



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

Luiz Fernando Segato dos Santos

Entropias *Fuzzy* e Multidimensionais
para Quantificação de Imagens
Histológicas Colorretais

São José do Rio Preto
2018

Luiz Fernando Segato dos Santos

Entropias *Fuzzy* e Multidimensionais
para Quantificação de Imagens
Histológicas Colorretais

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.

Orientador:

Prof. Dr. Leandro Alves Neves

**São José do Rio Preto
2018**

Santos, Luiz Fernando Segato dos.

Entropias Fuzzy e multidimensionais para quantificação de imagens histológicas colorretais / Luiz Fernando Segato dos Santos. -- São José do Rio Preto, 2018

116 f. : il., tabs.

Orientador: Leandro Alves Neves

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

1. Computação - Matemática. 2. Processamento de imagens - Técnicas digitais. 3. Entropia (Teoria da informação). 4. Função Fuzzy. 5. Cólon (Anatomia) - Câncer. I. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. II. Título.

CDU – 518.72:76

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

Luiz Fernando Segato dos Santos

Entropias *Fuzzy* e Multidimensionais
para Quantificação de Imagens
Histológicas Colorretais

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.

Comissão Examinadora:

Prof. Dr. Leandro Alves Neves
UNESP – Câmpus de São José do Rio Preto
Orientador

Prof. Dr. Adriano Mauro Cansian
UNESP – Câmpus de São José do Rio Preto

Prof. Dr. Alessandro Santana Martins
Instituto Federal do Triângulo Mineiro – Câmpus Ituiutaba

São José do Rio Preto
14 de maio de 2018

À minha esposa

À minha família

À ciência

"A conquista é um acaso que talvez dependa mais das falhas dos vencidos do que do gênio do vencedor".

-Madame de Staël

Agradecimentos

Agradeço a Deus primeiramente pela sabedoria e paciência concedidas em momentos críticos desta caminhada.

Agradeço à minha esposa pelo companheirismo e por me apoiar em todos os momentos, além de abdicar da minha atenção para que eu pudesse me dedicar ao trabalho. Agradeço também a minha querida enteada pelo carinho e compreensão deste trabalho em nossas vidas.

Agradeço também à minha família, principalmente minha mãe e avós, por terem me ensinado valores que fizeram com que eu persistisse na minha batalha diária.

Agradeço ao prof. Dr. Leandro Alves Neves pela paciência, pelos ensinamentos e por estar meu lado na construção deste pilar em minha vida e em minha carreira.

Agradeço aos colegas de laboratório pelo auxílio e dicas com a metodologia e a utilização de ferramentas, principalmente ao Guilherme Freire e ao Guilherme Bottazzo. Também agradeço aos colegas de trabalho e chefes que colaboraram com as ausências e trocas de aulas para que eu pudesse cursar as disciplinas do programa e também participar de eventos.

Agradeço também a meu pai (em memória), pois sua "partida" fortaleceu ainda mais meus ideais.

Por último, agradeço à minha filha, Maria Luiza, que irradia os meus dias com muita felicidade e faz eu querer me esforçar cada vez mais.

Resumo

Neste trabalho são apresentados dois novos métodos para calcular as entropias de imagens coloridas. Os métodos foram baseados nas entropias Amostral e de Shannon com associações de abordagens multiescala e multidimensional. Os modelos foram aplicados para quantificar e classificar imagens histológicas H&E de câncer colorretal. A abordagem multiescala foi definida com observações em janelas de diferentes tamanhos e variações da tolerância para determinar a similaridade de padrões. A estratégia multidimensional foi definida considerando cada pixel da imagem colorida no padrão RGB como um vetor n-dimensional, o qual foi analisado a partir da distância de Minkowski. A entropia Amostral também contemplou uma estratégia *Fuzzy*, uma função gaussiana para verificar as pertinências das distâncias das janelas. Os resultados foram métodos capazes de computar as similaridades entre *pixels* contidos em janelas de diversos tamanhos, bem como as informações presentes nos canais de cores. O poder de quantificação de cada modelo foi testado em uma base pública de imagens colorretais H&E, composta pelas classes benigna e maligna. Os resultados foram dados como entradas para classificadores de diferentes categorias e analisados via validação cruzada *k-folds*, com $k=10$. Os desempenhos indicaram que a associação proposta foi capaz de distinguir os grupos benigno e maligno, com valores que superaram os resultados obtidos com importantes técnicas disponíveis na literatura. O melhor desempenho foi uma AUC de 0,952, resultado que é relevante principalmente quando consideramos as dificuldades da prática clínica para separar os grupos estudados.

Palavras-chave: Entropia Amostral. Entropia de Shannon. Abordagem multiescala. Abordagem multidimensional. Função *Fuzzy*. Câncer colorretal.

Abstract

In this study, two new methods are presented to calculate the entropies of colour images. The methods were based on the combination of Sample and Shannon Entropy with multiscale and multidimensional approaches. The model was applied to quantify and classify histological images H&E of colorectal cancer. The multiscale approach was defined by analysing windows of different sizes and variations in tolerance for determining pattern similarity. The multidimensional strategy was established through a consideration of each pixel in the colour image through the RGB standard, as an n -dimensional vector, which was analysed from the Minkowski distance. A Fuzzy strategy was also considered by the Sample Entropy, a Gaussian function used to verify the pertinence of the distances between windows. The results were methods capable of computing similarities between pixels contained in windows of various sizes, as well as the information present in the colour channels. The power of quantification of each model was tested in a public colorectal H&E image database, which was composed of both benign and malignant classes. The results were given as inputs for classifiers of different categories and analysed by applying k -fold cross-validation, with $k=10$. The derived performances indicate that the proposed association was capable of distinguishing the benign and malignant groups, with values that surpassed those results obtained with important techniques available in the Literature. The best performance was an AUC of 0.952, a result considered relevant, mainly when we consider the difficulties of clinical practice, when it comes to separating the groups under study.

Keywords: Sample Entropy. Shannon Entropy. Multiscale approach. Multidimensional approach. Fuzzy functions. Colorectal cancer.

Lista de Figuras

2.1	Representação do processo de granulação.	33
2.2	Exemplos de imagens das classes benigna (a) e maligna (b).	40
4.1	Ilustração das etapas do método proposto.	58
4.2	Ilustração do deslocamento de janelas que representam os padrões de tamanho $m \times m$, de cor vermelha, realizado em imagens, considerando as escalas $m = 2$ e $m = 3$	62
4.3	Ilustração do modelo utilizado para definir sub-regiões no processo de seleção de janelas aleatórias.	63
4.4	Ilustração do processo de rotulação de uma janela $b_{i,j}^g$ a partir da aplicação da função de Minkowski na janela $a_{i,j}^g$	67
4.5	Exemplos de imagens histológicas colorretais do grupo benigno.	69
4.6	Exemplos de imagens histológicas colorretais do grupo maligno.	69
4.7	Imagens geradas a partir do <i>software</i> Matlab R2012b, (a) imagem colorida (a) e a correspondente em níveis de cinza (b), para exemplificar as qualificações realizadas com os modelos de entropias propostos.	71
4.8	Curvas de entropias Amostral ($SampEn_{MF}$) das imagens de controle, com escalas variando de $m = 1$ a $m = 3$	71
4.9	Curvas de entropia de Shannon (H_M) das imagens de controle.	73
4.10	Ilustração da estrutura adotada para a primeira composição de atributos, um vetor com 24 descritores para a entropia Amostral ($SampEn_{MF}$).	76

4.11	Ilustração da estrutura para a segunda composição de atributos, um vetor definido com 13 descritores para a entropia Amostral ($SampEn_{MF}$).	76
4.12	Ilustração da estrutura do vetor completo, definido por 37 atributos da entropia Amostral ($SampEn_{MF}$).	76
4.13	Ilustração da estrutura do vetor com 31 atributos da entropia de Shannon (H_M).	77
4.14	Ilustração da estrutura do vetor com cinco atributos da entropia de Shannon (H_M).	77
4.15	Ilustração da estrutura do vetor composto por 36 atributos para entropia de Shannon (H_M).	77
5.1	Exemplos de imagens das classes benigna, (a) e (b), e maligna, (c) e (d).	80
5.2	Curvas da entropia Amostral ($SampEn_{MF}$), com variações dos parâmetros ϵ e m , para as imagens indicadas nas Figuras 5.1a (a) e 5.1b (b).	81
5.3	Curvas de entropia Amostral ($SampEn_{MF}$), com variações dos parâmetros ϵ e m , para as imagens ilustradas nas Figuras 5.1c (a) e 5.1d (b).	82
5.4	Curvas médias de entropia Amostral ($SampEn_{MF}$) com as variações dos parâmetros ϵ e m referentes aos grupos benigno e maligno.	84
5.5	Taxas médias de AUC de DF, lacunaridade, da $SampEn_{MF}$ e da $SampEn_{2D}$ (SILVA et al., 2016), considerando as diferentes configurações dos vetores de características.	89
5.6	Curvas de entropia de Shannon (H_M) em função das diversas variações de escala do parâmetro g para as imagens representadas nas Figuras 5.1a (a) e 5.1b (b).	90

5.7	Curvas de entropia de Shannon (H_M) em função das diversas variações de escala do parâmetro g para as imagens representadas nas Figuras 5.1c (a) e 5.1d (b).	91
5.8	Curvas médias de entropia de Shannon (H_M) com as variações do parâmetro g referentes aos grupos benigno e maligno.	92
5.9	Taxas médias de AUC de DF, lacunaridade, H_M e H , considerando as diferentes configurações dos vetores de características.	94
5.10	Taxas médias de AUC de DF, lacunaridade, $SampEn_{2D}$, $SampEn_{MF}$, H e H_M , considerando as diferentes configurações dos vetores de características.	96

Lista de Tabelas

3.1	Resumo dos principais trabalhos correlatos, com informações sobre as imagens, grupos investigados, características e métodos de classificação.	55
5.1	Ilustração de vetores de características para as imagens mostradas na Figura 5.1, considerando a configuração dos 24 atributos obtidos pelo cálculo da entropia Amostral ($SampEn_{MF}$).	83
5.2	Ilustração de vetores de características para as imagens ilustradas na Figura 5.1, considerando a configuração dos 13 atributos obtidos pela curva de entropia Amostral ($SampEn_{MF}$).	84
5.3	Ilustração de vetores de características para as imagens representadas na Figura 5.1, considerando a configuração dos 37 atributos obtidos pela junção dos atributos do cálculo da entropia Amostral ($SampEn_{MF}$) e os extraídos da curva.	84
5.4	Valores médios da entropia Amostral ($SampEn_{MF}$) e desvios padrões (DP) dos grupos benigno e maligno.	85
5.5	Valores e médias de AUC calculados a partir da entropia Amostral proposta ($SampEn_{MF}$), considerando diferentes classificadores e composições dos vetores de características.	86
5.6	Valores e médias de AUC calculados a partir da técnica $SampEn_{2D}$ (SILVA et al., 2016).	86

5.7	Valores e médias de AUC calculados a partir de diferentes classificadores e as composições dos vetores de características, os quais foram definidos por meio da dimensão fractal (DF) apresentada por Ivanovici e Richard (2011) e por Nikolaidis, Nikolaidis e Tsouros (2011) e lacunaridade (LAC) apresentada por Ivanovici, Richard e Decean (2009).	88
5.8	Ilustração de vetores de características para as imagens representadas pela Figura 5.1, considerando a configuração dos 31 atributos obtidos pelo cálculo da entropia de Shannon (H_M).	91
5.9	Ilustração de vetores de características para as imagens indicadas na Figura 5.1, considerando a configuração dos 5 atributos obtidos pela curva da entropia de Shannon (H_M).	92
5.10	Ilustração de vetores de características para as imagens mostradas na Figura 5.1, considerando a configuração dos 36 atributos obtidos pela junção dos atributos do cálculo da entropia de Shannon (H_M) e os extraídos da curva.	92
5.11	Valores e médias de AUC calculados a partir de diferentes classificadores e de diferentes composições dos vetores de características obtidos pelo método proposto (H_M), considerando imagens coloridas.	93
5.12	Valores e médias de AUC calculados a partir de diferentes classificadores e de diferentes composições dos vetores de características obtidos pelo método multiescala tradicional da entropia de Shannon (H), considerando imagens em níveis de cinza.	93
5.13	Teste de Friedman para as combinações entre classificadores e técnicas, considerando taxas de AUC obtidas via vetores de características compostos por valores calculados das curvas.	97

5.14	Teste de Friedman para as combinações entre classificadores e técnicas, considerando taxas de AUC obtidas com vetores de características completos.	98
5.15	Identificação dos 10 atributos mais relevantes para distinguir os grupos de câncer colorretal e os valores obtidos de AUC, sensibilidade (Se) e especificidade (Es).	100
5.16	Desempenhos (valores de AUC) do método proposto (<i>SampEn_{MF}</i>) e de trabalhos correlatos, com informações sobre os atributos selecionados, classificadores e condições testadas entre os grupos avaliados. . .	102

Lista de Abreviações

ANOVA *Análise de Variância*

ApEn *Approximate Entropy*

AQUA *Automated Quantification Analyses*

AUC *Area Under Curve*

CCLSC *Color Component Level Spatial Correlation*

CMSE *Composite Multiscale Entropy*

DF *dimensão fractal*

DP *Desvio Padrão*

EPCO *Evaluation of Posterior Capsule Opacification*

Es *Especificidade*

FCE *Fuzzy Classification Entropy*

fn *falso negativo*

fp *falso positivo*

H&E *Hematoxylin&Eosin*

IFS *Intuitionistic Fuzzy Set*

LDF *Laser Doppler Fluxometry*

LoG *Laplaciano da Gaussiana*

LSCI *Laser Speckle Constrast Images*

MLP *Multilayer Perceptron*

MMSE *Multivariate Multiscale Entropy*

mRMR *minimum Redundancy Maximum Relevance*

MSC *Multiscale Complexity*

MSE *Multiscale Entropy*

PCO *Posterior Capsule Opacification*

RBF *Radial Basis Function*

RCMSE *Refined Composite Multiscale Entropy*

RGB *Red Green Blue*

RMSE *Refined Multiscale Entropy*

ROC *Receiver Operating Characteristic*

ROI *Region of Interest*

SampEn *Sample Entropy*

Se *Sensibilidadade*

SFFS *Sequential floating forward selection*

SFS *Sequential forward selection*

SMO *Sequential Minimal Optimization*

SNK *Student-Newman-Keuls*

SRM *Statistical Region Merging*

SVM *Support Vector Machine*

vn *verdadeiro negativo*

vp *verdadeiro positivo*

WSI *whole-slide images*

Sumário

Lista de Figuras	8
Lista de Tabelas	11
Lista de Abreviações	14
Sumário	17
1 Introdução	20
1.1 Motivação e justificativas	22
1.2 Objetivos	24
1.3 Organização do trabalho	25
2 Fundamentação Teórica	26
2.1 Entropia	26
2.1.1 Tipos de entropias: Aproximada, Amostral e associação <i>Fuzzy</i>	28
2.1.2 Abordagem multiescala	32
2.1.3 Abordagem multidimensional	34
2.2 Métricas de desempenho	35
2.3 Classificadores	36
2.4 Características do câncer colorretal	39
3 Trabalhos Relacionados	41

3.1	Aplicações dos diferentes modelos de entropias e comparações	41
3.2	Entropia com abordagens multiescalas	43
3.3	Entropia com estratégias <i>Fuzzy</i>	45
3.4	Entropia com abordagens multidimensionais	46
3.5	Extração de características em imagens colorretais	48
3.6	Considerações parciais	52
4	Metodologia	56
4.1	Estruturação do método e justificativas	56
4.2	Aplicação da abordagem multiescala	57
4.2.1	Definição da entropia Amostral com abordagem multiescala	57
4.2.2	Definição da entropia de Shannon com a abordagem multiescala	63
4.3	Definição da abordagem multidimensional	64
4.3.1	Definição da entropia Amostral com a abordagem multidimensional	64
4.3.2	Definição do modelo de entropia de Shannon com a abordagem multidimensional	65
4.4	Definição da abordagem <i>Fuzzy</i> para o modelo de entropia Amostral	67
4.5	Contexto de aplicação: imagens colorretais H&E	68
4.5.1	Definição do método considerando sub-imagens	69
4.6	Curvas de valores de entropia	70
4.6.1	Curvas de valores de entropia Amostral	71
4.6.2	Curvas de valores de entropia de Shannon	72
4.7	Extração de atributos da curva	73
4.8	Composição dos vetores de características	75
4.8.1	Estruturas dos vetores da entropia Amostral	75
4.8.2	Estruturas dos vetores da entropia de Shannon	76
4.9	Classificação dos padrões e obtenção das taxas de desempenho	77

4.10	Materiais	78
5	Resultados e Discussão	79
5.1	Resultados do método proposto para a entropia Amostral	79
5.2	Resultados do método proposto para a entropia de Shannon	89
5.3	Visão geral das diferenças entre os métodos	95
5.4	Desempenho da entropia Amostral frente aos trabalhos correlatos . . .	101
5.5	Desempenhos das técnicas: tempos estimados	103
6	Conclusões	105
6.1	Contribuições obtidas	106
6.2	Trabalhos futuros	107
	Referências Bibliográficas	109

Capítulo 1

Introdução

A quantificação de textura pode ser definida como um processo para representar numericamente as propriedades visuais contidas em uma imagem. Os valores podem definir conjuntos de elementos e são úteis para o reconhecimento de padrões (BENDER, 2003; OLIVEIRA; ASSIS; BRAGA, 2012). Os resultados podem ser modelos compreensíveis do domínio estudado para apoiar decisões.

A teoria da informação contribuiu significativamente com o desenvolvimento de abordagens para o processo de quantificação de imagens, principalmente com técnicas baseadas em entropia, aplicadas para quantificar sinais unidimensionais (DAMES; MATTOS; MELO, 2014; LI et al., 2013) e bidimensionais (em níveis de cinza) (SILVA et al., 2014; HUMEAU-HEURTIER et al., 2013; YEH; LIN; SHIEH, 2011). Esta teoria pode ser tratada a partir de diferentes fundamentações conceituais e é caracterizada pelo estudo de medidas de informação em problemas de telecomunicações, incluindo suas propriedades e aplicações (COVER; THOMAS, 2012). Um exemplo foi a fundamentação baseada em probabilidades, descrita por Claude Shannon em 1948 (SANTOS, 2013; FIGUEIREDO, 2011).

No contexto de imagens, a entropia pode ser entendida como uma medida de quantidade de informação e está diretamente ligada a distribuição ou arranjo da informação presente na intensidade luminosa de cada *pixel* (RONG, 2015). Para isso, a análise

pode considerar comparações entre *pixels* ou blocos de *pixels* para computar a irregularidade da quantidade de informação. Existem diferentes modelos de entropias para análises de imagens, sendo que as mais exploradas são as entropias de Shannon, Aproximada e Amostral (SILVA et al., 2014; HANMADLU et al., 2013). A entropia Aproximada foi aplicada com sucesso para avaliar informações sobre doenças (DAMES; MATTOS; MELO, 2014). A entropia Amostral é uma versão aprimorada da Aproximada, com aplicações para quantificar imagens médicas de linfomas (YEH; LIN; SHIEH, 2011) e analisar, com uma abordagem bidimensional, imagens do sistema nervoso de ratos (SILVA et al., 2016).

A entropia também foi associada com abordagens *Fuzzy* para minimizar a influência de ruídos em séries pequenas de dados (DAMES; MATTOS; MELO, 2014). Esta associação tem sido aplicada com sucesso em diferentes contextos, por exemplo na segmentação de imagens de lesões cerebrais (BIJAR; KHAYATI; BENAVENT, 2013) ou de imagens de núcleos de leucócitos (JATI et al., 2014), bem como na quantificação de imagens geológicas (HANI; SATHYAMOORTHY; ASIRVADAM, 2014) ou de imagens infravermelhas de diferentes contextos (YE et al., 2015).

Apesar do sucesso indicado em cada aplicação, um apontamento comum foi a necessidade de alterar as escalas de observações dos dados, visando minimizar os ruídos presentes nos dados ou contornar problemas envolvendo os tamanhos das séries. Diferentes abordagens multiescalas foram associadas com sucesso aos modelos de entropias, tais como o modelo multiescala denominado *Multiscale Entropy* (MSE), explorado por Humeau-Heurtier et al. (2013) e por Vandendriessche et al. (2014), o modelo *Multivariate Multiscale Entropy* (MMSE), discutido por Li et al. (2013), e a estratégia *Multiscale Complexity* (MSC), descrita por Yeh, Lin e Shieh (2011).

Além dos aspectos apresentados, outro avanço importante foi associar a entropia ao conceito multidimensional para potencializar o poder de quantificação de regiões de interesse de imagens coloridas (OSWAL et al., 2013). A ideia básica foi realizar

o processo de quantificação que considera janelas sobrepostas aos canais de cores da imagem e, então, verificar a similaridade dos *pixels* para construir um histograma tridimensional. Os autores estabeleceram correlações espaciais entre os componentes de cores para realizar segmentações e quantificações.

Apesar desta relevante contribuição da entropia, com diferentes modelos e aplicações, não foram encontrados trabalhos que exploraram modelos de entropia associados com abordagens multiescalas, multidimensionais e estratégias *Fuzzy*. Portanto, técnicas capazes de quantificar imagens com essa combinação podem proporcionar análises mais efetivas sobre as texturas de interesse em imagens coloridas, consequentemente favorecer a identificação e a classificação de padrões.

1.1 Motivação e justificativas

As diversas aplicações de entropia para avaliar dados representativos dos mais distintos contextos indicam sua importância no cenário científico, tais como aplicações envolvendo a segmentação de imagens em níveis de cinza e definição de estratégias *Fuzzy* (ANANTHI; BALASUBRAMANIAM; LIM, 2014), a análise de imagens geológicas coloridas (CONVERTINO et al., 2012), sinais fisiológicos causados por aplicação de anestesia (LIANG et al., 2015), investigação da massa cardíaca realizada por imagens de ultrassom (TSIAPARAS et al., 2012), aprimoramento da qualidade de imagens coloridas com técnicas *Fuzzy* (VERMA et al., 2012), sinais de curvas respiratórias (DAMES; MATTOS; MELO, 2014), comportamentos de mercado financeiro (RODRÍGUEZ; BÁRDOSSY, 2015) e de sinais mecânicos (XIONG et al., 2010).

Os estudos com entropia envolvem diferentes associações, tais como abordagens *Fuzzy* para o realce de imagens coloridas (HANMADLU et al., 2013) ou segmentação de imagens agrícolas (HUANG; HE; YANG, 2013); multiescalas, para quantificar linfomas em imagens médicas (YEH; LIN; SHIEH, 2011); e, multidimensional, para segmentar imagens (LU et al., 2014; RONG, 2015) ou para prover mecanismos de

segurança a partir de imagens digitais coloridas (SEYEDZADEH; NOROUZI; MIR-ZAKUCHAKI, 2014).

Apesar da relevância destacada, não foram encontrados métodos para quantificar imagens por meio de análises realizadas a partir da combinação de todas as abordagens apresentadas previamente: *Fuzzy*, multiescala e multidimensional. Esta associação pode aprimorar o processo de quantificação de regiões de interesse em imagens, com destaque para as imagens médicas. Nesta área de aplicação, a entropia indicou ser relevante para o estudo de padrões (DAMES; MATTOS; MELO, 2014; SILVA et al., 2016; JATI et al., 2014; YEH; LIN; SHIEH, 2011).

Neste contexto, a análise de imagens colorretais H&E para o entendimento do câncer colorretal pode contribuir com o desenvolvimento de técnicas direcionadas para o diagnóstico desta doença. É importante destacar que o câncer é um dos principais problemas de saúde pública em todo o mundo, sendo considerado a segunda principal causa de mortes nos Estados Unidos (SIEGEL; MILLER; JEMAL, 2018). Entre os tipos de câncer, o câncer colorretal é o terceiro mais comum, excluindo o câncer de pele, em homens e mulheres nos Estados Unidos. A estimativa é de 97.220 novos casos de câncer no cólon e de 43.030 novos casos de câncer retal, com uma estimativa de 50.630 óbitos para o ano de 2018, conforme dados apresentados por Siegel, Miller e Jemal (2018). No Brasil, de acordo com o Instituto Nacional de Câncer (INCA) (INCA, 2018), a estimativa para o ano de 2018 é de 36.360 novos casos, sendo 17.380 homens e 18.980 mulheres. O número de mortes registrado em 2013 foi de 15.415, com 7.387 homens e 8.024 mulheres.

O câncer colorretal é um tumor maligno que se desenvolve na parede interna do intestino grosso (cólon) ou reto (ALTERI; KRAMER; SIMPSON, 2014) e seu diagnóstico pode ser feito por sigmoidoscopia ou por colonoscopia com a confirmação por biópsia de tecido. Diversos métodos foram propostos para avaliação de imagens colorretais, explorando técnicas baseadas em matrizes de co-ocorrência (JØRGENSEN

et al., 2017; AKBAR; GOPI; BABU, 2015; IFTIKHAR; HASSAN; ALQUHAYZ, 2016) e combinações de abordagens estruturais, espectrais e de textura (FU et al., 2014). Também, quantificações de imagens colorretais H&E propostas em Jørgensen et al. (2017), Masood e Rajpoot (2009) e Altunbay et al. (2010) não exploraram as abordagens de forma associada, como a combinação proposta neste trabalho.

1.2 Objetivos

Neste trabalho, o objetivo principal é definir um novo método para quantificar imagens coloridas de câncer colorretal, a partir de modelos de entropias associados com estratégias *Fuzzy*, multiescala e multidimensional. Para isso, os objetivos específicos foram:

- I Apresentar as associações de estratégias multiescala, multidimensional e *Fuzzy* para compor novos métodos baseados nas Entropias de Shannon e Amostral, visando quantificar imagens histológicas colorretais H&E;
- II Testar o poder discriminativo de cada método em imagens histológicas colorretais H&E representativas dos grupos benigno e maligno;
- III Identificar o modelo de entropia mais relevante para quantificar e classificar os grupos indicados no item II;
- IV Comparar os desempenhos dos métodos com os fornecidos por modelos consolidados na literatura, entropias e técnicas fractais, bem como enquadrar os resultados frente aos indicados por trabalhos focados no estudo do câncer colorretal.

