casos de câncer colorretal localizam-se no cólon esquerdo, principalmente no cólon sigmóide e no reto (Figura 3). A presença de anormalidades causadas por esse câncer pode ser detectada principalmente pelos exames de sigmoidoscopia e colonoscopia. Ao encontrar anormalidades, é feita uma coleta de tecido da região afetada a fim de informar ao clínico a natureza, gravidade, extensão, evolução e a intensidade do problema no tecido. A histologia é a área de estudo que analisa os tecidos biológicos a nível microscópico. Por meio da técnica de coloração de tecidos em histologia Hematoxilina-Eosina (H&E) (SANTOS et al., 2018), as estruturas que compõem o tecido epitelial são

Figura 3 – Segmentos do Cólon.



Fonte: Yamane (2013).

destacadas. Por exemplo, a hematoxilina têm atração por substâncias ácidas (basófilas) dos tecidos, como os núcleos e o retículo endoplasmático rugoso e ácidos nucleicos. Já a eosina, sendo ácida, cora predominantemente o citoplasma, as fibras de colágeno e outras estruturas compostas por substâncias com caráter básico (acidófilas). Uma imagem de histologia com coloração H&E do tecido colorretal foi apresentada na Figura 1. O tecido apresenta diferentes regiões que podem incluir: tumor epitelial, estroma simples, estroma complexo, células imunes, detritos, glândulas mucosas, tecido adiposo e fundo. A identificação correta de cada região pode ajudar o patologista a definir o tipo ou grau de complexidade do câncer (GURCAN et al., 2009).

3.2 Técnicas de extração de características

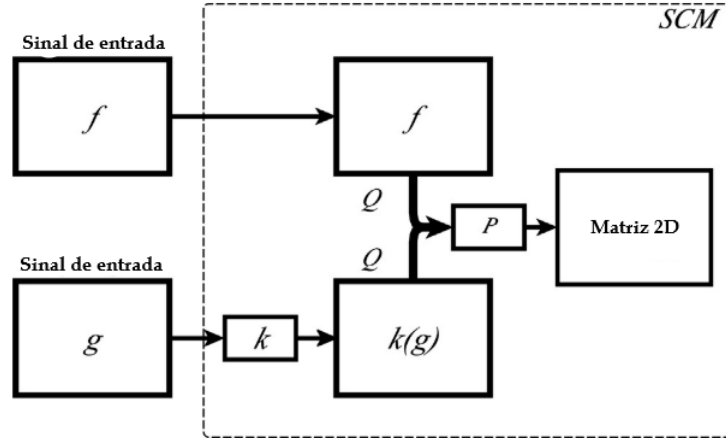
A etapa de extração de características consiste em identificar informações presentes em um sinal, seja ele 1D ou 2D. Essas informações, chamadas de atributos, descrevem quantitativamente ou qualitativamente um sinal. Essas técnicas foram desenvolvidas de diferentes perspectivas para reduzir a dimensionalidade com informações que representam as estruturas que compõem uma imagem.

3.2.1 Matriz de Coocorrência Estrutural

O método Matriz de Coocorrência Estrutural, SCM (do inglês, *Structural Co-occurrence Matrix*), desenvolvido por Ramalho et al. (2016), é um descritor que analisa o relacionamento entre estruturas de baixo nível de sinais no espaço N -dimensional. A informação resultante da análise do relacionamento estrutural entre os sinais é armazenado em uma matriz. O detalhamento do modelo do método é apresentado na Figura 4.

Considerando f e g sinais discretos, com os mesmos elementos e dimensões, eles podem assumir valores dos tipos inteiro, racional, real ou complexo. $k(g)$ é uma função de

Figura 4 – Modelo genérico do SCM.



Fonte: Adaptado de Ramalho et al. (2016).

degradação do sinal g , têm o objetivo de garantir diferença estrutural entre f e g . O que define o tipo de k é o problema a ser resolvido, ou seja, depende da informação estrutural do sinal de entrada que precisa ser capturada. Supondo que a entrada do SCM seja uma imagem (sinal 2D), então pode-se dizer que um valor na posição p é representado por f_p , e $k(g)_{p+d}$ dado um deslocamento d . Q é uma função que transforma a escala de tons de cinza da imagem para o domínio D_f com $I = 0, \dots, N - 1$ valores. A matriz 2D de coocorrência estrutural dos sinais $Q(f)$ e $Q(g)$, representada por $M = m_{ij}$ é calculada usando a seguinte expressão:

$$m_{ij} = \#\{(i, j) | P(i, j), i = Q(f)_p, j = Q(k(g))_{p+d}\}, \quad (3.1)$$

por sua vez $\#\{\cdot\}$ representa a cardinalidade do subconjunto de pares (i, j) que satisfaz a propriedade P . Essa propriedade retrata o relacionamento entre os valores $Q(f)_p$ e $Q(k(g))_{p+d}$ nas posições p e $p + d$, respectivamente.

Quando esse método é usado para analisar imagens, a intensidade dos pixels mudam as probabilidades da matriz SCM de diferentes formas. A Figura 5 indica como as informações estruturais ficam armazenadas na matriz de acordo com as características da imagem. Regiões de transições escura/clara e clara/escura descrevem as principais diferenças entre as estruturas das imagens de entrada. Por exemplo, se k for uma função que represente um filtro passa-baixa, a concentração das probabilidades ficarão em regiões que representem tons de cinza claro ou escuro. Se a imagem f for igual a $k(g)$ a concentração será diagonal. Para cada imagem analisada pelo método é gerado uma matriz M . A partir de M algumas maneiras podem ser usadas para representar/interpretar a informação estrutural da imagem, por exemplo: a distribuição numérica da própria matriz M e suas distribuições marginais, como também, os atributos SCM, detalhados na seção 3.1.1.1. Demonstradas na Figura 6, as distribuições marginais são dadas pela soma das linhas

Figura 5 – Aspectos da disposição das probabilidades na matriz M de acordo com as diferenças estruturais de f e $k(g)$.

Região Escura	Transição Escura/Clara
Transição Clara/Escura	Região Clara

Fonte: Adaptado de Ramalho et al. (2016).

ou colunas de M . Nesse trabalho, adotou-se chamar as marginais X e Y , de MX e MY , respectivamente. Um detalhamento prático do cálculo das marginais é apresentado no APÊNDICE A.

3.2.1.1 Atributos do SCM

A partir da matriz M , um conjunto de atributos numéricos podem ser calculados. Esses atributos são divididos em três grupos, grupo estatístico, grupo de informação e grupo divergente.

Os atributos do grupo estatístico são representados por COR , IDM e CHI . COR (do inglês, *Correlation*) é uma medida que calcula o nível de correlação da informação estrutural de M , usando

$$COR = \sum_{i,j=0}^{N-1} m_{ij} \frac{(i - \mu_i)(j - \mu_j)}{\sqrt{\sigma_i^2 \sigma_j^2}} \in [-1, 1], \quad (3.2)$$

onde μ_i e μ_j são a média dos valores das linhas e colunas de M . O σ_i e σ_j são os desvios padrão das linhas e colunas de M .

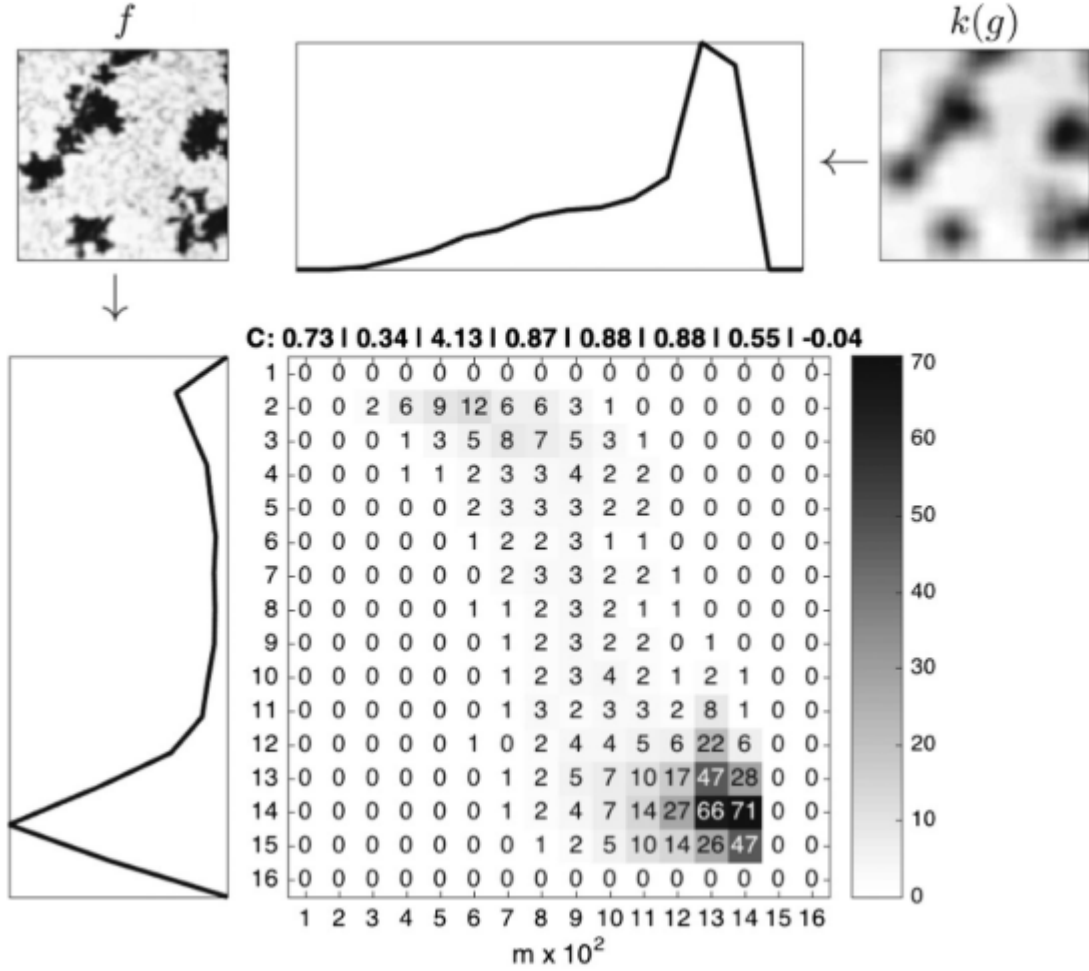
O atributo IDM (do inglês, *Inverse Difference Moment*) mede a homogeneidade da informação estrutural de M . Ele é definido por

$$IDM = \sum_{i,j=0}^{N-1} \frac{m_{ij}}{1 + (|i - j|)} \in [0, 1]. \quad (3.3)$$

O grupo de informações compreende um atributo que classifica a ordenação dos dados em M . Mede a aleatoriedade dos dados estruturais.

$$ENT = - \sum_{i,j=0}^{N-1} m_{ij} \log(m_{ij}) \quad (3.4)$$

Figura 6 – Matriz SCM ($N=16$) e suas marginais X e Y resultante da entrada de uma imagem f e função de degradação $k(g)$.



Fonte: Ramalho et al. (2016).

O grupo de divergência descreve a informação mútua ou dissimilaridade entre as distribuições marginais de M . CSD (do inglês, *Chi-Square Distance*) é definido como

$$CSD = \sum_{i=0}^{N-1} \frac{(P_i^0 - P_i^e)^2}{P_i^e}, \quad (3.5)$$

onde $P_i^0 = m_{ij} \forall i = j$ e $P_i^e = \sum_{j=0}^{N-1} m_{ij} \forall j \neq i$.

O atributo CSR (do inglês, *Chi-Square Distance Ratio*) mede as distribuições entre os quadrantes I, $P_i^I = \sum_{j=0}^{N/2-1} \sum_{k=N/2}^{N-1} m_{ijk}$ e III, $P_i^{III} = \sum_{j=N/2}^{N-1} \sum_{k=0}^{N/2-1} m_{ijk}$. Essa medida calcula a relação entre as transições de regiões escura e escura/clara. CSR é dado por

$$CSR = \sum_{i=0}^{N-1} \frac{(P_i^I - P_i^m)^2}{P_i^m}, \quad (3.6)$$

onde $P^m = (P^I + P^{III})/2$.

Os atributos a seguir medem a divergência entre as distribuições marginais de M definidas pela soma das linhas $P^r = \sum_{j=0}^{N-1} m_j$ e a soma das colunas $P^c = \sum_{i=0}^{N-1} m_i$. MDR é definido pela relação

$$MDR = \frac{MD_{min}}{MD_{max}}, \quad (3.7)$$

onde $MD_p = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P(S_i)P(S_j)|S_i - S_j|$ têm o valor esperado da diferença absoluta de duas variáveis aleatórias com a mesma distribuição P . S_i , $i = 0, \dots, N-1$ são índices de valores diferentes de zero das distribuições marginais. $MD_{min} = \min(MD_p^c, MD_p^r)$ e $MD_{max} = \max(MD_p^c, MD_p^r)$.

O D_{kl} (do inglês, *divergence of Kullback and Leibler*) mede a informação divergente ou informação ganha entre as distribuições marginais. Essa divergência é obtida usando

$$D_{kl} = \sum_{i=0}^{N-1} \log\left(\frac{P_i^c}{P_i^r}\right)P_i^c, \quad (3.8)$$

onde $P_i^c/P_i^r = 0$ onde $P_i^r = 0$.

