UNIVERSIDADE FEDERAL DE PERNAMBUCO
DEPARTAMENTO DE ENERGIA NUCLEAR

COMISSÃO NACIONAL DE ENERGIA NUCLEAR
CENTRO REGIONAL DE CIÊNCIAS NUCLEARES DO NORDESTE

PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS ENERGÉTICAS E NUCLEARES

Análise Quantitativa de Imagens Médicas Funcionais Utilizando Métodos de Reconhecimento de Padrões e Inteligência Artificial

Igor Fagner Vieira

Orientadores:

Dr. Fernando R. de A. Lima(UFPE)
Dr. Michel Koole(KU Leuven)

Co-orientadores:

Dr. José Wilson Vieira (UPE/IFPE)
Dr. Vincent Vandecaveye (KU Leuven)

Recife-PE
Fevereiro, 2018

Igor Fagner Vieira

Análise Quantitativa de Imagens Médicas Funcionais Utilizando Métodos de Reconhecimento de Padrões e Inteligência Artificial

Tese submetida ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares para obtenção do título de Doutor em Ciências pela UFPE e KU Leuven, Bélgica, Área de Concentração: Dosimetria e Instrumentação Nuclear.

Orientador: Prof. Dr. Fernando Roberto de Andrade Lima

Dr. Michel Koole

Co-orientador: Prof. Dr. José Wilson Vieira

Dr. Vincent Vandecavey

Catalogação na fonte
Bibliotecário Carlos Moura, CRB-4 / 1502

V658a Vieira, Igor Fagner.
Análise quantitativa de imagens médicas funcionais utilizando métodos de reconhecimento de padrões e inteligência artificial. / Igor Fagner Vieira. - Recife: O Autor, 2018.
104 f. : il., tabs.

Orientador: Dr. Fernando R. de A. Lima.
Orientador: Dr. Michel Koole.
Coorientador: Dr. José Wilson Vieira.
Coorientador: Dr. Vincent Vandecaveye.

Tese (doutorado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares, 2018.
Inclui referências bibliográficas.

1. Ressonância magnética. 2. DWI. 3. Inteligência artificial. 4. Textura não-adaptativa. 5. Textura adaptativa. I. Lima, Fernando R. de A., orientador. II. Koole, Michel, orientador. III. Vieira, José Wilson, coorientador. IV. Vandecaveye, Vincent, coorientador. V. Título.

UFPE
CDD 621.48 (21. ed.) BDEN/2018-01

Igor Fagner Vieira

Análise Quantitativa de Imagens Médicas Funcionais Utilizando Métodos de Reconhecimento de Padrões e Inteligência Artificial

Tese Aprovada em: 28 de Fevereiro de 2018,

Orientador

Dr. Fernando R. de A. Lima(UFPE)
Dr. Michel Koole(KU Leuven)

Co-Orientador

Dr. José Wilson Vieira (UPE/IFPE)
Dr. Vincent Vandecaveye (KU Leuven)

Professor
Dr. Sílvio de Barros Melo

Professor
Dr. Hector Raúl Montagne Dugrós

Professor
Dr. Tiago Alessandro Espínola Ferreira

Recife-PE
Fevereiro, 2018

AGRADECIMENTOS

A todos os que direta ou indiretamente contribuíram para o desenvolvimento desse trabalho: familiares, orientadores, colegas de trabalho e amigos. A lista é enorme, tanto que não vou listar tal extensiva lista aqui. E o melhor: tem gente de diferentes lugares do mundo. A todos vocês, minha eterna gratidão!

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADC	Coeficiente de Difusão Aparente
RM	Ressonância Magnética
DW-MRI ou DWI	Ressonância Magnética por Difusão Ponderada
VOI	Volume de interesse
CAVH	Volume-Histograma acumulativo de intensidades ADC
AUC-CAVH	Área sobre a curva CAVH
GLCM	Matriz de co-ocorrência de níveis de cinza
GLRLM	Matriz dos comprimentos de agrupamentos de pixels/voxels com mesmo nível de cinza
SRE	Ênfase nos comprimentos de agrupamento curto (GLRLM)
LRE	Ênfase nos comprimentos de agrupamento longo (GLRLM)
LGRE	Ênfase nos comprimentos de agrupamento de baixo nível de cinza (GLRLM)
HGRE	Ênfase nos comprimentos de agrupamento de alto nível de cinza (GLRLM)
GLNU	Não-uniformidade de nível de cinza (GLRLM)
RLNU	Não-uniformidade de comprimentos de agrupamento de nível de cinza (GLRLM)
NGTDM	Matriz da diferença dos tons de cinza da vizinhança
GLRLM	Matriz do tamanho da zona de tonalidade de cinza
GLDM	Matriz de dependencia do nível de cinza
CV	Coeficiente de variância
ROC	Características operacionais do receptor (Curva)
PPV	Predição de Valor Positivo (%)
NPV	Predição de Valor Negativo (%)
AUC	Área sobre a curva ROC

MW	Teste não-paramétrico de Mann-Whitiney
PCA	Análise de Componentes Principais
LDA	Análise de Discriminante Linear
SVM	Support Vector Machine (Classificador)
RF	Random Forest (Classificador)
LR	Logistic Regression (Classificador)
ErT	Extremely randomized trees (Classificador)
GB	Gradient Boosting (Classificador)
KNN	K-Nearest Neighbour (Classificador)

Análise Quantitativa de Imagens Médicas Funcionais Utilizando Métodos de Reconhecimento de Padrões e Inteligência Artificial

Autor: Igor Fagner Vieira
Orientadores: Dr. Fernando R. de A. Lima(UFPE) & Dr. Michel Koole(KUL)
Dr. José Wilson Vieira (UPE/IFPE) & Dr. Vincent Vandecaveye (KUL)

RESUMO

Os métodos quantitativos para análises de imagens médicas se limitam a uma descrição não-adaptativa dos volumes de interesse (VOI) e tendem a levar em conta apenas o desempenho discriminativo, quer seja das variáveis de interesse, quer seja dos modelos de aprendizagem de máquina utilizados na tarefa de classificação. O objetivo deste trabalho foi propor uma metodologia para avaliar conjuntamente o desempenho e estabilidade de modelos avançados de classificação e de métodos adaptativos de extração de características descrevendo textura de um dado VOI, visando caracterizar, respectivamente, linfonodos pélvicos e efeito do tratamento de tumores gastrointestinais. Para isso, utilizou-se em ambos os casos imagens de ressonância magnética por difusão ponderada (DWI) e o desempenho de cada modelo foi medido em termos da área sobre a curva ROC (AUC), enquanto a estabilidade foi medida em função do coeficiente de variação (CV), ambos estimados em amostras bootstrap geradas a partir das imagens iniciais. Testes estatísticos não-paramétricos projetados para comparar o desempenho das diferentes abordagens, em diferentes amostras, treinados e testados sob as mesmas condições, foram utilizados para determinar quais modelos apresentaram maior performance em termos de AUC e menor CV durante a classificação. Como resultado, dos sete modelos de classificação, quatro mostraram AUC e CV dentro da margem de aceite, entre eles Logistic Regression, Support Vector Machine, Random Forest e Gradient Boosting com duas ou três características. Por outro lado, em termos de características, entre as cento e três características não-adaptativas avaliadas e desenvolvidas neste trabalho, variáveis como assimetria e curtose apresentaram melhor desempenho. Todavia, no segundo caso clínico, os métodos adaptativos foram mais sensitivos e estáveis na captura do efeito do tratamento ao selecionar automaticamente as regiões discriminativas associadas a cada VOI, sobretudo quando comparado ao parâmetro utilizado na rotina clínica.

Palavras-chave: RM. DWI. inteligência artificial. textura não-adaptativa. textura adaptativa.

Quantitative Analysis of Functional Medical Images Using Methods of Pattern Recognition and Artificial Intelligence

Author: Igor Fagner Vieira
Advisors: Dr. Fernando R. de A. Lima(UFPE) & Dr. Michel Koole(KUL)
Dr. José Wilson Vieira (UPE/IFPE) & Dr. Vincent Vandecaveye (KUL)

ABSTRACT

Quantitative methods for medical image analysis are limited to a non-adaptive description of volumes of interest (VOI) and tend to take into account only the discriminative performance, either of the variables of interest, or of the machine learning models used in the classification task. The objective of this work was to propose a methodology to jointly evaluate the performance and stability of advanced classification models and adaptive methods of extraction of characteristics describing texture of a given VOI, aiming to characterize, respectively, pelvic lymph nodes and treatment effect of gastrointestinal tumors. For this, DWI images were used in both cases and the performance of each model was measured in terms of the area on the ROC curve (AUC), while the stability was measured as a function of the coefficient of (CV), both estimated in bootstrap samples generated from the initial images. Non-parametric statistical tests designed to compare the performance of the different approaches, in different samples, trained and tested under the same conditions, were used to determine which models presented higher performance in terms of AUC and lower CV during classification. As a result, of the seven classification models, four showed AUC and CV within the accepted range, among them Logistic Regression, Support Vector Machine, Random Forest and Gradient Boosting either with two or three features. On the other hand, in terms of texture features, among the one hundred and three non-adaptive features evaluated and developed in this study, features such as skewness and kurtosis presented better performance. However, in the second clinical case, the adaptive methods were more sensitive and stable in capturing the treatment effect by automatically selecting the discriminative regions associated with each VOI, especially when compared to the parameter used in the clinical routine.

Keywords: MR. DWI. artificial intelligence. non-adaptative textura. adaptative textura.

SUMÁRIO

1 INTRODUÇÃO

1.1 Imagens Médicas Funcionais: Novas Tendências

Os esforços na era moderna, no que diz respeito a pesquisas na área médica, incluem criar métodos quantitativos que auxiliem o desenvolvimento de uma medicina personalizada, de tal maneira que as estratégias terapêuticas sejam melhor aplicadas e avaliadas considerando os aspectos de cada indivíduo, de cada tipo de doença. Naturalmente, esta é uma tarefa de aprendizado constante, dada a complexidade dos sistemas biológicos, cujas soluções tendem a exigir abordagens multidisciplinares. No caso do diagnóstico, isto se reflete na busca de maneiras de aumentar a precisão e exatidão, minimizando o erro associado ao fator humano, no processo de decisão final. O ideal para isto são imagens e métodos de análises destas, que forneçam o máximo de poder diagnóstico, tornando o processo de análise mais objetivo.

Nessa tarefa, as imagens funcionais são cada vez mais importantes, ante os limites das imagens anatômicas em situações em que a mudança fisiológica é muito mais rápida do que, por exemplo, alteração da forma ou do volume tumoral.

Entre as modalidades de imagens funcionais, a mais conhecida é a Tomografia por Emissão de Pósitron (*Positron Emission Tomography- PET*). A PET é utilizada para visualizar alterações fisiológicas em indivíduos vivos, de modo não-invasivo, potencialmente auxiliando o diagnóstico precoce do câncer e o acompanhamento do efeito de tratamentos complexos como o radioterápico ou quimioterápico (GREGOIRE; CHITI, 2010). A principal razão do interesse no PET pela comunidade médica se deve a sua capacidade de fornecer essas informações funcionais de modo semi-quantitativo. Um parâmetro internacionalmente utilizado é usado para isso: o valor padronizado de captação, SUV (*Standard Uptake Value*) (WEBER, 2010). Na prática clínica, a estimativa do SUV é realizada no *software* dedicado de cada empresa que comercializa o equipamento. Porém, apesar de ser uma medida mundialmente usada e difundida, diversos trabalhos apontam os limites inerentes à mesma (PAQUET et al., 2004; WEBER, 2005a; BOELLAARD, 2009). Entre os principais limites, destacam-se a dificuldade de com o SUV estabelecer limiares entre atividade benigna e maligna, sua sensibilidade ante os efeitos do volume parcial, composição corporal, etc (PAUL; JAMES, 2010). As consequências destas limitações se mostram presentes tanto no diagnóstico de determinadas doenças (PAQUET et al., 2004; WEBER, 2005a; BOELLAARD, 2009; WEBER, 2010), quanto em toda cadeia do planejamento e acompanhamento do tratamento radioterápico (THORWARTH; GEETS; PAIUSCO, 2010; SATTLER et al., 2010). Outro limite de consequência prática é sua baixa resolução espacial, requerendo seu uso integrado com imagens anatômicas, tais como a tomografia computadoriza (CT, *Computed Tomography*) ou, mais recentemente, a ressonância magnética (RM, *Magnetic Resonance*). No caso do PET/MR, a presença de campo magnético pode interagir com os pósitrons de alta energia, o que

aumenta o grau de incerteza do local exato onde ocorreu a aniquilação do pósitron, sobretudo na fronteira com tecidos moles, o que impacta a reconstrução das imagens (REBEKKA; GASPAR; SIBYLLE, 2012).

Embora fármacos marcados com materiais radioativos sejam de grande utilidade, sabe-se, por outro lado, que a água está presente em grande parte da composição do peso corporal, sob a forma de fluidos intra e extra-celulares no corpo humano. Além disso, nos tecidos biológicos, a difusão das moléculas de água segue um padrão de acordo com a estrutura e as propriedades dos tecidos. Em algumas condições patológicas, como por exemplo, acidente vascular cerebral agudo, este padrão de difusão é perturbado e a taxa de difusão muda na área afetada. Através do estudo dessas mudanças na difusão, as anormalidades podem ser detectadas.

Tal detecção pode ser realizada usando uma técnica especializada de ressonância magnética chamada difusão ponderada (*Diffusion Weighted Magenetic Imaging*, DW-MRI) ou DWI, onde a difusão das moléculas de água é explorada para visualizar a fisiologia interna. O contraste da imagem em DWI reflete a diferença na taxa de difusão entre os tecidos. A primeira quantificação das mudanças de difusão, em um estudo pioneiro, foi feita por (STEJSKAL; TANNER, 1965), em que eles fizeram o contraste da imagem depender da difusão. No entanto, foi (BIHAN; BRETON, 1985), (TAYLOR; BUSHELL, 1985) e (MERBOLDT; HANICKE; FRAHM, 1985) quem efetivamente implementaram o primeiro DWI em 1985. Posteriormente, (BIHAN et al., 1986) aplicou DWI no cérebro humano pela primeira vez em 1986.

Do ponto de vista físico, para gerar difusão ponderada em uma imagem de MRI, o sinal de leitura é feito dependente de gradientes de difusão aplicados, que podem ser adicionados a sequencias de MR convencionais usando o principio conhecido como *spin-echo*: dois gradientes de difusão de magnitude G, em fases opostas, são adicionados a uma sequência de MR, simétrica ao pulso de radio frequência de 180°. O primeiro gradiente ($G1$) introduz a mudança de fase aos prótons, dependendo de suas posições, enquanto o segundo gradiente ($G2$) inverterá as mudanças feitas pelo primeiro gradiente. Se houver movimentos de prótons, $G2$ não será capaz de desfazer completamente as mudanças induzidas por $G1$. Como resultado, haverá atenuação do sinal. Essa perda de sinal do movimento líquido das partículas fornece informação quantitativa por meio do chamado coeficiente de difusão aparente (ADC) e vem sendo usado para caracterizar o estado fisiopatológico dos tecidos (WANG et al., 2001). Avanços nas técnicas de MR permitiram que a difusão ponderada fosse usada como um protocolo de imagem de rotina para diversas doenças (TOZER et al., 2007; Koh, DM, 2010). Essas aplicações cobrem a detecção, estadiamento do câncer, assim como caracterização de tecidos malignos e benignos (JUST, 2011).

Todavia, embora todo o potencial das imagens como as aqui discutidas, na rotina clínica, procedimentos invasivos como biópisias ainda são frequentes. O objetivo com as biópisias é extrair características moleculares dos tumores usando, hoje em dia, análise genética dos tecidos. Ainda que muitas dessas análises genéticas venham sendo aplicadas com sucesso em oncologia clínica, evidentemente há limites em procedimentos invasivos como este. O principal

motivo é que os tumores são espacial e temporalmente heterogêneo, o que na prática implica que repetidas biopsias se fazem necessárias para capturar a heterogeneidade molecular dos tumores, aumentando o risco para os pacientes.

Assim sendo, esses desafios clínicos e éticos relacionados ao uso excessivo de biopsias podem potencialmente ser resolvido, na maioria dos casos clínicos, com o uso de métodos quantitativos e, quando for o caso, métodos de reconhecimento de padrões e mineração de dados, que usados em conjunto com aprendizagem de máquina oferecem outras perspectivas ao diagnóstico por imagem.

1.2 Análise Quantitativa de Imagens Médicas Funcionais

Quando o assunto é quantificação de imagens, a maioria dos estudos e procedimentos na rotina da radiologia se limitam a medir o tamanho ou volume tumoral (PALHARES et al., 2017).

Ainda que esse seja um procedimento que indica efeito de tratamento, ou mesmo caracteriza tumores em alguns casos, sabe-se que há casos em que, dado um tratamento, quer seja quimioterapia ou radioterapia, somente se observa mudanças morfométricas após alguns meses. Ou ainda, no caso de linfonodos, cujo maior diâmetro, em geral, é da ordem de milímetros, sabe-se que análises baseadas puramente na forma falham. A principal razão é que os tumores são biologicamente complexos e exibem um comportamento heterogêneo em todos os níveis.

Para quantificar esse comportamento heterogêneo, um conjunto de métodos tem sido investigado. Esses métodos cobrem uma área emergente e promissora, chamada *Radiomics*, que parte da hipótese de que a imagem médica fornece informações cruciais sobre a fisiologia do tumor e que estas podem ser melhor exploradas para melhorar o diagnóstico do câncer, assim como acompanhar o desenvolvimento de tratamentos como radioterapia e quimioterapia. Ou seja, *Radiomics* explicitamente é a reunião de um conjunto de métodos projetados em um fluxo de procedimentos para extrair um grande número de características quantitativas de imagens médicas para posterior mineração.

Entre estes métodos quantitativos de imagem estão aqueles baseados na (i) forma, tamanho ou volume, (ii) intensidade, assim como (iii) textura. Tais características oferecem informações sobre o fenótipo e o microambiente (ou habitat) do tumor que são distintos dos fornecidos por relatórios clínicos, resultados de testes laboratoriais ou mesmo ensaios genômicos (AERTS et al., 2014). Esses recursos ainda apresentam a vantagem de potencialmente serem combinados com dados de resultados clínicos e utilizados para dar suporte à decisão clínica baseando-se em evidências quantitativas.

Diante desse cenário, para um único volume de interesse, estes métodos permitem construir, por meio de abordagens diferentes, uma matriz de características ou descritores. Essa matriz de características consiste em uma matriz com dimensões Nxd, onde N é o número de volumes de interesse (VOI) e d o número de descritores que representam quantitativamente cada VOI.

Contudo, embora nem sempre explicito na literatura, esses descritores tendem a compartilhar comportamentos estatísticos similares, quer seja um descritor de primeira, segunda ou maior ordem estatística (SCHWARTZ; PEDRINI, 2012). Uma análise mais detalhada de cada um deles mostra que todos esses descritores exploram relação entre valores de intensidades de cinza entre pares ou sequências de *voxels*, armazenando o resultado sob a forma de vetores ou matrizes. Exemplos são *gray level cooccurrence matrix* (GLCM) e *gray level run length matrix* (GLRLM) (SCHWARTZ; PEDRINI, 2012). Em todos os casos, esse(a) vetor/matriz é então normalizado(a) e características são por fim estimadas a partir da distribuição de probabilidade dentro do(a) mesmo(a).

Desse modo, só indiretamente tais métodos descrevem a textura presente na imagem. Além disso, vale a pena enfatizar que todos esses descritores são inerentemente *ad hoc* [1] e, como tais, não se adaptam às imagens em estudo, atribuindo mesmo peso a todas as regiões das imagens, independente se estas são discriminativas ou não. Matematicamente falando, esses descritores se comportam como funções que expressam somas ponderadas, sendo o fator de peso a distribuição de probabilidade dos elementos que compõem a(o) matriz/vetor. Como resultado tem-se descritores *não-adaptativos* baseados quer no valor dos elementos da matriz ou em posições dentro da mesma. No último caso, a ponderação varia apenas ao longo de um eixo por vez. Além disso, essa extração de características deve ser repetida para várias configurações de alguns parâmetros livres (por exemplo, número de níveis de cinza na imagem, distância intervoxel, orientação), resultando em uma dimensionalidade relativamente alta do espaço de características.

Ou seja, uma vez o vetor $\mathbf{X} = (X_1, X_2, ..., X_d)$, *d-dimensional,* $\mathbb{R}^d$ de características tenham sido extraídas dos VOIs, o passo seguinte é a construção de modelos capazes de capturar o comportamento associado a cada classe discreta, c, com conjunto de rótulos (binários) que caracterizam os casos clínicos de interesse, como por exemplo, benigno e maligno. É disso que trata a aprendizagem de máquina no contexto deste trabalho.