1.3 Organização do trabalho

Este trabalho foi organizado em seis capítulos: capítulo 2 com a fundamentação teórica para o entendimento dos conceitos sobre entropia e as abordagens utilizadas; capítulo 3 com detalhes sobre o levantamento bibliográfico realizado no contexto de entropia e aplicações; capítulo 4 contém a metodologia proposta para o desenvolvimento do método; capítulo 5 com os resultados obtidos a partir da aplicação do método: as curvas de entropia e os vetores de características, bem como a comparação com outras técnicas indicadas na literatura; capítulo 6 com as conclusões sobre o trabalho.

Capítulo 2

Fundamentação Teórica

Este capítulo tem como objetivo apresentar os conceitos necessários para a elucidação da metodologia proposta, tais como propriedades envolvendo entropia, abordagens multiescalas e multidimensional. Outros aspectos importantes e necessários para avaliação de atributos, tais como medidas de desempenho e classificadores, também são apresentados em detalhes.

2.1 Entropia

Entropia representa a irregularidade ou desordem de um sistema. Entretanto, nem sempre esse foi o sentido de entropia, que surgiu com as leis da termodinâmica para explicar as alterações de energia, massa e volume das moléculas em diferentes estados que um sistema pode assumir (PINEDA, 2006). Também, entropia era sinônimo de perda de calor em um sistema (ARTUSO, 2012). Com o passar do tempo, a entropia foi calculada estatisticamente por Boltzman, como exposto por Bawden e Robinson (2015) e por Pineda (2006), e formalizada por Claude Shannon (ARTUSO, 2012; BAWDEN; ROBINSON, 2015; PAVÃO, 2011).

De acordo com Shannon, a entropia é o grau de incerteza presente em um sistema, dado por uma distribuição de probabilidades de eventos i . A entropia permite quanti-

ficar essa incerteza. Com o aumento da probabilidade p_i presente no sistema, maior o nível de incerteza e maior o valor da entropia. A partir da formalização apresentada por Shannon, Equação 2.1, foi possível perceber que quanto maior o número de eventos ou estados possíveis, maior a incerteza presente no sistema. Esta abordagem permitiu a definição de complexidade ou desordem, sendo que a informação contida em menos eventos é mais fácil de ser representada ou quantificada.

$$H = \sum p_i \cdot \log \left(\frac{1}{p_i} \right). \quad (2.1)$$

Neste contexto, existem diferentes modelos de entropias que podem ser aplicados para o estudo de sistemas (SILVA et al., 2014; HANMADLU et al., 2013). Por exemplo, segundo Liang et al. (2015), a entropia de *wavelet* de Renyi permite analisar a quantidade de informação em diversos eventos com distintas probabilidades cada, no domínio de tempo-frequência. Há também variações na entropia de Tsallis, que analisa a quantidade de informação gerada nas ocorrências de eventos frequentes e raros nas séries. A entropia Aproximada surgiu com o intuito de analisar padrões em séries de dados unidimensionais e, assim, calcular os padrões similares. Entretanto, essa análise gera auto-similaridades (DAMES; MATTOS; MELO, 2014), fato este que não acontece na entropia Amostral, que é considerada como um refinamento da Aproximada (SILVA et al., 2014; XIONG et al., 2010). A partir dessas técnicas, outros modelos de entropias surgiram, tais como os que associam conceitos *Fuzzy*: o núcleo desses modelos está em verificar os graus de pertinência dos dados em relação as várias intensidades luminosas presente em uma imagem, por exemplo. Consequentemente, é possível calcular a entropia para cada *pixel* e obter um valor representativo para a imagem (sistema) (HUANG; HE; YANG, 2013).

2.1.1 Tipos de entropias: Aproximada, Amostral e associação *Fuzzy*

A entropia Aproximada é uma modalidade de entropia que visa considerar padrões de segmentos de séries¹ com determinada similaridade dentro de uma tolerância conhecida. Esses padrões são representados em um vetor de tamanho m , que é comparado com outro vetor $m + 1$. A medida gerada pela entropia é probabilística e conforme a complexidade da série varia, a entropia também varia (DAMES; MATTOS; MELO, 2014; LIANG et al., 2015; XIONG et al., 2010).

A representação dos vetores de tamanho m é dada pela Equação 2.2. A distância (d) entre os pares de vetores X_i e X_j , comparados nas posições i e j , com limites dados por $1 \leq i \leq S - m + 1$ e $1 \leq j \leq S - m + 1$, é dada pela Equação 2.3, na qual k representa o deslocamento de posições dos elementos dos vetores, S representa o tamanho da série, $s(i+k)$ e $s(j+k)$ representam os valores dos elementos nas diversas posições dos vetores X_i^m e X_j^m , respectivamente.

$$X_i^m = \{s(i), s(i+1), \dots, s(i+m-1)\}, \quad (2.2)$$

$$d_{ij} = d[X_i^m, X_j^m] = \max(|s(i+k) - s(j+k)|) \quad 0 \leq k \leq m-1. \quad (2.3)$$

Os valores das variáveis contidas nas séries podem sofrer alterações para diferentes tamanhos de padrões ou vetores (m), fato este não considerado no modelo de Shannon. Outro fator que torna a entropia Aproximada mais robusta é a tolerância (r), que é uma maneira de determinar a similaridade de um valor em comparação com os presentes em um vetor da série de tamanho S . No caso de uma imagem abstraída em uma matriz T , o tamanho total é dado por $N_{linha} \times N_{coluna}$, em que N_{linha} representa o número de linhas de T e N_{coluna} , o número de colunas de T . A tolerância pode definir se o valor de um *pixel*, por exemplo, pertence ao padrão tomado como referência. Desta forma,

¹Uma série representa um conjunto de valores obtidos (pontos) em um determinado período de tempo. Este termo é utilizado em contextos de sinais unidimensionais. (DAMES; MATTOS; MELO, 2014).

permite descartar elementos indesejáveis para compor o valor final da entropia. A entropia Aproximada (ApEn) pode ser calculada de acordo com a Equação 2.4:

$$ApEn(m,r,S) = \Phi_m(r) - \Phi_{m+1}(r). \quad (2.4)$$

O coeficiente de similaridade para o padrão de tamanho m é expresso por:

$$\Phi_m(r) = \frac{1}{S-m+1} \sum_{i=1}^{S-m+1} \ln[C_r^m(i)]. \quad (2.5)$$

Assim, $C_r^m(i)$ representa uma medida que expressa o valor normalizado da similaridade entre um vetor tomado como referência e todos os outros comparados e pode ser calculada por:

$$C_r^m(i) = \frac{1}{S-m+1} \sum_{j=1}^{S-m+1} \theta(d_{ij} - r). \quad (2.6)$$

A função degrau θ permite avaliar se a distância entre um vetor e outro está dentro de uma tolerância determinada, considerando:

$$\theta(d_{ij} - r) = \begin{cases} 1, & \text{se } d_{ij} - r \leq 0, \\ 0, & \text{se } d_{ij} - r > 0. \end{cases} \quad (2.7)$$

Pelas formulações expostas, nota-se que a principal característica da entropia Aproximada está em realizar a contagem da própria ocorrência (Equação 2.3) para evitar que o valor de $C_r^m(i)$ seja zero e garantir que o logaritmo da série sempre exista. Porém, este fato não atende o conceito básico da informação, tornando o resultado inconsistente e dependente do tamanho da série (DAMES; MATTOS; MELO, 2014; HUMEAU-HEURTIER et al., 2013; LIANG et al., 2015; XIONG et al., 2010). Uma alternativa é a entropia Amostral, tratada como um refinamento da entropia Aproximada: a similaridade entre os vetores comparados dentro de um padrão não considera a comparação dos vetores entre si, apenas com vetores distintos. Tal modificação faz

com que sejam eliminadas as auto-similaridades e redundância nas comparações entre os vetores. A entropia Amostral (SampEn) é calculada por:

$$SampEn(m,r,S) = \ln \varphi_m(r) - \ln \varphi_{m+1}(r). \quad (2.8)$$

Os coeficientes de similaridade para os padrões avaliados são dados pelas Equações 2.9 e 2.10:

$$\varphi_m(r) = \frac{1}{S-m} \sum_{i=1}^{S-m} C_r^m(i), \quad (2.9)$$

$$C_r^m(i) = \frac{1}{S-m-1} \sum_{j=1}^{S-m-1} \theta(d_{ij} - r). \quad (2.10)$$

Nos trabalhos de Dames, Mattos e Melo (2014), de Liang et al. (2015) e de Xiong et al. (2010) são destacadas questões envolvendo a função degrau θ , a qual é utilizada para os cálculos das entropias Aproximada e Amostral. A principal questão indica os problemas envolvendo o tratamento ou detecção de mudanças na complexidade da série analisada. Um desses problemas é apresentar valores indefinidos, comuns em séries pequenas, ou ainda sofrer variações elevadas pela presença de ruídos.

Para evitar esses problemas, nas abordagens apresentadas por Dames, Mattos e Melo (2014), por Liang et al. (2015) e por Xiong et al. (2010), a função degrau foi trocada por uma função *Fuzzy*. O objetivo foi definir o grau de pertinência da distância dos elementos em relação a tolerância r para, portanto, deixar os cálculos das entropias menos suscetíveis para computar ruídos e valores indefinidos. Isso ocorre pelo fato da função *Fuzzy* minimizar o problema da representação de similaridade em relação ao vetor e , com isso, evitar apenas valores zero ou um.

Na literatura é possível verificar diferentes tipos de funções *Fuzzy* (CHEN et al., 2009; DAMES; MATTOS; MELO, 2014; XIE; GUO; ZHENG, 2010; XIONG et al., 2010; ZHANG et al., 2010). A substituição da função degrau pela *Fuzzy* implica em alterar algumas etapas do cálculo da entropia.

A primeira alteração é a remoção da chamada linha base de cada vetor comparado da série em questão, como exposto na Equação 2.11, na qual $s0(i)$ representa a linha de base do vetor, que é substituída via Equação 2.12:

$$X_i^m = \{s(i), s(i+1), \dots, s(i+m-1)\} - s0(i) \quad 1 \leq i \leq S - m + 1, \quad (2.11)$$

$$s0(i) = \frac{1}{m} \sum_{j=0}^{m-1} s(i+j). \quad (2.12)$$

Tal mudança reflete no cálculo da distância entre os vetores, que deve ser computado por meio de:

$$d_{ij}^m = d[X_i^m, X_j^m] = \max(|(s(i+k) - s0(i)) - (s(j+k) - s0(j))|) \quad 0 \leq k \leq m-1. \quad (2.13)$$

Além disso, há ainda a mudança na própria função degrau, substituída por uma função *Fuzzy*. Um exemplo de função *Fuzzy* foi a descrita por Dames, Mattos e Melo (2014), conforme Equação 2.14. A principal característica dessa função é definir graus de pertinência de similaridade entre os pontos analisados com o comportamento de uma função gaussiana, que além de evitar valores indefinidos de entropia causados por ruídos, também aproveitam valores que apresentavam certa similaridade, mas que foram desprezados pela margem de tolerância na função degrau. Nesta formulação, $u(d_{ij}, r)$ representa o grau de pertinência da similaridade (calculada pela distância d_{ij}^m) entre os vetores comparados de um padrão analisado em relação a tolerância r considerada.

$$u(d_{ij}, r) = \exp\left(-\frac{d_{ij}^2}{r}\right). \quad (2.14)$$

A função descrita na Equação 2.14 assume valores $0 \leq u(d_{ij}, r) \leq 1$, na qual 0 representa a não-similaridade entre os vetores comparados e 1 representa a total similaridade entre os vetores comparados.

2.1.2 Abordagem multiescala

A entropia Multiescala (*Multiscale Entropy* - MSE) visa analisar objetos em diferentes escalas de observação e de obter diferentes quantificações de uma mesma imagem. Quando abordada em sinais 1-D, como em Humeau-Heurtier et al. (2013), em Verma, Sarangi e Kolekar (2014) e em Zhang et al. (2010). Neste caso, a abordagem multiescala expressa a reorganização dos pontos da série, de modo que granulações mais grossas sejam formadas, conseqüentemente, partes do sinal podem ser analisadas ao invés do sinal todo. Em imagens, os *pixels* podem formar granulações maiores e representar um objeto específico. Este processo ajuda em uma quantificação mais precisa e livre de ruídos presentes em outras partes da imagem.

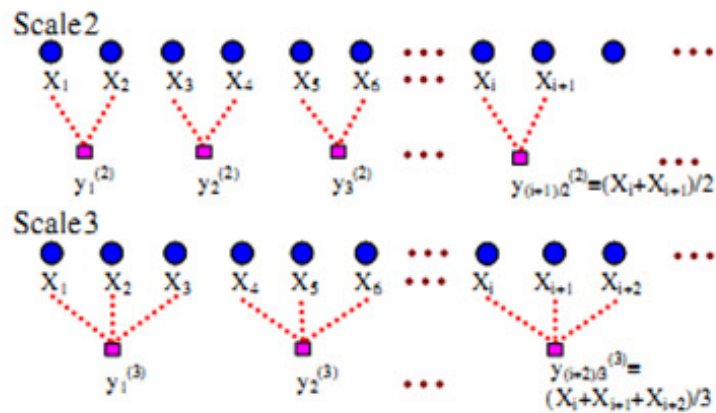
Verma, Sarangi e Kolekar (2014) descreveram que a granulação da abordagem multiescala pode ser construída a partir da Equação 2.15, no qual τ representa a escala considerada para a formação da granulação; Q_i , os pontos da série que compõe a granulação; e, $q_j^{(\tau)}$, o valor correspondente da granulação para dada escala τ .

$$q_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} Q_i. \quad (2.15)$$

Os autores observaram que a escala original é dada por $\tau = 1$ e, para variações de $\tau > 1$, os valores da série original são combinados para gerar a granulação desejada. Para $\tau = 2$, os valores da imagem são combinados dois a dois e geram um novo valor. Essas granulações são representadas por $1 \leq j \leq \frac{S}{\tau}$, no qual S representa o total de pontos em uma série. Essa abordagem é ilustrada por meio da Figura 2.1.

Neste contexto, a abordagem proposta por Yeh, Lin e Shieh (2011) nomeada de *Multiscale Complexity* (MSC) destaca-se por permitir observações importantes nas quantificações de imagens médicas que superam a MSE, com valores que expressam mais a distinção de tipos de linfomas encontrados em níveis sub-celulares, celulares e histológicos. Esta abordagem consistiu em adaptar a proposta multiescala, de séries

Figura 2.1: Representação do processo de granulação.



Fonte: (ZHANG et al., 2010).

unidimensionais para imagens, considerando que a tolerância r , a qual foi computada sobre a escala original da imagem, não foi mais aplicada no cálculo das granulações em todas as escalas. Sendo assim, a cada alteração de escala, uma nova tolerância r foi calculada e permitiu maior distinção entre os dados avaliados.

Outra abordagem que foi explorada em Da Silva et al. (2014) foi a variação do tamanho da janela, definida pelo parâmetro m , para aferir padrões similares presentes em imagens. Neste caso, foram obtidos diversos valores de entropia Amostral, denominada $SampEn_{2D}$, sendo que cada um foi associado ao tamanho de m e envolveu um conjunto de *pixels* contidos em uma janela de tamanho $m \times m$, com limites dados em $1 \leq m \leq N_{menor} - 1$. O valor de N_{menor} corresponde ao tamanho do menor lado da imagem.

Segundo os autores, esta abordagem permitiu determinar distinções mais efetivas entre grupos de imagens. Da Silva et al. (2014) indicaram que m pode variar entre 1 e 4 para a obtenção dos valores mais expressivos. Por fim, os autores concluíram que para valores altos de tolerância, a quantificação pode incluir ruídos.

2.1.3 Abordagem multidimensional

A abordagem multidimensional é definida como observações realizadas nas diferentes dimensões presentes no sistema dado como entrada. Para imagens, por exemplo, possíveis canais de cores são tratados como um espaço dimensional. Assim, uma imagem considerando o modelo de cores RGB é tratada como hipersuperfície (5D). Essa hipersuperfície é dada por $f(i,j) = (R,G,B)$, em que cada *pixel* pode ser visto como um vetor 5D (i,j,R,G,B) , sendo (R,G,B) os componentes de cor p de um *pixel* com coordenadas espaciais (i,j) , no espaço de cor RGB (IVANOVICI; RICHARD, 2011).

A distância entre os canais de cor, incluindo a informação espacial, utilizada por Ivanovici e Richard (2011) é a distância de Minkowski, representada na Equação 2.16. A distância de Minkowski é dada pela distância máxima obtida entre um *pixel* $F = f(i,j,R,G,B)$ e o *pixel* central da hipersuperfície $F_c = f_c(i_c,j_c,R_c,G_c,B_c)$ em todas as dimensões, sendo que f_p e f_{c_p} representam os valores dos *pixels* em cada canal de cor p ; L representa o tamanho do lado da hipersuperfície.

$$|F - F_c| = \max(|f_p - f_{c_p}|) \leq L \quad 1 \leq p \leq 5. \quad (2.16)$$

Esse fato expande o potencial de quantificações com modelos de entropia, pois permite considerar as características presentes em cada canal cor. Alguns trabalhos fazem uso desta abordagem no cálculo de entropia considerando imagens coloridas e em níveis de cinza (SILVA et al., 2014; DHIEB et al., 2014; OSWAL et al., 2013). Nas aplicações e testes realizados por esses autores, observou-se que as quantificações foram mais precisas e livres de fatores externos, pois os autores conseguiram utilizar técnicas especiais de histogramas para separar as informações de cada canal, tratando-os independentemente.

2.2 Métricas de desempenho

As métricas para aferir desempenhos de atributos ou sistemas são amplamente exploradas nos mais diferentes contextos (ANANTHI; BALASUBRAMANIAM; LIM, 2014; OLIVO; SANTIN; OLIVEIRA, 2013; YANG et al., 2015; ZOU et al., 2015). As métricas podem ser aplicada para confirmar uma hipótese, por exemplo. Na área de processamento de imagens ou reconhecimento de padrões, as métricas são extraídas a partir de uma rotulagem dos dados investigados. A rotulagem é realizada ao analisar, por meio de algoritmos de classificação de dados, os valores atribuídos aos conjuntos. É importante destacar que os valores atribuídos são definidos durante a etapa de quantificação de características, com técnicas específicas, como por exemplo, os diferentes modelos de entropias descritos previamente. Com isso, os rótulos atribuídos aos dados são: verdadeiro positivo (vp), falso positivo (fp), verdadeiro negativo (vn) e falso negativo (fn), com os seguintes significados:

- verdadeiro positivo (vp): se uma instância é positiva e é classificada como positiva;
- verdadeiro negativo (vn): se uma instância é negativa e é classificada como negativa;
- falso positivo (fp): se uma instância é negativa e é classificada como positiva;
- falso negativo (fn): se uma instância é positiva e é classificada como negativa.

Com esses rótulos, é possível indicar a sensibilidade e a especificidade dos sistemas. A sensibilidade (Se) indica a taxa de instâncias positivas em relação ao total de amostras positivas dada pela Equação 2.17 e a especificidade (Es) define a taxa de instâncias negativas em relação ao total de amostras negativas dada pela Equação 2.18.

$$Se = \frac{vp}{vp + fn}, \quad (2.17)$$

$$Es = \frac{vn}{vn + fp}. \quad (2.18)$$

Para essa rotulagem, uma técnica aplicada é a análise da curva ROC (*Receiver Operating Characteristic*), que é uma representação gráfica da relação entre sensibilidade (Se) e o complemento da especificidade (Es), dada por $1 - Es$ (OLIVO; SANTIN; OLIVEIRA, 2013). A representação dessa relação permite classificar uma variável, com uma decisão categórica. Este operador é comumente utilizado em análises de classificadores sob várias condições (YANG et al., 2015). Nesta relação, o ponto que mais se aproxima do valor $Se = 1$ e $1 - Es = 0$ indica o melhor ponto de corte para distinguir os grupos comparados. A área sob a curva ROC (*Area Under Curve* - AUC) indica um valor escalar entre as relações. Esta métrica é amplamente utilizada na área de inteligência artificial, pois representa com maior precisão a distinção entre classes (OSHIRO, 2013), principalmente se as classes estiverem desbalanceadas (YANG et al., 2015). A curva com maior área representa a melhor classificação: uma AUC=1 é dita como classificação ótima; e, uma AUC=0,5 é dita como não relevante.

2.3 Classificadores

Os classificadores são técnicas que podem ter a função de prever valores para variáveis categóricas discretas ou contínuas em um espaço amostral (ALVES et al., 2015; CARVALHO, 2015; LLADÓ et al., 2012; MARXER; PURWINS, 2015; PEREIRA, 2013; WANG; WANG; LIU, 2016; ZHOU et al., 2016). O desempenho de um classificador pode ser verificado por meio da AUC. Para isso, uma etapa importante é conhecer a capacidade de generalização do teste realizado. Essa capacidade é obtida com a validação cruzada dos dados analisados. A validação cruzada é aplicada para estimar o quão preciso é o modelo para um novo conjunto de dados. O núcleo desta validação é a divisão do conjunto de dados em subconjuntos mutualmente exclusivos, sendo alguns destes utilizados para treinamento e os demais para validação ou teste. A

divisão do conjunto de dados pode ser com as abordagens *holdout*, *leave-one-out* ou *k-fold*. Esta última é a mais aplicada (KOHAVI et al., 1995; LAUREANO; CAETANO; CORTEZ, 2014; SATO et al., 2013).

A validação cruzada por diversas pastas ou *k-fold cross-validation* utiliza vários agrupamentos ou pastas que separam as amostras em parcelas de dados iguais (CARVALHO, 2015). O intuito é que, independentemente da quantidade de pastas utilizadas, o classificador utilize uma destas pastas para testes e as demais para treinamento. Em seguida, o método adota uma segunda pasta para teste e as demais para treinamento, e assim sucessivamente, passando por todas as pastas (HORNUNG et al., 2014; TRIBA et al., 2015). Após todas as passagens, pode-se calcular a taxa média de AUC do sistema, visando definir uma medida mais confiável sobre a capacidade do modelo. A validação cruzada pode ser aplicada em diferentes classificadores, tais como *Support Vector Machine*, *Decision Tree*, *Rotation Forest* e outros.

Há diversas categorias de classificadores: baseados em árvores de decisão (CARVALHO, 2015), em funções (JUNIOR, 2010; PEREIRA, 2013; EKSTRÖM et al., 2018), *meta*-classificadores (RODRIGUEZ; KUNCHEVA; ALONSO, 2006), dentre outros.

O classificador *Support Vector Machine* (SVM) é baseado em função e consiste em encontrar um hiperplano ótimo capaz de separar classes linearmente, com a máxima distância entre elas. Para tanto, em muitas situações, os dados são transportados para altas dimensões (espaços de características) (JUNIOR, 2010). A função que separa as classes é conhecida como *kernel*, e pode assumir qualquer formato, como por exemplo, linear (LORENA; CARVALHO, 2003; PINEDA et al., 2015).

O MLP *Multilayer Perceptron* também se enquadra na categoria de funções e baseia-se em redes neurais artificiais que utiliza funções ponderadas em camadas em uma rede de *perceptrons* para comunicar com a próxima camada. Este tipo de classificador fornece resultados similares aos de métodos estatísticos, porém com entradas

não correlacionadas e com atribuições de pesos adequados aos dados, fato que proporciona classificações representativas (PEREIRA, 2013). Outros métodos baseados em funções são: *Radial Basis Function* (RBF) e *Logistic*, em que utilizam outras funções. Por exemplo, o RBF utiliza um *kernel* radial e não linear como o SVM (JUNIOR, 2010). O método *Logistic*, baseado em regressão logística, é uma técnica similar à regressão linear, mas quando a variável dependente é categórica ou binária. Nesta técnica uma função sigmoïdal é utilizada para separar grupos e indicar uma relação probabilística entre a variável dependente e as variáveis independentes do modelo por meio da máxima verossimilhança entre seus coeficientes (EKSTRÖM et al., 2018).

O classificador *Decision Tree* é baseado em árvores de decisão e adota uma estrutura de árvore, na qual os atributos, que representam as características extraídas de objetos, são os nodos da árvore, e as classes são as folhas. A classificação se dá por testes sobre vários “atributos-nodos”, da raiz em direção às folhas. Normalmente, os atributos são elencados com um cálculo de ganho de entropia, no qual o atributo que possui o maior valor de ganho é usado na raiz (CARVALHO, 2015). Se o “atributo-raiz” não for capaz de distinguir todas as classes então outros “atributos-nodos” serão necessários e serão elencados sob o mesmo critério de ganho de entropia do maior para o menor (CARVALHO, 2015; NEVES et al., 2014; PRADHAN, 2013; SATO et al., 2013).

Por fim, a categoria *meta* apresenta classificadores com a característica iterativa, em que modelos são gerados a partir das instâncias e os melhores conjuntos são escolhidos para a próxima iteração. Ao final das iterações obtém-se o melhor modelo. Os classificadores *Rotation Forest* e *Multiclass Classifier* se enquadram nesta categoria. O classificador *Rotation Forest* é um tipo de método combinado (*ensemble*), que consiste em dividir o conjunto de características em subconjuntos que serão submetidos à técnica PCA (RODRIGUEZ; KUNCHEVA; ALONSO, 2006), formando novos subconjuntos com base nos componentes principais extraídos de cada subconjunto ori-

ginal, considerando cada classificador do método combinado. Os novos subconjuntos são guardados em uma matriz diagonal com lado igual a quantidade de subconjuntos definidos. Com a matriz formada, árvores de decisão atuam em cada novo subconjunto em um processo de rotação dentro da matriz, mantendo todos os componentes principais para obter o melhor classificador. As árvores são utilizadas por serem mais sensíveis ao processo de rotação e por proporcionarem boas taxas de acurácia.

