



UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Matheus Gonçalves Ribeiro

Classificação do câncer colorretal baseada na associação de características multidimensional e multirresolução

São José do Rio Preto
2018

Matheus Gonçalves Ribeiro

Classificação do câncer colorretal baseada na associação de características multidimensional e multirresolução

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Financiadora: CAPES

Orientador:

Prof. Dr. Leandro Alves Neves

**São José do Rio Preto
2018**

Ribeiro, Matheus Gonçalves.

Classificação do câncer colorretal baseada na associação de características multidimensional e multirresolução / Matheus Gonçalves Ribeiro. -- São José do Rio Preto, 2018

90 p. : il. ; grafs. , tabs.

Orientador: Leandro Alves Neves

Dissertação (Mestrado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Ciência da computação. 2. Fractais. 3. Cólon (Anatomia) – Câncer. I. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. II. Título.

CDU – 518.72:61

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Matheus Gonçalves Ribeiro

Classificação do câncer colorretal baseada na associação de características multidimensional e multirresolução

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Financiadora: CAPES

Comissão Examinadora

Prof. Dr. Leandro Alves Neves
Unesp - Câmpus São José do Rio Preto
Orientador

Prof. Dr. Geraldo Francisco Donegá Zafalon
Unesp - Câmpus São José do Rio Preto

Prof. Dr. Bruno Augusto Nassif Travençolo
UFU - Universidade Federal de Uberlândia

São José do Rio Preto
06 de Agosto de 2018

A ciência atua na fronteira entre o conhecimento e a ignorância sem medo de admitir que não sabemos. Não há nenhuma vergonha nisso. A única vergonha é fingir que temos todas as respostas.

Neil deGrasse Tyson

Agradecimentos

Gostaria de agradecer primeiramente à minha mãe Maria Cleuza e meu pai Joaquim, que me deram suporte incondicional durante esses anos de graduação e mestrado.

Agradeço à orientação do Prof. Dr. Leandro Alves Neves, que me ajudou e aconselhou desde a época da graduação, além de ser um dos maiores incentivadores para eu iniciar minha carreira acadêmica. Sou grato também por todos os professores da graduação e pós-graduação que por meio do que me ensinaram, puderam contribuir para que eu chegar à essa etapa de minha vida.

Agradeço também à todos amigos do laboratório, Álvaro, Bruna, Camila, Daniela, Gabriela, Guilherme Botazzo, Guilherme Freire, Leonardo, Matheus Carreira, Matheus Popp e Rafaela, por todos os momentos e vivências incríveis, que levarei por toda a vida.

Agradeço ao meu amigo Leandro Bertini Lara Gonçalves, pelo companheirismo e por dividir alegrias, tristezas e sucessos desde a graduação.

Por fim, agradeço ao apoio financeiro da CAPES, que possibilitou o desenvolvimento deste trabalho.

Resumo

Neste trabalho, um método foi proposto a partir da combinação entre técnicas fractais multiescalas e multidimensionais, transformada curvelet e descritores de Haralick para o estudo e reconhecimento de padrões do câncer colorretal nos estágios T3 e T4. Esta combinação ainda não foi explorada na literatura. O método foi baseado em uma abordagem para seleção de características e diferentes técnicas de classificação para avaliar as combinações, tais como árvore de decisão, floresta aleatória, máquina de vetores de suporte, *naive bayes*, *k*-estrela e um método polinomial. Os testes foram avaliados a partir da validação cruzada com *10-folds*. A estratégia proposta permitiu realizar interpretações mais precisas ao considerar as melhores associações para separar as classes benigno e maligno de imagens histológicas colorretais. O melhor resultado foi uma taxa de *AUC* de 0,994, com características baseadas principalmente em lacunaridade e percolação obtidas a partir de subimagens curvelet, aplicando um classificador polinomial. O desempenho obtido com uma análise detalhada envolvendo diferentes tipos de características e classificadores são contribuições importantes para patologistas e especialistas interessados no estudo do câncer colorretal.

Palavras-chave: curvelet, dimensão fractal, lacunaridade, descritores de haralick.

Abstract

In this study, a method is proposed that associates multidimensional fractal techniques, curvelet transforms and Haralick features for the study and pattern recognition of colorectal cancer in stages T3 and T4, a combination not yet explored in the literature. The method considered a feature selection approach and different classification techniques for evaluating combinations, such as decision tree, random forest, support vector machine, naive bayes, k-star and a polynomial method. This strategy allowed for more precise interpretations regarding the best associations for the separation into benign and malignant classes concerning colorectal histological images. The best result was reached with features based mainly on lacunarity and percolation obtained from curvelet sub-images, using a polynomial classifier. The tests were evaluated by applying the 10-fold cross-validation and the best result was a rate of $AUC = 0.994$. The obtained performance with a detailed analysis involving different types of features and classifiers are important contributions both to pathologists and specialists interested in the study of colorectal cancer.

Keywords: *curvelet, fractal dimension, lacunarity, haralick features.*

Lista de Figuras

2.1	Ilustração da curva de Koch para evidenciar a propriedade de autossimilaridade.	20
2.2	Exemplo de imagem com tamanho 4×4 <i>pixels</i> e 5 níveis de cinza. . . .	25
2.3	Exemplo de uma matriz de co-ocorrência da Figura 2.2 com os parâmetros $d = 1$ e $\theta = 0^\circ$	25
2.4	Exemplo de uma matriz de co-ocorrência da Figura 2.2 com os parâmetros $d = 1$ e $\theta = 90^\circ$	26
4.1	Um resumo do método para avaliar imagens histológicas colorretais H&E.	50
4.2	<i>Wrapping</i> de uma imagem.	51
4.3	Ilustração do vetor de características obtido após a aplicação dos métodos fractais e características de Haralick às imagens colorretais e sub-imagens curvelet correspondentes.	58
4.4	Ilustração do processo de validação cruzada e seleção por <i>folds</i> aplicado no método proposto.	60
4.5	Fluxograma do processo de contagem de características sem repetições.	61
4.6	Exemplos de imagens histológicas colorretais do grupo benigno. . . .	63
4.7	Exemplos de imagens histológicas colorretais do grupo maligno. . . .	63

5.1	Exemplos de uma imagem de cólon do grupo benigno (a) e as subimagens obtidas com uma variação de escalas e rotações: escala 1 e rotação 1 (b), escala 2 e rotação 2 (c), escala 2 e rotação 4 (d), escala 2 e rotação 8 (e), escala 3 e rotação 2 (f), escala 3 e rotação 6 (g), escala 3 e rotação 8 (h), escala 4 e rotação 6 (i), escala 4 e rotação 7 (j), escala 4 e rotação 14 (k), escala 4 e rotação 15 (l).	65
5.2	Exemplos de uma imagem de cólon do grupo maligno (a) e as subimagens obtidas com uma variação de escalas e rotações: escala 1 e rotação 1 (b), escala 2 e rotação 2 (c), escala 2 e rotação 4 (d), escala 2 e rotação 8 (e), escala 3 e rotação 2 (f), escala 3 e rotação 6 (g), escala 3 e rotação 8 (h), escala 4 e rotação 6 (i), escala 4 e rotação 7 (j), escala 4 e rotação 14 (k), escala 4 e rotação 15 (l).	66
5.3	Taxas médias de AUC calculadas a partir dos <i>folds</i> e aplicando os métodos DT, SVM, NB, RaF, K* e PL. As taxas foram obtidas com subconjuntos de 1 a 100 características, limites aplicados em cada <i>fold</i>	67
5.4	Visualização de uma imagem do grupo benigno com regiões marcadas em azul para indicar a preservação da organização e da distribuição das criptas (a) e (b), bem como das células caliciformes (c).	69
5.5	Imagem do grupo benigno com regiões marcadas em azul para indicar a degradação da organização e da distribuição das criptas (a) e (b), bem como das células caliciformes (c).	70
5.6	Taxas médias de AUC obtidas a partir de diferentes níveis de ruído (0%, 20% e 40%) sobre as características, considerando as combinações C vs D, D vs C e D vs D.	71
B.1	Exemplos de imagens de colorretais do grupo tumor da base pública.	79
B.2	Exemplos de imagens de colorretais do grupo estroma da base pública.	80

Lista de Tabelas

3.1	Sumarização dos trabalhos correlatos, com informações sobre métodos de extração de características, classificação e desempenho obtido.	48
4.1	Exemplo de representação de uma matriz de probabilidade.	53
4.2	Exemplo de tabela de partição de uma imagem de tamanho L_x por L_y .	54
5.1	Discriminação das características selecionadas pelo método proposto.	67
5.2	Características selecionadas em todos os <i>folds</i> da abordagem de validação cruzada <i>k-folds</i> com médias para os grupos benigno e maligno. As métricas de curva para a lacunaridade e percolação foram: área sob a curva (<i>ARC</i>), obliquidade (<i>OBL</i>), razão de área (<i>RA</i>), ponto máximo (<i>PM</i>) e escala do ponto máximo (<i>EPM</i>).	68
5.3	Desempenhos dos algoritmos de classificação com a combinação de características obtida em nosso modelo.	70
5.4	Desempenho do método proposto e estudos correlatos considerando os valores de <i>AUC</i>	72
5.5	Comparação do método proposto com a base colorretal disponibilizado por (KATHER et al., 2016).	73
A.1	Exemplos das características obtidas a partir do método proposto. Considerando os grupos benigno e maligno.	77

Lista de Abreviações

AC Acurácia

ARC Área sob a curva

AUC *Area under the receiver operating characteristic curve*

DBC *Differential Box Counting*

DF Dimensão Fractal

DT *Decision Tree*

EPM Escala do Ponto Máximo

FDCT *Fast Discrete Curvelet Transform*

K* K-estrela

Lac Lacunaridade

NB *Naive Bayes*

OBL Obliquidade

Perc Percolação

PL Polinomial

PM Ponto Máximo

RaF *Random Forest*

ROC *Receiver Operating Characteristic*

RZA Razão de área

SVM *Support Vector Machine*

Sumário

Lista de Figuras	6
Lista de Tabelas	8
Lista de Abreviações	9
Sumário	11
1 Introdução	15
1.1 Motivação e Justificativas	17
1.2 Objetivos	17
1.3 Organização do Trabalho	18
2 Fundamentação Teórica	19
2.1 Descritores de Características de Imagens	19
2.2 Abordagens Fractais	20
2.2.1 Contagem de Caixas	22
2.2.2 Contagem de Caixas Diferencial	22
2.2.3 Dimensão Fractal Multiescala e Multidimensional	23
2.2.4 Percolação	24
2.2.5 Lacunaridade	24
2.3 Descritores de Haralick	25
2.4 Transformada Curvelet	29

2.5	Técnicas para Seleção de Características	30
2.6	Métodos para Classificação de Padrões	31
2.6.1	Árvore de Decisão	33
2.6.2	Máquina de vetores de suporte	33
2.6.3	Random Forest	33
2.6.4	Naive Bayes	34
2.6.5	K-estrela	34
2.6.6	Polinomial	35
2.7	Medidas de Desempenho para Classificação	35
3	Levantamento Bibliográfico	37
3.1	Técnicas Fractais	37
3.2	Transformada Curvelet	39
3.3	Descritores de Haralick	41
3.4	Combinação de Técnicas de Extração de Características	42
3.5	Abordagens para Extração de Características em Imagens Colorretais	45
3.6	Uma Visão Geral Sobre os Trabalhos Relacionados	47
4	Metodologia	50
4.1	Transformada Curvelet	50
4.2	Aplicação das Técnicas Fractais	52
4.2.1	Dimensão fractal baseada na abordagem probabilística	53
4.2.2	Dimensão fractal baseada na abordagem de fusão de caixas	54
4.2.3	Percolação multiescala e multidimensional	54
4.2.4	Lacunaridade	57
4.3	Aplicação dos Descritores de Haralick	58
4.4	Composição do Vetor de Características	58
4.5	Seleção das Características	59

4.6	Classificação e avaliação das características	60
4.6.1	Análise de ruído	62
4.7	Contexto de Aplicação: Banco de Imagens Histológicas Colorretais .	63
5	Resultados	65
5.1	Análise do impacto do ruído em diferentes características	71
5.2	Comparação com Trabalhos Correlatos	73
5.3	Comparação com base de imagens pública	74
6	Conclusões	75
6.1	Contribuições Obtidas	76
6.2	Trabalhos Futuros	76
A	Exemplo de vetores de características após obtenção dos atributos	77
B	Exemplos de imagens colorretais da base pública	79
	Referências Bibliográficas	82

Capítulo 1

Introdução

O câncer colorretal é um tumor maligno que se desenvolve na parede interna do intestino grosso (cólon) ou reto (ALTERI; KRAMER; SIMPSON, 2014). Em 2012, a Agência Internacional de Pesquisa sobre o Câncer (IARC) apresentou um estudo global em que o câncer colorretal foi definido como o terceiro câncer mais comum em homens (746.000 casos) e o segundo em mulheres (614.000 casos), com 55% dos casos diagnosticados em regiões mais desenvolvidas. A mortalidade foi de 694.000 casos e com maior incidência (52% das mortes) nas regiões menos desenvolvidas do mundo (IARC, 2012). No Brasil, a estimativa para cada ano do biênio de 2018-2019 é de 36.360 novos casos e 16.696 mortes por ano (INCA, 2017). Nos Estados Unidos, no ano de 2018 a ocorrência foi de 140 mil casos novos e 50 mil mortes (SIEGEL; MILLER; JEMAL, 2018). O diagnóstico de câncer colorretal pode ser feito por sigmoidoscopia ou por colonoscopia com a confirmação por biópsia de tecido. Os resultados são amostras de tecidos coradas com hematoxilina e eosina (H&E), as quais são avaliadas visualmente por patologistas. Essa tarefa é complexa e demanda tempo do especialista, o que pode levar a interpretações equivocadas. Para minimizar esses problemas, métodos computacionais foram propostos para apoiar os patologistas na tarefa de classificação e reconhecimento de padrões de tecidos colorretais corados com H&E (MASOOD; RAJPOOT, 2009; KALKAN et al., 2012; RATHORE et al., 2013; NAIYAR; ASIM; SHAHID, 2015; JØRGENSEN et al., 2017). Parte do sucesso de cada método pode ser atribuído a etapa de extração de características.

Diferentes abordagens podem ser utilizadas para extrair características (HARALICK, 1982), como as técnicas exploradas neste trabalho. Por exemplo, os descritores de Haralick (HARALICK; SHANMUGAM et al., 1973) foram aplicados para investigações de doenças do fígado (SUGANYA; RAJARAM, 2013), detecção de atividade epilética (BOUBCHIR; AL-MAADEED; BOURIDANE, 2014), diagnóstico

de glaucoma (SAMANTA et al., 2014) e detecção de câncer bucal (CHAKRABORTY et al., 2016). As abordagens fractais são úteis para quantificação de propriedades de autossimilaridade e alguns exemplos de descritores fractais são dimensão fractal (DF), lacunaridade (Lac) e Percolação (Perc). DF quantifica o nível de preenchimento do espaço na imagem, além de indicar a complexidade e irregularidade de um fractal (IVANOVICI; RICHARD, 2011). Lac indica como os *pixels* estão distribuídos e organizados em uma imagem. Perc consiste na ocorrência de um caminho que conecta uma extremidade à outra de uma imagem (GOULD et al., 1996). Sistemas de reconhecimento de padrões a partir de técnicas fractais obtiveram êxitos com lesões psoriáticas (IVANOVICI; RICHARD; DECEAN, 2009), detecção de glaucoma (LAMANI; MANJUNATH et al., 2014) e diagnóstico de doenças mamárias (DOBRESCU et al., 2014), estudo do comportamento do câncer de próstata (NEVES et al., 2014) e a quantificação de linfomas non-Hodgkin (ROBERTO et al., 2017), por exemplo. Outra categoria destacada aqui permite representar os métodos baseados em frequência (TANG; HE, 2013). Nesta, a transformada curvelet é útil por considerar escalas e ângulos de observações para analisar singularidades de curvas e linhas. A transformada curvelet foi aplicada com sucesso no estudo de úlceras em imagens endoscópicas (EID et al., 2013), classificação dos diferentes níveis de câncer prostático (LIN et al., 2015), diagnóstico do câncer de mama (SARASWATHI; DHARANI; SRINIVASAN, 2016) e detecção de doenças cerebrovasculares e neoplásticas (NAYAK et al., 2017).

Apesar da relevância no processo de quantificação de texturas com os descritores de Haralick, métodos fractais e transformada curvelet, a combinação destas abordagens ainda foi pouco explorada na literatura. Estes métodos podem ser combinados para fornecer características multidimensionais e multirresolução. Por exemplo, a transformada curvelet foi combinada com análises realizadas por meio de DF ou Lac, separadamente, tais como as apresentadas nos trabalhos de (EID et al., 2013; LIU et al., 2016; CHARISIS; HADJILEONTIADIS, 2016; SARASWATHI; SRINIVASAN, 2017). Eid et al. (2013) propuseram a combinação da transformada curvelet com a lacunaridade em níveis de cinza. Liu et al. (2016) exploraram a transformada curvelet com DF sobre subimagens geradas a partir de quatro escalas e oito rotações. Saraswathi and Srinivasan (2017) combinaram transformada curvelet com DF para análises de microcalcificação em mamogramas digitais. Charisis et al. (2016) exploraram transformada curvelet e Lac para estudar lesões causadas pela doença de Chron. Porém, não foram encontrados trabalhos que exploraram a transformada curvelet com análises considerando DF, Lac e Perc com estratégias multiescalas e multidimensionais, especialmente para auxiliar o diagnóstico de câncer colorretal.

1.1 Motivação e Justificativas

As técnicas fractais, transformada curvelet e os descritores de Haralick são métodos amplamente aplicados para avaliar texturas (BOUBCHIR; AL-MAADEED; BOURIDANE, 2014; DOBRESCU et al., 2014; LIN et al., 2015; SARASWATHI; SRINIVASAN, 2017). As técnicas fractais são vantajosas por fornecerem propriedades estocásticas ou de autossimilaridade presentes em imagens. A transformada curvelet realiza a extração de singularidades curvas e as representa em subimagens, caracterizando-se assim como uma técnica multirresolução. Os descritores de Haralick são obtidos ao definir dependências espaciais e de intensidades de níveis cinza. A vantagem dessa abordagem é permitir aplicações em diferentes contextos de imagens, extraindo propriedades como contraste, correlação e entropia. Apesar de algumas combinações destas abordagens constar em (EID et al., 2013), (LIU et al., 2016), (CHARISIS; HADJILEONTIADIS, 2016) e (SARASWATHI; SRINIVASAN, 2017), no contexto de imagens histológicas colorretais não foram encontrados trabalhos que explorem o reconhecimento de padrões a partir da combinação das características previamente citadas.

A quantificação com técnicas fractais pode ser realizada com abordagens baseadas em níveis de cinza ou com estratégia multidimensional (colorida) (IVANOVICI; RICHARD, 2011). A abordagem multidimensional expande o potencial das técnicas fractais (dimensão fractal, lacunaridade e percolação), por fornecer quantificações mais amplas e precisas, uma vez que envolve informações presentes em diferentes canais de cores de uma imagem. A transformada curvelet permite extrair informações com escalas e ângulos distintos, gerando subimagens que podem ser quantificadas com a dimensão fractal, lacunaridade, percolação e descritores de Haralick. Os descritores de Haralick, por sua vez, complementam o processo de quantificação por fornecer uma análise probabilística completa. Assim, a combinação de características proposta neste trabalho não está presente em outras propostas disponíveis na literatura (EID et al., 2013; LIU et al., 2016; CHARISIS; HADJILEONTIADIS, 2016; ALTUNBAY et al., 2010; KALKAN et al., 2012; AKBAR; GOPI; BABU, 2015; NAIYAR; ASIM; SHAHID, 2015). Os diferenciais desta combinação consistem em análises mais completas das imagens, explorando inclusive quantificações com a dimensão fractal, lacunaridade e percolação de maneira concomitante, via abordagens multidimensionais.

1.2 Objetivos

O objetivo geral deste trabalho consistiu em realizar a combinação de características para classificar o câncer colorretal a partir de imagens H&E. Para tanto, os objetivos

específicos foram:

- I Explorar as melhores associações entre técnicas fractais, multiescala e multidimensional, com a transformada discreta curvelet (abordagem multirresolução) e descritores de Haralick;
- II Analisar o poder discriminativo de cada associação a partir de diferentes métodos de classificação de características;
- III Aplicar as estratégias em imagens colorretais H&E;
- IV Identificar o conjunto de características capaz de fornecer taxas relevantes para distinguir os grupos do câncer colorretal, mesmo sob a influência de altos níveis de ruído.

1.3 Organização do Trabalho

No Capítulo 2 é apresentada a fundamentação teórica, em que estão descritos métodos para extrair características de imagens, tais como técnicas fractais, descritores de Haralick e transformada curvelet. Informações para o entendimento sobre abordagens para selecionar características, métodos para classificar dados, medidas de desempenho também estão no Capítulo 2. No Capítulo 3 estão os trabalhos relacionados mais relevantes para embasar a proposta de combinação de técnicas. No Capítulo 4 é descrita a metodologia para a combinação de técnicas. No Capítulo 5 estão os resultados obtidos, bem como as conclusões e contribuições a partir do método proposto.

Capítulo 2

Fundamentação Teórica

Neste capítulo são apresentados os principais conceitos sobre técnicas para quantificação de texturas, com detalhes na Seção 2.1 sobre a organização das abordagens. As técnicas para calcular a dimensão fractal, lacunaridade e percolação multiescala e multidimensional são apresentadas na Seção 2.2. Informações sobre as matrizes de coocorrência e descritores de Haralick encontram-se na Seção 2.3. As técnicas para calcular a transformada curvelet são apresentadas na Seção 2.4. Os conceitos e técnicas para seleção de características estão na Seção 2.5. Os métodos para classificação de padrões são descritos na Seção 2.6. As medidas de desempenho para avaliação dos métodos de classificação são indicadas na Seção 2.7.

2.1 Descritores de Características de Imagens

Um dos principais problemas no desenvolvimento de um sistema para análise de imagens consiste na obtenção de um conjunto de características extraídas do objeto de interesse para a realização da classificação. Um objeto pode ser representado a partir de suas características externas (bordas) ou internas (PEDRINI; SCHWARTZ, 2008).

A forma de um objeto consiste em uma informação visual importante e tem sido utilizada como um dos tipos de características básicas para descrever uma imagem. Entretanto, esse tipo de descrição pode sofrer distorções e outros tipos de degradações devido à presença de ruído. Uma alternativa consiste em obter informações sobre as texturas dos objetos presentes em imagens (TANG; HE, 2013). Não há uma definição precisa sobre o que é textura, mas segundo TAMURA, MORI e YAMAWAKI (1978), a textura pode ser considerada como uma região macroscópica, em que sua estrutura é atribuída simplesmente à padrões repetitivos, cujos elementos ou primitivas são dispostas com base em uma regra de colocação. Assim, o modelo de textura

é comumente definido por primitivas em níveis de cinza ou coloridas, denominados *textons* (JULESZ; BERGEN, 1983) ou *texels* (HARALICK, 1979). Exemplos de técnicas para obter informações sobre texturas são abordagens fractais, como dimensão fractal, lacunaridade e percolação, transformada curvelet e os descritores de Haralick (LIU; WANG, 2003).

2.2 Abordagens Fractais

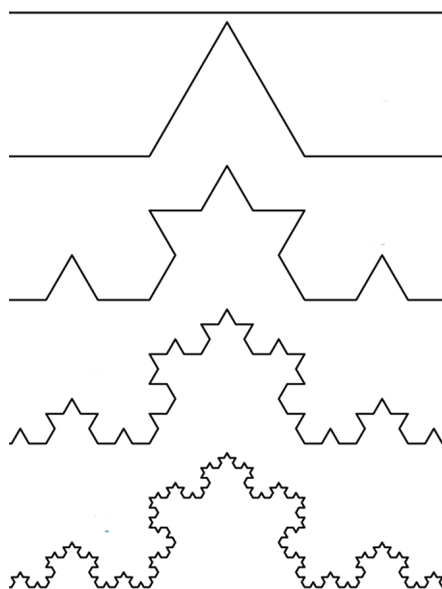
A geometria fractal foi proposta por MANDELBROT (1983) com o objetivo de minimizar as lacunas da geometria Euclidiana, no que diz respeito as restrições para representar formas presentes na natureza. A principal característica dos fractais consiste no estudo de formas autossimilares, que intuitivamente podem ser definidas como uma figura geométrica ou objeto natural que é composto de duas características: (1) suas partes têm a mesma forma ou estrutura do todo. Esses elementos característicos cobrem uma ampla gama de escalas e podem estar levemente deformados; (2) sua forma é irregular, interrompida ou fragmentada, em qualquer escala de observação.