O CAD (do inglês, *Complementary Absolute Difference*) compara duas distribuições de probabilidade, em que uma variável é derivada da outra por pequenas perturbações não-uniformes (aleatórias ou determinísticas). Esse atributo é calculado por

$$CAD = 1 - \sum_{i=0}^{N-1} |P_i^c - P_i^r| \quad (3.9)$$

3.2.2 Matriz de Coocorrência de Níveis de Cinza

Haralick, Shanmugam e Dinstein (1973) apresentaram um procedimento geral para extrair propriedades de textura de uma imagem. As características são calculadas no domínio espacial, de onde parte a premissa de que a informação de textura de uma imagem I está contida na relação geral de espaçamento que os tons de cinza têm entre si. Por meio de uma Matriz de Coocorrência de Níveis de Cinza, GLCM (do inglês, *gray level co-occurrence matrix*), calcula-se a dependência espacial dos tons de cinza de uma imagem, e da matriz pode ser abstraído 4 atributos de informações. Os atributos são descritos abaixo. Para cada item nas equações a seguir:

- $p(i, j)$: (i, j) – ésimas entrada em uma matriz normalizada de tons de cinza $g(i, j)/N$;
- $p_x(i)$: i – ésimas entrada em uma matriz das probabilidades marginais obtidas pela soma das linhas de $p(i, j)$, $\sum_{j=1}^L p(i, j)$;
- L : Número de tons de cinza da imagem quantizada;
- $\sum i$ e $\sum j$: forma simplificada de $\sum_i^L$ e $\sum_j^L$
- $p_{x+y}(k) = \sum_{i=1}^L \sum_{j=1}^L p(i, j)$ para $k = 2, 3, \dots, 2L$;

- $p_{x-y}(k) = \sum_{i=1}^L \sum_{j=1}^L p(i, j)$ para $k = 0, 1, \dots, 2L - 1$.
- Segundo Momento Angular:

$$f_1 = \sum_i \sum_j [p(i, j)]^2. \quad (3.10)$$

- Contraste:

$$f_2 = \sum_{k=0}^{L-1} k^2 [p_{x-y}(k)]. \quad (3.11)$$

- Correlação:

$$f_3 = \frac{\sum_i \sum_j (i, j) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}, \quad (3.12)$$

onde μ_x , μ_y , σ_x e σ_y são média e desvio padrão de p_x e p_y .

- Homogeneidade:

$$f_4 = \sum_{i,j} \frac{p(i, j)}{1 + |i - j|}. \quad (3.13)$$

3.2.3 Estatística de baixa e alta ordem

Estatísticas de baixa ordem podem ser usadas para descrever textura e são bastante aplicadas em problemas de classificação e recuperação de imagens. Características do histograma como média, variância, assimetria e curtose são calculados usando a distribuição de probabilidade dos níveis de intensidade de um histograma (MALIK; BAHARUDIN, 2013).

Dado que $P(b)$ é a distribuição de probabilidade de um intervalo b . Ele é definido como:

$$P(b) = \frac{H(b)}{M} \quad (3.14)$$

onde M é o total do número de pixels de uma imagem I . O valor obtido a partir do cálculo da média pode fornecer características do brilho de uma imagem. Se a média é alta, significa que a imagem é clara, caso contrário, é escura. A média pode ser definida como:

$$Média = \sum_{b=1}^L b P(b) \quad (3.15)$$

O desvio padrão é o momento de segunda ordem, e mostra o contraste das intensidades dos tons de cinza. O desvio padrão alto indica alto contraste, caso contrário, o contraste é baixo. Pode ser calculado como:

$$Desvio = \sqrt{\sum_{b=1}^L (b - Média)^2 P(b)} \quad (3.16)$$

O momento de terceira ordem indica a assimetria dos valores de intensidade do histograma. O valor negativo mostra que grande parte dos valores estão do lado direito da média. O valor positivo mostra que as intensidades estão boa parte do lado esquerdo da média. O valor zero indica igual distribuição de intensidade em ambos os lados. A assimetria é definido como:

$$Assimetria = \frac{1}{(Desvio)^3} \sum_{b=1}^L (b - Média)^3 P(b) \quad (3.17)$$

O momento de quarta ordem é representado pelo curtose; ele é usado para medir o pico da distribuição dos valores de intensidade em torno da média.

$$Curtose = \frac{1}{(Desvio)^4} \sum_{b=1}^L (b - Média)^4 P(b) \quad (3.18)$$

Para as estatísticas de alta ordem o conjunto de características é representado por uma composição do 2° ao 11° momentos centrais. No restante do trabalho a referência aos dois conjuntos de características segue o padrão HL (do inglês, *histogram-lower*) e HH (do inglês, *histogram-higher*).

3.2.4 Padrões Binários Locais

Proposto por Ojala, Pietikäinen e Harwood (1996), o método Padrões Binários Locais, LBP (do inglês, *Local Binary Patterns*) é um extrator que analisa a vizinhança de pixels, classificando cada pixel de uma forma binária. A anotação de cada vizinho é feita a partir do valor do pixel central onde ele é visto como um limiar. Assim, o valor de pixel acima do limiar pode ser levado para 0 ou 1.

Considere uma imagem $I(x, y)$ e g_c o valor de tons de cinza de um pixel qualquer (x, y) , por exemplo $g_c = I(x, y)$. Além disso, g_p denota o valor de um ponto de amostragem em uma vizinhança circular uniformemente espaçada de pontos de amostragem P e raio R ao redor do ponto (x, y) :

$$g_p = I(x_p, y_p), \quad p = 0, \dots, P-1 \quad \text{e} \quad (3.19)$$

$$x_p = x + R \cos(2\pi p/P), \quad (3.20)$$

$$y_p = y - R \sin(2\pi p/P). \quad (3.21)$$

Assumindo que a textura local de uma imagem $I(x, y)$ é caracterizada pela distribuição conjunta dos valores de tons cinza dos pixels $P + 1$ ($P > 0$):

$$T = t(g_c, g_0, g_1, \dots, g_{p-1}). \quad (3.22)$$

Sem perda de informação, o valor do pixel central pode ser subtraído da vizinhança:

$$T = t(g_c, g_0 - g_c, g_1 - g_c, \dots, g_{p-1} - g_c). \quad (3.23)$$

No próximo passo, a distribuição conjunta é aproximada assumindo que o pixel central seja estatisticamente independente das diferenças, o que permite a fatoração da distribuição:

$$T \approx t(g_c)t(g_0 - g_c, g_1 - g_c, \dots, g_{p-1} - g_c). \quad (3.24)$$

Agora o primeiro fator $t(g_c)$ é a intensidade de distribuição sobre $I(x, y)$. Do ponto de vista da análise de padrões texturais locais, ele não contém informações úteis. Em vez disso, a distribuição conjunta das diferenças

$$t(g_0 - g_c, g_1 - g_c, \dots, g_{p-1} - g_c) \quad (3.25)$$

pode ser usado como modelo da textura local. A abordagem baseada na quantização de vetores de aprendizagem ainda possui certas propriedades:

$$t(s(g_0 - g_c), s(g_1 - g_c), \dots, s(g_{p-1} - g_c)) \quad (3.26)$$

onde $s(z)$ é dado pela função limiar

$$s(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0. \end{cases} \quad (3.27)$$

Por fim, o LBP genérico é definido como:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p. \quad (3.28)$$

3.2.5 Filtro de Gabor

O sexto conjunto de características usado nesse trabalho foi baseado no filtro de Gabor (DUNN; HIGGINS, 1995). Ele têm sido largamente utilizado na construção de

aplicações de reconhecimento de padrões. A vantagens desse método são as invariações a escala, iluminação e translação. No domínio espacial, um filtro Gabor bidimensional é uma função de núcleo gaussiano modulada por uma onda plana senoidal complexa, definida como (HAGHIGHAT; ZONOUZ; ABDEL-MOTTALEB, 2013):

$$G(x, y) = \frac{f^2}{\pi\gamma\eta} \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp(j2\pi f x' + \phi) \quad (3.29)$$

onde $x' = x \cos \theta + y \sin \theta$, $y' = -x \sin \theta + y \cos \theta$, onde, f é a frequência do fator senoidal, θ representa a orientação do normal para as listras paralelas de uma função Gabor, ϕ é o deslocamento de fase, σ é o desvio padrão Gaussiano e γ é a razão de aspecto espacial que especifica a elipticidade do suporte da função Gabor.

3.2.6 Características perceptuais do sistema visual humano

Mais um método utilizado neste trabalho é baseado na percepção de imagens a partir do sistema visual humano. Tamura, Mori e Yamawaki (1978) mostraram que o sistema visual humano discrimina textura por meio de vários atributos específicos que foram depois aprimorados e testados por Bianconi, Álvarez-Larrán e Fernández (2015). As características usadas foram:

- **Aspereza:** O conceito de aspereza está relacionado ao tamanho dos elementos de textura, quanto maior o tamanho, mais grossa é a textura e vice-versa. O cálculo desse recurso ocorre primeiramente aplicando um conjunto de filtros da média à imagem de entrada. Cada filtro é definido por uma janela quadrática de dimensão $2^k \times 2^k$, onde $k \in \{1, \dots, K\}$ e A_k indica a k -ésima transformação de imagem resultante nesse passo. No segundo passo, aplica-se, a cada A_k , uma máscara de diferença vertical e horizontal, que atribui, a cada pixel, a diferença entre os valores dos dois pixels simétricos que se encontram vertical ou horizontalmente na distância 2^{k-1} . Sejam as matrizes resultantes correspondentes às direções horizontal e vertical para cada valor de k , $\mathbf{E}_{k,h}$ e $\mathbf{E}_{k,v}$, respectivamente. Agora, pesquisa-se o valor de k que minimiza, em cada ponto, o valor de $\mathbf{E}_{x,y}$ em qualquer direção, por exemplo:

$$\bar{k}(x, y) = \arg \max_{k \in \{1, \dots, K\}} \{\mathbf{E}(x, y)_{k,h}, \mathbf{E}(x, y)_{k,v}\}. \quad (3.30)$$

Finalmente a Aspereza é definida como:

$$F_{crs} = \frac{1}{2^K} \frac{1}{WH} \sum_{x=1}^H \sum_{y=1}^W 2^{\bar{k}(x,y)}, \quad (3.31)$$

onde H e W são dimensões da imagem de entrada.

- **Contraste:** De acordo com Tamura, Mori e Yamawaki (1978) o contraste depende da distribuição (histograma) dos níveis de cinza, da nitidez das bordas e do período de repetição dos padrões. É definido através da seguinte expressão:

$$F_{con} = 2 \frac{\sigma}{\alpha_4^n}, \quad (3.32)$$

onde σ e α_4 , são, respectivamente, o desvio padrão e a curtose da distribuição dos níveis de cinza.

- **Direcionalidade:** A direcionalidade está relacionada à probabilidade de que a variação das intensidades dos pixels ocorra ao longo de certas orientações predefinidas. Uma imagem composta principalmente por linhas paralelas terá direcionalidade forte, por outro lado, um composto de pontos quase aleatoriamente dispersos terá direcionalidade fraca. Para estimar esse parâmetro, primeiro é aplicado um filtro de Sobel, vertical e horizontal, para calcular o gradiente da imagem em cada ponto, referenciados como G_x e G_y . Depois é calculado a orientação do gradiente em cada pixel $\theta(x, y)$ através da Eq. 4.40, descartando os pixels em que a magnitude do gradiente é zero:

$$\theta(x, y) = \tan \left[\frac{G_y(x, y)}{G_x(x, y)} \right]. \quad (3.33)$$

É feito uma quantização do valor de θ em N_o valores discretos, cada um representando um intervalo angular de amplitude $2\pi/N$. Dessa forma, a distribuição de probabilidade é aproximada da orientação do gradiente por meio de um histograma de N_o intervalos. Uma distribuição quase horizontal será sugestiva de uma imagem fracamente orientada, enquanto uma distribuição concentrada em apenas um ou poucos picos indicará uma imagem fortemente orientada. Para quantificar esse valor, é calculada a distância euclidiana entre a distribuição estimada e a distribuição uniforme, que idealmente representa uma imagem não direcional: quanto maior a distância, maior a direcionalidade e vice-versa. A equação é definida como:

$$F_{dir} = \sqrt{\frac{N_o}{N_o - 1} \sum_{i=1}^{N_o} \left[p(\theta_i) - \frac{1}{N_o} \right]^2}, \quad (3.34)$$

onde $p(\theta_i)$ é a probabilidade de cada orientação θ_i . O fator $N_o/(N_o - 1)$ é o inverso da distância teórica máxima para a distribuição uniforme; portanto, normaliza-se F_{dir} no intervalo $[0,1]$

- **Probabilidade de linha:** Este parâmetro está relacionado à probabilidade do gradiente manter a mesma direção nos pixels vizinhos. Quando a direção permanece quase constante, considera-se esse grupo de pixels como uma linha. Portanto, a probabilidade de linha pode ser avaliada através da probabilidade conjunta das

direções de gradiente calculadas em pixels separados por um determinado vetor de deslocamento. Quanto maior a probabilidade dos dois vetores de gradiente serem colineares, maior a probabilidade de linha e vice-versa. Numericamente esse parâmetro é expressado como:

$$F_{lin,d} = \frac{\sum_{i=1}^{N_o} \sum_{j=1}^{N_o} |\cos \left[(i-j) \frac{2\pi}{N_o} \right]| \mathbf{CM}_{\theta,d}}{\sum_{i=1}^{N_o} \sum_{j=1}^{N_o} \mathbf{CM}_{\theta,d}}, \quad (3.35)$$

onde $\mathbf{CM}_{\theta,d}$ é a matriz de coocorrência da direção do gradiente para um dado deslocamento de vetor $\mathbf{d}$. O cosseno na Eq. 4.42 atua como peso que assume valor um quando os dois vetores de gradiente são colineares. Para obter invariância contra a rotação, $F_{lin,d}$ é calculado em quatro vetores circularmente orientados e os resultados são calculados como uma média:

$$F_{lin} = \frac{1}{4} \sum_{l=1}^4 F_{lin,d_l}, \quad (3.36)$$

onde por exemplo, $\mathbf{d}_1 = (0, d)$, $\mathbf{d}_2 = (-d, d)$, $\mathbf{d}_3 = (-d, 0)$ e $\mathbf{d}_4 = (-d, -d)$.

- Rugosidade: Essa característica é geralmente considerada um conceito relacionado ao acabamento de uma superfície. Com o termo "áspero", quer dizer uma superfície marcada por protuberâncias, cordilheiras e vales - em oposição a uma superfície "lisa", livre de tais irregularidades. Quando se trata de imagens, esse conceito indica o grau e a taxa de variabilidade da intensidade em uma vizinhança: uma superfície áspera mostrará mudanças acentuadas e abruptas na intensidade; uma suave apresenta-se uniforme. Esse parâmetro é definido como:

$$F_{rgh} = 2 \sqrt{\frac{1}{WH} \sum_{x=1}^H \sum_{y=1}^W [\mathbf{I}(x, y) - R]^2}, \quad (3.37)$$

onde $\mathbf{I}$ é a imagem de entrada em tons de cinza, e R um fator de correção.