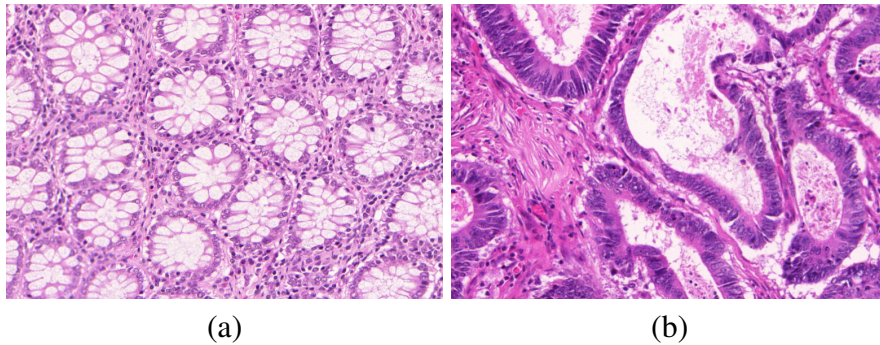
2.4 Características do câncer colorretal

O câncer colorretal é uma doença que se espalha pelo trato do intestino grosso (cólon) e reto. A mutação genética das células em regiões epiteliais do tecido pode provocar o surgimento de estruturas conhecidas como pólipos, porém nem todos os pólipos tem pré-disposição cancerígena (neoplasia) (COTTI et al., 2000). Os pólipos neoplásicos são conhecidos como adenomas, tumores benignos que podem evoluir para carcinomas (tumores malignos) (GRIGGS et al., 2017). Com o passar do tempo e o aumento das mutações, os adenomas tornam-se visíveis com as alterações histológicas bem acentuadas, permitindo identificar os estágios existentes na sequência adenoma-carcinoma (SOUZA et al., 2014). A intensificação da proliferação celular provoca também deformações das criptas, formato anormal do *lumen* e alterações da mucosa, característica que diferencia mais o adenoma do carcinoma (COTTI et al., 2000). Os principais sintomas decorrentes do câncer colorretal são sangramento retal, dores abdominais inferiores, perda de peso, mudança do hábito intestinal e até mesmo anemia.

A comprovação da doença somente pode ser realizada por meio de diagnósticos e exames médicos como a colonoscopia que, inclusive, possibilita a extração de pólipos. As extrações de estruturas maiores podem ser realizadas também via procedimento cirúrgico, como a videolaparoscopia. Os segmentos do intestino acometidos pela doença são comumente corados com hematoxilina & eosina (H&E) e analisados por patologistas, visando indicar o tipo e estágio do câncer. Na Figura 2.2 são apresen-

tadas exemplos de imagens colorretais H&E dos grupos benigno e maligno, fornecidas por (SIRINUKUNWATTANA et al., 2017). As principais características histológicas avaliadas pelos patologistas na detecção da doença são: localização do tumor, profundidade, número e estágio dos linfonodos, presença de metástases à distância, tipo histológico e diferenciação tumoral (NAHAS et al., 2015).

Figura 2.2: Exemplos de imagens das classes benigna (a) e maligna (b).



Fonte: (SIRINUKUNWATTANA et al., 2017).

Capítulo 3

Trabalhos Relacionados

Os modelos de entropias representam um papel importante em processamento de imagens, com aplicações para segmentar, quantificar, classificar e reconhecer padrões. Diversas aplicações de entropia são descritas neste capítulo, com indicações de associações com algumas abordagens, tais como *Fuzzy*, multiescala e multidimensional. Outro ponto importante do capítulo é a identificação de possíveis trabalhos que utilizam a entropia com o objetivo de quantificar imagens histológicas de câncer colorretal ou regiões de interesse das mesmas.

3.1 Aplicações dos diferentes modelos de entropias e comparações

As entropias de Shannon, Aproximada e Amostral são amplamente aplicadas para quantificações de sinais unidimensionais (1-D) (LIANG et al., 2015; NARDELLI et al., 2015; XIONG et al., 2010). Essas entropias são constantemente comparadas, como descrito por Xiong et al. (2010), os quais indicam que a entropia Amostral tem vantagens sobre a Aproximada, no que tange um número ótimo de escalas de observação. Os autores mencionam que a entropia Amostral é um refinamento da entropia Apro-

ximada, afirmação também relatada por Yeh, Lin e Shieh (2011) e confirmada por suas vantagens, principalmente em relação aos possíveis ruídos presentes nas fontes de dados (HUMEAU-HEURTIER et al., 2013; LI et al., 2013; VERMA; SARANGI; KOLEKAR, 2014; ZHANG et al., 2010).

No contexto de sinais bidimensionais, Convertino et al. (2012) compararam as entropias de Shannon e Kullback-Leibler (entropia relativa), a partir de experimentos com imagens multiespectrais RGB de ecossistemas. Essas entropias foram associadas a uma abordagem multiescala e os resultados indicaram que a entropia de Shannon foi mais relevante para distinguir as classes estudadas.

A informação espacial também foi considerada para o cálculo da entropia, conforme descrito em Lu et al. (2014). Os autores calcularam a entropia de Shannon da imagem completa e aplicaram, ainda, conjuntos *Fuzzy* para definir a pertinência dos *pixels*. Outro conceito explorado pelos autores foi o de relacionar os *pixels* com a menor intensidade luminosa da imagem. Com isso foi possível obter por meio da entropia de Shannon a medida de incerteza da maior intensidade e também da menor, nas quais favoreceram a separação de objetos (*pixels* mais claros na imagem) e fundo (*pixels* mais escura na imagem). Os autores concluíram que os resultados para os testes foram relevantes e que o método pode ser aplicado em diferentes tipos de imagens.

Da Silva et al. (2014) apresentaram uma abordagem da entropia Amostral para imagens em níveis de cinza e aplicaram a informação espacial entre as dimensões envolvidas. O cálculo de distância entre os padrões também foi realizado a partir de duas dimensões espaciais. Os autores testaram o método sobre imagens histológicas do sistema nervoso de dois ratos, um com idade de 30 dias e outro com 720 dias. Os valores da entropia Amostral foram comparados com índices referentes da área, sendo o perímetro, o diâmetro e a ocupação das fibras do nervo analisado. Os autores concluíram que o método proposto, denominado *SampEn_{2D}*, permitiu separar as classes estudadas e estabelecer uma correlação entre entropia e os comportamentos observados nos

tecidos.

3.2 Entropia com abordagens multiescalas

A abordagem multiescala foi associada aos diferentes modelos de entropias, com indicativos claros de que foi possível obter resultados mais relevantes para análises de texturas.

Boltz, Nielsen e Soatto (2010) analisaram um conjunto de 32 imagens, de texturas de diferentes contextos, que estão presentes em *Berkeley segmentation dataset*, com a entropia de Kullback-Leibler (entropia relativa) e o método fusão de regiões estatísticas, *Statistical Region Merging* (SRM). Os autores exploraram ainda valores de entropias mínima e máxima. Os autores concluíram que a entropia de Kullback-Leibler permitiu interpretações mais apropriadas com variações entre escalas, inclusive com resultados mais relevantes que os fornecidos pela entropia de Shannon.

Yeh, Lin e Shieh (2011) apresentaram um método de entropia, nomeado de complexidade multiescala, *Multiscale Complexity* (MSC). O objetivo foi quantificar a complexidade existente em 58 imagens de linfomas de vários tipos. Os autores utilizaram a métrica de curvas ROC para a metodologia proposta, a MSC, e realizaram o teste de análise de variância (ANOVA) para comparação entre classes. De maneira complementar, o método *Student-Newman-Keuls* (SNK) avaliou a MSC entre duas categorias apenas, já que algumas diferenças entre as categorias não puderam ser bem significativas com os testes anteriores. Como resultado, os autores puderam detectar variações de escalas entre 1 a 6 como as mais representativas avaliações sub-celulares, de 30 a 50 para avaliações celulares e de 50 a 70 para avaliações histológicas.

Humeau-Heurtier et al. (2013) destacaram a importância da abordagem multiescala associada ao estudo das entropias, pois de acordo com os autores esta abordagem permite a detecção de padrões e comportamentos fisiológicos de fluxo sanguíneo. Humeau-Heurtier et al. (2013) constataram que métodos para aplicações de *laser*,

que contrasta manchas na pele, chamados de *Laser Speckle Contrast Images* (LSCI), criado por um mapa de imagem, e *Laser Doppler Fluxometry* (LDF), que verifica a fluxometria, via *Doppler*, apresentaram comportamentos distintos quando analisados sob a óptica de *pixels* únicos. Quando analisados sob uma óptica de regiões de interesse, *Region of Interest* (ROI), considerando outras escalas de sinais, apresentaram comportamentos semelhantes. Os autores concluíram a partir deste fato que é possível substituir exames realizados com equipamentos LDF pelos de modelo LSCI.

Vardhan et al. (2013) realizaram um trabalho sobre entropia multiescala aliada às operações morfológicas para detecção de granulometria direcional, no qual foram aplicados elementos estruturantes, nas direções norte-sul, leste-oeste, nordeste-sudoeste e noroeste-sudeste. As imagens analisadas foram em níveis de cinza e com formas variadas, obtidas de bacias hidrográficas da Índia. Os autores concluíram, a partir dos valores gerados por meio da entropia de Shannon, que quanto maior a entropia em uma dada direção do elemento estruturante, melhor a detecção desta direção e do elemento correspondente na imagem. Os resultados foram detecções mais efetivas das regiões de interesse, quando comparados aos obtidos com outros modelos tradicionais existentes.

Vivekanand, Werghi e Al-Ahmad (2014) utilizaram uma metodologia baseada em entropia multiescala para quantificar áreas de uma doença ocular chamada Opacificação da Capsula Posterior ou *Posterior Capsule Opacification* (PCO), que pode se fazer presente em pacientes após cirurgia de catarata. Para o processamento das imagens, foram utilizados os métodos de segmentação de Otsu, devido a sua complexidade e viabilidade de implementação, e posteriormente, operadores morfológicos para eliminações de ruídos. O método dos autores envolveu ainda a aplicação de círculos concêntricos para detecção da rugosidade, desde o interior da pérola até a sua borda, adotando tamanhos de raios ótimos. O método foi comparado com vários outros métodos tradicionais como: análise clínica, *software* EPCO, *software* AQUA (entropia

global), filtragem de entropia local e matrizes de co-ocorrência em níveis de cinza. Após a análise estatística via correlação Pearson, foi constatado que o método proposto forneceu resultados inferiores somente na comparação com os resultados do *software* EPCO. Apesar disso, os autores destacaram que a proposta foi melhor por quantificar PCO leves, de maneira automatizada, com boa taxa de robustez em relação a variação da iluminação.

Humeau-Heurtier (2015) apresentou detalhes sobre as principais entropias (Aproximada e Amostral) e indicou tendências referentes para melhorias desses modelos. O autor descreve que a associação de conjuntos *Fuzzy* com *Multiscale Entropy* favorece quantificações mais precisas e menos influenciadas por ruídos. Especificamente, os autores destacaram várias adaptações, tais como entropia multiescala refinada (RMSE), entropia multiescala composta (CMSE) e entropia multiescala composta refinada (RCMSE) que consideram tais melhorias.

3.3 Entropia com estratégias *Fuzzy*

A abordagem *Fuzzy* foi amplamente explorada com diferentes modelos *Fuzzy*. No entanto, foi possível constatar como aplicação principal a de segmentar região de interesse em imagens em níveis de cinza. Neste contexto, por exemplo, Liu, Jiang e Feng (2006), exploraram a entropia denominada *Fuzzy Classification Entropy* (FCE), composta pela entropia de Kapur e as funções *Fuzzy*: *S-function*, triangular e de Huang e Wang (1995), para definir *pixels* como classes de objetos. Em Ananthi, Balasubramaniam e Lim (2014), conjuntos *Fuzzy* Intuitivos ou *Intuitionistic Fuzzy Set* (IFS) foram aplicados para expressar o nível de pertinência de um grupo ou área de *pixels*, bem como o nível de incerteza para segmentar as imagens de interesse.

A associação de entropia com conjuntos *Fuzzy* para quantificações de imagens foram observadas em Dames, Mattos e Melo (2014) e Hani, Sathyamoorthy e Asirvadam (2014).

Hani, Sathyamoorthy e Asirvadam (2014) propuseram um método para quantificar a incerteza de características de relevos montanhosos. A quantificação foi obtida por meio da entropia de Shannon sob análise das imagens em várias escalas e complexidades das características por meio de domínios *Fuzzy*. O método forneceu bons resultados na detecção dos relevos e baixa sensibilidade a ruídos. Porém, os autores concluíram que a abordagem falhou para quantificar características acentuadas de deformações nos relevos.

Dames, Mattos e Melo (2014) apresentaram uma metodologia baseada em entropia Aproximada *Fuzzy* para quantificar representações de fluxos respiratórios de pacientes saudáveis e com um tipo de doença pulmonar crônica. O método consistiu em avaliar a variação da tolerância para a entropia e relacionar com os índices característicos dos grupos de pacientes. Os resultados mostraram alta correlação entre os valores do atributo e os índices analisados, o que permitiu a caracterização e reconhecimento da doença. Os autores concluíram que o método foi eficiente principalmente quando foram considerados valores baixos de tolerância para constituir os atributos.

3.4 Entropia com abordagens multidimensionais

A abordagem multidimensional consiste em observações realizadas nas diferentes dimensões presentes na imagem que é dada como entrada. Assim, os canais de cores são tratados como um espaço dimensional. Essa estratégia foi explorada principalmente para segmentação de imagens (ALARCÓN; MARROQUÍN, 2009; DHIEB et al., 2014; HANMANDLU; JHA, 2006; HANMADLU et al., 2013; OSWAL et al., 2013; XU et al., 2012).

Em Dhieb et al. (2014) e Rong (2015) a abordagem adotada para definir a entropia de Shannon foi considerar apenas os *pixels* pertencentes às regiões contempladas em um histograma 2-D. O histograma foi construído para separar uma imagem em quatro regiões. O cálculo da entropia propôs a separação entre objetos e fundo de maneira

a descartar regiões que apresentavam bordas ou ruídos. O método foi testado em diferentes imagens para obtenção dos melhores limiares para segmentação e obteve êxito com um custo computacional reduzido.

O núcleo básico desses trabalhos é o mesmo dos discutidos em Cheng e Li (2003), que contribuiu com a evolução da abordagem multidimensional por desenvolver um método capaz de calcular a entropia em cada canal de cor de uma imagem, no modelo de cores RGB. O interessante foi observar que tal abordagem foi constantemente relacionada com estratégias *Fuzzy*, principalmente para definir a qual classe um *pixel* pertencia.

Cheng e Li (2003) utilizaram a entropia de Shannon para quantificar regiões homogêneas de um objeto com estratégias *Fuzzy* e definir o melhor valor para um processo de segmentação. O processo foi realizado em imagens de diversos tipos e tamanhos e os autores constataram que além da redução do número de cores avaliadas, informações importantes das imagens eram preservadas, o que não ocorria com métodos que não apresentavam estratégias *Fuzzy*.

O conceito multidimensional foi mais bem explorado em Ivanovici e Richard (2011), trabalho em que os autores apresentaram uma proposta multidimensional como uma adaptação do método clássico de matriz de probabilidade de Voss. O método foi testado com associação da dimensão fractal (DF) e lacunaridade para analisar diferentes tipos de imagens coloridas. Os autores representaram um hiper-cubo (5-D) dentro do espaço de cores RGB para avaliar a relação de um *pixel* neste hiperespaço. O *pixel* neste espaço multidimensional foi representado por um vetor $F = f(x,y,R,G,B)$ e foi comparado com o *pixel* central do hiper-cubo, representado por um vetor $F_c = f_c(x_c,y_c,R_c,G_c,B_c)$. Se a distância entre F e F_c foi menor ou igual ao lado do hiper-cubo (L), então o *pixel* foi considerado na quantificação da dimensão fractal ou da lacunaridade. Os resultados observados no experimento com algumas imagens apontaram que o método era promissor.

A abordagem multidimensional e entropia para quantificação foi formalizada em Oswal et al. (2013). Os autores desenvolveram um método baseado em entropia de Shannon para segmentar e quantificar núcleos de células contidas em imagens coloridas, padrão RGB. O método baseado em entropia consistiu em segmentar regiões a partir de histogramas nomeados *Color Component Level Spatial Correlation* (CCLSC) e aplicar filtros para eliminação de ruídos, decorrentes da separação das regiões de interesse. A quantificação, por sua vez, foi realizada com a contagem de núcleos de cada componente de cor. No experimento foram analisadas 21 imagens histológicas de células cancerígenas. A análise se deu por meio da comparação dos resultados obtidos pelo método proposto e, manualmente, por um especialista. As métricas de desempenho utilizadas apontaram bons resultados para o método.

3.5 Extração de características em imagens colorretais

Altunbay et al. (2010) contribuíram significativamente para a quantificação e extração de características em imagens histológicas de cólon. Em seu método os autores desenvolveram um grafo colorido para representar os tipos de estruturas do tecido como estroma, núcleo e *lumen*. As arestas também possuíam cores que representavam os tipos de estruturas ligadas. A triangulação de *Delaunay* foi utilizada para a construção do grafo que permitia extrair características do tipo grau médio, coeficiente médio de agrupamento e diâmetro. Foram consideradas 213 imagens para o experimento e o método foi comparado à outras técnicas conhecidas na literatura como as baseadas em intensidade, em texturas e estruturais. Os resultados foram satisfatórios para a proposta dos autores com uma taxa de acurácia de 82,65%, aplicando o classificador SVM na separação dos grupos do tipo normal, maligno de baixa graduação e de alta graduação.

Kalkan et al. (2012) desenvolveram em seu trabalho um método para separação de grupos de imagens de cólon. As imagens H&E foram divididas em sub-imagens e alo-

cadadas por um especialista em 4 grupos. Primeiramente as imagens foram segmentadas pelo algoritmo *k-means clustering* com o principal foco na obtenção do núcleo celular, que foi detectado com auxílio de filtros como o Laplaciano da Gaussiana (LoG) . As imagens foram transformadas em níveis de cinza para que as características estruturais pudessem ser extraídas, tais como: número de núcleos por área do tecido; razão entre cada estrutura em relação à área do tecido; e razão entre as estruturas nas sub-áreas. Tais medidas permitiram obter alguns descritores globais. A extração também se deu no aspecto da textura, em que foram obtidas características por meio de descritores de Haralick e por filtros de Gabor. Em seguida, o total de características foi reduzido de 1127 para 120, devido a correlação entre algumas características. A distância de Mahalanobis foi utilizada como critério de seleção. Na última etapa, por decisões de trabalho, os autores preferiram confrontar os 4 grupos como sendo 2 grandes grupos: normal e anormal; câncer e não-câncer. Para a classificação foi aplicado o classificador *Logistic Regression*. Os resultados foram de 87,67% de acurácia média e 0,95 de AUC para a melhor combinação entre grupos (câncer e não-câncer).

Fu et al. (2014) propuseram um método capaz de extrair e selecionar as principais características de imagens obtidas por colonoscopia. O método foi dividido nas etapas de melhoramento da imagem, extração e classificação de padrões. Para a primeira etapa foi utilizada a transformação dos componentes principais para que os especialistas pudessem selecionar as áreas de interesse. Após a seleção das áreas, que tinham tamanho de 24 x 24 *pixels*, foram extraídas 180 características de cada área. As características extraídas foram dos tipos estruturais (variância de intensidade e de energia), espectrais (entropia) e de texturas (matrizes de co-ocorrência). Na etapa de seleção de características foram considerados dois métodos: o SFS (*Sequential forward selection*) e o SFFS (*Sequential floating forward selection*). Na etapa de classificação foi aplicado o algoritmo SVM que permitiu obter taxas de AUC de 0,887 para o conjunto de 180 características, enquanto que com o conjunto reduzido para 77 características

(pelo método SFS), as taxas foram de 0,913 e para o conjunto com apenas 47 características (pelo método SFFS), as taxas foram de 0,931. A acurácia obtida pelo classificador SVM atingiu 96% contra apenas 85% obtida com a classificação manual feita por especialistas.

Akbar, Gopi e Babu (2015) descreveram um método voltado para o auxílio da detecção de câncer de cólon. Foram investigadas 113 imagens divididas em 64 casos do tipo maligno e 49 do tipo benigno. A extração de características proposta pelo método foi baseada em estruturas de grafos, que representavam os atributos das imagens. Os nós e arestas dos grafos simbolizavam a distribuição espacial das imagens. Para facilitar o processo, os objetos foram definidos como sendo núcleo e não-núcleo. Após a construção de cada grafo, as consultas para extração das características foram feitas por buscas em profundidade, sendo que cada sub-grafo correspondia às características de determinada região de interesse. Da mesma maneira foram utilizados descritores de Haralick para extrair características estatísticas das mesmas regiões. Os classificadores testados foram *Multilayer Perceptron*, *KStar*, *Bayesian Logistic Regression* e *SVM*, com seus algoritmos disponibilizados na ferramenta Weka. Foram obtidas taxas de acurácia para o desempenho dos classificadores escolhidos com valor de 83,33% para o melhor classificador (*Multilayer Perceptron*).

Nayiar, Asim e Shahid (2015) propuseram em seu trabalho um método que permite diferenciar os grupos de imagens de cólon. Os autores reuniram um conjunto de 174 imagens dos tipos benigno (82 casos) e maligno (92 casos) para o experimento. O método consistia de pré-processamento, que era responsável pela separação das partes celulares (*lumens* e células epiteliais) por meio do algoritmo de *clustering*, já que as regiões de interesse eram os *lumens*. Outra etapa foi a de extração de características, obtidas por operadores morfológicos, tais como: área, excentricidade, compacidade, número de Euler, entre outros descritores. Além dos morfológicos também foram utilizados descritos de Fourier elípticos devido o formato dos *lumens*, bem como a

distância entre os quadrados mínimos dos *lumens*. Havia outra etapa, após a extração de características, que abordava a separação das imagens para treinamento e teste por meio da técnica *Jack-knife cross-validation*, considerando *10-folds*. Desta forma o classificador SVM foi aplicado com alguns tipos de funções, a saber: linear, sigmoidal e RBF. Os parâmetros destas funções também foram variados. Os melhores resultados foram de 96,55% de acurácia e 0,96 de AUC, considerando a combinação do vetor de características híbrido com a melhor configuração dos parâmetros das funções (escolhida via votação majoritária).

Iftikhar, Hassan e Alquhayz (2016) desenvolveram um método para trabalhar com imagens de câncer de cólon obtidas por meio de biópsia com o intuito de separar as classes de imagens: saudável, moderado e com câncer. O método foi organizado nas etapas de pré-processamento, extração de características, seleção de características e classificação de um conjunto de 92 imagens de cólon com resolução de 600 x 800 *pixels*. Na etapa inicial as imagens foram transformadas em níveis de cinza com a melhoria do contraste. Na próxima etapa foram extraídas as suas características por meio de técnicas baseadas em momentos estatísticos e descritores de Haralick. Depois, técnicas de balanceamento das amostras foram utilizadas para só então aplicar a seleção de características, que foi realizada por meio da mRMR (minimum Redundancy Maximum Relevance), técnica que seleciona as características mais relevantes eliminando as que tem efeito duplicado. No final, as taxas de acurácia foram obtidas a partir da aplicação do classificador SVM com validação cruzada do tipo *leave-one-out* e seu valor foi de 95,4%.

Jørgensen et al. (2017) descreveram um método capaz de distinguir grupos de imagens de câncer de cólon H&E do tipo benigno e maligno. O método primeiramente realizou a segmentação do núcleo celular (região de interesse) por meio de técnicas de agrupamento e deconvolução dos canais de cores RGB para avaliação do corante utilizado. Nesta etapa ainda foram utilizadas técnicas de limiarização e, na etapa seguinte,

foi realizada a extração de características, obtidas por meio das métricas de intensidade mínima, média, máxima dos canais de cores, matrizes de co-ocorrência e da matriz *run-length*. A seguir, a seleção de características foi adotada aplicando o classificador *Random Forest* com configurações específicas para estimar as principais características. Foram selecionadas 9 das 750 características e, por fim, realizada a classificação de um total de 87 imagens (44 benignas e 43 malignas) também aplicando o classificador *Random Forest*. As taxas de AUC e de acurácia obtidas foram de 0,96 e 91%, respectivamente.

3.6 Considerações parciais

As entropias são medidas que representam a quantidade de incerteza dos dados analisados e sua utilização no meio científico tem se mostrado importante (ANANTHI; BALASUBRAMANIAM; LIM, 2014; HUMEAU-HEURTIER et al., 2013; LI et al., 2013; VERMA; SARANGI; KOLEKAR, 2014; ZHANG et al., 2010). Pelos trabalhos analisados, foi possível verificar que existe uma relação entre valores de entropia e complexidade do objeto que foi explorado, sejam em contextos unidimensionais ou bidimensionais (DAMES; MATTOS; MELO, 2014; SILVA et al., 2014; XIONG et al., 2010). Em processamento de imagens, os diferentes modelos de entropias foram aplicados com sucesso para segmentar ou quantificar sinais (DHIEB et al., 2014; HANI; SATHYAMOORTHY; ASIRVADAM, 2014; VARDHAN et al., 2013; VIVEKANAND; WERGHI; AL-AHMAD, 2014). Entretanto, a entropia de Shannon (CONVERTINO et al., 2012; RODRÍGUEZ; BÁRDOSSY, 2015) e a entropia Amos-tral (LI et al., 2013; NARDELLI et al., 2015) merecem destaque, principalmente por conseguir quantificar a complexidade e a similaridade das características de um objeto de interesse.