Para capturar o comportamento associado a cada classe discreta, uma possibilidade é seguir como princípio aprendizagem supervisionada. Nessa abordagem, uma função, f_c, é inferida a partir de um conjunto de dados com categorias conhecidas *a priori*, o chamado conjunto treinamento ($\mathbf{X}_{train}$). Uma vez treinada, essa função (ou hipótese) f_c pode ser usada para classificar novas categorias, não observadas até então ($\mathbf{X}_{test}$):

$$y_c = f_c(\mathbf{X}_{test}, \theta_c)$$

$$y_c \in \mathbb{Z}$$

sendo θ_c o conjunto de parâmetros do modelo.

[1] solução especificamente elaborada para um problema ou fim específico e, portanto, não generalizável ou adaptável.

Todavia, é de suma importância que diferentes modelos de aprendizagem sejam comparados não somente em termos de seu poder discriminativo, mas também em termos de sua estabilidade de classificação, ante re-amostragens robustas que avaliem a generalização dos mesmos.

Nesse sentido, abordagens como essas, envolvendo métodos de reconhecimento de padrão e aprendizagem de máquina se tornam vitais e complementam as perspectivas levantadas nesse consenso (JUST, 2014), no que diz respeito a análise quantitativa do mapa ADC.

Assim sendo, este trabalho está organizado da seguinte maneira. No capítulo 2, os objetivos específicos deste trabalho são descritos. Em seguida, o capítulo 3 trata da fundamentação teórica dos métodos de imagens funcionais, assim como o estado da arte dos métodos de quantificação e análise dessas mesmas imagens, especificamente falando os métodos de extração e análise e textura, assim como dos modelos de aprendizagem de máquina. Por fim, os capítulos 4 e 5 tratam respectivamente da metodologia proposta e dos resultados obtidos, os quais incluem comparação com as variáveis usadas na prática clínica tanto para caracterização de linfonodos do cólon de útero, quanto para quantificação do efeito do tratamento de tumores gastrointestinais. Uma perspectiva dos próximos passos desse estudo é apresentada no capítulo 6.

2 OBJETIVO DESTE TRABALHO

Considerando o exposto anteriormente, há três desafios em aberto na literatura, no que diz respeito à análise quantitativa de imagens funcionais em geral, onde se inclui mapa ADC. O presente trabalho tem como meta resolver os seguintes problemas:

(a) Identificar na matriz multidimensional de características quais delas fornecem, consistentemente, maior poder discriminatório clinicamente falando, levando em conta o grau de correlação e, portanto, de similaridade entre elas. Para isso, usar-se-á aqui análise uni e multivariada, complementando os achados até então publicados envolvendo linfonodos de câncer de útero;

(b) Estimar o grau de acurácia, precisão e estabilidade (veja seções 3.2.4 e 3.2.4) de diferentes modelos de aprendizagem de máquina usando combinações das características que geralmente são analisadas univariadamente pelos médicos, de tal maneira a garantir o maior poder discriminatório possível entre linfonodos benignos e malignos;

(c) Desenvolver descritores *adaptativos* que usem informações sobre o grau de discriminação entre as classes representando os casos clínicos em estudo. Com isso, ser capaz de automaticamente destacar nos VOIs as regiões discriminativas, realizando análise intra-tumoral, tal como os médicos o fazem visualmente. Aplicar tais características na avaliação do efeito do tratamento de câncer gastrointestinal (GIST).

3 FUNDAMENTAÇÃO TEÓRICA

Esse capítulo está dividido em duas partes e nele é dada uma visão geral sobre os conceitos-chaves para o desenvolvimento do presente trabalho. Na primeira parte (item 3.1), são apresentadas em linhas gerais a técnica de imagens funcional utilizada nesse trabalho: ressonância por difusão (*Diffusion Weighted MRI*, DWI), com foco nos seus desafios no diagnóstico. Na segunda parte (item 3.2), o leitor encontra o atual estado da arte para extração de características, utilizando análise de textura. Muitas dessas características não haviam sido testadas em DW-MRI até o presente momento. Por fim, o conceito de aprendizagem de máquina é introduzido, e os modelos utilizados nesse trabalho discutidos.

Com isso, amplia-se o espectro de possibilidades quantitativas para caracterizar imagens médicas funcionais, sobretudo em situações em que a análise puramente visual é limitada.

3.1 Imagens Médicas Funcionais

Nas últimas duas décadas, a maneira de diagnosticar tem mudado na medicina. Atualmente há diferentes opções de imagens médicas, agrupadas em dois grandes grupos: métodos de aquisição que fornecem informação anatômica e informação fisiológica.

No primeiro grupo, as mais populares são aquelas que fazem uso de fontes de radiação externa para geração da imagem, tal como os tradicionais exames de Raio-X e as tomografias computadorizadas (CT). Vale salientar que, para alguns casos clínicos, tais modalidades fazem uso não somente de radiação, mas também de contrastes: substâncias que, quando injetadas, tornam visíveis determinadas regiões de interesse por alterar o coeficiente de atenuação à radiação de um dado meio.

No segundo grupo, por sua vez, o processo é dependente de uma fonte de radiação interna. Aqui destacam-se o SPECT (*Single-Photon Emission Computed Tomography*) e o PET (Positron Emission Tomography). Contudo, embora as fontes emissoras do SPECT e PET sejam diferentes, no final, a imagem é formada somente após a ingestão dos chamados radiofármacos, os quais, por serem marcados com material radioativo, passam a emitir fótons que, quando detectados, dão origem à imagem propriamente dita.

Todavia, diferentemente das modalidades anteriores que fazem uso de radiação ionizante, as imagens por ressonância magnética (MRI) são geradas a partir da propriedade física exibida por determinados núcleos como o hidrogênio: quando submetidos a um campo magnético forte e excitados por ondas de rádio (RF) em determinada frequência (Frequência de Larmor), emitem rádio-sinal, o qual pode ser captado por uma antena e finalmente transformado em imagem (ALESSANDRO, 2009).

Embora popularmente conhecida por suas imagens anatômicas com alta resolução espacial e atualmente de amplo uso, inclusive no planejamento radioterápico, a MRI também tem

o potencial de fornecer imagens com informações funcionais, por meio do processo físico chamado difusão. As imagens provenientes desse processo são conhecidas como *Diffusion Weighted MRI* (DWI), conforme abordado a seguir.

3.1.1 Diffusion Weighted MRI: visão geral

Difusão de moléculas de água no tecido

DWI explora o movimento de moléculas de água no corpo, também conhecido como movimento *Browniano*. Nos tecidos biológicos, esse movimento de água é restrito, uma vez que as interações com membranas celulares e macromoléculas criam certas barreiras ao livre movimento.

Como abordado em (GUO; CAI; CAI, 2002), o grau de restrição à difusão de água no tecido biológico é portanto inversamente correlacionado com a densidade celular do tecido e a integridade das membranas celulares.

Ou seja, o movimento de moléculas de água é mais restrito em tecidos com uma alta densidade celular com numerosas membranas celulares intactas (por exemplo, tecido tumoral), uma vez que membranas celulares lipofílicas atuam como barreiras ao movimento de moléculas de água nos espaços extra e intracelular. Por outro lado, em áreas de baixa celularidade ou onde a membrana celular foi rompida, o movimento de moléculas de água é menos restrito. Do mesmo modo, em um ambiente com células que sofreram redução de tamanho observa-se um espaço extracelular maior para a difusão de água. O mesmo se passa em membranas celulares defeituosas, onde é maior o fluxo de água do meio extracelular para o compartimento intracelular e vice-versa.

Uma vantagem aqui é que todo esse processo pode ser caracterizado e medido: (NEIL, 1997), por exemplo, a partir de medidas de difusão usando DWI-MRI, estimou o deslocamento quadrático médio das moléculas de água e encontrou algo da ordem de $8\mu m$. Em comparação, considerando que o tamanho médio das células no corpo humano é algo em torno de 10 μm, isso sugere que existe uma possibilidade de inferir fenômenos na escala celular a partir de tais imagens. Por esta razão, o DWI é considerado como um possível biomarcador para avaliar mudanças no microambiente do tumor, tanto antes quanto depois do tratamento. Para isso, o mapa ADC (*Apparent Diffusion Coefficient*) é de fundamental importância do ponto de vista da análise quantitativa.

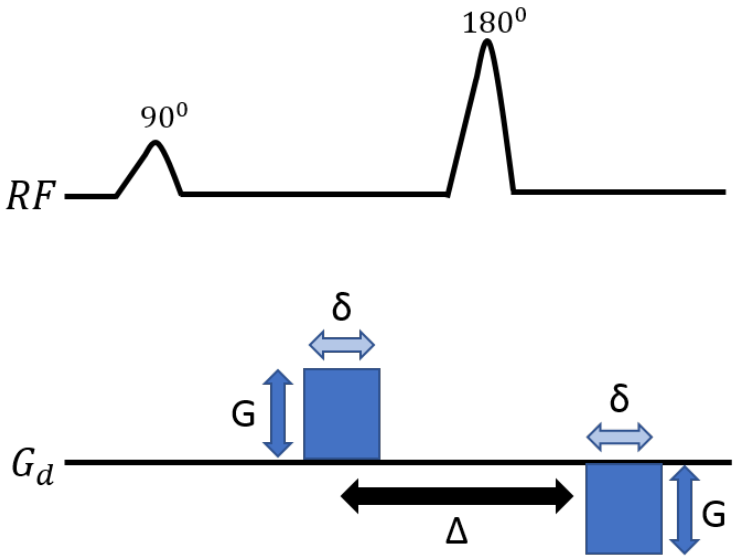
Medindo movimento de água usando DWI: mapa ADC

O principio básico que permeia a medição do grau de movimento da água é que este é proporcional ao grau de atenuação do sinal medido na imagem por difusão (DWI).

Para gerar difusão ponderada em uma imagem de MRI, o sinal de leitura irá depender dos gradientes de difusão aplicados, que podem ser adicionados a sequências de MR convencionais usando o principio conhecido como *spin-echo*: dois gradientes de difusão de magnitude G, em

fases opostas são adicionados a uma sequência de MR, simétrica ao pulso de radio frequência de 180° (Figura 1).

Figura 1 – Geração do DWI.



Adição da difusão ponderada à sequência de MRI. Dois pulsos de gradiente (denotados na linha de gradiente de difusão, G_d) de amplitude G e duração δ, iguais e em fases opostas, são aplicados em um intervalo de tempo Δ, simétrico ao pulso de radiofrequência (RF) de 180°.

Fonte: Adaptado de (Koh, DM, 2010)

O primeiro gradiente introduz a mudança de fase aos prótons, dependendo de suas posições, enquanto o segundo gradiente inverterá as mudanças feitas pelo primeiro gradiente. Se houver movimentos de prótons, o segundo gradiente não será capaz de desfazer completamente as mudanças induzidas pelo primeiro gradiente. Como resultado, haverá atenuação do sinal. Essa perda de sinal do movimento líquido das partículas segue o seguinte comportamento:

$$S(b) = S_0 e^{-bD} \tag{3.1}$$

onde $S(b)$ se refere ao sinal recebido para aquele valor particular de gradiente, que reflete o fator de sensibilização de difusão conhecido como $b-value$; S_0 é o sinal sem nenhuma difusão ponderada, equanto D é o denominado ADC (*Apparent Diffusion Coefficient*). A Equação 3.1 também é conhecida como equação de *Stejskal-Tanner*.

O fator ”b”, na Equação 3.1 é dada por:

$$b = \gamma^2.G^2.\delta^2\left(\Delta - \frac{\delta}{3}\right) \tag{3.2}$$

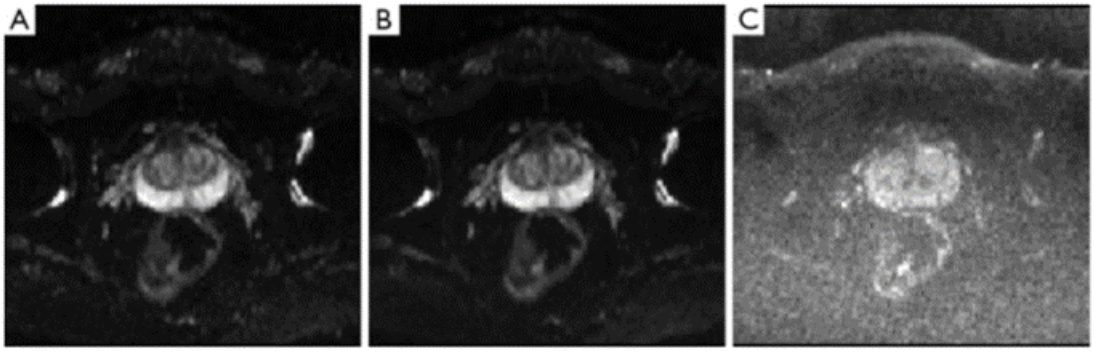
aonde γ é a constante giromagnética (42.57 MHz/T para prótons).

Das equações 3.1 e 3.2, é evidente que essa perda do sinal depende de três fatores: a amplitude do gradiente (G), a duração do gradiente aplicado (δ) e o intervalo de tempo entre os gradientes emparelhados (Δ). Todavia, variações no produto ($G\delta$) predominam sobre os

outros fatores. Por exemplo, um aumento de fator 2 em $G\delta$, equivale a um aumento de ordem 4 no $b-value$. Assim, considerando que geralmente os intervalos de tempo, ou tempo entre os gradientes são mantidos constante, a amplitude do gradiente passa a ser o parâmetro principal na geração do sinal de difusão (S), e é o que diferencia um *scanner* do outro. Através desta dependência do sinal ao movimento de partículas, DWI é capaz de gerar imagens funcionais ao explorar a propriedade de difusão de moléculas de água nos tecidos. Para descrição física detalhada de toda a mecânica de difusão em geral, e DWI em particular, aos leitores sugerimos (JONES, 2010).

Clinicamente, várias imagens de DWI podem ser obtidas alterando a magnitude dos gradientes aplicados ($b-value$), que são referenciados como imagens por difusão ponderada com um valor particular de $b-value$. Em valores $b-value$ elevados, o efeito da difusão é mais pronunciado nas imagens. Tecidos com alta difusão são vistos como regiões hipo-intensas, enquanto os tecidos com difusão restrita são vistos como regiões hiper-intensas. A Figura 2 mostra três exemplos de exames de próstata com imagens de difusão obtidas com diferente $b-values$: b0, b50 e b1000.

Figura 2 – Exemplo de imagens DWI.



DWI da prostata obtida a (A) $b-value$ = 0, (B) $b-value$ = 50 and (C) $b-value$ = 1000. Todas as imagens foram obtidas no plano transversal.

Fonte: Autor

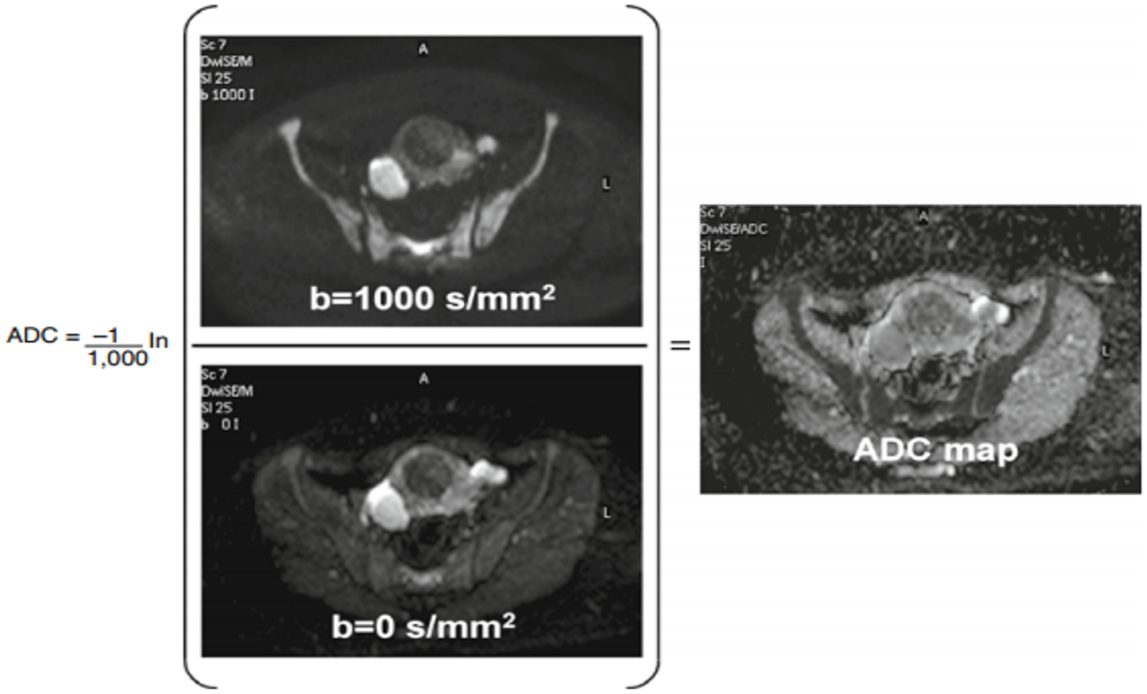
Nesse sentido, moléculas de água com um grande grau de movimento ou uma grande distância de difusão (por exemplo, dentro do espaço intravascular) mostrarão a atenuação do sinal com $b-values$ pequenos (por exemplo, b = 50-100 s/mm^2). Em contraste, valores grandes (por exemplo, b = 1.000 s/mm^2) são geralmente necessários para perceber moléculas de água de movimento lento ou pequenas distâncias de difusão, uma vez que estas mostram uma atenuação de sinal mais gradual com o aumento de b.

Uma vez que a sequência DWI é uma sequência MR modificada, todas as imagens DW têm contraste tipo T1 e T2 (imagens anatômicas), além do contraste de difusão pretendido (imagens funcionais). Portanto, às vezes, mesmo que exista um sinal hiper-intenso na imagem DW, tal sinal pode ser devido ao alto sinal de T2 (efeito de brilho T2) em vez de difusão restrita (por exemplo, no caso de acidente vascular cerebral sub-agudo). Para evitar esses efeitos, os mapas

de difusão completos chamados mapas ADC são derivados de pelo menos duas imagens de difusão (S_1 e S_2, Figura 3) usando a Equação 3.1:

$$ADC = -\frac{1}{b_2 - b_1} ln\left(\frac{S_2}{S_1}\right) \tag{3.3}$$

Figura 3 – Cálculo do mapa ADC *voxel* a *voxel* para S_2 e S_1 correspondentes a $b = 1000$ e $b = 0$, respectivamente.

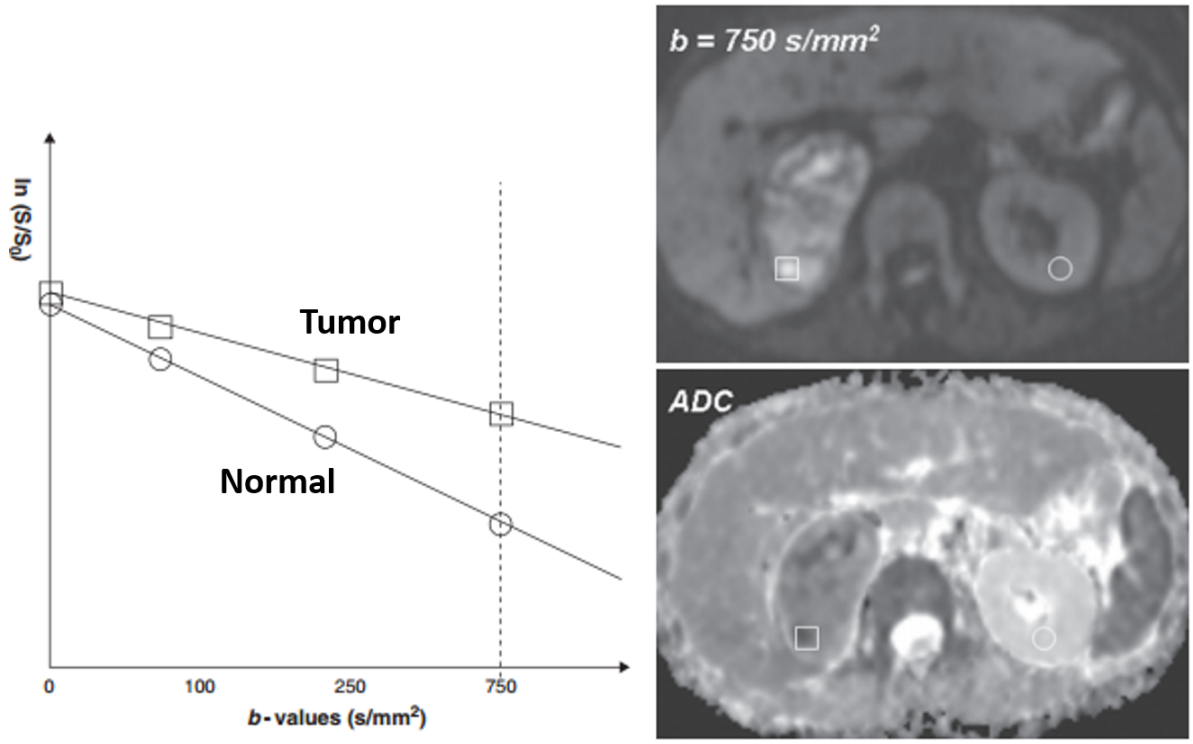


Fonte: Adaptado de (Koh, DM, 2010)

É importante notar que, enquanto as áreas de verdadeira difusão restrita de água aparecerão com intensidade de sinal elevada na imagem de $b-value$ alto, no mapa ADC, o valor do *voxel* correspondente será baixo.

