

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

João Marcos de Freitas Moreira Gomes

**Classificação de Núcleos com Diferentes Níveis  
de Displasia em Imagens Histológicas**

**Uberlândia, Brasil**

**2021**

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

João Marcos de Freitas Moreira Gomes

**Classificação de Núcleos com Diferentes Níveis de  
Displasia em Imagens Histológicas**

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Marcelo Zanchetta do Nascimento

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2021

*Dedico este trabalho à minha família e amigos.*

# Agradecimentos

Agradeço aos meus pais pelos sacrifícios feitos para que eu pudesse chegar até aqui, aos meus amigos por me ajudarem nos momentos difíceis, ao professor Marcelo por todo o conhecimento transmitido que tornou possível o desenvolvimento deste trabalho e à minha irmã.



*“Há coisas muito melhores à frente do que qualquer uma das que deixamos pra trás.”*

*C. S. Lewis*

# Resumo

Com os inúmeros avanços tecnológicos que ocorreram nos últimos anos, os computadores passaram a ser usados como auxiliares nos processos de diagnósticos de doenças, sendo uma delas, o câncer. O câncer é uma doença que ainda mata milhares de pessoas todos os anos, sendo o câncer de boca, o quinto tipo de câncer mais comum em homens no país. Com técnicas de processamento de imagem, extração e seleção de características e algoritmos de classificação, este trabalho avalia a eficiência de descritores na representação das imagens de diversos graus de displasia epitelial oral. O trabalho propõe um sistema capaz de avaliar as imagens e classificá-las entre os grupos saudável, leve, moderada e severa, de acordo com o grau de displasia presente na imagem, usando associações entre as técnicas apresentadas na literatura. Com o uso do classificador *Random Forest* e descritores morfológicos, descritores de Haralick, matriz de co-ocorrência estrutural, índice de Moran e entropias, foi possível obter um classificador de lesões multi-classe com acurácia de 82,5%.

**Palavras-chave:** Displasia, Classificação, Extração de Características, Random Forest.

# Lista de ilustrações

Figura 1 – Alterações ocorridas nos tecidos em diferentes graus de displasia. (a) Tecido saudável. (b) Tecido com displasia leve. (c) Tecido com displasia moderada. (d) Tecido com displasia severa. Fonte: (NAG; DAS, 2018) . . .	16
Figura 2 – Micrótopo para cortar tecidos incluídos em parafina. Acionando-se a manivela (à direita da figura), o bloco contendo o fragmento de tecido sobe e desce. Após cada volta da manivela, o bloco avança uma distância definida (geralmente de um a dez $\mu\text{m}$ ) e, ao passar pela navalha, deixa uma fatia do tecido (JUNQUEIRA; CARNEIRO, 2018). . . . .	19
Figura 3 – Microscópios de luz: (a) desenho esquemático de um microscópio de luz. (JUNQUEIRA; CARNEIRO, 2018), (b) microscópio Leica DM500, usado na obtenção das imagens do banco usado neste trabalho (MICROSYSTEMS, ). . . . .	20
Figura 4 – Diagrama esquemático do fluxo de funcionamento de um sistema CAD (KAUSHAL et al., 2019) . . . . .	22
Figura 5 – Etapas do sistema de classificação de lesões da cavidade bucal. . . . .	32
Figura 6 – Imagens de tecido com displasia moderada: (a) Imagem do tecido corado, pronta para a extração de características, (b) Imagem segmentada, pronta para a extração de características morfológicas. . . . .	34
Figura 7 – Figura esquemática do funcionamento de um algoritmo de floresta de decisão (VERIKAS et al., 2016) . . . . .	36
Figura 8 – Função de distribuição acumulada empírica do <i>p-value</i> das classes de displasia com os descritores morfológicos. . . . .	38
Figura 9 – Função de distribuição acumulada empírica do <i>p-value</i> das classes de displasia com os descritores de Haralick. . . . .	39
Figura 10 – Função de distribuição acumulada empírica do <i>p-value</i> das classes de displasia com os descritores da SCM. . . . .	40
Figura 11 – Função de distribuição acumulada empírica do <i>p-value</i> das classes de displasia com os descritores de entropia. . . . .	40
Figura 12 – Função de distribuição acumulada empírica do <i>p-value</i> das classes de displasia com todos os descritores (exceto entropias). . . . .	41
Figura 13 – Mapa de calor da correlação dos descritores morfológicos. . . . .	41
Figura 14 – Mapa de calor avaliando a correlação dos descritores de Haralick. . . . .	42
Figura 15 – Investigação da correlação dos descritores da SCM. . . . .	42
Figura 16 – Investigação da correlação dos descritores obtidos com entropia. . . . .	43
Figura 17 – Frequência de representação dos descritores. . . . .	46
Figura 18 – Representação da métrica acurácia por número de descritores. . . . .	46

Figura 19 – Função de distribuição acumulada com os descritores selecionados com L1. . . . .	47
Figura 20 – Mapa de calor de correlação dos descritores selecionados com L1. . . . .	47
Figura 21 – Matriz de confusão do classificador <i>Random Forest</i> . . . . .	49

# Lista de tabelas

Tabela 1 – Alterações histológicas que caracterizam displasias epiteliais (Traduzido de (RANGANATHAN; KAVITHA, 2019)). . . . .	16
Tabela 2 – Acurácia obtida com uso dos descritores morfológicos. . . . .	44
Tabela 3 – Acurácia obtida com uso dos descritores não morfológicos baseados em matrizes de co-ocorrências. . . . .	44
Tabela 4 – Acurácia obtida com uso do classificador Random Forest. . . . .	45
Tabela 5 – Acurácia obtida com uso do classificador <i>Random Forest</i> . . . . .	48
Tabela 6 – Métricas para avaliação com o classificador Random Forest. . . . .	49

# Lista de abreviaturas e siglas

AUC	<i>Area Under Curve</i>
CAD	<i>Computer Aided Detection and Diagnosis</i>
CADe	<i>Computer-Aided Detection</i>
CADx	<i>Computer-Aided Diagnosis</i>
GLCM	<i>Gray Level Co-occurrence Matrix</i>
H&E	Hematoxilina e Eosina
LASSO	<i>Least Absolute Shrinkage and Selection Operator</i>
RaFs	<i>Random Forests</i>
RGB	<i>Red Green Blue</i>
ROC	<i>Receiver Operating Characteristic</i>
ROI	<i>Region Of Interest</i>
RSS	Soma Residual dos Quadrados
SCM	<i>Structural Co-occurrence Matrix</i>
TIFF	<i>Tagged Image File Format</i>

# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>12</b>
<b>1.1</b>	<b>Objetivos</b>	<b>14</b>
1.1.1	Geral	14
1.1.2	Específicos	14
<b>1.2</b>	<b>Organização desse Trabalho de Conclusão de Curso</b>	<b>14</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>15</b>
<b>2.1</b>	<b>Câncer de Boca</b>	<b>15</b>
<b>2.2</b>	<b>Histologia</b>	<b>16</b>
2.2.1	Preparação de Tecidos	17
<b>2.3</b>	<b>Microscopia</b>	<b>18</b>
<b>2.4</b>	<b>Processamento Digital de Imagens</b>	<b>20</b>
<b>2.5</b>	<b>Sistemas de Diagnóstico Auxiliado por Computador</b>	<b>21</b>
<b>2.6</b>	<b>Extração de Características</b>	<b>23</b>
2.6.1	Descritores Morfológicos	24
2.6.1.1	Área	24
2.6.1.2	Excentricidade	24
2.6.1.3	Orientação	24
2.6.1.4	Perímetro	25
2.6.1.5	Solidez	25
2.6.2	Descritores Não-Morfológicos	25
2.6.2.1	Índice de Moran	25
2.6.2.2	Entropia	25
2.6.2.3	Matriz de Co-ocorrência de Níveis de Cinza	26
2.6.2.4	Matriz de Co-ocorrência Estrutural	28
<b>2.7</b>	<b>Trabalhos Correlatos</b>	<b>30</b>
<b>2.8</b>	<b>Considerações Finais</b>	<b>31</b>
<b>3</b>	<b>MATERIAIS E MÉTODOS</b>	<b>32</b>
<b>3.1</b>	<b>Visão Geral</b>	<b>32</b>
<b>3.2</b>	<b>Linguagens e Ferramentas</b>	<b>32</b>
<b>3.3</b>	<b>Banco de Imagens</b>	<b>33</b>
<b>3.4</b>	<b>Extração de características</b>	<b>33</b>
<b>3.5</b>	<b>Redução de dimensionalidade</b>	<b>34</b>
<b>3.6</b>	<b>Classificação</b>	<b>35</b>
<b>3.7</b>	<b>Avaliação dos Métodos</b>	<b>36</b>

<b>4</b>	<b>RESULTADOS</b>	<b>38</b>
4.1	Análise dos Descritores Morfológicos e Não Morfológicos	38
4.2	Classificação com os Descritores	43
4.3	Classificação com Associação de Descritores	45
4.4	Avaliação do Regularizador L1	45
4.5	Avaliação do Classificador	48
4.6	Comparação com a Literatura	50
<b>5</b>	<b>CONCLUSÃO</b>	<b>51</b>
5.1	Trabalhos Futuros	51
	<b>REFERÊNCIAS</b>	<b>53</b>



# 1 Introdução

Os estudos na área de sistemas computacionais de diagnóstico não são uma novidade, datando até mesmo do ano de 1963, apenas 16 anos após a criação do primeiro computador digital eletrônico de grande escala, demonstrando já um esboço do que se evoluiria para os sistemas de auxílio ao diagnóstico e à detecção (*Computer-Aided Detection and Diagnosis*, ou CAD) atuais. As primeiras versões desses sistemas não tiveram muito sucesso, em partes por se basearem em uma visão otimista do poder de processamento computacional, que era limitado na época, e também por supor que esses sistemas seriam capazes de automatizar por completo o processo de diagnóstico, incorporando o pensamento humano em um sistema computadorizado (LODWICK et al., 1963; JR, 1992).

A partir dos anos 80, o pensamento deixou de ser o de substituir profissionais para passar a auxiliá-los, com os sistemas de computadores sendo usados como uma confirmação do diagnóstico, criando uma sinergia entre o sistema e o profissional. Tal sistema permitiria então uma soma de conhecimentos, mas deixaria ainda a cargo do profissional a escolha de concordar ou não com o diagnóstico. Apesar de sistemas CAD incorporarem num software único um sistema para detecção e um de diagnóstico (FIRMINO et al., 2016), essas soluções podem também ser separadas em dois grupos: sistemas computacionais de detecção (do inglês, *computer aided detection* - CAdE), que são responsáveis por auxiliar no processo de detecção, desempenhando as etapas de segmentação, pós processamento e extração de características; e sistemas computacionais de auxílio ao diagnóstico (do inglês, *computer aided diagnosis* - CAdx), que operam no nível de classificação, usando as informações que a ele são dadas e ajudando no processo de diagnóstico, como por exemplo, na definição de lesões como malignas ou benignas.

Um bom uso de sistemas CAD pode ser observado na análise de lesões de cânceres como o de boca. Por ser geralmente assintomático nos estágios iniciais, onde o tecido sofre várias alterações - entre elas, as displasias - um dos desafios encontrados em relação ao câncer de boca é que boa parte dos casos fatais pode ser associada não necessariamente à agressividade do câncer, mas à demora na busca por auxílio médico (INCA, 2020), que em muitos casos ocorre já em estágios avançados do tumor. Tal negligência reduz a chance de sobrevivência do paciente de 91% (em casos de diagnóstico precoce) para 50% (FOUNDATION, 2018) demandando uma maior urgência no processo de classificação da doença. Alguns dos principais fatores de risco para o câncer de boca são o tabagismo e o etilismo, sendo o fumo responsável por cerca de 42% dos óbitos, e o álcool, aproximadamente 16%. Outros fatores também apontados são dieta pobre em frutas e vegetais, o consumo de erva mate *Ilex paraguariensis* e as infecções pelo HPV (BRASIL; CÂNCER, 2011; WARNAKULASURIYA, 2009).

Analisando as estatísticas, percebe-se que o estudo da doença, assim como a classificação das alterações no tecido em um estágio inicial, é de extrema importância, mas mesmo com os diversos avanços tecnológicos introduzidos à área médica nos últimos anos, ao se analisar as estatísticas referentes ao câncer de boca, percebe-se que ele já é o quinto tipo de câncer mais comum entre homens no Brasil (INCA, 2020). Nota-se também que o número esperado de novos casos para o biênio 2018-2019 era de aproximadamente 10457, e 4316 mortes de câncer por ano, representando cerca de 5,2% de todos os casos de câncer entre homens no país, ficando atrás apenas dos cânceres de próstata (31,7%), pulmão (8,7%), intestino (8,1%) e estômago (6,3%) (INCA, 2020).

Para que possa ser feito o diagnóstico da lesão cancerosa, é necessário um estudo histológico realizado por um profissional habilitado. O profissional deve efetuar uma raspagem da mucosa oral, seguida pelo uso de uma técnica de coloração, como por exemplo o uso de hematoxilina e eosina, que são respectivamente: um corante catiônico azul-violeta, rico em cargas positivas; e um corante aniônico rosa, rico em cargas negativas. Essa técnica permite a correta visualização e facilita a classificação das organelas presentes no tecido (MONTANARI, 2016) que depois serão analisadas pelo patologista por meio de um microscópio ou de forma digitalizada. As alterações na mucosa poderão então ser classificadas corretamente de acordo com o grau de displasia, ou seja, do quanto do tecido epitelial foi afetado. Fatores como acuidade visual, cansaço, desatenção por parte do profissional e seu nível de experiência podem levar a diferentes classificações do tecido. Por esses motivos, o diagnóstico acaba sendo subjetivo, permitindo discrepância entre classificações de diferentes patologistas.

Caso o profissional chegue a um diagnóstico que não o deixe convicto, pode ser que ele sugira uma contra-análise a outro patologista para confirmação, o que pode demorar, dependendo da carga de trabalho, da disponibilidade e da existência de um outro profissional no local. Caso o último requisito não possa ser cumprido, a análise da lesão por meio de uma amostra digitalizada do tecido permite que o especialista envie somente as imagens para outro profissional, de forma mais rápida e segura do que com o uso de lâminas, uma vez que, se as imagens forem danificadas na entrega, seu reenvio pode ser feito facilmente. Persiste ainda o problema de encontrar um segundo profissional que esteja disponível, mesmo que remotamente, para atender a demanda do patologista no mesmo momento. Nesses casos, o uso de sistemas CAD pode ser benéfico. Com diversos algoritmos de extração de características e classificação disponíveis para esses sistemas CAD, um grande desafio é encontrar quais algoritmos conseguem melhor acurácia na descrição de características e no processo de classificação das displasias orais. Este trabalho se propõe a investigar algoritmos que permitam descrever informações das lesões na etapa de extração de características e métodos de classificação, fazendo uma análise comparativa dos resultados obtidos com seus usos.

## 1.1 Objetivos

### 1.1.1 Geral

Desenvolvimento de métodos para classificação de lesões de displasia da cavidade bucal usando descritores morfológicos e não-morfológicos. Pretende-se investigar esses descritores no processo de classificação dos diversos níveis de displasia, de forma a auxiliar no avanço dos estudos sobre essa doença.

### 1.1.2 Específicos

- Investigação de algoritmos de extração de características morfológicas e não-morfológicas sobre os diversos níveis de displasia;
- Investigação do algoritmo de redução da dimensionalidade baseado em regularização  $L1$  (Lasso) para classificação das lesões;
- Investigação da influência do algoritmo de florestas aleatórias no desempenho dos descritores para os grupos de tecidos histológicos;
- Discretização da importância das características para o classificador de acordo com os tipos de lesões nas imagens histológicas.

## 1.2 Organização desse Trabalho de Conclusão de Curso

Neste capítulo, foram apresentadas as considerações iniciais, motivação e objetivos do desenvolvimento deste trabalho. O restante deste trabalho possui a seguinte organização:

- Capítulo 2: Introduce os principais conceitos para a compreensão e validação do trabalho proposto e apresenta os trabalhos correlatos relacionados com a proposta;
- Capítulo 3: Descreve o trabalho proposto, bem como as técnicas e metodologias desenvolvidas;
- Capítulo 4: Apresenta os resultados alcançados e as respectivas análises do método de classificação, além de um comparativo com a literatura;
- Capítulo 5: São apresentadas as contribuições obtidas, considerações finais e trabalhos futuros.