Outro aspecto observado foi a associação sistemática de modelos de entropias com estratégias *Fuzzy*, multiescalas ou multidimensional, separadamente, com o propósito

de aprimorar os processos de quantificações de elementos, nos mais diferentes domínios (DAMES; MATTOS; MELO, 2014; SILVA et al., 2014; HANMADLU et al., 2013; LU et al., 2014; OSWAL et al., 2013). A aplicação da abordagem *Fuzzy* foi justificada por permitir reduções de ruídos nas fontes de dados analisados (DAMES; MATTOS; MELO, 2014; HANI; SATHYAMOORTHY; ASIRVADAM, 2014). Por exemplo, a entropia de Shannon com uma abordagem denominada *Fuzzy 2-D Entropy* foi descrita por Huang, He e Yang (2013) com o intuito de melhorar a segmentação de objetos. A entropia Amostral foi explorada com um modelo nomeado *Fuzzy SampEn*, conforme descrito por Xiong et al. (2010): o objetivo foi tornar o cálculo da similaridade de regiões menos sensível a ruídos. Pelos trabalhos pesquisados, foi possível concluir que a associação de modelos de entropias, principalmente de Shannon e Amostral, com funções *Fuzzy* permitiu melhorar os desempenhos de tarefas de segmentação (ANANTHI; BALASUBRAMANIAM; LIM, 2014; CHENG; LI, 2003; LIU; JIANG; FENG, 2006) ou de quantificação (DAMES; MATTOS; MELO, 2014; HANI; SATHYAMOORTHY; ASIRVADAM, 2014; XIONG et al., 2010).

No que tange entropia com abordagens multiescalas, foi possível constatar que observações dos elementos em imagens e em várias escalas propiciam quantificações mais precisas, fato que favorece processos de segmentação (BOLTZ; NIELSEN; SOATTO, 2010), caracterização (SILVA et al., 2014; VARDHAN et al., 2013; VIVEKANAND; WERGHI; AL-AHMAD, 2014; XIONG et al., 2010), classificação e reconhecimento de padrões (HUMEAU-HEURTIER et al., 2013; YEH; LIN; SHIEH, 2011). Para tanto, foram observadas diferentes estratégias multiescalas, tais como granulação de *pixels* (HUMEAU-HEURTIER et al., 2013; XIONG et al., 2010), variações dos tamanhos (m) das janelas e tolerância (r) (SILVA et al., 2016; YEH; LIN; SHIEH, 2011). Cada abordagem tem vantagens e desvantagens, mas variações dos tamanhos das janelas e da tolerância mostraram-se mais utilizados e vantajosos em processos de quantificações (SILVA et al., 2016; YEH; LIN; SHIEH, 2011). Essas abordagens fo-

ram exploradas principalmente em estimativas com a entropia Amostral (SILVA et al., 2016; XIONG et al., 2010; YEH; LIN; SHIEH, 2011). Para a entropia de Shannon, as abordagens multiescalas que proporcionaram melhores resultados foram associadas com variações dos tamanhos de janelas quadradas, as quais se expandem gradativamente até atingir uma área de cobertura apropriada na imagem (CONVERTINO et al., 2012).

Quando a abordagem multidimensional é considerada, a conclusão foi que possíveis propriedades presentes individualmente em cada canal de cor podem compor os valores finais das entropias, conseqüentemente também favorecendo a caracterização, classificação e reconhecimento dos padrões investigados (SILVA et al., 2014). A abordagem multidimensional mais explorada com entropia foi a que envolveu canais de cores analisados isoladamente (DHIEB et al., 2014; RONG, 2015). No entanto, há também explorações em imagens coloridas envolvendo todos os canais como a abordagem denominada *Color Component Level Spatial Correlation* (OSWAL et al., 2013). Outras abordagens multidimensionais também foram encontradas, com resultados relevantes quando associadas aos atributos dimensão fractal e lacunaridade (IVANOVICI; RICHARD, 2011), mas não explorada com o conceito de entropia. Nesse caso, por exemplo, a estratégia descrita por Ivanovici e Richard (2011) aponta que a pertinência do *pixel* no hiper-cubo pode ser confirmada pela distância de Minkowski entre os *pixels*.

Apesar da relevância reconhecida de cada estratégia, não foram encontrados trabalhos que associaram propriedades *Fuzzy*, multiescalas e multidimensional com as entropias de Shannon e Amostral. Assim, um método contendo estas características pode contribuir significativamente com a área de processamento de imagens de diversos contextos, tais como o de imagens de câncer colorretal, que não possuem associações com os modelos de entropia (FU et al., 2014; IFTIKHAR; HASSAN; ALQUHAYZ, 2016; JØRGENSEN et al., 2017) e tem sua importância reconhecida na

literatura (KALKAN et al., 2012; NAIYAR; ASIM; SHAHID, 2015; AKBAR; GOPI; BABU, 2015). Desta forma, o método pode permitir quantificações mais livres de ruídos e fatores externos, por avaliar os *pixels* de modo combinado e não isolados na imagem, o que pode favorecer o processo de classificação e reconhecimento de padrões. Além disso, pode proporcionar observações das propriedades em todos os canais da imagem e, possivelmente, escolher qual contém a informação mais significativa para ser computada.

Um resumo dos principais trabalhos focados na classificação de câncer colorretal com as principais técnicas adotadas, bem como os tipos de imagens utilizadas e seus melhores desempenhos são apresentados na Tabela 3.1.

Tabela 3.1: Resumo dos principais trabalhos correlatos, com informações sobre as imagens, grupos investigados, características e métodos de classificação.

Métodos	Base de Imagens	Extração de características	Classificador	Acurácia
Esgiar et al. (1998)	102 imagens (44 benigno, 58 maligno), Níveis de cinza	Matriz de co-ocorrência	<i>K-nearest neighbor</i>	90,20%
Esgiar et al. (2002)	102 imagens (44 benigno, 58 maligno), Níveis de cinza	Matriz de co-ocorrência e dimensão fractal	<i>K-nearest neighbor</i>	94,10%
Masood et al. (2006)	Imagens hiperespectrais	Matriz de co-ocorrência e Características morfológicas	SVM	90,00%
Masood, Rajpoot (2009)	32 Imagens hiperespectrais	Características LBP circulares	SVM	90,60%
Altunbay et al. (2010)	213 imagens histológicas coloridas	Grafos coloridos	SVM	82,65%
Kalkan et al. (2012)	2000 amostras (benigno, maligno, adenomatoso e inflamado)	Características de Haralick e Gabor	Regressão Logística	87,67%
Fu et al. (2014)	365 imagens (214 com pólipos hiperplásico e 151 com pólipos adenomatoso)	Matriz de co-ocorrência, variância de intensidade e de energia e entropia	SVM	96,00%
Akbar, Gopi e Babi (2015)	113 imagens (49 benigno, 64 maligno)	Características estruturais e Haralick	<i>Multilayer Perceptron</i>	83,33%
Naiyar, Asim e Shahid (2015)	174 imagens (82 benigno, 92 maligno)	<i>Least square lumen distance</i> , descritores elípticos de Fourier e características morfológicas	SVM	96,55%
Iftikhar, Hassan and Alquhayz (2016)	92 imagens (pobre, moderado e saudável)	Haralick e momentos estatísticos	SVM	96,13%
Jorgensen et al. (2017)	87 imagens (44 benigno, 43 maligno)	Características de intensidade, matrizes de co-ocorrência, matriz {run-length} e características estatísticas de primeira ordem	<i>Random Forest</i>	91,00%

Capítulo 4

Metodologia

Neste capítulo é apresentada uma descrição detalhada sobre o método proposto para quantificar imagens coloridas, a partir de modelos de entropias combinados com estratégias *Fuzzy*, multiescalas e multidimensional. Detalhes sobre a modelagem matemática que constitui o método são apresentados nas próximas seções, com o objetivo de associar modelos de entropias Amostral e Shannon com cada uma das estratégias. Além disso, são apresentados os tratamentos dados às curvas obtidas para composição dos vetores de características.

4.1 Estruturação do método e justificativas

O método foi estruturado em três etapas. A primeira etapa consistiu em escolher a abordagem multiescala para as entropias de Shannon e Amostral. Essas entropias foram escolhidas a partir de apontamentos existentes na literatura pesquisada, o que indicou que esses modelos são os mais relevantes para quantificações de imagens (BALASUBRAMANIAN; NAIR; NAGARAJ, 2015; CONVERTINO et al., 2012; HUMEAU-HEURTIER et al., 2013; LIANG et al., 2015; XIONG et al., 2010). A abordagem multiescala foi definida com variações dos tamanhos das janelas, g (Shannon) e m (Amostral), bem como com alterações da tolerância r , somente para a en-

tropia Amostral. A segunda etapa permitiu aplicar o conceito multidimensional, com objetivo de avaliar os componentes de cores mais relevantes da imagem. O modelo escolhido foi o descrito por Ivanovici e Richard (2011). Na terceira etapa foi considerada uma função *Fuzzy*, porém aplicada para a entropia Amostral. A função *Fuzzy* explorada foi amplamente discutida em diversos trabalhos no contexto de sinais fisiológicos (DAMES; MATTOS; MELO, 2014; XIE; GUO; ZHENG, 2010) e mecânicos (XIONG et al., 2010).

O resultado obtido com a aplicação do método foi um conjunto de atributos que define uma curva de entropia para cada imagem. A partir de cada curva foi possível obter os vetores de características. Os vetores se tornaram entradas para diferentes classificadores. A relevância e a robustez do método foram apresentadas a partir de taxas de AUC e testes estatísticos. Um resumo do método está na Figura 4.1. Nas próximas seções cada etapa foi detalhada apropriadamente.

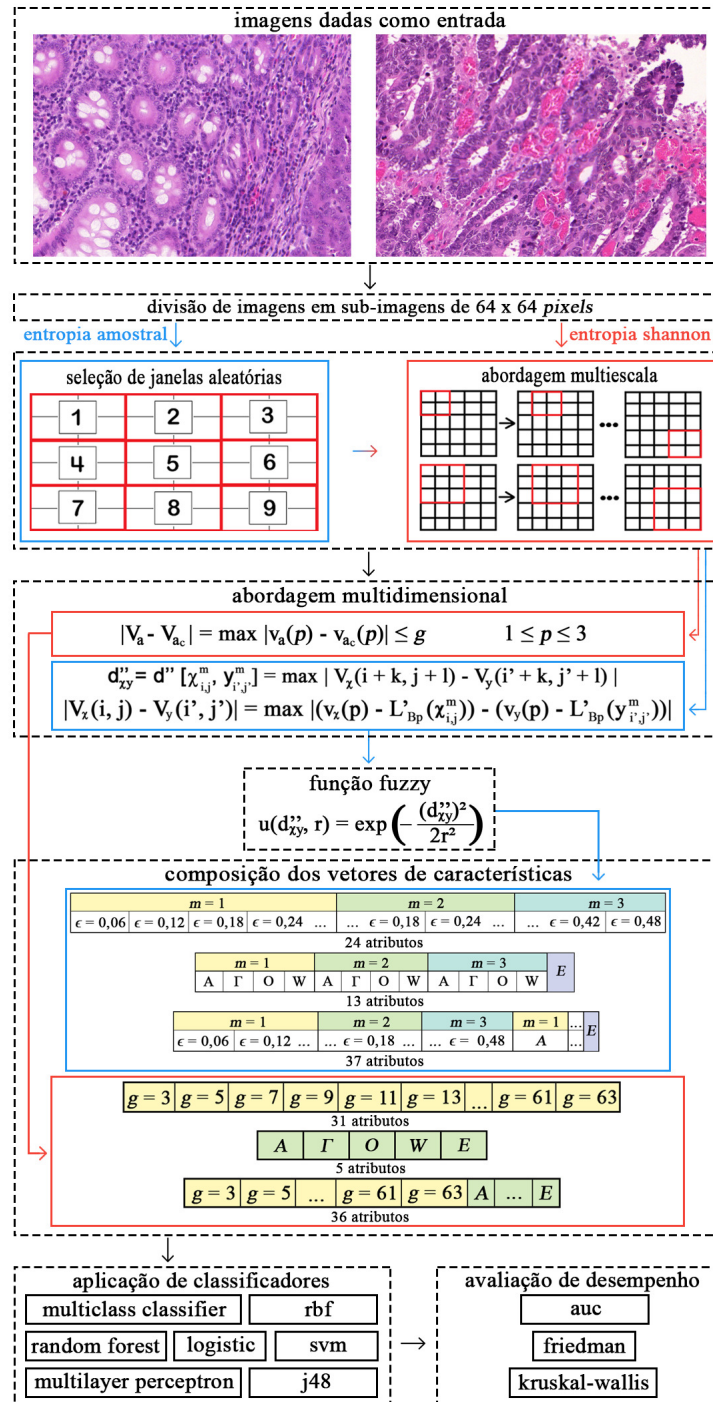
4.2 Aplicação da abordagem multiescala

A primeira etapa consistiu em combinar a abordagem multiescala com cada modelo de entropia. Para isso, a estratégia foi variar o tamanho dos padrões comparados para computar o valor de similaridade entre *pixels* (SILVA et al., 2014).

4.2.1 Definição da entropia Amostral com abordagem multiescala

A primeira etapa consistiu em escolher uma abordagem multiescala para o modelo de entropia Amostral. A estratégia definida foi variar o tamanho dos padrões comparados para computar o valor de similaridade entre *pixels*, procedimento reconhecidamente aplicado em diferentes estudos (YEH; LIN; SHIEH, 2011; HUMEAU-HEURTIER et al., 2013; SILVA et al., 2014; XIONG et al., 2010). A entropia Amostral foi calculada com base no modelo apresentado por Da Silva et al. (2014), a *SampEn_{2D}*, consi-

Figura 4.1: Ilustração das etapas do método proposto.



Fonte: Elaborado pelo autor.

derando janelas (submatrizes) para as comparações dos padrões. Portanto, dada uma imagem, abstraída em uma matriz bidimensional T de tamanho $N_{img} = N_{linha} \times N_{coluna}$, é possível definir janelas $x_{i,j}^m$ que representam padrões de tamanho $m \times m$:

$$x_{i,j}^m = \{t(i,j), t(i,j+1), \dots, t(i+1,j), \dots, t(i+m-1, j+m-1)\}, \quad (4.1)$$

em que $1 \leq i \leq N_{linha} - m + 1$ e $1 \leq j \leq N_{coluna} - m + 1$.

Cada janela $x_{i,j}^m$ é fixada em uma posição (i,j) e comparadas com janelas $y_{i',j'}^m$, definidas de maneira similar ao apresentado na Equação 4.1, porém posicionadas (i',j') por toda a extensão da imagem e respeitando a condição de que $(i,j) \neq (i',j')$.

A similaridade entre dois padrões $x_{i,j}^m$ e $y_{i',j'}^m$ pode ser determinada a partir de uma função degrau da forma $\theta(d_{xy} - r)$, com o objetivo de avaliar se a distância d_{xy} entre os padrões é menor ou igual a uma tolerância r predefinida (SILVA et al., 2014). A distância d_{xy} trata-se da máxima diferença de intensidade luminosa entre os *pixels* correspondentes das janelas avaliadas no processo. Entretanto, no método proposto foi aplicada uma função *Fuzzy u* para permitir analisar a similaridade entre padrões e minimizar a presença de ruído no processo de quantificação. Este aprimoramento está descrito em detalhes na seção 4.4, Equação 4.15. Com a definição de u houve a necessidade de subtrair as linhas de base das janelas. Uma linha de base $L_B(x_{i,j}^m)$ de uma janela $x_{i,j}^m$ é definida pela média aritmética dos *pixels* contidos em $x_{i,j}^m$:

$$L_B(x_{i,j}^m) = \frac{1}{m^2} \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} t(i+k, j+l). \quad (4.2)$$

O resultado foi uma nova distância d'_{xy} para atender estes requisitos, conforme Equação 4.3:

$$d'_{xy} = d'[x_{i,j}^m, y_{i',j'}^m] = \max |(t(i+k, j+l) - L_B(x_{i,j}^m)) - (t(i'+k, j'+l) - L_B(y_{i',j'}^m))|, \quad (4.3)$$

em que $0 \leq k, l \leq m - 1$.

A distância d'_{xy} entre duas janelas consiste em comparar os valores de intensidades luminosas dos *pixels* correspondentes. Entretanto, para imagens coloridas, as intensidades luminosas constituem um vetor de cores, definidos a partir do modelo de cor da imagem analisada. Neste trabalho, as imagens colorretais foram adquiridas no padrão RGB (SIRINUKUNWATTANA et al., 2017). Portanto, o cálculo da distância entre as janelas com vetores de intensidades luminosas foi obtido a partir da Equação 4.3 expandida para o contexto multidimensional (d''_{xy}), descrita detalhadamente na seção 4.3 e representada via Equação 4.12.

A probabilidade de uma janela $y_{i',j'}^m$ ser similar a uma janela fixa $x_{i,j}^m$ é determinada a partir da média da contagem de todas as janelas $y_{i',j'}^m$ que estão a uma distância de no máximo r de $x_{i,j}^m$, expressa a partir da Equação 4.4:

$$C_m(x_{i,j}^m) = \frac{1}{Z-1} \sum_{i'=1}^{N_{linha}-m} \sum_{j'=1}^{N_{coluna}-m} u(d''[x_{i,j}^m, y_{i',j'}^m], r), \quad (4.4)$$

com $(i, j) \neq (i', j')$, $Z = (N_{linha} - m)(N_{coluna} - m)$ e a função *Fuzzy* $u(d''[x_{i,j}^m, y_{i',j'}^m], r)$, descrita na seção 4.4 via Equação 4.15.

A entropia Amostral ($SampEn_{2D}$) (SILVA et al., 2016) pode ser calculada pelo logaritmo natural negativo da razão entre coeficiente de similaridade de todos os padrões analisados de tamanho $(m+1) \times (m+1)$ e o coeficiente de similaridade referente aos padrões $m \times m$ (SILVA et al., 2016):

$$SampEn_{2D}(m,r,N_{img}) = -\ln \frac{\Phi_{m+1}}{\Phi_m}, \quad (4.5)$$

em que Φ_{m+1} e Φ_m são os coeficientes de similaridade definidos respectivamente por:

$$\Phi_{m+1} = \frac{1}{Z} \sum_{i=1}^{N_{linha}-(m+1)} \sum_{j=1}^{N_{coluna}-(m+1)} C_{m+1}(x_{i,j}^{m+1}), \quad (4.6)$$

e

$$\Phi_m = \frac{1}{Z} \sum_{i=1}^{N_{linha}-m} \sum_{j=1}^{N_{coluna}-m} C_m(x_{i,j}^m). \quad (4.7)$$

No entanto, na abordagem *Fuzzy*, implementada neste trabalho, as pertinências quantificadas para Φ_{m+1} podem ser maiores do que para Φ_m , resultando em um valor de entropia negativo na Equação 4.5. Portanto, o cálculo da $SampEn_{2D}$ (SILVA et al., 2016) foi modificado para acomodar a estratégia *Fuzzy* utilizada e manter as propriedades da entropia como uma medida positiva. A entropia Amostral Multidimensional e *Fuzzy* ($SampEn_{MF}$) foi dada por:

$$SampEn_{MF}(m,r,N_{img}) = \left| \ln \frac{\Phi_{m+1}}{\Phi_m} \right|. \quad (4.8)$$

A partir do modelo apresentado, a abordagem multiescala foi obtida com variações nas dimensões m das janelas e da tolerância r . A tolerância foi computada com variações dos valores definidos por ε , conforme Equação 4.9.

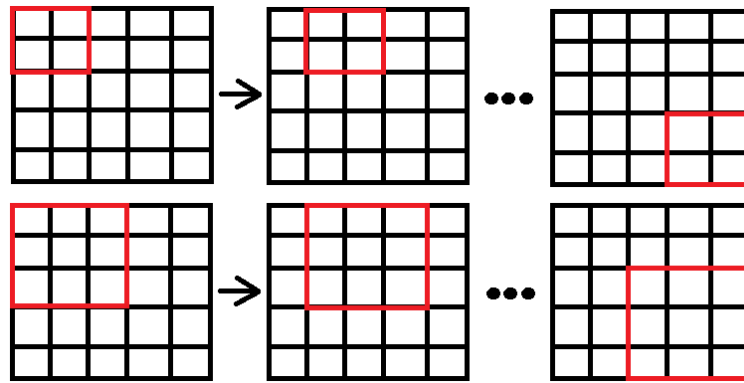
$$r = \varepsilon \times \sigma(T), \quad (4.9)$$

em que $\sigma(T)$ representa o desvio padrão na imagem T .

É importante ressaltar que o intervalo de ε foi dado por $0,06 \leq \varepsilon \leq 0,5$, com variações de 0,06. Para as dimensões das janelas, o valor de m foi limitado em $1 \leq m \leq 3$, com incrementos de uma unidade a cada mudança de escala e com $m \in \mathbb{N}$. Uma ilustração de janelas definidas a partir de diferentes tamanhos m está na Figura 4.2. Os

valores escolhidos foram indicados como os mais relevantes (SILVA et al., 2016).

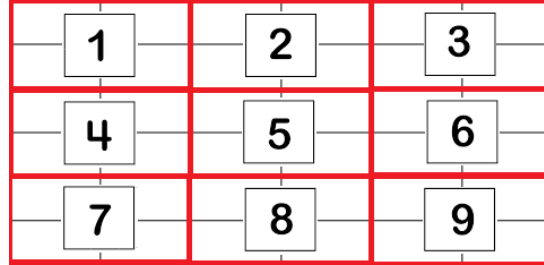
Figura 4.2: Ilustração do deslocamento de janelas que representam os padrões de tamanho $m \times m$, de cor vermelha, realizado em imagens, considerando as escalas $m = 2$ e $m = 3$.



Fonte: Elaborado pelo autor.

O processo de quantificação por meio da entropia Amostral não depende de um número elevado de janelas e escolher janelas aleatórias pode ser um procedimento aceitável para calcular o valor de entropia de uma imagem (YEH; LIN; SHIEH, 2011). No modelo proposto, cada imagem colorretal foi dividida em regiões de tamanho 64×64 pixels, totalizando 96 sub-imagens analisadas. Cada sub-imagem foi dividida em 9 sub-regiões e, em cada uma das sub-imagens, 90 janelas foram definidas aleatoriamente e distribuídas pelas sub-regiões de forma balanceada e sem sobreposições. Esta estratégia foi escolhida considerando procedimentos similares existentes na literatura (SILVA et al., 2014; YEH; LIN; SHIEH, 2011), com o objetivo de permitir uma análise mais homogênea da imagem. O valor final de entropia para cada imagem foi definido como o valor médio entre suas sub-imagens. Uma ilustração do modelo utilizado para distribuir as janelas está na Figura 4.3.

Figura 4.3: Ilustração do modelo utilizado para definir sub-regiões no processo de seleção de janelas aleatórias.



Fonte: Elaborado pelo autor.

4.2.2 Definição da entropia de Shannon com a abordagem multi-escala

A entropia de Shannon (H) pode ser calculada para uma dada imagem, também abstraída em uma matriz bidimensional T . Para o cálculo foi definida uma janela (submatriz) de tamanho $g \times g$, representada via Equação 4.10, em que $1 \leq i \leq N_{linha} - g + 1$ e $1 \leq j \leq N_{coluna} - g + 1$ representam as posições da janela $a_{i,j}^g$.

$$a_{i,j}^g = t(i+k, j+l) \quad 0 \leq k, l \leq g-1. \quad (4.10)$$

A variação multiescala foi em função do parâmetro g , com limites $3 \leq g \leq N_{menor}$ e incrementos em duas unidades, em que N_{menor} corresponde ao menor lado da imagem T . O valor inicial $g = 3$ foi necessário para definir um *pixel* central em cada janela $a_{i,j}^g$. O *pixel* central é utilizado como referência no processo de comparação multidimensional (IVANOVICI; RICHARD, 2011), descrito na seção 4.3.

Para uma dada escala g , a entropia proposta H_M de uma janela $a_{i,j}^g$ é calculada para cada posição (i, j) em que se encontra na imagem T . Dada uma posição (i, j) da janela $a_{i,j}^g$, uma nova janela $b_{i,j}^g$ é obtida a partir das informações multidimensionais da janela $a_{i,j}^g$, descrita em mais detalhes na seção 4.3. A entropia H_M é calculada a partir da entropia de Shannon indicada na Equação 2.1 para cada janela $b_{i,j}^g$. A adaptação do cálculo de entropia para toda a imagem, $H_M(T)$, descrito por Cheng e Li (2003), com

a abordagem proposta neste trabalho, em que os parâmetros N_{linha} e N_{coluna} definem as dimensões da imagem T e g indica o tamanho de cada janela, foi dada por:

$$H_M(T) = \frac{1}{(N_{linha} - g + 1)(N_{coluna} - g + 1)} \sum_{i=1}^{N_{linha}-g+1} \sum_{j=1}^{N_{coluna}-g+1} H_M(b_{i,j}^g). \quad (4.11)$$

O processo do cálculo é repetido para todas as variações de g .

4.3 Definição da abordagem multidimensional

A segunda etapa consistiu em associar a abordagem multidimensional aos modelos de entropias. Para cada janela definida nas entropias Amostral ou de Shannon, o componente de cor mais relevante foi obtido por meio da abordagem discutida em Ivanovici e Richard (2011). Assim, cada *pixel* das janelas foi projetado no espaço de cor RGB e seu formato neste espaço foi definido para cada modelo de entropia a seguir.

4.3.1 Definição da entropia Amostral com a abordagem multidimensional

A abordagem multidimensional foi aplicada sobre pares de *pixels* correspondentes das janelas $x_{i,j}^m$ e $y_{i',j'}^m$, com o objetivo de computar as possíveis similaridades existentes entre as intensidades luminosas. Para atender o cálculo da entropia Amostral, cada *pixel* contido em $x_{i,j}^m$ e $y_{i',j'}^m$ foi representado por vetores n-dimensionais tais como $V_x(i,j) = v_x(R_x, G_x, B_x)$ e $V_y(i',j') = v_y(R_y, G_y, B_y)$, respectivamente. Estes vetores permitem indicar as intensidades luminosas presentes nos canais R, G e B dos pixels analisados, considerando o espaço de representação RGB. As comparações entre vetores $V_y(i',j')$ e $V_x(i,j)$ foram realizadas por meio da distância de Minkowski, descrita na Equação 4.12, com o objetivo de verificar a máxima diferença entre as cores de

cada canal RGB:

$$d''_{xy} = d'' [x_{i,j}^m, y_{i',j'}^m] = \max |V_x(i+k, j+l) - V_y(i'+k, j'+l)|, \quad (4.12)$$

em que $0 \leq k, l \leq m-1$ e $|V_x(i, j) - V_y(i', j')|$ a diferença entre os valores de intensidades dos canais de cores que estão sob análise:

$$|V_x(i, j) - V_y(i', j')| = \max |(v_x(p) - L'_{B_p}(x_{i,j}^m)) - (v_y(p) - L'_{B_p}(y_{i',j'}^m))| \quad 1 \leq p \leq 3, \quad (4.13)$$

em que $L'_{B_p}(x_{i,j}^m)$ e $L'_{B_p}(y_{i',j'}^m)$ são as linhas de base referentes ao canal p de $x_{i,j}^m$ e $y_{i',j'}^m$, respectivamente.