3.3 Técnicas de aprendizagem de máquina

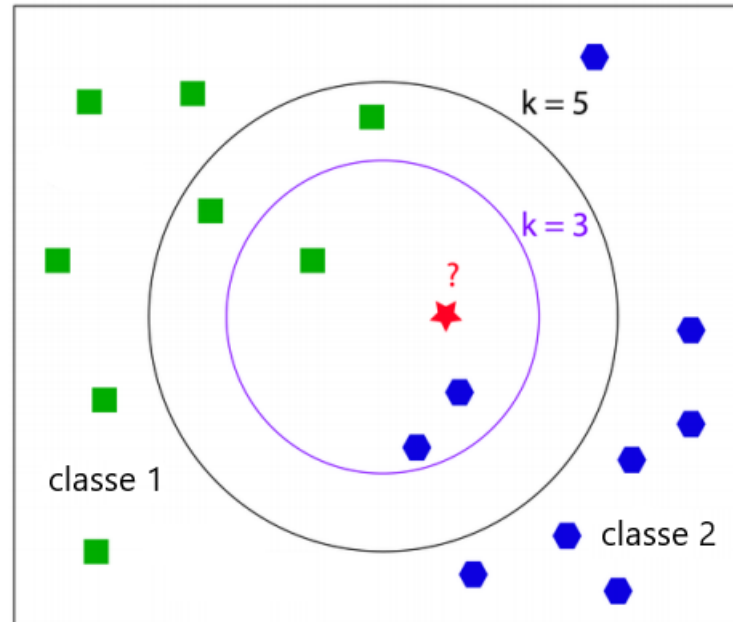
Nesse trabalho, a escolha do classificador foi baseada em métodos utilizados por outros autores (SALAMA; ELTRASS; ELKAMCHOUCI, 2018; SINGH; SRIVASTAVA; SRIVASTAVA, 2017; PANDA et al., 2015) para resolução de problema de classificação de imagens médicas e no princípio de funcionamento desses métodos.

3.3.1 K-vizinhos mais próximos

O K-vizinhos mais próximos, KNN (do inglês, *K-Nearest Neighbors*) é um dos métodos mais simples e antigos na área de classificação de padrões. Esse método de aprendizado é baseado na distância dos seus vizinhos e pode ser do tipo supervisionado e não

supervisionado. O aprendizado supervisionado divide-se em dois tipos: classificação e regressão.

Figura 7 – Ilustração do processo de classificação do KNN.



Fonte: Adaptado de Fisher (1922).

A classificação do KNN acontece da seguinte forma: dado um conjunto C de treinamento, cada elemento de C é um padrão $x = (x_1, x_2, \dots, x_n, c)$, onde c é classe de um padrão x qualquer. O padrão x pode ser considerado como um ponto em um espaço n -dimensional $\mathbb{R}_n$. Seja y um padrão ainda não classificado, o KNN calcula as distâncias de y para todos os padrões de treinamento e considera os k padrões mais próximos de y . Dentre os k padrões mais próximos, será o vencedor, o padrão com o maior número de ocorrências. A classe do padrão vencedor será atribuída a y (FISHER, 1922).

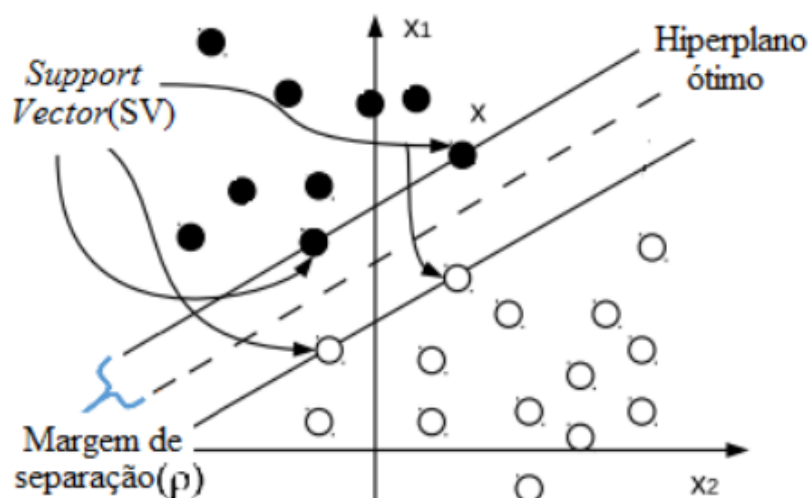
A Figura 7 apresenta um exemplo de como o processo de classificação do KNN é abordado. Seja o ponto vermelho um padrão y , no espaço 2-dimensional, de classe ainda não identificada. A distância entre y e os demais pontos é calculada por meio de uma métrica de similaridade. Se o valor é de $k=3$, serão observadas as classes dos 3 pontos mais próximos de y . A classe atribuída a y será a que se repetir o maior número de vezes.

As métricas de similaridades mais comuns utilizadas pelo KNN são: distância Euclidiana, distância de *Minkowsky*, distância *Mahalanobis* e distância de *Hamming*.

3.3.2 Máquina de Vetores de Suporte

A técnica Máquina de Vetores de Suporte, SVM (do inglês, *Support Vector Machine*) busca criar um hiperplano representado por uma função de aproximação que separa classes de dados através da classificação ou regressão. Esse classificador de aprendizado

Figura 8 – O hiperplano ótimo, construído pelo SVM, separando os dados com a máxima margem ρ . Os pontos apontados pelas setas são os vetores de suporte.



Fonte: Adaptado de Andreola (2009).

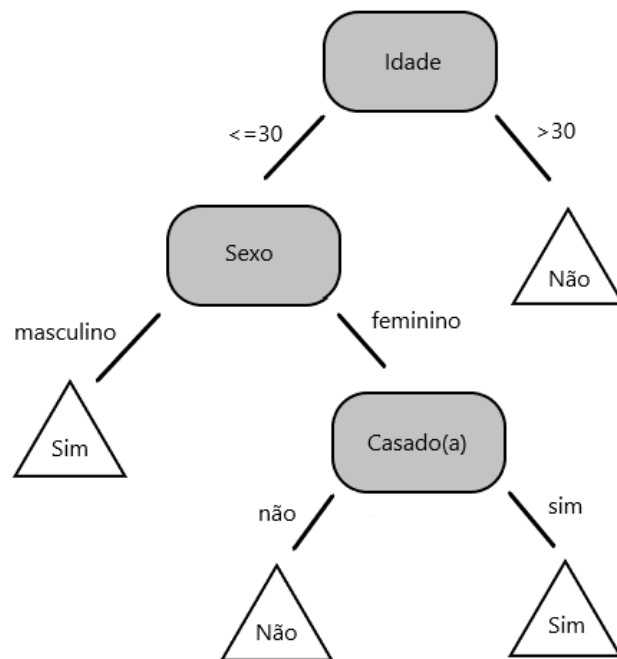
supervisionado, analisa os dados e reconhece padrões, adquirindo na etapa de treinamento a capacidade de generalização (HAYKIN, 2007). Desenvolvida por Vapnik (1999), a Teoria de Aprendizado Estatístico (TAE), é o princípio de funcionamento do SVM. O método visa a minimização do risco estrutural (do inglês, Structural Risk Minimization - SRM) e estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização.

Muitos classificadores não encontram a fronteira ótima para separação das classes do problema. O SVM, por sua vez, classifica os dados encontrando o melhor hiperplano que diferencie duas classes por meio da separação dos seus pontos. O melhor hiperplano para um SVM significa aquele com a maior margem entre as duas classes (THEODORIDIS; KOUTROUMBAS, 1999). A definição da fronteira entre os pontos de duas classes diferentes é ilustrada na Figura 8.

No contexto de SVM, existem também as funções *kernel*. Essas funções têm a finalidade de projetar os vetores de características em um espaço de características de alta dimensão para classificação de problemas que se encontram em espaços não separáveis linearmente. Vários tipos de *kernels* podem ser usados, porém os mais conhecidas são: Polinomial, RBF (do inglês, *Radial basis function*) e linear.

As principais vantagens desse método são: robustez para lidar com a presença de ruído nos dados; e a indução de um hiperplano de decisão que pode ser descrito em termos de uma quantidade relativamente pequena de exemplos de entrada. (CRAMMER; SINGER, 2001).

Figura 9 – Ilustração do processo de classificação do DT.



Fonte: Adaptado de Rokach e Maimon (2005).

3.3.3 Árvores de Decisão

Árvores de Decisão, DT (do inglês, *Decision Tree*) é um classificador expresso como uma partição recursiva de um espaço amostral. Ele consiste em nós que formam uma árvore enraizada, ou seja, é uma árvore direcionada com um nó chamado raiz que não possui arestas de entrada (ROKACH; MAIMON, 2005). Todos os outros nós têm exatamente uma entrada. Um nó com arestas de saída é chamado nó interno. Todos os outros nós são chamados de folhas, também conhecidos como nós terminais. Cada nó interno divide o espaço em dois ou mais subespaços de acordo com uma função discreta dos valores dos atributos de entrada. No caso mais simples, cada nó interno representa um atributo, ao modo que o espaço amostral é particionado de acordo com o seu valor. No caso de atributos numéricos, a condição refere-se a um intervalo. Cada folha representa uma classe do problema.

Cada amostra é introduzida na árvore por meio da raiz e navega até um nó folha obedecendo as condições de cada aresta visitada no caminho. A Figura 9 descreve uma árvore de decisão que classifica se um candidato é adequado ou não a uma vaga de emprego. Observe que essa árvore de decisão incorpora atributos nominais e numéricos. Nesse caso, o analista pode prever a resposta de um funcionário em potencial e entender as características comportamentais dos candidatos. Cada nó é rotulado com o atributo do conjunto amostral e suas ramificações são rotuladas com seus valores correspondentes.

Algumas vantagens fazem com que esse classificador seja bastante utilizado: capaz de lidar com dados numéricos e categóricos; simples de entender e interpretar (JAMES et al., 2013).

3.4 Conclusão

Neste capítulo foram apresentados os conceitos básicos de sete extratores de características e três métodos de classificação. Estes métodos foram objetos de estudo para o desenvolvimento da abordagem de classificação multiclasse do tecido CRC. Em particular, o SCM é um método que vêm apresentando bom desempenho como descritor de textura em diversos problemas. Peixoto e Filho (2018), Filho et al. (2017), Filho et al. (2018) aplicaram essa técnica para caracterizar imagens de acidente vascular cerebral, câncer de pele e câncer de pulmão, respectivamente. Nos trabalhos estudados o método apresentou bom desempenho como descritor de textura e mostrou ser computacionalmente eficiente em relação a outras técnicas. Em relação aos métodos de classificação, o SVM foi usado pois alcançou desempenho superior a outras técnicas na resolução de problemas similares ao abordado nesse trabalho, e os métodos KNN e DT devido a simplicidade dos algoritmos, indicados para trabalhar com grandes volumes de dados.

4 Metodologia

Nesse capítulo são explicados os materiais usados e os procedimentos necessários para o desenvolvimento desta pesquisa. Para isso, é descrita a base de imagens analisada, o pré-processamento aplicado as imagens, o processo de construção da abordagem de classificação multiclasse do tecido colorretal e como os resultados são avaliados.

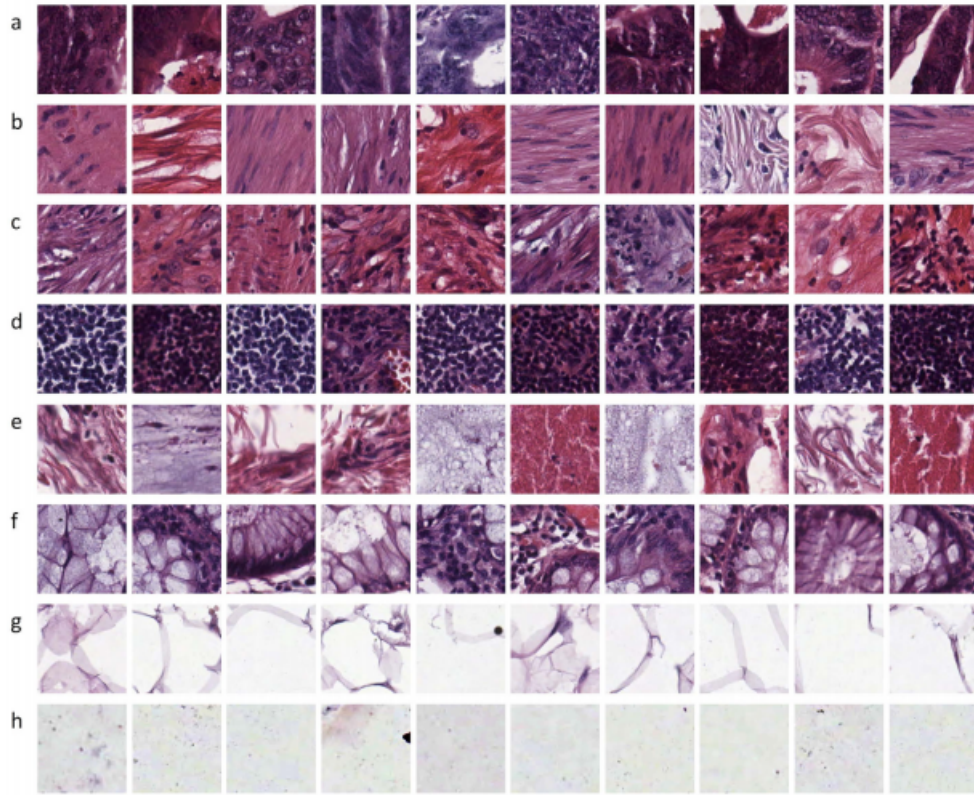
4.1 Base de imagens

A base de imagens utilizada nesse trabalho foi apresentada e discutida por Kather et al. (2016). Ela é composta por recortes realizados de diferentes partes de exames H&E de câncer colorretal do arquivo patológico da University Medical Center Mannheim. Os recortes incluem oito tipos de classes de tecidos: tumor epitelial; estroma simples; estroma complexo; células imunes; detritos (incluindo necrose, hemorragia e muco); glândulas mucosas normais; tecido adiposo e fundo do exame. Cada linha da Figura 10 apresenta dez amostras dos diferentes tecidos colorretal. Para cada tipo de tecido a base possui 625 recortes, não sobrepostos, de dimensões $150\text{px} \times 150\text{px}$ ($74\mu\text{m} \times 74\mu\text{m}$), juntos, totalizando $8 \times 625 = 5000$ imagens. Os grupos de tecidos diferem entre si em iluminação, intensidade de manchas e texturas. Nos trabalhos relacionados, as classes estroma simples e estroma complexo são responsáveis por apresentar a maior parte do erro de classificação, portanto, nesse trabalho são usadas todas as imagens, considerando as oito classes de tecidos, incluindo estroma simples e estroma complexo. Os tumores do câncer colorretal (CRC) são estruturas complexas que normalmente contêm vários tipos de tecidos distintos. Esses diferentes tecidos podem ser distinguidos pela avaliação manual histopatológica em imagens com coloração Hematoxilina e Eosina (H&E). A arquitetura do tumor muda durante sua progressão no corpo, e é de extrema importância verificar seu estado durante o prognóstico do paciente. Quantificar a composição do tecido do CRC é, portanto, uma tarefa relevante na histopatologia.