O padrão de autossimilaridade é encontrado com frequência na natureza, como em folhas, raios, flocos de neve e montanhas. Um exemplo apresentado na literatura consiste na curva de Koch (MANDELBROT; PIGNONI, 1983), em que as retas de cada geração são divididas em quatro retas similares, reduzidas a $1/3$ de seu tamanho. Para exemplificar, na Figura 2.1 são apresentadas cinco gerações da curva de Koch.

A dimensão é uma das características mais importantes na definição de um fractal. Diferentemente do que é observado na geometria Euclidiana, em que valor de dimensão indica a dimensionalidade do espaço utilizado para representação do objeto, a dimensão fractal representa o nível de irregularidade de uma forma (objeto) (MANDELBROT; PIGNONI, 1983). Portanto, a dimensão fractal (DF) mensura o quanto do espaço encontra-se preenchido e pode assumir valores fracionários. Há uma relação de que quanto maior a irregularidade de uma forma ou objeto, maior é sua dimensão fractal.

Na literatura foram apresentadas diversas abordagens para determinar a dimensão fractal, como a dimensão de similaridade (MANDELBROT; PIGNONI, 1983), dimensão de informação (FALCONER, 2004), entre outras. Entretanto, a maioria das abordagens se baseia na dimensão de Hausdorff (GRASSBERGER, 1981). A dimensão de Hausdorff pode ser entendida considerando duas retas, uma de comprimento L e outra de comprimento u , de modo que $L > u$. O número de retas N necessário para cobrir L é obtido sobrepondo a reta u sobre a reta L , definido como $N = L/u$. Analogamente

Figura 2.1: Ilustração da curva de Koch para evidenciar a propriedade de autossimilaridade.



Fonte: Adaptado de (MANDELBROT; PIGNONI, 1983, p. 44).

com a reta, é possível medir a dimensão de um quadrado de lado L sobrepondo-o com pequenos quadrados de lado u . Essa relação é dada por $N = (L/u)^2$ (TRICOT, 1994). Generalizando para D dimensões, obtém-se $N = (L/u)^D$, ao aplicar o logaritmo em ambos os lados:

$$D = \frac{\log(N)}{\log(L/u)}, \quad (2.1)$$

em que D é a dimensão fractal de Hausdorff do objeto analisado.

Existem diversos métodos para estimativa da dimensão fractal, entretanto nem todos podem ser aplicados a qualquer tipo de estrutura. Isso ocorre porque cada método foi proposto para quantificar categorias diferentes de objetos, consequentemente gerando diferentes resultados para um mesmo objeto (BACKES; BRUNO, 2006). Para imagens, a dimensão fractal pode ser obtida a partir de estratégias que quantificam o objeto em níveis de cinza ou que consideram as informações dos canais de cores, conhecidas como abordagens multidimensionais.

2.2.1 Contagem de Caixas

Um método amplamente conhecido para estimar a dimensão fractal é o algoritmo de contagem de caixa (*box counting*), apresentado por (GAGNEPAIN; ROQUES-CARMES, 1986). Nesta abordagem, um espaço 3D é considerado a partir das coordenadas (x,y) , representando a posição de cada *pixel* e da coordenada z , definida pela intensidade da imagem. Assim, dada um escala L , a partição do espaço tridimensional é realizada em caixas de tamanhos $L \times L \times L'$, em que L é o múltiplo do comprimento lateral de um pixel (x,y) e L' é o múltiplo do comprimento da intensidade em níveis de cinza na direção z . Considerando G o número total de níveis de cinza e $M \times M$ o tamanho da imagem, então $L' = \lfloor L \times G/M \rfloor$. Portanto, se uma caixa contém pelo menos uma amostra na superfície de intensidade de níveis de cinza, essa caixa é denominada caixa contável. Ao fixar um valor para L , é possível realizar uma varredura sem sobreposições de modo a obter o número total de caixas contáveis (N_L). A cada nova iteração, diferentes valores de L são escolhidos. Ao final desse processo, os valores obtidos em L e N_L são utilizados para obter a dimensão fractal por meio de um ajuste linear $\log(L)$ versus $\log(N_L)$.

2.2.2 Contagem de Caixas Diferencial

O método contagem de caixas diferencial (*Differential Box Counting* - DBC) foi proposto por (SARKAR; CHAUDHURI, 1992). A abordagem DBC baseia-se na definição matemática de autossimilaridade estabelecida por MANDELBROT ((MANDELBROT; PIGNONI, 1983)), em que é considerado um conjunto delimitado A no espaço Euclidiano n -dimensional. O conjunto é dito autossimilar quando A é a união de N_r cópias distintas entre si, cada uma sendo similar a A reduzidos à uma razão r , Eq. 2.2:

$$1 = N_r r^D \text{ ou } D = \frac{\log(N_r)}{\log(1/r)}. \quad (2.2)$$

O método DBC realiza a contagem de N_r de forma diferente do método de contagem de caixas. Para tanto, o tamanho da imagem definido como $M \times M$ *pixels* deve ser reduzido a um tamanho $s \times s$ em que $M/2 \geq s > 1$ e s é um número inteiro. Dessa forma, SARKAR e CHAUDHURI (1992) propuseram uma estimativa para o valor da razão, dada por $r = s/M$.

A etapa seguinte da abordagem DBC é similar à abordagem de contagem de caixas, em que a imagem analisada é considerada em um espaço tridimensional, com (x,y) representando a posição e z obtido a partir da intensidade do nível de cinza. O espaço (x,y) é particionado em grades de tamanho $s \times s$, em cada grade há uma coluna de

caixas de tamanho $s \times s \times s$. Um número sequencial é atribuído a cada caixa da base ao topo, iniciando a contagem no canto superior esquerdo da grade analisada. As intensidades mínima (k) e máxima (l) dos níveis de cinza são obtidas em cada grade (i, j) . Assim, com base nos valores de intensidade k e l , a contribuição de cada grade (i, j) pode ser calculada com base na Eq. 2.3.

$$n_r(i, j) = l - k + 1. \quad (2.3)$$

A contribuição de todas as grades N_r é obtida a partir da contribuição individual de cada grade $N_r(i, j)$, conforme Eq. 2.4.

$$N_r = \sum_{i, j} n_r(i, j). \quad (2.4)$$

Diferentes valores de r são considerados para a contagem de N_r . A Eq. 2.2 é aplicada para a estimativa da dimensão fractal D , a partir dos ajustes lineares de $\log(N_r)$ contra $\log(1/r)$.

2.2.3 Dimensão Fractal Multiescala e Multidimensional

Com o objetivo de se obter mais informações na quantificação da dimensão fractal, as técnicas em níveis de cinza foram expandidas para considerar as informações presentes em cada canal de cor da imagem analisada. O padrão de cores mais utilizado é baseado nos canais *Red* (vermelho), *Green* (verde) e *Blue* (azul). Assim, cada *pixel* da imagem é representado por meio de um vetor de 5 dimensões (x, y, r, g, b) : x e y consideram a posição do pixel no plano; r , g e b consideram os valores dos níveis de intensidade presentes nos canais R, G e B, respectivamente. As principais abordagens para extração da dimensão fractal multiescala e multidimensional foram propostas por IVANOVICI, RICHARD e DECEAN (2009) e NIKOLADIS I., NIKOLAIDS N. e TSOUROS (2011).

Na abordagem de IVANOVICI, RICHARD e DECEAN (2009), um método probabilístico é proposto para quantificação da dimensão fractal colorida. Nessa técnica, uma matriz de probabilidade é definida para mensurar a contribuição dos *pixels* observados em cada escala de observação, considerando a distância infinita de Minkowski (KRUSKAL, 1964).

NIKOLADIS I., NIKOLAIDS N. e TSOUROS (2011) propuseram a abordagem de fusão de caixas (*box merge*), que consistiu na expansão da quantificação da dimensão fractal para espaços n -dimensionais, baseado método de contagem de caixas. Diferentemente do método proposto em (IVANOVICI; RICHARD; DECEAN, 2009),

os autores propuseram a construção de uma tabela de partição para mensurar as contribuições dos *pixels* em cada escala de observação.

2.2.4 Percolação

A percolação (Perc) consiste em um fenômeno físico de transporte de fluídos através de um meio poroso, como óleo ou gás através de uma rocha, ou água através do pó de café (FONTES, 1996). O modelo de percolação foi inicialmente proposto em (BROADBENT; HAMMERSLEY, 1957), com o objetivo de representar esse fenômeno físico.

A base para a definição da percolação consiste na formação de aglomerados de forma probabilística e ao considerar um caminho que conecta duas extremidades de um sistema com base nos aglomerados. A partir de um valor específico de p , denominado limiar percolante, é garantida a existência de um aglomerado percolante (DEAN, 1963).

Existem dois tipos de percolação: de lugar e de ligação. A percolação de lugar pode ser definida a partir da probabilidade de ocupação, determinando se um lugar no sistema está ou não ocupado (MALARZ, 2015). Este tipo de percolação é a mais comum para o processamento de imagens em razão da analogia feita entre *pixels* e lugares (YAMAGUCHI; HASHIMOTO, 2010). O modelo de percolação comumente utilizado para imagens consiste em definir matrizes quadradas com base em uma grade G de tamanho $M \times M$ espaços (*pixels*) e uma probabilidade p de que cada espaço não esteja ocupado, denominados poros, a percolação ocorre quando um aglomerado de poros conecta duas extremidades paralelas da grade (GOULD et al., 1996).

2.2.5 Lacunaridade

A lacunaridade (Lac) consiste em outra característica que pode ser observada em um fractal. VOSS (1991) definiu a lacunaridade como a entropia dos pontos de uma superfície discreta que representa a imagem. Intuitivamente, essa métrica compreende a maneira como os *pixels* estão distribuídos e organizados em uma determinada imagem. A lacunaridade é considerada como uma métrica complementar à dimensão fractal, uma vez que formas que apresentam as mesmas dimensões fractais podem possuir lacunaridades diferentes (PLOTNICK et al., 1996). A lacunaridade está intrinsecamente ligada à organização topológica dos objetos em uma imagem. Assim, valores pequenos de lacunaridades indicam que as imagens são mais homogêneas, ao considerar o tamanho da distribuição e o arranjo espacial das lacunas. Por outro lado, imagens com

lacunaridades maiores são mais heterogêneas.

Assim como a dimensão fractal, a lacunaridade pode ser calculada a partir de imagens coloridas ou em níveis de cinza. A abordagem comumente aplicada para quantificação da lacunaridade em níveis de cinza é denominada de *gliding-box*, que é um algoritmo similar ao de contagem de caixas. O funcionamento básico da abordagem de *gliding-box* consiste em posicionar uma caixa de comprimento de lado r no canto superior esquerdo da imagem, calculando o número de pontos presentes na caixa. A principal diferença em relação ao processo de contagem de caixas reside na leitura dos *pixels*, pois as partições são sobrepostas, avançando *pixel a pixel* de modo a realizar a varredura em toda a imagem. O escaneamento continua até o canto inferior direito da imagem. Ao final dessa etapa o valor de r é incrementado para uma nova varredura (GRAU et al., 2006).

Outra maneira para se obter a lacunaridade consiste na abordagem probabilística, proposta por (VOSS, 1991) para quantificar imagens em níveis de cinza. Uma extensão dessa abordagem para imagens coloridas foi apresentada por IVANOVICI e RICHARD (2009). Para tanto, uma matriz de probabilidade é elaborada para definir a lacunaridade colorida, de modo similar à abordagem da dimensão fractal multiescala e multidimensional.

2.3 Descritores de Haralick

Os descritores de Haralick (HARALICK; SHANMUGAM et al., 1973) foram definidos a partir das matrizes de co-ocorrência, sendo um dos métodos mais utilizados na literatura para extração de características em texturas (CHADDAD et al., 2011). A obtenção dos descritores de Haralick é baseada na matriz de co-ocorrência. Nessa matriz, cada elemento representa a frequência em que os diferentes níveis de cinza ocorrem na imagem, considerando uma distância d que separa esses *pixels* e uma direção θ . Um aspecto importante da matriz de co-ocorrência diz respeito ao número de elementos da matriz, que é definido pela quantidade de intensidade G presentes na textura avaliada. Dessa forma, uma imagem é representada como uma matriz de tamanho $G \times G$, em que cada elemento representa a probabilidade de ocorrência de dois *pixels*, com níveis de cinza i e j , separados por uma distância d e por um ângulo θ entre esses *pixels*. Essa relação pode ser representada como $matco(i, j, d, \theta)$ ou como $matco(i, j)$, para o caso em que a distância e direção já tenham sido explicitadas.

Diferentes matrizes de probabilidade são obtidas com alterações dos valores de d e θ . A Figura 2.2 representa uma imagem com 4×4 *pixels* e 5 níveis de intensidades

de cinza distintos. A partir dessa imagem, duas matrizes de co-ocorrência distintas podem ser obtidas, a matriz apresentada na Figura 2.3 avalia os *pixels* vizinhos horizontalmente e na Figura 2.4 verticalmente. Dessa forma, os parâmetros considerados foram $d = 1$ e $\theta = 0^\circ$ (Figura 2.3) e $d = 1$ e $\theta = 90^\circ$ (Figura 2.4).

Figura 2.2: Exemplo de imagem com tamanho 4×4 *pixels* e 5 níveis de cinza.

0	2	2	4
0	3	1	0
1	1	3	2
4	2	0	2

Fonte: elaborado pelo autor.

Figura 2.3: Exemplo de uma matriz de co-ocorrência da Figura 2.2 com os parâmetros $d = 1$ e $\theta = 0^\circ$.

	0	1	2	3	4
0	0	0	2	1	0
1	1	1	0	1	0
2	1	0	1	0	1
3	0	1	1	0	0
4	0	0	1	0	0

Fonte: elaborado pelo autor.

A partir das Figuras 2.3 e 2.4, é possível notar que a distinção entre as matrizes de co-ocorrência obtidas. Outra constatação interessante residiu no fato que a matriz de co-ocorrência da Figura 2.3 apresentou o valor 2 em $matco(0,2,1,0^\circ)$, devido à duas ocorrências dos níveis de cinza $i = 0$ e $j = 2$ horizontalmente.

Haralick et al. (1973) propuseram 14 medidas de textura para facilitar a análise das matrizes de co-ocorrência. Estas medidas indicam características gerais das texturas presentes nas imagens.

O primeiro descritor é denominado segundo momento angular e indica a energia ou uniformidade da imagem. A característica é obtida por meio da soma dos elementos elevados ao quadrado na matriz de co-ocorrência normalizada, Eq. 2.5. A energia

Figura 2.4: Exemplo de uma matriz de co-ocorrência da Figura 2.2 com os parâmetros $d = 1$ e $\theta = 90^\circ$.

	0	1	2	3	4
0	1	1	1	0	0
1	0	0	1	1	1
2	0	1	1	1	0
3	1	1	0	0	0
4	1	0	0	0	0

Fonte: elaborado pelo autor.

possui valor 1 para uma imagem com intensidades similares dos níveis de cinza, o que indica que a imagem analisada é homogênea.

$$f_1 = \sum_i \sum_j [matco(i,j)]^2. \quad (2.5)$$

O segundo descritor considerado calcula o contraste entre as intensidades de um *pixel* analisado e do *pixel* vizinho, Eq. 2.6. Para uma imagem com poucas variações de intensidade, ou seja, uma imagem homogênea, o contraste resultante é zero.

$$f_2 = \sum_i \sum_j (i - j)^2 matco(i,j). \quad (2.6)$$

O terceiro descritor apresentado por Haralick et al. é a correlação, que define o quão correlacionado um *pixel* está ao seu vizinho, Eq. 2.7, em que σ_x , σ_y , μ_x , μ_y representam o desvio-padrão e a média de p_x e p_y , respectivamente. O valor de $p_x(i)$ consiste na soma das linhas da matriz de co-ocorrência *matco*, analogamente, $p_y(j)$ consiste na soma das colunas de *matco*. Uma imagem é considerada totalmente correlacionada caso $f_3 = 1$. Caso contrário, $f_3 = -1$.

$$f_3 = \sum_i \sum_j \frac{(i - \mu_x)(j - \mu_y) matco(i,j)}{\sigma_x \sigma_y}. \quad (2.7)$$

A variância da imagem é calculada a partir do descritor soma dos quadrados, que considera a média μ da imagem analisada e as entradas da matriz de co-ocorrência, Eq. 2.8.

$$f_4 = \sum_i \sum_j (i - \mu)^2 matco(i,j). \quad (2.8)$$

O quinto descritor, momento inverso das diferenças, é uma característica complementar ao contraste f_2 , apresentado na Eq. 2.9.

$$f_5 = \sum_i \sum_j \frac{matco(i,j)}{1 + (i-j)^2}. \quad (2.9)$$

O sexto descritor é denominado o somatório médio (Eq. 2.10). A quantidade de níveis de cinza distintos na imagem quantificada é representado por N_g e p_{x+y} é calculado na Eq. 2.11.

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i). \quad (2.10)$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j), \quad i + j = k = 2, 3, \dots, 2N_g. \quad (2.11)$$

O somatório da variância é o sétimo descritor, para tal utiliza-se o atributo f_8 de Haralick (somatório da entropia), Eq. 2.12.

$$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i). \quad (2.12)$$

O oitavo descritor, somatório da entropia é obtido conforme Eq. 2.13.

$$f_8 = - \sum_{i=2}^{2N_g} p_{x+y}(i) \log(p_{x+y}(i)). \quad (2.13)$$

A entropia é o descritor de Haralick de número 9, que tem por característica medir a informação contida em *matco*: muitos valores nulos representam pouca informação, Eq. 24.

$$f_9 = - \sum_i \sum_j matco(i,j) \log(matco(i,j)). \quad (2.14)$$

O descritor 10 consiste na diferença da variância, Eq. 2.15. O descritor de número 11 consiste na diferença da entropia, Eq. 2.16. O valor de p_{x-y} é calculado na Eq. 2.17.

$$f_{10} = var(p_{x-y}). \quad (2.15)$$

$$f_{11} = - \sum_{i=0}^{N_g-1} p_{x-y}(i) \log(p_{x-y}(i)). \quad (2.16)$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \text{matco}(i,j), \quad |i-j| = k = 0, 1, \dots, N_g - 1. \quad (2.17)$$

Os descritores 12 e 13 são medidas de informação de correlação, os quais são definidos por HX e HY como as entropias de p_x e p_y , respectivamente. Os cálculos destes descritores são realizados a partir das Eqs. 2.18 e 2.19. Para tanto, é necessário estimar HXY , Eq. 2.20, $HXY1$, Eq. 2.21, e $HXY2$, Eq. 2.22.

$$f_{12} = \frac{HXY - HXY1}{\max\{HX, HY\}}, \quad (2.18)$$

$$f_{13} = \sqrt{(1 - \exp[-2(HXY2 - HXY)])}, \quad (2.19)$$

$$HXY = - \sum_i \sum_j \text{matco}(i,j) \log(p(i,j)), \quad (2.20)$$

$$HXY1 = - \sum_i \sum_j \text{matco}(i,j) \log[p_x(i)p_y(j)], \quad (2.21)$$

$$HXY2 = - \sum_i \sum_j p_x(i)p_y(j) \log[p_x(i)p_y(j)]. \quad (2.22)$$

O último descritor foi denominado de coeficiente de correlação máximo, sendo apresentado na Eq. 2.23. A matriz Q é obtida a partir do somatório da matriz de co-ocorrência dos *pixels* analisados sobre a i -ésima e j -ésima entradas na matriz de probabilidade marginal, Eq. 2.24.

$$f_{14} = (\text{segundo maior autovalor de } Q)^{1/2}, \quad (2.23)$$

$$Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(j)}. \quad (2.24)$$

2.4 Transformada Curvelet

A transformada curvelet representa os resultados das pesquisas mais recentes na análise de multirresolução. Por combinar os métodos espacial e espectral, a obtenção de informações sobre as bordas do objeto são capturadas com mais precisão do que comparado aos métodos espectrais convencionais, como filtro de Gabor e wavelet (SUMANA et al., 2008). Além disso, segundo LIN et al. (2015) a transformada curvelet

permite uma representação esparsa para singularidades curvas, permitindo a obtenção de características cruciais acerca de contornos e curvas. Esta técnica é aplicada principalmente para análises de imagens médicas (EID et al., 2013; LIN et al., 2015; LIU et al., 2016).

A transformada curvelet foi introduzida por CANDÉS e DONOHO (2004). A versão utilizada atualmente é conhecida como transformada curvelet discreta rápida (*Fast Discrete Curvelet Transform* - FDCT), que possui um número menor de redundâncias. Segundo CANDÉS et al. (2006) existem duas implementações para FDCT: (1) transformada rápida de Fourier desigualmente espaçada (*unequally spaced Fast Fourier transforms*); e, (2) função de invólucro (ou envelopamento) (*wrapping function*).

A principal diferença entre as implementações de FDCT consiste nas diferenças em relação à escolha da grade espacial utilizada para transladar as curvelets em cada escala e ângulo. A transformada baseada em invólucro baseia-se na reunião de amostras selecionadas de Fourier, sendo esta mais simples de entender e implementar (ALZUBI; ISLAM; ABBOD, 2011).

2.5 Técnicas para Seleção de Características

A combinação de características pode resultar em um número elevado de informações para cada imagem analisada. A seleção dos atributos pode ser aplicada para possibilitar a análise de dados, além de proporcionar a classificação mais precisa. Diversas são as razões para realizar a redução de características, tais como: (1) evitar o sobreajuste e aprimorar o desempenho do modelo; (2) fornecer modelos mais rápidos e com consumo menor de recursos, (3) prover melhor compreensão dos processos que geraram os dados. Para tanto, uma estratégia é selecionar as características mais relevantes do contexto estudado. As técnicas disponíveis para esta tarefa podem ser agrupadas em três categorias: métodos de envelopamento, métodos incorporados e métodos de filtragem (SAEYS; INZA; LARRAÑAGA, 2007). As abordagens baseadas em envelopamento integram o aprendizado supervisionado para validar os subconjuntos de características gerados, utilizando estratégias de busca (GNANA; BALAMURUGAN; LEAVLINE, 2016). As técnicas incorporadas têm por objetivo buscar por um subconjunto de características ótimo durante a construção do classificador (SAEYS; INZA; LARRAÑAGA, 2007).

As técnicas de filtragem avaliam a relevância das características ao analisar apenas as propriedades intrínsecas dos dados. Na maioria dos casos uma escala de relevância é calculada, o que permite eliminar características com baixos valores de relevância.

Técnicas de filtragem são facilmente escaláveis para conjuntos de dados de dimensão alta, sendo computacionalmente simples e rápidos, e não dependentes do algoritmo de classificação (SAEYS; INZA; LARRAÑAGA, 2007).

Alguns exemplos de algoritmos de seleção baseados em filtragem consistem no algoritmo *Relief* e suas variações (KIRA; RENDELL, 1992; URBANOWICZ et al., 2017). O algoritmo *Relief* foi inspirado pelo aprendizado baseado em instância. Essa abordagem tem por objetivo calcular a estatística das características, que consiste na “qualidade” ou “relevância”. A estatística de uma característica é referida como $W[A]$, peso da característica A ou como *score*, que pode variar de -1 (pior) a +1 (melhor). O algoritmo *Relief* original era limitado para problemas de classificação binários, além de não ser capaz de lidar com dados faltantes (URBANOWICZ et al., 2017).

No pseudocódigo ilustrado no Algoritmo 1, o método *Relief* avança por m instâncias aleatórias de treinamento (R_i), selecionadas sem substituições, em que m é definido pelo usuário. A cada ciclo, R_i é a instância “alvo” e o vetor de *score* de características W é atualizado, baseado no valor das diferenças das características das instâncias alvo e da vizinhança. Assim, a cada ciclo, a distância entre a instância “alvo” e todas as outras instâncias é calculado. A abordagem *Relief* identifica as duas instâncias vizinhas mais próximas do alvo; uma da mesma classe, denominada *nearest hit* (H) e outra da classe oposta, denominada *nearest miss* (M). A última etapa do ciclo atualiza o peso de uma característica A em W , caso o valor da característica se diferencie da instância alvo R_i e também de *nearest hit* (H) ou *nearest miss* (M). A lógica para atualizar os pesos consiste no fato que um bom atributo deve ter o mesmo valor para instâncias a partir da mesma classe (subtraindo a diferença $\text{diff}(A, R_i, H)$) e deve diferenciar entre as instâncias de diferentes classes (adicionando a diferença $\text{diff}(A, R_i, M)$) (KIRA; RENDELL, 1992).