O nível de pertinência da distância d''_{xy} para definir a entropia foi dada por meio de uma função *Fuzzy*. A pertinência é avaliada pela distância entre d''_{xy} e o parâmetro r , por meio de uma função gaussiana. Este processo é apresentado na seção 4.4.

4.3.2 Definição do modelo de entropia de Shannon com a abordagem multidimensional

Para calcular a entropia de Shannon de uma janela $a_{i,j}^g$, cada *pixel* foi representado por vetores n -dimensionais do tipo $V_a = v_a(i, j, R_a, G_a, B_a)$. O *pixel* central desta janela foi definido como $V_{a_c} = v_{a_c}(i_c, j_c, R_{a_c}, G_{a_c}, B_{a_c})$. De acordo com a abordagem multiescala proposta, a entropia de Shannon foi computada para cada janela $a_{i,j}^g$ posicionada em (i, j) na imagem T . O cálculo realizado em cada janela depende de informações multidimensionais obtidas de acordo com a proposta de Ivanovici e Richard (2011). Neste caso, todos os *pixels* de uma janela em questão foram definidos como vetores no espaço RGB. A partir disso, o *pixel* V_{a_c} foi comparado aos demais da mesma janela pela distância de Minkowski, considerando o tamanho da janela g como limite para a

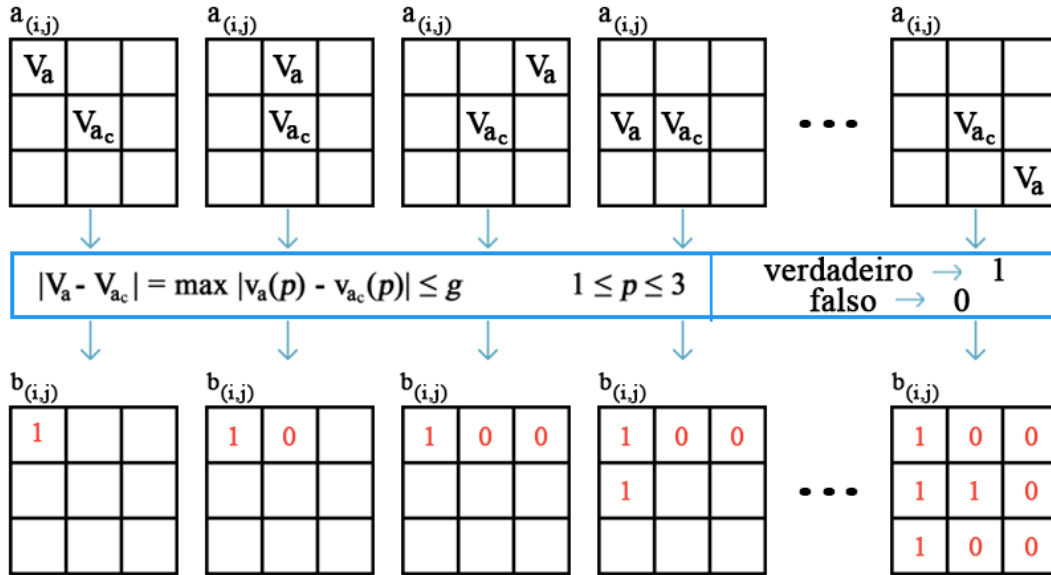
função estabelecida na Equação 4.14.

$$|V_a - V_{a_c}| = \max |v_a(p) - v_{a_c}(p)| \leq g \quad 1 \leq p \leq 3, \quad (4.14)$$

em que p corresponde a cada canal dos vetores V_a e V_{a_c} da janela $a_{i,j}^g$.

Após a aplicação da função na janela $a_{i,j}^g$ dada como entrada obteve-se uma nova janela $b_{i,j}^g$, de mesmo tamanho, como saída. Os *pixels* desta nova janela foram rotulados com valores 0 ou 1. Estes valores representaram se os *pixels* da janela dada como entrada estavam (valor 1) ou não (valor 0) dentro do limite g da função. Os rótulos foram escolhidos como símbolos para a entropia de Shannon e substituíram as intensidades dos *pixels*. A partir desta nova janela $b_{i,j}^g$, a entropia de Shannon foi calculada baseada nestes dois valores (0 e 1), os quais permitem indicar a variação de frequência das intensidades dos canais de cores da janela original $a_{i,j}^g$. Assim, o cálculo da entropia de Shannon $H_M(b_{i,j}^g)$ para cada janela $b_{i,j}^g$ foi dada a partir destas distribuições de frequência. O processo de comparação e geração da nova janela $b_{i,j}^g$ com seus rótulos (0 ou 1) é apresentado na Figura 4.4.

Figura 4.4: Ilustração do processo de rotulação de uma janela $b_{i,j}^g$ a partir da aplicação da função de Minkowski na janela $a_{i,j}^g$.



Fonte: Elaborado pelo autor.

4.4 Definição da abordagem *Fuzzy* para o modelo de entropia Amostral

O objetivo na terceira etapa foi verificar as pertinências de distâncias entre janelas, por meio da estratégia *Fuzzy*, substituindo a abordagem binária implementada com a função degrau $\theta(d_{xy} - r)$, presente na proposta de Da Silva et al. (2016). Na abordagem binária a pertinência é contabilizada como 1 quando uma janela $y_{i',j'}^m$ é considerada similar (condição definida por $d_{xy} - r \leq 0$) à uma janela $x_{i,j}^m$ fixa e 0 quando não há similaridade (condição indicada por $d_{xy} - r > 0$). Na abordagem *Fuzzy* é empregado o conceito de pertinência, em que é possível obter diferentes níveis de similaridades entre duas janelas. A pertinência foi avaliada por meio de uma função $u(d_{xy}^'', r)$ que associou a distância entre duas janelas à um número real no intervalo $[0,1]$, sendo 1 o grau máximo de pertinência (totalmente similar) e 0 o grau mínimo de pertinência (não similar). Esta função foi amplamente discutida por Xiong et al. (2010) e por Xie et al. (2010) para calcular a entropia Amostral no contexto de sinais unidimensionais

e foi concluído que a abordagem *Fuzzy* pode minimizar a influência de ruídos.

Portanto, neste trabalho foi realizada a expansão da função *Fuzzy*, utilizada no contexto unidimensional (XIONG et al., 2010; XIE; GUO; ZHENG, 2010), para o contexto bidimensional. No modelo proposto, o objetivo foi avaliar a similaridade de *pixels* próximos, por meio de uma função gaussiana, como definido na Equação 4.15:

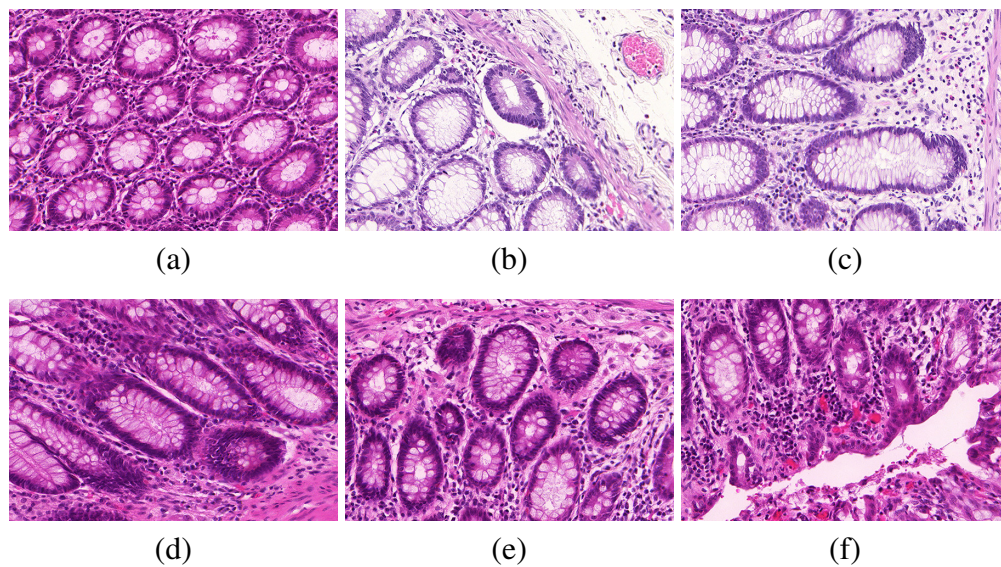
$$u(d''_{xy}, r) = \exp\left(-\frac{(d''_{xy})^2}{2r^2}\right), \quad (4.15)$$

em que d''_{xy} é a distância entre duas janelas descrita na Equação 4.12 e r o valor de tolerância.

4.5 Contexto de aplicação: imagens colorretais H&E

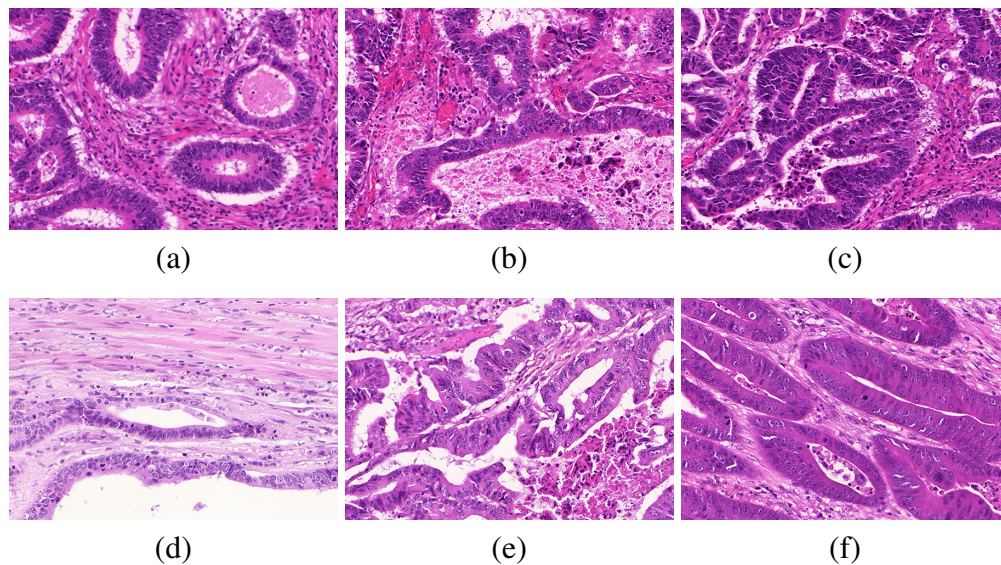
Os testes foram realizados a partir de um banco de imagens disponibilizado por (SIRINUKUNWATTANA et al., 2017). O banco foi constituído por imagens histológicas derivadas de 16 seções coradas em H&E (*Hematoxylin&Eosin*) do estágio T3 ou T4 de câncer colorretal. Cada seção pertencia a um paciente, das 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 um resolução de *pixel* $0,465\mu\text{m}$. As WSIs foram reescaladas para uma resolução de *pixel* de $0,620\mu\text{m}$, 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 Figuras 4.5 e 4.6 são apresentados exemplos de imagens dos grupos benigno e maligno, respectivamente.

Figura 4.5: Exemplos de imagens histológicas colorretais do grupo benigno.



Fonte: (SIRINUKUNWATTANA et al., 2017).

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



Fonte: (SIRINUKUNWATTANA et al., 2017).

4.5.1 Definição do método considerando sub-imagens

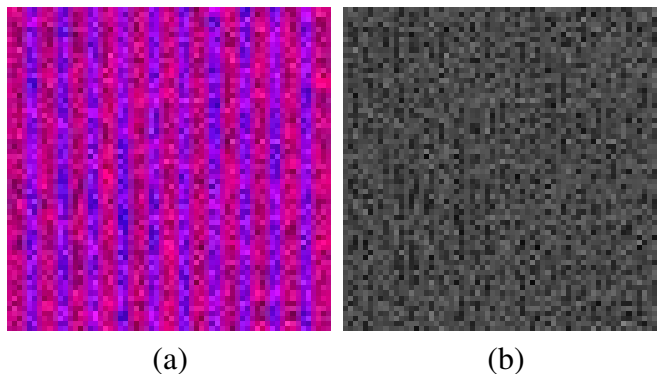
As imagens colorretais testadas de tamanho 775×522 pixels foram divididas em regiões de tamanho 64×64 pixels, totalizando 96 sub-imagens, uma estratégia utilizada de forma empírica para permitir a aplicação dos modelos. Esta estratégia foi baseada

em indicações descritas no trabalho de Da Silva et al. (2014), que utilizou em seus testes sub-imagens de tamanho 128×128 *pixels* com a finalidade de melhorar o tempo de processamento quando comparado ao de imagens com tamanhos maiores. A entropia em cada sub-imagem foi calculada aplicando os métodos descritos previamente. O valor final de entropia para cada imagem foi definido como o valor médio entre suas respectivas sub-imagens.

4.6 Curvas de valores de entropia

Os resultados obtidos com as entropias foram curvas de valores para cada imagem colorretal. As curvas foram geradas em função de variações dos parâmetros requeridos em cada método. Para a entropia Amostral ($SampEn_{MF}$), as curvas foram definidas por meio da aplicação da abordagem multiescala, variações sobre os parâmetros ϵ e m , fornecendo três curvas independentes, uma para cada valor de m (de 1 a 3), porém variando ϵ , de $0,06 \leq \epsilon \leq 0,5$, com incrementos de 0,06. No caso da entropia de Shannon (H_M), cada curva foi dada pela variação do parâmetro g (de 3 a 63, considerando os tamanhos das sub-imagens utilizadas). Para exemplificar, duas imagens controladas foram geradas com os mesmos padrões, porém uma colorida (Figura 4.7a) no padrão RGB e a correspondente em níveis de cinza (Figura 4.7b). As imagens possuem dimensões de 256×256 *pixels*. Na imagem colorida, Figura 4.7a, os padrões de cores são identificados pela alternância, linhas verticais, das cores azul e magenta. A imagem colorida foi convertida para níveis de cinza (Figura 4.7b), na qual é possível observar que os padrões foram comprometidos pela restrição do espaço de representação. Os resultados são irregularidades nas variações das intensidades dos *pixels*. Estas características das imagens permitem fornecer as assinaturas dos padrões a partir dos modelos propostos.

Figura 4.7: Imagens geradas a partir do *software* Matlab R2012b, (a) imagem colorida (a) e a correspondente em níveis de cinza (b), para exemplificar as quantificações realizadas com os modelos de entropias propostos.

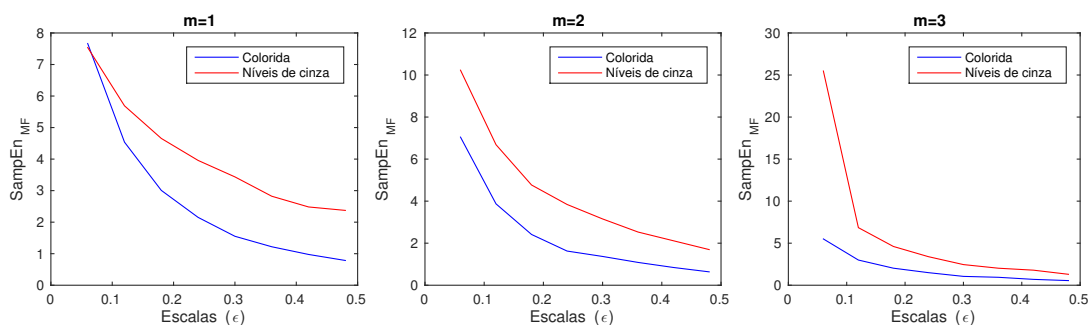


Fonte: Elaborado pelo autor.

4.6.1 Curvas de valores de entropia Amostral

Para a entropia Amostral ($SampEn_{MF}$) foram extraídas 3 curvas, uma para cada variação de m em cada imagem de controle, representadas na Figura 4.8.

Figura 4.8: Curvas de entropias Amostral ($SampEn_{MF}$) das imagens de controle, com escalas variando de $m = 1$ a $m = 3$.



Fonte: Elaborado pelo autor.

Em cada uma das curvas foi possível observar que o valor de entropia ($SampEn_{MF}$) da imagem colorida é menor do que o valor da imagem em níveis de cinza. Este fato ocorre devido a identificação dos padrões. Na imagem colorida os padrões identificados forneceram uma assinatura de homogeneidade. Os valores de entropia Amostral tendem a 0 quando a técnica é aplicada em uma imagem que apresenta estruturas homogêneas. Na imagem em níveis de cinza os padrões não foram identificados, fato que

representa a ausência de estruturas homogêneas. Para este caso, o valor de entropia tende a ser maior do que o valor calculado para a imagem colorida.

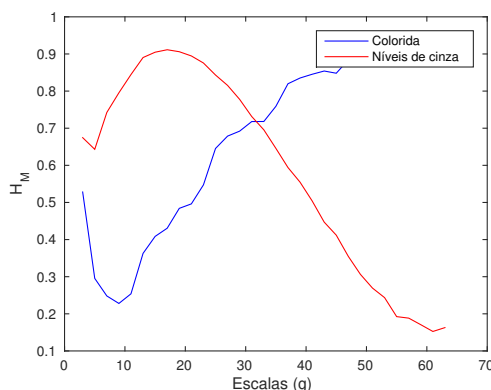
Para as curvas que consideram a escala $m = 1$ é comum que os valores de entropia das imagens de controle sejam próximos, principalmente na primeira escala de variação ϵ , pois um único *pixel* é avaliado. Quando as janelas apresentam tamanho maior ($m = 2$, o qual possui 4 *pixels* avaliados ou $m = 3$, que possui 9 *pixels* avaliados) os padrões começam a ser identificados e os valores para as imagens consideradas tendem a divergir, considerando também a primeira escala de variação ϵ .

Em todas as curvas de entropia Amostral o comportamento foi decrescente quando analisado o eixo dos valores de tolerância ϵ , fato que permitiu constatar que o método proposto teve um comportamento compatível com os de outros trabalhos como o de Da Silva et al. (2016). Com o aumento das tolerâncias, as quantificações também aumentam, bem como a similaridade entre os padrões analisados, desta forma o valor final de entropia diminui.

4.6.2 Curvas de valores de entropia de Shannon

Para a entropia de Shannon (H_M) uma curva de valores também foi extraída das imagens de controle por meio da variação do parâmetro g e está representada na Figura 4.9.

Para a entropia de Shannon (H_M) os valores da curva da imagem colorida representada pela Figura 4.9 aumentam conforme a escala (eixo x) também aumenta. Isso ocorre devido ao comportamento deste tipo de entropia, que está diretamente relacionado com a proporção de *pixels* com intensidades diferentes presentes em uma determinada janela. Para pequenas janelas a proporção de intensidades diferentes é quase nula. Este comportamento pode ser explicado pela homogeneidade das cores presentes na janela, conseqüentemente o valor de entropia tende a 0. Estas características indicam que conforme os padrões de cores são identificados, maior a variabilidade de

Figura 4.9: Curvas de entropia de Shannon (H_M) das imagens de controle.

Fonte: Elaborado pelo autor.

cores presentes na imagem. Outra constatação para este caso é que os padrões são repetidos ao longo da imagem, o que permite que as quantidades de *pixels* de cada cor sejam mais balanceadas e o valor de entropia aumente conforme a escala também aumenta.

Em relação a curva da imagem em níveis de cinza, o valor de entropia é oposto ao da colorida. Para valores pequenos de g a entropia é alta e tende a diminuir com o aumento da escala. Este comportamento ocorre devido a janelas pequenas apresentarem proporções balanceadas entre a quantidade de intensidades presentes, pois a quantidade de *pixels* é pequena. Em decorrência disso, o valor de entropia tende ao valor máximo (valor 1), pois as intensidades tendem a ser equiprováveis. Como não há padrões de cores definidos nesta imagem, a complexidade (variabilidade de cores) diminui conforme o tamanho das janelas aumenta, pois a quantidade de *pixels* de algumas intensidades se tornam predominantes em comparação as demais e a janela tende a ser mais homogênea. Com isso, o valor de entropia tende ao valor mínimo (valor 0).

4.7 Extração de atributos da curva

Os resultados obtidos com a entropia Amostral ($SampEn_{MF}$) foram curvas de valores para cada imagem colorretal. As curvas foram geradas em função de variações dos

parâmetros ϵ e m , fornecendo três curvas independentes, uma para cada valor de m (de 1 a 3), porém variando ϵ , de $0,06 \leq \epsilon \leq 0,5$, com incrementos de 0,06. Cada curva de entropia foi analisada a partir de métricas indicadas em Căliman e Ivanovici (2012) e Roberto et al. (2017), resultando em novos atributos, tais como área sob a curva, obliquidade, razão da área, escala do ponto máximo e valor do ponto máximo. Para este trabalho, as curvas foram definidas considerando os valores da entropia Amostral (eixo y) em função da constante de tolerância (ϵ) (eixo x) para cada escala m . Para a entropia de Shannon (H_M) os valores (eixo y) variam em função do tamanho das janelas (g) (eixo x).

A métrica área sob a curva A foi calculada via o método de integração numérica trapezoidal, considerando uma integração de Ω pontos igualmente espaçados. A aproximação A foi dada pela Equação 4.16, na qual o espaço entre cada ponto, representado pelo incremento de ϵ ($0,06$ com limites $0,06 \leq \epsilon \leq 0,48$), corresponde ao valor escalar $\frac{U-V}{\Omega-1}$, com Ω representando o total de pontos utilizados na integração, U representando o menor valor de ϵ e V o maior valor. (MATHWORKS, 2017).

$$A = \frac{U - V}{2(\Omega - 1)} \sum_{\omega=1}^{\Omega-1} (f(\epsilon_{\omega}) + f(\epsilon_{\omega+1})), \quad (4.16)$$

em que $f(\epsilon_{\omega})$ e $f(\epsilon_{\omega+1})$ definem os valores de entropia para pares consecutivos de ϵ .

A obliquidade (O) permitiu definir a assimetria contida nas amostras dos dados analisados. Resultados positivos indicaram que as amostras estavam mais concentradas à direita da média. Para concentrações à esquerda, os resultados foram negativos. Uma distribuição perfeita indica que não há assimetria, portanto a obliquidade é igual a zero (MATHWORKS, 2017). A obliquidade foi calculada aplicando a Equação 4.17:

$$O = \frac{\frac{1}{\Omega} \sum_{\omega=1}^{\Omega} (f(\epsilon_{\omega}) - \psi)^3}{\left(\sqrt{\frac{1}{\Omega} \sum_{\omega=1}^{\Omega} (f(\epsilon_{\omega}) - \psi)^2} \right)^3}, \quad (4.17)$$

em que ψ define a média da amostra considerada; Ω define o total de pontos da amostra e $f(\epsilon_\omega)$ indica o valor de entropia em função do ω -ésimo valor da constante da tolerância ϵ .

A razão da área Γ correspondeu a área calculada da metade direita da função sobre a área calculada da metade esquerda da função. A razão da área (Γ) utilizada neste trabalho foi definida como:

$$\Gamma = \frac{A_{(\frac{\Omega}{2}+1, \Omega)}}{A_{(1, \frac{\Omega}{2})}}. \quad (4.18)$$

Outra métrica foi o valor da escala do ponto máximo (E), indicada por um valor do eixo x que forneceu o maior valor (ponto máximo) de entropia (eixo y). Neste caso, o valor obtido foi uma tolerância ϵ para a entropia Amostral. Além desta, também foi considerado como métrica o próprio valor do ponto máximo (W).

4.8 Composição dos vetores de características

Os vetores de características foram obtidos a partir dos valores de entropia (Amostral e de Shannon). Foram testadas três composições de vetores: a) vetores constituídos pelos valores das entropias; b) vetores constituídos pelos valores extraídos das curvas de entropias; c) união das duas primeiras composições. Esta estratégia foi adotada para conhecer quais foram as composições de características mais relevantes para o estudo do contexto de imagens colorretais.

4.8.1 Estruturas dos vetores da entropia Amostral

Para a entropia Amostral ($SampEn_{MF}$), a primeira composição (vetores constituídos pelos valores das entropias) foi dada por 24 atributos, obtidos em função de variações dos parâmetros m e ϵ . A estrutura desta composição é ilustrada na Figura 4.10.

A segunda composição de vetor foi definida pelos valores extraídos das três curvas, aplicando as métricas área (A), razão da área (Γ), obliquidade (O), ponto máximo (W)

Figura 4.10: Ilustração da estrutura adotada para a primeira composição de atributos, um vetor com 24 descritores para a entropia Amostral ($SampEn_{MF}$).

$m = 1$					$m = 2$			$m = 3$			
$\epsilon = 0,06$	$\epsilon = 0,12$	$\epsilon = 0,18$	$\epsilon = 0,24$	$\epsilon = 0,18$	$\epsilon = 0,24$	$\epsilon = 0,42$	$\epsilon = 0,48$

Fonte: Elaborado pelo autor.

e escala do ponto máximo (E), totalizando 13 atributos. A estrutura deste vetor é ilustrada na Figura 4.11.

Figura 4.11: Ilustração da estrutura para a segunda composição de atributos, um vetor definido com 13 descritores para a entropia Amostral ($SampEn_{MF}$).

$m = 1$				$m = 2$				$m = 3$				E
A	Γ	O	W	A	Γ	O	W	A	Γ	O	W	

Fonte: Elaborado pelo autor.

A terceira composição foi dada pela união dos atributos, constituindo o vetor completo, com 37 atributos. O exemplo da estrutura para este vetor está na Figura 4.12.