4.2 Pré-processamento

A análise estrutural de imagens é caracterizada por analisar os tons dos pixels de uma imagem e sua vizinhança (RAMALHO et al., 2016). Nesse trabalho, os descritores adotados caracterizam imagens analisando tons de cinza dos pixels. Portanto, o primeiro passo experimental é converter todas as imagens da base, descrita na seção anterior, do

Figura 10 – Amostras H&E de diferentes estruturas do tecido colorretal. (a) Tumor epitelial, (b) Estroma simples, (c) Estroma complexo, (d) Células imunes, (e) Detritos, (f) Glândulas mucosas, (g) Tecido adiposo e (h) Fundo.



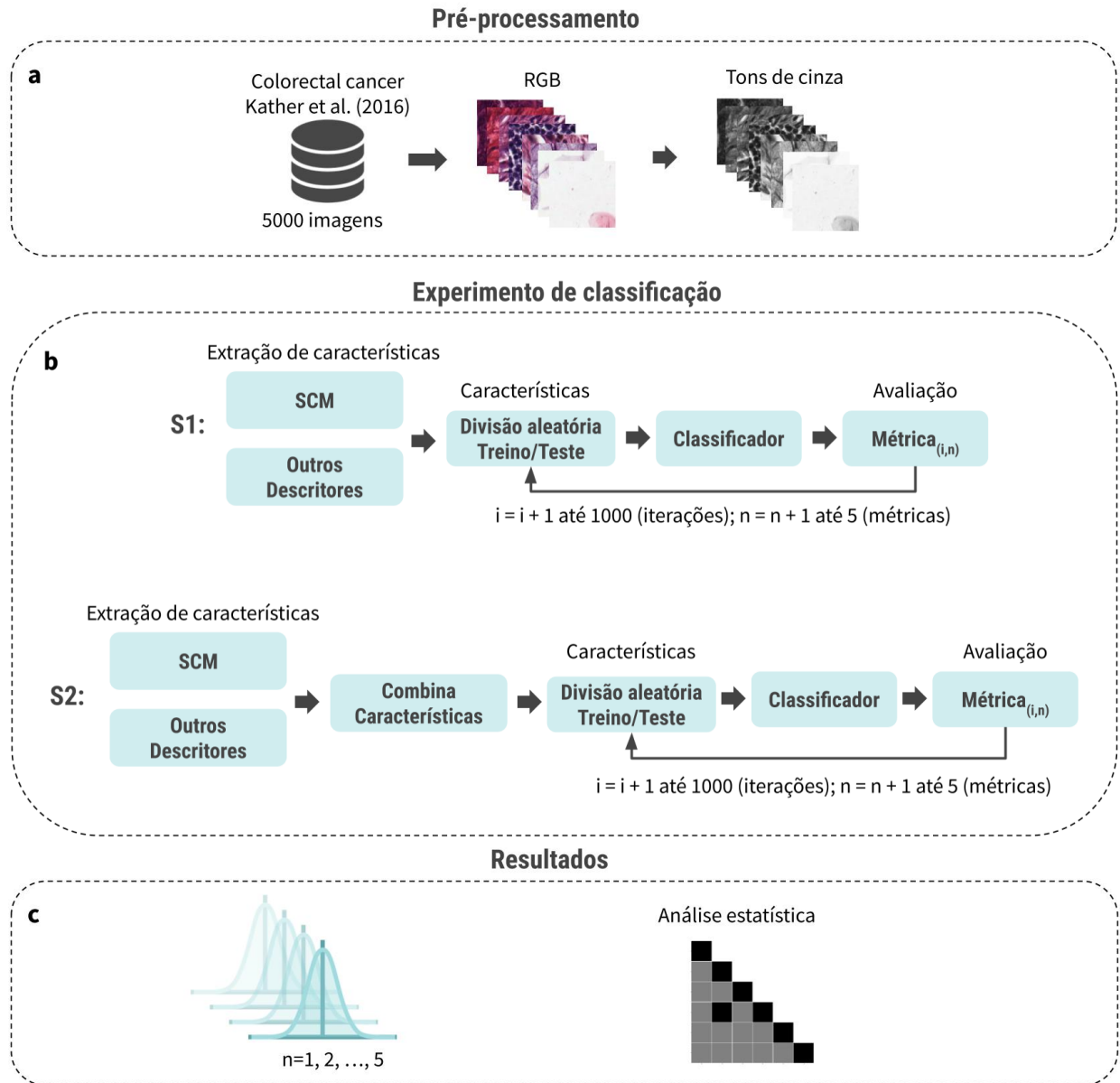
Fonte: Kather et al. (2016).

formato RGB para tons de cinza, usando a equação:

$$GS = 0.298 * R + 0.5870 * G + 0.1140 * B \quad (PRATT, 1978), \quad (4.1)$$

onde GS é a imagem em tons de cinza e R (*Red*), G (*Green*), B (*Blue*) são componentes das imagens de entrada. O pré-processamento é descrito na Figura 11a.

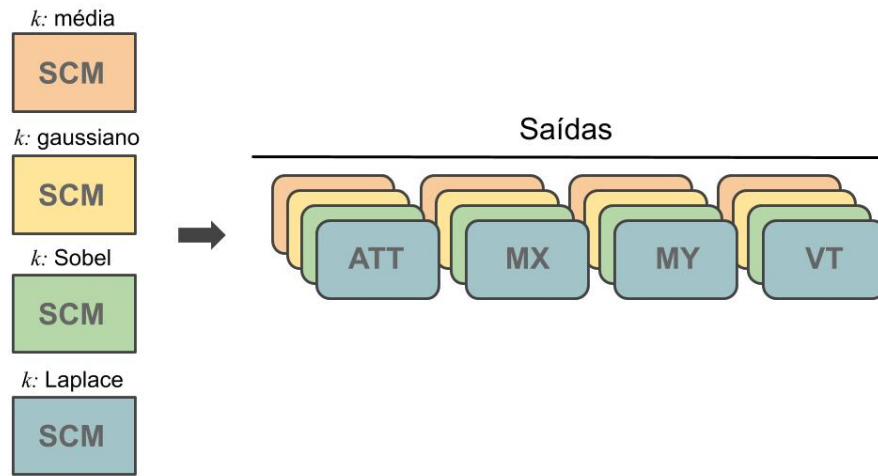
Figura 11 – Abordagem proposta para classificação multiclasse do CRC. n representa as diferentes métricas de avaliação empregadas. S1 e S2 corresponde aos dois cenários explorados.



4.3 Experimento de classificação

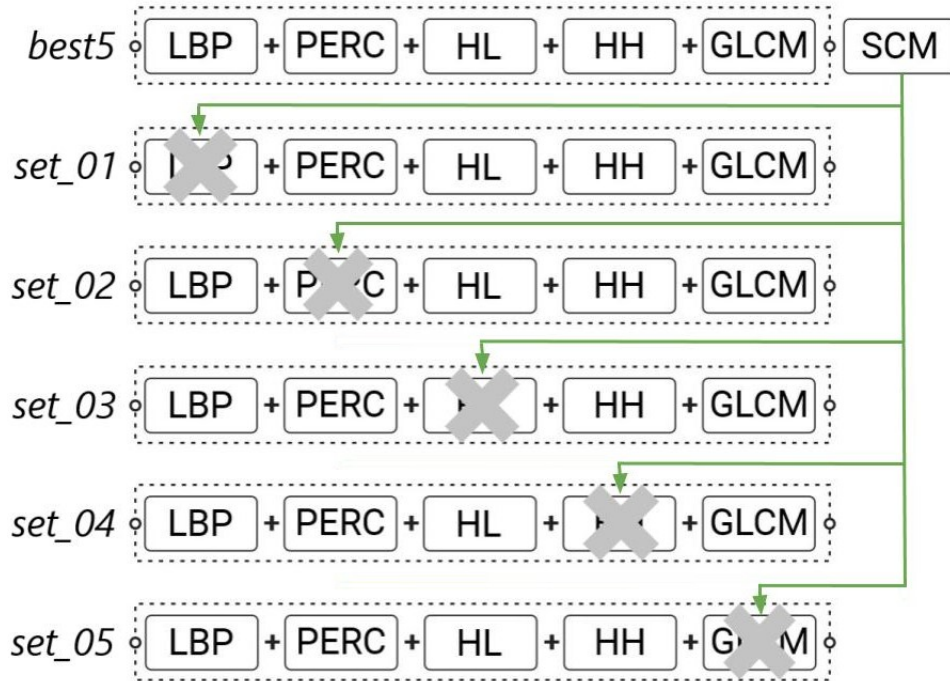
O experimento de classificação, detalhado da Figura 11b, é dividido em dois momentos, os cenários S1 e S2. No cenário S1 a classificação do tecido é feita com base nas características extraídas com o SCM, e posteriormente o desempenho de caracterização individual desse extrator é comparado com outros seis métodos. Nesse cenário, antes da comparação entre os métodos, um estudo é desenvolvido com o SCM buscando investigar os parâmetros desse método que maximizam o desempenho de discriminação dos tecidos colorretais. Portanto, experimentos são desenvolvidos variando a função de degradação

Figura 12 – Abordagem desenvolvida para buscar a melhor configuração do SCM na caracterização multiclasse do CRC.



k com objetivo de identificar um filtro que melhor revele saliências estruturais no tecido colorretal. Para isso, os seguintes filtros são analisados como função k : {passa-baixa: média e gaussiano}, e {passa-alta: sobel e laplace}. Esses filtros foram usados juntos com o SCM em outros trabalhos (PEIXOTO; FILHO, 2018; FILHO et al., 2017; FILHO et al., 2018) e se adequaram bem a alguns problemas, dependendo das imagens analisadas. Diferentes saídas do SCM também são usadas para verificar quais proporcionam a melhor representação das características extraídas das imagens. Então, são criados diferentes conjuntos de características a partir da aplicação do SCM: os atributos da SCM (ATT); as marginais X e Y da SCM (MX e MY), propostos em Ramalho et al. (2016); e a vetorização da matriz SCM (VT). A forma como cada um desses conjuntos foram criados é detalhado no APÊNDICE A. Com a vetorização, a matriz SCM (d, d), onde d representa a dimensão dos eixos, é transformada em uma matriz de dimensão $(1, d*d)$. Detalhes da abordagem de busca da melhor configuração SCM são apresentados na Figura 12. Cada saída apresentada na Figura 12 constitui um conjunto de características e são investigados de forma individual. Vale ressaltar que o tempo computacional do SCM está relacionado a função Q (Equação 3.1), que define uma nova escala de tons de cinza para as imagens analisadas pelo método. Para avaliar três bases imagens com características bem distintas entre si, Ramalho et al. (2016) usou uma função Q que mapeou a escala de tons de cinza das imagens de entrada para $N=8$ níveis, e obteve resultados comparáveis ao estado da arte. Nesse trabalho adota-se usar também $N=8$, pois se lhe for atribuído um valor menor, a sensibilidade do método pode ser diminuída. Caso esse valor seja maior, o número de operações para calcular a matriz de coocorrência aumenta, como também, influencia no tamanho final de alguns conjuntos de características, por exemplo, os conjuntos MX, MY e VT, apresentados na Figura 12. O tamanho dos conjuntos de características pode impactar no custo computacional de uso do método.

Figura 13 – Criação dos conjuntos de características combinados com base no SCM.



Após a extração de características das 5000 imagens com diferentes configurações do SCM, e com os métodos de referência para comparação, cada conjunto de características gerado é dividido de forma aleatória nos subconjuntos de treino (75%) e teste (25%). Esses subconjuntos são submetidos individualmente a um classificador para o processo de aprendizagem das características e depois testado com o conjunto de teste. A etapa de teste do classificador é avaliada por um conjunto de métricas. Os passos do cenário S1, divisão treino/teste, classificação e avaliação são repetidos 1000 vezes a fim de diversificar a cada rodada as amostras presentes nos conjuntos de treino e teste. Os resultados são analisados a partir da média obtida das 1000 amostras de cada métrica.

O passo **Combina Características**, descrito na Figura 11b, é a diferença entre os cenários S1 e S2. No cenário S2, o melhor conjunto de características do SCM, obtido no S1, é combinado com o melhor conjunto de características encontrado na literatura para classificação do tecido colorretal em oito classes, chamado de *best5*. O *best5* é gerado a partir da combinação de características de cinco extratores de textura. Os conjuntos de características gerados da combinação entre o SCM e o *best5* são descritos na Figura 13. O objetivo é verificar se a combinação de características de outros extratores com o SCM influenciam no aumento de desempenho de classificação. A Tabela 2 também ajuda no entendimento do experimento desenvolvido nos cenários S1 e S2.

Alguns algoritmos de aprendizagem de máquina, bem conhecidos na literatura, são usados para classificar as características extraídas. São eles: Árvores de Decisão (ROKACH; MAIMON, 2005), KNN (FISHER, 1922), e SVM com *Linear kernel*, *Polynomial kernel*

Tabela 2 – Formato experimental para os cenários S1 e S2.