2.6 Métodos para Classificação de Padrões

O reconhecimento de padrões refere-se a qualquer aspecto que possa ser padronizado em uma estrutura armazenável, como uma base de dados de características. A importância do reconhecimento de padrões se dá pela busca de métodos informatizados e automatizados para tarefas humanas repetitivas e altamente sujeitas a falhas. Além do processamento de imagens, o reconhecimento de padrões é utilizado nas áreas de inteligência artificial, aprendizado de máquina, mineração de dados, teoria dos autômatos, entre outros (CHEN, 2015).

A classificação ou reconhecimento é denominado como o processo de atribuição

Algoritmo 1 Proposta original do método de seleção *Relief*

```

1:  $n \leftarrow$  número de instâncias de treinamento
2:  $a \leftarrow$  número de características
3: Parâmetro:  $m \leftarrow$  número de instâncias de treinamento aleatórias de  $n$  utilizadas
   para atualizar  $W$ 
4: Inicializar todos os pesos  $W[A] = 0,0$ 
5: para  $i = 1$  até  $m$  faça
6:   Aleatoriamente selecionar uma instância “alvo”  $R_i$ 
7:   Encontra um nearest hit “H” e nearest miss “M” (instâncias)
8:   para  $A=1$  até  $a$  faça
9:      $W[A] = W[A] - \text{diff}(A, R_i, H)/m + \text{diff}(A, R_i, M)/m$ 
10:  fim para
11: fim para
12: retorna o vetor  $W$  de scores que estimam a qualidade das características

```

de um rótulo a um objeto, tomando por base as características extraídas do mesmo, representadas por seus descritores. As técnicas de classificação devem ter a capacidade de diferenciar os grupos analisados. Para isso, a eficiência da classificação é observada quando uma imagem com n elementos, cada um pertencendo a n classes diferentes, sendo identificadas corretamente (WEISS; KULIKOWSKI, 1990). Duas abordagens básicas são utilizadas para as técnicas de classificação: supervisionada e não supervisionada.

Na classificação supervisionada, um conjunto de objetos conhecidos e representativos de diferentes classes é analisado por meio de funções discriminantes. Segundo CONCI, AZEVEDO e LETA (2008), o processo de classificação supervisionada é dividido nas seguintes etapas: (1) escolha de um conjunto de treinamento, observando se tal conjunto representa corretamente a população; (2) escolha dos parâmetros utilizados na discriminação, buscando-se evitar a redundância de informação, além de verificar a existência de correlação entre os parâmetros escolhidos e as classes do conjunto de treinamento; (3) obtenção de uma função discriminante a partir de método estatístico ou não-estatístico; (4) eliminação de parâmetros irrelevantes; (5) realização de testes com objetos que não pertençam ao conjunto de treinamento; e (6) atualização dos dados ou realimentação.

A classificação não-supervisionada é utilizada para contextos em que não há informações sobre os rótulos ou classes. Dessa forma, tenta-se estabelecer *clusters* ou agrupamentos naturais no espaço de características, a partir da medida de similaridade dos dados. Para tanto, é analisado a semelhança de um determinado dado com os outros *clusters* com base em distâncias, tais como euclidiana, probabilística, por exemplo (MATHWORKS, 2017). Um *cluster* pode ser definido como um conjunto de

pontos que possui densidade local máxima, quando comparada à densidade de pontos ao redor (JAIN; MURTY; FLYNN, 1999).

Uma breve descrição a respeito dos algoritmos supervisionados de classificação, como árvore de decisão (*decision tree-DT*), máquina de vetores de suporte (*support vector machine*), *random forest*, *naive Bayes* (NB), K-estrela (K*) e polinomial (PL), considerados nesse trabalho, são apresentados nas subseções 2.6.1 a 2.6.6.

2.6.1 Árvore de Decisão

A árvore de decisão (*decision tree-DT*) consiste em um método estatístico baseado em uma estrutura de árvore, composto por vértices e arestas. Tal estrutura é estabelecida a partir do ganho de informação, em que há um critério de separação, consistindo de uma característica que diferencia o grupo analisado em subgrupos. O processo de separação é testado em cada subconjunto de forma recursiva, constituindo uma estrutura hierárquica. As decisões são definidas como condições nos nós internos da árvore e a saída consiste nos conteúdos dos nós folha (SNOUSY et al., 2011).

2.6.2 Máquina de vetores de suporte

O algoritmo de classificação máquina de vetores de suporte (*support vector machine - SVM*) foi proposto por VAPNIK (1999). O modelo de máquina de vetores de suporte considera a representação das amostras para o treinamento como um conjunto de pontos no espaço, em que essas amostras são mapeadas para separação dos rótulos previamente atribuídos. A separação é realizada a partir de um hiperplano que separa totalmente os dados de cada rótulo. O conjunto de teste é mapeado no mesmo espaço e a predição do grupo correto consiste na análise de qual localização espacial do hiperplano esses elementos se enquadram. O mapeamento utilizado no treinamento e teste dos elementos é realizado a partir de uma função *kernel*, que calcula o produto escalar para definir os hiperplanos adequados para separação dos conjuntos analisados.

2.6.3 Random Forest

Na abordagem de floresta aleatória (*random forest-RaF*) é realizada a combinação de várias árvores de decisão para fazer as predições. O objetivo dessa abordagem é melhorar o processo de aprendizagem e consequentemente o resultado geral.

O funcionamento do algoritmo RaF consiste em selecionar aleatoriamente k elementos do conjunto total de treinamento. Na etapa seguinte, esses k elementos são utilizados para encontrar o nó raiz, que consiste no elemento que melhor separa os

grupos considerados. Na etapa seguinte novos elementos são acrescentados em cada nó de separação da árvore, até ser obtido o nó folha, que consiste no resultado da classificação para aquela árvore específica. Esse processo é repetido, criando aleatoriamente n árvores de decisão (BREIMAN, 2001).

Na etapa de teste, os conjuntos são submetidos às regras criadas para cada árvore de decisão, obtendo-se resultados para cada árvore. É realizada uma análise de todas as saídas obtidas pelas árvores de decisão e o resultado final consiste na saída que foi predita por mais árvores.

2.6.4 Naive Bayes

As redes *naive Bayes* (NB) são pertencentes à família de classificadores probabilísticos. Essa categoria é denominada *naive* (ingênuo) devido à simplificação de duas hipóteses, é assumido que os atributos preditivos são condicionalmente independentes de uma dada classe e, assume-se que nenhum atributo escondido ou latente influencie no processo de predição (JOHN; LANGLEY, 1995).

As hipóteses simplificadas possibilitam um algoritmo eficiente para classificação e aprendizado. Uma classe de interesse C é considerada como uma variável aleatória e X é considerado como um vetor de variáveis aleatórias, indicando os valores nos atributos observados. Assim, no método NB, c representa um rótulo particular da classe e x representa um valor de atributo particular do vetor. A regra de Bayes é utilizada para calcular a probabilidade de ocorrência de cada classe a partir do vetor dos valores observados, que são considerados os atributos preditivos, Eq. 2.25.

$$p(C = c|X = x) = \frac{p(C = c)p(X = x|C = c)}{p(X = x)}. \quad (2.25)$$

2.6.5 K-estrela

O algoritmo k-estrela (K^*) realiza a classificação baseada em instância. Assim, a classe de teste baseia-se nas instâncias de treinamento similares à ela, determinadas por uma função de similaridade (CLEARY; TRIGG, 1995). Diferente de outros algoritmos baseados em instância, o algoritmo K^* utiliza uma função de distância de entropia, que pode ser obtida a partir da soma das distâncias dos atributos analisados.

A complexidade da transformação de uma instância em outra é dada pela distância entre as instâncias, obtida em duas etapas: (1) determinar um conjunto finito de transformações que vão mapear uma instância em outra e, (2) transformar uma instância a para b (MAHMOOD; HUSSEIN, 2013).

2.6.6 Polinomial

A abordagem polinomial (PL) se assemelha à função *kernel* do algoritmo SVM, uma vez que são realizados ajustes nos coeficientes e o grau de um polinômio, de modo a separar as classes analisadas. Para isso, os vetores de características são fornecidos como entradas, definidos por $\mathbf{x} = [x_1 \dots x_d]^T$ e são expandidos em termos da base polinomial do vetor $\mathbf{pol}_n(\mathbf{x})$. Tal processo possibilita o mapeamento de um vetor de características d –dimensional em um vetor L –dimensional. Dado um vetor de entrada bi-dimensional $\mathbf{x} = [x_1, x_2]^T$, o vetor $\mathbf{pol}_2(\mathbf{x})$ resulta em parâmetros similares ao mostrado em CAMPBELL et al. (2002):

$$\mathbf{pol}_2(\mathbf{x}) = [1, x_1, x_2, x_1^2, x_1x_2, x_2^2]^T. \quad (2.26)$$

Assim, a dimensionalidade do vetor expandido $\mathbf{p}_n(\mathbf{x})$ pode ser expresso em termos da ordem polinomial e da dimensionalidade do vetor de entrada $\mathbf{x}$ (CAMPBELL; ASSALEH; BROUN, 2002). A saída y_i é obtida pela combinação linear do vetor expandido $\mathbf{pol}_n(\mathbf{x})$ como:

$$y_i = \mathbf{a}_i^T \mathbf{P}_n(\mathbf{x}), \quad (2.27)$$

em que $\mathbf{a}_i$ é um vetor de coeficientes da função polinomial discriminante para a classe ω_i . Este modelo pode ser dividido nas etapas de treinamento e teste, conforme apresentado em Nascimento et al. (NASCIMENTO et al., 2013).

2.7 Medidas de Desempenho para Classificação

A curva *Receiver Operating Characteristic* (ROC) é um método que permite analisar a capacidade de discriminação dos classificadores (BRADLEY, 1997). Essa métrica consiste em um gráfico que apresenta os valores de sensibilidade no eixo y e no eixo x a especificidade. Essas métricas são calculadas a partir da taxas de: verdadeiros positivos (VP), valores positivos que o sistema classificou corretamente como positivos; verdadeiros negativos (VN), valores negativos que o sistema classificou corretamente como negativos; falsos positivos (FP), valores negativos que o sistema classificou incorretamente como positivos; e falsos negativos (FN), valores positivos que o sistema classificou incorretamente como negativos. A partir dessas taxas é possível obter as medidas de sensibilidade, especificade e acurácia para um teste.

A sensibilidade (S) consiste na proporção de verdadeiros positivos que foram corretamente identificados. O cálculo é realizado utilizando a taxa de verdadeiros posi-

tivos VP e o número total de casos positivos, Eq. 2.28. A especificidade (E) define a proporção de verdadeiros negativos corretamente identificados, Eq. 2.29. Para se obter a especificidade pode-se calcular o inverso de S , uma vez que S e E são complementares. A acurácia consiste em uma das métricas mais utilizadas para avaliação de dados na literatura, uma vez que seu cálculo é simples e a interpretação dos dados é intuitiva. O cálculo da acurácia (AC) baseia-se na proporção de predições realizadas corretamente pelo número total de amostras, Eq. 2.30.

$$S = \frac{VP}{VP + FN}. \quad (2.28)$$

$$E = \frac{VN}{VN + FP}. \quad (2.29)$$

$$AC = \frac{VP + VN}{VP + VN + FP + FN}. \quad (2.30)$$

A área sob a curva ROC, *area under the receiver operating characteristic curve* (AUC) consiste em outra métrica para avaliar o desempenho de um classificador. Testes com taxas de AUC próximas de 1, indicam resultados ideais e valores próximos de 0,5 indicam um poder discriminativo baixo. A AUC é comumente calculada por meio do método dos trapézios (MARTINEZ; LOUZADA-NETO; PEREIRA, 2003).

Capítulo 3

Levantamento Bibliográfico

Neste capítulo é apresentada uma seleção de artigos que aplicaram técnicas fractais, curvelet, Haralick em diferentes contextos de imagens. Além disso, foram apresentadas abordagens que propuseram a extração de características a partir de imagens colorretais. As contribuições envolvendo as técnicas fractais estão na seção 3.1, transformada curvelet na seção 3.2, descritores de Haralick na seção 3.3, combinações de técnicas para extração de características encontram-se na seção 3.4. Na seção 3.5 foi apresentada uma seleção dos trabalhos relacionados à extração de características de imagens cólon. Algumas considerações sobre este capítulo encontram-se na seção 3.6.

3.1 Técnicas Fractais

As técnicas fractais são aplicadas para extrair características de textura, uma vez que a principal motivação é explorar propriedades de autossimilaridade dos tecidos. Essas propriedades são encontradas em imagens médicas histológicas. SILVEIRA et al. (2009) propuseram a utilização da lacunaridade para identificar doenças mamárias, com quantificações realizadas a partir de imagens térmicas. Segundo os autores, a motivação para extrair características desse tipo de imagem se deu pelo fato de que o processo de aquisição das imagens térmicas não é invasiva. A lacunaridade foi obtida com a abordagem de deslizamento de caixas e aplicada sobre os grupos de imagens normal e com doença. SILVEIRA et al. (2009) observaram que o método foi eficaz para separar os grupos estudados a partir das diferenças de temperatura das regiões afetadas pela doença.

IVANOVICI, RICHARD e DECEAN (2009) propuseram o uso da dimensão fractal e a lacunaridade multiescala e multidimensional, em que foram consideradas in-

formações sobre os canais de cores para análise dos diferentes níveis de severidade das lesões psoriáticas. Os descritores fractais foram obtidos com base na abordagem probabilística de MANDELBROT (1983), porém estendidos para análise de imagens coloridas. A abordagem colorida foi escolhida pela necessidade de se analisar diferentes aspectos da lesão, como eritema e espessura que determinam variações aleatórias no formato e tamanho destas. Os autores concluíram que quanto maior o valor da dimensão fractal, maior é a severidade da lesão analisada. O mesmo foi observado para a lacunaridade.

LAMANI, MANJUNATH e RAMEGOWDA (2014) aplicaram a dimensão fractal para o diagnóstico de glaucoma. As principais motivações para extração de características e estudos do glaucoma foram pela complexidade do diagnóstico, pelo protocolo executado manualmente e imprecisão. Além disso, a visão pode ser comprometida rapidamente caso o diagnóstico seja equivocado. A dimensão fractal foi calculada para imagens do aro neuroretinal. A abordagem de contagem de caixas foi utilizada em conjunto com uma etapa de pré-processamento, para converter as imagens para o padrão binário. LAMANI, MANJUNATH e RAMEGOWDA (2014) observaram que a dimensão fractal era menor para os casos de glaucoma, devido à perda do nervo óptico e vasos sanguíneos, resultando em imagens mais regulares e com menos complexidade. Os intervalos de dimensão fractal de 1,8743 a 2,3027 foram definidos para os casos sem a presença da doença e intervalos de 1,4242 a 1,5986 para a presença de glaucoma. Com base nos resultados obtidos, os autores determinaram que o diagnóstico antecipado pode ser realizado, quando a dimensão fractal atinge o limiar de 1,6.

A lacunaridade foi utilizada em sua abordagem multiescala e multidimensional no método proposto por NEVES et al. (2014) para quantificar o comportamento do câncer de próstata. A etapa inicial do método consistiu na segmentação das regiões de interesse: células epiteliais, lúmen e estroma. A lacunaridade foi aplicada para imagens dos grupos normal, hiperplasia e câncer considerando o método de deslizamento de caixas. As acurácias obtidas foram de 90,6% na diferenciação dos grupos câncer versus normal; 89,1% dos grupos câncer versus hiperplásico e 85,9% para os grupos câncer versus normal. Os autores concluíram que a classe de imagens que forneceu a melhor discriminação foi o lúmen. Além disso, foi observado que lacunaridades menores foram observadas em tecido normal. NEVES et al. (2014) estabeleceram um modelo de regras para avaliação das imagens prostáticas por meio da classificação por árvore de decisão, possibilitando o auxílio à patologistas no diagnóstico da doença.

FARAHIBOZORG, HASHEMI-GOLPAYEGANI e J. ASHBURNER (2015) rea-

lizaram um estudo com o objetivo de analisar a variação da matéria branca cerebral relacionado com idade e sexo a partir da dimensão fractal. Imagens de ressonância magnética do cérebro foram utilizadas com a técnica contagem de caixas para cumprir a etapa de quantificação. Os resultados foram relevantes, em que foi observado que indivíduos do sexo masculino apresentaram maior dimensão fractal na maioria das regiões analisadas. Apenas áreas do hemisfério esquerdo e fronteiras, obtidas de indivíduos do sexo feminino, apresentaram dimensão fractal superior. Com isso, os autores concluíram que a dimensão fractal foi um excelente descritor para analisar a matéria branca cerebral.

3.2 Transformada Curvelet

GUPTA e TIWARI (2014) propuseram o uso da transformada curvelet para o diagnóstico de câncer de pulmão utilizando imagens de tomografia computadorizada para a extração das características. Os autores argumentaram que esse tipo de imagem proporcionou um volume maior de dados em menor tempo, quando comparado a outros métodos de aquisição de imagens, como raios X ou ressonância magnética. As imagens foram submetidas a uma etapa de pré-processamento para melhoria do contraste, por meio da equalização de histograma. A etapa de quantificação consistiu na abordagem de envelopamento para extração das características curvelet. GUPTA e TIWARI (2014) propuseram descritores adicionais com o cálculo da média e desvio padrão do vetor de características curvelet. A última etapa dessa abordagem consistiu na classificação dos grupos câncer versus normal utilizando redes neurais. Os autores testaram o método em um conjunto de 120 imagens (60 diagnosticadas com câncer e 60 do grupo normal), obtendo acurácia de 90% na diferenciação dos grupos.

A transformada curvelet foi aplicada por FRANCIS, SASILAKA e SARANYA (2014) para a análise de anomalias mamárias. Imagens termográficas foram utilizadas para a extração de características. Este tipo de imagem foi considerado uma alternativa à mamografia, representando a variação de temperatura em formas de intensidades. Os autores justificaram o uso da transformada curvelet para imagens termográficas pelo fato das fontes de calor das imagens térmicas possuírem aspecto de círculos concêntricos, havendo a necessidade de extrair características curvas e de bordas. O método foi testado em um conjunto de 22 imagens (11 casos normais e 11 anormais), obtidas de pacientes com média de 42 ± 8 anos. A etapa inicial do método consistiu na segmentação das regiões de interesse, garantindo maior precisão no processo de extração de características. Após a separação das regiões de interesse, a transformada curvelet foi

aplicada utilizando 4 escalas e 8 ângulos, produzindo 9 subimagens. Características estatísticas foram extraídas a partir das subimagens, a saber: média, mediana, moda, desvio padrão e variância. A última etapa consistiu na classificação dos grupos a partir do classificador máquina de vetores de suporte e a acurácia obtida no sistema foi de 90,91%.

LIN et al. (2015) propuseram um método para extração de características utilizando a transformada curvelet para o diagnóstico do câncer prostático. Os descritores obtidos foram classificados com base no sistema de pontuação de Gleason (GLEASON, 1977), amplamente utilizado pelos especialistas em diagnóstico de doenças de tecidos microscópicos por meio de biópsia. Os autores utilizaram imagens previamente classificadas com casos de alteração no tecido. A transformada curvelet foi aplicada com a abordagem de envelopamento com parâmetros de quantificação de 4 escalas e 8 ângulos. Foram obtidas 9 características a partir da extração características estatísticas das subimagens curvelet, considerando a média, variância, entropia e energia. A classificação dos grupos foi realizada com máquina de vetores de suporte e a acurácia geral obtida foi de 93,68%.

SARASWATHI, DHARANI e SRINIVASAN (2016) aplicaram a transformada curvelet para extração de características em imagens mamográficas para o diagnóstico do câncer de mama. O método foi dividido nas etapas: pré-processamento das imagens; extração de características utilizando a transformada curvelet; redução do vetor de características baseada na otimização por enxame de partículas; classificação das imagens por meio do classificador máquina de vetores de suporte. Os autores utilizaram 56 imagens normais, 70 anormais, 37 benignas e 19 malignas. Na primeira etapa as regiões de interesse foram extraídas, o processo consistiu na remoção do fundo e rótulos presentes nas imagens mamográficas. Na segunda etapa, a transformada curvelet discreta foi aplicada com 16 ângulos e um número de escalas determinado por $\log_2(\min(M,N)) - 3$, sendo M e N a largura e altura da imagem, respectivamente. A redução de características foi realizada na etapa 3, com base na otimização por enxame de partículas. Nesse método, um indivíduo no enxame, denominado de partícula, consistiu em uma solução em potencial, com os parâmetros de aptidão e velocidade. Cada partícula teve a função de aprender o envolvimento do enxame para avaliar o ótimo global. A classificação foi realizada na etapa 4, por meio do vetor de características reduzido e máquina de vetores de suporte. A transformada curvelet associada ao método de seleção de características otimização por enxame de partículas resultou em acurácia de 96% na classificação dos grupos analisados.

3.3 Descritores de Haralick

SUNGANYA e RAJARAM (2013) utilizaram os descritores de Haralick para analisar doenças do fígado. A obtenção de descritores com maior poder discriminativo para o diagnóstico da doença consistiu no objeto do trabalho. Foram utilizadas 40 imagens de ultrassom para cada contexto: fígado normal, fígado com esteatose hepática, cistos e cirrose. Na etapa de pré-processamento foram removidos ruídos nas imagens. Os autores extraíram 12 características de Haralick, como contraste, correlação, média, segundo momento angular, somatório da variância e somatório da entropia. A classificação consistiu na última etapa do sistema e, para essa tarefa, foi aplicado o classificador máquina de vetores de suporte. Os autores observaram que cinco descritores apresentaram maior discriminação na análise dos grupos: contraste, correlação, segundo momento angular, somatório da variância e somatório da entropia. A acurácia obtida foi de 81,7%.

Os descritores de Haralick foram utilizados em (BOUBCHIR; AL-MAADEED; BOURIDANE, 2014) para detecção de atividades epiléticas. As imagens utilizadas foram extraídas a partir de eletroencefalogramas. A motivação dos autores consistiu no fato da detecção manual de atividades epiléticas, via eletroencefalogramas, pode ser um processo demorado e totalmente dependente da experiência do especialista. A análise dos autores foi realizada com base em imagens representativas de tempo-frequência dos sinais de eletroencefalograma. Foram utilizadas 4096 amostras obtidas de pacientes normais e diagnosticados com epilepsia. Os descritores de Haralick foram aplicados para quantificar os padrões normal e de atividades epiléticas. A classificação dos grupos normal versus epilepsia foi realizada a partir do classificador máquina de vetores de suporte. A acurácia obtida foi superior a 98% para todos os casos analisados.

SAMANTA et al. (2014) propuseram a utilização dos descritores de Haralick para diagnóstico de glaucoma, analisando os casos com a presença e ausência de glaucoma. Foram utilizadas imagens do fundo do olho, também denominadas *fundus*. A quantificação foi realizada em 455 imagens, 200 afetadas por glaucoma e 255 de pacientes saudáveis. Os descritores de Haralick foram extraídos em conjunto com os valores médios correspondentes, totalizando 28 descritores. Para a etapa de classificação e análise dos grupos, os autores utilizaram um modelo de redes neurais de *backpropagation*. A acurácia obtida no sistema para detecção de glaucoma foi de 96%. Uma das maiores vantagens do método proposto consistiu em extrair características sem a necessidade de uma etapa de pré-processamento das imagens.

CHAKRABORTY et al. (2016) propuseram a extração de características em ima-

gens térmicas infravermelhas digitais para detecção de câncer de boca. O conjunto de imagens infravermelhas analisado pelos autores continha 81 casos malignos, 59 casos pré-cancerosos e 63 casos normais, extraídos de um grupo com idade entre 23 e 65 anos. A etapa 1 compreendeu o pré-processamento da imagem: extrair a região de interesse da imagem, realizar a transformação das imagens e relacionar as intensidades das temperaturas aos níveis de cinza. A segunda etapa do trabalho consistiu na extração de características. Os descritores de Haralick foram extraídos a partir da matriz de co-ocorrência. A seleção de características foi realizada na etapa 3, a partir de uma função lógica. Para tanto, foram selecionadas apenas as características que apresentaram diferenças significativas entre os grupos normais e com alterações. A etapa 4 consistiu na classificação dos grupos normal versus maligno e normal versus pré-canceroso. Foram aplicados os classificadores *k-means* e *fuzzy k-means*, os quais forneceram acurácias superiores a 85%. Os autores destacaram o poder discriminativo presente em imagens térmicas para o diagnóstico do câncer de boca, utilizando apenas características de texturas.