A Figura 4 ilustra isso no caso de um homem de 53 anos com carcinoma renal (direito). Na imagem de difusão com b = 750 s/mm^2, a área do tumor (caixa) mostra maior intensidade de sinal em comparação com o rim normal (círculo). O gráfico mostra a relação de logaritmo das intensidades de sinal vs. $b-values$, onde S_0 é a intensidade do sinal em b = 0 s/mm^2. Note as parcelas de atenuação do sinal para o tumor (caixa) *versus* tecido renal normal (círculo). A inclinação dessas linhas representa o ADC para os tecidos individuais. Desse modo, embora a área do tumor retorna maior intensidade do sinal em b = 750 s/mm^2 em comparação com tecido renal normal (marcado pelo gráfico de linha pontilhada), a inclinação ou gradiente da linha é mais acentuada para o tecido normal do que para o tumor, o que explica o fato do tumor renal ter um baixo ADC em comparação com o tecido renal normal. Esta diferença pode ser facilmente visualizada na imagem do mapa ADC.

Figura 4 – Diagrama mostrando a relação entre $b-value$ e ADC.



Fonte: Adaptado de (Koh, DM, 2010)

Assim sendo, no paradigma padrão de diagnóstico, puramente baseado na análise visual, quando o médico investiga imagens de alto $b-value$, os olhos estão treinados a focar em regiões de alta intensidade, contra um fundo de baixa intensidade. Porém, uma série de estruturas anatômicas normais também podem mostrar restrição de difusão de água: cérebro, glândulas salivares, linfonodos normais, baço, útero, ovários, testículos, a medula espinhal e os nervos periféricos se mostram relativamente brilhantes em DW-MRI nas imagens de alto $b-value$ (TAKAHARA et al., 2004). A parede normal do intestino também demonstra graus variados de sinal de alta intensidade.

Considerando o exposto, um importante ponto a ser destacado é que na rotina os médicos são instruídos a usar imagens DW-MRI e mapas ADC em conjunto com informações provenientes extras de procedimentos clínicos e imagens MR morfológicas convencionais (T2), visando diagnóstico ótimo.

O custo disso na prática é que o médico necessita analisar ainda mais variáveis ao mesmo tempo. Isto torna todo o processo um tanto quanto exaustivo e aumenta a probabilidade de interpretações subjetivas ou mesmo procedimentos invasivos como biopsia.

Para lidar com esses desafios, uma alternativa possível é o uso de métodos quantitativos, a fim de explorar todo o potencial do mapa ADC na tarefa do diagnóstico e acompanhamento do tratamento.

3.2 Análise Quantitativa

Uma vez que a região de interesse tenha sido definida [1], a análise quantitativa da mesma consiste em dois principais passos: estimativa das características fundamentais que potencialmente descreve o estado em análise, aqui estaremos tratando do conceito de textura; seguido por treinamento e classificação dessas mesmas regiões de interesse, usando métodos de aprendizagem de máquina.

Assim sendo, esta seção trata desses dois passos com algum detalhe, conforme descrito a seguir.

3.2.1 Análise de Textura

Toda e qualquer imagem, como a de um tumor por exemplo, reflete, em linhas gerais, uma representação da realidade do sistema em análise. Em outras palavras, a imagem pode ser vista como um campo de informações provenientes de um conjunto de experimentos aleatórios. Assim sendo, pensando no diagnóstico ou classificação mais preciso, por meio de atividades de alto nível como reconhecimento de padrões, é imprescindível um processo de extração de características que forneça o máximo poder discriminativo, quantitativamente falando.

Contudo, uma imagem médica pode ser caracterizada de diferentes formas. A forma clássica é a partir de análise visual de regiões de interesse, o que é um processo subjetivo e nem sempre o mais eficiente em alguns casos clínicos (TAKAHARA et al., 2004). Nesse sentido, uma alternativa mais objetiva é análise quantitativa de características que medem parâmetros outros associados à forma ou à textura dessas mesmas regiões de interesse, tal como sumarizado na Figura 5.

Enquanto a análise da forma, explorando aspectos geométricos variados, parece ser um caminho natural, até intuitivo, há casos clínicos em que a mudança fisiológica é mais rápida que alterações morfométricas. Ou seja, para uma análise complementar, faz-se necessário o uso de métodos que levem em conta mais do que a geometria do sistema. Uma alternativa é seguir a lógica do sistema visual humano e explorar a possibilidade de análise de textura. A principal vantagem dessa estratégia é que textura está entre as características que contém informações tanto sobre a distribuição espacial e variação de intensidade, quanto acerca do arranjo estrutural de superfícies e relações entre regiões vizinhas (SCHWARTZ; PEDRINI, 2012).

Nesse terreno, o desafio é formalizar matematicamente a ideia de textura e, até o presente momento, não há na literatura conceito único. Na sequencia, todavia, o conceito proposto por Haralick, 1979, será o abordado ao tratar de características de segunda ordem, ou seja, aquelas em que a relação espacial entre *voxels* são levadas em conta. Sob este ponto de vista, a textura é definida como uma medida da interação entre as primitivas tonais que compõem a imagem, ocorrendo estas primitivas de diferentes formas. *Voxels ou pixels* contíguos com propriedades

[1] Não se abordará aqui métodos de segmentação de imagens para obtenção dos VOIs, uma vez que, nesse trabalho, as delineações dos mesmos foram realizadas pelos radiologistas usando as ferramentas de rotina do hospital UZ Leuven, Bélgica.

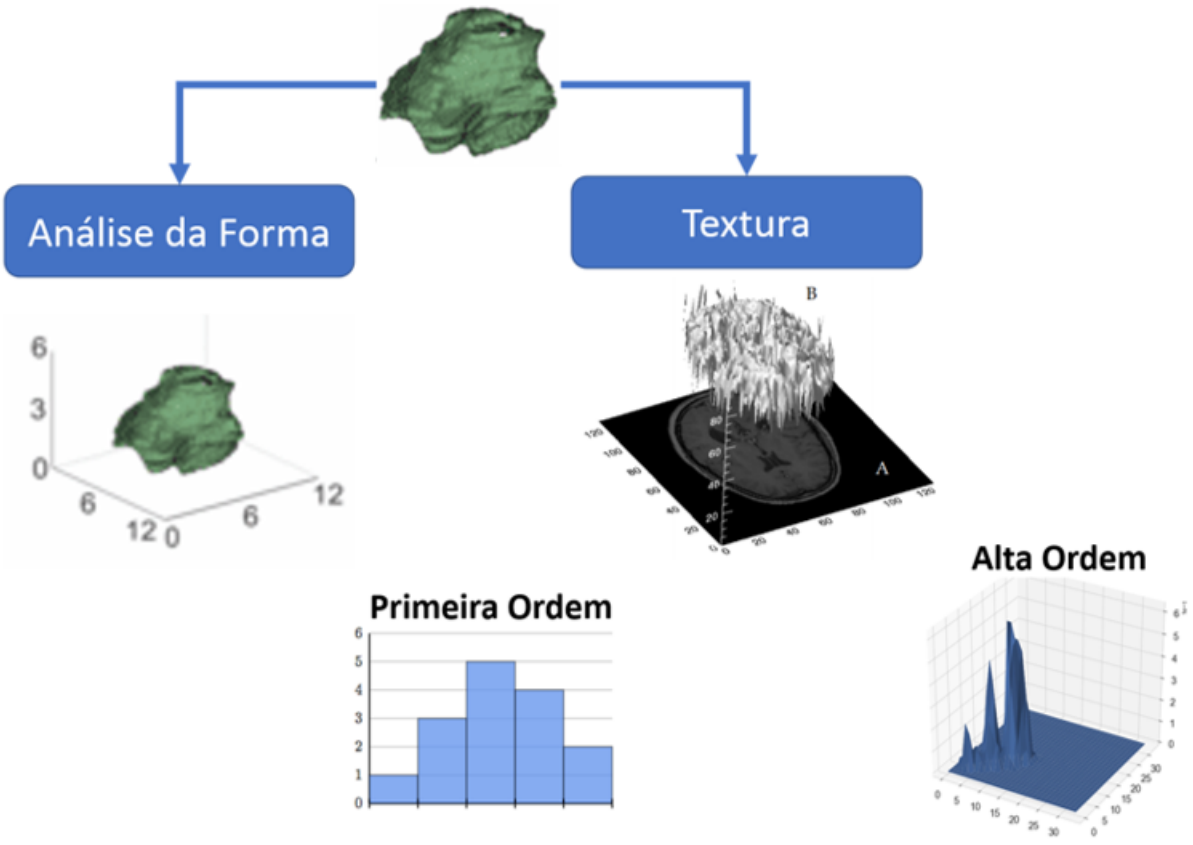
semelhantes formam estas primitivas, dentre as quais podem ocorrer interações aleatórias ou com algum grau de determinismo (HARALICK, 1973).

Os principais métodos para extração de características são aqueles baseados em: (i) descrição geométrica, (ii) métodos estatísticos, (iii) processamento de sinais e (iv) modelos paramétricos.

Até o momento na literatura, a maioria dos trabalhos recentes sobre extração de características, restringe-se, predominantemente, à aplicação dos métodos de análise estatística e, em menor número, métodos de processamento de sinais (SULLIVAN; ROY; EARY, 2003; NAQA et al., 2009; TIXIER et al., 2011; TAN et al., 2012; YANG et al., 2013; ORLHAC et al., 2014). Em DW-MRI, os métodos estatísticos descritos em (ii) são de longe predominantemente utilizados (Koh, DM, 2010).

A seguir, a atenção estará voltada para os métodos (i),(ii) e (iii).

Figura 5 – Tipos de características para quantificação de imagens.



Fonte: autor

3.2.1.1 Descrição Geométrica: análise da forma

Esse grupo de características descrevem forma e tamanho do volume de interesse e, portanto, não levam em conta, *por sí*, a distribuição de intensidade de níveis de cinza presente no VOI.

Seja o volume V, definido como a soma de cada voxel, V_i, presente no VOI,

$$V = \sum_{i=1}^{N} V_i \tag{3.4}$$

Para cada volume, existe uma área superficial, A, dada por:

$$A = \sum_{i=1}^{N} \frac{1}{2} |a_i b_i x a_i c_i| \tag{3.5}$$

onde N é o número de triângulos formando a superfície *mesh* do VOI, enquanto $a_i b_i$ e $a_i c_i$ são as faces do i^{th} triângulo formado pelos pontos a_i,b_i e c_i (LORENSEN; CLINE, 1987).

Baseado nesses dois conceitos geométricos, três outras características são consideradas no presente estudo: razão área superficial volume (Λ), esfericidade (Esf) e máximo diâmetro (D_{max}).

Na razão área superficial volume (Λ) (Eq. 3.5), valores baixos indicam formas mais compactas (exemplo uma esfera).

$$\Lambda = \frac{A}{V} \tag{3.6}$$

Contudo, para medir o quão próximo uma dada região tumoral está de uma esfera, a medida adimensional, chama esfericidade (Esf), é utilizada. Tal medida além de ser independente de escala e orientação, está definida no intervalo $]0,1]$, onde 1 indica uma esfera perfeita. A esfericidade é definida como:

$$Esf = \frac{\sqrt[3]{36\pi V^2}}{A} \tag{3.7}$$

Por fim, uma outra variável é o máximo diâmetro do VOI, D_{max}, normalmente utilizada pelo médicos apenas em um corte da imagem. Para tal, primeiramente cada par de *voxel*, i e j, presente na superfície tumoral, é identificado. Em seguida, calcula-se a distância euclideana $D_{i,j}$ entre cada par de *voxel* da superfície e o diâmetro máximo é definido como D_{max},

$$D_{max} = \max_{\mathbf{D}} \left[D_{i,j} \right] \tag{3.8}$$

Uma série de outros métodos derivados destes parâmetros fundamentais podem ser calculados, conforme mostrado no capitulo de metodologia. Todavia, quando a informação fisiológica é uma variável importante, outras estratégias se mostram fundamentais.

3.2.1.2 Métodos Estatísticos

Estes métodos visam representar a textura indiretamente assumindo que as imagens são geradas a partir de fenômenos aleatórios. Para isso, são definidos distribuições e relacionamentos entre os níveis de cinza dos *pixels* [2] em uma imagem.

Medidas Baseadas na Distribuição de Níveis de Cinza

Em geral, as medidas baseadas na distribuição consideram a intensidade de cada *pixel* de maneira isolada, tornando-as sensíveis a variações em um único tom de cinza.

Seja x_i a variável aleatória denotando intensidade, e $p(x_i)$, o correspondente histograma normalizado da imagem (ou de uma região da mesma), para os níveis de cinza $i = 0, 1, 2, ..., L-1$, onde L o número de diferentes tons de cinza (GONZALEZ; WOODS, 2007).

Dentre as possíveis medidas para descrever textura estão as de tendência central (primeiro momento) e dispersão (segundo, terceiro e quarto momentos, por exemplo). O primeiro momento, mostrado na equação 3.9, como o próprio nome diz, fornece o primeiro valor esperado da distribuição dos níveis de cinza presentes na textura; enquanto o segundo, fornece o quanto as intensidades com L níveis estão dispersos em torno da média, em outras palavras, permite avaliar a rugosidade ou não de determinada região.

$$\mu = < x > = \sum_{i=0}^{L-1} x_i p(x_i) \tag{3.9}$$

$$\sigma(x)^2 = \sum_{i=0}^{L-1} (x_i - \mu)^2 . p(x_i) \tag{3.10}$$

Outro descritor possível é o que fornece a *assimetria*, ou seja, a concentração de valores em relação à mediana (Equação 3.11), fornecendo assim o grau intensidade claro/escuro comparada à média. A curtose, por sua vez, definida na Equação 3.12, indica o achatamento da função de distribuição como mostra a Figura 6, e mede o quão uniforme é a distribuição do nível de cinza da região.

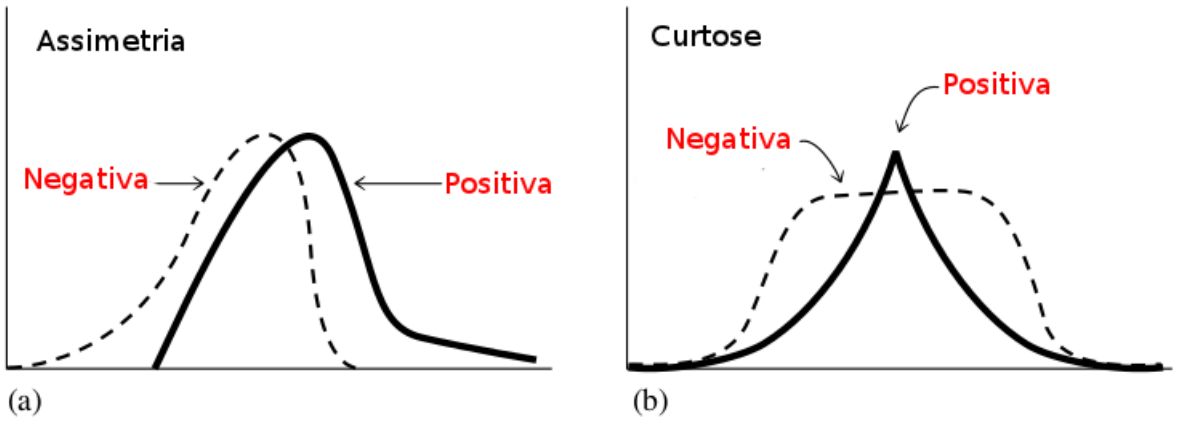
$$s = \sum_{i=0}^{L-1} (x_i - \mu)^3 . p(x_i) \tag{3.11}$$

$$k = \sum_{i=0}^{L-1} (x_i - \mu)^4 . p(x_i) \tag{3.12}$$

A Figura 7 mostra uma avaliação quantitativa de ADC para medir a resposta ao tratamento. Neste exemplo, um homem de 65 anos com doença óssea metastática do câncer de próstata. Mapas ADC da pelve antes e 1 mês após tratamento. Uma região de interesse (ROI) foi desenhada em torno do local da doença no ílio esquerdo (contorno verde). A avaliação visual mostra aumento no ADC ($x10^{-3} mm^2/s$) dentro da área da doença após o tratamento. O

[2] Embora a título didático os exemplos aqui mostrem matrizes 2D, como o objeto de estudo deste trabalho são imagens médicas, todos os métodos utlizados foram implementados para lidar com matrizes 3D.

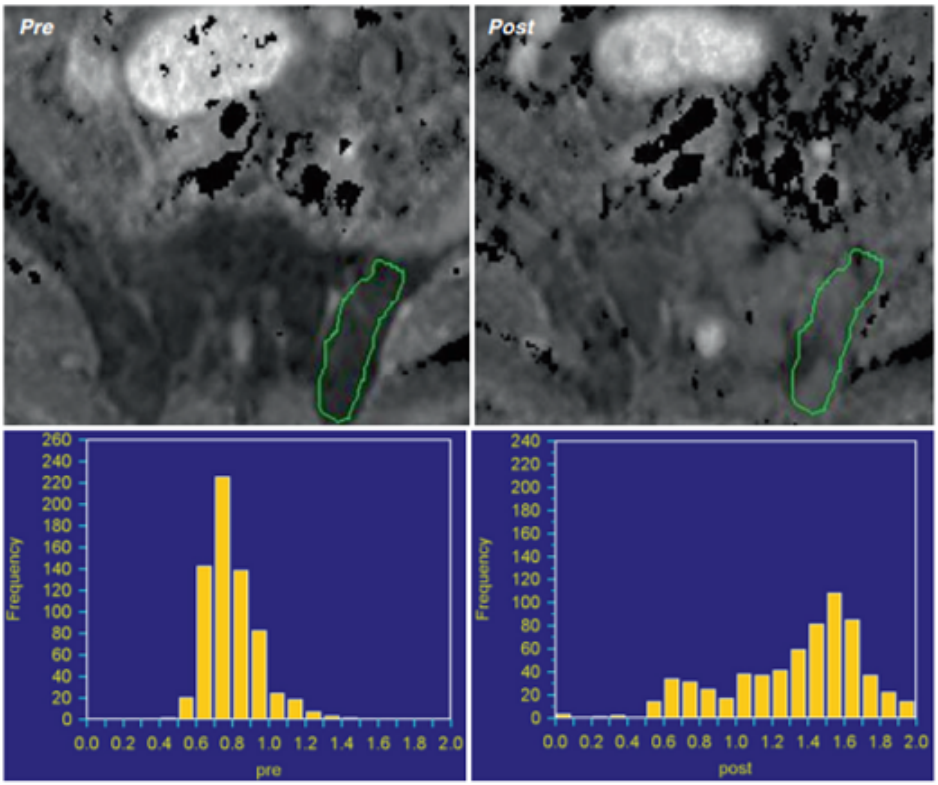
Figura 6 – (a) Assimetria e (b) Curtose para os casos de distribuições que diferem significativamente de uma distribuição normal.



Fonte: autor

histograma de frequência dos valores de ADC no ROI confirma um aumento no valor médio após o tratamento, com uma mudança da distribuição de ADC para a direita.

Figura 7 – Diferença na (a) Assimetria e (b) Curtose antes (Pre) e após (Post) o tratamento.



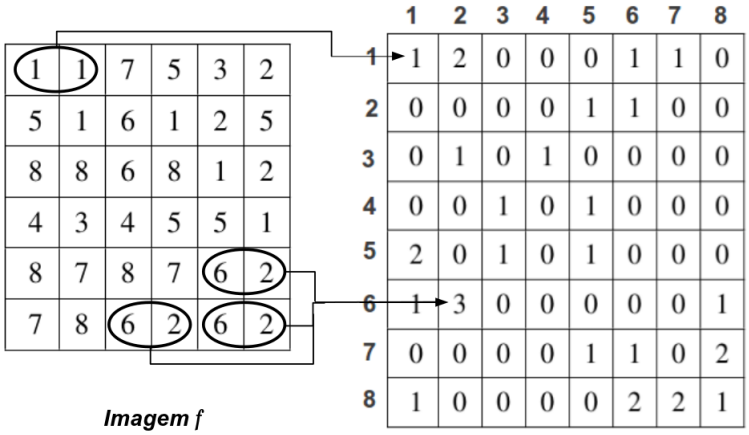
Fonte: (Koh, DM, 2010)

Matriz de Co-ocorrência

Para adquirir informações sobre transições de níveis de cinza entre dois *voxels*, utiliza-se a chamada matriz de co-ocorrência.