## 2 Fundamentação Teórica

Neste capítulo, são apresentados conceitos fundamentais para o entendimento deste trabalho. Como ele aborda temas de duas áreas distintas — computação e biologia — esses temas são apresentados aqui de forma introdutória.

### 2.1 Câncer de Boca

O câncer de boca já é o quinto tipo de câncer mais comum entre homens no Brasil (INCA, 2020), representando cerca de 5,2% de todos os casos de câncer entre homens no país segundo o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA), e de de todas as displasias malignas, a oral representa cerca de 30 a 40% dos casos no subcontinente indiano (NAG; DAS, 2018). Segundo a Oral Health Foundation, 17% dos cânceres de cavidade oral são causados diretamente pelo cigarro, e as chances de um fumante ser diagnosticado com câncer de boca são 91% maiores que as de um não-fumante (FOUNDATION, 2018). Outros fatores também apontados como causadores são: dieta pobre em frutas e vegetais, consumo de erva mate (*Ilex paraguariensis*) e as infecções pelo HPV (BRASIL; CÂNCER, 2011; WARNAKULASURIYA, 2009).

Quando a doença é detectada em estágios iniciais, a sobrevivência é significante-mente maior quando comparada aos casos de detecção tardia, em que a doença já está em um estado avançado. Infelizmente, dois a cada três pacientes (cerca de 62%) são diagnosticados em estágios avançados da doença, e isso acontece principalmente pelo fato da doença ser assintomática nos seus estágios iniciais (FOUNDATION, 2018). Sendo assim, o diagnóstico precoce por meio da análise de lesões orais pré-cancerosas como as displasias epiteliais orais se torna um assunto de muita importância, que deve ser estudado.

Desde 2017, a organização mundial de saúde (OMS) utiliza um sistema de pontuação baseado em alterações arquitetônicas e citológicas presentes nos tecidos para classificar as lesões displásicas, apresentadas na Tabela 1. Dividindo-se o tecido em três partes, na displasia leve, as células da parte mais inferior (camada basal) apresentam alterações, que são mínimas nas camadas mais superiores, enquanto que nos casos de displasia moderada e severa, tais alterações se apresentam também nas camadas mais superiores (RANGA-NATHAN; KAVITHA, 2019). A Figura 1 mostra um comparativo de quatro tecidos que apresentam níveis de displasia diferentes.

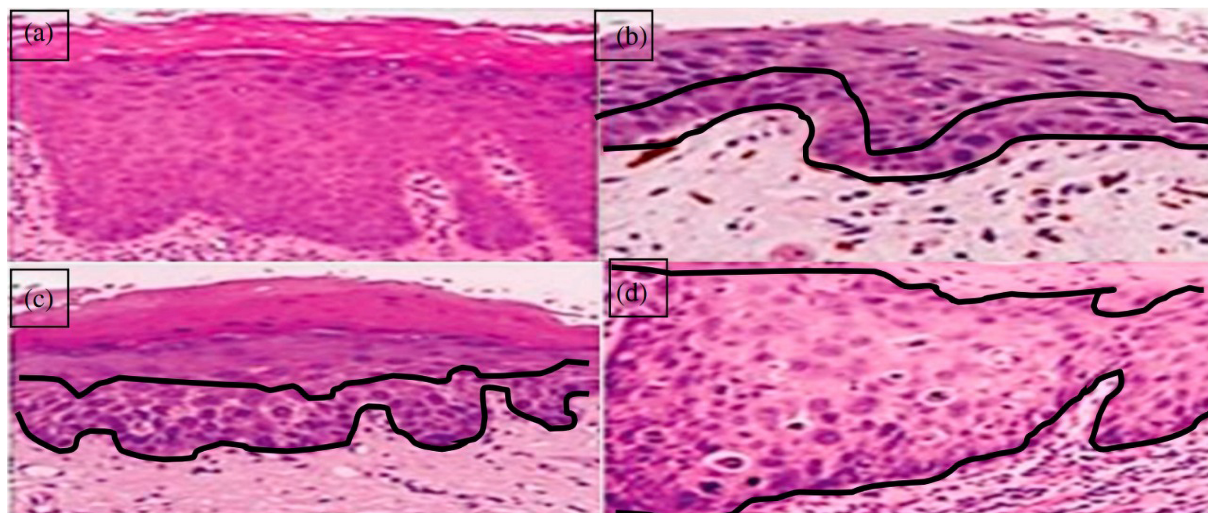


Figura 1 – Alterações ocorridas nos tecidos em diferentes graus de displasia. (a) Tecido saudável. (b) Tecido com displasia leve. (c) Tecido com displasia moderada. (d) Tecido com displasia severa. Fonte: (NAG; DAS, 2018)

Alterações Arquiteturais	Alterações Celulares
Estratificação epitelial irregular	Varição anormal no tamanho nuclear (aniso-nucleose)
Perda de polaridade das células da camada basal	Varição anormal das formas nucleares (pleomorfismo nuclear)
Cristas epiteliais em formato de gota	Varição anormal no tamanho celular
Aumento do número de figuras mitóticas	Varição anormal na forma celular
Superfície mitótica anormal	Aumento da proporção núcleo-citoplasmática
Queratinização prematura singular das células	Mitose atípica
Pérolas de queratina nas cristas epiteliais	Aumento do número e tamanho dos nucléolos
Perda de coesão das células epiteliais	Hipercromasia
Queratinização individual	

Tabela 1 – Alterações histológicas que caracterizam displasias epiteliais (Traduzido de (RANGANATHAN; KAVITHA, 2019)).

## 2.2 Histologia

Histologia pode ser definida como o estudo das células, dos tecidos e da forma como elas se organizam em estruturas para constituir os órgãos. O termo surgiu em 1819, em um livro do autor alemão Karl Meyer, porém o estudo de tecidos começou bem antes. Em 1798, o cirurgião e anatomista francês Marie François Xavier Bichat, considerado pai da histologia, passou a publicar artigos descrevendo o corpo humano como sendo formado por diferentes texturas e tecidos. Bichat considerou o “tecido” como a unidade morfológica

e fisiológica dos seres vivos, e conseguiu identificar vinte e um tecidos, sendo 7 gerais e 14 localizados (CALADO, 2019).

Durante o século XIX, vários avanços técnicos ocorreram no desenvolvimento de metodologias e reagentes que auxiliaram na observação, processamento e preservação de amostras biológicas. Somado aos progressos que ocorriam na área da microscopia, o período viu extraordinários avanços no desenvolvimento de reagentes e metodologias para análise, preservação e processamento das amostras biológicas. No final do século XIX e no início do século XX, melhorias em outros instrumentos como os micrótomos e adição de novas ferramentas e de técnicas de coloração, além da criação do microscópio eletrônico, melhoraram a observação e análise das estruturas biológicas.

Com os múltiplos avanços ocorridos nas últimas décadas, a histologia passou a ser cada vez mais adotada na medicina e em outras ciências biológicas, e a digitalização do material - usado na microscopia virtual - passou a ser amplamente aceita até mesmo no meio acadêmico (CALADO, 2019). No processo da geração das imagens, diversas etapas são realizadas para preparação da lâmina com o tecido: fixação, desidratação, inclusão, corte e coloração. Nas próximas seções são descritos os detalhes dessas etapas.

### 2.2.1 Preparação de Tecidos

No microscópio de luz (também chamado de microscópio óptico ou fotônico), a imagem é formada a partir de raios luminosos que atravessam as estruturas do corte histológico. Por essa razão, antes que os tecidos possam ser examinados no microscópio, eles devem ser fatiados em seções delgadas e colocados sobre lâminas de vidro (JUNQUEIRA; CARNEIRO, 2018). O instrumento usado no corte é chamado micrótomo, e permite cortes com grande precisão. Antes do corte, os tecidos e órgãos devem passar por uma série de tratamentos que garantem sua durabilidade e facilitam a visualização de suas estruturas. Os procedimentos mais comuns na preparação estão detalhados a seguir, como explicitados em (JUNQUEIRA; CARNEIRO, 2018):

- **Fixação:** após sua remoção do corpo, células ou fragmentos de tecidos e órgãos devem ser submetidos a um processo chamado fixação, para evitar a digestão dos tecidos por enzimas existentes nas próprias células (autólise) e preservar o máximo possível de sua estrutura e composição molecular. A fixação pode ser feita usando métodos químicos ou físicos, sendo os fixadores mais usados para microscopia de luz a solução de formaldeído a 4% e o glutaraldeído;
- **Desidratação e clareamento:** para o processo de desidratação, a água contida nos tecidos é extraída ao passar os fragmentos por diversos banhos em soluções crescentes de etanol. Após a desidratação, o etanol é substituído por um solvente orgânico miscível tanto no etanol quanto no meio escolhido para inclusão (parafina ou resina),



deixando o tecido transparente ou translúcido. A parafina é habitualmente utilizada para microscopia de luz, e as resinas, para microscopia de luz e eletrônica;

- **Inclusão:** nesse estágio, o tecido é colocado em uma substância derretida (parafina ou resina), e o calor da substância causa a evaporação do solvente orgânico. Com a evaporação, os espaços existentes dentro do tecido são preenchidos com a substância, que ao se solidificar, mantém o tecido em uma consistência rígida, formando um bloco a ser fatiado;
- **Microtomia ou corte:** como na maioria dos casos os tecidos e órgãos são espessos e não possibilitam a passagem adequada da luz para a formação de uma imagem, os tecidos devem ser fatiados usando-se um micrótomo (ver Figura 2). O bloco de parafina formado na inclusão é então fatiado, gerando cortes finos de 1 a 10 micrômetros de espessura;
- **Reidratação:** como a coloração é realizada sobre soluções à base de água, é necessária a substituição da parafina por água. Assim, além de serem colocados para flutuar sobre uma superfície de água aquecida, os tecidos são dispostos sobre lâminas de vidro, onde são aderidos e corados em seguida;
- **Coloração:** para que possam ser analisados no microscópio, a maioria dos cortes histológicos deve ser corada, pois nesse ponto, salvo raras exceções, os tecidos são incolores. A combinação de Hematoxilina e Eosina (H&E) é a mais usada no preparo de lâminas histológicas. A eosina é responsável por corar os componentes básicos das células e tecidos de rosa, enquanto a hematoxilina cora os componentes ácidos de azul ou violeta.

## 2.3 Microscopia

No final do século XVI, um negociante de óculos holandês chamado Zaccharias Janssen (1587-1638) inseriu duas lentes, uma convexa e uma côncava nas extremidades de um tubo e descobriu assim que objetos apareciam ampliados. No século XVII, o curador Robert Hooke (1635-1703) da London Royal Society desenvolveu o seu próprio microscópio composto, levando-o a observações microscópicas e à primeira utilização do termo “célula” para indicar as unidades microscópicas que observou na cortiça ([CALADO, 2019](#)).

O holandês Antoni van Leeuwenhoek (1632-1723) fez melhorias no microscópio de Hooke, proporcionando uma imagem mais ampliada e de melhor qualidade. Tais melhorias permitiram que Leeuwenhoek observasse espermatozoides de várias espécies, células sanguíneas, o aspecto estriado dos músculos, penas, escamas, cabelos, insetos, a estrutura



Figura 2 – Micrótomo para cortar tecidos incluídos em parafina. Acionando-se a manivela (à direita da figura), o bloco contendo o fragmento de tecido sobe e desce. Após cada volta da manivela, o bloco avança uma distância definida (geralmente de um a dez  $\mu\text{m}$ ) e, ao passar pela navalha, deixa uma fatia do tecido (JUNQUEIRA; CARNEIRO, 2018).

de folhas, leveduras, e as bactérias de seus próprios dentes, designando-os de “animalcules” (CALADO, 2019). Porém, foi com o trabalho de Marcello Malpighi (1628-1694) que, com o uso do microscópio, observou e descreveu órgãos e estruturas anatômicas humanas, como o cérebro, fígado, língua, retina, pulmões, gânglios linfáticos, pele e glândulas, que o microscópio teve o seu primeiro uso verdadeiramente científico (CALADO, 2019).

Durante o século XIX, Ernst Abbe (1840-1906), Carl Zeiss (1816-1888) e Otto Schott (1851-1935) trabalharam na criação de microscópios de alta qualidade, com lentes mais homogêneas e corpos mais estáveis (ERNST, 1900). O primeiro protótipo de microscópio a usar um feixe de elétrons foi construído pelos alemães Max Knoll (1897-1969) e Ernst Ruska (1906-1988) em 1931. Embora Ruska e Knoll tenham os créditos da invenção do microscópio eletrônico, a patente foi obtida por Reinhold Rudenberg (1883-1961), diretor científico da Siemens em 1931 (CALADO, 2019).

Com o avanço de novas tecnologias, o microscópio de luz (ou microscópio ótico), apresentado na figura 6b, se tornou cada vez mais popular. Composto de partes mecânicas e óticas, ele permite a análise das imagens coradas por meio de uma iluminação que atravessa o tecido (transiluminação). O componente ótico do microscópio possui três sistemas de lentes: condensador, objetivas e oculares. O condensador projeta um feixe de luz sobre o tecido, em direção à objetiva. A objetiva recebe a luz que chega ao atravessar o tecido e repassa uma versão ampliada da imagem em direção à ocular, que aumenta a imagem mais uma vez e a projeta na retina, que pode ser uma câmera, um detector eletrônico ou uma tela. O aumento total da imagem é dado pela multiplicação entre o aumento da ocular e o da objetiva.

O uso de câmeras de alta sensibilidade e resolução torna possível a ampliação da



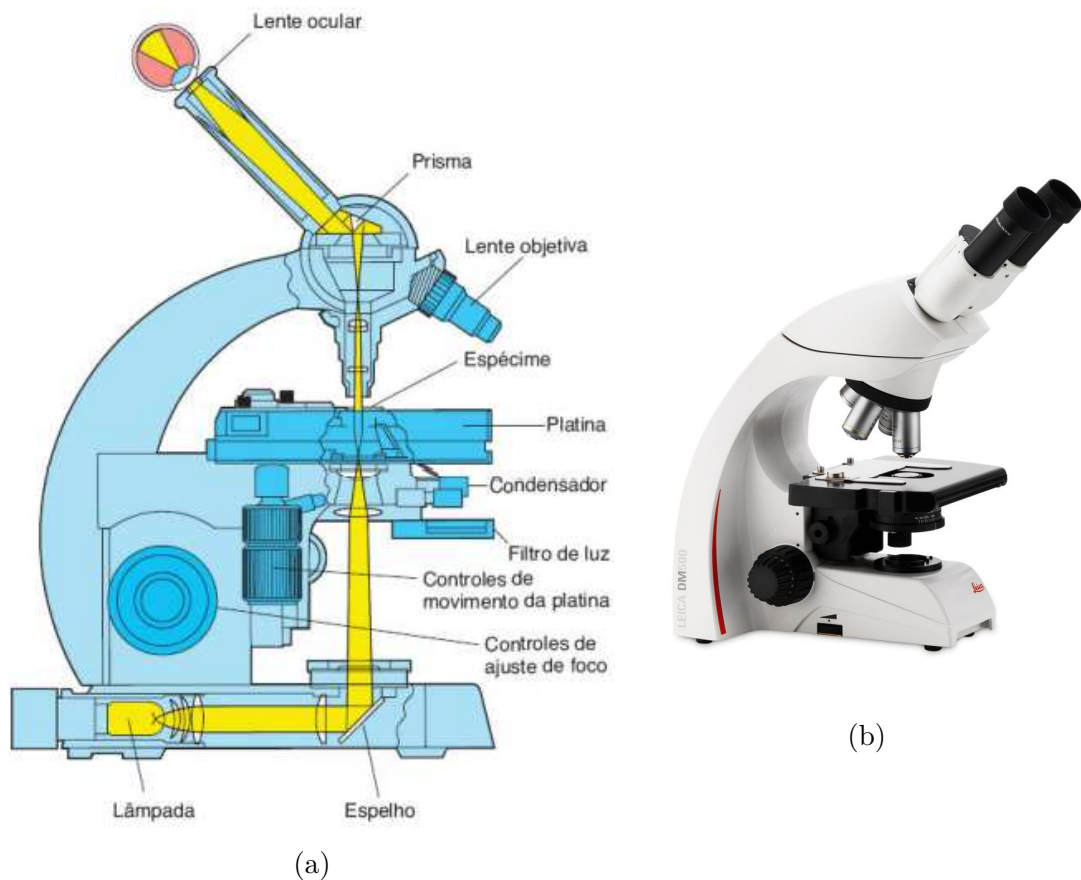


Figura 3 – Microscópios de luz: (a) desenho esquemático de um microscópio de luz. (JUNQUEIRA; CARNEIRO, 2018), (b) microscópio Leica DM500, usado na obtenção das imagens do banco usado neste trabalho (MICROSYSTEMS, ).