Figura 4.12: Ilustração da estrutura do vetor completo, definido por 37 atributos da entropia Amostral ($SampEn_{MF}$).

$m = 1$			$m = 2$			$m = 3$			$m = 1$...	E
$\epsilon = 0,06$	$\epsilon = 0,12$	$\epsilon = 0,18$	$\epsilon = 0,48$	A	...		

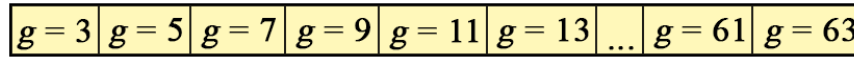
Fonte: Elaborado pelo autor.

4.8.2 Estruturas dos vetores da entropia de Shannon

Para a entropia de Shannon (H_M), os vetores de características foram estruturados de maneira similar ao adotado para a Amostral. A diferença está no número de atributos: a primeira composição foi dada por 31 atributos, obtidos em função de variações aplicadas sobre o parâmetro g . A estrutura do vetor é representada na Figura 4.13.

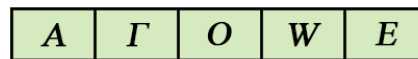
A segunda composição foi definida via valores extraídos das curvas, considerando as métricas área, razão da área, obliquidade, ponto máximo e escala do ponto máximo, totalizando 5 atributos. A estrutura definida para este vetor é ilustrada na Figura 4.14.

Figura 4.13: Ilustração da estrutura do vetor com 31 atributos da entropia de Shannon (H_M).



Fonte: Elaborado pelo autor.

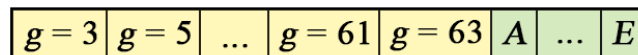
Figura 4.14: Ilustração da estrutura do vetor com cinco atributos da entropia de Shannon (H_M).



Fonte: Elaborado pelo autor.

A terceira composição é a união dos descritores indicados previamente, fornecendo 36 atributos, conforme estrutura ilustrada na Figura 4.15.

Figura 4.15: Ilustração da estrutura do vetor composto por 36 atributos para entropia de Shannon (H_M).



Fonte: Elaborado pelo autor.

4.9 Classificação dos padrões e obtenção das taxas de desempenho

Os conjuntos de atributos para as configurações dos vetores de características descritos na seção 4.8 foram considerados como entradas para os classificadores *Decision Tree*, *SVM*, *Multilayer Perceptron*, *Logistic*, *RBF*, *Multiclass Classifier* e *Rotation Forest*, conforme descrição apresentada na seção 2.3 do capítulo 2. As classificações foram a partir da plataforma *weka v3.6.10* (GARNER et al., 1995), na qual os algoritmos correspondentes são *J48*, *SMO (Sequential Minimal Optimization)*, *MultilayerPerceptron*, *Logistic*, *RBFNetwork*, *MulticlassClassifier* e *RotationForest*. Os algoritmos foram aplicados com as configurações padrões indicadas pela plataforma. Os resultados preliminares foram obtidos considerando as amostras para treinamento e testes,

aplicando o método de validação cruzada *k-fold* ($k=10$), conforme descrito na seção 2.3 da Fundamentação Teórica. Os resultados desta etapa foram taxas de AUC extraídas das melhores combinações entre as diferentes configurações dos vetores de características e classificadores. Os resultados foram comparados aos disponíveis em trabalhos relacionados com as técnicas abordadas (modelos de entropia com associações de abordagens) e aos contextos de imagens explorados na proposta aqui descrita.

4.10 Materiais

Os modelos foram executados em computadores com processador Intel Xeon 2,40 GHz e 16 GB de memória RAM, com Windows 7 Professional 64 bits, Matlab R2012b e Weka v3.6.10.

Capítulo 5

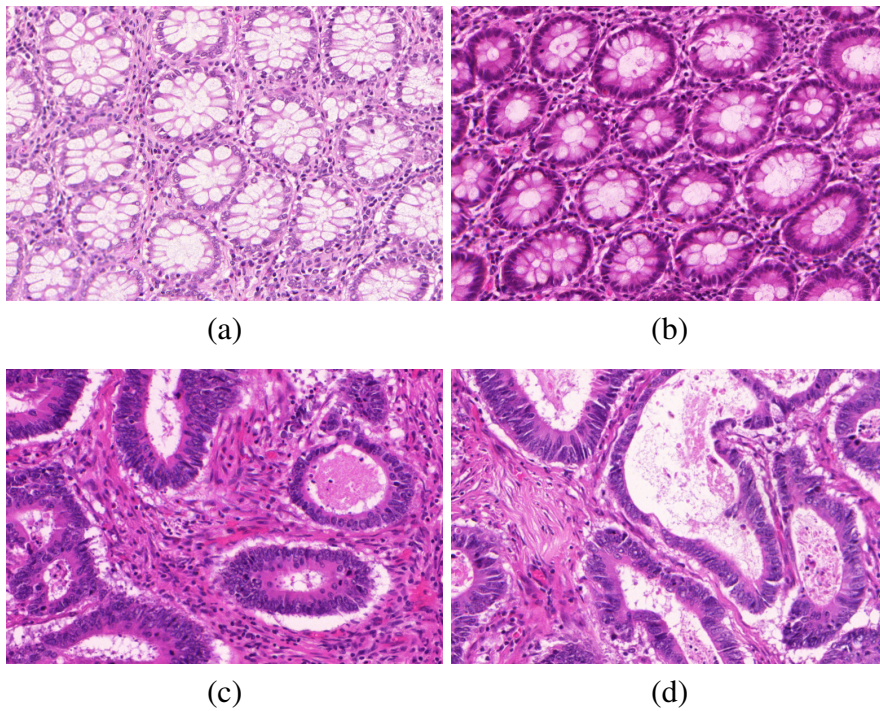
Resultados e Discussão

As técnicas propostas foram executadas em imagens histológicas colorretais e os resultados foram valores que compõem curvas de entropias. As melhores combinações entre características e classificadores foram investigadas para distinguir os grupos benigno e maligno de câncer colorretal. Detalhes dos resultados obtidos estão descritos nas próximas seções.

5.1 Resultados do método proposto para a entropia Amostral

O método de entropia Amostral com as abordagens multiescalas, *Fuzzy* e multidimensional foi aplicado sobre as amostras de H&E. Exemplos de algumas imagens das classes benigna e maligna estão na Figura 5.1.

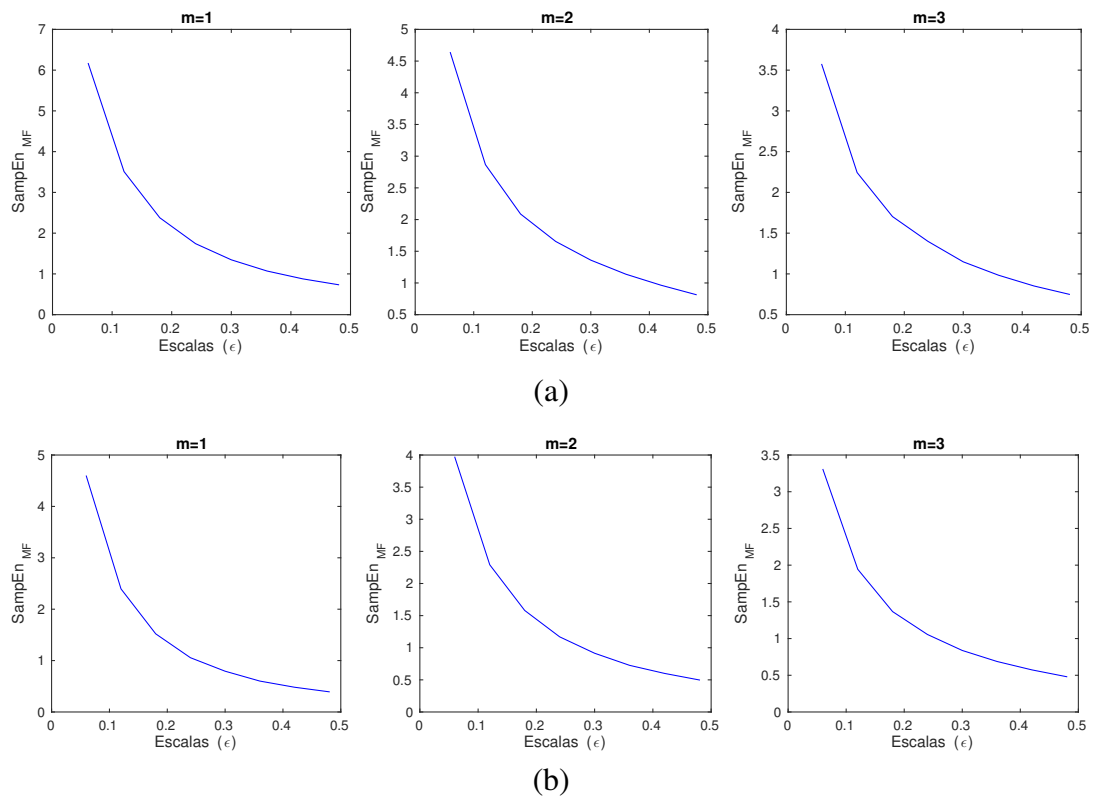
Figura 5.1: Exemplos de imagens das classes benigna, (a) e (b), e maligna, (c) e (d).



Fonte: (SIRINUKUNWATTANA et al., 2017).

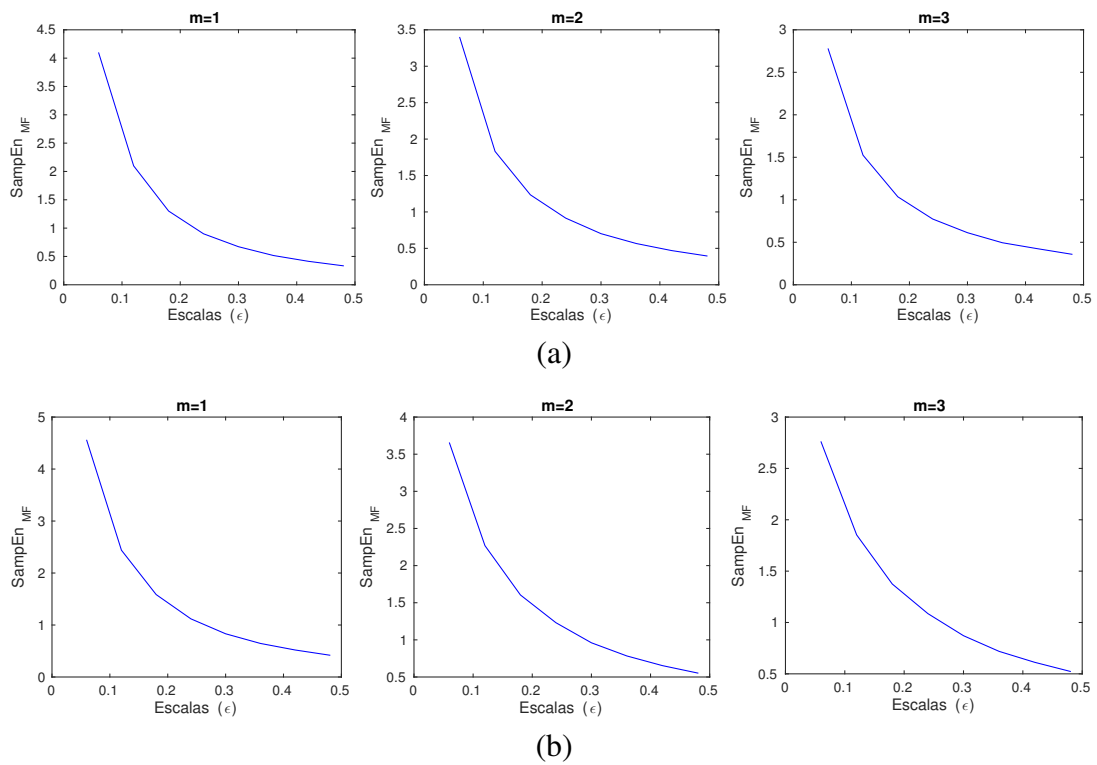
As curvas de entropias para estas imagens estão nas Figuras 5.2 e 5.3, bem como os valores correspondentes estão nas Tabelas 5.1 a 5.3.

Figura 5.2: Curvas da entropia Amostral ($SampEn_{MF}$), com variações dos parâmetros ϵ e m , para as imagens indicadas nas Figuras 5.1a (a) e 5.1b (b).



Fonte: Elaborado pelo autor.

Figura 5.3: Curvas de entropia Amostral ($SampEn_{MF}$), com variações dos parâmetros ϵ e m , para as imagens ilustradas nas Figuras 5.1c (a) e 5.1d (b).



Fonte: Elaborado pelo autor.

Tabela 5.1: Ilustração de vetores de características para as imagens mostradas na Figura 5.1, considerando a configuração dos 24 atributos obtidos pelo cálculo da entropia Amostral ($SampEn_{MF}$).

Imagem	$m = 1$			$m = 2$			$m = 3$		
	$\epsilon = 0,06$	$\epsilon = 0,12$	$\epsilon = 0,18$	$\epsilon = 0,24$	$\epsilon = 0,18$	$\epsilon = 0,24$	$\epsilon = 0,42$	$\epsilon = 0,48$	
Figura 5.1a	6,1616	3,5109	2,3776	1,7423	2,0875	1,6550	0,8519	0,7492	
Figura 5.1b	4,6347	2,4628	1,5666	1,0910	1,6303	1,2315	0,6023	0,5098	
Figura 5.1c	4,2568	2,2011	1,3785	0,9515	1,4978	1,1241	0,5492	0,4647	
Figura 5.1d	4,5534	2,4392	1,5856	1,1187	1,6046	1,2311	0,6122	0,5235	

Tabela 5.2: Ilustração de vetores de características para as imagens ilustradas na Figura 5.1, considerando a configuração dos 13 atributos obtidos pela curva de entropia Amostral ($SampEn_{MF}$).

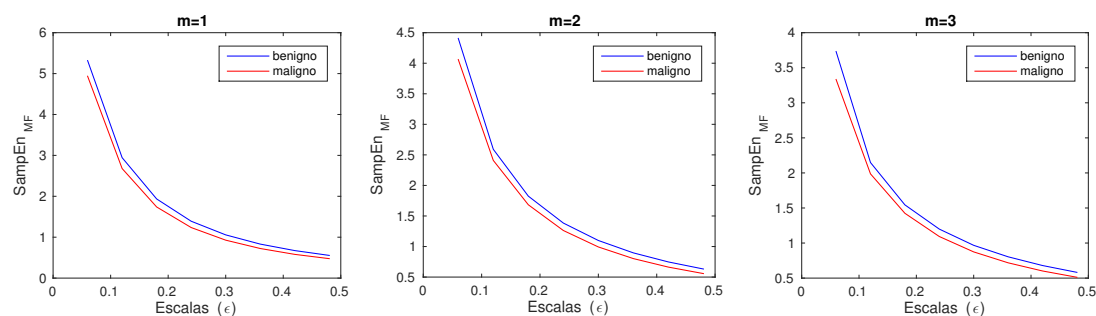
Imagem	$m = 1$				$m = 2$			$m = 3$			E
	A	Γ	O	W	A	\dots	W	A	\dots	W	
Figura 5.1a	14,373	0,30372	1,3578	6,1616	12,795	\dots	4,6342	10,490	\dots	3,5699	0,06
Figura 5.1b	9,598	0,25392	1,4472	4,6347	9,7818	\dots	3,9015	8,4906	\dots	3,1317	0,06
Figura 5.1c	8,501	0,24056	1,4788	4,2568	8,974	\dots	3,6325	7,9392	\dots	2,9375	0,06
Figura 5.1d	9,627	0,26106	1,5025	4,5534	9,6039	\dots	3,6513	8,1572	\dots	2,7586	0,06

Tabela 5.3: Ilustração de vetores de características para as imagens representadas na Figura 5.1, considerando a configuração dos 37 atributos obtidos pela junção dos atributos do cálculo da entropia Amostral ($SampEn_{MF}$) e os extraídos da curva.

Imagem	$m = 1$			$m = 2$		$m = 3$		$m = 1$	\dots	E	
	$\epsilon = 0,06$	$\epsilon = 0,12$	\dots	\dots	$\epsilon = 0,18$	\dots	$\epsilon = 0,48$	A	\dots		
Figura 5.1a	6,1616	3,5109	\dots	\dots	2,0875	\dots	\dots	0,7492	14,373	\dots	0,06
Figura 5.1b	4,6347	2,4628	\dots	\dots	1,6303	\dots	\dots	0,5098	9,598	\dots	0,06
Figura 5.1c	4,2568	2,2011	\dots	\dots	1,4978	\dots	\dots	0,4647	8,501	\dots	0,06
Figura 5.1d	4,5534	2,4392	\dots	\dots	1,6046	\dots	\dots	0,5235	9,627	\dots	0,06

Para observar os comportamentos dos grupos benigno e maligno a partir dos valores das entropias, as curvas médias foram calculadas para as 67 amostras benignas e 84 malignas. As curvas estão na Figura 5.4 e os valores médios de entropia dos grupos com seus respectivos desvios padrões estão na Tabela 5.4.

Figura 5.4: Curvas médias de entropia Amostral ($SampEn_{MF}$) com as variações dos parâmetros ϵ e m referentes aos grupos benigno e maligno.



Fonte: Elaborado pelo autor.

Tabela 5.4: Valores médios da entropia Amostral ($SampEn_{MF}$) e desvios padrões (DP) dos grupos benigno e maligno.

Grupos	$\epsilon = 0,06$	$\epsilon = 0,12$	$\epsilon = 0,18$	$\epsilon = 0,24$	$\epsilon = 0,30$	$\epsilon = 0,36$	$\epsilon = 0,42$	$\epsilon = 0,48$
		$m = 1$						
Benigno	5,3177±0,6394	2,9440±0,4684	1,9351±0,3519	1,3905±0,2802	1,0568±0,2305	0,8315±0,1954	0,6729±0,1674	0,5550±0,1436
Maligno	4,9353±0,7974	2,6781±0,5382	1,7389±0,3970	1,2362±0,3095	0,9276±0,2477	0,7238±0,2033	0,5803±0,1708	0,4757±0,1447
	$m = 2$							
Benigno	4,4062±0,5279	2,5913±0,3284	1,8247±0,2539	1,3842±0,2094	1,0988±0,1771	0,8960±0,1547	0,7476±0,1360	0,6321±0,1209
Maligno	4,0603±0,5824	2,4094±0,3852	1,6833±0,2961	1,2623±0,2433	0,9929±0,2031	0,8024±0,1736	0,6632±0,1492	0,5572±0,1309
	$m = 3$							
Benigno	3,7310±0,7655	2,1481±0,2956	1,5445±0,2081	1,1991±0,1692	0,9675±0,1395	0,8039±0,1217	0,6795±0,1086	0,5822±0,0974
Maligno	3,3324±0,5265	1,9864±0,2970	1,4245±0,2371	1,0931±0,1965	0,8737±0,1661	0,7182±0,1417	0,6018±0,1233	0,5120±0,1079

Conforme descrição apresentada na seção 4.8, os vetores de características foram definidos e fornecidos como entradas para os classificadores indicados na seção 4.9. As taxas de AUC foram coletadas considerando validação cruzada *10-fold* e mostradas na Tabela 5.5. Para evidenciar que as alterações definidas na entropia Amostral forneceram bons resultados, a abordagem multiescala (*SampEn_{2D}*) descrita por Da Silva et al. (2016) foi aplicada sobre o mesmo conjunto de imagens e os resultados estão na Tabela 5.6. Os melhores valores de AUC foram destacados em negrito, assim como a média e desvio padrão (DP) entre os classificadores.

Tabela 5.5: Valores e médias de AUC calculados a partir da entropia Amostral proposta (*SampEn_{MF}*), considerando diferentes classificadores e composições dos vetores de características.

Classificadores	10-fold cross-validation		
	Vetor com 24 atributos	Vetor com 13 atributos	Vetor com 37 atributos (24 + 13)
J48	0,720	0,651	0,742
SMO	0,603	0,687	0,700
<i>RotationForest</i>	0,923	0,896	0,913
<i>MultiClassClassifier</i>	0,889	0,940	0,866
<i>MultilayerPerceptron</i>	0,891	0,917	0,907
<i>Logistic</i>	0,889	0,940	0,866
<i>RBFNetwork</i>	0,641	0,682	0,675
Média±DP	0,794±0,135	0,816±0,135	0,810±0,101

Tabela 5.6: Valores e médias de AUC calculados a partir da técnica *SampEn_{2D}* (SILVA et al., 2016).

Classificadores	10-fold cross-validation		
	Vetor com 8 atributos	Vetor com 5 atributos	Vetor com 13 atributos (8 + 5)
J48	0,459	0,581	0,637
SMO	0,500	0,494	0,494
<i>RotationForest</i>	0,803	0,758	0,789
<i>MultiClassClassifier</i>	0,775	0,662	0,776
<i>MultilayerPerceptron</i>	0,775	0,740	0,784
<i>Logistic</i>	0,775	0,662	0,776
<i>RBFNetwork</i>	0,543	0,703	0,622
Média±DP	0,661±0,153	0,657±0,093	0,697±0,115

É importante observar que os vetores de características definidos com o método *SampEn_{2D}* (SILVA et al., 2016) foram compostos seguindo as mesmas regras de composições. Apesar disso, o método descrito por estes autores forneceu vetores compostos por um número menor de atributos. A primeira composição foi com oito valores

de entropia. A entropia Amostral com as alterações propostas ($SampEn_{MF}$) forneceu quantificações mais completas, 24 atributos, permitindo avaliar um número maior de combinações entre ϵ e m . Uma explicação é que o modelo proposto permite avaliar informações presentes nos diferentes canais de cores.

Observando os resultados é possível verificar que a $SampEn_{MF}$ forneceu resultados superiores aos conquistados com a técnica $SampEn_{2D}$ (SILVA et al., 2016). A maior taxa de AUC foi de **0,940** no método $SampEn_{MF}$, via classificadores *Multi-class Classifier* e *Logistic*, contra **0,803** no método $SampEn_{2D}$ (SILVA et al., 2016) (classificador *Rotation Forest*). A taxa média de AUC com 13 atributos, **0,816**, também superou a fornecida pela $SampEn_{2D}$ (SILVA et al., 2016), valor de **0,697**, bem como a quantidade de classificações com melhores taxas. Para a primeira e terceira composições de vetores de características, o método proposto forneceu taxas de AUC superiores para todos os classificadores. Para a segunda composição, os atributos extraídos da curva, o método proposto foi superior em seis classificadores, considerando o total de sete possíveis. Para verificar se as diferenças entre as taxas de AUC são estatisticamente significativas, o teste Kruskal-Wallis foi aplicado para comparação de todos os pares, via o método Dwass-Steel-Christlow-Fligner (com P resultando em 0,0495) e pelo método Conover-Inman (com P calculado em 0,0213). É importante destacar que este é um método não-paramétrico para comparação de amostras independentes. O teste de Kruskal-Wallis é o mais adequado para aplicação em amostras não-balanceadas (MCKIGHT; NAJAB, 2010 apud COLLINS, 2017). Portanto, é possível constatar que os valores para P indicaram que as diferenças de AUC são significativas entre a $SampEn_{MF}$ e a $SampEn_{2D}$ (SILVA et al., 2016). Este é mais um indicativo importante para mostrar que as associações com as abordagens *Fuzzy* e multidimensional na entropia Amostral forneceram resultados promissores, principalmente considerando um método correlato e aplicados em condições de igualdade.

Com o objetivo de observar o comportamento da $SampEn_{MF}$ frente aos desem-

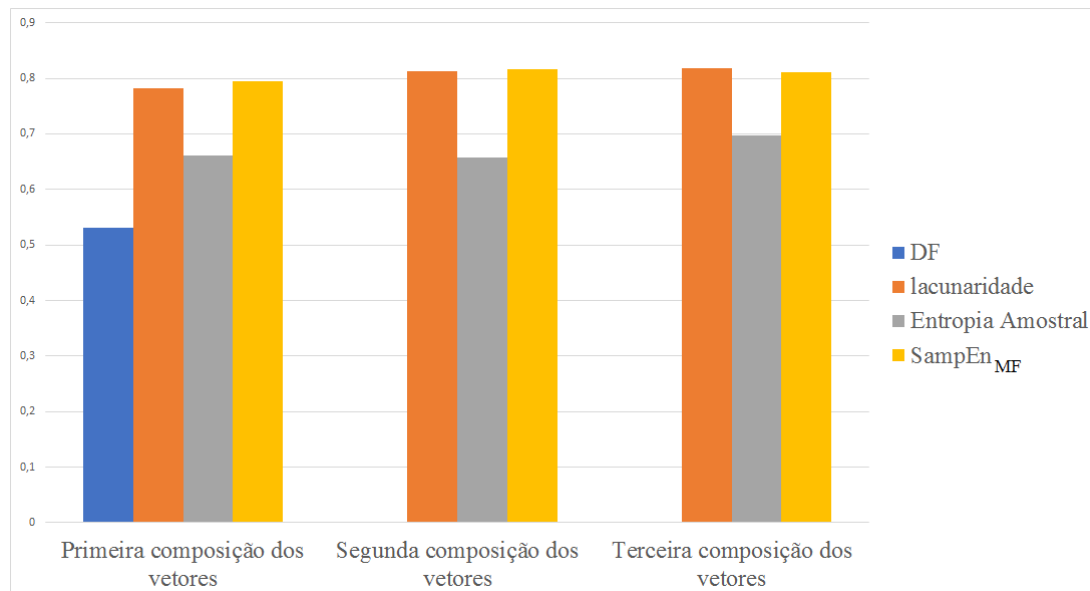
penhos de técnicas multiescalas e multidimensionais reconhecidas na literatura, na Tabela 5.7 estão os resultados para a dimensão fractal e lacunaridade. A dimensão fractal contou com dois atributos obtidos via técnicas descritas por Ivanovici e Richard (2011) e por Nikolaidis et al. (2011). A abordagem de lacunaridade foi apresentada por Ivanovici, Richard e Decean (2009). Os melhores valores de AUC foram destacados em negrito, bem como a média e desvio padrão entre os classificadores. Complementarmente, na Figura 5.5 estão ilustradas as taxas médias de AUC calculadas com as técnicas de DF (IVANOVICI; RICHARD, 2011; NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011), lacunaridade (IVANOVICI; RICHARD; DECEAN, 2009), $SampEn_{2D}$ (SILVA et al., 2016) e $SampEn_{MF}$.