Para descrever o relacionamento espacial entre os *voxels* que compõem a textura, define-se um operador Q [3]. A matriz de co-ocorrência é formada por elementos que descrevem a frequência com que ocorrem as transições de níveis de cinza entre pares de *voxels*. A Figura 8 mostra um exemplo de como obter a matriz de co-ocorrência para uma imagem 2D, cujas dimensões desta matriz são proporcionais à quantidade de níveis de cinza, independentemente do tamanho da imagem. Neste exemplo, a imagem tem 8 níveis de cinza e o operador Q está definido como um *pixel* imediatamente à direita, com o par $(x_i, x_j) = (6, 2)$ ocorrendo $g_{i,j} = 3$ vezes.

Figura 8 – Construção de uma matriz de co-ocorrência usando como regra a definição de um pixel imediatamente a direita.



Para facilitar entendimento, essa figura mostra como obter GLCM para imagem 2D. Neste trabalho, esse algorítimo foi estendido para imagens 3D.

Fonte: Adaptado: (GONZALEZ; WOODS, 2007)

Assim sendo, pode-se definir a probabilidade do par de pontos (x_i, x_j), que satisfaz a regra contida no operador Q, como:

$$p_{ij} = \frac{g_{ij}}{\sum_{i,j}^{K} g_{ij}} \tag{3.13}$$

onde g_{ij} é o número de vezes que o par de *pixel* (x_i, x_j), com o número total de K pares, ocorrem na imagem, dada as condições especificadas por Q.

A partir desse histograma bidimensional, diversas características podem ser extraídas tal como descrito originalmente em (HARALICK, 1973) e aplicado em imagens médicas nos trabalhos de (SULLIVAN; ROY; EARY, 2003; NAQA et al., 2009; TIXIER et al., 2011; TAN et al., 2012; YANG et al., 2013; ORLHAC et al., 2014).

[3] Este operador define a posição dos dois *voxels* com relação aos demais da vizinhança, ou seja, determina quais *voxels* e quais transições de nível de cinza serão considerados.

Embora este método seja bastante utilizado em análise de textura, ele apresenta alguns limites, tais como: a seleção da distância *d* entre os *voxels*; (2) a não captura dos aspectos da forma das primitivas das imagens e (3) a geração de características altamente correlacionadas, conforme mostrado no segundo caso clínico desse trabalho (vide Tabela 5).

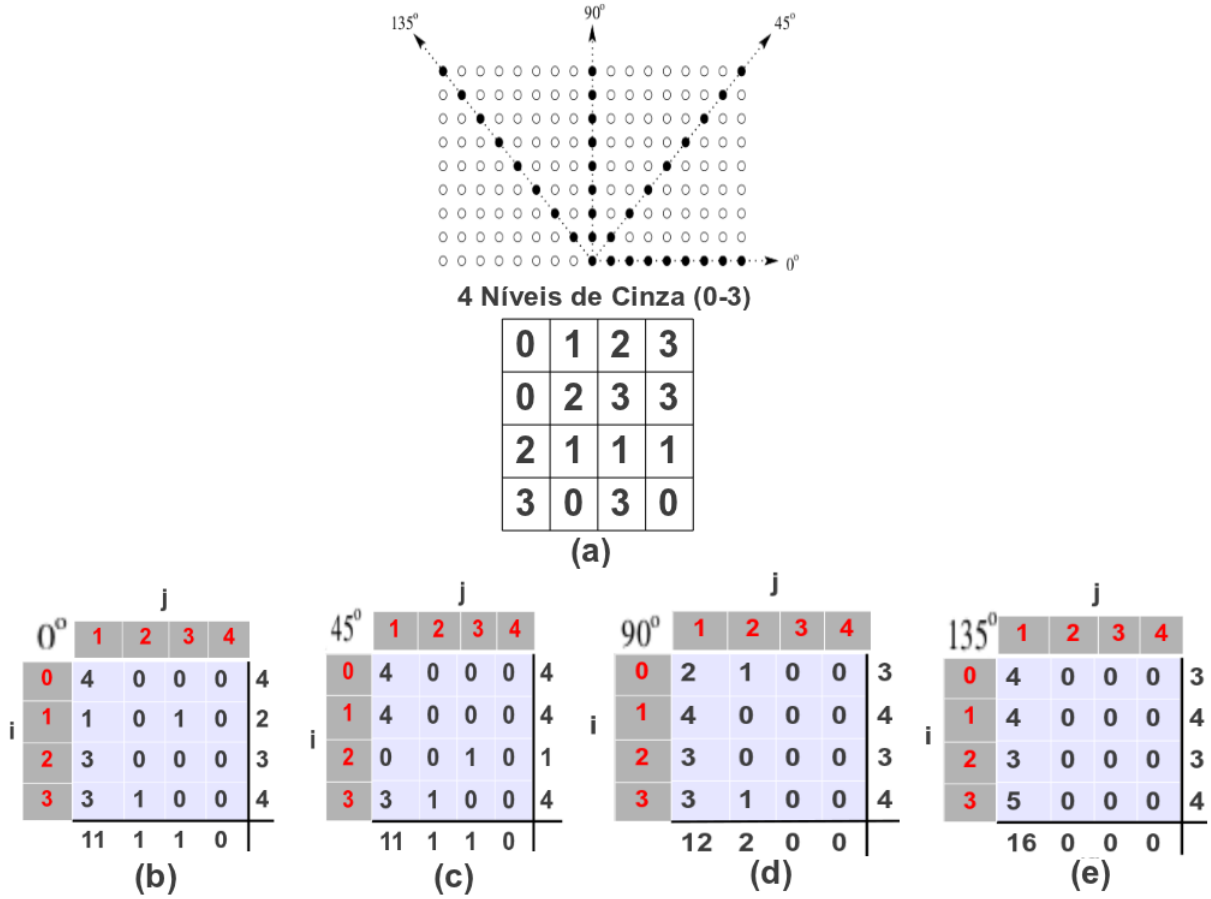
Matriz dos comprimentos de agrupamentos de pixels com mesmo nível de cinza

Uma outra abordagem possível, a chamada matriz dos comprimentos de agrupamentos de *pixels* com mesmo nível de cinza, do inglês, *Gray Level Run-Length Matrices (GLRLM)*, apresenta elementos que contêm o número de comprimentos de agrupamento, para um determinado nível de cinza. A partir dessas matrizes podem ser obtidas informações relevantes sobre as características da textura que estão sendo analisadas (SCHWARTZ; PEDRINI, 2012). Para construir a matriz GLRLM, *P*, primeiramente obtém-se o conjunto ordenado dos *i-ésimos* níveis de cinza dos *pixels* presentes na imagem; na sequência, contam-se os *j-ésimos* comprimentos de agrupamento do *i-ésimo* nível de cinza. A contagem é realizada para todos os níveis de cinza e obedece à seguinte regra: os *pixels* contabilizados devem ser colineares em uma direção pré-estabelecida. Cada elemento da matriz, simbolizado por $P(i, j|\theta)$, representa, para o pixel com o *i-ésimo* nível de cinza, a frequência do agrupamento com tamanho j. O parâmetro θ determina a direção de contagem. A dimensão da matriz *P*, por sua vez, é dada por $N_g x N_r$, onde N_g e N_r representam o número de níveis de cinza na imagem e o tamanho do agrupamento mais longo, respectivamente.

A Figura 9 mostra um exemplo de construção da GLRLM para a imagem mostrada em (a), considerando corridas nas diferentes direções (0^o, 45^o, 90^o e 135^o). As matrizes resultantes são mostradas em (b),(c),(d) e (e). A soma dos elementos de cada linha, na última coluna separada, representa o número de agrupamentos de um dado nível de cinza; enquanto a soma dos elementos de cada coluna, na última linha separada, representa o número de agrupamentos com um tamanho específico.

A Tabela 1 mostra as características que podem ser extraídas a partir das matrizes GLRL, tal como proposto por *Galloway, 1975*: ênfase nos comprimentos de agrupamento curto (SRE) e longo (LRE). O SRE tende a ser grande em agrupamentos curtos (texturas finas), enquanto o LRE nos longos (texturas grossas). Quando as texturas apresentam valores similares com relação as características SRE e LRE, mas diferem na distribuição dos níveis de cinza presente nas corridas, dois outros descritores podem ser estimados: o LGRE (*Low Gray Level Runs Emphasis*) e o HGRE (*High Gray Level Runs Emphasis*), ambos usados para enfatizar o tamanho dos agrupamentos de níveis de cinza alto e baixos, respectivamente. Para estimar a não-uniformidade do tom de cinza e a não-uniformidade do tamanho do agrupamento, utiliza-se, respectivamente, as medidas GLNU (*Gray Level Non-uniformity*) e RLNU (*Run Length Non-uniformity*). (GALLOWAY, 1975).

Figura 9 – Exemplo de construção da GLRL para uma imagem com 4 níveis de cinza, avaliada em todas as direções.



Adaptado: (SCHWARTZ; PEDRINI, 2012)

Tabela 1 – Descritores de textura que podem ser extraídos a partir das matrizes de GLRL com número total de agrupamentos n_c.

Descritor	Expressão Matemática	Característica
SRE	$SRE = \frac{1}{n_c}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{j^2}$	Valor grande em texturas finas.
LRE	$LRE = \frac{1}{n_c}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} j^2 P(i,j\|\theta)$	Valor grande em texturas grossas.
LGRE	$LGRE = \frac{1}{n_c}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{P(i,j\|\theta)}{i^2}$	Valor grande em imagens com baixo nível de cinza.
HGRE	$HGRE = \frac{1}{n_c}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i^2 P(i,j\|\theta)$	Valor grande em imagens com alto nível de cinza.
GLNU	$GLNU = \frac{1}{n_c}\sum_{i=1}^{N_g}\left[\sum_{j=1}^{N_r} P(i,j\|\theta)\right]^2$	Mede a similaridade dos níveis de cinza. Será grande para níveis de cinza diferentes.
RLNU	$RLNU = \frac{1}{n_c}\sum_{i=1}^{N_r}\left[\sum_{j=1}^{N_g} P(i,j\|\theta)\right]^2$	Mede a similaridade do comprimento de agrupamentos. Será pequeno para comprimentos uniformes.

Matriz da Diferença dos Tons de Cinza da Vizinhança

Um outro método potencialmente útil para analisar características de imagens funcionais é aquele proveniente da matriz da diferença dos tons de cinza da vizinhança - NGTDM (*Neighborhood Gray Tone Difference Matrix*). Esta matriz permite avaliar características tais como aspereza, contraste, etc. Tal matriz será unidimensional, sendo composta por N_g elementos, correspondentes ao número de tons de cinza distintos presentes na imagem (AMADASUN; KING, 1989). Para calcular cada elemento da matriz NGTDM, obtém-se a média conforme definida na equação 3.14, representando o parâmetro d o tamanho da vizinhança. Esta média, $\bar{A}(x,y)$, é obtida considerando todos os *pixels* com intensidade $f(x,y)$ na imagem, exceto aqueles situados em regiões de borda, e consiste no módulo da diferença de cada *pixel*, localizado na coordenada (x,y), com relação à soma dos *pixels* de sua respectiva vizinhança efetiva.

$$\bar{A}(x,y) = \left[\frac{1}{(2d+1)^2 - 1} \left| f(x,y) - \sum_{m=-d}^{d} \sum_{n=-d}^{d} f(x+m, y+n) \right| \right] \tag{3.14}$$

Uma vez conhecendo o valor de d, estabelece-se na imagem a região de interesse e obtém-se então a lista de tons de cinza presentes nesta região, bem como suas frequências (N_i). O tamanho desta lista estabelece a dimensão da matriz NGTDM. O *i-ésimo* tom de cinza e $\bar{A}(x,y)$ são usados para calcular o *j-ésimo* elemento de S, $s_j(i)$, dado por:

$$s_j(i) = \begin{cases} \sum_{k=1}^{N_i} |(i - \bar{A}_k(i)| & se \quad N_i > 0 \\ 0 & se \quad N_i = 0 \end{cases} \tag{3.15}$$

No exemplo da Figura 10, há uma imagem com 4 níveis de cinza diferentes, $i = \{0,1,2,3\}$, e região de vizinhança para $d = 1$. O terceiro elemento, $j = 2$, da matriz S, para o tom de cinza $i = 2$, é obtido determinando-se a frequência $N_{i=2} = 3$, e suas respectivas médias $\bar{A}(x,y)$. Por fim, utiliza-se a equação 3.15. Os demais elementos são obtidos do mesmo modo.

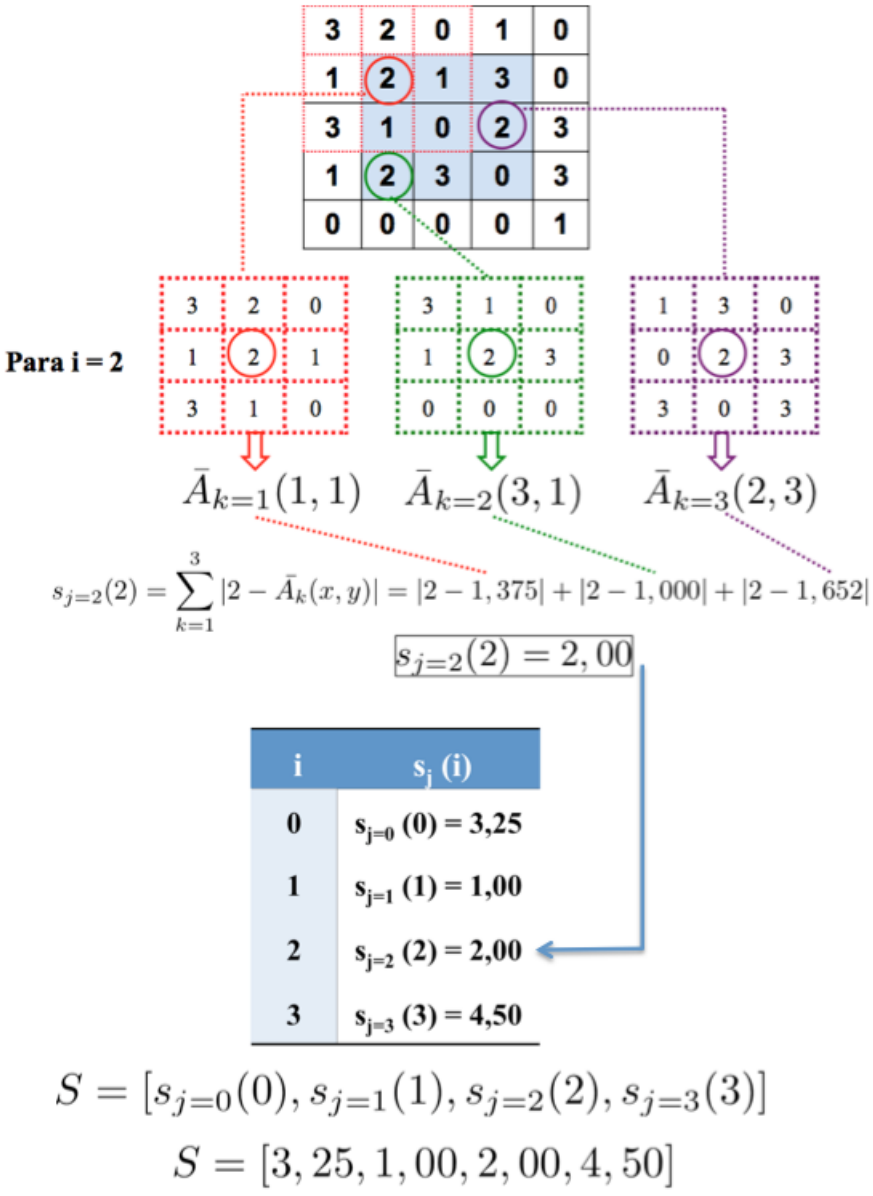
Os elementos da matriz S podem ser usados para caracterizar: aspereza (f_{asp}), contraste (f_{con}), fineza (f_{fin}), complexidade (f_{com}) e força da textura (f_{str}) (SCHWARTZ; PEDRINI, 2012).

A primeira delas, a aspereza, é definida por:

$$f_{asp} = \left[\varepsilon + \sum_{i=0}^{L_{max}} P(i) s(i) \right]^{-1} \tag{3.16}$$

onde ε é um número pequeno, $\approx 10^{-7}$, para evitar valor nulo no denominador, L_{max} o valor do maior tom de cinza e $P(i)$ a probabilidade de ocorrência do *i-ésimo* tom de cinza, que depende de N_i. Dessa maneira, a textura é mais áspera, quando o somatório da equação 3.16 for pequeno. Em outras palavras, quando a vizinhança efetiva do pixel de intensidade i tiver aproximadamente a mesma intensidade.

Figura 10 – Construção da matriz S para uma imagem com 4 níveis de cinza $i = \{0,1,2,3\}$ para a região de vizinhança $d = 1$.



Fonte: Adaptado de (SCHWARTZ; PEDRINI, 2012)

A segunda medida de de interesse é o contraste. Admite-se alto contraste quando regiões com níveis de cinza distintos podem ser claramente perceptíveis. Em outras palavras, quando existe grande diferença de intensidade entre regiões vizinhas. Matematicamente é expressa por:

$$f_{con} = \left[\frac{1}{N_g(N_g-1)} \sum_{i=0}^{L_{max}} \sum_{j=0}^{L_{max}} P(i)P(j)(i-j)^2 \right] \left[\frac{1}{n^2} \sum_{i=0}^{L_{max}} s(i) \right] \tag{3.17}$$

onde $n = N - 2d$, sendo N um fator da dimensão, NxN, da região de interesse a ser analisada na imagem.

Um terceiro descritor a ser explorado é a fineza, que é dada por:

$$f_{fin} = \left[\sum_{i=0}^{L_{max}} P(i)s(i)\right] . \left[\sum_{i=0}^{L_{max}}\sum_{j=0}^{L_{max}} |iP(i) - jP(j)|\right]^{-1} \tag{3.18}$$

Diferente da propriedade de aspereza, a fineza utiliza informação sobre a frequência espacial e sobre a variação de intensidade na imagem. Ela indica a existência de variações acentuadas dos níveis de cinza, quando considerada uma região relativamente pequena da textura (YU, 2010; SCHWARTZ; PEDRINI, 2012).

Por fim, numa tentativa de medir o conteúdo visual da informação de textura presente em uma imagem, a seguinte definição da chamada complexidade é dada por:

$$f_{Comp} = \frac{1}{N_p^2}\sum_{i=1}^{L_{max}}\sum_{j=1}^{L_{max}} |i-j|\frac{P(i)s(i)+P(j)s(j)}{P(i)+P(j)}, \text{ onde } P(i) \neq 0, P(j) \neq 0 \tag{3.19}$$

Esses descritores, quando usados em conjunto, formam uma espaço vetorial multidimensional e servem como entrada para construção dos modelos de classificação conforme abordado a seguir.

3.2.2 Aprendizado de Máquina

Aprendizado de máquina é parte da ampla área de inteligência artificial e tem como base a construção de modelos capazes de aprender a partir de exemplos (aprendizado supervisionado), diretamente a partir da experiência (não-supervisionado) ou ambos (semi-supervisionado). Em linhas gerais, esses algorítimos aprendem a construir fronteiras de decisões inteligentes a partir da possibilidade de reconhecimento de padrões complexos com aplicações que vão desde reconhecimento de escrita, *marketing*, até o diagnóstico médico, sendo esta última a área de interesse do presente trabalho. Além disso, com relação a diagnostico médico, o interesse principal aqui é classificação de imagens a partir de uma dado conjunto de características.

Assim sendo, esse capítulo fornece uma visão geral dos modelos de aprendizado de máquina mais populares para classificação de imagens.

3.2.3 Modelos para classificação de imagens

Uma imagem médica pode ser representada por um vetor, $\mathbf{x} = (x_1, x_2, ..., x_D)$, *d-dimensional* de características extraídas de volumes de interesse (VOI) da mesma. O objetivo geral na construção de um modelo de classificação é atribuir esse vetor de características a uma das K classes discretas C_k. Para os propósitos desse trabalho, $K = 2$ e as classes são consideradas disjuntas, de tal modo que, cada vetor de características pertence a uma, e somente uma das duas classes [4]. Dessa forma, o modelo de classificação é uma função $f(\mathbf{x})$ que retorna um valor que indica a classe a qual pertence o vetor de característica $\mathbf{x}$. Os parâmetros da função $f(\mathbf{x})$ são

[4] No contexto desse trabalho, classes significam estado clínico dos VOIs analisados. Por exemplo, C_1 :tumor e C_2 : tecido sadio.