Cenário	Estratégia	Métodos de Referência
S1	Classificação de amostras de tecido colorretal com base apenas no SCM.	O desempenho do SCM é comparado com GLCM; LBP; Gabor; informação perceptual do sistema visual humano (PERC); e características de histograma de baixa ordem (HL) e alta ordem (HH).
S2	Classificação de amostras do tecido colorretal baseado na combinação do SCM com outros descritores. São realizados testes estatísticos para identificar descritores em potencial para uma solução baseada no SCM.	Baseado na combinação de cinco técnicas chamada de <i>best5</i> (PERC + HH + HL + LBP + GLCM). O <i>best5</i> foi proposto por Kather et al. (2016) para classificação multiclasse do tecido colorretal e corresponde ao estado da arte.

(Poly) e *Radial Basis Function kernel* (RBF) (BURGES, 1998; CORTES; VAPNIK, 1995). A escolha do SVM foi baseada na observação do seu bom desempenho para problemas similares, e os métodos KNN e Árvores de Decisão devido a simplicidade dos algoritmos, indicados a trabalhar com grandes volumes de dados.

4.4 Avaliação dos resultados

As métricas utilizadas para avaliar e quantificar os resultados desse trabalho são baseadas nos valores de Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN) obtidos a partir da matriz de confusão. As métricas computadas são a acurácia (ACC), precisão ($PREC$), revocação (REC), especificidade ($SPEC$), e F_1 :

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}, \quad (4.2)$$

$$PREC = \frac{VP}{VP + FP}, \quad (4.3)$$

$$REC = \frac{VP}{VP + FN}, \quad (4.4)$$

$$SPEC = \frac{VN}{VN + FP}. \quad (4.5)$$

A métrica F_1 combina $PREC$ e REC para calcular um único número que indica a qualidade geral do modelo.

$$F_1 = 2 * \frac{PREC * REC}{PREC + REC}. \quad (4.6)$$

O conjunto de métricas apresentado permite avaliar a abordagem em vários aspectos, diferentemente do estado da arte, onde o padrão de avaliação observado foi apenas com a ACC . De acordo com a metodologia proposta (Figura 11), é obtido uma população com um total de 1000 amostras para cada métrica. Baseado nesses valores, aplicou-se o método estatístico one-way ANOVA (CARDINAL; AITKEN, 2013) com o *post-hoc Turkey test* ($\alpha = 0.05$) (ABDI; WILLIAMS, 2010) para cada métrica, com o objetivo de revelar os resultados que diferem significativamente um do outro.

5 Resultados e Discussões

Este capítulo apresenta os resultados e discussões da proposta de classificação multiclasse do CRC, detalhada no capítulo anterior e demonstrada na Figura 11. São apresentados os resultados do experimento desenvolvido nos cenários S1 e S2.

5.1 Cenário S1

O primeiro passo experimental desenvolvido nesse cenário foi buscar a configuração do SCM que alcançasse o melhor resultado de classificação multiclasse do CRC. Os resultados são apresentados na Tabela 3. Os resultados revelaram que utilizar filtro passa-baixa (média), como função de degradação, em conjunção com a saída VT obteve a melhor caracterização dos tecidos. Em particular, observa-se que a degradação de altas frequências de imagens de tecidos colorretais é um valioso descritor para o pré-diagnóstico em oito classes do CRC. Portanto, o filtro da média e VT é a melhor configuração encontrada para o SCM. Para esses parâmetros, foram encontrados ACC : 90.84%, $PREC$: 91.10%, REC : 90.70%, $SPEC$: 98.69%, e F_1 : 90.79%. O classificador SVM RBF alcançou os melhores resultados. Adotou-se usar a melhor configuração do SCM para o desenvolvimento do cenário S2.

O cenário S1 analisa o desempenho do SCM usado de forma isolada. Quando comparado o resultado isolado da classificação das características de cada método de extração, o classificador SVM RBF novamente oferece o melhor desempenho para todos os descritores pesquisados, independentemente da métrica de avaliação. Os resultados quantitativos da avaliação são sintetizados na Figura 14. Observou-se que o desempenho do *best5* é superior ao desempenho apresentado por cada um de seus componentes, incluindo Gabor. O experimento ainda revela que o SCM apresenta melhor resultado que o *best5*, melhorando o estado da arte. Para todos os extratores, $SPEC$ foi acima de 94%, mostrando a habilidade dos métodos de identificar casos verdadeiros negativos. O SCM contribuiu ainda mais, melhorando o desempenho em outras métricas (ACC , $PREC$, REC e F_1). Assim, aplicar um filtro passa-baixa como parâmetro do SCM mostra-se ser um valioso descritor para a caracterização de imagens de tecidos colorretais. O tempo médio de execução, durante o processo de extração, dos algoritmos pesquisados são apresentados na Tabela 4. O SCM teve o melhor desempenho, com um ganho de 66,66% e 98,83% no tempo de execução da extração de recursos quando comparado com os métodos HL (segundo algoritmo mais rápido) e Gabor (algoritmo menos rápido), respectivamente.

Tabela 3 – Resultado da classificação, em oito classes do CRC, usando diferentes configurações do SCM.

Filtro	Saída	Classificador	ACC(%)	PREC(%)	REC(%)	SPEC(%)	F_1 (%)
Média	VT	KNN	87.62±1.38	87.90±1.35	87.59±1.28	98.23±0.20	87.64±1.28
		SVM Linear	84.34±0.70	84.75±0.78	84.36±0.66	97.76±0.10	84.44±0.70
		SVM Poly	52.92±36.28	54.77±36.74	53.60±35.76	93.34±5.09	49.94±40.56
		SVM RBF	90.84±1.47	91.10±1.36	90.70±1.44	98.69±0.21	90.79±1.39
	MX	DT	81.36±1.39	81.49±1.40	81.38±1.32	97.34±0.20	81.37±1.34
		KNN	80.80±1.25	81.04±1.20	80.82±1.34	97.26±0.18	80.81±1.30
		SVM Linear	75.88±1.20	76.56±1.01	75.89±1.22	96.55±0.17	75.96±1.12
		SVM Poly	68.14±19.85	67.39±22.39	68.44±19.08	95.46±2.81	66.63±22.91
	MY	SVM RBF	82.38±1.49	82.48±1.54	82.41±1.48	97.48±0.21	82.35±1.51
		DT	77.52±1.78	77.66±1.90	77.56±1.87	96.79±0.25	77.54±1.89
	ATT	KNN	78.06±2.08	78.34±1.81	78.13±1.76	96.87±0.30	78.05±1.80
		SVM Linear	72.94±1.74	73.78±1.87	72.95±1.77	96.13±0.25	72.98±1.77
		SVM Poly	77.92±2.52	77.96±2.38	77.93±2.30	96.85±0.36	77.78±2.36
		SVM RBF	80.18±2.03	80.21±2.16	80.14±2.11	97.17±0.29	80.09±2.12
Gaussiano	VT	DT	74.56±1.94	74.54±1.84	74.63±1.77	96.37±0.28	74.49±1.82
		KNN	73.66±1.29	73.81±0.99	73.71±0.90	96.24±0.20	73.56±0.92
		SVM Linear	71.06±1.07	71.40±0.96	71.08±0.70	95.87±0.17	70.91±0.76
		SVM Poly	75.28±1.56	76.15±1.76	75.31±1.54	96.47±0.23	75.21±1.69
	MX	SVM RBF	77.92±0.77	78.10±1.06	77.92±0.86	96.85±0.11	77.87±0.98
		DT	66.16±2.71	66.47±2.29	66.17±2.22	95.16±0.40	66.27±2.24
	MY	KNN	87.16±0.86	87.58±0.57	87.18±0.69	98.17±0.13	87.25±0.67
		SVM Linear	84.12±1.45	84.65±1.19	84.11±1.39	97.73±0.21	84.24±1.31
		SVM Poly	65.60±30.93	65.55±33.12	65.97±29.99	95.14±4.29	63.98±34.35
		SVM RBF	89.98±0.69	90.08±0.69	90.00±0.70	98.57±0.10	90.00±0.69
	ATT	DT	81.20±1.62	81.30±1.25	81.22±1.44	97.31±0.24	81.18±1.36
		KNN	81.48±1.15	81.86±1.27	81.43±1.35	97.35±0.16	81.48±1.35
		SVM Linear	76.58±1.68	77.18±1.82	76.62±1.80	96.66±0.24	76.62±1.86
		SVM Poly	77.98±1.03	78.39±1.25	78.00±1.18	96.85±0.14	77.95±1.25
	MY	SVM RBF	82.84±1.90	82.95±2.07	82.80±1.97	97.55±0.27	82.82±2.02
		DT	76.86±1.34	76.9±1.62	76.86±1.53	96.69±0.19	76.79±1.53
	ATT	KNN	80.38±1.28	80.62±1.19	80.46±1.04	97.20±0.19	80.39±1.14
		SVM Linear	75.00±0.71	75.67±0.69	75.05±0.84	96.43±0.10	75.11±0.80
		SVM Poly	74.52±8.57	77.18±3.68	74.61±8.50	96.36±1.22	74.22±9.25
		SVM RBF	81.84±1.43	81.91±1.23	81.94±1.24	97.41±0.21	81.82±1.30
Laplace	VT	DT	77.08±0.93	77.19±0.91	77.11±0.89	96.73±0.13	77.07±0.90
		KNN	71.74±0.59	71.94±1.01	71.75±0.75	95.96±0.08	71.60±0.93
		SVM Linear	67.94±1.64	67.82±1.75	67.93±1.73	95.42±0.23	67.68±1.81
		SVM Poly	73.92±2.67	73.99±2.64	73.94±2.62	96.28±0.39	73.70±2.64
	MX	SVM RBF	74.00±1.21	73.96±1.06	74.01±1.00	96.29±0.18	73.79±1.08
		DT	65.74±1.35	65.98±0.77	65.80±1.19	95.11±0.20	65.79±0.99
	MY	KNN	86.06±1.63	86.53±1.89	86.11±1.77	98.01±0.24	86.13±1.68
		SVM Linear	82.82±0.83	83.34±0.83	82.83±0.86	97.55±0.12	82.93±0.85
		SVM Poly	75.28±4.43	78.08±2.36	75.37±4.08	96.47±0.63	75.53±4.02
		SVM RBF	88.54±1.45	88.75±1.57	88.60±1.49	98.36±0.21	88.55±1.46
	ATT	DT	79.78±1.29	79.84±1.48	79.86±1.36	97.11±0.18	79.73±1.44
		KNN	80.78±1.16	81.02±1.20	80.78±1.24	97.26±0.17	80.80±1.21
		SVM Linear	76.40±1.54	76.97±1.12	76.47±1.12	96.63±0.23	76.47±1.14
		SVM Poly	75.12±8.37	77.75±3.64	75.45±7.73	96.45±1.18	75.08±8.35
	MY	SVM RBF	80.96±3.12	81.18±2.92	80.98±3.13	97.28±0.44	80.89±3.29
		DT	77.66±1.38	77.61±1.39	77.65±1.32	96.81±0.20	77.57±1.36
	ATT	KNN	55.82±2.39	56.31±1.67	55.92±1.67	93.69±0.35	55.58±1.69
		SVM Linear	51.52±1.50	52.80±1.63	51.63±0.92	93.08±0.22	50.23±1.18
		SVM Poly	57.50±3.85	60.77±4.48	57.53±3.90	93.93±0.53	56.37±4.43
		SVM RBF	59.60±1.90	61.80±1.29	59.60±1.39	94.23±0.28	58.89±1.56
Sobel	VT	DT	48.74±1.93	48.71±2.15	48.73±2.10	92.68±0.27	48.64±2.18
		KNN	78.20±1.29	78.63±1.49	78.17±1.44	96.88±0.18	78.24±1.48
		SVM Linear	75.60±0.99	75.68±1.11	75.65±1.11	96.51±0.14	75.52±1.11
		SVM Poly	79.56±1.84	79.76±2.26	79.58±2.14	97.08±0.26	79.52±2.21
	MX	SVM RBF	81.62±1.64	81.98±1.60	81.60±1.64	97.37±0.24	81.66±1.62
		DT	73.92±0.82	73.96±1.00	73.87±0.95	96.27±0.12	73.86±0.96
	MY	KNN	85.50±1.56	85.79±1.56	85.37±1.53	97.93±0.22	85.47±1.59
		SVM Linear	83.06±1.58	83.54±1.48	82.98±1.46	97.58±0.23	83.10±1.51
		SVM Poly	78.96±9.90	80.21±7.97	78.94±9.69	97.00±1.40	78.75±10.12
		SVM RBF	88.66±1.08	88.79±1.25	88.57±1.11	98.38±0.15	88.62±1.17
	ATT	DT	78.96±1.53	79.00±1.33	78.89±1.40	96.99±0.22	78.87±1.39
		KNN	80.90±0.66	81.12±0.81	80.89±0.68	97.27±0.09	80.91±0.73
		SVM Linear	76.34±0.63	76.83±0.89	76.35±0.74	96.62±0.09	76.37±0.77
		SVM Poly	78.46±4.87	78.87±4.10	78.59±4.80	96.93±0.69	78.35±5.04
	MY	SVM RBF	82.04±1.92	82.10±1.93	82.04±1.79	97.44±0.28	81.99±1.87
		DT	76.44±1.23	76.48±1.17	76.45±1.14	96.64±0.18	76.38±1.14
	ATT	KNN	52.96±2.27	52.78±2.26	53.01±2.15	93.29±0.32	52.18±2.17
		SVM Linear	53.24±2.15	54.36±1.84	53.36±1.76	93.32±0.29	51.69±1.86
		SVM Poly	56.48±3.21	59.31±2.35	56.55±3.18	93.79±0.47	55.71±3.56
		SVM RBF	57.40±3.24	59.39±2.55	57.49±3.10	93.92±0.47	57.23±3.23
	ATT	DT	48.12±2.62	48.15±2.20	48.14±2.22	92.59±0.38	48.06±2.21
		KNN	76.24±1.39	76.53±1.31	76.3±1.64	96.61±0.20	76.27±1.44
		SVM Linear	73.96±1.38	73.85±1.30	73.96±1.23	96.28±0.20	73.67±1.30
		SVM Poly	76.92±0.66	76.82±0.73	76.97±0.89	96.71±0.10	76.70±0.66
	ATT	SVM RBF	77.78±2.28	77.73±2.25	77.89±2.13	96.83±0.33	77.63±2.22
		DT	71.2±1.49	71.45±1.05	71.17±1.32	95.89±0.22	71.21±1.19

Figura 14 – Avaliação quantitativa e individual para cada descritor pesquisado. Todos os valores apresentados estão em (%). Os resultados estão ordenados da menor para a maior *ACC*. Resultados obtidos usando o classificador SVM RBF.