3.4 Combinação de Técnicas de Extração de Características

A combinação de técnicas fractais, Haralick e curvelet ainda são pouco exploradas na literatura. Os resultados obtidos em trabalhos que combinaram esses descritores possibilitaram resultados relevantes (EID et al., 2013; LIU et al., 2016; CHARISIS; HADJILEONTIADIS, 2016). Além disso, a combinação de técnicas permite a extração de características a partir de diferentes abordagens, possibilitando analisar aspectos de autossimilaridade, contraste, singularidades curvas e distribuição dos *pixels*.

EID et al. (2013) propuseram o diagnóstico de úlceras utilizando imagens de cápsula endoscópica. Esse processo de aquisição de imagem é considerado uma técnica não invasiva para inspeção do trato digestivo de pacientes. Os autores quantificaram grupos saudável e com úlcera, utilizando a transformada curvelet discreta, associada com a lacunaridade para obtenção de características. A primeira etapa do método consistiu na decomposição dos canais de cores e conversão do padrão RGB para o YCbCr. Segundo EID et al. (2013), o padrão YCbCr possibilita desempenho superior na classificação de imagens endoscópicas. A obtenção da transformada discreta curvelet para os canais de cores escolhidos consistiu na segunda etapa. Os parâmetros de escala e ângulo foram determinados por meio de testes exaustivos. A melhor combinação consistiu na configuração de quatro escalas e oito ângulos para decomposição, totali-

zando 41 subimagens. Na terceira etapa, a quantificação da lacunaridade foi aplicada em cada subimagem obtida. Um vetor de características de alta dimensionalidade foi o resultado dessa etapa. De modo a tratar esse problema, EID et al. (2013) modelaram uma função baseada no problema dos mínimos quadrados. Características adicionais foram obtidas por meio de medidas estatísticas, como média, desvio padrão, entropia, energia, assimetria e achatamento. A classificação foi realizada utilizando o classificador máquina de vetores de suporte e validação cruzada. Os experimentos foram realizados em uma base de 130 imagens normais e 130 com úlcera. O desempenho geral do sistema na abordagem proposta foi de 86,5% de acurácia na classificação dos grupos.

LIU et al. (2016) utilizaram a transformada curvelet associada à dimensão fractal para extração de características de imagens de cápsula endoscópica sem fio, contexto de aplicação similar ao trabalho proposto por (EID et al., 2013). Na primeira etapa do método, o pré-processamento das imagens foi realizado para compor imagens sintetizadas a partir das informações mais relevantes da textura. Os resultados foram duas subimagens geradas para cada imagem de entrada, além disso, imagens sintetizadas foram obtidas para cada canal de cor, totalizando 6 imagens sintéticas para cada imagem de cápsula endoscópica. O método de envelopamento foi aplicado com quatro escalas e oito ângulos. Segundo LIU et al. (2016), as informações mais relevantes acerca de tumores foram observadas na segunda e terceira escalas. A quantificação da dimensão fractal nas imagens sintetizadas foi realizada na segunda etapa. Para reduzir o número de características, os autores aplicaram seis descritores estatísticos de texturas: desvio padrão, energia, assimetria, momento da diferença inversa, correlação e entropia. Os resultados obtidos foram de 126 momentos de segunda ordem e 84 momentos de ordem elevada, constituindo o vetor de características. Na terceira etapa, a classificação e seleção de características foram realizadas com um Algoritmo Genético associado ao classificador máquina de vetores de suporte. O método foi aplicado em 900 imagens normais e 900 com tumores típicos, previamente diagnosticados por especialistas. A acurácia obtida foi de 97,3% na identificação dos tumores.

CHARISIS e HADJILEONTIADIS (2016) propuseram uma nova técnica para extrair características para detecção de lesões inflamatórias das mucosas gastrointestinais causadas pela doença de Crohn. O método explorou imagens de cápsula endoscópica sem fio, nomeada filtragem adaptativa híbrida (*Híbrid Adaptive Filtering*). O método foi testado em uma base de imagens com 800 casos. Na etapa 1, as imagens contendo a região de interesse foram pré-processadas, convertendo-as do padrão RGB para o espaço YCbCr. A conversão foi realizada para solucionar o problema de alta correlação

entre os canais RGB, que poderia degradar o processo de quantificação. A segunda etapa consistiu na aplicação da transformada curvelet discreta. A quantificação foi realizada com 4 escalas e 8 ângulos. Essa configuração foi determinada de forma experimental, via testes exaustivos. A quantificação obtida resultou em um número elevado de características e a abordagem de filtragem adaptativa híbrida foi aplicada para selecionar as características mais relevantes. Para tanto, os autores optaram por duas abordagens com função de aptidão (*fitness function*) no Algoritmo Genético, a saber: função de aptidão baseada em energia; e, função de aptidão baseado na curva gradiente da lacunaridade. A filtragem baseada em energia foi aplicada para selecionar as subimagens com baixa energia associada. A filtragem baseada em lacunaridade foi utilizada para capturar as variações das características estruturais e texturais nas imagens endoscópicas. Na terceira etapa, o vetor de características foi estabelecido com base na quantificação da lacunaridade nas subimagens. A lacunaridade foi escolhida devido à sua tolerância a luminosidade não uniforme e invariância em diferentes rotações e translações. Um total de cinco vetores de características foi estabelecido, realizando diferentes combinações. Na etapa 4 foi realizada a classificação dos grupos moderado e severo. O classificador escolhido foi máquina de vetores de suporte, considerando os vetores de características calculados para os diferentes canais de cores. Os melhores resultados obtidos na classificação consistiram na análise dos casos severos utilizando a filtragem baseada em lacunaridade, com acurácia superior a 90%.

SARASWATHI e SRINIVASAN (2017) propuseram a associação da transformada curvelet com a dimensão fractal em níveis de cinza. O método foi aplicado para detectar microcalcificações em imagens mamográficas. Os autores utilizaram a base de imagens mamográficas MIAS, composta por 322 imagens digitalizadas. Na etapa 1, as imagens mamográficas foram pré-processadas a partir da transformada do cosseno discreta, com o objetivo de aguçar os contornos das imagens. A segmentação foi aplicada na etapa 2, aplicando a abordagem de *fuzzy C-means*. Na etapa 3, as imagens segmentadas foram decompostas em subimagens via transformada curvelet por envelopamento. Em seguida, a dimensão fractal em níveis de cinza foi aplicada para cada subimagem gerada. Na etapa 4, o classificador baseado em redes neurais *ensembled FCRN* foi aplicado com validação cruzada de 10 *folds*. O resultado foi uma acurácia média de 98,18%.

3.5 Abordagens para Extração de Características em Imagens Colorretais

Ao considerar as contribuições no contexto da quantificação de imagens histológicas de cólon, ALTUNBAY et al. (2010) propuseram um novo método estrutural. O método construiu um grafo com as várias estruturas presentes nas imagens histológicas, rotulando cores nas arestas representativas dos tipos de estruturas em cada vértice. Após a construção dos grafos coloridos, foram extraídas as características estruturais grau médio, coeficiente médio de agrupamento e diâmetro. Os autores consideraram 213 imagens histológicas de cólon previamente rotuladas como benigno, maligno de baixa graduação e maligno de alta graduação. O classificador máquina de vetores de suporte foi aplicado para diferenciação dos grupos, utilizando validação cruzada com 10-*folds*, obtendo-se a acurácia de 82,65%.

KALKAN et al. (2012) propuseram a obtenção de características em imagens segmentadas de cólon, aplicando características estruturais e de textura. A primeira etapa consistiu na segmentação das imagens em estroma, núcleo e lúmen. Na etapa seguinte, foram extraídas características a partir de subdivisões da imagem original: número de núcleos por área do tecido, razão entre cada estrutura em relação à área do tecido e razão entre as estruturas nas subdivisões. A média e a variância das características extraídas foram calculadas para obtenção de descritores globais. Na terceira etapa, as características de textura foram obtidas: distribuição das cores, descritores de Haralick e filtro de Gabor. Ao final desta etapa foram extraídas 1127 características para cada imagem. A quarta etapa consistiu na seleção das características baseada em um método de distância. As 120 melhores características foram utilizadas no processo de classificação. A última etapa do método proposto consistiu na classificação dos grupos normal versus anormal e câncer versus não-câncer. O classificador de regressão logística foi utilizado considerando a validação cruzada com 10-*folds*. O melhor resultado na classificação consistiu na comparação dos grupos câncer e não-câncer, atingindo a acurácia de 87,67%.

AKBAR, GOPI e BABU (2015) propuseram a detecção do câncer de cólon utilizando características estruturais e estatísticas. As características estruturais foram definidas com base na representação de grafos, em que as estruturas histológicas consistem nos vértices do grafo e as arestas consistem na relação espacial entre cada estrutura. As relações espaciais foram codificadas ao construir o grafo do tecido utilizando a triangulação de *Delaunay*. As características foram extraídas a partir de regiões de maior relevância, mensurando os níveis de similaridade com os subgrafos obtidos. Os des-

critores de Haralick foram utilizados para extrair características estatísticas nas regiões de maior relevância. A última etapa do método consistiu na classificação dos grupos normal versus câncer, com um total de 113 imagens. Os autores utilizaram os classificadores *Multi Layer Perceptron* (MLP), máquina de vetores de suporte, K-estrela e regressão logística bayesiana. O melhor resultado foi obtido com o classificador MLP, em que foi obtida a acurácia de 83,33%.

NAYIAR, ASIM e SHAHID (2015) propuseram a detecção do câncer de cólon utilizando características estruturais e morfológicas. Na etapa 1 foi realizado o pré-processamento, para aumentar o contraste das imagens e a realizar a separação do lúmen e células epiteliais, utilizando o algoritmo *K-means clustering*. A etapa 2 do método consistiu na extração das características estruturais e morfológicas. O formato da estrutura do lúmen foi obtido utilizando a distância entre os mínimos quadrados. As características estruturais foram extraídas a partir da quantificação dos descritores de Fourier elípticos. As características morfológicas foram: perímetro, área, excentricidade, número de Euler, área convexa, compacidade, orientação, tamanho do menor e maior eixo extraído do lúmen. Na última etapa, a classificação foi realizada a partir do classificador SVM com validação cruzada de 10-*folds*, aplicado em 174 imagens, com 82 casos normais e 92 casos malignos. O resultado foi uma acurácia de 96,55%.

No trabalho de IFTIKHAR, HASSAN e ALQUHAYZ (2016) um sistema de predição de câncer de cólon foi proposto extraindo características de textura e momentos estatísticos de 92 imagens histológicas de cólon, divididas nos grupos saudável, moderado e câncer. Na primeira etapa, o pré-processamento das imagens foi realizado a partir da melhoria no contraste e conversão para níveis de cinza. A etapa seguinte consistiu na extração de características, aplicando os descritores de Haralick e características estatísticas: média, desvio padrão, aspereza e achatamento. Na terceira etapa foi realizada a seleção das características utilizando a técnica *mRMR*, em que a seleção é realizada a partir da minimização da redundância entre as características. A classificação consistiu na etapa final do método, em que o classificador máquina de vetores de suporte foi utilizado com validação cruzada *leave-one-out*. A acurácia obtida para o conjunto de dados completo foi de 95,4%.

No trabalho de KATHER et al. (2016) foi proposto um sistema para detectar os casos de câncer *versus* não câncer e para diferenciar 8 subclasses do câncer colorretal: (1) tumor, (2) estroma simples, (3) estroma complexo, (4) células imunes, (5) detritos, (6) glândula normal mucosal, (7) tecido adiposo e (8) segundo plano (sem tecidos), utilizando 5000 imagens histológicas. As abordagens para extração de atributos foram características de histograma (*lower-order* e *high-order*), *local binary pattern* (LBP),

matriz de co-ocorrência, filtros de Gabor e características *perceptron-like*. O conjunto de características foi utilizado para classificar os grupos câncer *versus* não câncer, atingindo uma acurácia de 98,6% ao utilizar o classificador máquina de vetores de suporte. Para a classificação entre os oito subgrupos de imagens, a acurácia obtida foi de 87,4%. As características que mais se destacaram no processo de diferenciação foram: *lower-order* e *high-order*, matriz de co-ocorrência e LBP.

WANG et al. (2017) propuseram a aplicação de redes neurais convolucionais bilineares para classificar o câncer colorretal considerando as classes câncer e não-câncer. O método realizou a decomposição dos componentes das imagens H&E. Posteriormente, as redes neurais convolucionais bilineares foram construídas sobre as imagens decompostas para realizar a fusão e melhora da representação das características. A base de imagens utilizada foi a mesma apresentada no trabalho de (KATHER et al., 2016), com 5000 imagens de 8 subclasses distintas. A acurácia na comparação entre os grupos benigno e maligno obtida foi de 92,6%, com o classificador SVM e validação cruzada de 5-*folds*.

JØRGENSEN et al. (2017) propuseram um método para detecção do câncer de cólon a partir de características do núcleo celular. Os autores aplicaram o método em 87 imagens (44 benignas e 43 malignas) previamente rotuladas. Na primeira etapa foi realizada a segmentação das estruturas celulares do núcleo, utilizando os métodos *color deconvolution*, *k-means clustering* e limiarização adaptativa local. Na etapa 2, foram extraídas 750 características: intensidade mínima, média e máxima de cada canal de cor, matrizes de co-ocorrência, matriz *run-length* e estatísticas de primeira ordem. Na etapa 3 a seleção das características foi realizada utilizando o algoritmo *random forest* com a métrica de desempenho importância variável. A classificação dos grupos benigno contra maligno foi realizada com o algoritmo *random forest*, utilizando as 9 características mais discriminativas obtidas na etapa de seleção. O resultado foi uma acurácia de 91%.

3.6 Uma Visão Geral Sobre os Trabalhos Relacionados

Neste capítulo foi apresentada uma visão detalhada sobre os principais trabalhos vinculados ao tema combinação de técnicas para extração de características, bem como trabalhos para extração de características em imagens histológicas de cólon. Nota-se que as técnicas fractais, transformada curvelet e descritores de Haralick forneceram resultados relevantes para o diagnóstico de doenças em diferentes contextos, como imagens de ultrassom (SUGANYA; RAJARAM, 2013), endoscópicas (LIU et al., 2016),

térmicas (CHAKRABORTY et al., 2016) e mamográficas (SARASWATHI; DHARANI; SRINIVASAN, 2016).

As técnicas fractais se mostraram interessantes por quantificar o nível de ocupação e complexidade das imagens (IVANOVICI; RICHARD; DECEAN, 2009). A transformada curvelet permitiu extrair contornos e singularidades curvas (EID et al., 2013). Os descritores de Haralick permitiram indicar características estatísticas na distribuição dos níveis de cinza dos *pixels* nas imagens (SUGANYA; RAJARAM, 2013). A partir disso, uma abordagem interessante consiste em combinar as diferentes técnicas mencionadas.

A combinação de técnicas fractais, Haralick e curvelet ainda foi pouco explorada na literatura. A principal razão para este fato consiste na dificuldade em determinar os melhores parâmetros em cada técnica. Por exemplo, para transformada curvelet a dificuldade está em definir a escala e ângulo adequados, que são cruciais para o sucesso de definição das subimagens. Nos trabalhos pesquisados, foi observada a aplicação de testes exaustivos para determinar a melhor combinação de parâmetros. Nas imagens endoscópicas os parâmetros utilizados foram de quatro escalas e oito ângulos (EID et al., 2013), (LIU et al., 2016), (CHARISIS; HADJILEONTIADIS, 2016).

Apesar dos resultados interessantes e das diferentes abordagens descritas na literatura especializada no tema, a combinação proposta neste trabalho para auxílio no diagnóstico de câncer de cólon ainda não foi explorada. Um resumo dos principais trabalhos focados em imagens H&E colorretais é apresentado na Tabela 3.1

Tabela 3.1: Sumarização dos trabalhos correlatos, com informações sobre métodos de extração de características, classificação e desempenho obtido.

Referência	Extração de características	Classificador	AC
Esgiar et al. (1998)	Matriz de co-ocorrência	<i>K-nearest neighbor</i>	90,2
Esgiar et al. (2002)	Matriz de co-ocorrência e dimensão fractal	<i>K-nearest neighbor</i>	94,1
Masood et al. (2006)	Matriz de co-ocorrência e Características morfológicas	SVM	90
Masood, Rajpoot (2009)	Características LBP circulares	SVM	90,6
Altunbay et al. (2010)	Grafos coloridos	SVM	82,65
Kalkan et al. (2012)	Características de Haralick e Gabor	Regressão Logística	87,67
Akbar, Gopi e Babi (2015)	Características estruturais e Haralick	<i>Multi-Layer Perceptron</i>	83,33
Naiyar, Asim e Shahid (2015)	<i>Least square lumen distance</i> , descritores elípticos de Fourier e características morfológicas	SVM	96,55
Iftikhar, Hassan e Alquhayz (2016)	Haralick e momentos estatísticos	SVM	96,13
Kather et al. (2016)	<i>higher-order e lower-order</i> , LBP, matriz de co-ocorrência, filtros de Gabor e <i>perceptron-like</i>	SVM	98,6
Wang et al. (2017)	Redes neurais convolucionais bilineares	SVM	92,6
Jorgensen et al. (2017)	Características de intensidade, matrizes de co-ocorrência, matriz <i>run-length</i> e características estatísticas de primeira ordem	<i>Random forest</i>	91

Capítulo 4

Metodologia

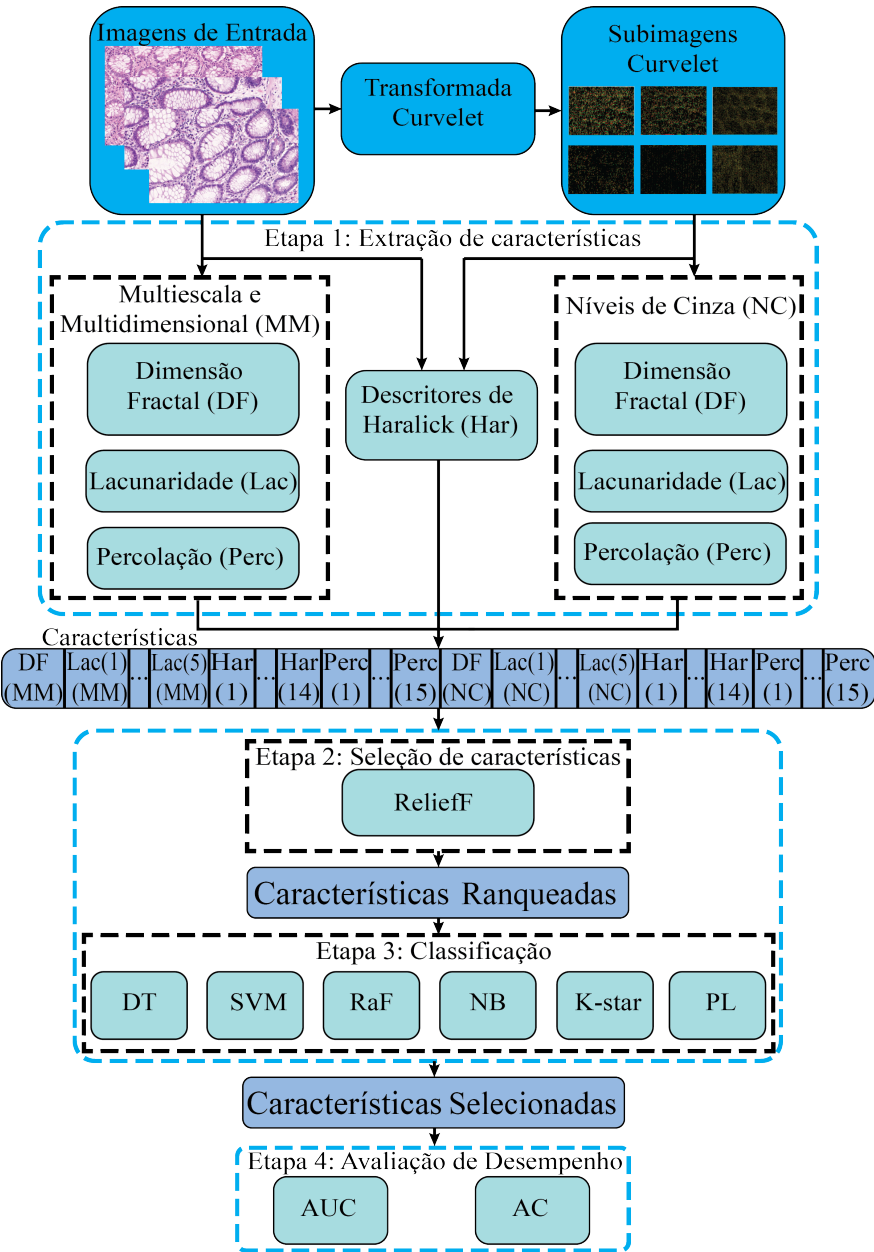
O método proposto com a associação de características multidimensionais e multirresolução foi organizado em etapas. A etapa 1 consistiu na extração das características, obtidas de cada imagem colorida bem como das subimagens curvelet correspondentes. Na etapa 2, uma seleção das características mais relevantes foi realizada para a classificação do câncer colorretal. Na etapa 3, os classificadores DT, SVM, RaF, NB, K* e PL foram aplicados para diferenciar os grupos de interesse. Na etapa 4 foi realizada a avaliação de desempenho, indicando as melhores combinações e resultados. Uma visão geral do método é apresentado na Figura 4.1.

4.1 Transformada Curvelet

A transformada curvelet foi aplicada para gerar sub-imagens a partir de cada imagem histológica H&E. A abordagem utilizada neste trabalho foi a transformada curvelet discreta rápida (*fast discrete curvelet transform*) com a função de envelopamento *wrapping function*, por reunir amostras selecionadas de Fourier e facilitar as representações com diferentes escalas e ângulos (ALZUBI; ISLAM; ABBOD, 2011). Para tanto, a implementação da transformada curvelet foi baseada no invólucro das amostras de Fourier, processo que considera a imagem como uma matriz cartesiana $f[m,n]$, em que $0 \leq m < M$, $0 \leq n < N$, M e N são as dimensões da matriz. As saídas obtidas foram conjuntos de coeficientes curvelet $c^D(j,l,\rho_1,\rho_2)$ indexados por uma escala j , orientação l e parâmetros de localização espacial ρ_1 e ρ_2 , conforme indicado na Eq. 4.1.

$$c^D(j,l,\rho_1,\rho_2) = \sum_{n=1}^N \sum_{m=1}^M f[m,n] \phi_{j,l,\rho_1,\rho_2}^D[m,n]. \quad (4.1)$$

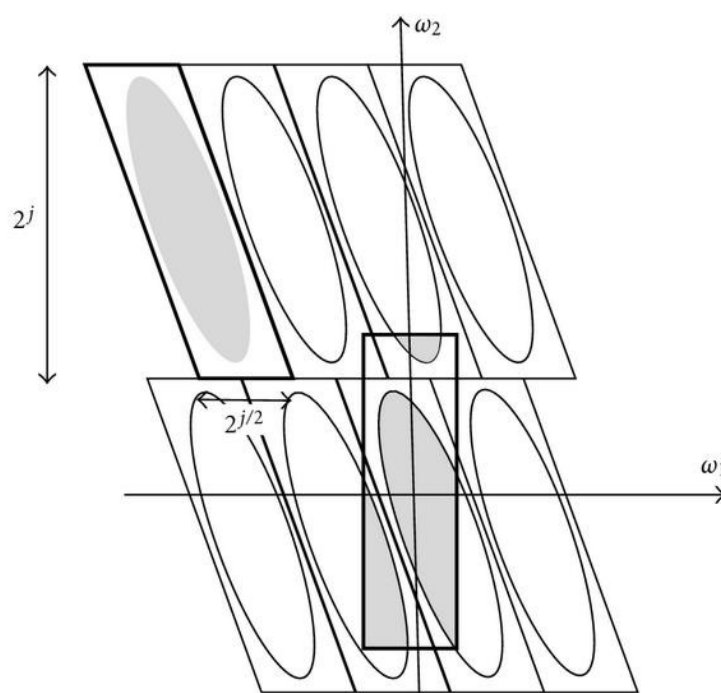
Figura 4.1: Um resumo do método para avaliar imagens histológicas colorretais H&E.