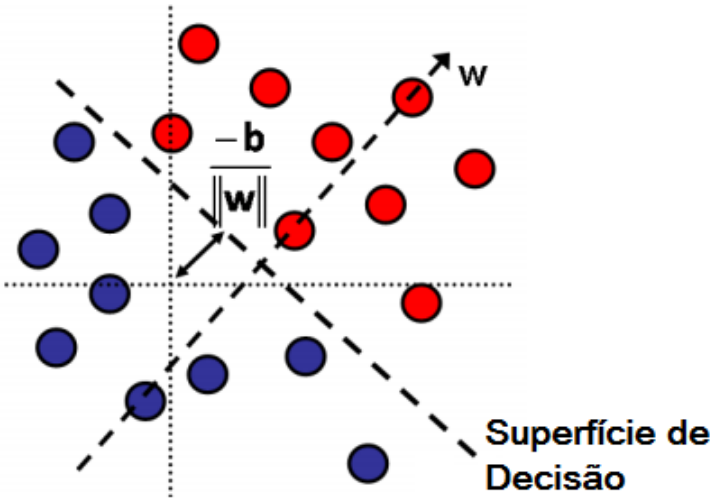
otimizados durante a fase de treinamento. Nessa fase, são fornecidos ao modelo um conjunto de N exemplos, para os quais o estado clínico é conhecido [5], $\{(\mathbf{x_i}, t_i)\,|\mathbf{x_i} \in \mathbb{R}^D, t_i \in \{-1,1\}\}_{i=1}^{N}$, e o modelo constrói um campo de hipótese sobre como as classes se distribuem naquele espaço vetorial de características. Seguida a fase de treinamento, a performance da classificação é avaliada usando novos pacientes, diferentes daqueles utilizados durante o treinamento, a fim de se estimar a capacidade do modelo de generalizar para novos casos futuros.

Todavia, há uma ampla variedade de modelos possíveis em tarefas como as brevemente descritas anteriormente. Assim sendo, os modelos relevantes para o presente trabalho serão apresentados a seguir incluindo o tratamento matemático inerente a cada um deles.

Linear Discriminant Analysis- LDA

LDA tem como objetivo determinar a combinação linear de características que resulte na máxima separação entre as médias das classes com relação à soma da variância intra-classes. Como consequência, essa transformação no espaço de características pode ser expressa como $y(\mathbf{x}) = \mathbf{w^T x} - b$, sendo a superfície de decisão $y(\mathbf{x}) = 0$ um hiperplano com $(K-1)$-dimensões como ilustrado na Figura 11. Desse modo, a seguinte regra de decisão é definida: um vetor $\mathbf{x}$ pertence a classe C_1 se $y(\mathbf{x}) \geq 0$, e a classe C_2 se $y(\mathbf{x}) < 0$. O valor de $y(\mathbf{x})$ fornece uma medida (com sinal) da distância entre cada ponto e a superfície de decisão, o que pode ser interpretado como uma medida de certeza de que um dado vetor $\mathbf{x}$ pertence à classe C_k.

Figura 11 – Ilustração 2-D da superfície de decisão obtida usando LDA.



Vetores de características associados a classe C_1 são mostrados em vermelho, e aqueles associados a C_2 estão em azul. A superfície de decisão é definida pela sua ortogonalidade com o vetor $\mathbf{w}$ (peso das características) e sua distância da origem, o que depende do limiar b.

Fonte: autor

[5] Nesse estudo foram utilizadas imagens de pacientes que foram submetidos a biopsia. O resultado da mesma foi utilizado para definir os grupos de pacientes a cada uma das classes C_1 e C_2.

Esse critério de maximização pode ser escrito como:

$$\max_{\mathbf{w}} J(\mathbf{w}) = \max_{\mathbf{w}} \frac{\mathbf{w^T S_B w}}{\mathbf{w^T S_w w}} \tag{3.20}$$

onde $\mathbf{S_B}$ é a matriz de espalhamento inter-classes, e $\mathbf{S_w}$ a matriz intra-classes. Para duas classes C_1 e C_2, tais matrizes são definidas como:

$$S_B = \sum_{i=1}^{k=2} N_i \left(\mu_\mathbf{i} - \mu\right)\left(\mu_\mathbf{i} - \mu\right)^T = \frac{N_1 N_2}{N}\left(\mu_i - \mu\right)\left(\mu_i - \mu\right)^T$$

$$S_w = \sum_{i=1}^{k=2} \sum_{j \in C_i} \left(\mathbf{x_i} - \mu_\mathbf{i}\right)\left(\mathbf{x_i} - \mu_\mathbf{i}\right)^T$$

onde N_i é o número de vetores de características na classe C_i, N número total de vetores características, $\mu_\mathbf{i}$ a média da classe C_i, e μ média total.

Contudo, uma vez que estamos interessados na direção $\mathbf{w}$, e não na sua magnitude, a maximização deve ser equivalentemente expressa como

$$\max_{\mathbf{w}} \mathbf{w^T S_B w}$$

restrito a

$$\mathbf{w^T S_w w} = 1$$

Após manipulação matemática usando os multiplicadores de Lagrange, o vetor $\mathbf{w}$ pode ser expresso como $\mathbf{w} \propto \mathbf{S_w^{-1}}\left(\mu_\mathbf{1} - \mu_\mathbf{2}\right)$, e o limiar como $b = \mathbf{w^T}\mu$ (THEODORIDIS; KOUTROUMBAS, 2008).

Logistic Regression- LR

No contexto de modelagem do efeito do tratamento, quer seja em radioterapia ou medicina nuclear, a resposta no tecido segue uma curva sigmoide (tipo-S) (NAQA et al., 2015). Comportamento similar é observado em problemas de classificação de tumores. Em ambos os casos, um modelo com forma sigmoidal que tem sido amplamente utilizado é *logistic regression* (LR). Por definição, LR é um modelo dado por:

$$f(\mathbf{x_i}) = \frac{1}{1 + e^{-g(\mathbf{x_i})}}, i = 1, 2..., N \tag{3.21}$$

onde N é o número de casos (pacientes) descritos pelo vetor de variáveis $\mathbf{x}$ usadas para estimar $f(\mathbf{x_i})$. Essa função recebe valores no intervalo dos números reais e transforma tais valores no intervalo $[0, 1]$, expressando a probabilidade de ocorrência de um certo evento, dado certas características $(\mathbf{x_i})$. Além disso, tal transformação assume uma combinação linear das variáveis $\mathbf{x_i}$, expressa por:

$$g(\mathbf{x_i}) = w_0 + \sum_{j=1}^{S} w_j x_i = \mathbf{w^T x}, i = 1, ..., N; j = 1, ..., S \tag{3.22}$$

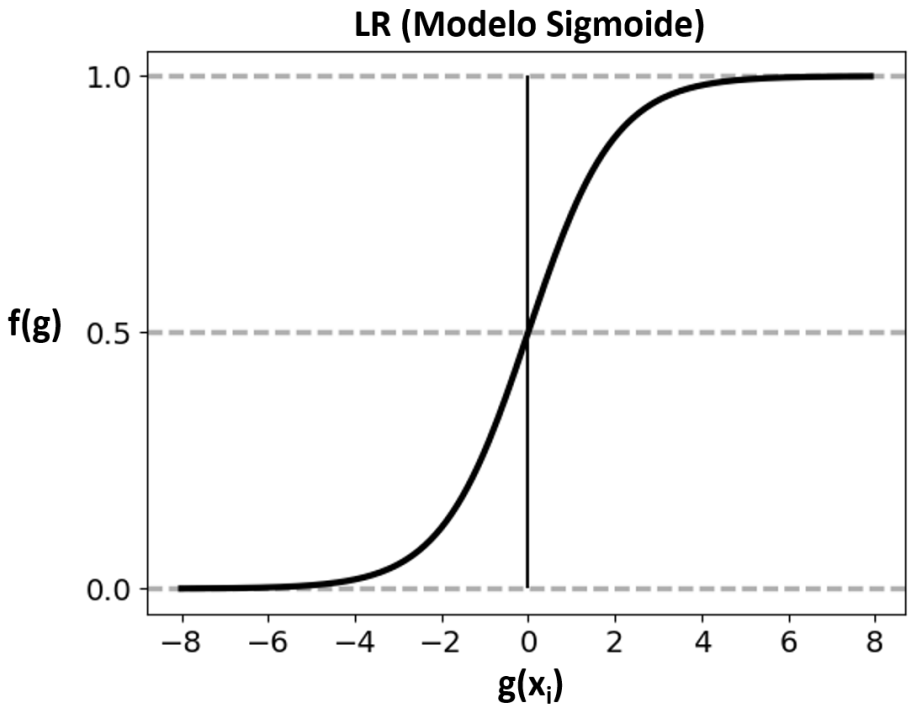
sendo S o número de variáveis e w's o conjunto de coeficientes que são obtidos durante o treinamento do modelo. A Figura 12 mostra o comportamento geral de $f(x)$ que permite estimar a probabilidade de uma certo vetor pertencer a uma dada classe y_i (tumoral ou benigna, por exemplo).

Na prática, para parametrizar tal modelo, ou seja, obter os coeficientes ($\mathbf{w}$) associados a cada variável ($\mathbf{x}$), faz-se uso do princípio de verossimilhança, assumindo que os vetores $\mathbf{x_i}$ usados na fase de treinamento do modelo são independentes. Após algumas manipulações matemáticas, a seguinte função, a ser minimizada com relação a $\mathbf{w}$, é obtida (DUDA; HART; STORK, 2000):

$$J(\mathbf{w}) = -logP(\mathbf{y}|\mathbf{x};\mathbf{w}) = \sum_{i=1}^{N} -y_i log\left(f(g_i)\right) - (1-y_i) log\left(1-f(g_i)\right)) \tag{3.23}$$

A minimização de $J(\mathbf{w})$ é frequentemente obtida usando métodos numéricos tais como *gradient descent- GD* (DUDA; HART; STORK, 2000; THEODORIDIS; KOUTROUMBAS, 2008).

Figura 12 – Função sigmoide f(g) usada em logistic regression(LR).



No contexto de classificação de tumores, um dado VOI é considerado tumoral se $f(g) \geq 0.5$, e benigno caso contrário.

Fonte: autor

Support Vector Machine- SVM

Um classificador SVM tem como objetivo construir um hiperplano que maximiza a distância entre os pontos mais próximos em ambos os lados da fronteira (ou margem) entre as classes.

Esses pontos são conhecidos como vetores de suporte (*support vectors*), e o como eles contribuem para construir a margem ótima é ilustrado na Figura 13. A primeira versão desse modelo foi proposto para lidar com classes linearmente separáveis. Todavia, modificações foram propostas para lidar com dados não linearmente separáveis. A primeira delas consiste na formulação de uma margem suave (soft-margin), enquanto a segunda usa a manipulação de *kernels* criando classificadores não lineares (CORTES; VAPNIK, 1995; CHANG; LIN, 2011). Essas formulações serão discutidas com mais detalhes a seguir.

Contudo, antes de utilizar SVM, é importante que ambos os dados de treinamento e teste estejam normalizados de tal maneira que as variáveis com alta variância não dominem aquelas com baixa variância na construção do hiperplano (NAQA et al., 2015). Para esse trabalho, todas as variáveis que compõem os vetores de características estão normalizados com média zero e desvio padrão unitário.

SVM Linear

Assim como em LDA, a superfície de decisão criada pelo SVM é descrita por $y(\mathbf{x}) = \mathbf{w^T x} - b$. O vetor peso $\mathbf{w}$ e o limiar b são escolhidos de tal modo que a distância entre os vetores de suporte é máxima. Como ilustrado na Figura 13, os vetores de suporte originam dois hiperplanos descritos por $y(\mathbf{x}) = 1$ e $y(\mathbf{x}) = -1$, de tal modo que a distância entre eles é $d = \frac{2}{\|\mathbf{w}\|}$. A maximização da margem pode ser expressa como a seguinte otimização:

$$\min_{\mathbf{w},b} = \frac{1}{2}\mathbf{w^T w}$$

restrito a

$$t_i \left(w^T x_i - b\right) \geq 1,$$

onde a restrição cuida que não haverá pontos dentro da margem. Usando multiplicadores de Lagrange, isso pode ser reescrito como

$$\min_{\mathbf{w},b} \max_{\alpha} \left\{ \frac{1}{2}\mathbf{w^T w} - \sum_{i=1}^{N} \alpha_i \left[t_i \left(\mathbf{w^T x_i} - b \right) - 1 \right] \right\}$$

restrito a

$$\alpha_i \geq 0$$

a partir do qual uma expressão para o vetor peso $\mathbf{w}$ pode ser obtida como uma combinação linear das variáveis $\mathbf{x}$

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i t_i \mathbf{x_i}$$

Assim, a superfície de decisão é expressa em termos dos vetores de suporte, uma vez que apenas eles correspondem a α_i diferente de zero. Uma solução robusta para o limiar b pode ser então obtido pela média dos N_{vs} vetores de suporte,

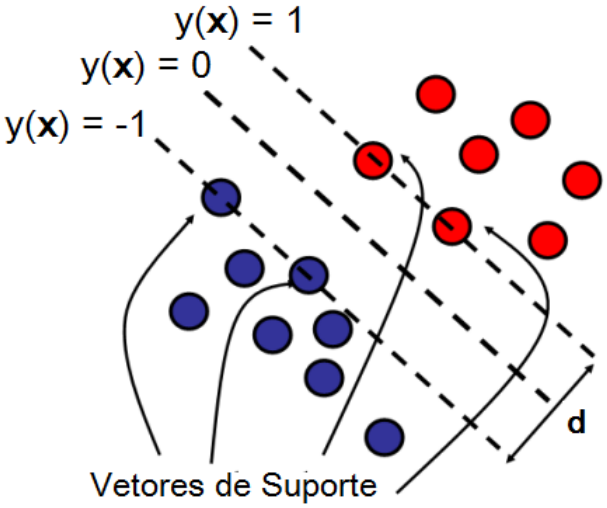
$$b = \frac{1}{N_{vs}} \sum_{i=1}^{N_{vs}} \left(w^T x_i - t_i \right)$$

A forma da Lagrangiana $L(\mathbf{w}, b, \alpha)$ pode ser equivalentemente escrita na forma dual substituindo a expressão acima obtida para $\mathbf{w}$:

$$\max_{\alpha} \tilde{L}(\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j t_i t_j \mathbf{x_i^T x_j} \right\}$$

Escrita dessa forma, o critério de otimização passa a ser definido em termos do produto escalar das variáveis. Essa propriedade é fundamental para a criação de classificadores SVM não lineares.

Figura 13 – Ilustração 2-D da construção da margem máxima ou hiperplano em SVM.



Essa superfície de decisão maximiza a distância entre os vetores de suporte, indicados pelas setas. Pontos associados a classe C_1 são mostrados em vermelho, e aqueles associados a C_2 estão em azul.

Fonte: autor

SVM Não-Linear

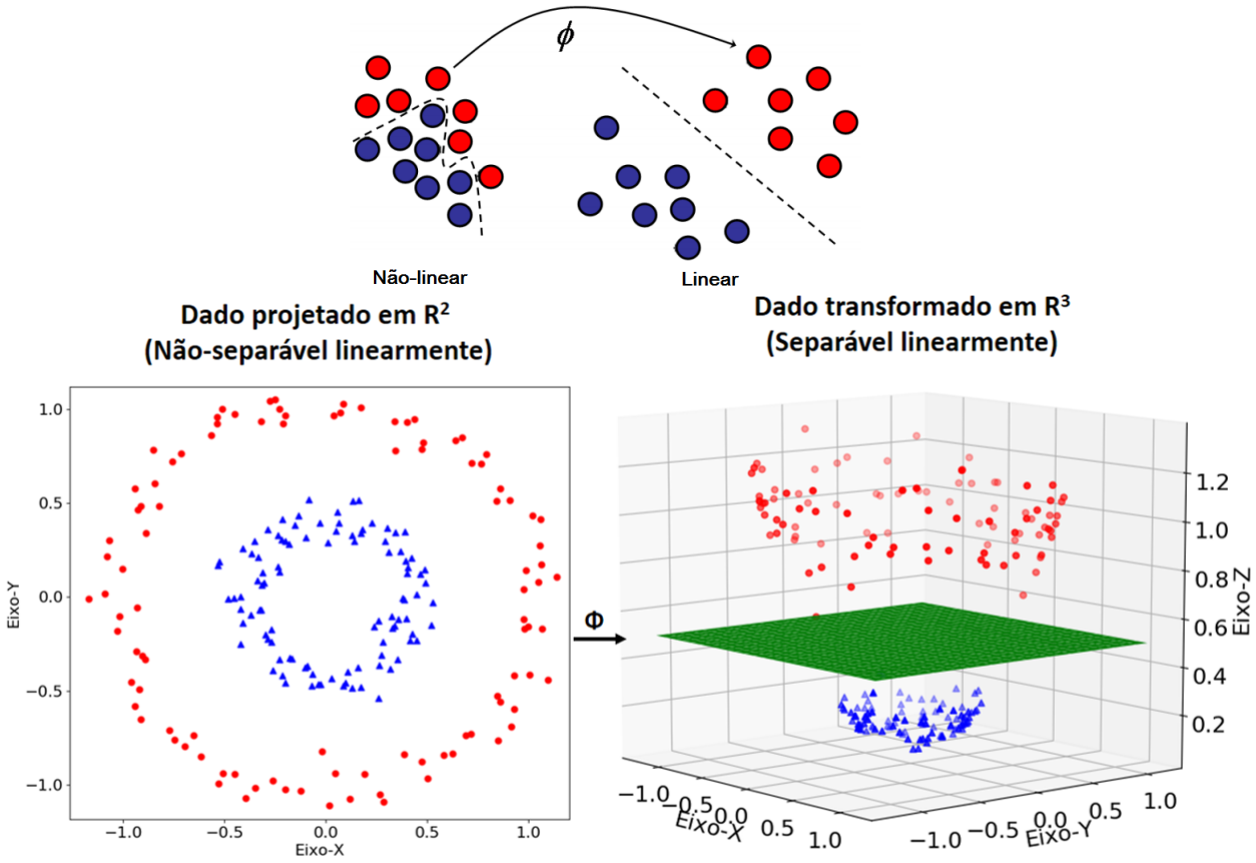
Em casos onde as classes não são linearmente separáveis no espaço de características original, uma função não linear, $\phi(\mathbf{x})$, pode ser usada para projetar cada vetor de características em um espaço de dimensão maior. Como ilustra a Figura 14, um problema não linear pode ser separado por um hiperplano linear nesse novo espaço. Para isso, o problema de otimização passa a ser definido pela seguinte Lagrangiana:

$$\tilde{L}(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j t_i t_j \phi(\mathbf{x_i})^{\mathbf{T}} \phi(\mathbf{x_j})$$

Em outras palavras, o critério de otimização é então expresso em termos do produto interno dos vetores características transformados. Um passo fundamental aqui é a escolha de ϕ,

que deve ser feita de tal modo que os produtos internos possam ser expressos em termos de um *Kernel* $\mathbf{k}(\mathbf{x_i}, \mathbf{x_j}) \equiv \phi(\mathbf{x_i})^{\mathbf{T}} \phi(\mathbf{x_j})$ e, desse modo, não é necessário realizar a transformação explicitamente. Isto tem diversas vantagens do ponto de vista computacional considerando problemas não lineares com alta dimensão. Um exemplo de *Kernel* comumente utilizado é um *Kernel* gaussiano (*RBF- radial basis function*), dado por $\mathbf{k}(\mathbf{x_i}, \mathbf{x_j}) = exp\left(-\gamma \left\|\mathbf{x_i} - \mathbf{x_j}\right\|^2\right)$, onde $\gamma > 0$ é um parâmetro que descreve a largura de K.

Figura 14 – Exemplo de manipulação de *Kernels* para obtenção de separação linear usando SVM.