área de atuação da microscopia através da digitalização de imagens usadas em computadores para análise quantitativa por meio de aplicativos. Esses sistemas são úteis para estudar células vivas por períodos longos, pois usam luz de baixa intensidade e evitam o dano celular que pode resultar de uma iluminação intensa (JUNQUEIRA; CARNEIRO, 2018). Além disso, nossa visão limitada pode ser falha em algumas situações, e o uso de sistemas computacionais diminui a influência de fenômenos ópticos na classificação final, além de permitir a visualização de objetos que não são visíveis diretamente pelo sistema ocular.

## 2.4 Processamento Digital de Imagens

O processamento digital de imagens é todo tipo de alteração e análise feita na imagem através de um computador. Diferente dos seres humanos, que são limitados à banda visual do Espectro Eletromagnético (EM), os aparelhos de processamento de imagens cobrem quase todo o EM, variando de ondas gama a ondas de rádio, podendo trabalhar com imagens geradas por fontes que os humanos não estão acostumados a associar

com imagens, como ultrassom, microscopia eletrônica e imagens geradas por computador. Dessa forma, o processamento digital de imagens inclui um amplo e variado campo de aplicações (GONZALEZ; WOODS, 2009).

Tradicionalmente, as etapas empregadas no processamento digital de imagens são: pré-processamento, segmentação, extração de características e classificação. O pré-processamento é utilizado para eliminar ruídos e variações de pigmentação, realçar objetos e contraste, entre outras operações (GONZALEZ; WOODS, 2009). A etapa de segmentação é usada para identificar e rotular os diferentes objetos presentes na imagem, separando-a em regiões. Após as etapas anteriores, a extração de características é feita para identificar os atributos referentes aos núcleos segmentados e à imagem como um todo (SILVA et al., 2019). A extração de características permite que atributos relativos aos núcleos celulares sejam utilizados para identificar o surgimento de lesões e progressão de câncer, tornando-a um importante aliado na tarefa de comparação quantitativa das estruturas celulares (SILVA et al., 2019).

A última etapa, a classificação, tem o objetivo de identificar os padrões nas características extraídas na etapa anterior, identificando estruturas, padrões ou propriedades de uma imagem para associá-la a uma classe predefinida (SILVA et al., 2019). Para isso, é necessário que um especialista execute a classificação manual de algumas imagens, que serão usadas de referência na aprendizagem do sistema classificador, que identificará o restante dos objetos de acordo com a proximidade entre suas características e as das classes já definidas.

## 2.5 Sistemas de Diagnóstico Auxiliado por Computador

Nos primeiros estudos sobre análise quantitativa de imagens médicas por computador, realizados na década de 60, era assumido que os radiologistas poderiam ser substituídos pelos computadores na tarefa de detecção de anomalias (DOI, 2007). Dado o baixo poder de processamento das máquinas da época, os resultados foram desanimadores e não corresponderam à alta expectativa que se tinha dos computadores. Já na década de 80, uma nova abordagem passou a ser estudada. Dessa vez, foi sugerido que os computadores passassem a ser usados para auxiliar os especialistas, ao invés de substituí-los.

O conceito ficou conhecido como *Computer Aided Detection and Diagnosis* (CAD), e se espalhou rapidamente. Os sistemas CAD, cujo diagnóstico leva em consideração os papéis do computador e do especialista, são diferentes dos sistemas de *Automated Computer Diagnosis*, em que o diagnóstico independe da opinião de um especialista, ficando todo a encargo do sistema. Na Figura 4, são apresentadas as principais etapas envolvidas em um sistema CAD. Essas etapas vão do melhoramento de imagens (fase do pré-processamento) até a fase de classificação de imagens.

Em sistemas CAD, não é necessário que seu desempenho seja melhor ou comparável ao dos especialistas, pois o sistema opera de forma complementar, auxiliando na tomada de decisão. Assim, os especialistas podem concordar ou não com o resultado apresentado pelo computador, escolhendo descartar resultados incoerentes quando estão confiantes com seus diagnósticos ou usar os resultados apresentados para reforçar suas decisões. Essa sinergia entre a competência do profissional e o poder de processamento da máquina é valiosa para diagnósticos ainda mais precisos. Enquanto um sistema de *Automated Computer Diagnosis* tem seu desempenho analisado com base na performance do sistema na realização dos diagnósticos, o sistema CAD tem seu nível de eficácia igual ao do especialista que utiliza essa abordagem como uma segunda opinião (DOI, 2007).

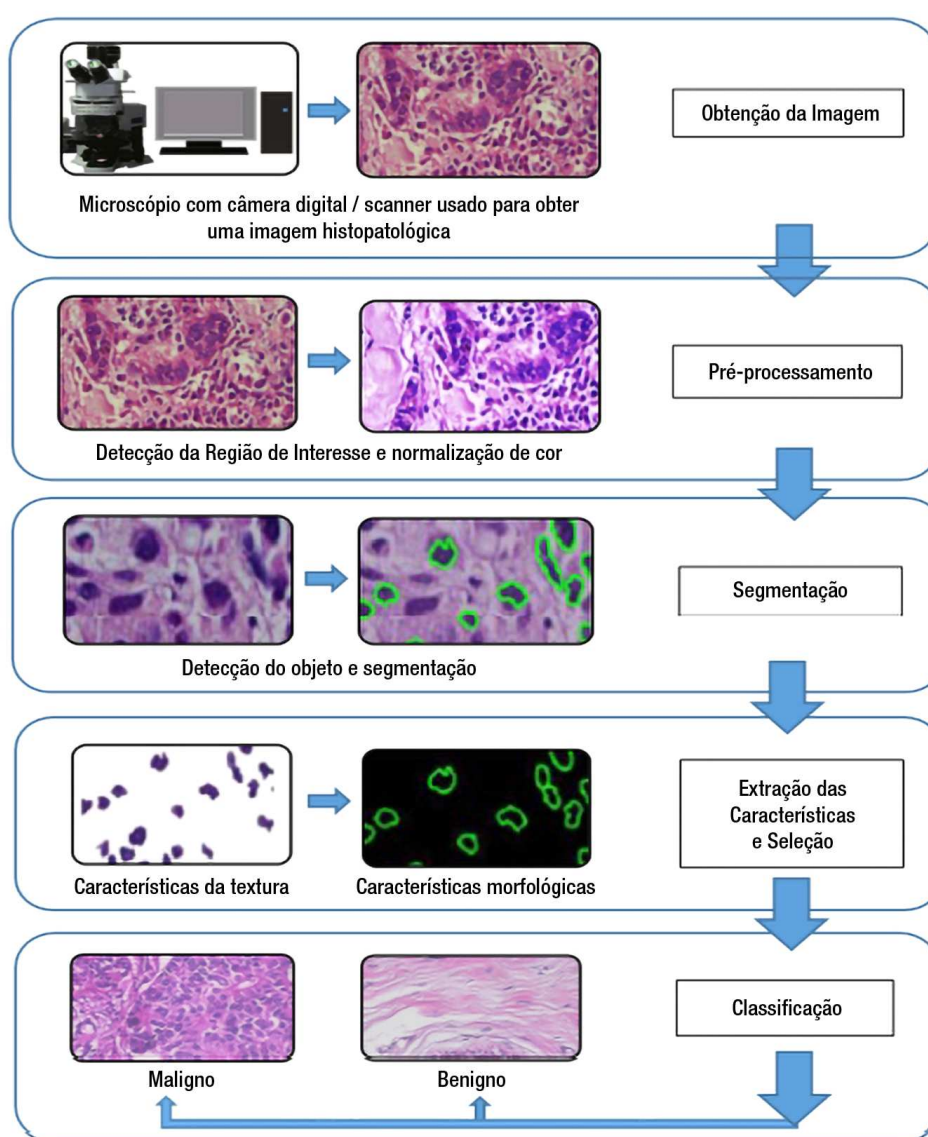


Figura 4 – Diagrama esquemático do fluxo de funcionamento de um sistema CAD (KAUSHAL et al., 2019)

Um sistema CAD pode ser enquadrado em dois tipos de aplicações: sistemas

*Computer-Aided Detection* (CADE) e sistemas *Computer-Aided Diagnosis* (CADx). Enquanto os sistemas CADE são desenvolvidos para auxiliar o especialista a detectar lesões em imagens médicas, os sistemas CADx são utilizados para auxiliar no diagnóstico de tais lesões, quantificando características da imagem e analisando padrões para chegar a uma classificação, por exemplo, se um tumor é maligno ou benigno (FIRMINO et al., 2016).

Um dos objetivos dos sistemas CAD é a redução de casos benignos enviados para biópsia. Em diagnósticos com o uso de sistema CAD, já foram relatados aumentos de sensibilidade, de 73% para 87%, e de especificidade, de 32% para 40% conforme (COOK; FOX, 1989). Tais resultados indicam o potencial de uso do computador no auxílio de radiologistas, reduzindo também o número de biópsias feitas de forma desnecessária (AZEVEDO-MARQUES, 2001).

Dentre as diversas etapas do sistema CAD, a extração de características é uma etapa muito importante, pois afeta diretamente desempenho do sistema. Uma variedade de métodos de processamento de imagem são utilizados na obtenção de informações texturais para representação de informações não-morfológicas. Também existem métodos que exploram informações morfológicas em imagens como área, perímetro, etc. A investigação de novas abordagens de descritores robustos continua sendo um desafio no desenvolvimento de sistemas computacionais para investigação de lesões em imagens histológicas.

## 2.6 Extração de Características

Existem muitas formas de representar as imagens no computador, com formatos de arquivos e cores diferentes, porém mesmo que a representação dos dados seja diferente dependendo do formato, na maior parte dos casos, a imagem é exibida da mesma forma, em um conjunto de pixels. Graças às diferentes formas de representação, é possível fazer várias análises que não são possíveis a olho nu, pois nem mesmo a menor diferença de cor entre dois pixels passa despercebida pelo computador, mesmo que seja invisível a nossos olhos. Tais informações permitem que objetos e classes sejam distinguidos a partir de imagens, por meio de sistemas reconhedores de padrões capazes de reduzir complexas e subjetivas análises a meros cálculos aritméticos e matriciais. Para isso, é necessário que seja feito o processo de descrição da imagem.

Na descrição da imagem, os descritores (chamados frequentemente pelo termo *características* na literatura de reconhecimento de padrões) são adicionados ao que é chamado de *vetor de características*, para que o sistema reconhedor de padrões possa analisar as propriedades comuns à classe de padrões e fazer a atribuição de elementos às suas respectivas classes de forma automática e com a menor intervenção humana possível (GONZALEZ; WOODS, 2009). Existem inúmeras características que podem ser analisadas, porém a maior parte delas se encaixa em dois grupos: Descritores Morfológicos,

também chamados de descritores de fronteira, e Descritores Não-Morfológicos, chamados também de descritores regionais.

### 2.6.1 Descritores Morfológicos

O processamento de informações morfológicas de imagens lida com ferramentas para a extração de componentes da imagem úteis na representação e descrição da forma (GONZALEZ; WOODS, 2009). O estudo morfológico da imagem busca maneiras de agrupar e identificar grupos e objetos nelas, e seus descritores representam as informações referentes à forma dos objetos. Os descritores mais comuns são os descritores de fronteira, que representam informações sobre o formato dos objetos da imagem, os descritores regionais, que descrevem as regiões da imagem - como área e circularidade - e os descritores topológicos, que representam propriedades que não são afetadas por deformações (por exemplo, uma rotação ou um alongamento da figura não altera o número de buracos na superfície de um objeto) (GONZALEZ; WOODS, 2009). Exemplos de descritores morfológicos são apresentados aqui:

#### 2.6.1.1 Área

Refere-se à quantidade de pixels dentro de uma fronteira.

#### 2.6.1.2 Excentricidade

Relaciona a proximidade da forma do objeto a uma circunferência, se mantendo invariante mesmo com operações de escalonamento, rotação, e translação. Sendo  $\mathbf{a}$  uma linha que conecta os dois pontos mais extremos que compõem o diâmetro (ou seja, o eixo maior da fronteira), e  $\mathbf{b}$ , uma linha perpendicular ao eixo maior (que será chamada então de eixo menor), a excentricidade será definida pela razão entre os eixos menor e maior:

$$E = \sqrt{1 - \frac{b^2}{a^2}} \quad (2.1)$$

Quanto mais próxima de 0 for a excentricidade, mais parecida com um círculo é a figura, ao passo que uma figura cuja excentricidade se aproxima de 1 é mais semelhante a uma linha.

#### 2.6.1.3 Orientação

A orientação ( $O$ ) é o ângulo entre o eixo x e o maior eixo da elipse que contém a região, variando, em graus, de  $-90$  a  $+90$ .

#### 2.6.1.4 Perímetro

O perímetro ( $P$ ) de uma região é o tamanho (comprimento) da sua fronteira em pixels.

#### 2.6.1.5 Solidez

É a proporção entre a área do objeto e a área do fecho convexo do menor polígono convexo que o contém. Um objeto com muitos furos terá uma solidez inferior à de um objeto maciço. Sua fórmula é:

$$S = \frac{\text{área do objeto}}{\text{área do fecho convexo}} \quad (2.2)$$

### 2.6.2 Descritores Não-Morfológicos

Assim como as características morfológicas da imagem representam as formas dos objetos presentes, existem características que são importantes para a análise da textura da imagem, ou seja, para verificar o comportamento dos pixels presentes. Exemplos de abordagens utilizadas no processamento de imagens para descrever as informações não-morfológicas de uma região são as abordagens estatísticas, a estruturais e a espectrais (GONZALEZ; WOODS, 2009). Para imagens médicas, a extração de características texturais é uma das mais importantes técnicas em diversas etapas de um sistema CAD, pois visa buscar a identificação de atributos ou características que diferenciam o comportamento da superfície representada na imagem (DUARTE; NASCIMENTO; OLIVEIRAS, 2013).

#### 2.6.2.1 Índice de Moran

O Índice de Moran mede a autocorrelação espacial a partir do desvio da intensidade de um pixel em comparação com a média da região. Ou seja, mede a nitidez da imagem ou a complexidade de sua textura.

#### 2.6.2.2 Entropia

A medida de Entropia, introduzida por Shannon (SHANNON, 1948), possibilita a análise da quantidade de informações gerada por uma fonte ou transferida por meio de um canal. Quanto maior for o valor da entropia, mais informação é associada ao canal (PEDRINI; SCHWARTZ, 2008). Nas imagens digitais, o cálculo da entropia se refere à energia, ou desordem de uma imagem, e é considerada uma característica importante na descrição de imagens histológicas. As entropias mais populares são a entropia de Shannon (Equação 2.3) (SHANNON, 1948), Renyi (Equação 2.4) (RÉNYI, 1961), Tsallis (Equação

2.5) (TSALLIS, 1988) e Kapur (Equação 2.6) (CHATTORAJ; VISHWAKARMA, 2018), permitindo trazer a informação da entropia sobre diferentes aproximações. As equações para o cálculo das métricas são apresentadas abaixo:

- Entropia de Shannon ( $E_S$ ):

$$E_S = \sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i) \quad (2.3)$$

Em que  $p(z_i)$  é a probabilidade do evento e  $L$  é o número de eventos. Em uma imagem com  $L$  pixels, é a probabilidade de um pixel  $z$  possuir um valor  $i$ .