Fonte: elaborado pelo autor.

O conjunto de coeficientes curvelet consistiu em paralelogramos contendo uma curvelet digital em forma de onda (*digital curvelet waveform*), definida como ϕ_{j,l,p_1,p_2}^D . O invólucro baseou-se na coleta de informação a partir dos paralelogramos dentro do retângulo centrado na origem, com largura de $2^{j/2}$ e altura de 2^j , como ilustrado na Figura 4.2.

Figura 4.2: *Wrapping* de uma imagem.



Fonte: (ALZUBI; ISLAM; ABBOD, 2011).

As curvas contidas em cada imagem foram representadas por lacunas ao longo dos níveis de cinza vizinhos. As curvelets foram valores numéricos que representam as curvas em coeficientes. Quanto menor a lacuna, maiores foram os coeficientes. As curvelets foram calculadas com observações realizadas em 4 escalas e 8 rotações. Estes valores forneceram resultados relevantes nos trabalhos de Eid et al. (2013), Liu et al. (2016) e Charisis e Hadjileontiadis (2016). Esta abordagem resultou em 41 sub-imagens curvelet para cada imagem colorretal dada como entrada.

4.2 Aplicação das Técnicas Fractais

A quantificação das imagens com as técnicas fractais foi aplicada após os cálculos das subimagens curvelet. A geometria fractal tem por principal característica o estudo de

formas autossimilares (MANDELBROT; PIGNONI, 1983), sendo a dimensão fractal e lacunaridade os atributos fractais mais conhecidos. Recentemente, a percolação tem sido aplicada com sucesso no contexto de imagens médicas (ROBERTO et al., 2017).

Na literatura existem abordagens de dimensão fractal que são multiescalas e multidimensionais, capazes de quantificar imagens coloridas ou em níveis de cinza. Neste contexto, as principais abordagens da DF são os métodos probabilísticos de (IVANOVICI; RICHARD; DECEAN, 2009) e de fusão de caixas de (NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011). O cálculo da lacunaridade foi também baseado na abordagem probabilística de (IVANOVICI; RICHARD; DECEAN, 2009). A percolação foi baseada na abordagem proposta por (ROBERTO et al., 2017), com três curvas de precolação distintas. Estes métodos foram os considerados para extrair as características das imagens histológicas colorretais coloridas e das sub-imagens correspondentes em níveis de cinza.

4.2.1 Dimensão fractal baseada na abordagem probabilística

O método probabilístico de (IVANOVICI; RICHARD; DECEAN, 2009) foi aplicado considerando uma caixa de lado L que percorreu todos os *pixels* da imagem colorida dada como entrada. As imagens foram representadas no modelo de cores *Red* (R), *Green* (G) e *Blue* (B). Assim, considerando uma caixa de tamanho L posicionada sobre a imagem com *pixel* central definido por F_c , o canal de cor mais relevante foi selecionado por meio de comparações entre F_c com os demais *pixels* F da caixa. O número de *pixels* $F = f(x, y, r, g, b)$ com distância Minkowski (Eq. 4.2) até o *pixel* central $F_c = f_c(x_c, y_c, r_c, g_c, b_c)$ menor ou igual a L foram totalizados e armazenados em uma matriz de probabilidade $P(m, L)$. Esta matriz forneceu a probabilidade de m pontos pertencerem a uma caixa de lado L . Por exemplo, na Tabela 4.1, a relação $P(2, 3)$ corresponde à probabilidade de uma caixa de lado $L = 3$ possuir dois *pixels* que satisfaçam a distância d de Minkowski.

É importante observar que para cada *pixel* verificado neste processo, os valores de x e y indicaram a localização do *pixel* na imagem e os valores de r , g e b corresponderam às intensidades dos canais. Após aplicar as abordagens multiescala e multidimensional, o número total de caixas $N_{DF}(L)$ para cobrir uma imagem foi definido considerando a Eq. 4.3. O valor de $N_{DF}(L)$ é computado para diferentes escalas de observações L . A dimensão fractal probabilística DF_p foi dada pelo coeficiente angular da regressão linear definida por $\log L \times \log N_{DF}(L)$. Os tamanhos das caixas foram de $L = 3$ a $L = 45$, valores que foram suficientes para indicar observações multiescalas das imagens colorretais com a técnica DF_p .

$$|F - F_c| = \max |f(i) - f_c(i_c)| \leq L, \forall i = \overline{1, 5}, \quad (4.2)$$

$$N_{DF}(L) = \sum_{m=1}^N \frac{P(m, L)}{m}. \quad (4.3)$$

Tabela 4.1: Exemplo de representação de uma matriz de probabilidade.

	3	5	...	L_{max}
1	P(1,3)	P(1,5)	...	P(1, L_{max})
2	P(2,3)	P(2,5)	...	P(2, L_{max})
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
L^2	P(L^2 ,3)	P(L^2 ,5)	...	P(L^2 , L_{max})

4.2.2 Dimensão fractal baseada na abordagem de fusão de caixas

A dimensão fractal baseada em fusão de caixas DF_f foi outro método aplicado para analisar as imagens de cólon e respectivas sub-imagens curvelet. A abordagem FD_f descrita por (NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011) complementa as quantificações obtidas com DF_p , pois é uma expansão para espaços n -dimensionais. O valor de DF_f foi determinado a partir de uma tabela de partição definida conforme Eq. 4.4, em que L é uma escala de observação, s indica o tamanho da partição sob análise, x é a coordenada de qualquer pixel na caixa e ϵ indica a relação de L/s . A tabela permite armazenar todas as coordenadas de partições t que tiveram ao menos um elemento do conjunto de dados. Na Tabela 4.2 é ilustrado um exemplo de uma partição t constituída a partir da primeira iteração para uma imagem considerando o modelo de cor *RGB*: cada linha indica o cálculo das partições t_x, t_y, t_r, t_b, t_g para cada pixel na imagem. A cada iteração a quantidade de linhas n_f na tabela de partição é contabilizada. O valor de DF_f foi obtido a partir do coeficiente angular de um gráfico de $\log n_f \times \log s$.

$$t_x = \left\lfloor \frac{x}{\epsilon_x} \right\rfloor = \left\lfloor \frac{xs}{L_x} \right\rfloor \quad (4.4)$$

4.2.3 Percolação multiescala e multidimensional

A percolação (*Perc*) foi aplicada para quantificar as imagens colorretais e as subimagens curvelet correspondentes, seguindo a mesma estratégia adotada para as imagens

Tabela 4.2: Exemplo de tabela de partição de uma imagem de tamanho L_x por L_y

	t_x	t_y	t_r	t_g	t_b
1	0	0	0	120	255
2	0	1	50	43	128
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$L_x \cdot L_y$	3	15	240	31	45

multiescalas e multidimensionais. A definição de percolação consistiu na identificação de um caminho conectado de *pixels* de uma extremidade à outra da imagem. O método aplicado foi descrito detalhadamente em (ROBERTO et al., 2017).

A primeira etapa consistiu em uma análise multiescala com a aplicação do algoritmo *gliding box*. Uma caixa quadrada de lado L foi utilizada para percorrer todos os *pixels* de uma imagem dada como entrada. O tamanho de L foi incrementado após percorrer completamente a imagem. O total de caixas T foi obtido considerando a largura W e altura H da imagem, em função do tamanho de uma caixa L , conforme Eq. 4.5.

$$T = (H - L + 1) \times (W - L + 1). \quad (4.5)$$

Na próxima etapa, a análise multidimensional foi realizada posicionando o *pixel* central F_c de uma caixa de lado L sobre cada *pixel* da imagem analisada, conforme abordagem descrita em (IVANOVICI; RICHARD, 2011; IVANOVICI; RICHARD; DECEAN, 2009). O canal mais relevante foi selecionado para realizar comparações entre o *pixel* central e os demais *pixels* existentes na caixa, aplicando a distância de Minkowski (Eq. 4.2). Este processo foi similar ao aplicado para calcular o atributo DF_p . Em seguida, quando a distância de Minkowski d foi menor ou igual à L , o valor de P foi definido como -1 para indicar que o *pixel* representa um poro na caixa analisada.

Em seguida, a terceira etapa consistiu em associar a teoria da percolação a partir dos resultados obtidos na análise multiescala e multidimensional. Assim, a altura e a largura de uma imagem, bem como probabilidade p , foram consideradas para definir um poro em que um fluido pode atravessar a partir de um caminho. A presença de um aglomerado percolante foi garantida para p maior que o limiar de percolação, definido como $p = 0,59275$ (BIRD; PERRIER, 2010). O algoritmo de rotulagem de aglomerados de Hoshen-Kopelman foi aplicado para obter os aglomerados percolantes baseados nos valores de $P = -1$ (HOSHEN; KOPELMAN, 1976). Após a rotulagem de cada poro, os aglomerados foram definidos considerando *pixels* com os mesmos

rótulos. Quando o primeiro aglomerado foi finalizado, o algoritmo avançou para o próximo poro ainda não rotulado.

Três diferentes funções foram consideradas para analisar as imagens colorretais H&E: média de aglomerados C , razão de caixas percolantes Q e razão de cobertura média do maior aglomerado Γ . O número de aglomerados em uma única caixa é dado por c_i e o número médio de aglomerados por caixa $C(L)$ é apresentado na Eq. 4.6, em que a soma do número de aglomerados em uma escala L foi dividida pelo número de caixas T .

$$C(L) = \frac{\sum_{i=1}^T c_i}{T} \quad (4.6)$$

A razão de caixas percolantes Q foi calculada ao contar o número de caixas percoladas em uma escala L . Uma caixa q_i foi considerada como uma caixa percolante caso a razão entre o número de *pixels* rotulados como poros (Ω_i) e o número total de *pixels* dentro da caixa (L^2) excedessem o limiar de percolação p (Eq. 4.7). Assim, a função $Q(L)$ foi obtida ao dividir o número total de caixas percolantes q_i pela quantidade total de caixas T em uma escala especificada L , conforme Eq. 4.8.

$$q_i = \begin{cases} 1, & \frac{\Omega_i}{L^2} \geq 0.59275 \\ 0, & \frac{\Omega_i}{L^2} < 0.59275 \end{cases} \quad (4.7)$$

$$Q(L) = \frac{\sum_{i=1}^T q_i}{T} \quad (4.8)$$

A função razão de cobertura média do maior aglomerado Γ foi calculada a partir da razão de cobertura do maior aglomerado em cada caixa avaliada de tamanho L . O cálculo da razão média de cobertura foi realizado ao dividir o somatório da cobertura em cada caixa pela quantidade de caixas T (Eq. 4.5). A divisão do maior aglomerado (γ_i) pelo número de *pixels* na caixa L^2 consiste na cobertura de uma caixa i (Eq. 4.9).

$$\Gamma(L) = \frac{\sum_{i=1}^T \gamma_i / L^2}{T} \quad (4.9)$$

O comportamento para cada função $C(L)$, $Q(L)$ e $\Gamma(L)$ foram obtidos considerando as métricas área sob a curva (*ARC*), obliquidade (*OBL*), razão de área (*RZA*), ponto máximo (*PM*) e escala do ponto máximo (*EPM*). A *ARC* foi obtida na integração a partir do método trapezoidal (Eq. 4.10), em que N é o número de amostras, a e b representam os valores máximo e mínimo de L . Neste trabalho, os valores 3 e 45 foram atribuídos a a e b , respectivamente.

$$ARC = \int_a^b f(x)dx \approx \frac{b-a}{2N} \sum_{n=a}^{b-1} (f(x_n) + f(x_{n+1})). \quad (4.10)$$

A *OBL* consistiu em uma indicação de assimetria comparando o valor médio. Os valores negativos de obliquidade indicaram que a amostra estava concentrada à esquerda do valor médio. Os valores positivos de *OBL* indicaram que a amostra estava concentrada à direita valor médio. O caso em que uma amostra é perfeitamente simétrica consiste em uma obliquidade de valor zero. Para uma amostra com N valores, a obliquidade *OBL* é calculada a partir da Eq. 4.11, em que $\bar{x}$ é o valor médio amostral e x_i indica o valor da amostra referente ao i -ésimo valor de x .

$$OBL = \frac{\frac{1}{N} \sum_{i=a}^b (x_i - \bar{x})^3}{\sqrt{[\frac{1}{N-1} \sum_{i=a}^b (x_i - \bar{x})^2]^3}}. \quad (4.11)$$

A *RZA* considerou a razão das áreas entre o lado direito e esquerdo sob a curva da função. Para o cálculo de *RZA*, foi considerada a área A entre os pontos a e b sobre o eixo x : $A_{a,b}$, apresentada na Eq. 4.12. A partir dessa relação, a razão de área *RZA* foi obtida conforme Eq. 4.13.

$$A = \int_a^b f(x)dx. \quad (4.12)$$

$$RZA = \frac{A(\frac{a}{2} + 1, b)}{A(a, \frac{b}{2})}. \quad (4.13)$$

O *PM* forneceu o valor do ponto máximo de cada função e *EPM* a escala do ponto máximo. Os tamanhos das caixas escolhidos para obter as características de percolação foram de $L = 3$ até $L = 45$, parâmetros similares aos utilizados no cálculo da DF_p .

4.2.4 Lacunaridade

Para quantificar como os *pixels* estão distribuídos e organizados em cada situação, conforme definições apresentadas por (VOSS, 1991) e (PLOTNICK et al., 1996), a lacunaridade $Lac(L)$ foi extraída das imagens coloridas e subimagens curvelet para complementar as observações realizadas com DF_f , DF_p e *Perc*. A lacunaridade foi definida com o método multidimensional descrito em (IVANOVICI; RICHARD; DECEAN, 2009), considerando os momentos de primeira ordem (Eq. 4.14) e segunda ordem Eq. 4.15. Estes momentos foram calculados com base na matriz de probabilidades que foi computada para determinar o valor de DF_f . A lacunaridade $Lac(L)$ foi

dada pela relação apresentada em Eq. 4.16. Os valores da curva de *Lac* considerando diferentes escalas *L* também foram utilizados para calcular as métricas *ARC*, *OBL*, *RZA*, *PM* e *EPM*, descritos na subseção 4.2.3. O uso dessas métricas possibilitou a extração de características globais da curva de lacunaridade para cada imagem.

$$\lambda(L) = \sum_{m=1}^N mP(m,L), \quad (4.14)$$

$$\lambda^2(L) = \sum_{m=1}^N m^2P(m,L), \quad (4.15)$$

$$\Lambda(L) = \frac{\lambda^2(L) - (\lambda(L))^2}{(\lambda(L))^2}. \quad (4.16)$$

4.3 Aplicação dos Descritores de Haralick

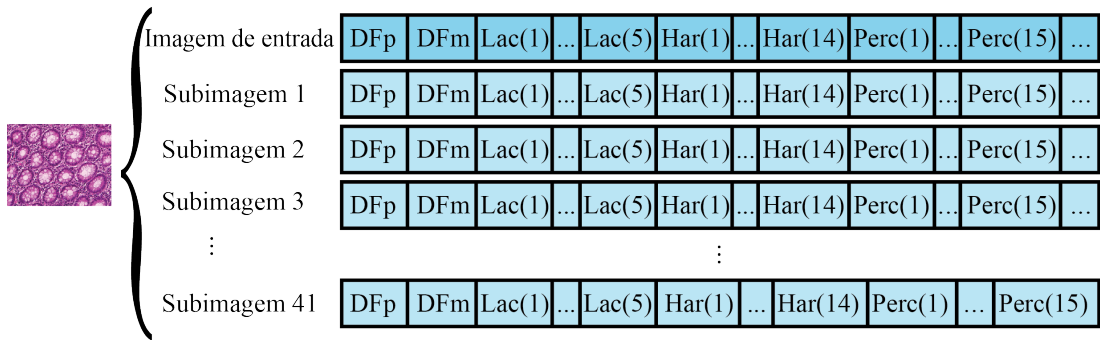
Os descritores de Haralick também foram utilizados para avaliar cada imagem H&E e as sub-imagens curvelet correspondentes. Os descritores foram calculados a partir das matrizes de co-ocorrência. Cada elemento da matriz representou a frequência que um nível de cinza *u* e outro com nível de cinza *v* ocorreram na imagem, conforme descrições apresentadas por (HARALICK; SHANMUGAM et al., 1973). O número de elementos da matriz de co-ocorrência foi em função da quantidade de intensidades *G* presentes na textura. Neste trabalho, os parâmetros considerados foram o valor de distância $\delta = 1$ e os valores de direção $\theta = 0^\circ$, $\theta = 45^\circ$, $\theta = 90^\circ$ e $\theta = 135^\circ$. A média dos valores das matrizes de co-ocorrência obtidas foram calculadas para definir os descritores de Haralick. Os descritores foram: (1) Segundo momento angular; (2) Contraste; (3) Correlação; (4) Soma dos quadrados; (5) Momento inverso das diferenças; (6) Somatório da média; (7) Somatório da variância; (8) Somatório da entropia; (9) Entropia; (10) Diferença da variância; (11) Diferença da entropia; (12) e (13) são medidas de informação de correlação; (14) Coeficiente de correlação maximal. Os descritores foram aplicados conforme proposto por (HARALICK; SHANMUGAM et al., 1973).

4.4 Composição do Vetor de Características

A partir das abordagens descritas previamente, o vetor de características foi constituído por um valor de DF_p , um valor de DF_f , cinco valores de *Lac*, obtidos a partir das métricas *ARC*, *OBL*, *RZA*, *PM* e *EPM*, representadas por *Lac*(1) até *Lac*(5); 14

características de Haralick, representadas por Har(1) até Har(14) e; 15 características de percolação, as características *ARC*, *OBL*, *RZA*, *PM* e *EPM* para cada função *C*, *Q*, Γ de percolação, nomeadas Perc(1) até Perc(15). As características mencionadas também foram calculadas para as sub-imagens curvelet, calculadas com 4 escalas e 8 ângulos, totalizando 41 sub-imagens para cada imagem colorretal. O resultado foi um vetor composto por 1512 valores, conforme ilustrado na Figura 4.3.

Figura 4.3: Ilustração do vetor de características obtido após a aplicação dos métodos fractais e características de Haralick às imagens colorretais e sub-imagens curvelet correspondentes.



Fonte: elaborado pelo autor.

4.5 Seleção das Características

O vetor de características foi avaliado a fim de reduzir sua dimensionalidade e garantir análises a partir das características mais relevantes para o entendimento do câncer colorretal, conforme descrições apresentadas por (GUYON et al., 2002). O vetor de características com 1512 valores foi submetido à análise com o algoritmo ReliefF (KONONENKO; ŠIMEC; ROBNIK-ŠIKONJA, 1997).

ReliefF é uma expansão do algoritmo Relief (KIRA; RENDELL, 1992) para possibilitar análises de duas ou mais classes e grandes conjuntos de dados (ROBNIK-ŠIKONJA; KONONENKO, 2003). Este método considera uma aproximação de probabilidades mais confiável, uma vez que analisa a contribuição média de características vizinhas (URBANOWICZ et al., 2017). O algoritmo ReliefF consiste em um processo iterativo para definir as instâncias de dados mais próximas da mesma classe e de outras classes. O critério utilizado é a distância de Manhattan. Em seguida, um ajuste de pesos W das características A é realizado por meio de uma aproximação das diferenças

de probabilidades P_r , as quais são calculadas entre a instância mais próxima da classe diferente (I_d) e a instância mais próxima pertencente à mesma classe (I_s) (Eq. 4.17). A saída do algoritmo ReliefF consiste em um vetor de características com os pesos, ranqueando as características em ordem de importância.

$$W[A] = P_r(\text{valor} \neq A|I_d) - P_r(\text{valor} \neq A|I_s). \quad (4.17)$$

4.6 Classificação e avaliação das características

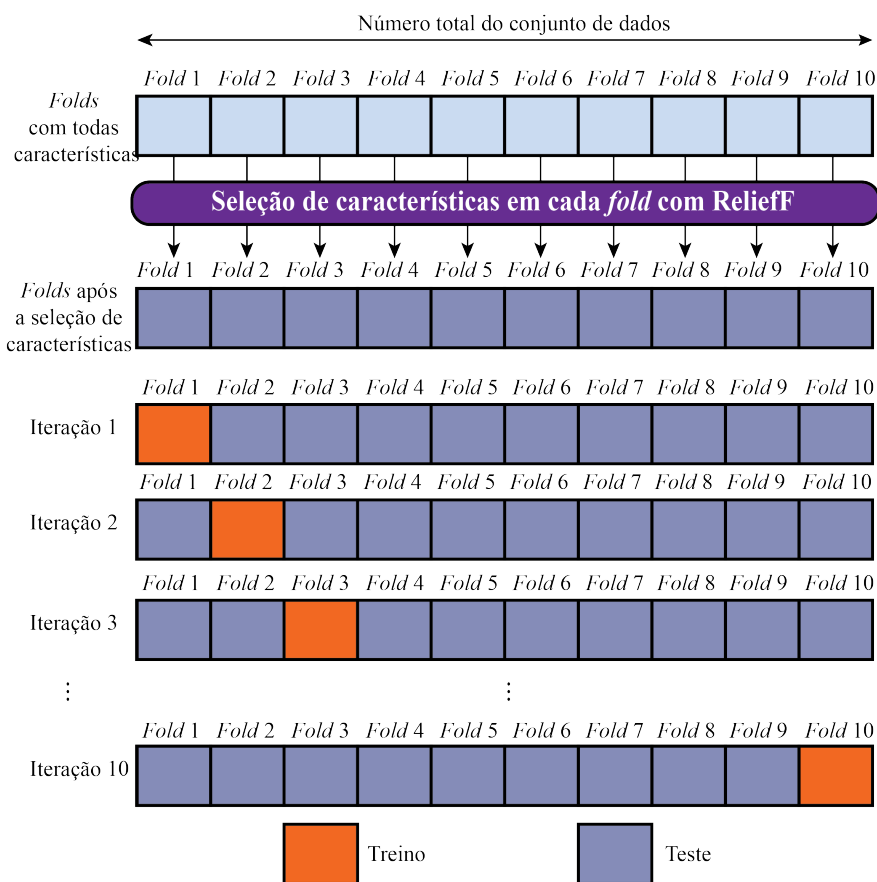
Os desempenhos dos algoritmos foram avaliados considerando a métrica área sob a curva ROC (AUC), amplamente utilizada na literatura (FAWCETT, 2004). Esta medida permite verificar a taxa de sucesso em relação à taxa de falha dos grupos (BECK; SHULTZ, 1986). Um teste de classificação é dito como perfeito quando o valor de AUC é 1. Por outro lado, se o valor de AUC é 0,5 ou menor, o teste é definido como não relevante.

O desempenho do método foi avaliado considerando algoritmos de classificação DT (SNOUSY et al., 2011), SVM (VAPNIK, 1999), RaF (BREIMAN, 2001), NB (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007), K^* (MAHMOOD; HUSSEIN, 2013) e o método PL (NASCIMENTO et al., 2013), aplicando a validação cruzada k -fold. Os classificadores DT, SVM, RaF, NB e K^* foram aplicados com as configurações padrões disponíveis na plataforma *Weka* 3.8.1 (WITTEN et al., 2016). O classificador PL foi aplicado com bases polinomiais de 4ª ordem. As amostras foram distribuídas em $k = 10$ subconjuntos exclusivos. Assim $k - 1$ subconjuntos foram utilizados para treinamento e os demais subconjuntos para testes. O procedimento foi repetido k vezes e a variação de subconjunto de teste resultou em vários valores de AUC (ARLOT; CELISSE et al., 2010). O valor final de AUC foi calculado considerando a média dos valores obtidos em cada *fold*. Uma ilustração da estratégia adotada está na Figura 4.4.

Neste trabalho, as abordagens de extração de características foram implementadas utilizando o software MATLAB R2015a. A seleção de atributos foi realizada com o auxílio da plataforma *Weka* 3.8.1 (WITTEN et al., 2016). Além disso, foram utilizados os pacotes *curvelab* (DEMANET, 2008) para o cálculo dos coeficientes *curvelet*.