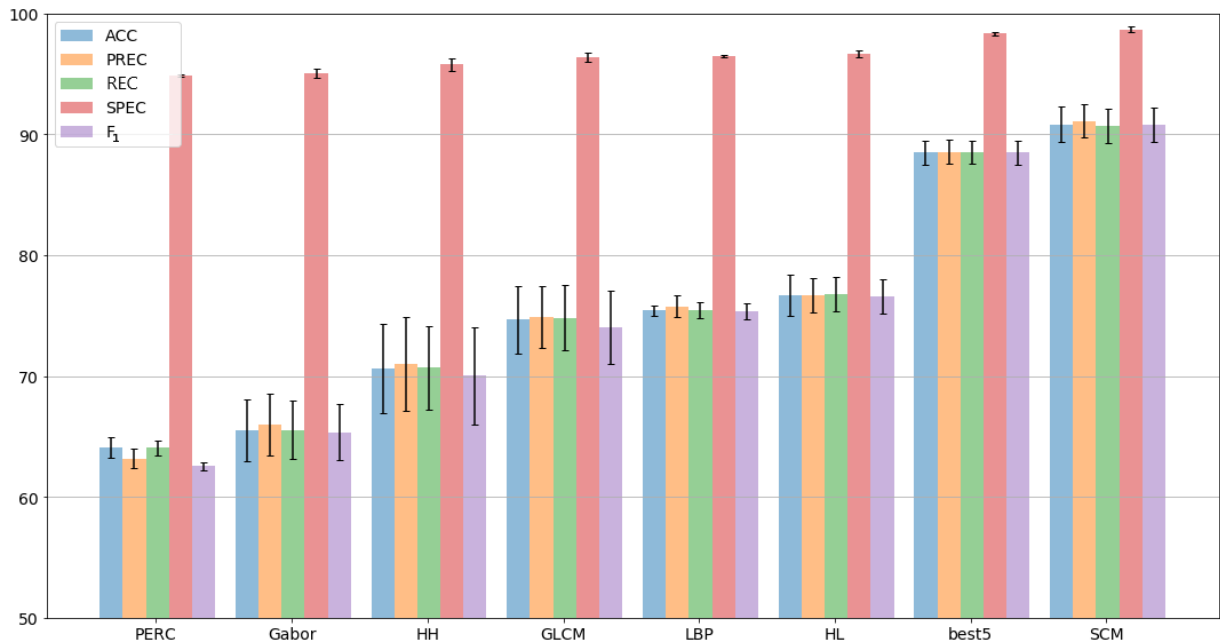


Tabela 4 – Tempo médio de execução para cada método de extração de características. Os valores apresentados estão ordenados do menor para o maior tempo médio.

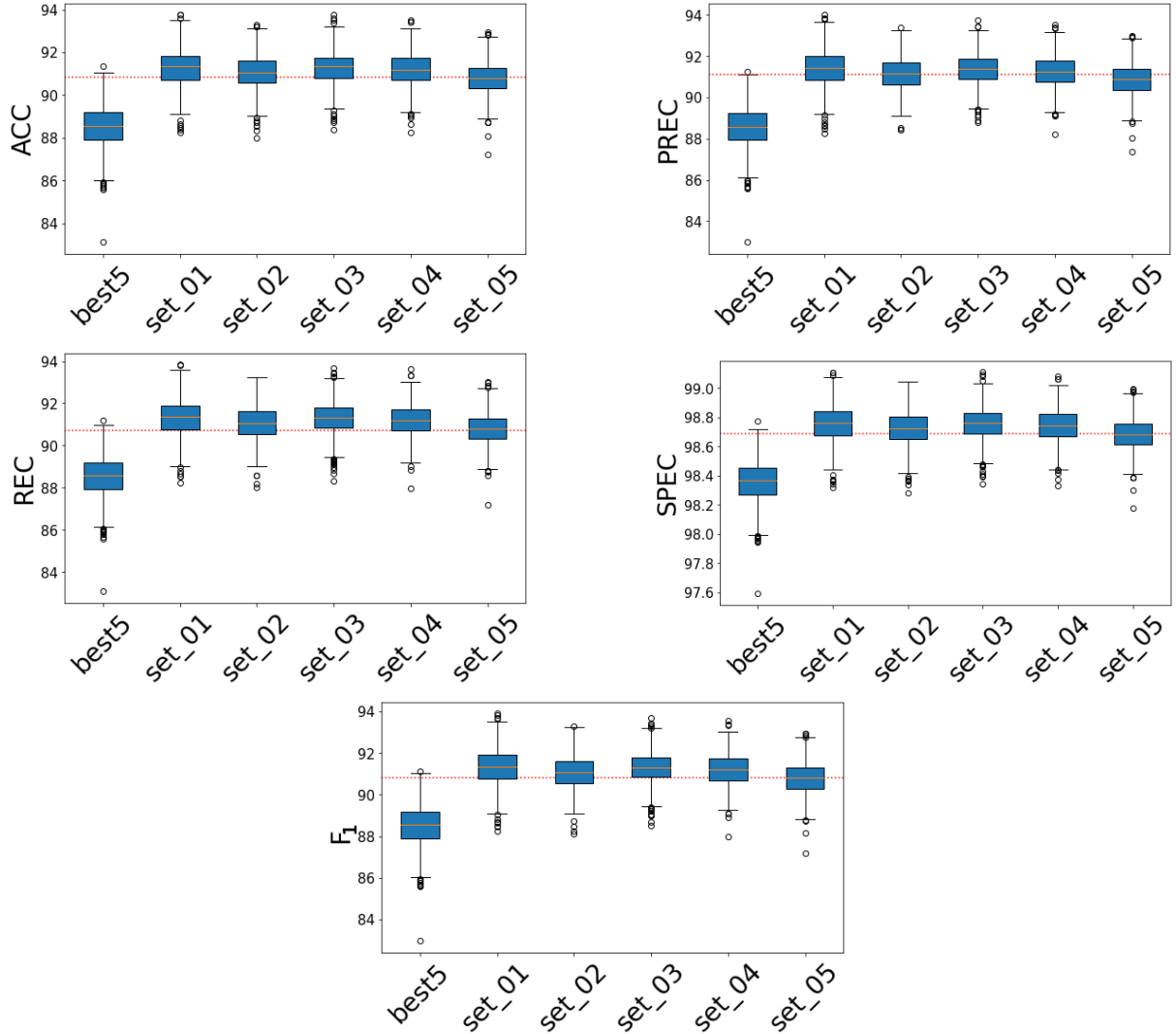
Extrator	SCM	HL	HH	GLCM	Perceptual	LBP	Gabor
tempo(s)	0.01	0.03	0.05	0.10	0.12	0.72	0.86

5.2 Cenário S2

A análise S2 combina características extraídas pelo SCM com outros descritores. O objetivo desse cenário é analisar o desempenho do SCM em conjunto com *best5*, como descrito na Tabela 2. Para isso, foi realizado o seguinte procedimento: cada método que compõe *best5* foi gradualmente substituído pelo SCM, formando 5 conjuntos de características diferentes. A condição de formação dos conjuntos é detalhada na Figura 13.

Os resultados da análise S2 são mostrados na Figura 15. Similar ao S1, o classificador SVM RBF forneceu os melhores resultados. Embora o desempenho dos descritores combinados sejam semelhantes, todos os conjuntos combinados com o SCM superam *best5* para todas as métricas de avaliação. A Figura 16 apresenta o resultado do cálculo da correlação entre as características que integram os conjuntos combinados. Pode-se perceber que existe uma alta correlação apenas entre as características do método HH e parte das características do GLCM, indicando que os demais métodos medem diferentes aspectos de textura das imagens analisadas e mostra que a combinação de características de diferentes métodos pode ser útil para o problema abordado nesse trabalho. Trabalhos futuros podem

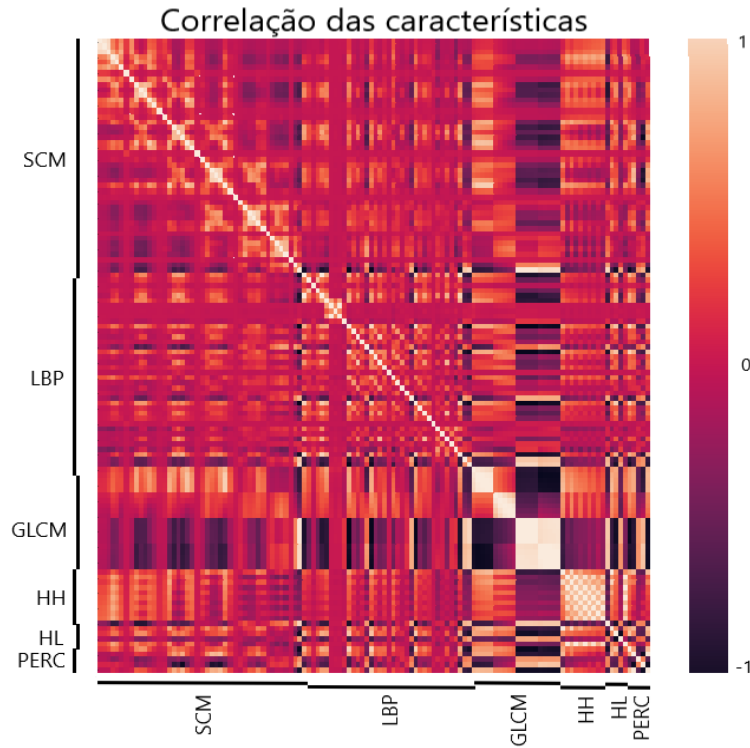
Figura 15 – Comparação quantitativa dos melhores resultados alcançados com os conjuntos de características combinados. A linha horizontal (pontilhada e na cor vermelha) representa a melhor média de cada métrica apresentada na Tabela 3, alcançadas com o uso da SCM de forma isolada.



abordar essa descoberta em relação à interpretabilidade do SCM.

Os resultados quantitativos em termos das médias das métricas são expostos na Tabela 5. Os maiores resultados alcançados foram com o *set_01*, *ACC*: 91.30%, *PREC*: 91.41%, *REC*: 91.31%, *SPEC*: 98.76% e *F1*: 91.31%, porém não pode-se afirmar que é o melhor resultado, pois outros conjuntos forneceram valores bem próximos. Nesta perspectiva, foi realizada uma busca dos conjuntos de características com resultados estatisticamente iguais ao *set_01*. Para isso, o teste ANOVA ($\alpha = 0.05$) foi usado para verificar se existe diferença significativa entre as médias de classificação dos conjuntos. Verificada a existência de diferença entre as médias, o teste *post-hoc* Turkey ($\alpha = 0.05$) foi usado para identificar entre quais pares de conjuntos a diferença é significativa. O resultado do teste de Turkey é apresentado na Figura 17. As interseções em cinza representam

Figura 16 – Resultado da correlação de Pearson (BENESTY; CHEN; HUANG, 2008) entre as características de diferentes métodos. Observa-se alta correlação apenas entre as características dos métodos GLCM e HH.



pares com diferença estatística. Por meio da figura percebeu-se que os conjuntos *set_03* e *set_04* ou apenas o *set_03* são estatisticamente iguais ao *set_01* para todas as métricas de avaliação. O conjunto *set_01* também teve a vantagem de ser a combinação baseada em SCM mais rápida. Como o tempo de execução do SCM é menor do que das outras técnicas pesquisadas, argumenta-se que o SCM melhorou o desempenho do agrupamento de descritores sem impactar o custo computacional.

Uma matriz de confusão resultante da classificação multiclasse CRC é apresentada na Figura 18. A análise foi baseada no *set_01*, pois ele é o melhor resultado encontrado nesse trabalho. A discriminação entre estroma simples e estroma complexo é difícil devido a similaridade dos tecidos. Quando a Figura 18 (a) e Figura 18 (b) são comparadas, observa-se que *set_01* contribuiu para diminuição do erro em todas as classes de tecido, principalmente nas classes mais difíceis de separação (estroma simples e estroma complexo).

Por fim, foi analisado o desempenho de classificação multiclasse do CRC, comparando a abordagem proposta nesse trabalho e os trabalhos relacionados que analisaram a mesma base de imagens. O resultado de cada abordagem é apresentado na Tabela 6. O melhor resultado encontrado para a abordagem proposta foi *ACC*: 91.30%, *PREC*: 91.41%, *REC*: 91.31%, *SPEC*: 98.76% e F_1 : 91.31% usando 5000 imagens. O resultado supera a abordagem de Kather *et al.* (2016) em 4,46% e Rachapudi e Devi (2020) em 18,10% considerando a métrica *ACC*. Por outro lado, a *ACC* alcançada por Wang *et al.*

Tabela 5 – Resultado da classificação, em oito classes do CRC, usando conjunto de características agrupadas. Os valores apresentados são médias resultantes de 1000 repetições do experimento.