- Entropia de Renyi ( $E_R$ ):

$$E_R = \frac{1}{1 - \alpha} \log_2 \sum_{i=1}^n p(z_i)^\alpha \quad (2.4)$$

- Entropia de Tsallis ( $E_T$ ):

$$E_T = \frac{1}{1 - \alpha} (1 - \log_2 \sum_{i=1}^n p(z_i)^\alpha) \quad (2.5)$$

Em que  $n$  é o número de eventos,  $p(z_i)$  é a probabilidade desse evento e o índice  $\alpha$  representa a ordem da entropia, assumindo que  $\alpha \neq 1$  e  $\alpha > 0$ .

- Entropia de Kapur ( $E_K$ ):

$$E_K = \frac{1}{\beta - \alpha} \log_2 \frac{\sum_{i=0}^{n-1} p(z_i)^\alpha}{\sum_{i=0}^{n-1} p(z_i)^\beta} \quad (2.6)$$

Em que  $\beta \neq \alpha$ ,  $\alpha > 0$  e  $\beta > 0$ .

### 2.6.2.3 Matriz de Co-ocorrência de Níveis de Cinza

Uma matriz de co-ocorrência representa a frequência de combinações de diferentes valores de intensidade dos pixels. Logo, ela possui o número de linhas e o número de colunas iguais à quantidade de níveis de cinza na imagem, sendo portanto sempre uma matriz quadrada. A matriz considera a relação entre pixels vizinhos, comparando-os dois a dois e segue uma direção  $\theta$  que pode ser leste ou oeste ( $0^\circ$ ), norte ou sul ( $90^\circ$ ), nordeste ou sudoeste ( $45^\circ$ ) e sudeste ou noroeste ( $135^\circ$ ), e uma distância  $d$  maior ou igual a 1. Assim, considerando um pixel de referência  $i$  e um pixel vizinho  $j$ , uma matriz de co-ocorrência pode ser definida como  $P(i, j, d, \theta)$ .

Em (HARALICK; SHANMUGAM; DINSTEN, 1973), foi analisada a capacidade da matriz de co-ocorrência de representar a textura de uma imagem corretamente. O estudo usou uma matriz de co-ocorrência de níveis de cinza (*Gray Level Co-occurrence*



*Matrix*, ou GLCM) e encontrou resultados promissores e de grande aplicabilidade, apresentando um conjunto de 14 descritores que ficaram conhecidos como descritores de Haralick. O conceito da GLCM vem sendo explorado de diversas maneiras, sendo amplamente usado para caracterizar imagens de diferentes aplicações de reconhecimento de padrões. Os descritores de Haralick serão explicitados abaixo:

- Energia:

$$\sum_i \sum_j p_{(i,j)}^2 \quad (2.7)$$

- Contraste:

$$\sum_i \sum_j |i - j|^k p_{(i,j)}^n \quad (2.8)$$

- Correlação:

$$\frac{\sum_i \sum_j (ij) p_{(i,j)} - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (2.9)$$

Em que  $\mu$  = média,  $\sigma$  = desvio padrão

- Variância:

$$\sum_i \sum_j (i - j)^2 p_{(i,j)} \quad (2.10)$$

- Homogeneidade:

$$\sum_i \sum_j \frac{p_{(i,j)}}{1 + |i - j|} \quad (2.11)$$

- Média da soma:

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

Em que  $N_g$  = níveis de cinza,  $i=(x+y)$

- Variância da soma:

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

Em que  $N_g$  = níveis de cinza,  $i=(x+y)$

- Entropia da soma:

$$-\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} \quad (2.14)$$

Em que  $N_g$  = níveis de cinza,  $i=(x+y)$



- Entropia:

$$- \sum_i \sum_j p_{(i,j)} \log\{p_{(i,j)}\} \quad (2.15)$$

- Variância da diferença:

$$\sum_i \sum_j (i - j)^2 p_{(i,j)} \quad (2.16)$$

Em que  $i=(x+y)$

- Entropia da diferença:

$$- \sum_{i=0}^{N_g-1} (p_{x-y}(i) \log\{p_{x-y}(i)\}) \quad (2.17)$$

Em que  $N_g =$  níveis de cinza,  $i=(x-y)$

- Medidas de correlação da informação I e II:

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

$$(1 - \exp[-2.0(HXY2 - HXY)])^{\frac{1}{2}} \quad (2.19)$$

Sendo  $HXY =$  Entropia (Equação 2.15),

$HX$  e  $HY =$  Entropias de  $p_x$  e  $p_y$ ,

$HXY1 = -\sum_i \sum_j p_{(i,j)} \log(p_x(i)p_y(j))$ ,

$HXY2 = -\sum_i \sum_j p_x(i)p_y(j) \log(p_x(i)p_y(j))$ .

- Coeficiente de correlação maximal:

$$(\text{Segundo maior autovalor (eigenvalue) de } Q)^{\frac{1}{2}} \quad (2.20)$$

$$\text{Onde } Q_{(i,j)} = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$$

#### 2.6.2.4 Matriz de Co-ocorrência Estrutural

Outro método baseado no conceito da GLCM é o da Matriz de Co-ocorrência Estrutural (*Structural Co-occurrence Matrix*, ou SCM), apresentado em (RAMALHO et al., ), que guarda informações de dissimilaridades das estruturas de dois sinais de entrada. Na variação apresentada para análise de imagens com a SCM, a imagem é tratada como um sinal e passa por um filtro passivo passa-baixa que suaviza as saliências da imagem ou um filtro passa-altas que realça tais saliências. O sinal degradado e o original passam por uma função de discretização que reduz a quantidade de níveis de cinza na imagem de 256 para 8. Assim, a matriz SCM pode ser representada por uma matriz  $\mathbf{M} = \{m_{ij}\}$ , em que  $m_{ij}$  representa a quantidade de pixels com valor  $i$  na imagem original possuem um vizinho com valor  $j$  na imagem degradada, a um certo deslocamento  $p$ . Ou seja,

$m_{ij} = \#\{(i,j) \mid P(i,j), i = Q(f_p), j = Q(k(g)_{p+d})\}$ ,  $f$  é o sinal da imagem,  $k$  é o filtro aplicado,  $Q$  uma função de partição que reduz a quantidade de níveis de cinza e  $d$  é um vetor que indica a distância do pixel vizinho. Os descritores mensuráveis a partir da SCM são:

- Correlação:

$$COR = \sum_{i,j=0}^{N-1} m_{ij} \frac{(i - \mu_i)(j - \mu_j)}{\sqrt{\sigma_i^2 \sigma_j^2}} \in [-1,1] \quad (2.21)$$

- Momento da Diferença Inversa:

$$IDM = \sum_{i,j=0}^{N-1} \frac{m_{ij}}{1 + (|i - j|)} \in [-1,1] \quad (2.22)$$

- Entropia:

$$ENT = - \sum_{i,j=0}^{N-1} m_{ij} \log(m_{ij}) \quad (2.23)$$

- Distância Qui-quadrado:

$$CSD = - \sum_{i=0}^{N-1} \frac{(P_i^O - P_i^e)^2}{P_i^e} \quad (2.24)$$

- Razão da distância Qui-quadrado:

$$CSR = \sum_{i=0}^{N-1} \frac{(P_i^I - P_i^m)^2}{P_i^m} \quad (2.25)$$

Em que  $P^m = (P^I + P^{II})/2$

- Razão da diferença média absoluta:

$$MDR = \frac{MD_1}{MD_2} \quad (2.26)$$

Em que  $P^m = (P^I + P^{II})/2$

- Divergência de Kullback & Leibler:

$$D_{KL} = \sum_{i=0}^{N-1} \log\left(\frac{P_i^c}{P_i^r}\right) P_i^c \quad (2.27)$$

Em que  $P_i^c/P_i^r = 0$  e  $P_i^r = 0$

- Diferença absoluta complementar:

$$CAD = 1 - \sum_{i=0}^{N-1} |P_i^c - P_i^r| \quad (2.28)$$

## 2.7 Trabalhos Correlatos

Dada a importância do diagnóstico precoce em lesões cancerosas, o estudo apresentado por (NAG; DAS, 2018) analisa mudanças ocorridas nos tecidos de imagens epiteliais displásicas orais em diferentes estágios da lesão. Algumas das várias alterações citológicas e arquitetônicas encontradas são: variação anormal no tamanho e forma do núcleo celular, variação anormal no tamanho e forma da célula, relação núcleo-citoplasmática aumentada, perda de polaridade e hiperplasia das células da camada basal, estratificação epitelial irregular, número atípico e aumentado de figuras mitóticas, mitose anormalmente superficial, cristas interpapilares em forma de gota e com formação de pérolas de queratina e queratinização prematura em células individuais. O estudo também avaliou que há uma dificuldade em fazer a discriminação entre os diferentes graus limítrofes, uma vez que é muito dificultosa a diferenciação entre displasia epitelial oral e carcinoma in situ, sugerindo a avaliação de novos métodos de análise de imagem para detecção de displasia epitelial oral.

Em (KAUSHAL et al., 2019), um estudo foi feito com as melhores técnicas apresentadas na literatura para operação de sistema CAD, das técnicas de pré-processamento às técnicas de classificação. Ao escolherem imagens histopatológicas de câncer de mama do *dataset* BreakHis, composto de 9109 imagens microscópicas de câncer de mama com diferentes fatores de magnificação, os pesquisadores puderam comparar técnicas publicadas usando um *dataset* único, tornando a comparação entre as técnicas mais precisa. Os resultados concluíram que o classificador *Random Forest* foi capaz de discriminar os níveis baixo, médio e alto da escala de Bloom Richardson, usada para graduar a malignidade dos tumores mamários, e identificar as características da imagem corretamente, com uma acurácia de 93% em uma magnificação de 10x. O estudo também conclui que a análise de características texturais com o uso de técnicas como GLCM pode ajudar na classificação dos tecidos cancerosos, auxiliando os patologistas em suas performances, reduzindo suas cargas de trabalho, agilizando o processo de diagnóstico e aumentando as taxas de sobrevivência por meio de um diagnóstico precoce.

Ao buscar estabelecer uma abordagem de um sistema CAD para detecção e classificação da displasia epitelial oral, o estudo apresentado em (ADEL et al., 2018) desenvolveu um método com base em diferentes seleções e ampliações de imagens microscópicas coradas. A abordagem extrai um conjunto de dezesseis características usando o algoritmo *Oriented FAST and Rotated Brief* (ORB) e utiliza o algoritmo de classificação *Support Vector Machine* (SVM) para classificar as imagens em dois grupos de acordo com a severidade das características extraídas. Testes foram feitos com um total de 46 imagens coradas com H&E e com ampliação de 100x em uma proporção de 7:3 respectivamente, alcançando uma acurácia de 92,8% na classificação de lesões displásicas orais e apresentando, no final, um relatório com múltiplos cálculos efetuados na imagem, para auxiliar

a acurácia e justificação dos médicos. Como trabalho futuro, o estudo sugere o uso de novas imagens com diferentes níveis de magnificação e o estudo de métodos capazes de determinar o nível da displasia, entre leve, moderado ou severo (carcinoma in-situ).

## 2.8 Considerações Finais

Com os inúmeros avanços tecnológicos aplicados na área da saúde nos últimos anos, as imagens microscópicas podem ser digitalizadas facilmente e com alta resolução por microscópios digitais, comumente encontrados em diversos laboratórios. As novas inovações tecnológicas aumentaram também o poder de processamento de máquinas comuns do dia-a-dia. Tal aumento no poder de processamento permitiu avanços nas áreas de processamento de imagem e de desenvolvimento de algoritmos de classificação cada vez mais robustos. Com isso, o uso de sistemas CAD por especialistas se tornou algo comum e aceito, permitindo a criação de uma sinergia que tem a capacidade de melhorar e agilizar os resultados de diagnósticos. Como constatado, existem sistemas capazes de verificar lesões causadas por diversos tipos de câncer, sendo um deles o câncer de boca.

O câncer de boca é uma doença silenciosa que atinge milhares de pessoas todos os anos, mas um diagnóstico precoce favorece o aumento da taxa de sobrevivência de pacientes acometidos com essa doença. Com o processamento das imagens, é possível fazer a extração de características que podem ajudar um algoritmo classificador a identificar as classes às quais as imagens pertencem, e tal algoritmo pode então ser usado como um sistema CAD que auxilie nos diagnósticos de especialistas. Existem descritores morfológicos e não-morfológicos que são citados na literatura como capazes de exprimir as características das imagens corretamente, e eles serão analisados aqui, assim como o algoritmo de classificação *Random Forest* que de acordo com a literatura, produziu bons resultados na classificação de outros tipos de lesão.

## 3 Metodologia

Neste capítulo são apresentadas as informações sobre o banco de imagens, as técnicas desenvolvidas para associação das informações das características morfológicas e não-morfológicas, assim como os classificadores empregados para avaliação dessas características para discretizar as classes de lesões.

### 3.1 Visão Geral

As principais etapas desses métodos são ilustradas na Figura 5,

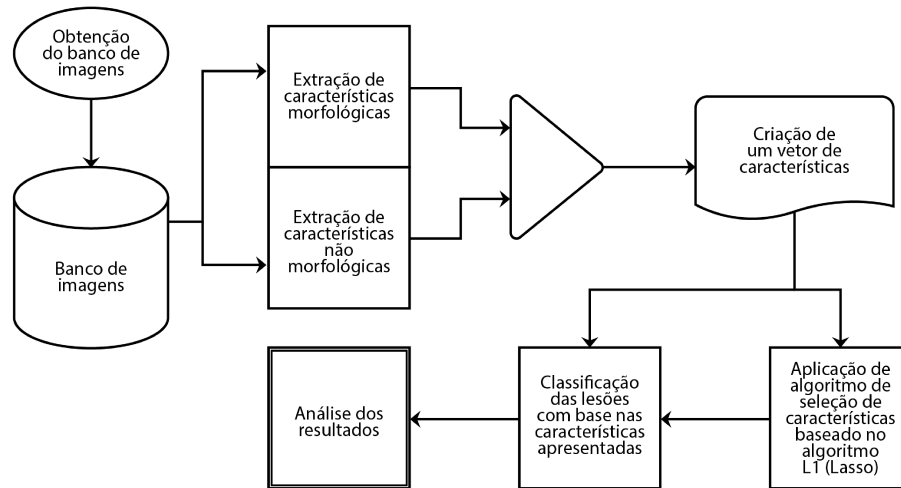


Figura 5 – Etapas do sistema de classificação de lesões da cavidade bucal.

### 3.2 Linguagens e Ferramentas

As imagens histológicas utilizadas no trabalho foram generosamente disponibilizadas pela Universidade Federal de Uberlândia, e as imagens segmentadas utilizadas, geradas no trabalho (SILVA et al., 2019) foram cedidas pelo professor orientador. Para a coleta dos descritores morfológicos e dos descritores da SCM, foi utilizada a ferramenta MATLAB®, na versão R2020a. A escolha da ferramenta foi motivada por seu amplo uso em trabalhos científicos e pela disponibilidade de funções já implementadas que agilizarão o desenvolvimento deste trabalho, como as funções dos descritores da SCM, que foram disponibilizados junto ao trabalho (RAMALHO et al., ). Para o desenvolvimento do restante do trabalho, as etapas de implementação dos descritores de Haralick, seleção de características e classificação foram implementadas com o uso do serviço em nuvem

Google Colaboratory (Colab) em conjunto com a linguagem Python, também pela ampla disponibilidade de bibliotecas disponíveis.

### 3.3 Banco de Imagens

O banco de imagens usado como amostra foi obtido a partir de lesões induzidas no experimento *in vivo* de número de protocolo 038/09 na Universidade Federal de Uberlândia. No experimento, realizado no período de 2009 - 2010, foram induzidas lesões na língua de 30 camundongos C57BL/6, de 6 a 8 semanas, entre 20 e 23g, com uso do carcinógeno 4-nitroquinolina 1-óxido (4NQO). Os animais foram separados em dois grupos de 15 indivíduos cada (denominados *semana 16* e *semana 32*), e foram mantidos no centro de Bioterismo a uma temperatura de 22°C, com um ciclo claro/escuro de 12h, em gaiolas de plástico contendo serragem. Foi feita a diluição do carcinógeno 4NQO em propilenoglicol na concentração de 5mg/ml e depois em água filtrada até atingir a concentração final de 100 ug/ml. Ambos os grupos tiveram acesso a essa solução, que foi preparada e trocada semanalmente nos frascos de beber dos camundongos por 16 semanas ininterruptas. No final desse período, os camundongos do grupo *semana 16* foram sacrificados imediatamente. Nas próximas 16 semanas, os 15 camundongos do grupo *semana 32* beberam só água filtrada, sendo sacrificados no final do período, com 32 semanas de experimento.