A seleção das características foi realizada aplicando o algoritmo ReliefF em cada *fold* k para ranquear as características mais relevantes. O número de características foi determinado a partir de classificações realizadas com composições das 100 características mais bem posicionadas. Esta estratégia foi adotada com o objetivo de escolher o

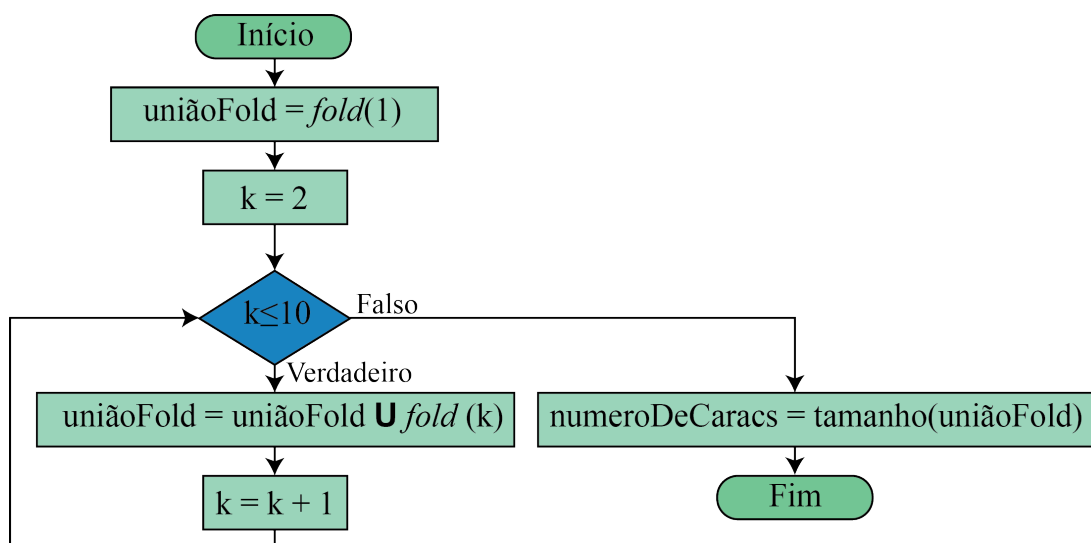
Figura 4.4: Ilustração do processo de validação cruzada e seleção por *folds* aplicado no método proposto.



Fonte: Elaborado pelo autor.

melhor desempenho médio para separar as classes de interesse com um número reduzido de características. O processo foi iniciado com a primeira característica do resultado fornecido pelo algoritmo ReliefF. As taxas de *AUC* foram coletadas para cada um dos métodos de classificação utilizados para separar as classes benigna e maligna de imagens colorretais. O resultado foi a taxa média de *AUC* entre os classificadores. Em seguida, uma nova etapa de testes foi realizada com as primeiras 5 características e, novamente, a taxa média de *AUC* foi definida pelo método. Este processo foi repetido com incrementos de 5 em 5 características, até avaliar as 100 características mais bem posicionadas. É importante destacar que o número total de características consistiu em contar as características utilizadas em cada *fold*, sem repetições. Esse processo consistiu em realizar a união dos *folds* e analisar o tamanho do conjunto obtido após esse processo (Figura 4.5). Além disso, a taxa média de *AUC* também foi calculada a partir de cada subconjunto e cada *fold*. O número de características escolhido para gerar os resultados foi a composição que forneceu a maior taxa média de *AUC*.

Figura 4.5: Fluxograma do processo de contagem de características sem repetições.



Fonte: Elaborado pelo autor.

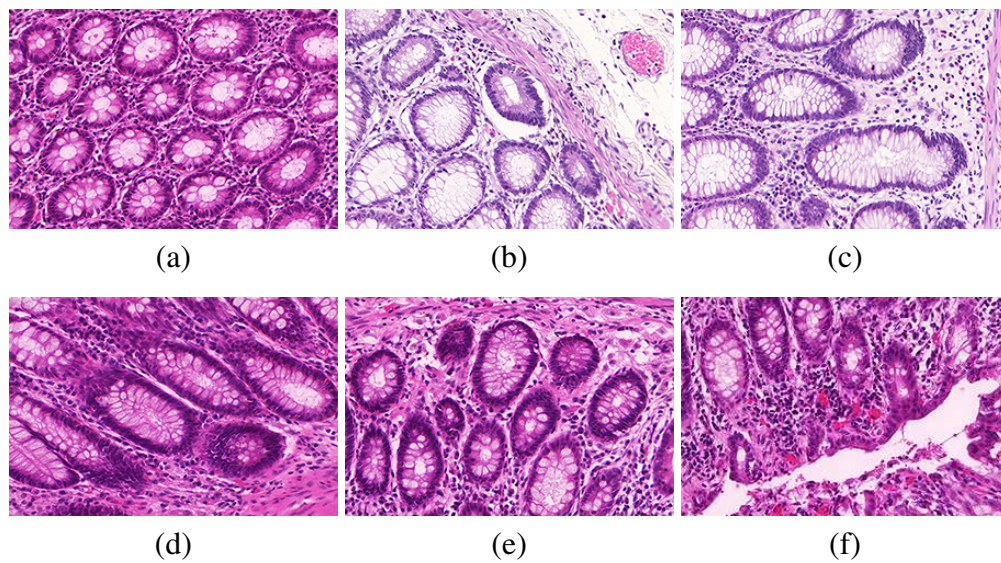
4.6.1 Análise de ruído

O desempenho da melhor associação de características foi verificado sob a influência de diferentes níveis ruído, a partir da proposta descrita por (ZHU; WU, 2004). Este teste é importante para conhecer a relevância da associação proposta em condições adversas para distinguir o câncer colorretal. O conjunto de dados foi dividido em treinamento X e teste Y . A melhor combinação de características foi analisada utilizando os conjuntos de treinos X e testes Y , em que o desempenho obtido (taxa de AUC) foi representado por conjunto de treinamento limpo vs conjunto de teste limpo (C vs C). As características foram corrompidas aplicando o método apresentado por (ZHU; WU, 2004). Assim, um novo conjunto de treinamento com ruído X' foi obtido a partir de X e aplicado para distinguir as instâncias do conjunto de teste Y . Essa comparação entre o conjunto de treino com ruído vs conjunto de teste limpo foi representada por (D vs C). Após a avaliação do conjunto de treino com ruído, os mesmos percentuais de ruídos foram incluídos nos conjuntos de testes Y . Os resultados foram conjuntos de características com ruídos Y' para o processo de avaliação. As taxas de AUC obtidas com esta metodologia foram indicadas por (C vs D) e (D vs D), as quais representam conjunto de características de treino limpo vs conjunto de características de teste com ruído, bem como conjuntos de características de treino e teste com ruídos, respectivamente. Os experimentos foram realizados com os níveis de ruído de 20% e 40%, conforme discutido por (ZHU; WU, 2004) e (XIAO et al., 2010).

4.7 Contexto de Aplicação: Banco de Imagens Histológicas Colorretais

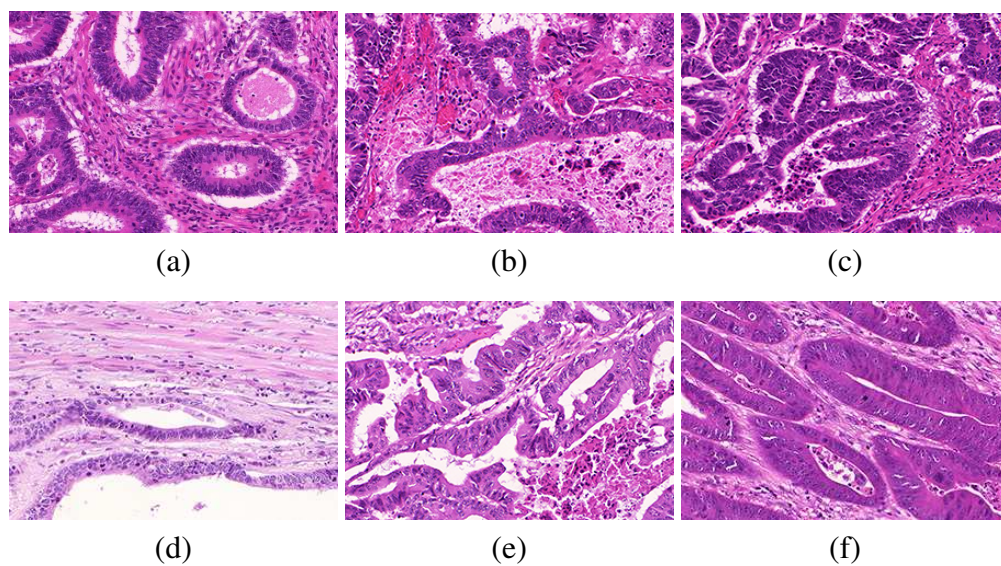
Os testes foram realizados a partir de um banco de imagens disponibilizado por (SIRINUKUNWATTANA et al., 2017). O banco é constituído por imagens histológicas derivadas de 16 seções coradas em H&E do estágio T3 ou T4 de câncer colorretal. Cada seção pertencia a um paciente, das quais foram extraídas áreas com diferentes arquiteturas histológicas. A digitalização das seções histológicas em *whole-slide images* (WSI) foi realizada com um escâner deslizante Zeiss MIRAX MIDI com uma resolução de *pixel* $0.465\mu m$. As WSIs foram reescaladas para uma resolução de *pixel* de $0.620\mu m$, sendo equivalente à magnificação 20x. As imagens foram rotuladas por um especialista na área de patologia, categorizadas nos grupos benigno ou maligno (SIRINUKUNWATTANA et al., 2017). Neste trabalho foram utilizadas 151 imagens com dimensões de 775×522 *pixels*, divididas em 67 casos benignos e 84 casos malignos. Nas Figs. 4.6 e 4.7 são apresentados exemplos de imagens dos grupos benigno e maligno, respectivamente.

Figura 4.6: Exemplos de imagens histológicas colorretais do grupo benigno.



Fonte: (SIRINUKUNWATTANA et al., 2017).

Figura 4.7: Exemplos de imagens histológicas colorretais do grupo maligno.



Fonte: (SIRINUKUNWATTANA et al., 2017)

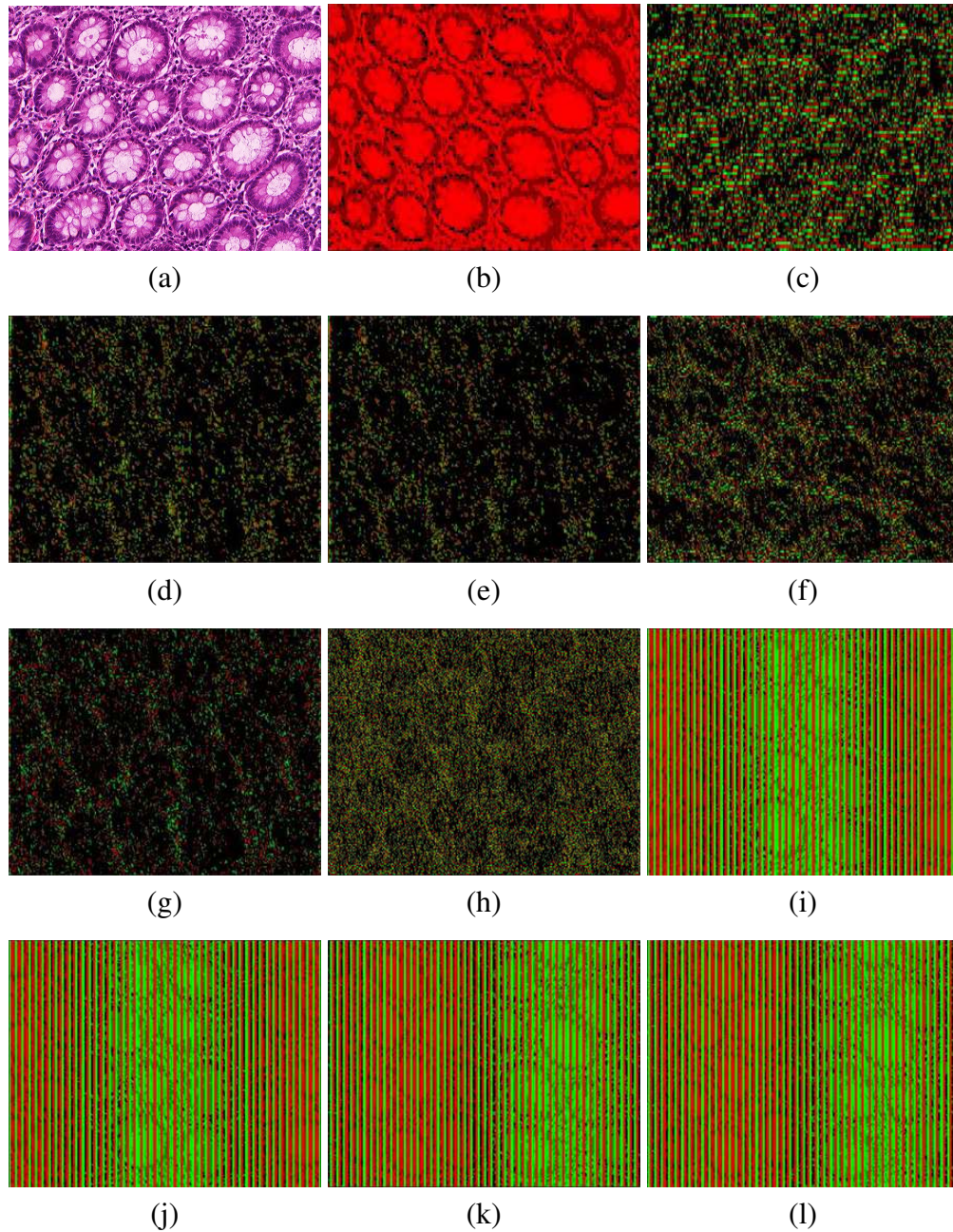
Capítulo 5

Resultados

A transformada curvelet foi aplicada em cada imagem H&E colorretal, com alterações de escala (1 a 4) e rotações (1 a 8). Os resultados foram subimagens como as ilustradas nas Figuras 5.1 para o grupo benigno e Figuras 5.2 para o grupo maligno. As cores apresentadas permitem ilustrar os coeficientes curvelets obtidos em cada subimagem, com a cor preta indicando a obtenção de coeficientes de valores zero, verde representando coeficientes baixos, amarelo definindo coeficientes intermediários e vermelho indica os valores mais altos de coeficientes.

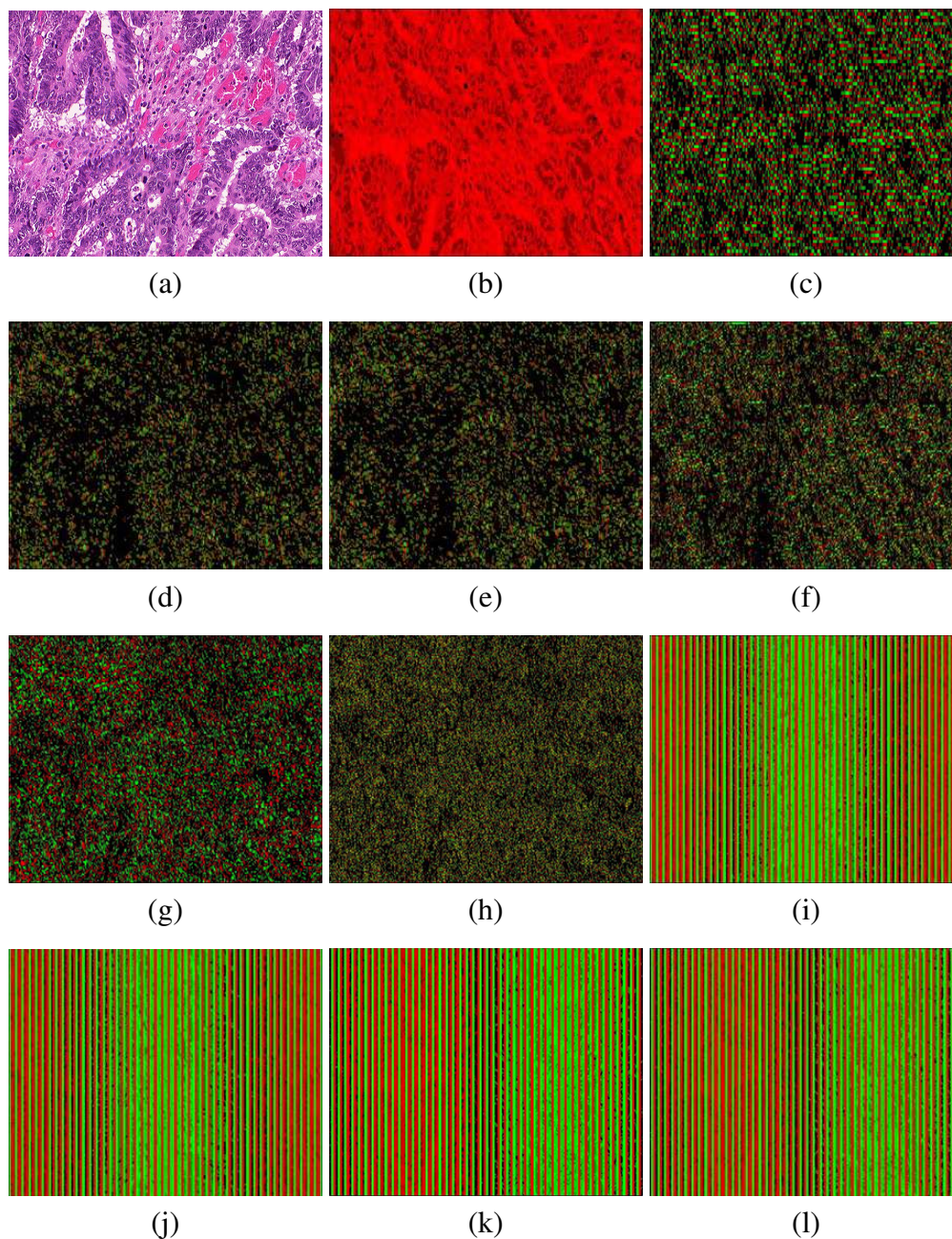
Os métodos para obter DF_f , DF_p , Lac , Har e $Perc$ foram aplicados para quantificar cada imagem dada como entrada e as subimagens curvelet correspondentes, resultando em 1512 valores: parte dos valores coletados são ilustrados na Tabela A.1.

Figura 5.1: Exemplos de uma imagem de cólon do grupo benigno (a) e as subimagens obtidas com uma variação de escalas e rotações: escala 1 e rotação 1 (b), escala 2 e rotação 2 (c), escala 2 e rotação 4 (d), escala 2 e rotação 8 (e), escala 3 e rotação 2 (f), escala 3 e rotação 6 (g), escala 3 e rotação 8 (h), escala 4 e rotação 6 (i), escala 4 e rotação 7 (j), escala 4 e rotação 14 (k), escala 4 e rotação 15 (l).



Fonte: Elaborado pelo autor.

Figura 5.2: Exemplos de uma imagem de cólon do grupo maligno (a) e as subimagens obtidas com uma variação de escalas e rotações: escala 1 e rotação 1 (b), escala 2 e rotação 2 (c), escala 2 e rotação 4 (d), escala 2 e rotação 8 (e), escala 3 e rotação 2 (f), escala 3 e rotação 6 (g), escala 3 e rotação 8 (h), escala 4 e rotação 6 (i), escala 4 e rotação 7 (j), escala 4 e rotação 14 (k), escala 4 e rotação 15 (l).

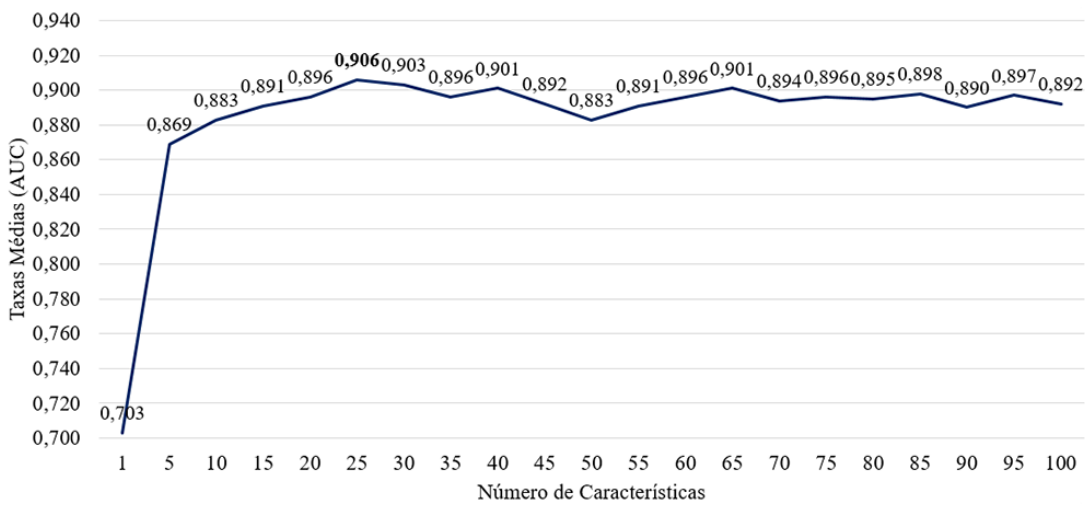


Fonte: Elaborado pelo autor.

A seleção de características foi realizada em cada *fold* para conhecer o número e as características mais relevantes entre as 1512 dadas como entrada. O processo de seleção seguiu a proposta descrita na subseção 4.6. O resultado obtido é mostrado na

Figura 5.3. Avaliando as taxas médias entre os classificadores, o número de características escolhido por *fold* foi a uma composição com 25 características, a qual resultou em uma taxa média de $AUC = 0,906$. Como o processo de validação cruzada fornece arquivos compostos por diferentes características em cada *fold*, como consequência, o método proposto utilizou o total geral de 43 características mais relevantes. Uma discriminação das características mais selecionadas pode ser observada na Tabela 5.1.

Figura 5.3: Taxas médias de AUC calculadas a partir dos *folds* e aplicando os métodos DT, SVM, NB, RaF, K* e PL. As taxas foram obtidas com subconjuntos de 1 a 100 características, limites aplicados em cada *fold*.



Fonte: Elaborado pelo autor

Tabela 5.1: Discriminação das características selecionadas pelo método proposto.

	DF _p		DF _f		Lac		Har		Perc		Total
	Imagem H&E	Sub imagem	Imagem H&E	Sub imagem	Imagem H&E	Sub imagem	Imagem H&E	Sub imagem	Imagem H&E	Sub imagem	
Número de características selecionadas	0	0	0	0	1	11	3	0	2	26	43
Número de características antes da seleção	1	41	1	41	5	205	14	574	15	615	1512

É importante notar que não foram selecionadas características obtidas a partir da dimensão fractal multiescala e multidimensional (DF_p e DF_f). Este fato pode indicar que a dimensão fractal e sua combinação com coeficientes curvelet não são relevantes para o reconhecimento de padrões de imagens colorretais H&E. Entre as características obtidas, os descritores de Haralick selecionados foram calculados diretamente das imagens de entrada, indicando que a combinação com subimagens curvelet não é relevante. Os três descritores selecionados foram variância, somatório médio e somatório

da variância, de um total de 14 descritores. Por outro lado, as características mais selecionadas (40 de um total de 43) foram as obtidas a partir da lacunaridade e percolação multiescalas e multidimensionais, com 12 e 28 características, respectivamente. Deste total de 40 características, 92,5% foram calculadas a partir de subimagens curvelet. Este fato indica um padrão importante de associação para distinguir os grupos investigados.

As características mais selecionadas foram analisadas para conhecer quais atributos e coeficientes curvelet forneceram as principais combinações. Na Tabela 5.2 é apresentada uma discriminação dos resultados obtidos, considerando o nome da característica, tipo de imagem em que o atributo foi obtido, valor médio e desvio padrão para os grupos benignos e malignos.

Tabela 5.2: Características selecionadas em todos os *folds* da abordagem de validação cruzada *k-folds* com médias para os grupos benigno e maligno. As métricas de curva para a lacunaridade e percolação foram: área sob a curva (*ARC*), obliquidade (*OBL*), razão de área (*RA*), ponto máximo (*PM*) e escala do ponto máximo (*EPM*).