(Superior) Fronteira de decisão não linear no espaço de características de entrada. Usando a função não linear,$\phi(\mathbf{x})$, o espaço original pode ser transformado em linearmente separável em outra dimensão mais alta. (Inferior) Exemplo de dados em R^2 não linearmente separável (esquerda). Mesmo dado transformado por $[\mathbf{x_1}, \mathbf{x_2}] = [\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_1^2} + \mathbf{x_2^2}]$ passa a ser linearmente separável,em R^3, pelo hiperplano em verde.
Fonte: autor

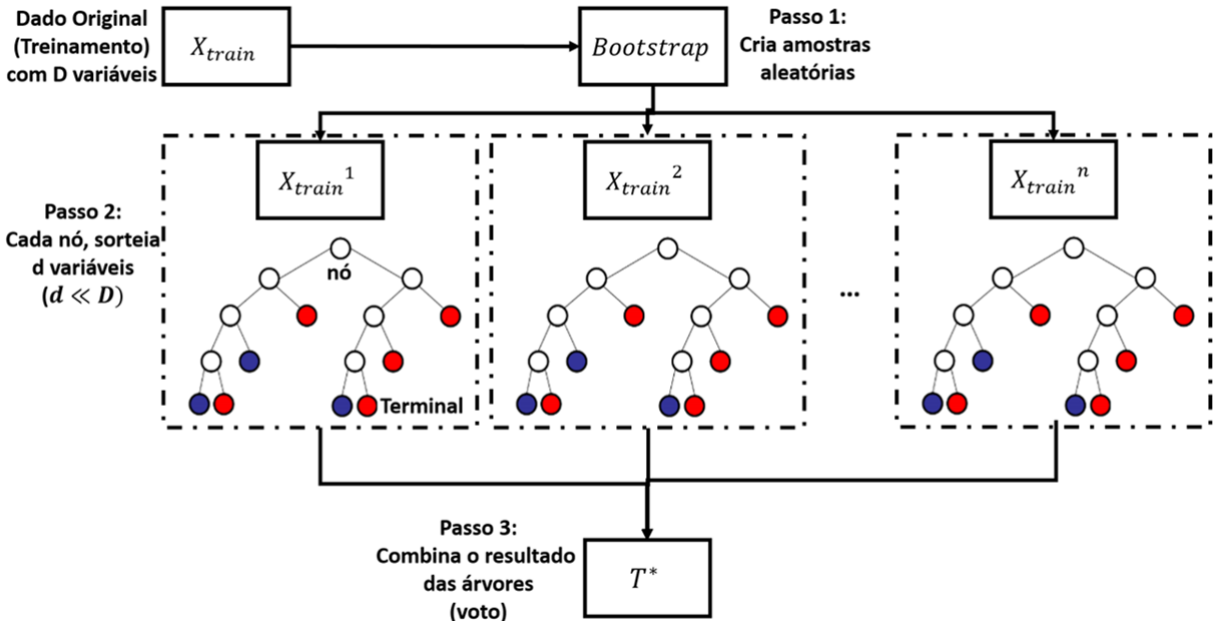
Boosting

Boosting é uma das estratégias de aprendizagem de máquina que combina classificadores com baixa performance, para criar um modelo com mais alta performance. Nesse sentido, há uma variedade de classificadores seguindo esse princípio, porém, abordar-se-á a seguir Random Forest (BREIMAN, 2001) e Extremelly Random Forest (GEURTS; ERNST; WEHENKEL, 2006).

Random Forest-RF

Uma floresta aleatória de árvores de decisão (*random forest-RF*) consiste do uso de várias árvores de decisão, ao invés de uma única, sendo a predição final da classe para um exemplo teste dada pelo voto da maioria [6] das predições de todas as árvores individuais. Desenvolvido por Breiman e Adele Cutler (BREIMAN, 2001), RF usa um subconjunto das observações originais (X_{train}) e um subconjunto de variáveis (d) que são escolhidos aleatoriamente para construir múltiplas árvores independentes, como ilustrado na Figura 16. Tal estratégia aumenta o desempenho de classificação se comparado a uma única árvore de decisão.

Figura 15 – Floresta de decisão aleatória.



Em cada árvore, cada nó é particionado baseado em uma única variável, e cada galho termina no chamado nó terminal o qual fornece a predição da classe para um dada caso teste. A predição final da classe para um dado caso teste é a moda (ou a média) das predições fornecidas por todas as árvores.

Fonte: autor

Esses subconjuntos de X_{train} são gerados usando reamostragem aleatória com reposição dos N exemplos originais por meio do *bootstrap* (vide seção 3.2.4). Por esse método, aproximadamente 36.8% dos N casos disponíveis não estão presentes no conjunto treinamento de cada árvore. Esses dados podem ser utilizados para testes de predições internas e, assim, podem ser úteis tanto para obtenção de uma estimativa da generalização da floresta de árvores, assim como medir a importância de cada variável.

Uma vez construídas as amostras aleatórias, segue-se a seleção das variáveis. Em cada nó da árvore, $d << D$ variáveis são aleatoriamente selecionadas, e o nó é dividido binariamente usando o melhor particionamento possível. Uma vantagem desse processo é a redução da

[6] Outros critérios pode ser utilizados como a média das predições de todas as árvores individuais.

correlação entre as árvores, implicando em maior diversidade no processo de classificação. O valor padrão recomendado para maioria das aplicações é $d = \sqrt{D}$ (BREIMAN, 2001).

Em seguida, o particionamento de nó-pai, n_p, em nós-filhos n_l e n_r, ocorre de acordo com um critério que visa maximizar a homogeneidade dos nós n_l e n_r com relação a n_p. Esse critério recebe o nome de *Gini index* (I_G) e mede a homogeneidade de um dada classe dentro do nó em análise. Em caso de duas classes, C_1 e C_2, para um nó n, esse índice de impureza pode ser expresso como:

$$I_G(n) = 1 - \sum_{K=1}^{2} p_K^2 \tag{3.24}$$

sendo p_K a proporção de casos pertencentes à classe K presentes no nó n, com I_G pertencendo ao intervalo $0 \leq I_G \leq (1 - \frac{1}{K})$. Um valor nulo indica que o nó contém apenas exemplos pertencentes a uma única classe, e o valor máximo indica que no nó há casos de ambas as classes com mesma proporção. A melhor separação é aquela com menor impureza, e equivale a maximizar a seguinte diferença, também conhecida como ganho de informação:

$$\Delta I_G(n_p) = I_G(n_p) - p_l I_G(n_l) - p_r I_G(n_r) \tag{3.25}$$

onde p_l e p_r são as proporções de casos no nó n_p atribuídos aos nós-filhos n_l e n_r, respectivamente.

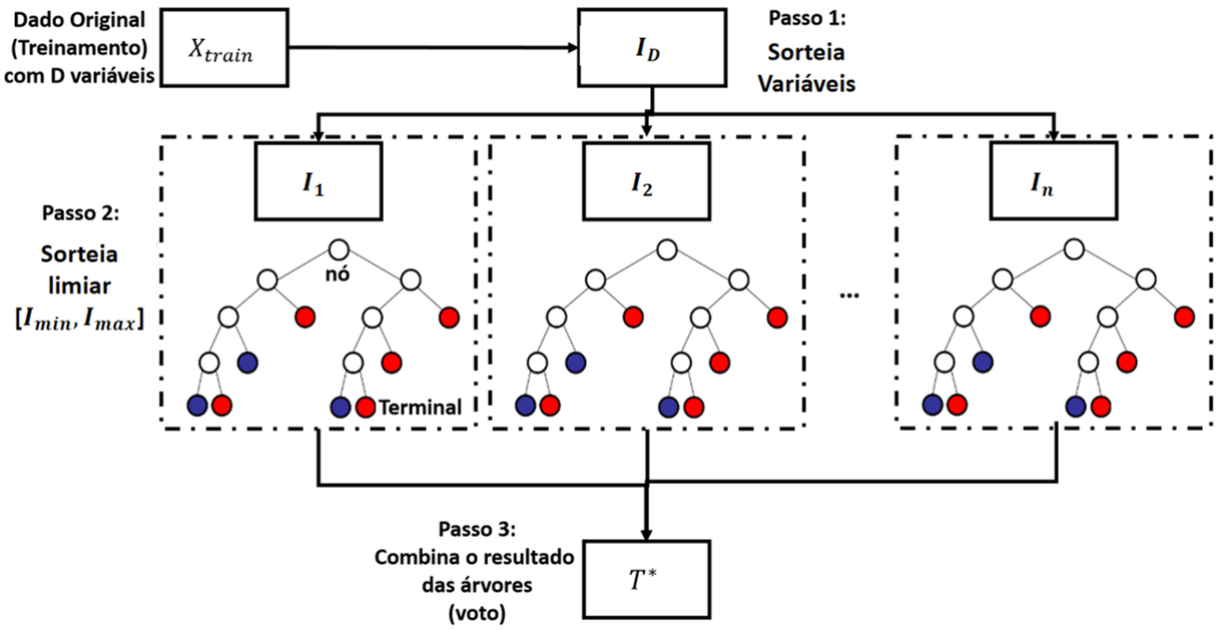
Uma importante propriedade de ΔI_G é a avaliação da importância relativa de cada variável que forma o vetor característica, $\mathbf{x_i}$, na tarefa de classificação. A importância de cada característica pode ser computada pela soma de ΔI_G para todos os nós particionados na floresta por aquela mesma característica, dividida pelo número total de árvores na floresta. Essa fração fornece a importância de cada característica e um filtro pode ser aplicado de tal maneira que apenas as variáveis com os maiores graus de importância passam a ser de fato utilizadas na classificação final dos exemplos testes.

Extremely randomized trees: ErT

Visando produzir ainda mais aleatoriedade, (GEURTS; ERNST; WEHENKEL, 2006) propôs o algoritmo de árvores extremamente aleatórias (ErT).

Esse passo a mais na aleatoriedade está presente na forma como as divisões dos nós são computadas. Como em RF, um subconjunto aleatório de características candidatas (d) é usado, todavia, em vez de procurar os limiares mais discriminatórios, estes limiares são aleatoriamente escolhidos a partir do intervalo de valores da característica selecionada. Isso geralmente permite reduzir a variância do modelo, à custa de um certo aumento no viés, comparado aos outros métodos baseados em árvore de decisão. Para gerenciar esse aumento no viés, toda a amostra de aprendizagem será usada para desenvolver as árvores de decisão, em vez de uma réplica *bootstrap* como no caso de RF.

Figura 16 – Floresta de decisão extremamente aleatória.



Combina árvores de decisão usando seleção aleatória de características (I_D) e limiares $[I_{min}, I_{max}]$.
Fonte: autor

Além disso, o processo de separação dos nós das árvores é realizado usando apenas dois parâmetros, aos quais o usuário precisa estar atento: o número de atributos a ser aleatoriamente selecionado (d) a cada nó e o número de árvores na floresta (n_{tree}). Quanto maior n_{tree} melhor, porém mais longo o tempo de processamento. Além disso, geralmente os resultados apresentam um valor estacionário a partir de um número crítico de árvores. Com relação a d, quanto menor melhor a redução na variança, mas também maior o viés. Esse parâmetros podem ser definidos de acordo com o problema a ser resolvido, quer seja definindo-os manualmente, quer seja automaticamente, usando por exemplo *cross-validation*, conforme abordado na próxima seção.

Por fim, a predição final para um conjunto teste é a agregação das predições fornecidas por todos as árvores, dada pela moda da distribuição de resultados, no caso de classificação.

3.2.4 Avaliação e seleção de modelos

K-fold cross-validation

Reamostragem é de fundamental importância para simular novas amostras em situações onde repetir experimentos é inviável ou mesmo impossível. Isto é uma realidade frequente no contexto médico. E, quando o assunto são modelos de aprendizagem de máquina, três são os principais objetivos:

1. estimar a generalização do desempenho preditivo de nosso modelo em dados futuros (não vistos).

2. aumentar o desempenho preditivo ajustando o algoritmo de aprendizagem e selecionando o modelo de melhor desempenho a partir de um determinado espaço de hipóteses

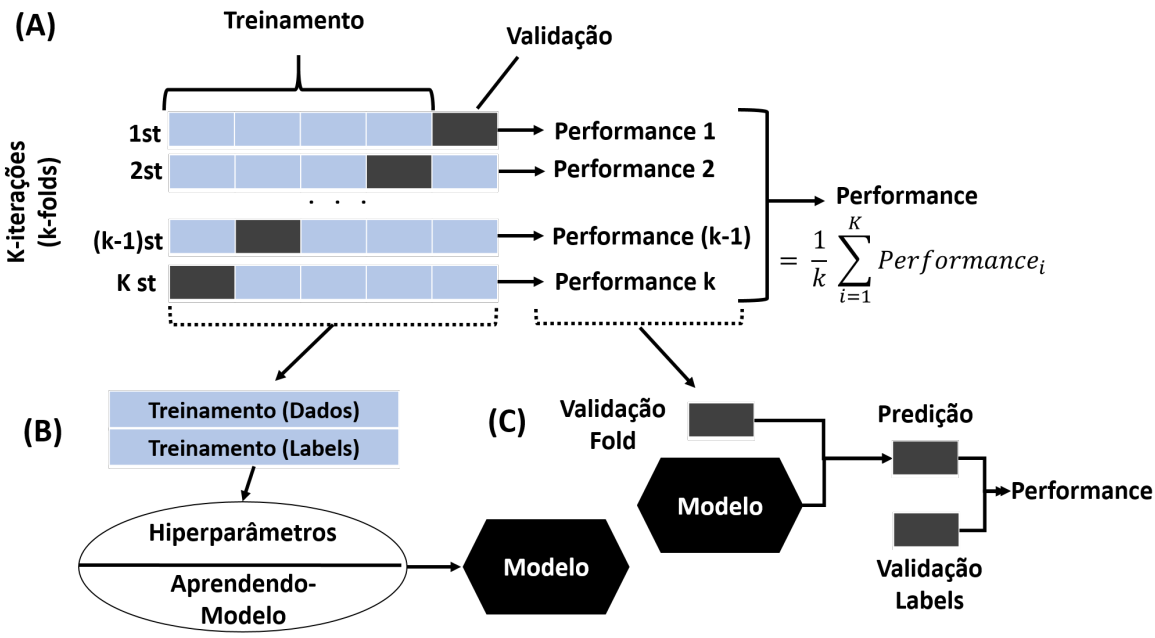
3. identificar o algoritmo de aprendizado de máquina que é mais adequado para o problema em questão. Isso implica comparar diferentes algoritmos, selecionando o que fornece o melhor desempenho, a partir de seu espaço de hipótese ótimo.

Desse modo, a pergunta que surge é como avaliar e selecionar modelos, dado um conjunto destes? Na literatura, a estratégia mais popular para avaliar modelos é o chamado *k-fold cross-validation* (KOUROU et al., 2015; NAQA et al., 2015). A Figura 17 mostra os passos envolvidos nessa estratégia. A ideia é iterar em um conjunto de dados k-vezes. A primeira linha corresponde ao Fold-1, a segunda ao Fold-2, etc. Em cada rodada, divide-se o conjunto de dados em k-partes: uma parte é usada para validação e as $k-1$ partes restantes são incorporadas em um subconjunto de treinamento para construção dos modelos. Este procedimento resultará em k-modelos diferentes. Esses modelos são obtidos em conjuntos de treinamento distintos, mas parcialmente sobrepostos, e avaliados em conjuntos de validação independentes. Por fim, calcula-se o desempenho dessa validação cruzada por meio da média aritmética das k-estimativas de desempenho dos conjuntos de validação.

Essa abordagem garante que (1) a cada *fold*, os conjuntos treinamento e teste são separados; (2) uma vez que todo esquema tenha sido executado, todos os casos terão sido usados para testar o modelo; (3) nenhum caso será usado mais de uma vez na fase de teste; e (4) cada caso terá sido usado $k-1$ vezes durante o treinamento. A ideia por trás dessa abordagem é reduzir o viés pessimista ao usar mais dados para o treinamento, assim como garantir o teste em um conjunto independente.

Além disso, quando o dado apresenta mais classes de um dado grupo, por exemplo, mais indivíduos sadios que acometidos com uma determinada doença, diz-se que os dados estão em desequilíbrio. Desse modo, existe o risco, neste caso, de a classe com menor representatividade não aparecer de maneira alguma no subconjunto em um dos *folds*. O desempenho do classificador em tal conjunto de dados estaria comprometido, assim como seria excessivamente otimista. Da mesma forma, se os dados de treinamento contém menor proporção de exemplos minoritários do que o conjunto de dados reais, o desempenho do classificador seria excessivamente pessimista. Para evitar ambos os problemas, um processo chamado *k-fold cross-validation* estratificada é usado para garantir que a distribuição de cada classe seja respeitada nos conjuntos de treinamento e teste criados em cada *fold*.

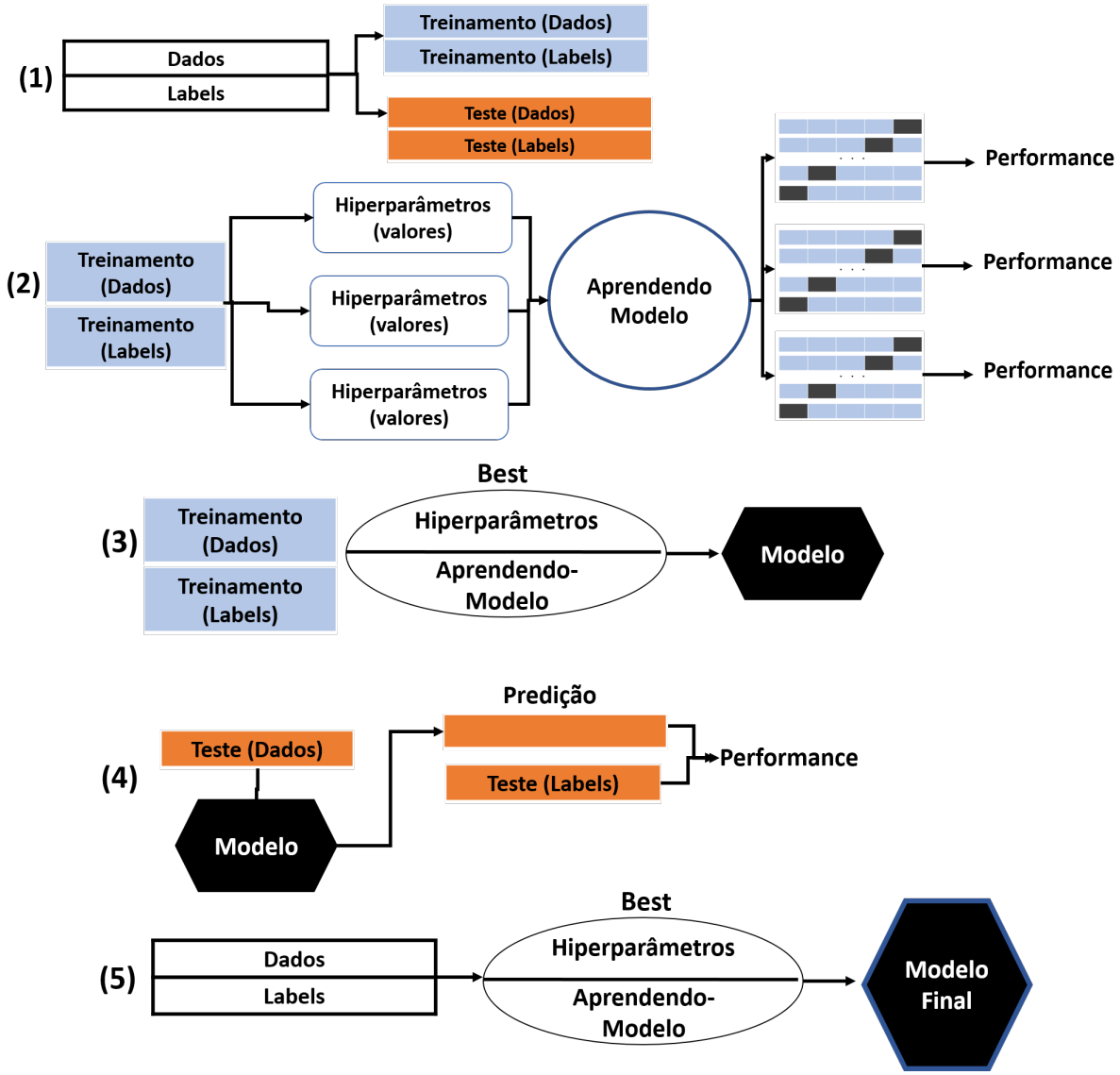
Figura 17 – *K-fold cross-validation para avaliação do modelo de aprendizagem.*



Fonte: autor

Porém, na prática, normalmente se dispõem de vários modelos com diferentes espaço de hipóteses, exemplo: LR, SVM, Random Forest, etc; e o objetivo final é selecionar um deles. Em situações como essa, *k-fold cross-validation* entra em ação como parte da mecânica de seleção que exige estratégias adicionais, como mostra a Figura 18.

Figura 18 – Seleção de modelos usando *cross-validation*.



Fonte: autor

Aqui, novamente, a idéia-chave é manter um conjunto de dados de teste independente, como ilustra o passo (1) na Figura 18. Esse conjunto teste é usado para avaliação final de cada modelo (LR, SVM, Random Forest, etc).

No segundo passo, experimentam-se várias configurações de hiper-parâmetros, usando uma das seguintes abordagens: otimização Bayesiana, pesquisa aleatória ou *grid search*. Para cada configuração de hiper-parâmetro, aplica-se a *k-fold cross-validation* no conjunto de treinamento, resultando em vários submodelos e estimativas de desempenho.

Tomando as configurações do hiper-parâmetro que correspondem ao submodelo de melhor desempenho, pode-se usar o conjunto de treinamento completo para a construção do submodelo ótimo (passo 3, Figura 18).