As línguas removidas foram colocadas por 24 horas em uma solução de formalina tamponada a 4%, e depois desse período, foram processadas e incluídas em parafina para a confecção dos blocos. Após tal processo, as línguas foram recortadas e coradas usando a técnica de coloração Hematoxilina-Eosina e foram classificadas conforme o modelo de Lumerman (LUMERMAN; FREEDMAN; KERPEL, 1995) em hiperplasias, displasias leve, moderada, grave, carcinoma *in situ* e carcinoma. Um total de 43 imagens histológicas foram obtidas com o uso do microscópio óptico LeicaDM500 com magnificação de 400x a salvas no formato TIFF utilizando o modelo de cores RGB e resolução de 2048x1536 pixels. Com o auxílio de um patologista, esses tecidos foram classificados em saudáveis, displasia leve, displasia moderada e displasia severa, e 74 regiões de interesse (*Region of Interest* - ROI) foram obtidas para cada classe, totalizando 296 ROI. Os núcleos de 40% das imagens foram marcados manualmente pelo patologista, gerando as imagens de referência conhecidas como padrão-ouro.

### 3.4 Extração de características

Os descritores morfológicos usados foram apresentados previamente em 2.6.1 e são: área, excentricidade, orientação, perímetro e solidez. Esses descritores foram obtidos sobre as imagens segmentadas obtidas com o método proposto por (SILVA et al., 2019), que fez a segmentação dos núcleos das regiões de fundo da imagem para uma análise das

informações internas. Foi usado um valor arbitrário de  $1 * 10^{-5}$  nas operações logarítmicas para evitar exceções ocorridas pelo cálculo do logaritmo natural de 0, mas por ser de tamanho pequeno, o valor não influenciou os resultados obtidos.

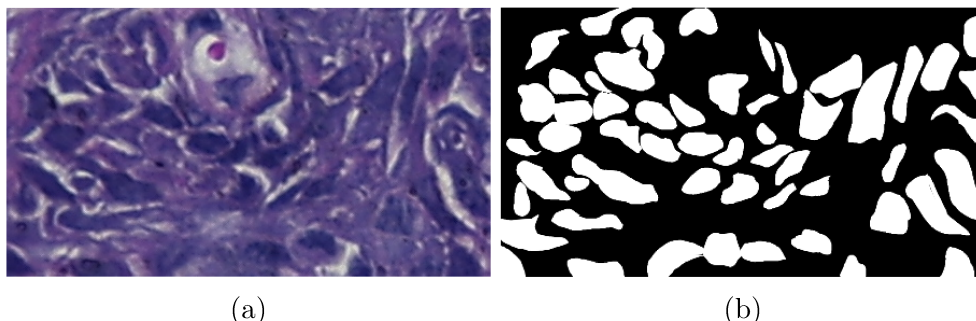


Figura 6 – Imagens de tecido com displasia moderada: (a) Imagem do tecido corado, pronta para a extração de características, (b) Imagem segmentada, pronta para a extração de características morfológicas.

Os descritores não-morfológicos investigados foram apresentados em 2.6.2 e são: índice de Moran, os descritores da SCM e os descritores de Haralick. Para o cálculo dos descritores de Haralick, a matriz de co-ocorrência GLCM foi criada a partir da junção normalizada de matrizes de ângulos  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  e  $135^\circ$  e de distâncias 1 a 5. Com a SCM, foi aplicado um ruído branco gaussiano de média e variância 0,01 nas imagens, como proposto no código fonte disponibilizado por (RAMALHO et al., ). O código foi empregado para a criação da matriz e extração das características. Os descritores permitiram investigar diversas associações entre os vetores de características. Para uma análise mais detalhada dos descritores, foram construídos os seguintes grupos (vetores de características): i) descritores morfológicos; ii) descritores não morfológicos compostos por Haralick e SCM; iii) morfológicos, Haralick e SCM, denominados por MHC; iv) MHC e entropia Shannon; v) MHC e entropia Renyi; vi) MHC e entropia Tsallis; vii) MHC e entropia Kapur; viii) MHC e índice de Moran.

### 3.5 Redução de dimensionalidade

Ao final do processo de extração de características, foi obtido um total de 27 descritores de cada imagem. Um passo importante para a criação de um classificador adequado é verificar quais descritores são úteis no processo de classificação. Um problema que pode ocorrer quando muitos descritores são usados é a presença de ruídos nos classificadores. Os atributos extraídos nem sempre contém informações que permitem a separação entre as classes, atrapalhando o classificador e comprometendo a acurácia do modelo, caso os descritores tenham pouca relevância.

Outro grande problema encontrado é o tempo de análise das características. Quanto maior o número de descritores, mais complexo pode ficar o sistema de classificação, e fre-

quentemente, mais custoso fica o processo como um todo, demandando mais tempo e mais recursos computacionais. Além dos problemas apontados, um grande número de características que se comportam de maneiras semelhantes entre si podem ser usadas para criar um sistema tendencioso e inconsistente, dando maior importância a certos aspectos da imagem.

Para evitar os obstáculos apresentados acima, existem vários algoritmos de redução da dimensionalidade dos dados, como seleção do melhor subconjunto, *stepwise*, *backward*, *forward* e o LASSO (*Least Absolute Shrinkage and Selection Operator*) (RODRIGUES, ), proposto por Tibshirani em 1996 (TIBSHIRANI, 1996). Enquanto alguns algoritmos preocupam-se em ajustar os modelos e selecionar o melhor com base em critérios próprios do regularizador, o LASSO L1 não foca em subconjuntos de descritores, analisando o conjunto de descritores como um todo e definindo uma operação de redução capaz de produzir coeficientes com valor 0.

O método apresentado pelo LASSO permite que sejam descartados atributos que tem alta correlação, de forma a reduzir o sobre-ajuste do modelo aos dados (*overfitting*), reduzir a complexidade do modelo e a dimensionalidade dos dados. O modelo é constituído por uma função de custo, representada pela soma residual dos quadrados (RSS) com um termo de regularização adicional, criando a função objetivo a ser minimizada:

$$\sum_{i=1}^M (y_i - \sum_{j=0}^p w_j * x_{ij})^2 + \alpha \sum_{j=0}^p |w_j| \quad (3.1)$$

O regularizador foi aplicado em subconjuntos de descritores, assim como no conjunto completo, para analisar a importância dos descritores quando pareados a grupos diferentes, e foi escolhido principalmente por sua capacidade não só de evitar o fenômeno de *overfitting*, mas também de remover características que podem ser negligenciadas.

## 3.6 Classificação

O algoritmo empregado no processo de classificação das imagens foi o de Florestas Aleatórias (do inglês, *Random Forest* - RF), que funciona criando uma coleção de árvores de decisão individuais, chamada de floresta, construindo vários classificadores durante o treinamento do modelo e gerando classes (classificação) ou previsões (regressão) nas árvores individuais (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Durante o processo de construção da floresta, cada árvore gerada se diferencia um pouco das outras, para que a decisão da floresta não seja tendenciosa. Um meta-algoritmo usado para que a floresta tenha tal característica, é o *bagging*, em que as árvores são criadas com base em subconjuntos do conjunto total de descritores. Assim, cada árvore da floresta fica responsável por analisar apenas uma parte dos descritores.



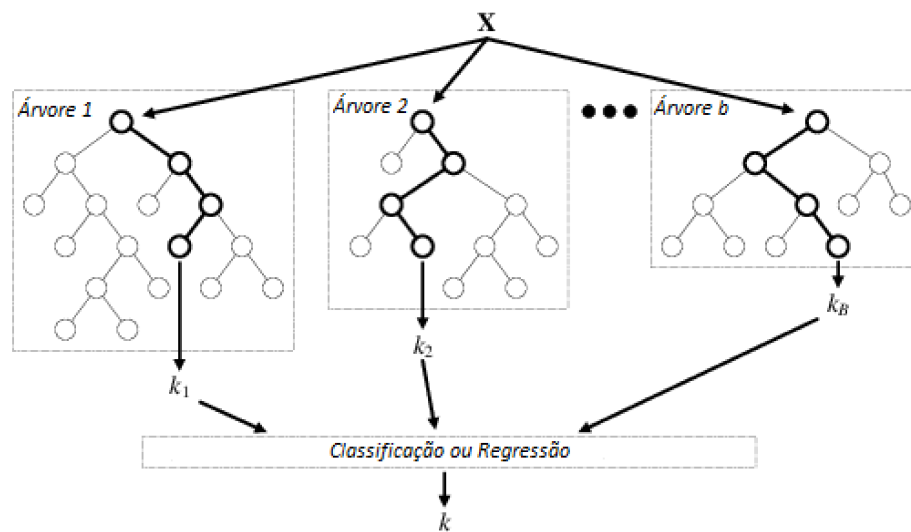


Figura 7 – Figura esquemática do funcionamento de um algoritmo de floresta de decisão (VERIKAS et al., 2016)

Os valores finais encontrados por cada uma das árvores é computado e o conjunto de soluções é analisado para se encontrar o resultado obtido pela floresta. Tal decisão pode ser feita por meio da média dos resultados individuais (regressão), ou pela moda (classificação). A Figura 7 representa o processo de tomada de decisão por uma floresta. A implementação do algoritmo de classificação não foi necessária visto que a biblioteca *Scikit-learn* também possui ótimas ferramentas para a criação de classificadores e regressores baseados em árvores de decisão. Após testes empíricos, o tamanho da floresta foi definida em duzentas árvores, entretanto, todas as outras configurações do classificador foram mantidas nos valores padrão da biblioteca.

### 3.7 Avaliação dos Métodos

A análise estatística baseada no teste de Wilcoxon-Mann-Whitney foi usada para medir a relevância das características para discriminação entre os dados. No teste, é analisada a hipótese de que as características de dois conjuntos (por exemplo, a entropia de Shannon das imagens de displasias moderadas e severas) pertencem a um único conjunto, com uma só média (MANN; WHITNEY, 1947). Um teste com que resulte em um *p-value* igual a 0,05 pode ter sua hipótese rejeitada com 95% de certeza, indicando que o conjunto de informações pertence a dois ou mais grupos com médias distintas. O teste foi usado primordialmente para verificar a porcentagem de características agrupáveis em cada grupo binário de displasias.

No caso dos descritores, também foram analisadas as correlações. Para isso, foi usado o coeficiente de correlação de Pearson, que mede o grau de correlação entre duas

variáveis X e Y, calculado por:

$$p = \frac{\text{covariância}(X,Y)}{\sqrt{\text{variância}(X) * \text{variância}(Y)}}$$

Nesse teste, uma correlação  $p = 1$  é considerada uma correlação positiva perfeita, indicando um comportamento único entre as duas variáveis. Uma correlação  $p = -1$  demonstra uma correlação negativa perfeita, em que as duas variáveis têm comportamentos completamente opostos, ou seja, o crescimento de uma variável simboliza a queda da outra. Quanto mais próximo de 0 for o valor de  $p$ , menor é a correlação entre as variáveis, sendo que valores com  $|p| < 0,3$  indicam uma correlação irrelevante.

Para apresentar o desempenho do classificador, foram aferidas taxas dependentes da matriz de confusão gerada à partir da execução do algoritmo. A matriz de confusão guarda os a quantidade de instâncias classificadas incorretamente como positivas (*false positives* - fp), classificadas corretamente como positivas (*true positives* - tp), as classificadas incorretamente como negativas (*false negatives* - fn) e as classificadas corretamente como negativas (*true negatives* - tn). De acordo com (SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006), as seguintes métricas foram aferidas:

- **Acurácia**, que é a proporção de classificações feitas corretamente, representando a probabilidade de que um teste de diagnóstico seja corretamente realizado, ou seja, o quão bem o classificador diagnostica os verdadeiros positivos e verdadeiros negativos;

$$Ac = \frac{tp + tn}{tp + fp + fn + tn} \quad (3.2)$$

- **Sensibilidade**, que é a probabilidade do resultado do diagnóstico ser positivo sendo que o teste é positivo, mostrando quantitativamente os graus de flutuação do classificador, ou seja, a capacidade do modelo identificar um positivo verdadeiro;

$$Se = \frac{tp}{tp + fn} \quad (3.3)$$

- **Especificidade**, que é a proporção de negativos verdadeiros, ou seja, a probabilidade do resultado do diagnóstico ser negativo quando o teste analisado é negativo.

$$Es = \frac{tn}{fp + tn} \quad (3.4)$$

- **F1-score**, que representa a média harmônica entre sensibilidade e acurácia, a fim de combinar essas duas métricas em uma única. Seu valor é calculado como informado em (SASAKI et al., 2007), através da equação:

$$F = 2 * \frac{Se * Ac}{Se + Ac} \quad (3.5)$$

## 4 Resultados

Neste capítulo são apresentados os resultados obtidos no estudo, assim como os experimentos realizados para análise dos descritores com algoritmo de aprendizagem de máquina.

### 4.1 Análise dos Descritores Morfológicos e Não Morfológicos

A primeira técnica usada na investigação dos descritores foi o uso do teste de Wilcoxon-Mann-Whitney, um teste estatístico não paramétrico que analisa duas amostras para verificar se pertencem ou não à mesma população. O teste permite avaliar a relevância dos descritores na distinção entre as classes, indicando amostras distintas e o nível de separação das classes do conjunto usadas como amostra. A hipótese analisada foi a de que descritores de duas classes de displasia pertencem a uma única classe, e o teste de um descritor gera um  $p$ -value que indica o grau de certeza de que a hipótese testada é verdadeira.

As Figuras 8, 9, 10, 11 e 12 mostram os gráficos da função de distribuição acumulada em relação a  $p$ -valores ( $p$ -values), indicando a porcentagem de descritores observados que são separáveis entre duas classes. Cada um dos seguintes grupos de características foi analisado separadamente: i) descritores morfológicos; ii) descritores de Haralick; iii) descritores da SCM; iv) descritores obtidos por entropia; v) associação de todos os descritores.

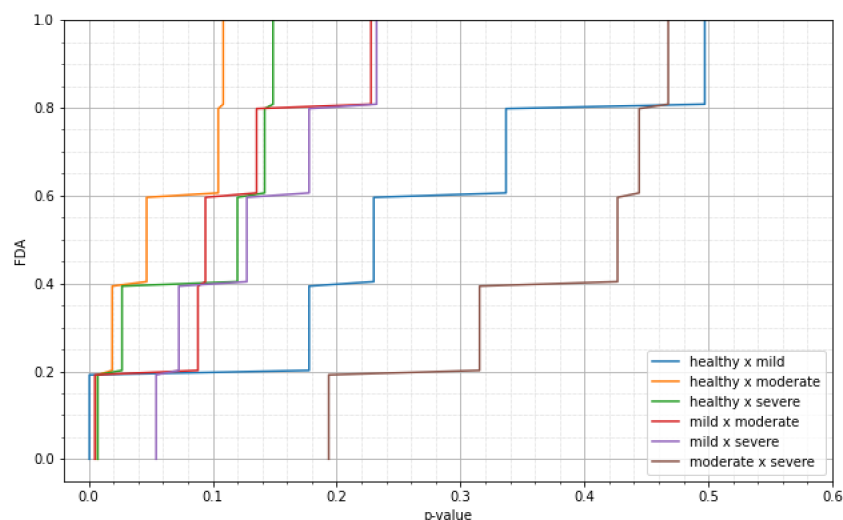


Figura 8 – Função de distribuição acumulada empírica do  $p$ -value das classes de displasia com os descritores morfológicos.

Como apresentado na Figura 8, o teste feito com os descritores morfológicos não encontrou bons resultados na comparação entre as classes moderada e severa. Analisando essas classes, não foi possível encontrar um único descritor morfológico que representasse bem as duas populações distintas, com  $p$ -value abaixo de 0,05. Isso pode indicar, por exemplo, a necessidade da análise de um conjunto de dois ou mais descritores para que essas populações possam ser discretizadas.