Tabela 5.7: Valores e médias de AUC calculados a partir de diferentes classificadores e as composições dos vetores de características, os quais foram definidos por meio da dimensão fractal (DF) apresentada por Ivanovici e Richard (2011) e por Nikolaidis, Nikolaidis e Tsouros (2011) e lacunaridade (LAC) apresentada por Ivanovici, Richard e Decean (2009).

Classificadores	10-fold cross-validation			
	Vetor com 2 atributos	Vetor com 22 atributos	Vetor com 5 atributos	Vetor com 27 atributos (22+5)
	DF	LAC	LAC	LAC
J48	0,470	0,652	0,731	0,763
SMO	0,500	0,711	0,775	0,769
<i>RotationForest</i>	0,451	0,912	0,897	0,918
<i>MultiClassClassifier</i>	0,527	0,857	0,837	0,862
<i>MultilayerPerceptron</i>	0,584	0,890	0,821	0,881
<i>Logistic</i>	0,527	0,857	0,837	0,862
<i>RBFNetwork</i>	0,652	0,597	0,786	0,672
Média±DP	0,530±0,069	0,782±0,126	0,812±0,053	0,818±0,086

Pelos resultados obtidos é possível constatar que as características fornecidas com a $SampEn_{MF}$ permitiu resultados superiores aos conquistados com a DF, mesmo considerando o seu melhor desempenho: a melhor taxa de AUC da $SampEn_{MF}$ foi de **0,923**, com o classificador *Rotation Forest*, contra **0,652** da DF com o classificador *RBFNetwork*. Em comparação com a lacunaridade, as taxas médias de AUC foram

Figura 5.5: Taxas médias de AUC de DF, lacunaridade, da $SampEn_{MF}$ e da $SampEn_{2D}$ (SILVA et al., 2016), considerando as diferentes configurações dos vetores de características.



Fonte: Elaborado pelo autor.

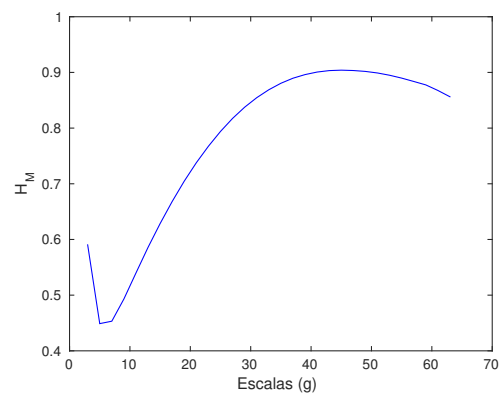
parecidas com as obtidas via $SampEn_{MF}$. A maior diferença foi de **0,012** para taxas de AUC calculadas com a primeira composição de vetor de características. Aplicando o teste de Kruskal-Wallis para comparação de todos os pares pelos métodos Dwass-Steel-Christchlow-Fligner e Conover-Inman, os valores P foram de 0,827 e 0,854, respectivamente. Essas diferenças não são estatisticamente relevantes. Porém, é importante destacar que o método proposto indicou a maior taxa de AUC (**0,940**), com 13 atributos, contra **0,918** da lacunaridade, com 27 atributos.

5.2 Resultados do método proposto para a entropia de Shannon

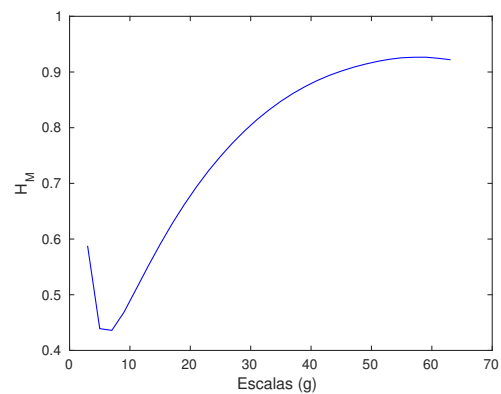
A entropia de Shannon (H), por sua vez, também foi associada com uma abordagem multiescala e multidimensional, conforme descrições apresentadas na seção 4.2 do capítulo 4. Após aplicar esta técnica para quantificar as imagens colorretais, os

resultados foram curvas compostas por 31 valores de entropia. Exemplos de curvas de valores de entropia de Shannon modificada (H_M) estão dispostas nas Figuras 5.6 e 5.7, considerando as imagens das classes benigna e maligna apresentadas previamente (Figura 5.1). Os valores para cada imagem estão nas Tabelas 5.8 a 5.10.

Figura 5.6: Curvas de entropia de Shannon (H_M) em função das diversas variações de escala do parâmetro g para as imagens representadas nas Figuras 5.1a (a) e 5.1b (b).



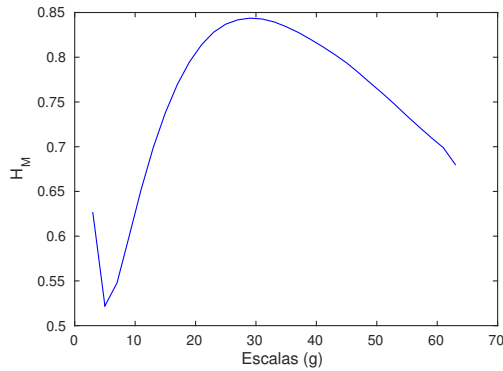
(a)



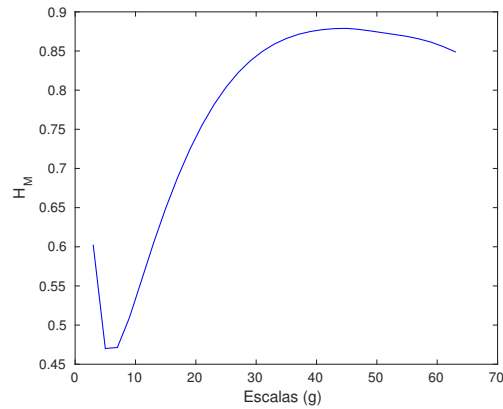
(b)

Fonte: Elaborado pelo autor.

Figura 5.7: Curvas de entropia de Shannon (H_M) em função das diversas variações de escala do parâmetro g para as imagens representadas nas Figuras 5.1c (a) e 5.1d (b).



(a)



(b)

Fonte: Elaborado pelo autor.

Tabela 5.8: Ilustração de vetores de características para as imagens representadas pela Figura 5.1, considerando a configuração dos 31 atributos obtidos pelo cálculo da entropia de Shannon (H_M).

Imagem	$g = 3$	$g = 5$	$g = 7$	$g = 9$	$g = 11$	$g = 13$...	$g = 61$	$g = 63$
Figura 5.1a	0,59034	0,44887	0,45321	0,49311	0,53984	0,58585	...	0,86759	0,85627
Figura 5.1b	0,59050	0,44008	0,43181	0,45883	0,49503	0,53303	...	0,93818	0,93701
Figura 5.1c	0,59920	0,46453	0,46827	0,50781	0,55621	0,60479	...	0,85578	0,84468
Figura 5.1d	0,60214	0,47007	0,47141	0,50957	0,55710	0,60487	...	0,85561	0,84890

Tabela 5.9: Ilustração de vetores de características para as imagens indicadas na Figura 5.1, considerando a configuração dos 5 atributos obtidos pela curva da entropia de Shannon (H_M).

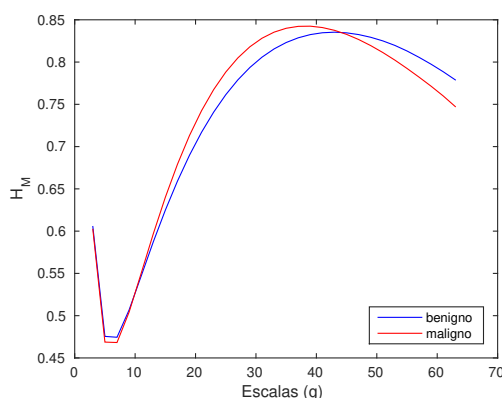
Imagem	A	Γ	O	W	E
Figura 5.1a	23,416	1,3566	-1,0365	0,90404	45
Figura 5.1b	22,449	1,2171	-1,0530	0,85281	39
Figura 5.1c	22,415	1,1056	-0,96486	0,85052	33
Figura 5.1d	21,883	0,99282	-0,78437	0,8088	25

Tabela 5.10: Ilustração de vetores de características para as imagens mostradas na Figura 5.1, considerando a configuração dos 36 atributos obtidos pela junção dos atributos do cálculo da entropia de Shannon (H_M) e os extraídos da curva.

Imagem	$g = 3$	$g = 5$...	$g = 61$	$g = 63$	A	...	E
Figura 5.1a	0,59034	0,44887	...	0,86759	0,85627	23,416	...	45
Figura 5.1b	0,59050	0,44008	...	0,93818	0,93701	22,449	...	39
Figura 5.1c	0,59920	0,46453	...	0,85578	0,84468	22,415	...	33
Figura 5.1d	0,60214	0,47007	...	0,85561	0,84890	21,883	...	25

Os comportamentos dos grupos benigno e maligno a partir da H_M estão indicadas na Figura 5.8: as curvas são os valores médios das entropias. Este tipo de representação permitiu constatar que escalas maiores de g (de 15 a 63) forneceram os valores de entropias mais distintos para os grupos de câncer colorretal.

Figura 5.8: Curvas médias de entropia de Shannon (H_M) com as variações do parâmetro g referentes aos grupos benigno e maligno.



Fonte: Elaborado pelo autor.

Para uma verificação mais detalhada sobre o potencial das características obtidas

com a H_M , o modelo de Shannon multiescala tradicional (H) (CONVERTINO et al., 2012) foi aplicado sobre o mesmo conjunto de imagens. Os vetores foram constituídos a partir das três composições descritas previamente e fornecidos como entradas para os classificadores. As classificações foram realizadas aplicando validação cruzada (10-*fold*) e os resultados a partir das técnicas estão nas Tabelas 5.11 e 5.12. As taxas de AUC mais relevantes foram destacadas em negrito, bem como as médias das taxas para cada conjunto testado.

Tabela 5.11: Valores e médias de AUC calculados a partir de diferentes classificadores e de diferentes composições dos vetores de características obtidos pelo método proposto (H_M), considerando imagens coloridas.

Classificadores	10- <i>fold cross-validation</i>		
	Vetor com 31 atributos	Vetor com 5 atributos	Vetor com 36 atributos (31 + 5)
J48	0,645	0,611	0,689
SMO	0,642	0,494	0,642
<i>RotationForest</i>	0,793	0,744	0,857
<i>MultiClassClassifier</i>	0,776	0,606	0,873
<i>MultilayerPerceptron</i>	0,793	0,686	0,787
<i>Logistic</i>	0,776	0,606	0,873
<i>RBFNetwork</i>	0,538	0,590	0,514
Média±DP	0,709±0,101	0,620±0,079	0,748±0,138

Tabela 5.12: Valores e médias de AUC calculados a partir de diferentes classificadores e de diferentes composições dos vetores de características obtidos pelo método multiescala tradicional da entropia de Shannon (H), considerando imagens em níveis de cinza.

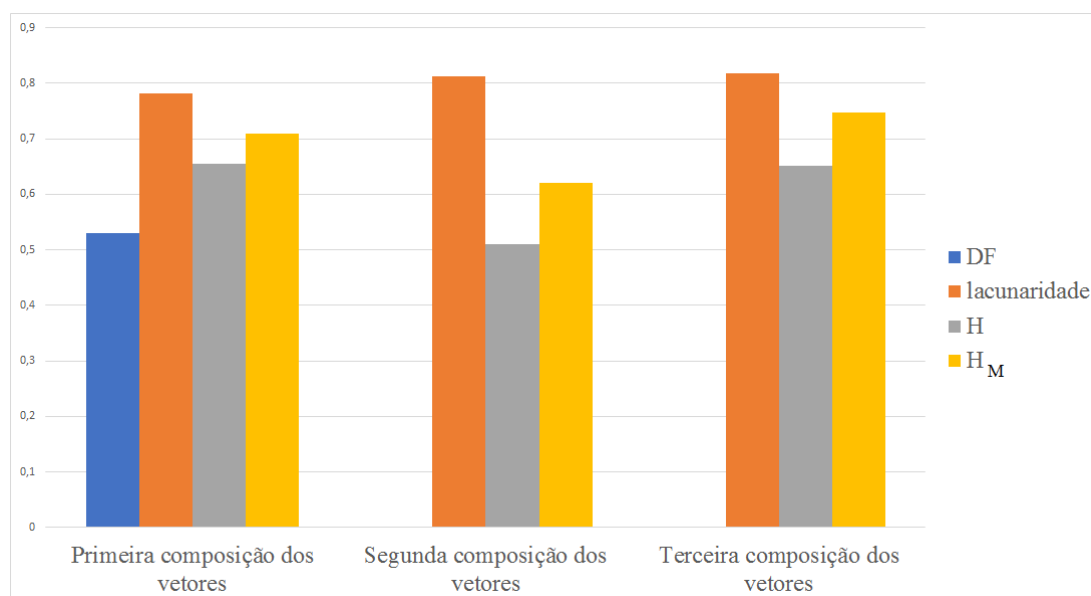
Classificadores	10- <i>fold cross-validation</i>		
	Vetor com 31 atributos	Vetor com 5 atributos	Vetor com 36 atributos (31 + 5)
J48	0,566	0,483	0,571
SMO	0,494	0,494	0,521
<i>RotationForest</i>	0,749	0,576	0,728
<i>MultiClassClassifier</i>	0,825	0,488	0,834
<i>MultilayerPerceptron</i>	0,546	0,501	0,533
<i>Logistic</i>	0,825	0,488	0,834
<i>RBFNetwork</i>	0,582	0,541	0,537
Média±DP	0,655±0,140	0,510±0,035	0,651±0,143

Os resultados obtidos permitem afirmar que as características fornecidas com a H_M permitiram desempenhos (valores médios) superiores aos coletados com a abordagem tradicional (H). Adicionalmente, as duas abordagens forneceram os melhores desempenhos com vetores completos (36 características). Nesta situação, a implementação

clássica proporcionou **0,834** como a maior taxa de AUC, com os classificadores *Multiclass Classifier* e *Logistic*, contra uma taxa de AUC de **0,873** via H_M e mesmos classificadores. Mesmo com estes destaques, as diferenças entre as taxas médias de AUC foram verificadas por meio do teste de Kruskal-Wallis, considerando comparações entre todos os pares. Para o teste Kruskal-Wallis com Dwass-Steel-Christchlow-Fligner o P foi de 0,2752 e via Conover-Inman o P foi de 0,3262, indicando que os resultados não são estatisticamente diferentes.

Os resultados obtidos com a H_M também foram confrontados com os fornecidos por outras técnicas com princípios multiescalas e multidimensionais: DF (IVANOVICI; RICHARD, 2011; NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011) e lacunaridade (IVANOVICI; RICHARD; DECEAN, 2009). Os resultados calculados com estas técnicas foram listados na Tabela 5.7. Um resumo dos desempenhos (taxas médias de AUC) é apresentado na Figura 5.9.

Figura 5.9: Taxas médias de AUC de DF, lacunaridade, H_M e H , considerando as diferentes configurações dos vetores de características.



Fonte: Elaborado pelo autor.

Neste contexto, é possível verificar que a taxa média obtida com a H_M superou os resultados fornecidos pela DF e H , porém os desempenhos foram inferiores aos con-

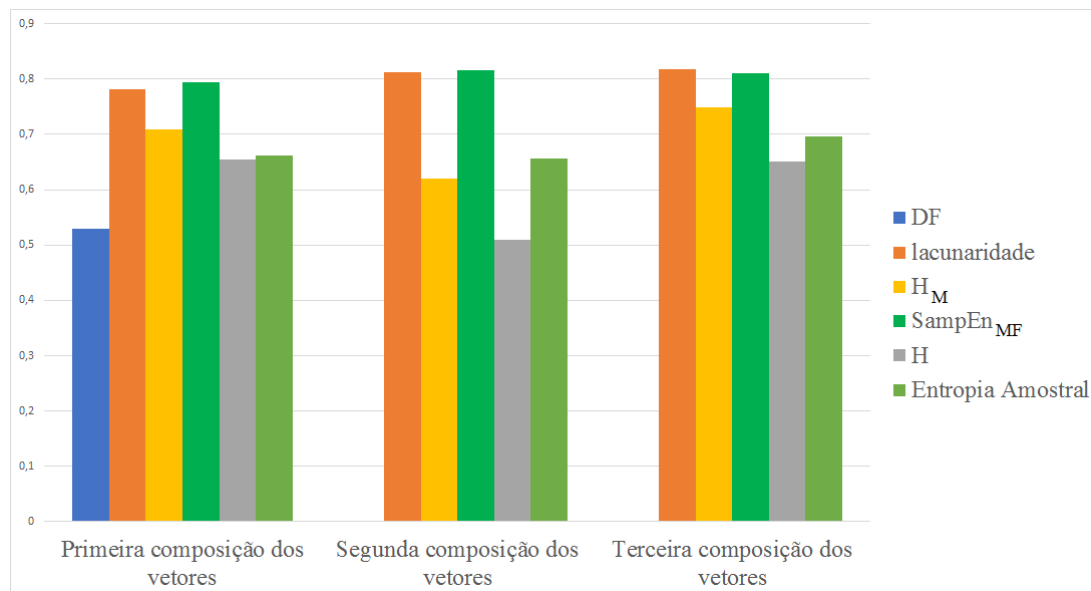
quistados com a lacunaridade (IVANOVICI; RICHARD; DECEAN, 2009). A média de AUC mais relevante foi de **0,748** para a entropia de Shannon (H_M) contra **0,818** da lacunaridade, taxas obtidas com vetores envolvendo todas as características. A maior diferença (**0,192**) entre as médias de AUC foi com os vetores compostos pelos cinco valores extraídos das curvas. Nesta situação, as características via H_M possibilitaram uma AUC média de **0,620**, contra uma taxa de **0,812** com atributos calculados a partir da lacunaridade, também pela composição de cinco atributos. Aplicando o teste de Kruskal-Wallis para comparações entre todos os pares de taxas médias, os resultados obtidos indicaram que as diferenças são estatisticamente relevantes, com $P = 0,0495$ (Dwass-Steel-Christchlow-Fligner) e $P = 0,0213$ (Conover-Inman). Assim, é possível atestar que as características fornecidas pela H_M não forneceram quantificações de imagens colorretais capazes de superar as realizadas com um método multiescala e multidimensional amplamente explorado em processos de quantificações de imagens, a lacunaridade.

5.3 Visão geral das diferenças entre os métodos

Para sumarizar os desempenhos obtidos com as entropias $SampEn_{MF}$ e H_M propostas neste trabalho, na Figura 5.10 são apresentados os desempenhos (taxas médias de AUC) de todas as abordagens discutidas previamente. É possível verificar que as características obtidas via $SampEn_{MF}$ forneceram melhores resultados do que os calculados com as técnicas DF (IVANOVICI; RICHARD, 2011; NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011), entropia de Shannon tradicional (H) (CONVERTINO et al., 2012), H_M e $SampEn_{2D}$ (SILVA et al., 2016). Porém, é evidente também que os desempenhos entre a $SampEn_{MF}$ e a lacunaridade (IVANOVICI; RICHARD; DECEAN, 2009) são equivalentes sob o critério taxa média. Por outro lado, considerando a maior taxa de AUC, é possível indicar que a $SampEn_{MF}$ foi a mais relevante para estudar as imagens colorretais da base pública disponibilizada por Sirinukunwattana

et al. (2017).

Figura 5.10: Taxas médias de AUC de DF, lacunaridade, $SampEn_{2D}$, $SampEn_{MF}$, H e H_M , considerando as diferentes configurações dos vetores de características.



Fonte: Elaborado pelo autor.

Apesar das constatações apresentadas, para verificar se existem diferenças entre as classificações realizadas a partir das características fornecidas pelas técnicas com os melhores desempenhos, o teste de Friedman foi aplicado para listar as melhores combinações, conforme indicações discutidas em Yu et al. (2016). O teste de Friedman foi aplicado sobre as combinações envolvendo os grupos benigno contra maligno, classificadores, lacunaridade (IVANOVICI; RICHARD; DECEAN, 2009), $SampEn_{2D}$ (SILVA et al., 2016) e entropia Amostral proposta neste trabalho. Além disso, as combinações foram testadas a partir das taxas de AUC obtidas com as duas melhores estratégias para compor os vetores de características (atributos calculados a partir das curvas e vetores completos). Os resultados obtidos estão nas Tabelas 5.13 e 5.14, com um ranqueamento das melhores combinações.

Observando os resultados da Tabela 5.13 é possível constatar que entre as primeiras cinco melhores combinações, quatro foram definidas a partir da entropia Amostral proposta, inclusive fornecendo as primeiras três melhores combinações. Este fato indica

Tabela 5.13: Teste de Friedman para as combinações entre classificadores e técnicas, considerando taxas de AUC obtidas via vetores de características compostos por valores calculados das curvas.

Técnica	Classificador	Ranking
SampEn_{MF}	<i>Logistic</i>	1,5
	<i>Multiclass Classifier</i>	1,5
	<i>Multilayer Perceptron</i>	3,0
lacunaridade	<i>Rotation Forest</i>	4,0
SampEn_{MF}	<i>Rotation Forest</i>	5,0
lacunaridade	<i>Logistic</i>	6,5
	<i>Multiclass Classifier</i>	6,5
	<i>Multilayer Perceptron</i>	8,0
	RBF	9,0
	SVM	10,0
<i>SampEn_{2D}</i>	<i>Rotation Forest</i>	11,0
	<i>Multilayer Perceptron</i>	12,0
lacunaridade	<i>Decision Tree</i>	13,0
<i>SampEn_{2D}</i>	RBF	14,0
SampEn_{MF}	SVM	15,0
	RBF	16,0
<i>SampEn_{2D}</i>	<i>Logistic</i>	17,5
	<i>Multiclass Classifier</i>	17,5
SampEn_{MF}	<i>Decision Tree</i>	19,0
<i>SampEn_{2D}</i>	<i>Decision Tree</i>	20,0
	SVM	21,0

Tabela 5.14: Teste de Friedman para as combinações entre classificadores e técnicas, considerando taxas de AUC obtidas com vetores de características completos.

Técnica	Classificador	Ranking
lacunaridade	<i>Rotation Forest</i>	1,0
SampEn_{MF}	<i>Rotation Forest</i>	2,0
	<i>Multilayer Perceptron</i>	3,0
lacunaridade	<i>Multilayer Perceptron</i>	4,0
SampEn_{MF}	<i>Logistic</i>	5,5
	<i>Multiclass Classifier</i>	5,5
lacunaridade	<i>Logistic</i>	7,5
	<i>Multiclass Classifier</i>	7,5
<i>SampEn_{2D}</i>	<i>Rotation Forest</i>	9,0
	<i>Multilayer Perceptron</i>	10,0
	<i>Logistic</i>	11,5
	<i>Multiclass Classifier</i>	11,5
lacunaridade	SVM	13,0
	<i>Decision Tree</i>	14,0
SampEn_{MF}	<i>Decision Tree</i>	15,0
	SVM	16,0
	RBF	17,0
lacunaridade	RBF	18,0
<i>SampEn_{2D}</i>	<i>Decision Tree</i>	19,0
	RBF	20,0
	SVM	21,0

a superioridade do método para distinguir os grupos de imagens colorretais, exploradas neste trabalho, com vetores compostos por características calculadas a partir das curvas de entropia (número reduzido de características). Combinações obtidas com a abordagem Amostral descrita por Da Silva et al. (2016) foram listadas somente a partir da undécima posição, confirmando a importância das alterações descritas neste trabalho. Quando verificados os testes obtidos a partir dos vetores completos, Tabela 5.14, é possível constatar que a melhor combinação foi definida com as características de lacunaridade (IVANOVICI; RICHARD; DECEAN, 2009). Porém, entre as cinco melhores combinações, três foram definidas a partir da entropia proposta. Este fato indica o potencial da abordagem em diferentes condições testadas.

Os atributos mais relevantes da entropia proposta também foram identificados para complementar o estudo sobre as classes benigna e maligna de câncer colorretal. Para este teste foi escolhido o método *Support Vector Machine Attribute Evaluation* (SV-MAE) (GUYON et al., 2002), técnica que fornece um *ranking* dos atributos por meio de pesos calculados e atribuídos por um classificador SVM. A identificação dos atributos mais relevantes foi a partir do vetor com 13 atributos extraídos da curva de entropia. Este vetor foi escolhido por apresentar os melhores resultados, conforme resumo mostrado na Tabela 5.5 e discutido previamente em detalhes. Os resultados obtidos com os 10 atributos mais relevantes estão na Tabela 5.15.

Tabela 5.15: Identificação dos 10 atributos mais relevantes para distinguir os grupos de câncer colorretal e os valores obtidos de AUC, sensibilidade (Se) e especificidade (Es).

Escalas	Atributos	Classificadores	AUC	Se	Es
$m = 1$	W O	<i>Multiclass Classifier e Logistic</i>	0,952	88,06%	89,29%
$m = 2$	W O Γ A				
$m = 3$	W O Γ A				

É importante destacar que mesmo com um número reduzido de atributos, a $SampEn_{MF}$ foi capaz de fornecer uma taxa de AUC de **0,952** em dois classificadores distintos (*Multiclass Classifier* e *Logistic*). Outro destaque é que as escalas mais relevantes da entropia foram definidas com $m = 2$ e $m = 3$, fornecendo 8 dos 10 atributos mais relevantes. Entre todos os atributos mais relevantes, W (indica o ponto máximo de cada função) e O (a obliquidade de cada curva da entropia) foram selecionados em todas as escalas de observações m , fato que merece destaque para o estudo do câncer colorretal. Quando as métricas sensibilidade (probabilidade do teste dar positivo na presença da doença) e especificidade (probabilidade do teste dar negativo na ausência da doença) são consideradas, é possível observar um certo equilíbrio para distinguir os grupos de interesse. A sensibilidade foi de 88,06% e a especificidade de 89,29%. Apesar do equilíbrio, uma taxa de sensibilidade sutilmente menor indica mais dificuldade para distinguir o grupo benigno com a $SampEn_{MF}$.