Obtido esse submodelo ótimo, o passo seguinte consiste em avaliar o mesmo fazendo uso do conjunto de testes independente, que aleatoriamente foi separado do conjunto de treinamento

no primeiro passo. Esse processo de obtenção do submodelo ótimo é repetido para cada um dos modelos de aprendizagem na lista de modelos (LR, SVM, Random Forest, etc).

Finalmente, concluído o estágio de avaliação, seleciona-se e treina-se o melhor modelo com todos os dados inicias. Esse modelo vai compor a fase de uso em larga escala para novos casos clínicos.

Bootstrap

O método bootstrap é uma técnica de reamostragem para estimar a distribuição de uma amostragem e, no contexto deste trabalho, estamos particularmente interessados em estimar a incerteza de nossa estimativa de desempenho.

O método bootstrap foi introduzido por Bradley Efron em 1979 (B. Efron and R. Tibshirani, 1993). A idéia principal do método é gerar novos dados a partir de uma população por meio de amostragem repetida, com reposição, do conjunto de dados original.

Em um passo a passo, o método bootstrap funciona como se segue:

1. Dado um conjunto de dados de tamanho N.
2. Para B rodadas de inicialização:
3. Uma única instância desse conjunto de dados é sorteada e atribuí-se a ela a jth ordem na amostra *bootstrap*. Repete-se este passo até que a amostra *bootstrap* tenha tamanho N - o tamanho do conjunto de dados original. Cada vez, sorteia-se amostras do mesmo conjunto de dados original de tal modo que certas amostras aparecerem mais de uma vez na amostra *bootstrap* e outras não.
4. Obtém-se o modelo para cada uma das amostras B *bootstrap* e calcula-se a performance da reamostragem.
5. Calcula-se a performance do modelo como a média sobre as B estimativas de performance.

Originalmente, o método *bootstrap* visa determinar as propriedades estatísticas de um estimador quando a distribuição subjacente é desconhecida e amostras adicionais não estão disponíveis. Assim, para explorar este método para a avaliação de modelos preditivos, faz-se necessária uma abordagem ligeiramente diferente: a chamada técnica *out-of-bag* (OOB) *bootstraping*. Nesse método, usa-se amostras OOB como conjuntos de teste (X_{test}) para avaliação do modelo em vez de avaliar o modelo nos dados de treinamento (X_{train}). As amostras OOB são os conjuntos exclusivos de instâncias que não são usados para obtenção do modelo, conforme mostrado na Figura 19. Nessa figura as amostras *bootstrap* aleatórias obtidas de um exemplo com dez amostras ($X_1, X_2, ...X_{10}$) e sua correspondente amostra OOB testes. Essa estratégia pode ser usada no passo 1 da Figura 18, generalizando o processo de seleção de modelos, ao inserir a possibilidade de avaliar a estabilidade dos mesmos.

A probabilidade de um dado caso aparecer em X_{test}, para uma amostra de tamanho N é:

$$P(X_{test}) = \left(1 - \frac{1}{N}\right)^N \tag{3.26}$$

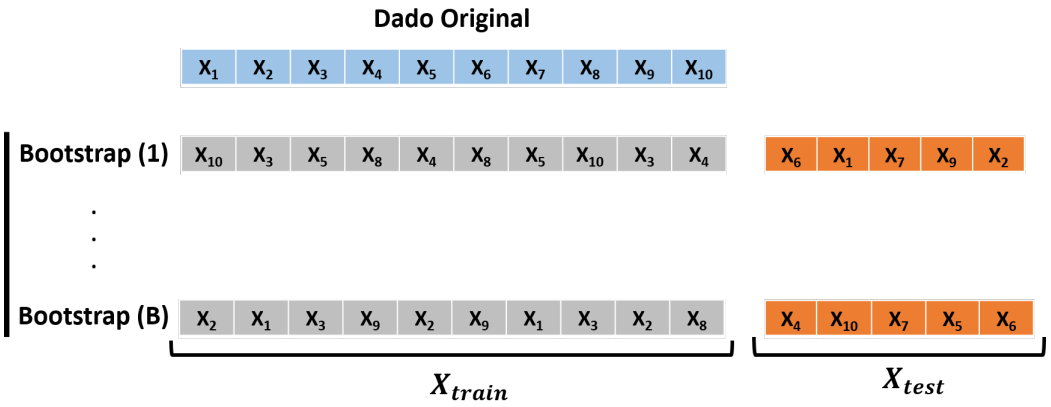
a qual é assintoticamente equivalente a $\frac{1}{N} \approx 0.368$, quando $n \to \infty$.

Vice-versa, a probabilidade de um caso aparecer em X_{train} é:

$$P(X_{train}) = 1 - \left(1 - \frac{1}{N}\right)^N \approx 0.632 \tag{3.27}$$

Isso significa que, aproximadamente $0.632xN$ casos únicos apareceriam no conjunto treinamento, e o restante, $0.368xN$ estariam reservados para teste a cada iteração.

Figura 19 – *Bootstrap Resampling*.



Fonte: autor

Medidas de desempenho de um modelo

Independente da estratégia usada, conforme discutido na seção anterior, diferentes opções existem para estimar a performance de um dado modelo.

Para isso, alguns termos como verdadeiros positivos (TP), verdadeiros negativos (TN), falso-positivos (FP) e falso-negativos (FN) comparam os resultados do classificador em teste com rótulos externos confiáveis. Importante salientar que, os termos positivo e negativo referem-se à predição do classificador (às vezes conhecida como a expectativa), e os termos verdadeiro e falso referem-se a se essa previsão corresponde ao rótulo externo (às vezes conhecido como a observação). Esses quatro parâmetros podem ser formuladas em uma matriz de contingência, como ilustra a Figura 20.

Alguns parâmetros comuns extraídos dessa matriz são: acurácia (Acc, Equação 3.28), precisão (PPV, Equação 3.29), sensitividade ou taxa de verdadeiro-positivo (Sens, Equação 3.30) e taxa de falso-positivo (FPR, Equação 3.31).

$$Acc = \frac{TP + TN}{P + N} \tag{3.28}$$

$$PPV = \frac{TP}{TP+FP} \tag{3.29}$$

$$Sens = \frac{TP}{P} \tag{3.30}$$

$$FPR = \frac{FP}{N} \tag{3.31}$$

Outras medidas possíveis incluem especificidade, taxa de valor preditivo negativo, etc, conforme detalhadamente abordado em (JAPKOWICZ; SHAH, 2011).

A Figura 20 também mostra dois casos, à título de exemplo. Nos dois casos, a acurácia é a mesma. Contudo, nos casos, os modelos apresentam diferentes comportamentos. No exemplo 1, o classificador apresenta baixa capacidade de reconhecimento de casos positivos, e alto reconhecimento dos negativos. Já no exemplo 2, observa-se o contrário.

Em casos como esses, e ainda mais quando as classes não estão balançadas, a acurácia não é a melhor métrica para selecionar modelos.

Figura 20 – ***Matriz de contingência para dois exemplos com a mesma acurácia*****.**

		Label da Predição	
		Pos	**Neg**
Label Verdadeira	**Sim**	TP	FP
	Não	FN	TN
		P = TP + FN	N = FP + TN

Exemplo 1

	Pos	**Neg**
Sim	200	100
Não	300	400
	P = 500	N = 500

Exemplo 2

	Pos	**Neg**
Sim	400	300
Não	100	200
	P = 500	N = 500

⇩

Mesma acurácia

$$Acc = \frac{TP+TN}{P+N} = \frac{600}{1000} = 60\,\%$$

Fonte: autor

Para superar os problemas encontrados pela acurácia, e outros parâmetros de performance dependentes da proporção de classes presentes na população, análise da curva ROC e seus sumários estatísticos tem se mostrado de grande importância em aprendizagem de máquina.

Para um modelo treinado em todos os possíveis níveis de limiar, duas medidas são tomadas para construção da curva ROC: a sensitividade e a taxa de falso-positivo (FPR) (ou taxa de alarme falso). Uma vez que todas as medidas foram feitas, os pontos representados por todos os pares obtidos são plotados no chamado espaço ROC, um gráfico que traça a sensitividade, ou taxa com que os verdadeiros positivos são classificados, como uma função da taxa de falso-positivos. Os pontos são então unidos em uma curva suave, que representa a curva ROC para um dado classificador. A Figura 21 mostra duas curvas ROC que representam o desempenho de dois classificadores f_1 e f_2 em todos os intervalos de limiares possíveis.

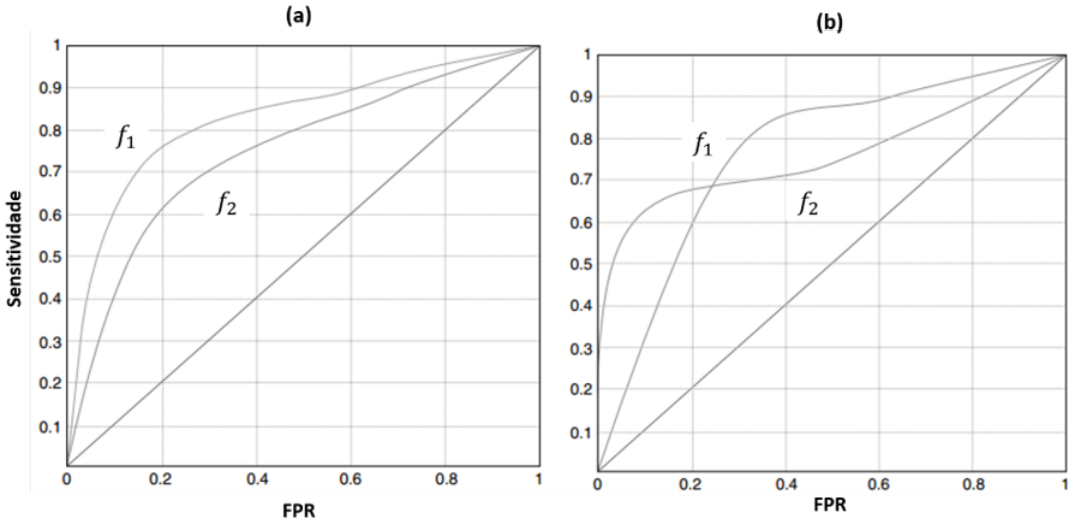
Quanto mais próxima uma curva que representa um classificador f está do canto superior esquerdo do espaço ROC (ou seja, pequena taxa de falso-positivo, e grande taxa de verdadeiros-positivos), melhor o desempenho desse classificador. No exemplo da Figura 21a, f_1 funciona melhor do que f_2. Na prática, porém, é frequente situações como a da Figura 21b: um classificador domina o outro em algumas partes do espaço ROC, e em outros não. Em casos como esse, a questão clínica a ser resolvida é quem define a região de interesse da curva ROC e, consequentemente, o limiar ótimo a partir do qual todos os parâmetros de interesse serão derivados. Esse limiar em geral é definido como o par especifidade i e sensitividade j que maximiza a soma da Equação 3.32, tal como proposto por *Youden* (YOUDEN, 1950).

$$J = \max_{i,j}(Sensitividade + Especificidade - 1) \tag{3.32}$$

Assim sendo, o motivo pelo qual análise da curva ROC [7] é adequado para comparar modelos em diferentes contextos é que a performance é decomposta em duas diferentes medidas e se aplica em casos onde uma classe é mais frequente que outra, conforme discutido anteriormente.

[7] Curva ROC é um gráfico que ilustra a capacidade de diagnóstico de um classificador binário à medida que o seu limiar de discriminação é variado. Ela mostra a taxa de positivo verdadeiro (TPR) contra a taxa de falso positivo (FPR) em várias configurações de limiar.

Figura 21 – ***Curva ROC comparando dois modelos (f_1 e f_2), em diferentes situações.***



Fonte: autor

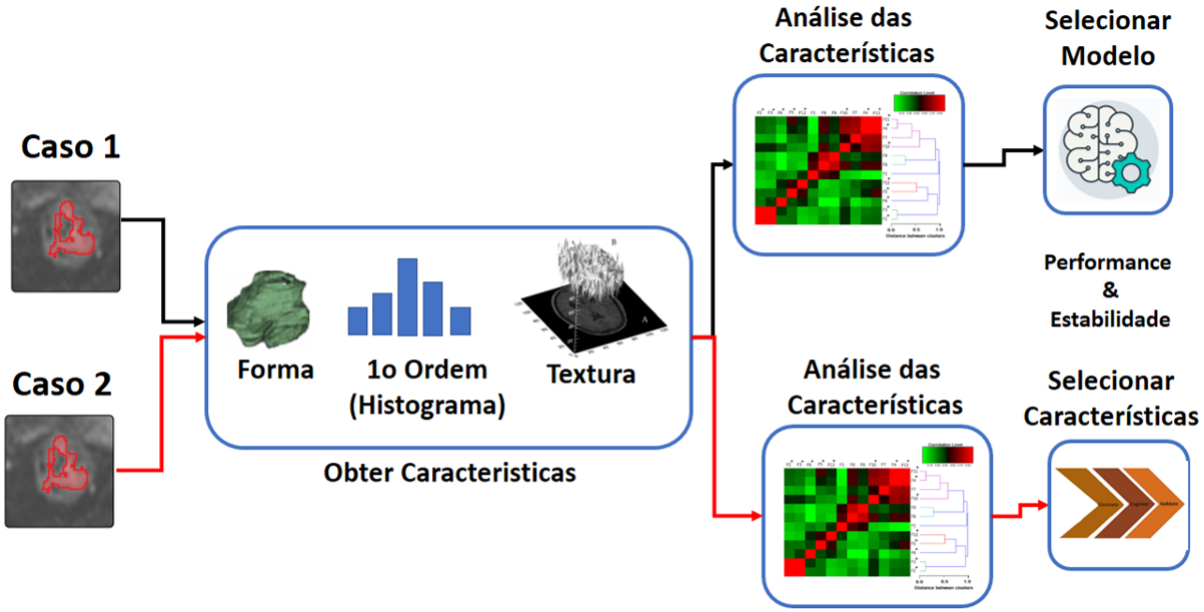
4 METODOLOGIA

Nesta seção, a metodologia será dividida em duas partes: casos clínicos 1 e 2, conforme ilustra a Figura 22. Caso clínico 1 trata da seleção de características e modelos de classificação para auxiliar o diagnóstico de linfonodos em pacientes com suspeita de câncer de colo do útero. Em seguida, o caso 2 discute o desenvolvimento e aplicação de métodos adaptativos para extração de características.

Em ambos os experimentos, cada característica (adaptativa e não-adaptativa) e modelo de aprendizagem foram avaliados em termos de seu valor adicional no diagnóstico e estabilidade de predição, quando comparado à variável usada na rotina clínica. Para isso, em conjunto com abordagens específicas a cada problema, a técnica de reamostragem *bootstrap*, foi utilizada tal como descrito em detalhe nas seções que se seguem.

Por fim, vale salientar que as imagens utilizadas neste trabalho pertencem ao banco de dados dedicado à pesquisa do departamento de Radiologia do Hospital Universitário da Universidade de Leuven (KU Leuven), Bélgica [1].

Figura 22 – Visão geral dos experimentos desenvolvidos na tese.



[1] Tais imagens são de uso exclusivo da KUL. Em caso de interesse em eventual colaboração, um projeto precisa ser submetido deixando clara finalidade de uso das mesmas.

4.1 Caso 1: Câncer de cólon do útero

4.1.1 Amostra de pacientes

Todos os linfonodos com diâmetros de pelo menos 4 mm foram delineados semi-automaticamente por dois radiologistas com mais de 12 anos de experiência em MRI. Um total de 129 linfonodos provenientes de cólon do útero (65 benignos e 64 malignos, confirmados por biópisia) foram inicialmente utilizados nesse estudo retrospectivo.

Os parâmetros de aquisições do *scanner*, assim como o processo de obtenção dos mapas de ADC seguem trabalho anterior publicado pelo grupo (VANDECAVEYE et al., 2009).

Obtidos os VOIs, os passos seguintes consistem em construir o espaço de características. Tal espaço de características é uma matriz 2D, cujas dimensões são da ordem do número de casos *versus* o número de variáveis, ou características, que quantitativamente descrevem cada caso.

4.1.2 Espaço de Características

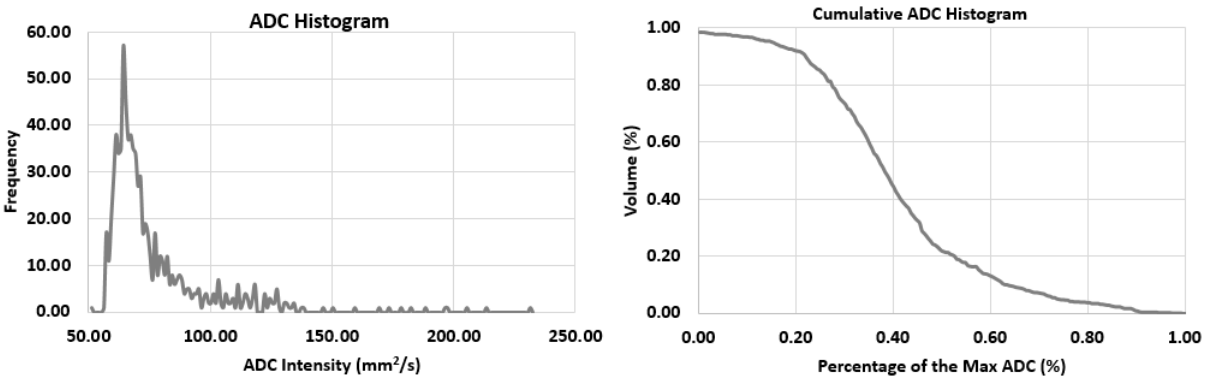
Dado o tamanho das lesões, as características utilizadas são compostas de variáveis descrevendo a forma, tal como o volume (F1) de cada lesão, seguido por 12 características de primeira ordem baseadas na distribuição de probabilidade simples e acumulada de ADCs. Desses doze, nove características se distribuem entre aqueles descrevendo tendência central (média-F2 e mediana-F3), forma da distribuição de probabilidade (segundo-F4, terceiro-F5 e quarto momentos-F6, incluindo coeficiente de variância do histograma-F7, entropia-F8 e energia-F9), cujo potencial em DW-MR tem sido discutido nesse consenso internacional (PADHANI et al., 2009).

As três demais características (F9, F10 e F11) são baseadas na distribuição acumulada (CAVH- cumulative ADC Volume-Histogram). Tais características baseadas na CAVH foram adaptadas do trabalho de (NAQA et al., 2009). Neste trabalho, baixos valores da área de CAVH (F9, AUC-CAVH) (Figura 23) estão associados a alta heterogeneidade. Além disso, as duas demais variáveis são definidas de maneira complementar, explorando diferentes aspectos da curva. Enquanto F10 ($I_{10-90} - CAVH$) mede a diferença entre os *limiares de ADC* contendo 10% e 90% do volume da lesão, F11 ($V_{10-90} - CAVH$) mede a diferença entre os *volumes* determinados pelos limiares de 10% e 90% do maior valor de ADC.

Além disso, cada variável (F_i) foi normalizada com relação a sua média ($\bar{F_i}$) e o desvio padrão (σ_{F_i}) (Eq. 4.1), a fim de colocar todas elas numa mesma escala. Essa transformação é conhecida como padronização e, por meio da mesma, cada variável será expressa com média zero e desvio padrão unitário.

$$F_i^{'} = \frac{F_i - \bar{F_i}}{\sigma_{F_i}} \tag{4.1}$$

Figura 23 – ADC histograma simples e acumulado.



As características (F1-F8) foram computadas a partir do histograma simples (esquerda), enquanto as derivadas de CAVH (F9-F11) do histograma a direita. Fonte: autor

Em seguida, aplicou-se um teste estatístico não-paramétrico, especificamente *Mann-Whitney* (MW) (MANN; WHITNEY, 1947), para avaliar individualmente o grau de significância de cada uma delas. Variáveis com $p < 0.05$ foram consideradas significativas. Adicionalmente, correlações entre todos os possíveis pares de características foram computadas usando coeficiente de correlação de *Spearman* (ρ) (MYERS; WELL, 1995) para construir a matriz de correlação. O nível de correlação entre as variáveis foi visualizado usando um dendograma (ROKACH; MAIMON, 2005), obtido por um *cluster* hierárquico usando o critério aglomerativo do vizinho mais próximo, assim como $1 - |\rho|$ como medida de similaridade. Dessa maneira, variáveis com alta correlação são agrupadas no mesmo *cluster*, uma vez que apresentam pequena distância entre sí. Tal matriz foi usada no processo de seleção de características conforme descrito na próxima seção.

4.1.3 Seleção de Características

Dado um conjunto de variáveis, um passo importante é a seleção da base do espaço de características, uma vez que mais variáveis em um modelo não necessariamente significa mais informação: tais variáveis normalmente compartilham certo grau de redundância o que diminui sua relevância quando considerada num contexto multivariado [2]. Desse modo, visando fazer o processo de seleção de características independente do modelo preditivo utilizado e, portanto, visando avaliar cada modelo usando um mesmo espaço de características ótimo, um filtro foi utilizado no processo de seleção do conjunto de variáveis (THEODORIDIS; KOUTROUMBAS, 2008).