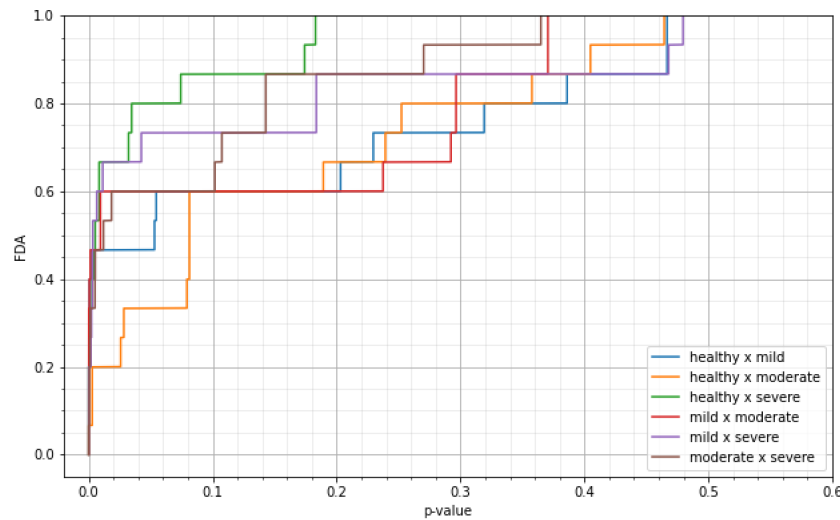


Figura 9 – Função de distribuição acumulada empírica do  $p$ -value das classes de displasia com os descritores de Haralick.

Analisando o gráfico calculado em função dos descritores de Haralick (Figura 9) e usando como base um  $p$ -value = 0,05, é possível observar que cerca de 80% dos descritores correspondem a esse valor para a comparação entre as classes saudável e severa. Isso permite considerar que cerca de 80% dos descritores testados descrevem as classes como pertencendo a populações diferentes. Contrariamente, cerca de apenas 34% dos descritores obtiveram resultado significativo na comparação entre as classes saudável e moderada.

Ao contrário dos descritores de Haralick (Figura 9), os descritores da SCM (Figura 10) não obtiveram bons resultados no teste para diferenciação entre as classes de displasia leve e moderada, pois apenas cerca de 34% dos descritores representaram as classes como populações distintas, considerando um  $p$ -value abaixo de 0,05. Em compensação, os descritores obtiveram seus melhores resultados nos comparativos envolvendo a classe leve, resultando em 100% de aproveitamento dos descritores quando a classe foi comparada às classes saudável e severa, e cerca de 90% no comparativo feito com a classe moderada.

Dados os bons resultados encontrados com os descritores não-morfológicos, foram adicionados às análises novos descritores de entropia, sendo eles: entropia de Shannon, entropia de Renyi, entropia de Tsallis, entropia de Kapur e índice de Moran, que foram comparados entre si para verificar a possibilidade de melhora que eles podem fornecer

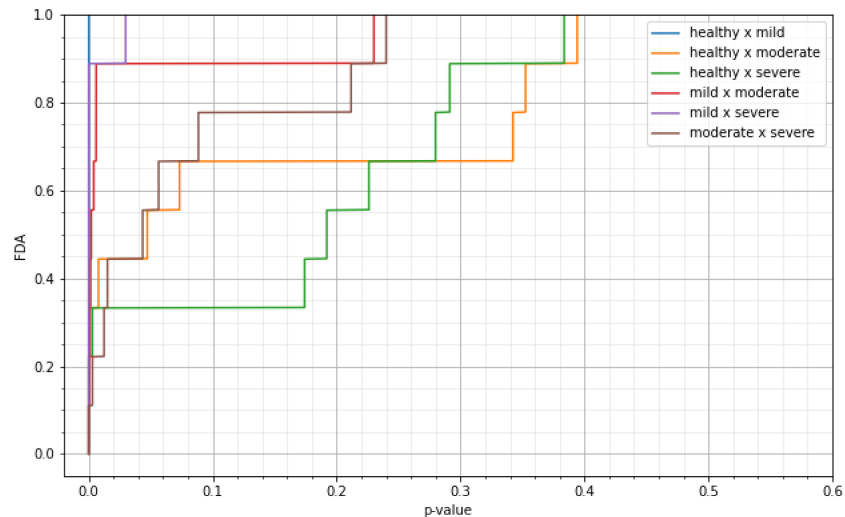


Figura 10 – Função de distribuição acumulada empírica do  $p$ -value das classes de displasia com os descritores da SCM.

ao conjunto de descritores já informados. Na Figura 11, é apresentado um comparativo dos valores dos descritores baseados em entropia. Nesse caso pode ser notado que cerca de 44% dos descritores já consegue diferenciar todas as classes com  $p < 0,05$ . Porém, em alguns comparativos de classes, esse valor chegou a 70%, sendo o maior valor encontrado para o comparativo moderada e severa, e 100% em metade dos comparativos.

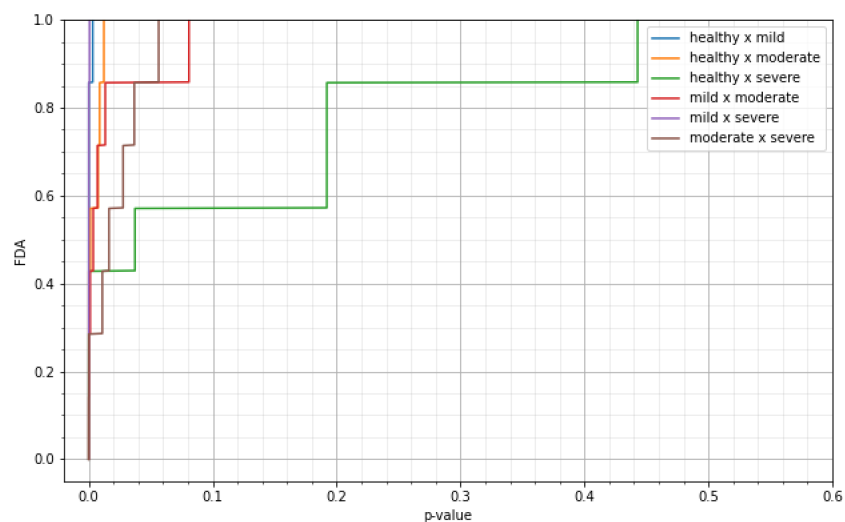


Figura 11 – Função de distribuição acumulada empírica do  $p$ -value das classes de displasia com os descritores de entropia.

Por tratarem do mesmo tipo de informação, os descritores de entropia não foram utilizados na comparação de todos os descritores, sendo analisados separadamente. Na Figura 12, os descritores (exceto os de entropia) foram agrupados para que pudesse ser feita uma análise do conjunto completo. Fica claro que a junção dos descritores colabora para aumentar o número de grupos que apresentam características separáveis, assim como

mantém a combinação de descritores morfológicas e não-morfológicas interessante para uma investigação na classificação. Isso pode ser notado pelo fato de que em todas as comparações binárias de classes (linhas do gráfico), pelo menos 60% dos descritores contém indícios de pertencerem a populações distintas, com cerca de 95% de certeza ( $p\text{-value} = 0,05$ ).

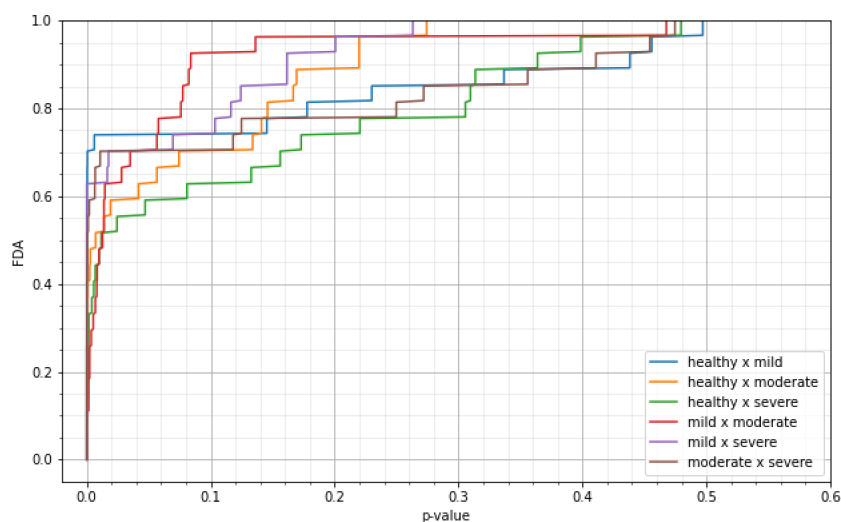


Figura 12 – Função de distribuição acumulada empírica do  $p\text{-value}$  das classes de displasia com todos os descritores (exceto entropias).

Nas Figuras 13, 14, 15 e 16, são apresentados os resultados da investigação de correlação de *Pearson* entre as informações dos descritores de Haralick, descritores da SCM e descritores morfológicos em uma representação de mapa de calor.

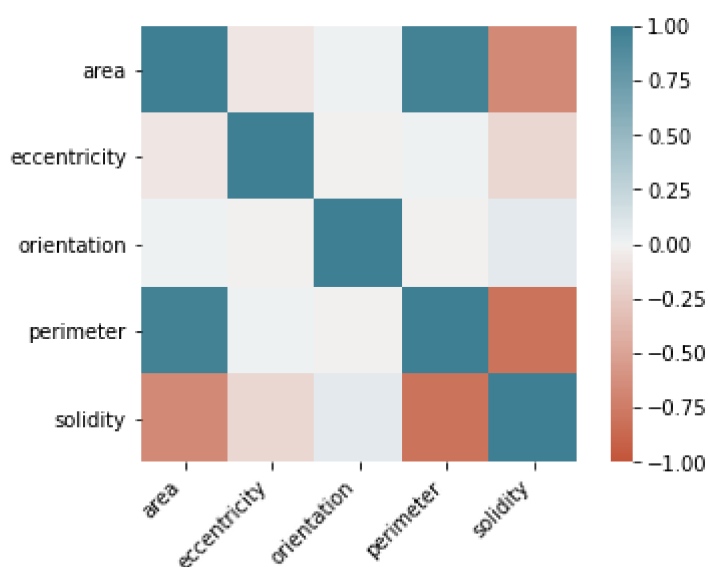


Figura 13 – Mapa de calor da correlação dos descritores morfológicos.

Como observado na Figura 13, dos cinco descritores morfológicos calculados, três

possuem alto nível de correlação. Os descritores *Perímetro* e *Área* possuem um comportamento semelhante, que é oposto ao comportamento do descritor *Solidez*.

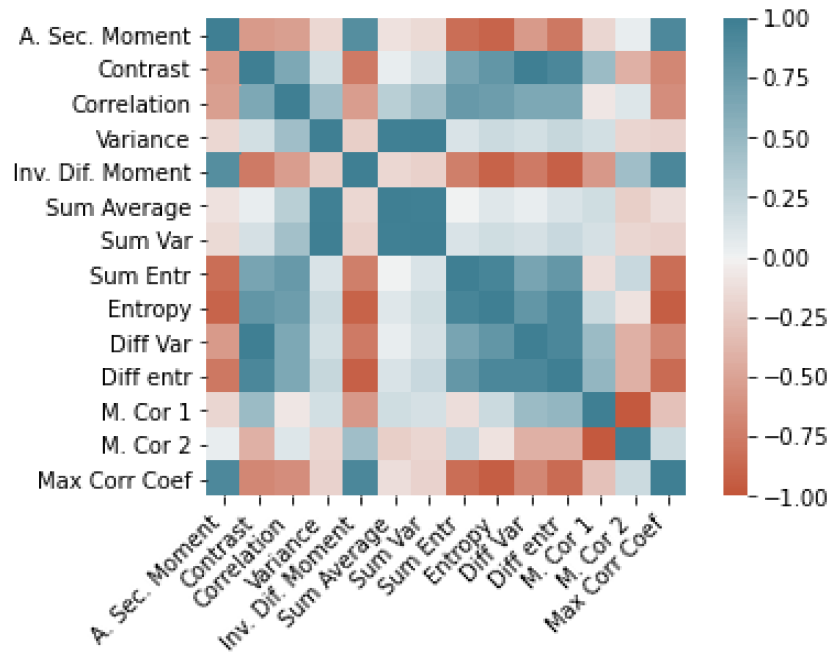


Figura 14 – Mapa de calor avaliando a correlação dos descritores de Haralick.

Entre os descritores de Haralick, foi possível notar um grande conjunto de descritores com comportamentos similares. Como podemos observar na Figura 14, os descritores entropia, entropia da soma, variância da diferença e entropia da diferença possuem correlação positiva de níveis médio e alto, indicando um comportamento oposto ao do conjunto de descritores que inclui o segundo momento amostral, momento inverso da diferença e coeficiente de correlação maximal, que são positivamente correlacionados entre si.

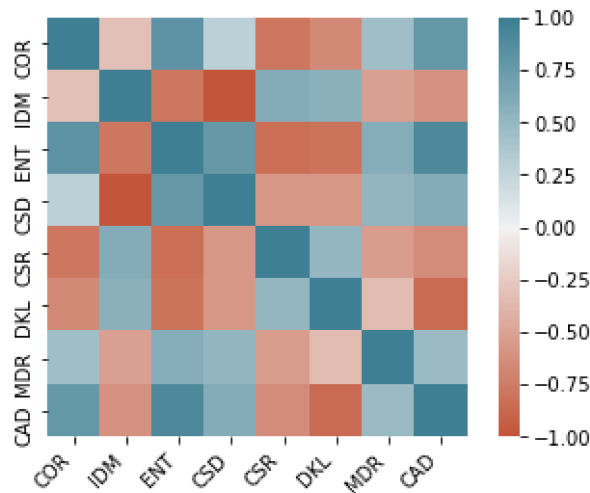


Figura 15 – Investigação da correlação dos descritores da SCM.

Entre todos os tipos de descritores analisados, a correlação dos descritores da SCM, apresentado na Figura 15, foi o que apresentou maiores índices de correlação, com poucos pontos em que  $|p| < 0,3$ . Esses valores mostram que parte das informações pode ser interessante para a classificação, e o uso de métodos de redução da dimensionalidade pode contribuir para obtenção de valores elevados de acurácia.

Na Figura 16, é apresentada a correlação entre os descritores de entropia e o índice de Moran. Como é possível verificar, a entropia de Tsallis e o índice de Moran são as características mais distintas de outras características, as entropias de Kapur, de Shannon e de Renyi possuem alta similaridade positiva.

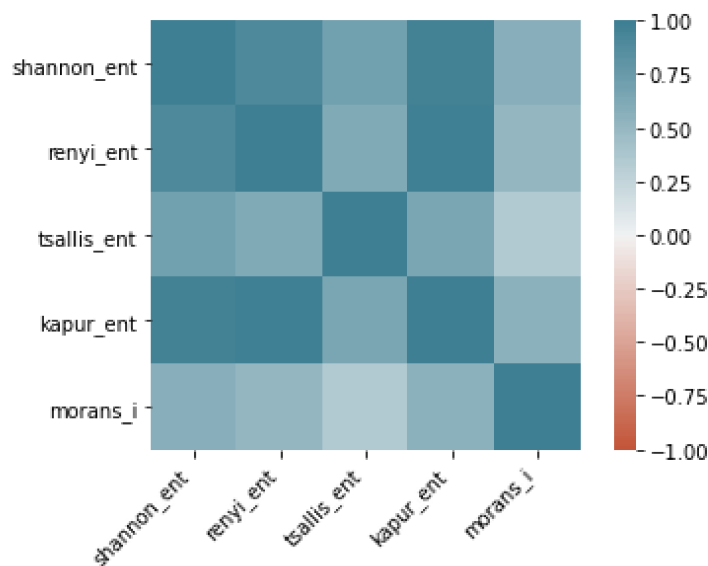


Figura 16 – Investigação da correlação dos descritores obtidos com entropia.