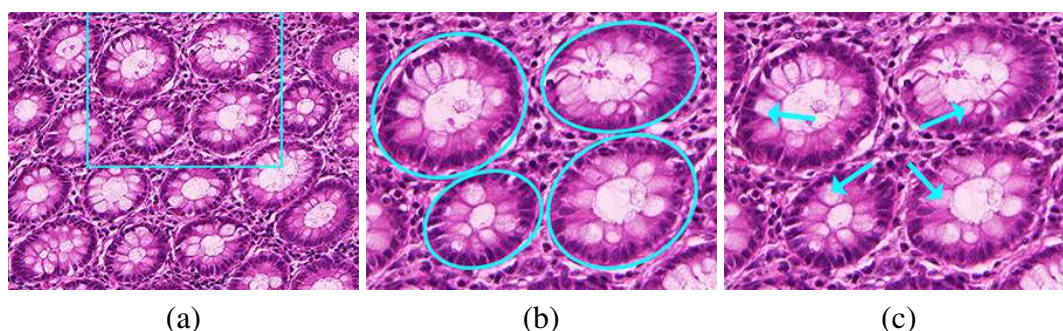
Característica	Imagem	Benigno	Maligno
RZA(<i>Lac(L)</i>)	H & E	0,832 ± 0,137	1,027±0,152
OBL(<i>Lac(L)</i>)	Curvelet: Escala 1, Rotação 1	0,404 ± 0,091	0,465 ± 0,071
ARC(<i>C(L)</i>) (Perc)	H & E	20,672±2,642	17,571±3,625
ARC(<i>Q(L)</i>) (Perc)	Curvelet: Escala 1, Rotação 1	0,457±0,246	0,239±0,241
EPM(<i>Q(L)</i>) (Perc)	Curvelet: Escala 1, Rotação 1	18,023±4,773	14,262±3,665
ARC($\Gamma(L)$) (Perc)	Curvelet: Escala 1, Rotação 1	0,433±0,157	0,619±0,126
RZA($\Gamma(L)$) (Perc)	Curvelet: Escala 1, Rotação 1	1,116±0,638	0,518±0,445
PM($\Gamma(L)$) (Perc)	Curvelet: Escala 1, Rotação 1	3,260±2,483	5,399±2,046
EPM($\Gamma(L)$) (Perc)	Curvelet: Escala 1, Rotação 1	275,281±66,178	222,293±52,896
PM($\Gamma(L)$) (Perc)	Curvelet: Escala 4, Rotação 6	2,529±0,161	2,704±0,157
PM($\Gamma(L)$) (Perc)	Curvelet: Escala 4, Rotação 7	2,530±0,162	2,709±0,159
RZA($\Gamma(L)$) (Perc)	Curvelet: Escala 4, Rotação 14	0,193±0,126	0,141±0,0761
PM($\Gamma(L)$) (Perc)	Curvelet: Escala 4, Rotação 14	2,771±0,196	2,958±0,168
RZA($\Gamma(L)$) (Perc)	Curvelet: Escala 4, Rotação 15	0,192±0,128	0,141±0,077
PM($\Gamma(L)$) (Perc)	Curvelet: Escala 4, Rotação 15	2,771±0,194	2,961±0,172

A função de percolação $\Gamma(L)$ foi a mais selecionada. Esta função é responsável por quantificar a razão média de cobertura do maior aglomerado. A partir de $\Gamma(L)$, o ponto máximo (*PM*) foi a característica mais utilizada. A maioria das características foram extraídas de subimagens curvelet com escala 1 e escala 4. A subimagem curvelet com mais características extraídas consistiu na combinação de escala 1 e rotação 1. Exemplos dessas subimagens curvelet foram apresentadas nas Figuras 5.1 (b), (i), (j), (k) e (l), bem como nas Figuras 5.2 (b), (i), (j), (k) e (l). Acreditamos que subimagens geradas com estas combinações possibilitaram destacar mais os aglomerados de *pixels*, permitindo ao atributo percolação e lacunaridade quantificar apropriadamente a

conectividade entre as regiões de cores similares, ou no caso das subimagens curvelet, coeficientes similares. Por exemplo, a lacunaridade é aplicada para quantificar como os *pixels* estão distribuídos e organizados em cada situação (VOSS, 1991; PLOTNICK et al., 1996). Essa análise é complementada ao aplicar a percolação. O método aqui proposto avalia a existência de um caminho de *pixels* conectados (aglomerado) e suas propriedades ao aplicar a teoria da percolação. Outros coeficientes curvelet não permitiram destacar tais propriedades. Por exemplo, nas Figuras 5.1 (c), (d), (e), (f), (g) e (h), bem como Figuras 5.2 (c), (d), (e), (f), (g) e (h) apresentam características de informações similares às observadas em um processo aleatório.

Os resultados apresentados possibilitam apresentar algumas interpretações a respeito dos atributos e possíveis relações com as estruturas histológicas presentes nas imagens. Observando os valores indicados na Tabela 5.2, é possível verificar que o ponto máximo (*PM*) de percolação é menor para imagens do grupo benigno do que para o grupo maligno. O aumento dos valores destes atributos, como observados entre os grupos benigno e maligno, podem estar associados a presença do câncer, que provoca um aumento não controlado de células e degrada as estruturas celulares, como as regiões de lúmen, criptas intestinais e células caliceformes. O aumento não controlado de células pode gerar mais informações no processo de quantificação, explicando os aumentos de valores dos atributos indicados previamente. Exemplos das estruturas preservadas do grupo benigno estão na Figura 5.4. As regiões degradadas pela presença do câncer, grupo maligno, podem ser visualizadas na Figura 5.5.

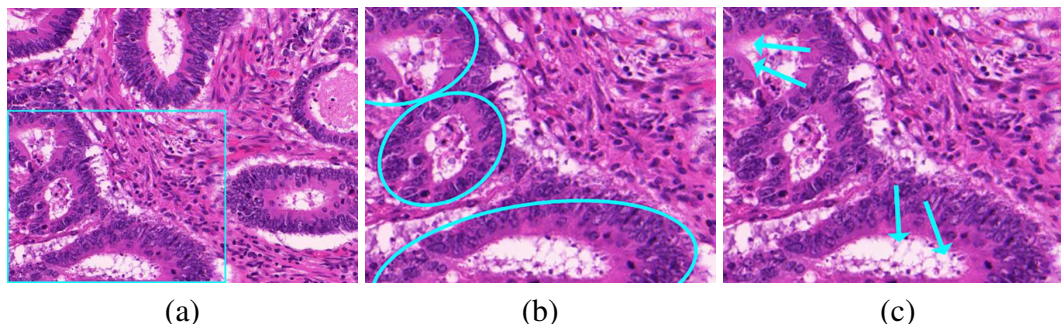
Figura 5.4: Visualização de uma imagem do grupo benigno com regiões marcadas em azul para indicar a preservação da organização e da distribuição das criptas (a) e (b), bem como das células caliceformes (c).



Fonte: Elaborado pelo autor.

O conjunto de características mais relevantes foi dado como entrada para os classificadores escolhidos para este estudo. Os resultados são apresentados na Tabela 5.3. Todos os métodos de classificação forneceram taxas de diferenciação superio-

Figura 5.5: Imagem do grupo benigno com regiões marcadas em azul para indicar a degradação da organização e da distribuição das criptas (a) e (b), bem como das células caliciformes (c).



Fonte: Elaborado pelo autor.

res a 0,820. É importante notar que as características foram suficientes para fornecer desempenhos relevantes em diferentes classificadores, com taxas médias de AUC superiores a 0,94, tais como RaF, $AUC = 0,945$, K^* , $AUC = 0,947$, e PL, $AUC = 0,994$. Este fato indica a relevância do poder discriminativo da combinação de características apresentada em nossa proposta. A partir destes destaques é possível concluir que a melhor combinação foi alcançada com os atributos fractais lacunaridade e percolação multiescala e multidimensional, calculados de imagens curvelet, aplicando o classificador PL (NASCIMENTO et al., 2013).

Tabela 5.3: Desempenhos dos algoritmos de classificação com a combinação de características obtida em nosso modelo.

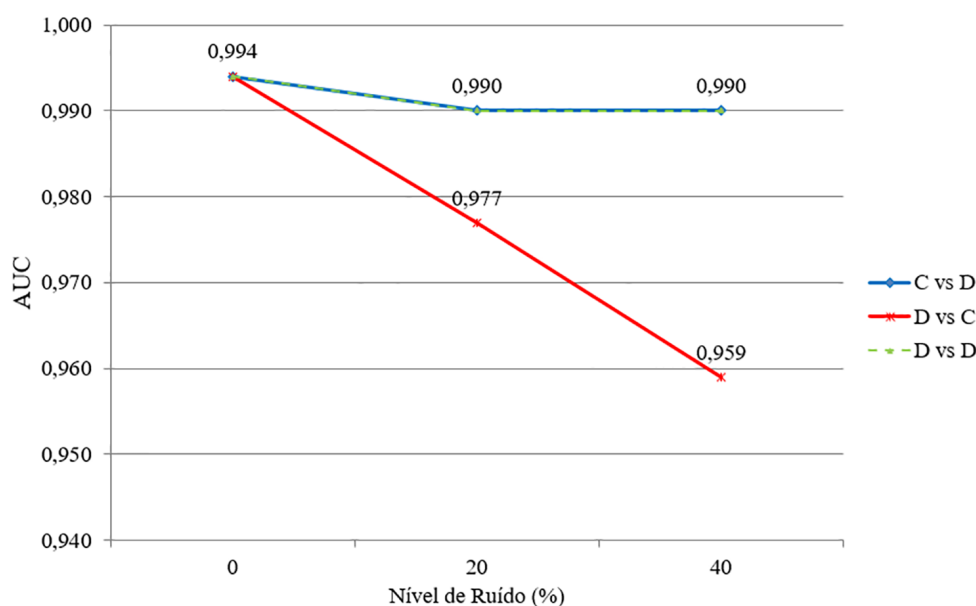
Classificador	AUC
DT	0,825
SVM	0,851
RaF	0,945
NB	0,873
K^*	0,947
PL	0,994

5.1 Análise do impacto do ruído em diferentes características

A melhor combinação de características e classificador para distinguir o câncer color-retal foi testada com a adição de diferentes níveis de ruído nos conjuntos de características. Este procedimento permitiu analisar o impacto deste processo na robustez

das características e combinação proposta. As taxas de ruído foram de 20% e 40%, conforme discutido em (ZHU; WU, 2004) e (XIAO et al., 2010). Para cada conjunto de dados foi considerado a *AUC* média com a validação cruzada *k-fold*. Na Figura 5.6 são apresentados as taxas médias de *AUC* fornecidas a partir da melhor combinação: atributos fractais lacunaridade e percolação, ambos multiescala e multidimensional, calculados de imagens curvelet, com três descritores de Haralick e aplicando o classificador PL (NASCIMENTO et al., 2013). As configurações consideradas foram: conjunto de treinamento limpo e conjunto de teste com ruído (C vs D), conjunto de treino com ruído e conjunto de teste limpo (D vs C) e ambos conjuntos de características com ruídos (D vs D).

Figura 5.6: Taxas médias de *AUC* obtidas a partir de diferentes níveis de ruído (0%, 20% e 40%) sobre as características, considerando as combinações C vs D, D vs C e D vs D.



Fonte: Elaborado pelo autor.

Considerando os resultados apresentados, é possível notar que mesmo no pior caso, conjunto de treino com ruído 40% de ruído e conjunto de teste sem ruído (D vs C), a taxa média de *AUC* foi superior à 0,959. Para as combinações C vs D e D vs D, a taxa média de *AUC* foi pouco afetada. A partir desses resultados, é possível notar que a combinação proposta é estável, mesmo sob condições adversas.

5.2 Comparação com Trabalhos Correlatos

Na combinação definida pelo sistema proposto, as técnicas fractais multiescalas e multidimensionais, descritores de Haralick e transformada curvelet possibilitaram um alto poder discriminativo para classificação e entendimento do câncer colorretal. Esse desempenho foi verificado em relação aos trabalhos correlatos. Uma variedade de estudos foram desenvolvidos para detecção de câncer colorretal partir de imagens histológicas, como nos trabalhos apresentados por (MASOOD; RAJPOOT, 2009; KALKAN et al., 2012; RATHORE et al., 2013; NAIYAR; ASIM; SHAHID, 2015; JØRGENSEN et al., 2017). Na Tabela 5.4 uma visão geral é apresentada com base nos valores de *AUC* observados pelos métodos citados e o obtido com a nossa abordagem, com informações sobre as características, etapas de pre-processamento e métodos de classificação considerando os grupos estudados.

Tabela 5.4: Desempenho do método proposto e estudos correlatos considerando os valores de *AUC*.

Referência	Número de características	Extração de característica	Etapas de Pré-processamento	Classificador	<i>AUC</i>
Massod, Rajpoot (2009)	59	Características LBP circulares	Não	SVM	0,900
Kalkan et. al (2012)	120	Haralick e Filtros de Gabor	Sim	Regressão Logística	0,950
Rathore et. al (2013)	26	<i>White run-length</i> e porcentagem de área de aglomerado	Sim	<i>Rotation Boost</i>	0,970
Naiyar, Asim and Shaid (2015)	57	Distância dos mínimos quadrados, descritores de Fourier elípticos e características morfológicas	Sim	SVM	0,960
Jorgensen et. al (2017)	9	Intensidade, primeira ordem estatística, matrizes de co-ocorrência em níveis de cinza e matrizes <i>run-length</i>	Sim	RaF	0,960
Método proposto	43(25)	Abordagens fractais multiescalas e multi-dimensionais, Haralick e coeficientes curvelet	Não	PL	0,994

Os métodos listados na Tabela 5.4 apresentaram *AUC* superiores à 0,900 ao aplicar diferentes metodologias. Cada autor propôs metodologias distintas que constituíram em contribuições relevantes observadas por meio dos resultados apresentados. A comparação direta não pode ser realizada para indicar a melhor abordagem, uma vez que tal tarefa é difícil e não adequada, principalmente pela aplicação de diferentes metodologias e bases de imagens. Mesmo assim, acreditamos que o método proposto é capaz de fornecer resultados relevantes para área de reconhecimento de padrões no contexto de imagens colorretais e compatíveis com os observados na literatura especializada no tema.

5.3 Comparação com base de imagens pública

No trabalho de (KATHER et al., 2016) foi proposto um sistema para detecção do câncer colorretal, representados pelos grupos tumor, comparando-se com o grupo sem a doença, representado pelo estroma. O conjunto de dados relativos à esses grupos consistiram em 500 imagens (250 de cada classe), provenientes do arquivo de patologia no Centro Médico Universitário de Mannheim e Universidade de Heidelberg. As amostras foram manualmente anotadas e tesselladas com porções não sobrepostas de tamanho 150px x 150px ($74\mu\text{m}$ x $74\mu\text{m}$).

Os autores Wang et. al (2017) utilizaram a mesma base disponibilizada em (KATHER et al., 2016) para classificação dos grupos tumor e estroma. A comparação direta foi possível devido ao fato da base de imagens ser disponibilizada publicamente. Exemplos de imagens do grupo tumor e estroma são apresentadas nas Figuras B.1 e B.2, respectivamente, Apêndice B. Na Tabela 5.5 são apresentados os resultados da comparação realizada, considerando as abordagens utilizadas, algoritmo de classificação e acurácia. Esses testes foram realizados considerando o mesmo conjunto de atributos selecionadas para a base de imagens apresentada em (SIRINUKUNWATTANA et al., 2017), conforme resumo apresentado nas Tabelas 5.1 e 5.2.

Tabela 5.5: Comparação do método proposto com a base colorretal disponibilizado por (KATHER et al., 2016).

Referência	Abordagem	Classificação	AC
Kather et al. (2016)	Característica de histograma, LBP, matriz de coocorrência, filtros de Gabor, características perceptron-like	SVM	98,30
Wang et al. (2017)	Redes neurais convolucionais bilineares	SVM	92,60
Proposto	Características fractais multiescala e multidimensionais, descritores de Haralick e curvelet	PL	97,68

Com base nos resultados apresentados na Tabela 5.5 é possível notar que a abordagem proposta neste trabalho obteve resultados superiores aos obtidos no trabalho de Wang et. al (2017) ($AC = 97,68\%$) e próximos ao trabalho proposto por Kather et. al (2016), com $AC = 98,30\%$. Assim, é possível destacar a relevância da combinação proposta com as técnicas fractais, Haralick e curvelet para o diagnóstico do câncer colorretal, uma vez que os resultados obtidos foram muito próximos àqueles apresentados aos trabalhos que utilizaram a mesma base de dados. Além disso, é possível destacar a generalidade do método, que forneceu taxas de diferenciação relevantes para diferentes bases de imagens.

Capítulo 6

Conclusões

Neste trabalho foi proposto um método capaz de fornecer a melhor combinação entre diferentes abordagens para extrair características multirresolução, assim sendo: abordagens fractais, multiescala e multidimensional, transformada curvelet discreta e três atributos de Haralick (variância, somatório médio e somatório da variância). O melhor resultado foi atingido com características de percolação multiescala e multidimensional extraídas de sub-imagens curvelet com escala 1 e escala 4, bem como quantificações realizadas com a lacunaridade multiescala e multidimensional, obtidas das imagens de entrada e suas sub-imagens curvelet calculadas com escala 1. O melhor resultado foi alcançado com o classificador PL com $AUC = 0,994$, sem a necessidade de uma etapa específica de pré-processamento, diferentemente das estratégias adotadas pro alguns importantes estudos na literatura.

A robustez da abordagem proposta foi analisada com diferentes taxas de ruído no conjunto de dados. Mesmo no pior caso, os resultados foram relevantes, com uma taxa média de AUC de 0,959. Os resultados obtidos foram interessantes quando comparados aos trabalhos correlatos, mostrando que o método proposto é capaz de fornecer desempenhos similares, mesmo sob a presença de ruído. Além disso, foi possível associar os valores de algumas propriedades de percolação, ponto máximo (PM) e razão de área (RA), com alguns características das estruturas celulares presentes nas imagens dos grupos benigno e maligno. Foi observado que os valores destes atributos são menores nos grupos benigno do que do grupo maligno, possivelmente pelo aumento de informações no processo de quantificação, em razão de degradações das estruturas celulares na presença do câncer.

Na comparação direta com os trabalhos relacionados, utilizando a base pública disponibilizada em (KATHER et al., 2016), a taxa de acurácia foi próxima às contribuições apresentadas na literatura. Uma observação é que os resultados obtidos

indicam que o método pode fornecer taxas relevantes em diferentes bases de imagens.

Acreditamos que os resultados obtidos são importantes para patologistas e especialistas interessados no estudo do câncer colorretal, por permitir análises detalhadas envolvendo diferentes tipos de atributos e métodos de classificação. As indicações das melhores associações com propriedades histológicas são contribuições relevantes.

6.1 Contribuições Obtidas

As principais contribuições com o desenvolvimento deste trabalho foram:

- I Definição da melhor combinação de características para a classificação e reconhecimento de padrões do câncer colorretal: lacunaridade e percolação, ambos multiescala e multidimensional, obtidas a partir de subimagens curvelet, com três atributos de Haralick extraídos diretamente das imagens coloridas;
- II Identificação dos parâmetros mais relevantes, como as funções e coeficientes curvelet para classificar os grupos benigno e maligno das imagens colorretais H&E;
- III Indicação dos descritores que forneceram as melhores taxas de classificação com as propriedades histológicas observadas nas imagens;
- IV Identificação das características que não foram relevantes para a classificação do câncer colorretal;
- V Uma combinação de características que possibilitou taxas de distinção relevantes para o câncer colorretal, mesmo sob influência de altos níveis de ruído no conjunto de dados.

6.2 Trabalhos Futuros

Em trabalhos futuros espera-se testar o comportamento do método, considerando:

- Diferentes modelos de cores;
- Diferentes parâmetros de escala e rotação da transformada curvelet;
- Testar diferentes técnicas de normalização;
- Analisar os pesos das melhores características na etapa de classificação e como cada uma destas pode induzir os resultados.

Apêndice A

Exemplo de vetores de características após obtenção dos atributos

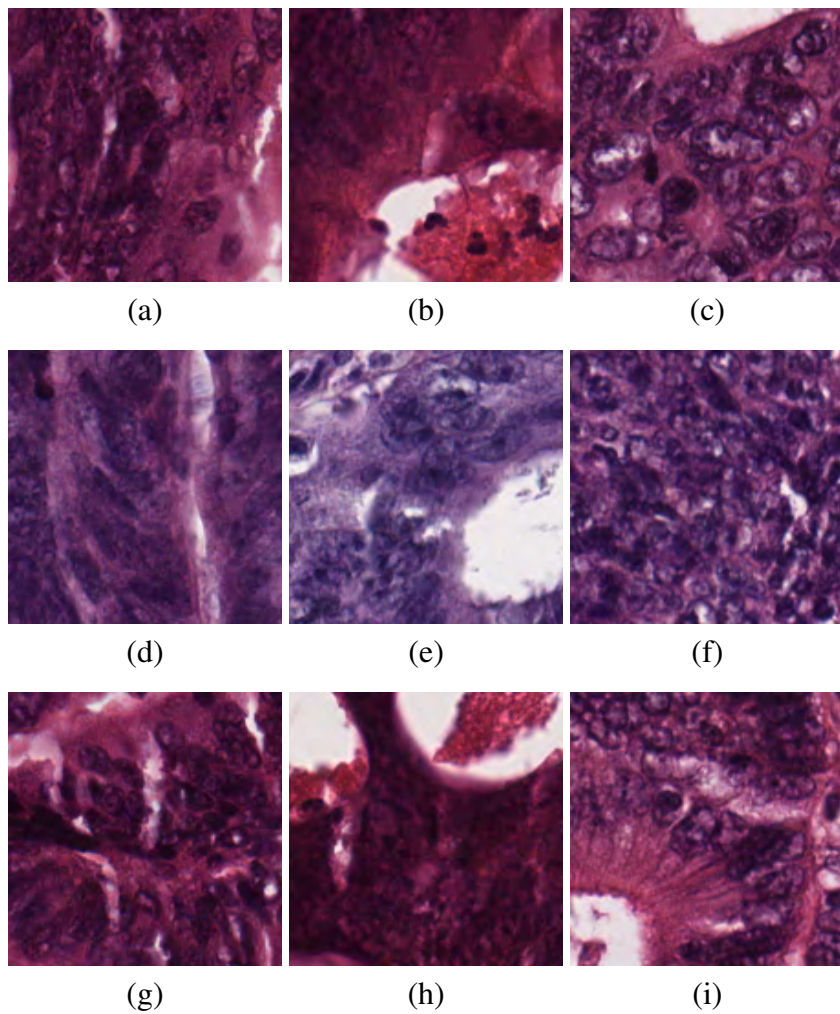
Tabela A.1: Exemplos das características obtidas a partir do método proposto. Considerando os grupos benigno e maligno.