5.4 Desempenho da entropia Amostral frente aos trabalhos correlatos

Uma visão geral do desempenho fornecido pela entropia Amostral proposta, em relação aos fornecidos por trabalhos descritos na literatura focada no tema, é útil para indicar o potencial do método no estudo de imagens colorretais H&E. O resultado utilizado para esta ilustração foi obtido a partir da melhor combinação mostrada na Tabela 5.15. Os valores utilizados para esta visão geral estão na Tabela 5.16, considerando como referências as taxas de AUC.

Tabela 5.16: Desempenhos (valores de AUC) do método proposto ($SampEn_{MF}$) e de trabalhos correlatos, com informações sobre os atributos selecionados, classificadores e condições testadas entre os grupos avaliados.

Métodos	Atributos	Pré-processamento ou Segmentação	Imagens coloridas	Seleção	Classificador	AUC
Masood, Rajpoot (2009)	59-135	Não	Não	Sim	SVM	0,900
Kalkan et al. (2012)	120	Sim	Sim	Sim	<i>Logistic</i>	0,950
Fu et al. (2014)	47-77	Sim	Não	Sim	SVM	0,931
Naiyar, Asim e Shahid (2015)	14	Sim	Não	Não	SVM	0,960
Jørgensen et al. (2017)	9	Sim	Sim	Sim	<i>Random Forest</i>	0,960
SampEn_{MF}	10	Não	Sim	Sim	<i>Multiclass Classifier e Logistic</i>	0,952

É possível indicar que o desempenho proporcionado pela $SampEn_{MF}$ é compatível com os observados na literatura, por exemplo os descritos por Masood e Rajpoot (2009), Kalkan et al. (2012), Fu et al. (2014), Naiyar, Asim e Shahid (2015) e Jørgensen et al. (2017). É importante destacar que este tipo de comparação é limitada, principalmente sob o ponto de vista que as condições de testes e bases de imagens não foram as mesmas. Entretanto, esta visão geral permite mostrar que o método proposto foi capaz de apresentar resultados importantes para o estudo do câncer colorretal. Por exemplo, a abordagem proposta indicou um desempenho melhor em relação aos métodos de Masood e Rajpoot (2009), Fu et al. (2014) e Kalkan et al. (2012). Por outro lado, os resultados de Naiyar, Asim e Shahid (2015) e Jørgensen et al. (2017) foram sutilmente melhor. Apesar dessas constatações, acreditamos que uma contribuição importante do modelo proposto foi utilizar poucos atributos de entropia e sem a necessidade de realizar etapas de pré-processamento ou segmentação para permitir uma distinção relevante de grupos do câncer colorretal em imagens coloridas H&E.

5.5 Desempenhos das técnicas: tempos estimados

É importante observar que o tempos para processar as imagens H&E foram coletados, considerando os valores médios para quantificar os grupos benigno e maligno. Para a $SampEn_{MF}$ foram analisadas 90 janelas e o tempo médio foi de aproximadamente 0,46 segundos para cada uma. É importante notar que o tempo médio final pode variar em função do número de janelas escolhidas e demais intervalos de valores necessários para cada parâmetro. Para a H_M , o tempo médio de processamento também foi coletado: 0,74 segundos para cada sub-imagem e com escalas variando de 3 a 63. Entre estas técnicas, fica evidente que o tempo requerido para a entropia Amostral ($SampEn_{MF}$) é o principal problema para aplicação desta técnica. No entanto, vale ressaltar que este problema está presente nas demais técnicas baseadas na abordagem Amostral (SILVA et al., 2016; YEH; LIN; SHIEH, 2011), em decorrência do número

de comparações que esta técnica requer: $2 \times (N_{linha} - m + 1)^2 \times (N_{coluna} - m + 1)^2$, tornando-se o principal limitador para o processamento de imagens com elevadas taxas de amostragem.

Capítulo 6

Conclusões

Neste trabalho foram desenvolvidos dois métodos baseados em modelos de entropia (Amostragem e Shannon) para quantificar e classificar imagens colorretais H&E. Os métodos unificaram abordagens multiescala, multidimensional e *Fuzzy* para permitir quantificações das imagens coloridas.

O método baseado na entropia de Shannon, e conhecido como H_M , forneceu resultados menos expressivos do que os obtidos com a $SampEn_{MF}$. A melhor condição foi uma taxa de AUC de **0,873** a partir dos classificadores *Multiclass Classifier* e *Logistic*, considerando 36 atributos fornecido pela técnica H_M . Estes resultados não superaram as taxas fornecidas por alguns modelos encontrados na literatura, porém foram melhores do que os obtidos com a entropia de Shannon clássica, H . Este fato é relevante por indicar que as alterações realizadas são promissoras e podem ser mais bem exploradas em outros tipos de imagens.

Em contrapartida, os resultados conquistados com a $SampEn_{MF}$ foram superiores aos fornecidos pelo método H_M , além de superar um modelo de entropia Amostral desenvolvido para quantificações de imagem monocromática, a $SampEn_{2D}$ (SILVA et al., 2016). O desempenho do nosso método também foi comparado aos fornecidos por outras técnicas multiescalas e multidimensionais, tais como a DF (IVANOVICI; RICHARD, 2011; NIKOLAIDIS; NIKOLAIDIS; TSOUROS, 2011) e lacunaridade

(IVANOVICI; RICHARD; DECEAN, 2009). Nos testes realizados foi possível identificar claramente que nossa proposta forneceu resultados superiores aos da técnica DF. Em comparação com a lacunaridade, apesar das taxas médias de AUC serem similares, a entropia Amostral proposta indicou a maior taxa de AUC (**0,940**), com 13 atributos, contra **0,918** da lacunaridade que foi obtida com 27 atributos. Uma explicação é que a entropia Amostral com as modificações foi capaz de quantificar mais apropriadamente padrões histológicos presentes nas imagens, identificando as similaridades entre regiões vizinhas a partir de informações presentes nos canais de cores das imagens H&E. Estas constatações indicam que o trabalho proposto fornece contribuições importantes para literatura especializada em temas como entropia e estudo do câncer colorretal.

O melhor desempenho fornecido pela nossa proposta foi uma taxa de AUC de **0,952** calculada com 10 atributos e em diferentes classificadores *Multiclass Classifier* e *Logistic*. Neste cenário, as escalas mais importantes da entropia proposta foram definidas com $m = 2$ e $m = 3$, fornecendo 8 dos 10 atributos. Além disso, os atributos que definem o ponto máximo de cada função W e a obliquidade de cada curva da entropia O foram selecionados em todas as escalas m . Por fim, no que tange uma visão geral da $SampEn_{MF}$ em relação aos métodos importantes disponíveis na literatura (MASOOD; RAJPOOT, 2009; KALKAN et al., 2012; FU et al., 2014; NAIYAR; ASIM; SHAHID, 2015; JØRGENSEN et al., 2017), o nosso método também foi capaz de fornecer um desempenho compatível, com a vantagem de utilizar poucos atributos e evitar o processo de conversão das imagens coloridas para níveis de cinza ou aplicação de etapas de pré-processamento.

6.1 Contribuições obtidas

Com o desenvolvimento deste trabalho, as contribuições obtidas foram:

- I Dois novos métodos para quantificar padrões presentes em imagens coloridas de câncer colorretal, por meio da associação de entropia com estratégias multiescala, multidimensional e *Fuzzy*;
- II A indicação da técnica mais apropriada, parâmetros e associações mais relevantes do vetor de características para classificar e distinguir os grupos benigno e maligno de imagens colorretais H&E;
- III Um método baseado no modelo de entropia Amostral capaz de constituir vetores de características que podem complementar quantificações realizadas com outras abordagens multiescalas e multidimensionais.

6.2 Trabalhos futuros

Em trabalhos futuros, é possível sugerir mais testes e pontos que podem aprimorar os modelos:

- I testar os métodos (Amostral e Shannon) em novas bases de imagens coloridas segmentadas ou não, bem como verificar a influência de diferentes métricas no processo de quantificação entre janelas;
- II testar diferentes modelos de cores nos métodos para constatar possíveis limitações;
- III definir tolerâncias específicas para cada canal de cor e observar os impactos no processo de quantificação em diferentes bases de imagens coloridas;
- IV realizar testes com diferentes quantidades de janelas aleatórias e diferentes tamanhos de sub-imagens, com esforços direcionados para a otimização do método proposto, problema também presente em outros modelos baseados em entropia Amostral;

V investigar os comportamentos dos vetores de características em um número maior de classificadores, bem como testar diferentes funções *Fuzzy* podem contribuir para o aprimoramento e o entendimento dos limites do modelo *SampEn_{MF}*.

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.

ALARCÓN, T. E.; MARROQUÍN, J. L. Linguistic color image segmentation using a hierarchical bayesian approach. *Color Research & Application*, Wiley Online Library, v. 34, n. 4, p. 299–309, 2009.

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

ALTUNBAY, D. et al. Color graphs for automated cancer diagnosis and grading. *IEEE Transactions on Biomedical Engineering*, IEEE, v. 57, n. 3, p. 665–674, 2010.

ALVES, R. F. et al. *Uma metodologia de binarização para áreas de imagens de cheque utilizando algoritmos de aprendizagem supervisionada*. 2015.

ANANTHI, V.; BALASUBRAMANIAM, P.; LIM, C. P. Segmentation of gray scale image based on intuitionistic fuzzy sets constructed from several membership functions. *Pattern Recognition*, Elsevier, v. 47, n. 12, p. 3870–3880, 2014.

ARTUSO, A. R. Entropias de shannon e rényi aplicadas ao reconhecimento de padrões. *Revista CIATEC-UPF*, v. 3, n. 2, p. 56–72, 2012.

BALASUBRAMANIAN, K.; NAIR, S. S.; NAGARAJ, N. Classification of periodic, chaotic and random sequences using approximate entropy and lempel–ziv complexity measures. *Pramana*, Springer, v. 84, n. 3, p. 365–372, 2015.

BAWDEN, D.; ROBINSON, L. “a few exciting words”: Information and entropy revisited. *Journal of the Association for Information Science and Technology*, Wiley Online Library, v. 66, n. 10, p. 1965–1987, 2015.

BENDER, T. C. *Classificação e recuperação de imagens por cor utilizando técnicas de inteligência artificial*. Dissertação (Mestrado) — Universidade do Vale do Rio dos Sinos, Curso de Pós-graduação em Computação Aplicada, 2003.

BIJAR, A.; KHAYATI, R.; BENAVENT, A. P. Increasing the contrast of the brain mr flair images using fuzzy membership functions and structural similarity indices in order to segment ms lesions. *PloS one*, Public Library of Science, v. 8, n. 6, p. e65469, 2013.

BOLTZ, S.; NIELSEN, F.; SOATTO, S. Texture regimes for entropy-based multiscale image analysis. *Computer Vision–ECCV 2010*, Springer, p. 692–705, 2010.

CĂLIMAN, A.; IVANOVICI, M. Psoriasis image analysis using color lacunarity. In: IEEE. *Optimization of Electrical and Electronic Equipment (OPTIM)*, 2012 13th International Conference on. [S.l.], 2012. p. 1401–1406.

CARVALHO, H. M. *Aprendizado de máquina voltado para mineração de dados: árvores de decisão*. 2015.

CHEN, W. et al. Measuring complexity using fuzzyen, apen, and sampen. *Medical Engineering & Physics*, Elsevier, v. 31, n. 1, p. 61–68, 2009.

CHENG, H.; LI, J. Fuzzy homogeneity and scale-space approach to color image segmentation. *Pattern recognition*, Elsevier, v. 36, n. 7, p. 1545–1562, 2003.

COLLINS, C. A. Impacts of enso on tornado frequency, intensity, and geography across the eastern united states. 2017.

CONVERTINO, M. et al. Inferring species richness and turnover by statistical multi-resolution texture analysis of satellite imagery. *PloS one*, Public Library of Science, v. 7, n. 10, p. e46616, 2012.

COTTI, G. C. de C. et al. Genética do câncer colorretal. *Revista de Medicina*, v. 79, n. 2-4, p. 65–72, 2000.

COVER, T. M.; THOMAS, J. A. *Elements of information theory*. [S.l.]: John Wiley & Sons, 2012.

DAMES, K. K.; MATTOS, R. A.; MELO, P. L. Aprimoramento da entropia fuzzy da vazão respiratória visando o diagnóstico precoce da doença pulmonar obstrutiva crônica. 2014.

DHIEB, M. et al. 2-d entropy image segmentation on thresholding based on particle swarm optimization (pso). In: IEEE. *Advanced Technologies for Signal and Image Processing (ATSIP)*, 2014 1st International Conference on. [S.l.], 2014. p. 143–147.

EKSTRÖM, M. et al. Logistic regression for clustered data from environmental monitoring programs. *Ecological Informatics*, Elsevier, v. 43, p. 165–173, 2018.

FIGUEIREDO, M. A. *Elementos de Teoria da Informação*. 2011.

FU, J. J. et al. Feature extraction and pattern classification of colorectal polyps in colonoscopic imaging. *Computerized Medical Imaging and Graphics*, Elsevier, v. 38, n. 4, p. 267–275, 2014.

GARNER, S. R. et al. Weka: The waikato environment for knowledge analysis. In: CITESEER. *Proceedings of the New Zealand computer science research students conference*. [S.l.], 1995. p. 57–64.

GRIGGS, R. K. et al. Challenging diagnostic issues in adenomatous polyps with epithelial misplacement in bowel cancer screening: 5 years' experience of the bowel cancer screening programme expert board. *Histopathology*, Wiley Online Library, v. 70, n. 3, p. 466–472, 2017.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. *Machine learning*, Springer, v. 46, n. 1, p. 389–422, 2002.

HANI, A. F. M.; SATHYAMOORTHY, D.; ASIRVADAM, V. S. Computing uncertainty of physiographic features extracted from multiscale digital elevation models. *Computers & Geosciences*, Elsevier, v. 64, p. 15–23, 2014.

HANMADLU, M. et al. A novel optimal fuzzy color image enhancement using particle swarm optimization. In: IEEE. *Contemporary Computing (IC3), 2013 Sixth International Conference on*. [S.l.], 2013. p. 41–46.

HANMANDLU, M.; JHA, D. An optimal fuzzy system for color image enhancement. *IEEE Transactions on image processing*, IEEE, v. 15, n. 10, p. 2956–2966, 2006.

HORNUNG, R. et al. Full versus incomplete cross-validation: measuring the impact of imperfect separation between training and test sets in prediction error estimation. 2014.

HUANG, L.; HE, D.; YANG, S. X. Segmentation on ripe fuji apple with fuzzy 2d entropy based on 2d histogram and ga optimization. *Intelligent Automation & Soft Computing*, Taylor & Francis, v. 19, n. 3, p. 239–251, 2013.

HUANG, L.-K.; WANG, M.-J. J. Image thresholding by minimizing the measures of fuzziness. *Pattern recognition*, Elsevier, v. 28, n. 1, p. 41–51, 1995.

HUMEAU-HEURTIER, A. The multiscale entropy algorithm and its variants: a review. *Entropy*, Multidisciplinary Digital Publishing Institute, v. 17, n. 5, p. 3110–3123, 2015.

HUMEAU-HEURTIER, A. et al. Multiscale entropy study of medical laser speckle contrast images. *IEEE Transactions on Biomedical Engineering*, IEEE, v. 60, n. 3, p. 872–879, 2013.

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, 2018.

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.

JATI, A. et al. Automatic leukocyte nucleus segmentation by intuitionistic fuzzy divergence based thresholding. *Micron*, Elsevier, v. 58, p. 55–65, 2014.

JØRGENSEN, A. S. et al. Using cell nuclei features to detect colon cancer tissue in hematoxylin and eosin stained slides. *Cytometry Part A*, Wiley Online Library, 2017.

JUNIOR, G. *Máquina de Vetores Suporte: estudo e análise de parâmetros para otimização de resultado*. 2010.

KALKAN, H. et al. Automated classification of local patches in colon histopathology. In: IEEE. *Pattern Recognition (ICPR), 2012 21st International Conference on*. [S.l.], 2012. p. 61–64.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: STANFORD, CA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

LAUREANO, R.; CAETANO, N.; CORTEZ, P. Previsão de tempos de internamento num hospital português: aplicação da metodologia crisp-dm. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*, Associação Ibérica de Sistemas e Tecnologias de Informação (AISTI), n. 13, p. 83–98, 2014.

LI, P. et al. Testing pattern synchronization in coupled systems through different entropy-based measures. *Medical & biological engineering & computing*, Springer, v. 51, n. 5, p. 581–591, 2013.

LIANG, Z. et al. Eeg entropy measures in anesthesia. *Frontiers in computational neuroscience*, Frontiers, v. 9, p. 16, 2015.

LIU, D.; JIANG, Z.; FENG, H. A novel fuzzy classification entropy approach to image thresholding. *Pattern Recognition Letters*, Elsevier, v. 27, n. 16, p. 1968–1975, 2006.

LLADÓ, X. et al. Segmentation of multiple sclerosis lesions in brain mri: a review of automated approaches. *Information Sciences*, Elsevier, v. 186, n. 1, p. 164–185, 2012.

LORENA, A. C.; CARVALHO, A. C. de. Introduções máquinas de vetores suporte. *Relatório Técnico do Instituto de Ciências Matemáticas e de Computação (USP/Sao Carlos)*, v. 192, 2003.

LU, Y. et al. Fuzzy entropy thresholding method using adaptive genetic algorithm. *Journal of applied science and engineering innovation Vol*, v. 1, n. 1, 2014.

MARXER, R.; PURWINS, H. Unsupervised incremental learning and prediction of music signals. *arXiv preprint arXiv:1502.00524*, 2015.

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.

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>>.

MCKIGHT, P. E.; NAJAB, J. Kruskal-wallis test. *Corsini Encyclopedia of Psychology*, Wiley Online Library, 2010.

NAHAS, S. C. et al. Prognostic factors of surgically-treated patients with cancer of the right colon: a ten years' experience of a single university institution. *ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo)*, SciELO Brasil, v. 28, n. 1, p. 03–07, 2015.

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.

NARDELLI, M. et al. Characterizing psychological dimensions in non-pathological subjects through autonomic nervous system dynamics. *Frontiers in computational neuroscience*, Frontiers Media SA, v. 9, 2015.

NEVES, L. A. et al. Multi-scale lacunarity as an alternative to quantify and diagnose the behavior of prostate cancer. *Expert Systems with Applications*, Elsevier, v. 41, n. 11, p. 5017–5029, 2014.

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

OLIVEIRA, H. R. S. de; ASSIS, L. B. de; BRAGA, E. Q. Análise estatística no reconhecimento automático de faces utilizando pca eigenfaces. *e-xacta*, v. 5, n. 2, 2012.

OLIVO, C. K.; SANTIN, A. O.; OLIVEIRA, L. S. Obtaining the threat model for e-mail phishing. *Applied soft computing*, Elsevier, v. 13, n. 12, p. 4841–4848, 2013.

OSHIRO, T. M. *Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica*. 2013.

OSWAL, V. et al. An entropy-based automated cell nuclei segmentation and quantification: application in analysis of wound healing process. *Computational and mathematical methods in medicine*, Hindawi Publishing Corporation, v. 2013, 2013.

PAVÃO, R. *Entropia informacional e aprendizagem de sequências*. Tese (Doutorado) — Universidade de São Paulo, 2011.

PEREIRA, L. A. d. A. *Classificação automática de áreas cafeeiras em imagens de satélite, utilizando Redes Neurais Artificiais*. 2013.

PINEDA, A. L. et al. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *Journal of biomedical informatics*, Elsevier, v. 58, p. 60–69, 2015.

- PINEDA, J. O. d. C. *A entropia segundo Claude Shannon: o desenvolvimento do conceito fundamental da teoria da informação*. 2006. 124 f. 2006.
- PRADHAN, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using gis. *Computers & Geosciences*, Elsevier, v. 51, p. 350–365, 2013.
- ROBERTO, G. F. et al. Features based on the percolation theory for quantification of non-hodgkin lymphomas. *Computers in biology and medicine*, Elsevier, v. 91, p. 135–147, 2017.
- RODRÍGUEZ, J.; BÁRDOSSY, A. Entropy measure for the quantification of upper quantile interdependence in multivariate distributions. *Journal of Multivariate Analysis*, Elsevier, v. 140, p. 317–324, 2015.
- RODRIGUEZ, J. J.; KUNCHEVA, L. I.; ALONSO, C. J. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 28, n. 10, p. 1619–1630, 2006.
- RONG, C. Image segmentation analysis based on maximum entropy algorithm. 2015.
- SANTOS, M. J. d. S. *Uso da Entropia de Shannon para classificação de zonas artificiais ou naturais em imagens*. 2013.
- SATO, L. Y. et al. Análise comparativa de algoritmos de árvore de decisão do sistema weka para classificação do uso e cobertura da terra. *XVI Simpósio Brasileiro de Sensoriamento Remoto*, p. 2353–2360, 2013.
- SEYEDZADEH, S. M.; NOROUZI, B.; MIRZAKUCHAKI, S. Rgb color image encryption based on choquet fuzzy integral. *Journal of Systems and Software*, Elsevier, v. 97, p. 128–139, 2014.
- SIEGEL, R. L.; MILLER, K. D.; JEMAL, A. Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, v. 68, n. 1, p. 7–30, 2018. ISSN 1542-4863. Disponível em: <<http://dx.doi.org/10.3322/caac.21442>>.
- SILVA, L. E. V. et al. Two-dimensional sample entropy: assessing image texture through irregularity. *Biomedical Physics & Engineering Express*, IOP Publishing, v. 2, n. 4, p. 045002, 2016.
- SILVA, L. E. V. da et al. Two-dimensional sample entropy analysis of rat sural nerve aging. In: IEEE. *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. [S.l.], 2014. p. 3345–3348.
- SIRINUKUNWATTANA, K. et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, Elsevier, v. 35, p. 489–502, 2017.
- SOUZA, W. F. d. et al. Sinalização celular em câncer. *Ciência e Cultura*, Sociedade Brasileira para o Progresso da Ciência, v. 66, n. 1, p. 30–33, 2014.

TRIBA, M. N. et al. Pls/opls models in metabolomics: the impact of permutation of dataset rows on the k-fold cross-validation quality parameters. *Molecular BioSystems*, Royal Society of Chemistry, v. 11, n. 1, p. 13–19, 2015.

TSIAPARAS, N. et al. Assessment of carotid atherosclerosis from b-mode ultrasound images using directional multiscale texture features. *Measurement Science and Technology*, IOP Publishing, v. 23, n. 11, p. 114004, 2012.

VANDENDRIESSCHE, B. et al. A multiscale entropy-based tool for scoring severity of systemic inflammation. *Critical care medicine*, LWW, v. 42, n. 8, p. e560–e569, 2014.

VARDHAN, S. A. et al. Automatic detection of orientation of mapped units via directional granulometric analysis. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 10, n. 6, p. 1449–1453, 2013.

VERMA, A.; SARANGI, S.; KOLEKAR, M. H. Stator winding fault prediction of induction motors using multiscale entropy and grey fuzzy optimization methods. *Computers & Electrical Engineering*, Elsevier, v. 40, n. 7, p. 2246–2258, 2014.

VERMA, O. P. et al. High dynamic range optimal fuzzy color image enhancement using artificial ant colony system. *Applied Soft Computing*, Elsevier, v. 12, n. 1, p. 394–404, 2012.

VIVEKANAND, A.; WERGHI, N.; AL-AHMAD, H. Multiscale roughness approach for assessing posterior capsule opacification. *IEEE journal of biomedical and health informatics*, IEEE, v. 18, n. 6, p. 1923–1931, 2014.

WANG, Y.; WANG, X.; LIU, W. Unsupervised local deep feature for image recognition. *Information Sciences*, Elsevier, v. 351, p. 67–75, 2016.

XIE, H.; GUO, J.; ZHENG, Y. Fuzzy approximate entropy analysis of chaotic and natural complex systems: detecting muscle fatigue using electromyography signals. *Annals of biomedical engineering*, Springer, v. 38, n. 4, p. 1483–1496, 2010.

XIONG, G. et al. A comparative study on apen, sampen and their fuzzy counterparts in a multiscale framework for feature extraction. *Journal of Zhejiang University SCIENCE A*, Springer, v. 11, n. 4, p. 270–279, 2010.

XU, S. et al. Similarity measures for content-based image retrieval based on intuitionistic fuzzy set theory. *JCP*, v. 7, n. 7, p. 1733–1742, 2012.

YANG, J. et al. On evaluating multi-class network traffic classifiers based on auc. *Wireless Personal Communications*, Springer, v. 83, n. 3, p. 1731–1750, 2015.

YE, Z.-W. et al. Fuzzy entropy based optimal thresholding using bat algorithm. *Applied Soft Computing*, Elsevier, v. 31, p. 381–395, 2015.

YEH, J.; LIN, C.; SHIEH, J. An approach of multiscale complexity in texture analysis of lymphomas. *IEEE Signal Processing Letters*, IEEE, v. 18, n. 4, p. 239–242, 2011.

YU, Z. et al. A new kind of nonparametric test for statistical comparison of multiple classifiers over multiple datasets. *IEEE transactions on cybernetics*, IEEE, 2016.

ZHANG, L. et al. Bearing fault diagnosis using multi-scale entropy and adaptive neuro-fuzzy inference. *Expert Systems with Applications*, Elsevier, v. 37, n. 8, p. 6077–6085, 2010.

ZHOU, N. et al. Global and local structure preserving sparse subspace learning: An iterative approach to unsupervised feature selection. *Pattern Recognition*, Elsevier, v. 53, p. 87–101, 2016.

ZOU, Q. et al. Improving trnascan-se annotation results via ensemble classifiers. *Molecular informatics*, Wiley Online Library, v. 34, n. 11-12, p. 761–770, 2015.