## 4.2 Classificação com os Descritores

Para os experimentos, foi usada validação cruzada (*cross-validation*) com o método *K-fold* estratificado, em que o conjunto de dados é separado em  $K$  grupos disjuntos, e a cada rodada,  $K-1$  grupos são usados para o treinamento do modelo e um grupo diferente é usado para teste. O valor  $K$  escolhido para os experimentos foi 5, e por isso, a cada experimento, o conjunto de dados foi usado na proporção 4:1, referente ao montante usado para treinamento e teste, respectivamente. Para cada experimento, foram usados os classificadores *Random Forest*, com florestas de 200 árvores de profundidade e quantidade de folhas máxima indefinidas, *splits* de no mínimo dois *samples*, sem associação de pesos e sem *threshold* de parada definido. Assim, o único critério de parada foi a criação de árvores com todos os nós-folha com apenas um *sample*.

Ao final de cada experimento, foram obtidos cinco valores de acurácia, referentes aos cinco testes de validação cruzada efetuados, e o valor considerado para a acurácia



do experimento foi a média dos cinco valores obtidos nesses testes. Para obter resultados menos dependentes de variáveis aleatórias, cada experimento foi realizado dez vezes, e portanto, cada valor apresentado abaixo equivale à acurácia média calculada a partir de um total de cinquenta testes. A Tabela 2 mostra os resultados obtidos com os descritores morfológicos. De forma semelhante ao observado nos gráficos da análise estatística, em que grande parte das informações morfológicas não se apresentaram separáveis entre as classes de lesões investigadas, os resultados da métrica têm valores baixos, proporcionando valores mais relevantes no grupo saudável x leve, e resultados muito baixos com o grupo moderado x severo.

Tabela 2 – Acurácia obtida com uso dos descritores morfológicos.

Classes	Acurácia
saudável x leve	67,67%
saudável x moderado	59,33%
saudável x severo	66,00%
leve x moderado	53,83%
leve x severo	60,67%
moderado x severo	33,67%
multi-classe	39,00%

A Tabela 3 mostra os resultados obtidos com os descritores não-morfológicos baseados nas matrizes de co-ocorrências. Como é possível notar, os conjuntos de classes obtiveram comportamentos semelhantes nos dois classificadores, com os grupos saudável x leve obtendo os melhores resultados e o grupo leve x moderado obtendo os piores. Além disso, nota-se que a junção das informações entre as matrizes contribui para a classificação multi-classes. No entanto, dentre os grupos binários de investigação, os resultados com a junção dos classificadores foram inferiores aos da SCM, com exceção dos grupos saudável x moderado e leve x severo.

Tabela 3 – Acurácia obtida com uso dos descritores não morfológicos baseados em matrizes de co-ocorrências.

Classes	Haralick	SCM	Haralick + SCM
saudável x leve	80,02%	85,29%	85,22%
saudável x moderado	77,84%	80,57%	80,57%
saudável x severo	72,25%	80,12%	79,85%
leve x moderado	65,81%	69,70%	69,28%
leve x severo	73,53%	75,66%	76,01%
moderado x severo	80,32%	79,00%	78,94%
multi-classe	52,86%	59,74%	60,26%

### 4.3 Classificação com Associação de Descritores

Com o resultado dos testes, foi possível perceber que os descritores não-morfológicos, que em grande parte representam as informações de textura, se mostraram instrumentos eficazes no processo de classificação das imagens. Eles alcançaram resultados individuais melhores do que os obtidos na classificação feita com os descritores morfológicos. Assim, outros descritores não-morfológicos baseados em entropia também foram usados para verificar a capacidade de melhora proporcionada por eles ao sistema. Nesse caso, o objetivo foi avaliar a influência de cada técnica baseada em entropia e sua influência sobre o modelo. Os resultados na Tabela 4 mostram as acurácias obtidas com as classificações feitas com o grupo de descritores anteriormente mencionados, compostos por Morfológicos, Haralick e SCM, denominados por MHC, assim como classificações feitas com a junção de novos descritores: entropia de Shannon, entropia de Renyi, entropia de Tsallis e índice de Moran. Associando o MHC aos novos descritores, foram obtidos melhores resultados do que os representados com o uso exclusivo do MHC. Além disso, observa-se que a combinação de MHC e entropia de Tsallis proporcionou o resultado mais expressivo no grupo multi-classe. Isso também ocorre em alguns grupos binários dessa investigação (saudável x leve e leve x severo).

Tabela 4 – Acurácia obtida com uso do classificador Random Forest.

Classes	MHC	Shannon	Renyi	Tsallis	Kapur	Moran
saudável x leve	92,42%	92,73%	93,03%	<b>94,88%</b>	93,03%	92,73%
saudável x moderado	92,41%	<b>94,05%</b>	93,69%	93,61%	93,08%	93,15%
saudável x severo	86,18%	85,82%	84,73%	<b>91,27%</b>	83,27%	84,00%
leve x moderado	92,86%	<b>93,71%</b>	92,86%	93,43%	93,43%	92,86%
leve x severo	86,21%	86,15%	85,56%	<b>86,44%</b>	85,18%	86,21%
moderado x severo	<b>89,30%</b>	89,27%	88,88%	86,85%	88,21%	87,82%
multi-classe	78,97%	80,25%	78,97%	<b>81,53%</b>	79,12%	80,57%

### 4.4 Avaliação do Regularizador L1

Para a avaliação com o regularizador, os descritores foram separados em três grupos de acordo com sua origem (descritores de Haralick, descritores da SCM e descritores morfológicos), e o regularizador foi aplicado em todos os conjuntos distintos formados por um ou mais grupos de descritores, para avaliar a capacidade do descritor de representar a imagem, quando em conjunto a outros tipos de descritores. As informações sobre entropia e índice de Moran não foram tratadas pelo regularizador devido ao fato de que, em alguns experimentos, o regularizador eliminou essas informações do conjunto. Na Figura 17, os descritores foram ordenados de acordo com a frequência normalizada em que foram selecionados pelo regularizador. Assume-se que um descritor com alta frequência tenha um alto potencial de representação da imagem, porém um descritor selecionado poucas

vezes não necessariamente tem um baixo poder de representação da imagem, mas pode apenas não ter se sobressaído tanto quanto os outros descritores usados na comparação entre os grupos.

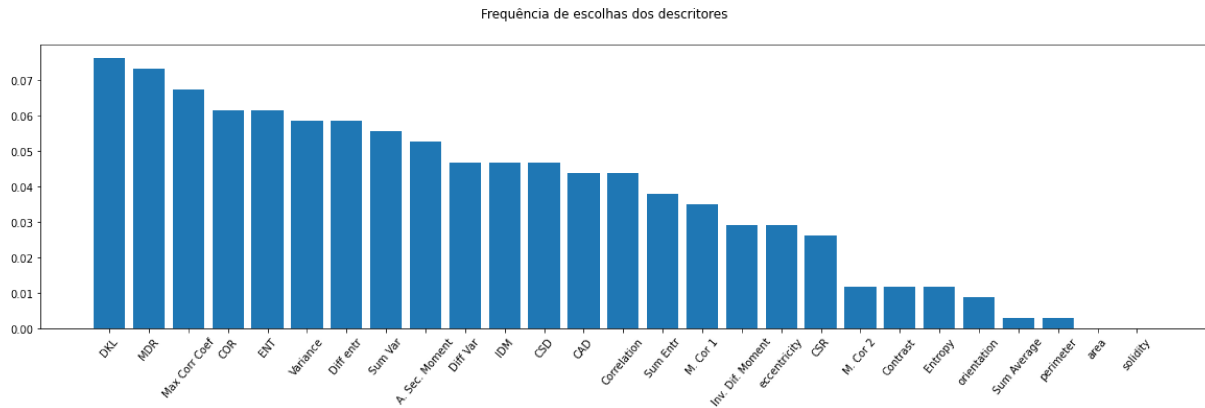


Figura 17 – Frequência de representação dos descritores.

Na Figura 18, é mostrada a influência dos descritores nos resultados alcançados. Os descritores foram selecionados um a um na ordem apresentada na Figura 17, do descritor com maior frequência para o com a menor (descritores DKL e *solidity*, respectivamente). Como é possível analisar, em um conjunto com poucos descritores, a inclusão de um novo descritor com alto potencial de representatividade promove uma melhora expressiva no resultado encontrado pelo classificador, como notado no intervalo entre o primeiro e o décimo descritor.

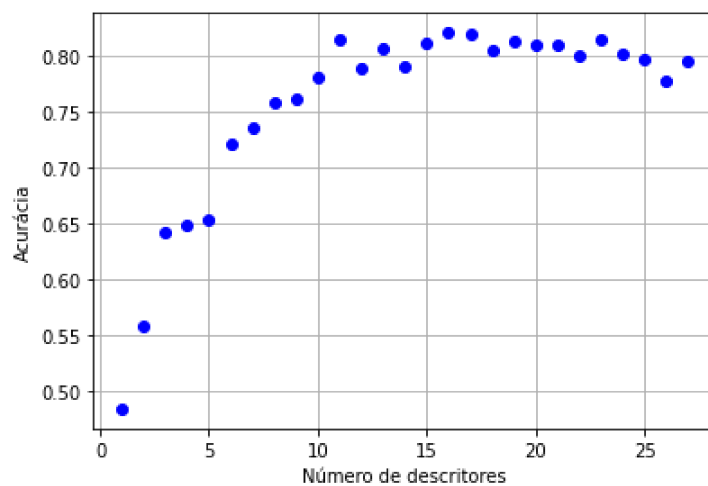


Figura 18 – Representação da métrica acurácia por número de descritores.

À medida que o conjunto de descritores aumenta, é possível notar que a melhoria nos resultados causada pela adição de um novo descritor é reduzida (intervalo entre o décimo e décimo quinto descritor). Ainda, caso o descritor adicionado tenha baixo poder de representação da imagem, é possível que ele provoque uma piora nos resultados, como

observado no gráfico a partir do décimo sétimo descritor. Assim, a redução dos descritores se deu pelo ponto mais alto alcançado no gráfico, com dezesseis descritores, promovendo a remoção de onze dos vinte e sete descritores usados até o momento, uma redução significativa de 40,74% na dimensionalidade dos dados.

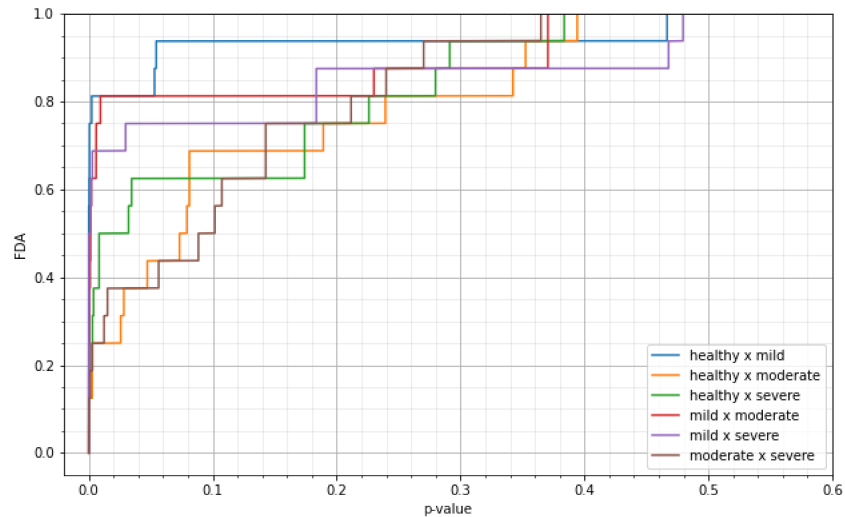


Figura 19 – Função de distribuição acumulada com os descritores selecionados com L1.

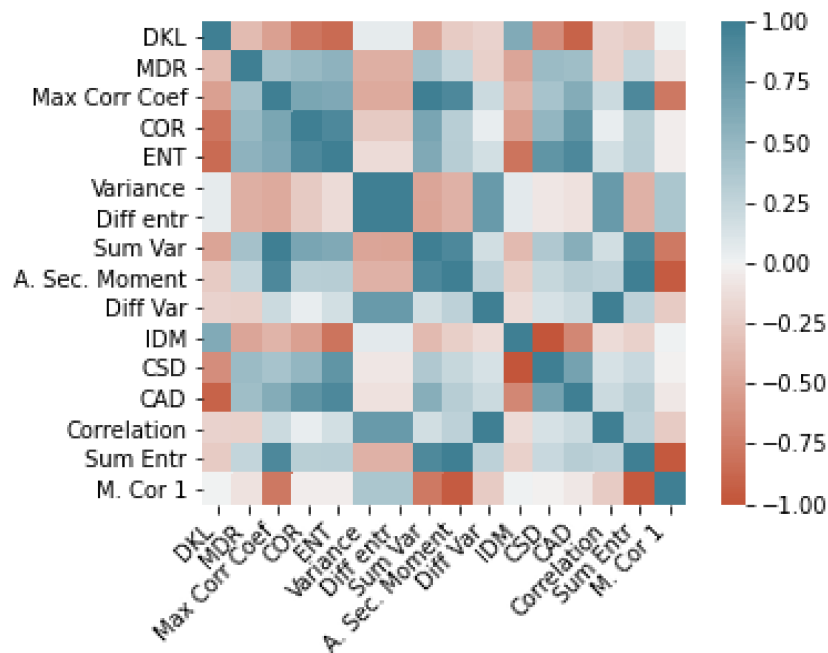


Figura 20 – Mapa de calor de correlação dos descritores selecionados com L1.

Mesmo após a eliminação de alguns descritores, foi notado que alguns grupos de classes ainda possuíam altos índices no teste de Wilcoxon-Mann-Whitney (ver Figura 19), os mais notáveis sendo os grupos de classes leve x saudável e leve x moderada, ambos com mais de 80% dos descritores com  $p \leq 0,05$ , mostrando a presença de muitas características capazes de diferenciar essas classes entre si. Porém, analisando a correlação do conjunto de

descritores como um todo (ver Figura 21), é possível ver que a maior parte dos descritores possui comportamentos heterogêneos em relação aos demais, o que pode ser averiguado visivelmente pelos tons claros no mapa de calor.

Utilizando os dezesseis primeiros descritores (DKL, MDR, coeficiente máximo de correlação, COR, ENT, variância, entropia da diferença, variância da soma, segundo momento amostral, variância da diferença, IDM, CSD, CAD, correlação, entropia da soma e medida de correlação da informação I), foram obtidos os resultados apresentados na Tabela 5. Deste momento em diante, os dezesseis descritores selecionados serão referenciados como MHC\*. Como é possível observar, o uso do regularizador não teve grande impacto nos resultados do sistema, com melhorias no resultado. No entanto, o maior benefício do uso do regularizador foi a eliminação de 40,74% dos descritores, reduzindo a dimensionalidade dos dados e o custo computacional, considerado um grande problema nos algoritmos baseados em árvores de decisão. Os resultados apresentados abaixo foram obtidos com a combinação de descritores que promoveu a melhor acurácia no classificador multi-classe (MHC\* e entropia de Tsallis).

Tabela 5 – Acurácia obtida com uso do classificador *Random Forest*.

Classes	MHC*	Shannon	Renyi	Tsallis	Kapur	Moran
saudável x leve	93,03%	92,42%	92,42%	<b>95,49%</b>	92,42%	92,42%
saudável x moderado	92,69%	93,37%	92,36%	<b>94,28%</b>	92,41%	92,41%
saudável x severo	85,45%	85,09%	85,82%	<b>92,36%</b>	87,27%	85,45%
leve x moderado	<b>93,41%</b>	92,57%	93,14%	93,14%	92,29%	91,69%
leve x severo	85,82%	86,18%	85,13%	<b>87,13%</b>	85,21%	84,49%
moderado x severo	91,45%	90,76%	91,45%	91,48%	<b>92,15%</b>	91,48%
multi-classe	81,37%	81,70%	82,33%	<b>82,49%</b>	81,69%	80,56%

## 4.5 Avaliação do Classificador

A Figura 21 mostra uma matriz de confusão obtida com 10 testes consecutivos, cada um deles com um sistema de *cross-validation K-fold* estratificado 4:1, como apresentado anteriormente na seção 4.2. No eixo horizontal da figura estão as classes reais das imagens, e no eixo vertical, as classes que foram atribuídas às imagens pelo classificador. Dentre as imagens da classe saudável, mil trezentas e quatorze foram classificadas corretamente, sessenta e nove foram classificadas como sendo pertencentes à classe leve, quatorze foram classificadas como pertencendo à classe moderada e trinta e quatro como pertencentes à classe severa. A partir desses dados, também são apresentadas outras métricas descritas na seção 3.7, que são usadas para aferir a qualidade do modelo proposto. Os cálculos foram feitos por classe, e o resultado dos testes se encontram na Tabela 6, assim como o resultado obtido com todas as classes. Algo a se notar nessa tabela é a sensibilidade das classes, pois um falso negativo pode trazer danos irreversíveis à saúde

de um paciente, o que torna indispensável que os sistemas CAD tenham sensibilidade elevada.