Imagem	Imagem H & E										Sub-imagem 1					Sub-imagem 41																	
	DF_m	DF_p	$\Lambda(1)$	...	$\Lambda(5)$	Har(1)	...	Har(14)	Perc(1)	...	Perc(15)	DF_m	DF_p	$\Lambda(1)$	...	$\Lambda(5)$	Har(1)	...	Har(14)	Perc(1)	...	Perc(15)	FD_m	DF_p	$\Lambda(1)$	...	$\Lambda(5)$	Har(1)	...	Har(14)	Perc(1)	...	Perc(15)
Benigno 1	3,304	2,643	7,0971	...	5	0,00133	...	0,01837	22	...	104,844	2,354	2,464	6,217	...	5	0,9998	...	0	22	...	189,709	2,604	2,668	2,293	...	5	0,9999	...	0	11	...	332,643
Benigno 2	3,462	2,663	10,326	...	5	0,00056	...	0,01148	22	...	139,576	2,449	2,489	6,717	...	5	0,9998	...	0	22	...	324,724	2,609	2,687	2,375	...	3	0,9999	...	0	10	...	334,618
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Benigno 67	3,659	2,697	8,419	...	7	0,00021	...	0,00744	22	...	150,613	2,462	2,500	3,759	...	5	0,9998	...	0	16	...	278,312	2,626	2,713	2,074	...	3	0,9999	...	0	10	...	377,085
Maligno 1	3,521	2,749	13,474	...	7	0,00024	...	0,00784	19	...	154,745	2,435	2,525	4,610	...	5	0,9998	...	0	16	...	167,989	2,624	2,711	2,616	...	5	0,9999	...	0	9	...	238,312
Maligno 2	3,388	2,766	8,665	...	5	0,00021	...	0,00809	18	...	170,529	2,343	2,492	6,058	...	5	0,9998	...	0	16	...	229,958	2,625	2,687	3,101	...	5	0,9999	...	0	12	...	491,926
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Maligno 82	3,506	2,766	8,079	...	7	0,00032	...	0,00988	16	...	176,823	2,373	2,508	4,122	...	5	0,9998	...	0	14	...	195,006	2,580	2,689	1,751	...	3	0,9999	...	0	9	...	242,631

Apêndice B

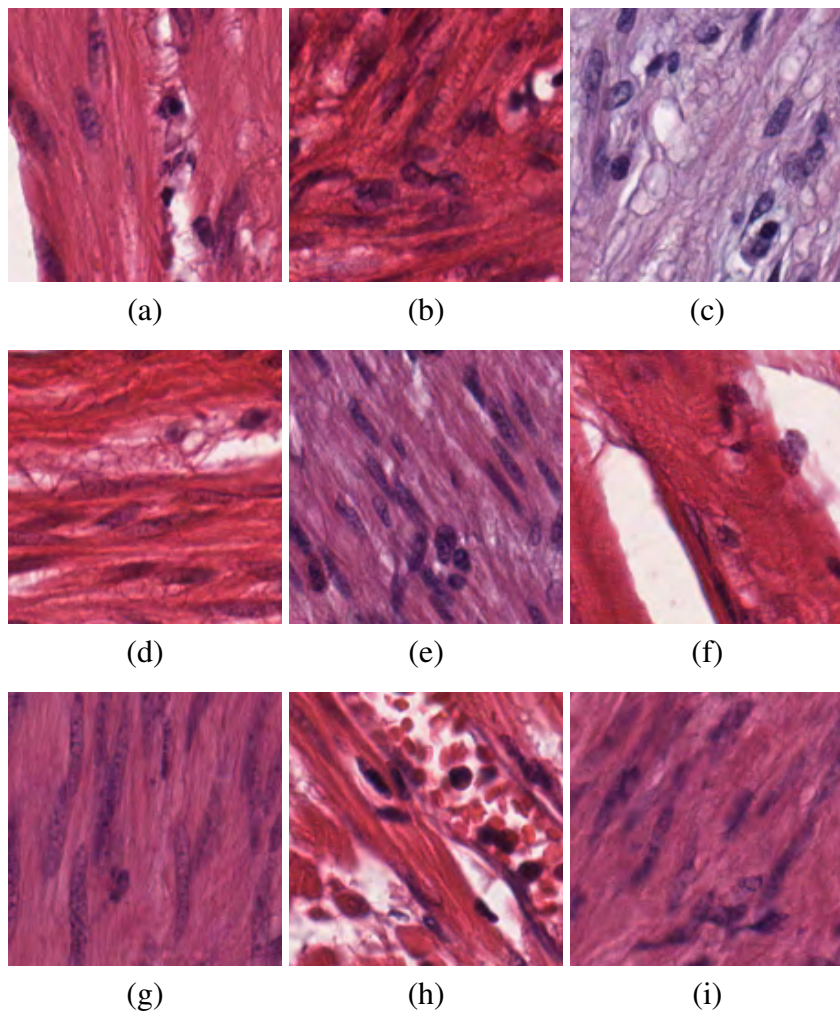
Exemplos de imagens colorretais da base pública

Figura B.1: Exemplos de imagens de colorretais do grupo tumor da base pública.



Fonte: (KATHER et al., 2016).

Figura B.2: Exemplos de imagens de colorretais do grupo estroma da base pública.



Fonte: (KATHER et al., 2016).

Referências Bibliográficas

AKBAR, B.; GOPI, V. P.; BABU, V. S. Colon cancer detection based on structural and statistical pattern recognition. In: IEEE. *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*. [S.l.], 2015. p. 1735–1739.

ALTERI, R.; KRAMER, J.; SIMPSON, S. Colorectal cancer facts and figures 2014-2016. *Atlanta: American Cancer Society*, p. 1–30, 2014.

ALTUNBAY, D.; CIGIR, C.; SOKMENSUER, C.; GUNDUZ-DEMIR, C. Color graphs for automated cancer diagnosis and grading. *IEEE Transactions on Biomedical Engineering*, IEEE, v. 57, n. 3, p. 665–674, 2010.

ALZUBI, S.; ISLAM, N.; ABBOD, M. Multiresolution analysis using wavelet, ridgelet, and curvelet transforms for medical image segmentation. *Journal of Biomedical Imaging*, Hindawi Publishing Corp., v. 2011, p. 4, 2011.

ARLOT, S.; CELISSE, A. et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, The author, under a Creative Commons Attribution License, v. 4, p. 40–79, 2010.

BACKES, A. R.; BRUNO, O. M. Segmentação de texturas por análise de complexidade. *INFOCOMP Journal of Computer Science*, v. 5, n. 1, p. 87–95, 2006.

BECK, J. R.; SHULTZ, E. K. The use of relative operating characteristic (roc) curves in test performance evaluation. *Archives of pathology & laboratory medicine*, v. 110, n. 1, p. 13–20, 1986.

BIRD, N.; PERRIER, E. Multiscale percolation properties of a fractal pore network. *Geoderma*, Elsevier, v. 160, n. 1, p. 105–110, 2010.

BOUBCHIR, L.; AL-MAADEED, S.; BOURIDANE, A. Haralick feature extraction from time-frequency images for epileptic seizure detection and classification of eeg data. In: IEEE. *Microelectronics (ICM), 2014 26th International Conference on*. [S.l.], 2014. p. 32–35.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, Elsevier, v. 30, n. 7, p. 1145–1159, 1997.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BROADBENT, S. R.; HAMMERSLEY, J. M. Percolation processes: I. crystals and mazes. In: CAMBRIDGE UNIVERSITY PRESS. *Mathematical Proceedings of the Cambridge Philosophical Society*. [S.l.], 1957. v. 53, n. 3, p. 629–641.

CAMPBELL, W. M.; ASSALEH, K. T.; BROUN, C. C. Speaker recognition with polynomial classifiers. *IEEE Transactions on Speech and Audio Processing*, IEEE, v. 10, n. 4, p. 205–212, 2002.

CANDES, E.; DEMANET, L.; DONOHO, D.; YING, L. Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, SIAM, v. 5, n. 3, p. 861–899, 2006.

CANDÈS, E. J.; DONOHO, D. L. New tight frames of curvelets and optimal representations of objects with piecewise c^2 singularities. *Communications on pure and applied mathematics*, Wiley Online Library, v. 57, n. 2, p. 219–266, 2004.

CHADDAD, A.; TANOUGAST, C.; DANDACHE, A.; BOURIDANE, A. Extraction of haralick features from segmented texture multispectral bio-images for detection of colon cancer cells. In: IEEE. *Informatics and Computational Intelligence (ICI), 2011 First International Conference on*. [S.l.], 2011. p. 55–59.

CHAKRABORTY, M.; MUKHOPADHYAY, S.; DASGUPTA, A.; PATSA, S.; ANJUM, N.; RAY, J. A new approach of oral cancer detection using bilateral texture features in digital infrared thermal images. In: IEEE. *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. [S.l.], 2016. p. 1377–1380.

CHARISIS, V. S.; HADJILEONTIADIS, L. J. Potential of hybrid adaptive filtering in inflammatory lesion detection from capsule endoscopy images. *World Journal of Gastroenterology*, Baishideng Publishing Group Inc, v. 22, n. 39, p. 8641, 2016.

CHEN, C.-h. *Handbook of pattern recognition and computer vision*. [S.l.]: World Scientific, 2015.

CLEARY, J. G.; TRIGG, L. E. K*. An instance-based learner using an entropic distance measure. In: *Proceedings of the 12th International Conference on Machine learning*. [S.l.: s.n.], 1995. v. 5, p. 108–114.

CONCI, A.; AZEVEDO, E.; LETA, F. R. *Computação Gráfica: Teoria e Prática*. Rio de Janeiro: Campus. [S.l.: s.n.], 2008. v. 407. 2 p.

DEAN, P. A new monte carlo method for percolation problems on a lattice. In: CAMBRIDGE UNIVERSITY PRESS. *Mathematical Proceedings of the Cambridge Philosophical Society*. [S.l.], 1963. v. 59, n. 2, p. 397–410.

DEMANET, L. *The curvelet organization*. 2008. <<http://www.curvelet.org/software.html>>. Accessed: 01.24.2018.

- DOBRESCU, R.; ICHIM, L.; MOCANU, S.; POPESCU, D. Benign and malignant breast tumors: Diagnosis using fractal measures. In: IEEE. *System Theory, Control and Computing (ICSTCC), 2014 18th International Conference*. [S.l.], 2014. p. 82–86.
- EID, A.; CHARISIS, V. S.; HADJILEONTIADIS, L. J.; SERGIADIS, G. D. A curvelet-based lacunarity approach for ulcer detection from wireless capsule endoscopy images. In: IEEE. *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. [S.l.], 2013. p. 273–278.
- ESGIAR, A. N.; NAGUIB, R. N.; SHARIF, B. S.; BENNETT, M. K.; MURRAY, A. 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.
- ESGIAR, A. N.; NAGUIB, R. N.; SHARIF, B. S.; BENNETT, M. K.; MURRAY, A. Fractal analysis in the detection of colonic cancer images. *IEEE transactions on information technology in biomedicine*, IEEE, v. 6, n. 1, p. 54–58, 2002.
- FALCONER, K. *Fractal geometry: mathematical foundations and applications*. [S.l.]: John Wiley & Sons, 2004.
- FARAHIBOZORG, S.; HASHEMI-GOLPAYEGANI, S.; ASHBURNER, J. Age- and sex-related variations in the brain white matter fractal dimension throughout adulthood: an mri study. *Clinical neuroradiology*, Springer, v. 25, n. 1, p. 19–32, 2015.
- FAWCETT, T. Roc graphs: Notes and practical considerations for researchers. *Machine learning*, v. 31, n. 1, p. 1–38, 2004.
- FONTES, L. R. G. Notas em percolação. 1996.
- FRANCIS, S. V.; SASIKALA, M.; SARANYA, S. Detection of breast abnormality from thermograms using curvelet transform based feature extraction. *Journal of medical systems*, Springer, v. 38, n. 4, p. 23, 2014.
- GAGNEPAIN, J.; ROQUES-CARMES, C. Fractal approach to two-dimensional and three-dimensional surface roughness. *Wear*, Elsevier, v. 109, n. 1-4, p. 119–126, 1986.
- GLEASON, D. Histologic grading and clinical staging of prostatic carcinoma. *Urologic pathology*, Lea and Febiger, 1977.
- GNANA, D. A. A.; BALAMURUGAN, S. A. A.; LEAVLINE, E. J. Literature review on feature selection methods for high-dimensional data. *International Journal of Computer Applications*, Foundation of Computer Science, v. 136, n. 1, 2016.
- GOULD, H.; TOBOCHNIK, J.; MEREDITH, D. C.; KOONIN, S. E.; MCKAY, S. R.; CHRISTIAN, W. An introduction to computer simulation methods: applications to physical systems. *Computers in Physics*, AIP, v. 10, n. 4, p. 349–349, 1996.

GRASSBERGER, P. On the hausdorff dimension of fractal attractors. *Journal of Statistical Physics*, Springer, v. 26, n. 1, p. 173–179, 1981.

GRAU, J.; MÉNDEZ, V.; TARQUIS, A. M.; DIAZ, M. C.; SAA, A. Comparison of gliding box and box-counting methods in soil image analysis. *Geoderma*, Elsevier, v. 134, n. 3-4, p. 349–359, 2006.

GUPTA, B.; TIWARI, S. Lung cancer detection using curvelet transform and neural network. *International Journal of Computer Applications*, Foundation of Computer Science, v. 86, n. 1, 2014.

GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. *Machine learning*, Springer, v. 46, n. 1, p. 389–422, 2002.

HARALICK, R. M. Statistical and structural approaches to texture. *Proceedings of the IEEE*, IEEE, v. 67, n. 5, p. 786–804, 1979.

HARALICK, R. M. 18 image texture survey. *Handbook of Statistics*, Elsevier, v. 2, p. 399–415, 1982.

HARALICK, R. M.; SHANMUGAM, K. et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, Ieee, v. 3, n. 6, p. 610–621, 1973.

HOSHEN, J.; KOPELMAN, R. Percolation and cluster distribution. i. cluster multiple labeling technique and critical concentration algorithm. *Physical Review B*, APS, v. 14, n. 8, p. 3438, 1976.

IARC. *Cancer fact sheets: Colorectal cancer*. Lyon, France, 2012.

IFTIKHAR, M. A.; HASSAN, M.; ALQUHAYZ, H. A colon cancer grade prediction model using texture and statistical features, smote and mrmr. In: IEEE. *Multi-Topic Conference (INMIC), 2016 19th International*. [S.l.], 2016. p. 1–7.

INCA. *Estimate/2018 – Cancer Incidence in Brazil*. Rio de Janeiro, Brazil, 2017.

IVANOVICI, M.; RICHARD, N. Fractal dimension of color fractal images. *IEEE Transactions on Image Processing*, IEEE, v. 20, n. 1, p. 227–235, 2011.

IVANOVICI, M.; RICHARD, N.; DECEAN, H. Fractal dimension and lacunarity of psoriatic lesions-a colour approach. *medicine*, v. 6, n. 4, p. 7, 2009.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, Acm, v. 31, n. 3, p. 264–323, 1999.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. [S.l.], 1995. p. 338–345.

JØRGENSEN, A. S.; RASMUSSEN, A. M.; ANDERSEN, N. K. M.; ANDERSEN, S. K.; EMBORG, J.; RØGE, R.; ØSTERGAARD, L. R. Using cell nuclei features to detect colon cancer tissue in hematoxylin and eosin stained slides. *Cytometry Part A*, Wiley Online Library, 2017.

JULESZ, B.; BERGEN, J. R. Human factors and behavioral science: Textons, the fundamental elements in preattentive vision and perception of textures. *The Bell system technical journal*, Alcatel-Lucent, v. 62, n. 6, p. 1619–1645, 1983.

KALKAN, H.; NAP, M.; DUIN, R. P.; LOOG, M. Automated classification of local patches in colon histopathology. In: IEEE. *Pattern Recognition (ICPR), 2012 21st International Conference on*. [S.l.], 2012. p. 61–64.

KATHER, J. N.; WEIS, C.-A.; BIANCONI, F.; MELCHERS, S. M.; SCHAD, L. R.; GAISER, T.; MARX, A.; ZÖLLNER, F. G. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, Nature Publishing Group, v. 6, p. 27988, 2016.

KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: *Proceedings of the ninth international workshop on Machine learning*. [S.l.: s.n.], 1992. p. 249–256.

KONONENKO, I.; ŠIMEC, E.; ROBNIK-ŠIKONJA, M. Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, Springer, v. 7, n. 1, p. 39–55, 1997.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. *Supervised machine learning: A review of classification techniques*. 2007.

KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, Springer, v. 29, n. 1, p. 1–27, 1964.

LAMANI, D.; MANJUNATH, T. et al. Fractal dimension with object rotation: A case study with glaucoma eye. In: IEEE. *Signal and Image Processing (ICSIP), 2014 Fifth International Conference on*. [S.l.], 2014. p. 111–116.

LIN, W.-C.; LI, C.-C.; CHRISTUDASS, C. S.; EPSTEIN, J. I.; VELTRI, R. W. Curvelet-based classification of prostate cancer histological images of critical gleason scores. In: IEEE. *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. [S.l.], 2015. p. 1020–1023.

LIU, G.; YAN, G.; KUANG, S.; WANG, Y. Detection of small bowel tumor based on multi-scale curvelet analysis and fractal technology in capsule endoscopy. *Computers in biology and medicine*, Elsevier, v. 70, p. 131–138, 2016.

LIU, X.; WANG, D. Texture classification using spectral histograms. *IEEE transactions on image processing*, IEEE, v. 12, n. 6, p. 661–670, 2003.

MAHMOOD, D. Y.; HUSSEIN, M. A. Intrusion detection system based on k-star classifier and feature set reduction. *IOSR Journal of Computer Engineering*, v. 15, n. 5, p. 107–12, 2013.

MALARZ, K. Simple cubic random-site percolation thresholds for neighborhoods containing fourth-nearest neighbors. *Physical Review E*, APS, v. 91, n. 4, p. 043301, 2015.

MANDELBROT, B. B.; PIGNONI, R. *The fractal geometry of nature*. [S.l.]: WH freeman New York, 1983. v. 1.

MARTINEZ, E. Z.; LOUZADA-NETO, F.; PEREIRA, B. d. B. A curva roc para testes diagnósticos. *Cad. saúde colet.*, (Rio J.), v. 11, n. 1, p. 7–31, 2003.

MASOOD, K.; RAJPOOT, N. Texture based classification of hyperspectral colon biopsy samples using clbp. In: IEEE. *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*. [S.l.], 2009. p. 1011–1014.

MASOOD, K.; RAJPOOT, N. M.; QURESHI, H. A.; RAJPOOT, K. Co-occurrence and morphological analysis for colon tissue biopsy classification. 2006.

MATHWORKS. *Machine Learning Method for Finding and Visualizing Natural Groupings and Patterns in Data*. 2017. Disponível em: <<https://www.mathworks.com/discovery/cluster-analysis.html>>.

NAIYAR, M.; ASIM, Y.; SHAHID, A. Automated colon cancer detection using structural and morphological features. In: IEEE. *Frontiers of Information Technology (FIT), 2015 13th International Conference on*. [S.l.], 2015. p. 240–245.

NASCIMENTO, M. Z. D.; MARTINS, A. S.; NEVES, L. A.; RAMOS, R. P.; FLORES, E. L.; CARRIJO, G. A. Classification of masses in mammographic image using wavelet domain features and polynomial classifier. *Expert Systems with Applications*, Elsevier, v. 40, n. 15, p. 6213–6221, 2013.

NAYAK, D. R.; DASH, R.; MAJHI, B.; PRASAD, V. Automated pathological brain detection system: A fast discrete curvelet transform and probabilistic neural network based approach. *Expert Systems with Applications*, Elsevier, v. 88, p. 152–164, 2017.

NEVES, L.; NASCIMENTO, M.; OLIVEIRA, D.; MARTINS, A.; GODOY, M.; ARRUDA, P.; NETO, D. de S.; MACHADO, J. Multi-scale lacunarity as an alternative to quantify and diagnose the behavior of prostate cancer. *Expert Systems with Applications*, v. 41, n. 11, p. 5017 – 5029, 2014. ISSN 0957-4174. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0957417414001250>>.

NIKOLAIDIS, N.; NIKOLAIDIS, I.; TSOUROS, C. A variation of the box-counting algorithm applied to colour images. *arXiv preprint arXiv:1107.2336*, 2011.

PEDRINI, H.; SCHWARTZ, W. R. *Análise de imagens digitais: princípios, algoritmos e aplicações*. [S.l.]: Thomson Learning, 2008.

PLOTNICK, R. E.; GARDNER, R. H.; HARGROVE, W. W.; PRESTEGAARD, K.; PERLMUTTER, M. Lacunarity analysis: a general technique for the analysis of spatial patterns. *Physical review E*, APS, v. 53, n. 5, p. 5461, 1996.

RATHORE, S.; IFTIKHAR, M. A.; HUSSAIN, M.; JALIL, A. Classification of colon biopsy images based on novel structural features. In: IEEE. *Emerging Technologies (ICET), 2013 IEEE 9th International Conference on*. [S.l.], 2013. p. 1–6.

ROBERTO, G. F.; NEVES, L. A.; NASCIMENTO, M. Z.; TOSTA, T. A.; LONGO, L. C.; MARTINS, A. S.; FARIA, P. R. Features based on the percolation theory for quantification of non-hodgkin lymphomas. *Computers in Biology and Medicine*, Elsevier, v. 91, p. 135–147, 2017.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Theoretical and empirical analysis of relief and rrelief. *Machine learning*, Kluwer Academic Publishers, v. 53, n. 1-2, p. 23–69, 2003.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. *bioinformatics*, Oxford University Press, v. 23, n. 19, p. 2507–2517, 2007.

SAMANTA, S.; AHMED, S. S.; SALEM, M. A.-M. M.; NATH, S. S.; DEY, N.; CHOWDHURY, S. S. Haralick features based automated glaucoma classification using back propagation neural network. In: *FICTA (1)*. [S.l.: s.n.], 2014. p. 351–358.

SARASWATHI, D.; DHARANI, D.; SRINIVASAN, E. An efficient feature extraction technique for breast cancer diagnosis using curvelet transform and swarm intelligence. In: IEEE. *Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on*. [S.l.], 2016. p. 441–445.

SARASWATHI, D.; SRINIVASAN, E. A high-sensitivity computer-aided system for detecting microcalcifications in digital mammograms using curvelet fractal texture features. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, Taylor & Francis, v. 5, n. 4, p. 263–273, 2017.

SARKAR, N.; CHAUDHURI, B. B. An efficient approach to estimate fractal dimension of textural images. *Pattern recognition*, Elsevier, v. 25, n. 9, p. 1035–1041, 1992.

SERRANO, R. C.; CONCI, A.; MELO, R. H. de; LIMA, R. C. et al. On using lacunarity for diagnosis of breast diseases considering thermal images. In: IEEE. *Systems, Signals and Image Processing, 2009. IWSSIP 2009. 16th International Conference on*. [S.l.], 2009. p. 1–4.

SIEGEL, R. L.; MILLER, K. D.; JEMAL, A. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, Wiley Online Library, v. 68, n. 1, p. 7–30, 2 2018. ISSN 1542-4863. Disponível em: <<https://doi.org/10.3322/caac.21442>>.

SIRINUKUNWATTANA, K.; PLUIM, J. P.; CHEN, H.; QI, X.; HENG, P.-A.; GUO, Y. B.; WANG, L. Y.; MATUSZEWSKI, B. J.; BRUNI, E.; SANCHEZ, U. et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, Elsevier, v. 35, p. 489–502, 2017.

- SNOUSY, M. B. A.; EL-DEEB, H. M.; BADRAN, K.; KHLIL, I. A. A. Suite of decision tree-based classification algorithms on cancer gene expression data. *Egyptian Informatics Journal*, Elsevier, v. 12, n. 2, p. 73–82, 2011.
- SUGANYA, R.; RAJARAM, S. Feature extraction and classification of ultrasound liver images using haralick texture-primitive features: Application of svm classifier. In: IEEE. *Recent Trends in Information Technology (ICRTIT), 2013 International Conference on*. [S.l.], 2013. p. 596–602.
- SUMANA, I. J.; ISLAM, M. M.; ZHANG, D.; LU, G. Content based image retrieval using curvelet transform. In: IEEE. *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*. [S.l.], 2008. p. 11–16.
- 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.
- TANG, L.; HE, C. Multiscale texture extraction with hierarchical (bv, gp, l2) decomposition. *Journal of mathematical imaging and vision*, Springer, v. 45, n. 2, p. 148–163, 2013.
- TRICOT, C. *Curves and fractal dimension*. [S.l.]: Springer Science & Business Media, 1994.
- URBANOWICZ, R. J.; MEEKER, M.; LACAVA, W.; OLSON, R. S.; MOORE, J. H. Relief-based feature selection: introduction and review. *arXiv preprint arXiv:1711.08421*, 2017.
- VAPNIK, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, IEEE, v. 10, n. 5, p. 988–999, 1999.
- VOSS, R. F. Random fractals: characterization and measurement. In: *Scaling phenomena in disordered systems*. [S.l.]: Springer, 1991. p. 1–11.
- WANG, C.; SHI, J.; ZHANG, Q.; YING, S. Histopathological image classification with bilinear convolutional neural networks. In: IEEE. *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. [S.l.], 2017. p. 4050–4053.
- WEISS, S.; KULIKOWSKI, C. Computer systems that learn', kaufman publishers inc., san mateo.. brusic, v., rudy, g. and harrison, lc (1994). *MHCPEP: a database of MHC-binding peptides*", *Nucleic Acids Res*, v. 22, n. 17, p. 3663–3665, 1990.
- WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016.
- XIAO, J.; HE, C.; JIANG, X.; LIU, D. A dynamic classifier ensemble selection approach for noise data. *Information Sciences*, Elsevier, v. 180, n. 18, p. 3402–3421, 2010.

YAMAGUCHI, T.; HASHIMOTO, S. Fast crack detection method for large-size concrete surface images using percolation-based image processing. *Machine Vision and Applications*, Springer, v. 21, n. 5, p. 797–809, 2010.

ZHU, X.; WU, X. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, Springer, v. 22, n. 3, p. 177–210, 2004.