Característica	Classificador	ACC(%)	PREC(%)	REC(%)	SPEC(%)	F_1 (%)	Tempo de extração(s)
best5	KNN	82.2±0.84	82.57±1.08	82.29±1.06	97.45±0.11	82.13±1.12	1.04
	SVM Linear	87.46±1.04	87.49±1.01	87.47±1.05	98.20±0.14	87.44±1.06	
	SVM Poly	88.50±1.49	88.52±1.47	88.51±1.58	98.35±0.21	88.46±1.56	
	SVM RBF	88.51±0.99	88.86±0.97	88.53±0.96	98.36±0.14	88.50±0.97	
	DT	78.02±1.14	78.04±1.38	77.97±1.49	96.85±0.15	77.94±1.43	
set_01	KNN	87.78±1.20	87.92±1.28	87.82±1.28	98.25±0.16	87.78±1.28	0.33
	SVM Linear	89.72±0.81	89.84±0.70	89.77±0.79	98.53±0.11	89.75±0.77	
	SVM Poly	86.00±6.51	86.68±5.79	86.04±6.49	98.00±0.92	86.13±6.39	
	SVM RBF	91.30±0.86	91.41±0.85	91.31±0.86	98.76±0.12	91.31±0.86	
	DT	81.36±1.10	81.38±1.12	81.47±1.06	97.33±0.15	81.33±1.08	
set_02	KNN	87.20±0.85	87.65±0.98	87.26±0.91	98.17±0.11	87.24±1.09	0.93
	SVM Linear	89.52±0.58	89.72±0.82	89.54±0.77	98.50±0.07	89.58±0.81	
	SVM Poly	88.85±1.06	89.30±1.02	88.85±1.28	98.40±0.14	88.93±1.24	
	SVM RBF	91.06±0.83	91.16±0.80	91.07±0.81	98.72±0.12	91.07±0.81	
	DT	82.20±0.83	82.14±1.24	82.19±1.23	97.45±0.11	82.07±1.22	
set_03	KNN	87.58±1.47	87.98±1.45	87.61±1.53	98.22±0.21	87.60±1.50	1.02
	SVM Linear	89.32±0.66	89.44±0.67	89.32±0.69	98.47±0.09	89.31±0.67	
	SVM Poly	89.62±2.06	89.89±1.79	89.61±2.10	98.51±0.29	89.64±2.00	
	SVM RBF	91.29±0.78	91.37±0.76	91.3±0.77	98.76±0.11	91.30±0.77	
	DT	81.52±1.50	81.50±1.49	81.66±1.48	97.36±0.21	81.46±1.46	
set_04	KNN	87.56±1.07	87.97±1.23	87.56±1.22	98.22±0.14	87.58±1.20	1.00
	SVM Linear	89.74±0.79	89.92±0.88	89.73±0.81	98.53±0.11	89.75±0.84	
	SVM Poly	88.66±2.69	89.02±2.33	88.58±2.84	98.38±0.38	88.66±2.69	
	SVM RBF	91.20±0.77	91.28±0.76	91.21±0.76	98.74±0.11	91.21±0.76	
	DT	81.82±0.83	81.92±0.88	81.83±0.65	97.40±0.12	81.80±0.76	
set_05	KNN	86.96±0.30	87.37±0.49	86.97±0.34	98.13±0.04	87.02±0.36	0.95
	SVM Linear	88.96±0.93	89.15±0.93	88.99±0.86	98.42±0.13	89.01±0.88	
	SVM Poly	82.90±15.21	86.00±8.96	83.04±14.99	97.55±2.17	82.62±16.00	
	SVM RBF	90.75±0.77	90.86±0.76	90.77±0.76	98.68±0.11	90.77±0.76	
	DT	82.16±1.58	82.21±1.67	82.15±1.50	97.45±0.22	82.14±1.56	

(2017) foi 1,4% maior, contudo eles usaram apenas 1000 imagens.

Tabela 6 – Comparação da abordagem proposta com trabalhos que utilizaram a mesma base de imagens para classificação multiclasse do CRC. Foram apresentados apenas os desvios padrões fornecidos pelos autores.

Abordagem	Quantidade de imagens	ACC(%)	PREC(%)	REC(%)	SPEC(%)	F_1 (%)
Abordagem proposta	5000	91.30 ± 0.86	91.41±0.85	91.31±0.86	98.76±0.12	91.31±0.86
Kather et al. (2016)	5000	87.40	-	-	-	-
Wang et al. (2017)	1000	92.60±1.20	-	-	-	-
Rachapudi e Devi (2020)	5000	77.30	-	-	-	-

Figura 17 – Comparações entre os conjuntos combinados usando o teste post-hoc Turkey. As interseções em cinza representam pares com diferença significativa em $\alpha = 0,05$.

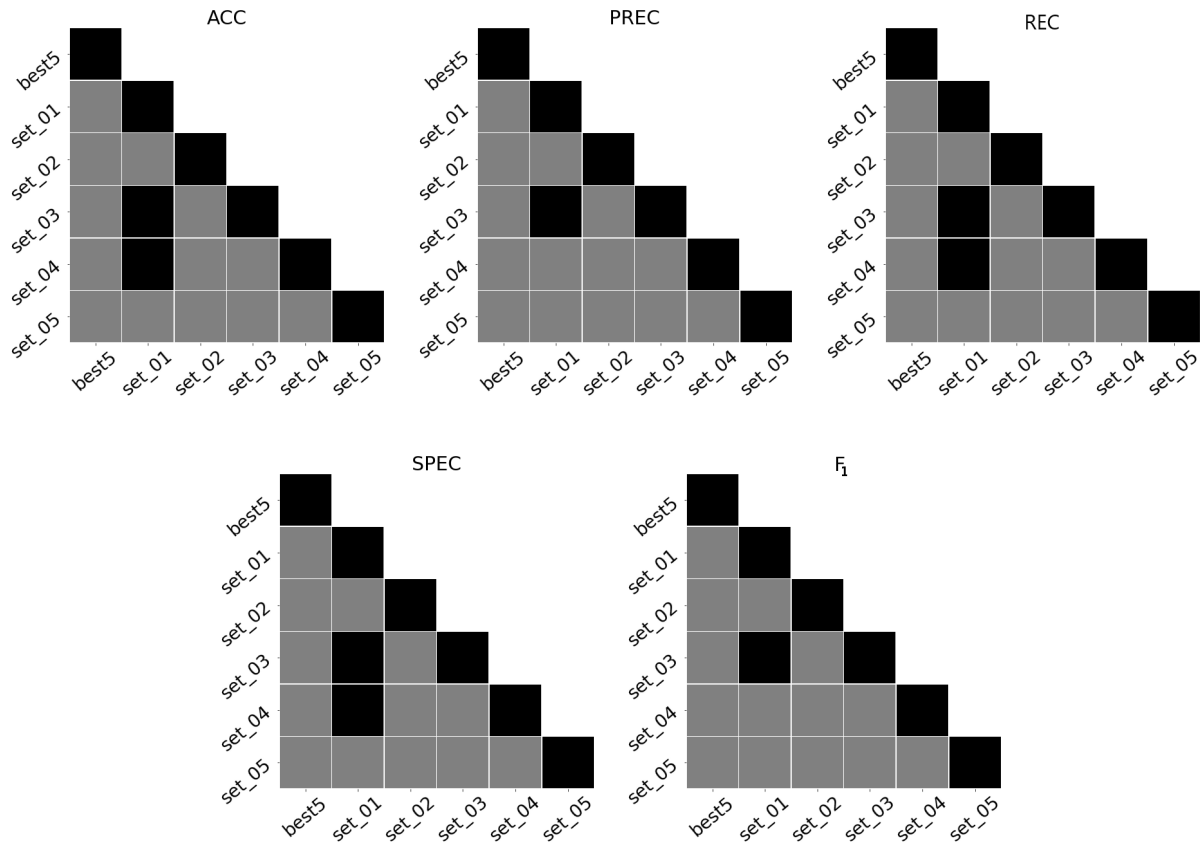
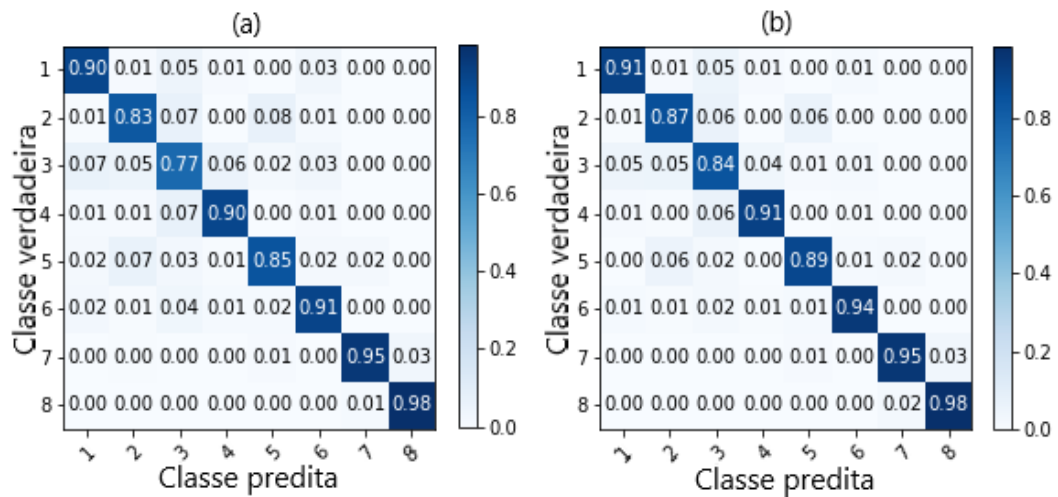


Figura 18 – Matriz de confusão da classificação multiclasse do CRC. a) Resultados obtidos do *best5* e b) *set_01*. As classes verdadeiras e preditas são descritas como: 1 - Tumor epitelial, 2 - Estroma simples, 3 - Estroma complexo, 4 - Células imunes, 5 - Detritos, 6 - Glândulas mucosas, 7 - Tecido adiposo e 8 - Fundo.



6 Conclusões e Trabalhos Futuros

Nesse trabalho, descritores de textura foram estudados para categorizar amostras de tecido colorretal em oito classes diferentes, incluindo tecido normal e anormal. Foram explorados características de sete descritores (SCM, HH, HL, PERC, Gabor, GLCM, LBP) em dois cenários: (S1) características analisadas de forma isoladas; e (S2) características analisadas de forma combinadas. Os experimentos mostraram que a degradação de altas frequências nas imagens do câncer colorretal e a descrição das diferenças estruturais resultantes com o SCM é relevante para o pré diagnóstico de oito classes. O SCM permite um ganho de 2,63% em termos de acurácia quando usado isolado, e 3,15 %, quando combinado com características de outros descritores de textura e comparado com o conjunto de características *best5*. Como houve um ganho no desempenho do processo de classificação do SCM isolado para o combinado, foi calculado a correlação entre as características dos diferentes métodos combinados e constatou-se baixa correlação entre eles, indicando que as características medem diferentes aspectos de textura e mostra que a combinação de métodos pode ser útil. Além disso, o melhor conjunto baseado em SCM consegue diminuir em até 68,26% o tempo médio da fase de extração de características. O experimento ainda mostrou uma diminuição do erro em todas as classes de tecido, principalmente nas classes mais difíceis de separação (estroma simples e estroma complexo). A literatura estudada padronizou a avaliação de suas abordagens apenas com a acurácia, o que não permite uma análise mais abrangente dos resultados de classificação. Portanto, por meio da avaliação em 5 métricas, percebeu-se que todos os extratores alcançaram SPEC acima de 94,00%, mostrando a habilidade dos métodos em identificar casos verdadeiros negativos. No entanto, a inclusão do SCM como extrator de características contribuiu melhorando o desempenho das outras métricas (ACC, PREC, REC e F_1), em mais de 19,00% (quando comparado de forma isolado com o HL - extrator com segundo melhor desempenho) e 3,10% (quando comparado com os extratores de forma agrupada - *best5*). Para todos os conjuntos de características avaliados, sejam na forma isolada ou agrupada, o classificador SVM RBF alcançou os melhores resultados.

Pesquisas podem se beneficiar da proposta de um descritor que atinja revocação de 91,31%, especificidade de 98,76%, com tempo de execução satisfatório à rotina clínica, para automatizar a tarefa de classificação dos tecidos através de um sistema CAD. Ainda não existe uma definição morfológica clara dos diferentes subtipos de estroma, portanto a análise de textura multiclasse pode ser usada para identificar subtipos dessa classe de tecido e investigar as suas implicações biológicas. A abordagem também pode ser aplicada a classificação de tecidos histológicos com estruturas similares ao colorretal, por exemplo: mama, ovário e próstata.

Referências

- ABDI, H.; WILLIAMS, L. J. Newman-keuls test and tukey test. *Encyclopedia of Research Design*. Thousand Oaks, CA: Sage, p. 1–11, 2010. Citado na página 50.
- AKBAR, B.; GOPI, V. P.; BABU, V. S. Colon cancer detection based on structural and statistical pattern recognition. *2nd International Conference on Electronics and Communication Systems (ICECS)*, p. 1735–1739, 2015. Citado 2 vezes nas páginas 19 e 23.
- ALTUNBAY, D. et al. Color graphs for automated cancer diagnosis and grading. *IEEE Transactions on Biomedical Engineering*, IEEE, v. 57, n. 3, p. 665–674, 2009. Citado na página 22.
- ANDREOLA, R. *Support vector machines na classificação de imagens hiperespectrais*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 08 2009. Citado na página 41.
- BEJNORDI, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, American Medical Association, v. 318, n. 22, p. 2199–2210, 2017. Citado na página 24.
- BENESTY, J.; CHEN, J.; HUANG, Y. On the importance of the pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 16, n. 4, p. 757–765, 2008. Citado 2 vezes nas páginas 9 e 55.
- BIANCONI, F.; ÁLVAREZ-LARRÁN, A.; FERNÁNDEZ, A. Discrimination between tumour epithelium and stroma via perception-based features. *Neurocomputing*, Elsevier, v. 154, p. 119–126, 2015. Citado 2 vezes nas páginas 19 e 37.
- 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. Citado 2 vezes nas páginas 17 e 28.
- BRENER, S. et al. Carcinoma de células escamosas bucal: uma revisão de literatura entre o perfil do paciente, estadiamento clínico e tratamento proposto. *Rev bras cancerol*, v. 53, n. 1, p. 63–9, 2007. Citado na página 28.
- BUCIU, I.; GACSADI, A. Directional features for automatic tumor classification of mammogram images. *Biomedical Signal Processing and Control*, Elsevier, v. 6, n. 4, p. 370–378, 2011. Citado na página 19.
- BURGES, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, Springer, v. 2, n. 2, p. 121–167, 1998. Citado na página 49.
- CARDINAL, R. N.; AITKEN, M. R. *ANOVA for the behavioral sciences researcher*. 1. ed. New York: Psychology Press, 2013. Citado na página 50.
- CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 49.