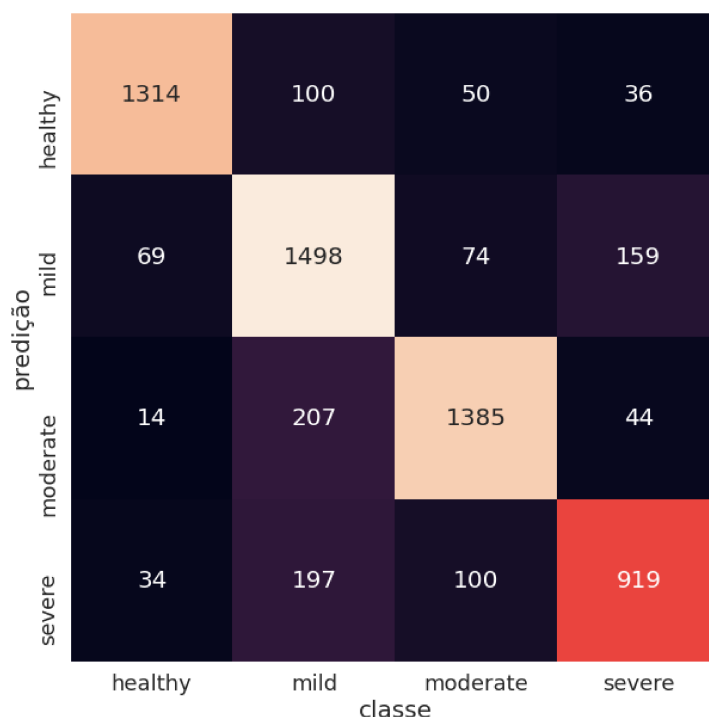


Figura 21 – Matriz de confusão do classificador *Random Forest*

Tabela 6 – Métricas para avaliação com o classificador Random Forest.

Classes	Sensibilidade	Precisão	Acurácia	Especificidade	F1-Score
saudável	91,82%	87,60%	95,11%	96,10%	0,912
leve	74,84%	83,22%	87,00%	92,81%	0,788
moderado	86,08%	83,94%	92,11%	94,23%	0,850
severo	79,36%	73,52%	90,81%	89,21%	0,763
geral	82,52%	82,52%	82,52%	94,17%	0,825

A sensibilidade da classe saudável possui o valor de 91,82%, o que significa dizer que de todas as classes que eram saudáveis, 91,82% delas foi classificada corretamente. Apesar das sensibilidades nas classes com displasia não terem sido tão altas, a sensibilidade das classes displásicas como um todo pode ser calculada pela quantidade de elementos displásicos que não foram considerados como saudáveis dividido pela quantidade total de elementos displásicos. Esse cálculo é equivalente ao da especificidade da classe saudável, que representa a porcentagem de elementos displásicos que não foi classificada como saudável. Assim, percebe-se que por mais que a sensibilidade das classes displásicas não tenha sido elevada, a probabilidade de um tecido displásico ser considerado como saudável (o pior caso possível, que coloca a vida do paciente em risco) é de menos de 4%, um resultado bem mais reconfortante.

## 4.6 Comparação com a Literatura

Como visto nos resultados, um grande esforço ainda é necessário para se alcançar acurácias mais adequadas para um sistema que possa ser disponibilizado em larga escala. Como demonstrado na seção 2.7, os sistemas CAD atuais possuem sensibilidade superior a 90% e um baixo número de falsos negativos. Analisando apenas a existência de displasias, o sistema conseguiu resultados comparáveis com os sistemas analisados, porém tais resultados não puderam ser alcançados no comparativo multi-classes. Como sugerido em (ADEL et al., 2018), este trabalho teve como proposta principal a análise de descritores no estudo e diferenciação de tipos diferentes de displasia e a criação de um método capaz de determinar o nível da displasia, entre leve, moderado, severo (carcinoma in-situ) ou saudável (sem displasia), e não somente na detecção de tecidos displásicos.

Os resultados encontrados neste trabalho estão em conformidade com o estudo de (NAG; DAS, 2018), sobretudo nos resultados encontrados na análise de descritores relacionados à morfologia dos objetos presentes nas imagens. Assim como citado no estudo, foi notado que à medida que o grau de displasia aumenta, a morfologia celular do tecido sofre alterações excessivas a ponto dos sistemas classificadores não conseguirem mais reconhecer os padrões corretamente, produzindo resultados considerados insatisfatórios em imagens de tecidos com displasia severa. Uma alternativa encontrada para esse problema foi a aplicação de descritores não-morfológicos baseados em matrizes de co-ocorrência, como a GLCM, usada na análise de imagens de lesões histopatológicas de câncer de mama feita em (KAUSHAL et al., 2019), e a SCM.

Outro problema discutido na literatura foi a dificuldade existente na diferenciação entre diferentes graus limítrofes, obstáculo que não foi encontrado neste trabalho, visto que todas as classificações entre classes limítrofes (saudável x leve, leve x moderado e moderado x severo) obtiveram acurácia superior a 90% com os descritores selecionados.

## 5 Conclusão

O câncer é uma doença que atinge milhares de pessoas todos os anos, e o câncer de boca se situa atualmente entre um dos dez tipos de câncer mais comuns no mundo. Porém, existem indicativos de que um diagnóstico precoce das lesões eleva a taxa de sobrevivência dos pacientes acometidos por essa doença, e é nessa situação que o trabalho desenvolvido se faz útil. O uso de sistemas CAD auxilia especialistas no diagnóstico de doenças, agilizando o processo de reconhecimento da lesão e aumentando a confiabilidade dos diagnósticos.

O trabalho desenvolvido buscou entender de quais formas dois tipos de descritores se comportam no processo de classificação de lesões de displasia oral. Para isso, foram aplicados descritores morfológicos e não morfológicos a processos estatísticos que permitem avaliar suas capacidades de representar as diferenças nas imagens de tecidos com classes de displasia diferentes. Ademais, foi feita uma análise da correlação entre os descritores para evitar altos custos computacionais e da análise dos resultados gerados por classificadores baseados em algoritmos de árvores de decisão no processo de diagnóstico do grau de displasia apresentado na imagem.

Como resultado, percebeu-se que os descritores de textura se mostram mais adequados ao processo de distinção binária entre as classes, com acurácias significativas que pouco variaram entre os diferentes conjuntos de classes analisados. No resultado das classificações multi-classe, o algoritmo desenvolvido conseguiu bons resultados, porém ainda um pouco longe dos encontrados em sistemas CAD atuais, e assim sendo, um próximo desafio seria o estudo de diferentes técnicas que proporcionem o aprimoramento do modelo de classificação multi-classe, além da aplicação do algoritmo em diferentes tipos de lesões, a fim de testar o grau de especificidade do algoritmo.

### 5.1 Trabalhos Futuros

Os descritores morfológicos analisados neste trabalho não alcançaram os resultados esperados, sendo descartados pelos regularizadores. Como é inegável a existência de diferenças morfológicas nos tecidos de diferentes graus de displasia, outros descritores morfológicos podem ser analisados para verificar suas eficácias no auxílio do processo de classificação das lesões, além do estudo de outros tipos de regularizadores, capazes de manter os descritores morfológicos no conjunto de características usado, permitindo a classificação com um conjunto de características diversificado. No estudo de (ADEL et al., 2018) são apresentados bons resultados alcançados com características extraídas com o algoritmo ORB (*Oriented FAST and Rotated BRIEF*), sendo assim um possível ponto



de partida.

Além disso, como os descritores não-morfológicos se mostraram relevantes para os resultados encontrados neste trabalho, estudos futuros podem ser feitos para analisar a interferência de novos descritores de textura no processo de classificação, pois o estudo de novas técnicas pode ajudar a encontrar modelos ou classificadores com acurácia ainda mais elevada.

# Referências

- ADEL, D. et al. Oral epithelial dysplasia computer aided diagnostic approach. In: IEEE. *2018 13th International Conference on Computer Engineering and Systems (ICCES)*. [S.l.], 2018. p. 313–318. Citado 3 vezes nas páginas 30, 50 e 51.
- AZEVEDO-MARQUES, P. M. d. DiagnÃauxiliado por computador na radiologia. *Radiologia Brasileira*, scielo, v. 34, p. 285 – 293, 10 2001. ISSN 0100-3984. Disponível em: <[http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S0100-39842001000500008&nrm=iso](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-39842001000500008&nrm=iso)>. Citado na página 23.
- BRASIL; CÂNCER, M. da Saúde e Instituto Nacional de. *Estimativa 2012: incidência de câncer no Brasil*. [S.l.]: Inca Rio de Janeiro, 2011. Citado 2 vezes nas páginas 12 e 15.
- CALADO, A. História do ensino de histologia. *História da Ciência e Ensino: construindo interfaces*, v. 20, n. 0, p. 455–466, 2019. ISSN 2178-2911. Disponível em: <<https://revistas.pucsp.br/index.php/hcensino/article/view/44779>>. Citado 3 vezes nas páginas 17, 18 e 19.
- CHATTORAJ, S.; VISHWAKARMA, K. Classification of histopathological breast cancer images using iterative vmd aided zernike moments & textural signatures. *arXiv preprint arXiv:1801.04880*, 2018. Citado na página 26.
- COOK, H.; FOX, M. Application of expert systems to mammographic image analysis. *American journal of physiologic imaging*, v. 4, n. 1, p. 16—22, 1989. ISSN 0885-8276. Disponível em: <<http://europepmc.org/abstract/MED/2923720>>. Citado na página 23.
- DOI, K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, v. 31, p. 198–211, 06 2007. Citado 2 vezes nas páginas 21 e 22.
- DUARTE, Y. A.; NASCIMENTO, M. Z. D.; OLIVEIRAS, D. L. Algoritmo de extração de textura baseado em wavelet e clbp para classificação de lesões em mamogramas. In: *XIII Workshop de Informática Médica*. [S.l.: s.n.], 2013. p. 174–183. Citado na página 25.
- ERNST, H. C. The development of the microscope. *Journal of the Boston Society of Medical Sciences*, American Society for Investigative Pathology, v. 4, n. 6, p. 148, 1900. Citado na página 19.
- FIRMINO, M. et al. Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy. *BioMedical Engineering OnLine*, v. 15, 01 2016. Citado 2 vezes nas páginas 12 e 23.
- FOUNDATION, O. H. *State of Mouth Cancer UK Report 2018/2019*. [S.l.], 2018. Disponível em: <<https://www.dentalhealth.org/Handlers/Download.ashx?IDMF=21dc592b-d4e7-4fb2-98a9-50f06bed71aa>>. Citado 2 vezes nas páginas 12 e 15.

- GONZALEZ, R. C.; WOODS, R. C. *Processamento digital de imagens*. [S.l.]: Pearson Educação, 2009. Citado 4 vezes nas páginas 21, 23, 24 e 25.
- HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, Ieee, n. 6, p. 610–621, 1973. Citado na página 26.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer Science & Business Media, 2009. Citado na página 35.
- INCA, M. d. S. *Estimativa 2020: incidência de câncer no Brasil*. 2020. Disponível em: <<https://www.inca.gov.br/publicacoes/livros/estimativa-2020-incidencia-de-cancer-no-brasil>>. Citado 3 vezes nas páginas 12, 13 e 15.
- JR, R. L. E. Attempts to use computers as diagnostic aids in medical decision making: a thirty-year experience. *Perspectives in biology and medicine*, Johns Hopkins University Press, v. 35, n. 2, p. 207–219, 1992. Citado na página 12.
- JUNQUEIRA, L. C. U.; CARNEIRO, J. *Histologia Básica Texto e Atlas*. 13. ed. [S.l.]: EDITORA GUANABARA KOOGAN LTDA, 2018. Citado 4 vezes nas páginas 6, 17, 19 e 20.
- KAUSHAL, C. et al. Recent trends in computer assisted diagnosis (cad) system for breast cancer diagnosis using histopathological images. *IRBM*, Elsevier, v. 40, n. 4, p. 211–227, 2019. Citado 4 vezes nas páginas 6, 22, 30 e 50.
- LODWICK, G. S. et al. Computer diagnosis of primary bone tumors: A preliminary report. *Radiology*, The Radiological Society of North America, v. 80, n. 2, p. 273–275, 1963. Citado na página 12.
- LUMERMAN, H.; FREEDMAN, P.; KERPEL, S. Oral epithelial dysplasia and the development of invasive squamous cell carcinoma. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, Elsevier, v. 79, n. 3, p. 321–329, 1995. Citado na página 33.
- MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, JSTOR, p. 50–60, 1947. Citado na página 36.
- MICROSYSTEMS, L. *Leica DM500*. Leica Microsystems. Disponível em: <<https://www.leica-microsystems.com/pt/produtos/microscopios-oticos/detalhes/leica-dm500/gallery/>>. Citado 2 vezes nas páginas 6 e 20.
- MONTANARI, T. *Histologia: texto, atlas e roteiro de aulas práticas*. Ed. do autor, 2016. Citado na página 13.
- NAG, R.; DAS, R. K. Analysis of images for detection of oral epithelial dysplasia: A review. *Oral oncology*, Elsevier, v. 78, p. 8–15, 2018. Citado 5 vezes nas páginas 6, 15, 16, 30 e 50.
- PEDRINI, H.; SCHWARTZ, W. R. *Análise de Imagens Digitais: princípios, algoritmos e aplicações*. [S.l.: s.n.], 2008. Citado na página 25.

- RAMALHO, G. L. B. et al. Rotation-invariant feature extraction using a structural co-occurrence matrix. *Measurement*. Citado 3 vezes nas páginas 28, 32 e 34.
- RANGANATHAN, K.; KAVITHA, L. Oral epithelial dysplasia: Classifications and clinical relevance in risk assessment of oral potentially malignant disorders. *Journal of Oral and Maxillofacial Pathology : JOMFP*, v. 23, p. 19 – 27, 2019. Citado 3 vezes nas páginas 8, 15 e 16.
- RÉNYI, A. *On measures of entropy and information*. [S.l.], 1961. Citado na página 25.
- RODRIGUES, K. A. S. Lasso clássico e bayesiano. Citado na página 35.
- SASAKI, Y. et al. The truth of the f-measure. *Teach Tutor mater*, v. 1, n. 5, p. 1–5, 2007. Citado na página 37.
- SHANNON, C. E. A mathematical theory of communication. *Bell system technical journal*, Wiley Online Library, v. 27, n. 3, p. 379–423, 1948. Citado na página 25.
- SILVA, A. B. et al. Métodos computacionais para análise e classificação de displasias em imagens da cavidade bucal. Universidade Federal de Uberlândia, 2019. Citado 3 vezes nas páginas 21, 32 e 33.
- SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: SPRINGER. *Australasian joint conference on artificial intelligence*. [S.l.], 2006. p. 1015–1021. Citado na página 37.
- TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Citado na página 35.
- TSALLIS, C. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, Springer, v. 52, n. 1-2, p. 479–487, 1988. Citado na página 26.
- VERIKAS, A. et al. Electromyographic patterns during golf swing: Activation sequence profiling and prediction of shot effectiveness. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 16, n. 4, p. 592, 2016. Citado 2 vezes nas páginas 6 e 36.
- WARNAKULASURIYA, S. Causes of oral cancer—an appraisal of controversies. *British dental journal*, Nature Publishing Group, v. 207, n. 10, p. 471, 2009. Citado 2 vezes nas páginas 12 e 15.