- CRAMMER, K.; SINGER, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, v. 2, n. Dec, p. 265–292, 2001. Citado na página 41.
- DEBELEE, T. G. et al. Classification of mammograms using convolutional neural network based feature extraction. *International Conference on Information and Communication Technology for Development for Africa*, Springer, v. 244, p. 89–98, 2017. Citado 2 vezes nas páginas 17 e 18.
- DELAUNAY, B. et al. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, v. 7, n. 793-800, p. 1–2, 1934. Citado na página 22.
- DOYLE, S. et al. Automated grading of prostate cancer using architectural and textural image features. *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, IEEE, p. 1284–1287, 2007. Citado na página 22.
- DUNN, D.; HIGGINS, W. E. Optimal gabor filters for texture segmentation. *IEEE Transactions on image processing*, IEEE, v. 4, n. 7, p. 947–964, 1995. Citado na página 36.
- ESGIAR, A. N. et al. Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa. *IEEE transactions on information technology in biomedicine*, IEEE, v. 2, n. 3, p. 197–203, 1998. Citado na página 22.
- FILHO, P. P. R. et al. Automatic histologically-closer classification of skin lesions. *Computerized Medical Imaging and Graphics*, Elsevier, v. 68, p. 40–54, 2018. Citado 3 vezes nas páginas 19, 43 e 47.
- FILHO, P. P. R. et al. Analysis of human tissue densities: A new approach to extract features from medical images. *Pattern Recognition Letters*, Elsevier, v. 94, p. 211–218, 2017. Citado 4 vezes nas páginas 18, 19, 43 e 47.
- FISHER, R. A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, The Royal Society, v. 222, n. 594-604, p. 309–368, 1922. Citado 2 vezes nas páginas 40 e 48.
- GURCAN, M. N. et al. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, IEEE, v. 2, p. 147–171, 2009. Citado 2 vezes nas páginas 17 e 29.
- HAGHIGHAT, M.; ZONOUZ, S.; ABDEL-MOTTALEB, M. Identification using encrypted biometrics. *International Conference on Computer Analysis of Images and Patterns*, Springer, p. 440–448, 2013. Citado na página 37.
- HAN, F. et al. Texture feature analysis for computer-aided diagnosis on pulmonary nodules. *Journal of digital imaging*, Springer, v. 28, n. 1, p. 99–115, 2015. Citado na página 19.
- HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, IEEE, n. 6, p. 610–621, 1973. Citado na página 33.

- HAYKIN, S. *Redes neurais: princípios e prática*. 2. ed. Porto Alegre: Bookman Editora, 2007. Citado na página 41.
- IFTIKHAR, M. A.; HASSAN, M.; ALQUHAYZ, H. A colon cancer grade prediction model using texture and statistical features, smote and mrmr. *19th International Multi-Topic Conference (INMIC)*, IEEE, p. 1–7, 2016. Citado na página 23.
- JAMES, G. et al. *An introduction to statistical learning with Applications in R*. New York: Springer, 2013. v. 112. Citado na página 43.
- JØRGENSEN, A. S. et al. Using cell nuclei features to detect colon cancer tissue in hematoxylin and eosin stained slides. *Cytometry Part A*, Wiley Online Library, v. 91, n. 8, p. 785–793, 2017. Citado na página 24.
- KALKAN, H. et al. Automated classification of local patches in colon histopathology. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, IEEE, p. 61–64, 2012. Citado 2 vezes nas páginas 19 e 22.
- KATHER, J. N. et al. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, Nature Publishing Group, v. 6, p. 27988, 2016. Citado 8 vezes nas páginas 10, 18, 20, 24, 26, 44, 45 e 49.
- KHAN, S. et al. Optimized gabor features for mass classification in mammography. *Applied Soft Computing*, Elsevier, v. 44, p. 267–280, 2016. Citado na página 19.
- KOOHBANANI, N. A. et al. Significance of hyperparameter optimization for metastasis detection in breast histology images. *Computational Pathology and Ophthalmic Medical Image Analysis*, Springer, p. 139–147, 2018. Citado na página 24.
- MALIK, F.; BAHARUDIN, B. The statistical quantized histogram texture features analysis for image retrieval based on median and laplacian filters in the dct domain. *The International Arab Journal of Information Technology*, v. 10, n. 6, p. 1–9, 2013. Citado na página 34.
- MITTAL, H.; SARASWAT, M. An automatic nuclei segmentation method using intelligent gravitational search algorithm based superpixel clustering. *Swarm and Evolutionary Computation*, Elsevier, v. 45, p. 15–32, 2019. Citado na página 17.
- NAIYAR, M.; ASIM, Y.; SHAHID, A. Automated colon cancer detection using structural and morphological features. *2015 13th International Conference on Frontiers of Information Technology (FIT)*, IEEE, p. 240–245, 2015. Citado na página 23.
- NEILAN, B. A. Adenocarcinoma of unknown origin. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 33, n. 4, p. 237–241, 1983. Citado na página 28.
- OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, Elsevier, v. 29, n. 1, p. 51–59, 1996. Citado na página 35.
- PANDA, R. N. et al. Efficient cad system based on glcm & derived feature for diagnosing breast cancer. *International Journal of Computer Science and Information Technologies*, Citeseer, v. 6, p. 3323–3327, 2015. Citado na página 39.

- PEIXOTO, S. A.; FILHO, P. P. R. Neurologist-level classification of stroke using a structural co-occurrence matrix based on the frequency domain. *Computers & Electrical Engineering*, Elsevier, v. 71, p. 398–407, 2018. Citado 3 vezes nas páginas 19, 43 e 47.
- PRATT, W. K. *Digital image processing. A wiley-interscience publication*. New York: Wiley, 1978. Citado na página 45.
- PS, S. K.; DHARUN, V. Extraction of texture features using glcm and shape features using connected regions. *International Journal of Engineering and Technology*, Engg Journals Publications, v. 8, n. 6, p. 2926–2930, 2016. Citado na página 18.
- RACHAPUDI, V.; DEVI, G. L. Improved convolutional neural network based histopathological image classification. *Evolutionary Intelligence*, Springer, p. 1–7, 2020. Citado na página 25.
- RAMALHO, G. L. B. et al. Rotation-invariant feature extraction using a structural co-occurrence matrix. *Measurement*, Elsevier, v. 94, p. 406–415, 2016. Citado 7 vezes nas páginas 19, 29, 30, 31, 32, 44 e 47.
- RATHORE, S. et al. A recent survey on colon cancer detection techniques. *IEEE/ACM Transactions on computational biology and bioinformatics*, IEEE, v. 10, n. 3, p. 545–563, 2013. Citado na página 24.
- RIBEIRO, M. G. et al. Classification of colorectal cancer based on the association of multidimensional and multiresolution features. *Expert Systems With Applications*, Elsevier, v. 120, p. 262–278, 2019. Citado na página 24.
- RIORDON, J.; MCCALLUM, C.; SINTON, D. Deep learning for the classification of human sperm. *Computers in Biology and Medicine*, Elsevier, p. 103342, 2019. Citado 2 vezes nas páginas 17 e 18.
- ROKACH, L.; MAIMON, O. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 35, n. 4, p. 476–487, 2005. Citado 2 vezes nas páginas 42 e 48.
- ROUHI, R.; JAFARI, M. Classification of benign and malignant breast tumors based on hybrid level set segmentation. *Expert Systems with Applications*, Elsevier, v. 46, p. 45–59, 2016. Citado na página 18.
- RUIFROK, A. C.; JOHNSTON, D. A. et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, St. Louis, MO: Science Printers and Publishers, 1985-, v. 23, n. 4, p. 291–299, 2001. Citado na página 24.
- SALAMA, M. S.; ELTRASS, A. S.; ELKAMCHOUCI, H. M. An improved approach for computer-aided diagnosis of breast cancer in digital mammography. *IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, IEEE, p. 1–5, 2018. Citado na página 39.
- SAN-SEGUNDO, R. et al. Classification of epileptic eeg recordings using signal transforms and convolutional neural networks. *Computers in biology and medicine*, Elsevier, v. 109, p. 148–158, 2019. Citado 2 vezes nas páginas 17 e 18.

- SANTOS, L. F. S. dos et al. Multidimensional and fuzzy sample entropy (sampenmf) for quantifying h&e histological images of colorectal cancer. *Computers in biology and medicine*, Elsevier, v. 103, p. 148–160, 2018. Citado 2 vezes nas páginas 17 e 28.
- SANTOS, M. K. et al. Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. *Radiologia Brasileira*, SciELO Brasil, v. 52, n. 6, p. 387–396, 2019. Citado na página 17.
- SERTEL, O. et al. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern recognition*, Elsevier, v. 42, n. 6, p. 1093–1103, 2009. Citado na página 22.
- SINGH, V. P.; SRIVASTAVA, S.; SRIVASTAVA, R. Effective mammogram classification based on center symmetric-lbp features in wavelet domain using random forests. *Technology and Health Care*, IOS Press, v. 25, n. 4, p. 709–727, 2017. Citado na página 39.
- SIRINUKUNWATTANA, K. et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, Elsevier, v. 35, p. 489–502, 2017. Citado 3 vezes nas páginas 10, 24 e 26.
- SOCIETY, A. C. *Breast Cancer Facts & Figures 2013–2014*. Atlanta: American Cancer Society Inc., 2013. Citado na página 17.
- SOCIETY, A. C. *Colorectal Cancer Facts & Figures 2017–2019*. Atlanta: American Cancer Society Inc., 2017. Citado na página 17.
- SONG, T.-H. et al. Simultaneous cell detection and classification in bone marrow histology images. *IEEE journal of biomedical and health informatics*, IEEE, v. 23, n. 4, p. 1469–1476, 2018. Citado na página 24.
- STREET, W. N.; WOLBERG, W. H.; MANGASARIAN, O. L. Nuclear feature extraction for breast tumor diagnosis. *Biomedical image processing and biomedical visualization*, Proceedings of SPIE, v. 1905, p. 861–870, 1993. Citado na página 22.
- SV, M. K.; GUNASUNDARI, R. Computer-aided diagnosis of anterior segment eye abnormalities using visible wavelength image analysis based machine learning. *Journal of medical systems*, Springer, v. 42, n. 7, p. 128, 2018. Citado na página 19.
- TAMAI, N. et al. Effectiveness of computer-aided diagnosis of colorectal lesions using novel software for magnifying narrow-band imaging: a pilot study. *Endoscopy international open*, © Georg Thieme Verlag KG, v. 5, n. 08, p. E690–E694, 2017. Citado 2 vezes nas páginas 17 e 18.
- TAMURA, H.; MORI, S.; YAMAWAKI, T. Textural features corresponding to visual perception. *IEEE Transactions on Systems, man, and cybernetics*, IEEE, v. 8, n. 6, p. 460–473, 1978. Citado 3 vezes nas páginas 19, 37 e 38.
- THEODORIDIS, S.; KOUTROUMBAS, K. Pattern recognition and neural networks. *Advanced Course on Artificial Intelligence*, Springer, p. 169–195, 1999. Citado na página 41.

- TSCHANDL, P.; SINZ, C.; KITTLER, H. Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation. *Computers in biology and medicine*, Elsevier, v. 104, p. 111–116, 2019. Citado 2 vezes nas páginas 17 e 18.
- TSOCHATZIDIS, L.; COSTARIDOU, L.; PRATIKAKIS, I. Deep learning for breast cancer diagnosis from mammograms—a comparative study. *Journal of Imaging*, Multidisciplinary Digital Publishing Institute, v. 5, n. 3, p. 37, 2019. Citado 2 vezes nas páginas 17 e 18.
- VAPNIK, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, Citeseer, v. 10, n. 5, p. 988–999, 1999. Citado na página 41.
- WANG, C. et al. Histopathological image classification with bilinear convolutional neural networks. *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, p. 4050–4053, 2017. Citado 2 vezes nas páginas 19 e 24.
- WEYN, B. et al. Value of morphometry, texture analysis, densitometry, and histometry in the differential diagnosis and prognosis of malignant mesothelioma. *The Journal of pathology*, Wiley Online Library, v. 189, n. 4, p. 581–589, 1999. Citado na página 22.
- WILTGEN, M.; GERGER, A.; SMOLLE, J. Tissue counter analysis of benign common nevi and malignant melanoma. *International journal of medical informatics*, Elsevier, v. 69, n. 1, p. 17–28, 2003. Citado na página 22.
- YAMANE, L. d. S. *Caracterização mutacional dos oncogenes KRAS e BRAF e análise de MSI nas lesões precursoras e do câncer colorretal*. Dissertação (Mestrado) — Fundação Pio XII – Hospital de Câncer de Barretos, Barretos, 2013. Citado na página 29.

APÊNDICE A – Detalhes da construção dos conjuntos de características obtidos a partir do SCM

Uma visão geral da maneira com que foram construídos os conjuntos de características obtidos a partir da aplicação do método SCM é exibida na Figura 19. Usou-se o modelo genérico do SCM (apresentado na Figura 4) para demonstrar um contexto específico de extração de características de uma imagem. No exemplo da Figura 19 (a) foram usados como configurações de entrada ao método, uma função de quantização Q , que muda o número de níveis de tons de cinza da imagem de entrada de um valor L (inicial) para $N = 8$; um filtro passa-baixa (filtro da média 5×5) como função de degradação; e uma imagem de entrada, da classe de tecido colorretal estroma complexo. A mesma figura, apresenta a matriz SCM gerada para essa configuração. Com base na matriz SCM, foi possível idealizar a criação de quatro conjuntos de características, chamados de MX, MY, ATT e VT. A Figura 19 (b) detalha a geração de cada um desses conjuntos. Os conjuntos 1 e 2 representam as marginais da matriz SCM. Cada elemento do vetor MX é o resultado da soma das colunas da matriz SCM, e cada elemento de MY, é o resultado da soma das linhas. Por outro lado, cada elemento do vetor ATT é o resultado dos cálculos dos diferentes atributos (discutidos na seção 3.2.1.1) que podem ser extraídos, também, da matriz SCM. E o conjunto 4, é o resultado da aplicação de uma transformação de matrizes, ou seja, VT é a matriz linha da matriz SCM.

Outra discussão que pode ser feita sobre a Figura 19 é com relação ao tamanho de cada conjunto de características. Pode-se observar que os conjuntos 1, 2 e 4 crescem em função do valor de N . MX e MY seguem uma ordem de crescimento linear e VT, uma ordem de crescimento polinomial. No entanto, o tamanho de ATT é fixo a quantidade de atributos calculados. Nesse trabalho foram usados oito atributos.

Figura 19 – (a) Aplicação do método SCM, com parâmetro $N = 8$, para extração de características de uma imagem da classe de tecido colorretal estroma complexo. (b) Conjuntos de características construídos a partir da matriz SCM.