Este filtro nada mais é do que uma função que cria um *rank* de variáveis de acordo com o grau de redundância e relevância entre cada uma delas usando uma avaliação *heurística*. Essa avaliação consiste em encontrar um subconjunto de características que contém variáveis

[2] Nesse contexto, uma característica é dita redundante se apresenta alto grau de correlação com outras variáveis presentes no conjunto. Ela será relevante em caso de baixa correlação com seus pares e alta correlação com as classes a serem discriminadas.

altamente correlacionadas com as classes ou grupos binários (ex. benigna e maligna) e o mínimo possível correlacionada entre sí. A razão é que se espera que variáveis irrelevantes e redundantes devem ser filtradas, uma vez que elas não adicionam poder discriminativo.

Assim, dado um conjunto S com k características , tal função, M_S, é definida como,

$$M_S = \frac{k.\bar{\rho}_{Fc}}{\sqrt{k+k(k-1)\bar{\rho}_{FF}}} \tag{4.2}$$

onde $\bar{\rho}_{Fc}$ é o grau médio de correlação entre cada característica $F \in S$ e as classes (c), enquanto $\bar{\rho}_{FF}$ é o grau médio de correlação entre cada par de características em S. Enquanto o numerador fornece uma indicação do grau de relevância de uma dado subconjunto de características, o denominador mede o grau de redundância.

A composição do conjunto S foi obtida a partir de uma busca exaustiva composta de dois principais passos. Primeiro, para cada possível subconjunto s com k variáveis, sendo $k \in [1,K]$, M_s é computada e s com o máximo local, M_s, é selecionado. O passo seguinte consiste em selecionar, neste caso entre os doze subconjuntos, aquele que fornece o máximo global,

$$S = \max_{\mathbf{s_i}} [M_{s_1}, M_{s_2} ... M_{s_K}]$$

4.1.4 Construção e teste dos classificadores

Uma vez selecionadas as variáveis de interesse, sete modelos de aprendizado de máquina foram comparados usando o mesmo conjunto de características. Esses classificadores foram treinados e avaliados usando uma abordagem de reamostragem aninhada (*nested resampling*), tal como descrito na Figura 24. A Tabela 2 contém o nome dos modelos utilizados assim como seus parâmetros de entrada.

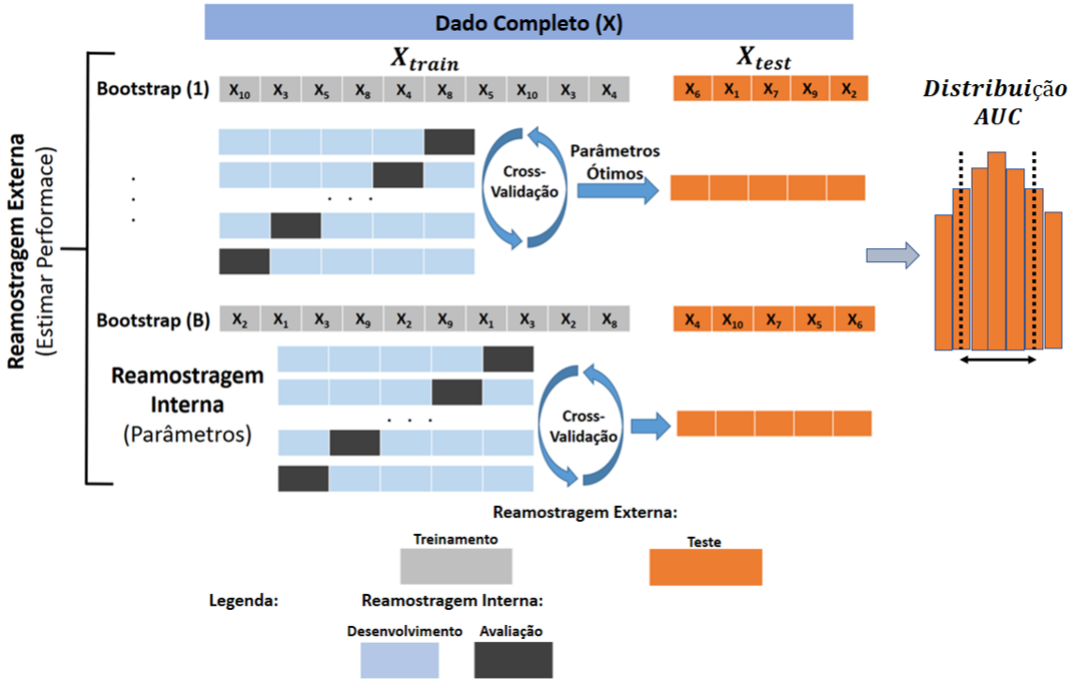
Tabela 2 – Modelos utilizados.

Sigla	Nome do Modelo	Parâmetros
RF	Random Forest	Número de árvores (n_{tree})
LR	Logistic Regression	Fator de regularização (λ)
ErT	Extremly Random Tree	Número de árvores (n_{tree})
GB	Gradient Boosting	taxa de aprendizado (γ)
SVM	Support Vector Machine	Fator de regularização (α) e fator do *kernel*
KNN	K-Nearest Neighbour	Número de vizinhos próximos(K)

Fonte: Autor.

Para um conjunto de casos $X = \{x_i : i = 1,2,...,N\}$, onde x_i é um VOI, descrito por k-características, um amostra de mesmo tamanho, $X^* = \{x_i^* : i = 1,2,...,N\}$ foi gerada usando *bootstrap*. Essa amostra compõe o conjunto de treinamento, denominado X_{train}, e como mostrado em (B. Efron and R. Tibshirani, 1993; GEORGES; MASATO, 2001), e discutido na seção 3.2.4, tal conjunto conterá algo em torno de 63.2% dos casos únicos do conjunto original. Nesta fase, para estimar o parâmetro ótimo de cada classificador, a partir de uma lista de parâmetros, um laço interno divide X_{train} em 5 diferente *folds*, sendo 4 deles usados

Figura 24 – Fluxo para treinar e testar cada modelo usando reamostragem por *bootstrap*.



Reamostragem por bootstrap foi usado para dividir o conjunto de dados em três conjuntos disjuntos. O conjunto dedicado a treinamento, que engloba obtenção dos parâmetros ótimos e avaliação de cada modelo (X_{train}, blocos em cinza) foi subdividido usando *cross-validation*. Nessa etapa, os parâmetros do modelo que forcem a maior performance na reamostragem interna são os selecionados, usando um método chamado *grid search*. Por fim, o outro conjunto para teste ou validação da performance dos mesmos foi obtida usando os dados não vistos (X_{test}, blocos laranjas). Esse processo foi repetido 100 vezes e a performance de cada modelo é medida em termos da distribuição de valores de área da curva ROC.

Fonte: autor

para construir o modelo e 1 *fold* restante é usado para avaliar o desempenho dos parâmetros na construção do modelo aprendido na etapa anterior. Esse procedimento descrito no laço interno na Figura 24 é então repetido (*cross-validation*) e via essa estratégia selecionam-se os parâmetros que fornecem maior desempenho médio durante o treinamento, nesse caso maior área sobre a curva ROC (AUC).

Contudo, tal modelo ótimo precisa ser testado em um conjunto de dados não utilizados durante a fase de treinamento. Assim, para garantir tal condição e checar a capacidade de generalização de cada modelo ante novos casos, o conjunto original de dados que *não aparece* em X_{train} (36.8% restantes), durante a reamostragem por *bootstrap*, chamado X_{test}, foi o conjunto usado na fase de teste.

Por fim, esse processo de reamostragem foi repetido $B = 100$ vezes e a cada vez AUC foi estimada para cada modelo. Desse modo, obtêm-se uma distribuição da estatística de interesse e, baseado nessa distribuição, têm-se uma ideia da performance de cada modelo na tarefa de diferenciar tumores benigno de maligno.

4.1.5 Avaliação dos modelos: teste estatísticos

Para comparar os métodos de classificação utilizaram-se neste trabalho dois critérios: poder discriminativo e estabilidade de predição.

Para avaliar poder discriminativo, medido em termos de AUC, conforme discutido anteriormente, utilizou-se o procedimento sugerido por (JANEZ, 2006), que permite considerar o desempenho de diferentes modelos em múltiplas amostras como um todo. Para isso, primeiro uma matriz ordenada é construída onde cada linha, b, representa uma amostra *bootstrap*, enquanto cada coluna c representa um classificador ou modelo. Nessa matriz, um elemento r_{bc} representa a ordem da performance de cada classificador, c, na amostra b em termos de AUC. Ou seja, para cada linha de r_{bc}, um classificador com a mais alta AUC receberá o valor 1 em sua coluna, o segundo 2, e assim sucessivamente. Em caso de empate, a cada um é atribuído a média entre seu antecessor e posterior. Em seguida, após todos os modelos terem sido ordenados, computa-se uma média e obtêm-se a ordem média de cada modelo. Por fim, teste de *Friedman* seguido por análise *post-hoc* usando teste de *Nemenyi* são usados para determinar qual classificador apresentou maior performance em termos de sua ordem média.

Por outro lado, para testar estabilidade de cada modelo, utilizou-se um outro parâmetro: o coeficiente de variação da distribuição de valores de AUC obtidos por cada modelo, expresso como porcentagem, tal como,

$$CV\% = 100x\frac{\sigma_{AUC}}{\mu_{AUC}} \tag{4.3}$$

onde σ_{AUC} e μ_{AUC} são, respectivamente, o desvio padrão e média dos valores de AUCs estimados em cada X_{test}.

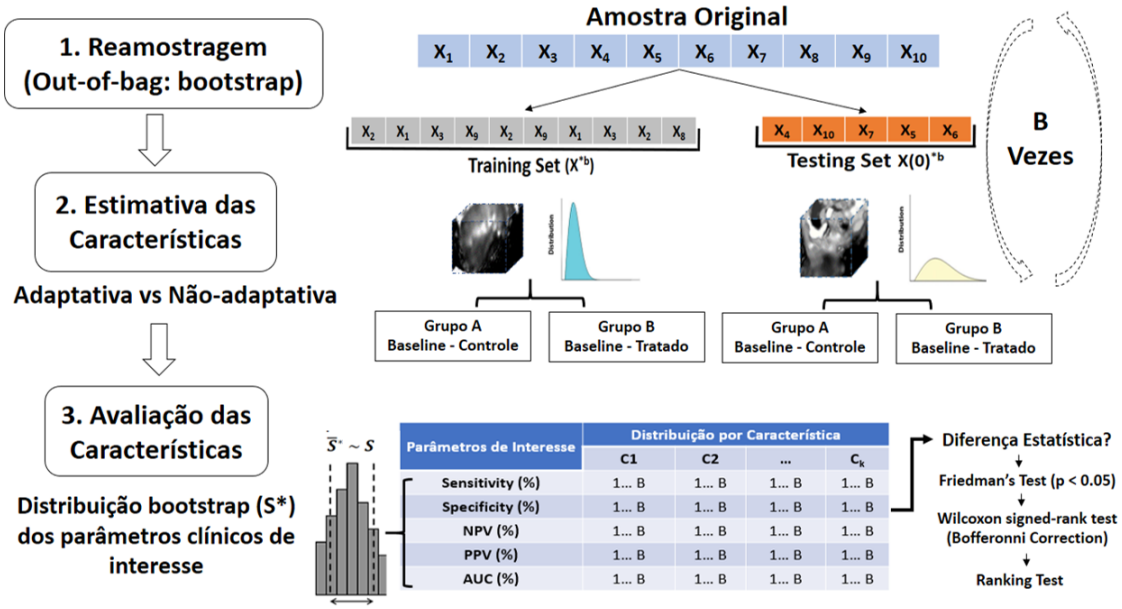
Os modelos que se mostraram em uma faixa aceitável desses critérios de desempenho e estabilidade, foram revalidados usando casos extras, que não fizeram parte do processo descrito anteriormente.

4.2 Caso 2: Câncer Gastrointestinal

Como discutido em (PADHANI et al., 2009; JUST, 2014), características de primeira ordem baseadas no histograma do mapa de ADC têm se mostrado útil em variadas aplicações clínicas. Todavia, até o presente momento, toda a análise tem se baseado no uso tradicional (não-adaptativo) dessas características. Nesse trabalho, um novo método é proposto e faz uso desses mesmos histogramas de maneira adaptativa, levando em conta apenas regiões discriminativas da lesão: nesse caso, regiões que marcadamente sofreram efeito do tratamento aplicado. A Figura 25 exibe os passos principais: (1) obtenção das distribuições de probabilidade e características (não-adaptativas e adaptativas) derivadas das mesmas, (2) uso do *bootstrap* para reamostrar o conjunto de imagens em dois conjuntos disjuntos, um de treinamento e outro de teste, os quais servem de base para comparação do poder diagnóstico de cada uma das características. Por fim,

(3) o processo de análise estatística dos parâmetros de interesse clínico. A seguir será abordado em detalhe cada um desses passos.

Figura 25 – Fluxograma para análise das características adaptativas propostas.



Fonte: autor

4.2.1 Modelo Animal

O modelo de tumor gastrointestinal (GIST) usado nesse estudo retrospectivo foi o GIST-DFR, com transplantes subcutâneos em ambos os lados. Esse modelo foi escolhido por apresentar baixa propensão de desenvolver espontaneamente necrose, ou mesmo processos degenerativos. O tempo levado para o tumor atingir o de $\pm$ 1cm maior diâmetro $[0,75-1,25cm]$ foi algo entre 3-4 meses. Esse procedimento foi realizado no laboratório de oncologia do hospital, que conta com um time com experiência nesse tipo de procedimento invasivo [3].

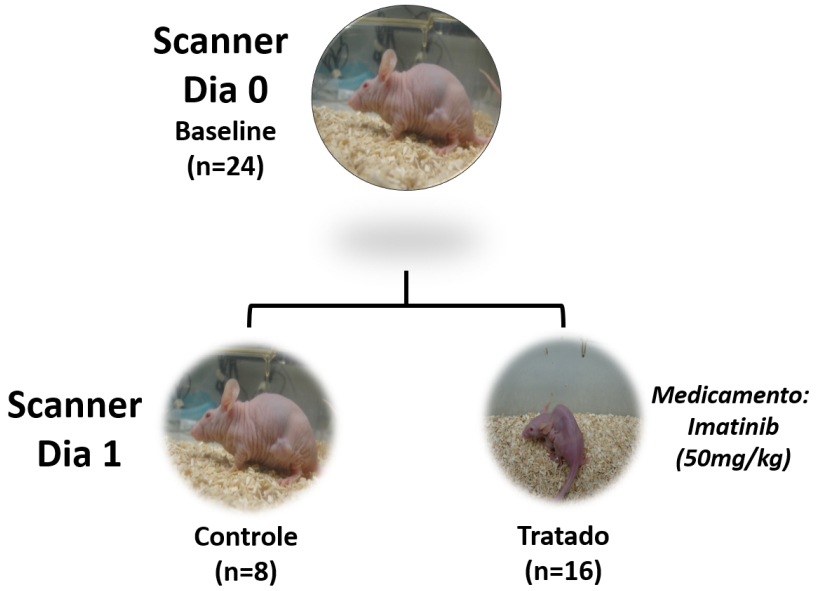
Em seguida à inserção tumoral, 24 animais foram submetidos ao *scanner* de 9.4 Tesla especialmente projetado para pequenos animais, gerando o conjunto *baseline* (Dia 0, Figura 26). Desses 24 animais, 8 foram aleatoriamente sorteados para formar o grupo controle, enquanto 16 foram sorteados para receber o tratamento com administração de *Imatinib* $50mg/kg$, duas vezes por dia.

Um dia após administração do tratamento, o *scanner* MRI foi conduzido. Em ambos os dias (Figura 26), sequências anatômicas (T2) e DWI usando 10 $b-values$ (0, 50, 100, 150, 200, 250, 300, 500, 750, e 1000 s/mm^2) em 3 direções (x, y, e z) foram obtidas. A fim de

[3] Esse experimento foi realizado com o devido consentimento do comitê de ética do hospital que segue regras definidas pela União Europeia.

reduzir suscetibilidade a artefatos, todos os camundongos foram *escaneados* conforme método desenvolvido por (CHEN et al., 2013).

Figura 26 – Fluxograma do modelo GIST.



Para calcular o mapa de ADC, o primeiro passo foi medir a intensidade do sinal (SI) a partir das imagens de difusão (DWI) originais com seus respectivos $b-values$. Em resumo, para cada animal, delineações semi-automáticas foram realizadas nos cortes transversais das DWI com $b-value$ igual a 1000 s/mm^2. Essas delineações foram então agrupadas para formar um volume 3D de interesse (VOI) por lesão. Esse VOI foi então automaticamente copiado para todas as outras imagens com diferente $b-values$.

Em seguida, o mapa ADC foi estimado pelo modelo mono-exponencial com todos os 10 $b-values$, usando o método dos mínimos quadrados tal como em (VANDECAVEYE et al., 2009).

4.2.2 Reamostragem

Dada uma amostra composta por um conjunto de VOIs, x_i, descrita por $X = \{x_i : i = 1,2,...,N\}$, uma amostra de mesmo tamanho, $X^* = \{x_i^* : i = 1,2,...,N\}$ foi gerada usando reamostragem com reposição, ou *bootstrap*. O conjunto de vetores de dados que não aparecem em X^* é denotado como $X^*(0)$. A geração aleatória de uma larga quantidade de amostras (B vezes), X_b^* para b= 1,2,...,B, pode ser usada para estimar a quantidade de interesse, no caso do presente trabalho, o poder discriminativo de um conjunto de características a fim de medir o efeito de um determinado tratamento. Tal avaliação é realizada ao se utilizar o processo descrito anteriormente no qual dois conjuntos disjuntos são criados, um denominado conjunto treinamento, S_{train}, composto por aproximadamente 63.2% dos dados, e outro conjunto teste, S_{test}, contendo os 36.8% restantes (Figura 25). Além disso, na fase de treinamento, para cada

amostra $X^*(b)$ gerada, a probabilidade de ocorrência de dada classe foi feita igual a ocorrência da outra classe, um procedimento conhecido como amostragem *bootstrap* baseada na instância (ou classe) (GEORGES; MASATO, 2001).

4.2.3 Estimativa das características

Não-adaptativa: ADC diferença relativa ao tratamento

Visando avaliar o efeito do tratamento, ou seja, medir a diferença entre os grupos controle e tratado com relação a um conjunto referência (*baseline*), uma diferença relativa entre casos foi computada.

Para cada caso j, $k = 103$ características foram estimadas quer a partir da geometria, ou forma dos VOIs (16 variáveis), quer a partir das distribuições de probabilidade (DP) do mapa de intensidades ADC, $P_n(i|c_j)$. Todavia, embora para cada característica a distribuição $P_n(i|c_j)$ seja estimada de diferentes formas - e aqui seguiram-se os trabalhos de (AERTS et al., 2014; GROVE et al., 2015; ZHANG et al., 2017)- todas essas características igualmente assumem que informações sobre aspectos da textura das imagens em análise estão contidas em $P_n(i|c_j)$. Para estimar cada distribuição associada a sua correspondente característica, utilizou-se *kernels* específicos com *bandwidths* escolhidos pelo método de *Scott* (SCOTT, 1992).

Desse modo, dado o *baseline* (F_{bas}) como referência, a diferença relativa ($\Delta F, 0 \leq \Delta F \leq 1$) descrita por cada uma dessas características (F) foi computada tal como:

$$\Delta F = \frac{F - F_{bas}}{F_{bas}} \tag{4.4}$$

Em seguida, testes de MW foram aplicados e apenas aquelas variáveis estatisticamente significativas foram utilizadas no passo seguinte. Nesta etapa, o poder de cada variável em discriminar os grupos controles e tratados foram avaliados como se segue (Figura 27): primeiramente um limiar ótimo, medindo a melhor separação univariada possível, é obtido no conjunto S_{train} e testado no conjunto disjunto S_{test}.Esse processo foi repetido $B = 1000$ vezes e uma distribuição de parâmetros de interesse derivados da curva ROC foram estimados, quais sejam: AUC, sensitividade, especificidade, PPV (positive predictive value) e NPV (negative predictive value).

A Tabela 3 sumariza os 103 descritores utilizados nessa etapa. Cada uma delas encontra-se descrita no seguinte documento publicado recentemente (ZWANENBURG et al., 2016). Todavia, nenhuma dessas características leva em conta se a região do mapa de ADC são discriminativas ou não. Tal procedimento recebe o nome de não-adaptativo. A seguir o método adaptativo que explora apenas as regiões discriminativas da imagem (análise intratumoral) será abordado.

Figura 27 – Obtenção do limiar não-adaptativo ótimo e geração dos parâmetros de interesse para discriminar os grupos controle e tratado.

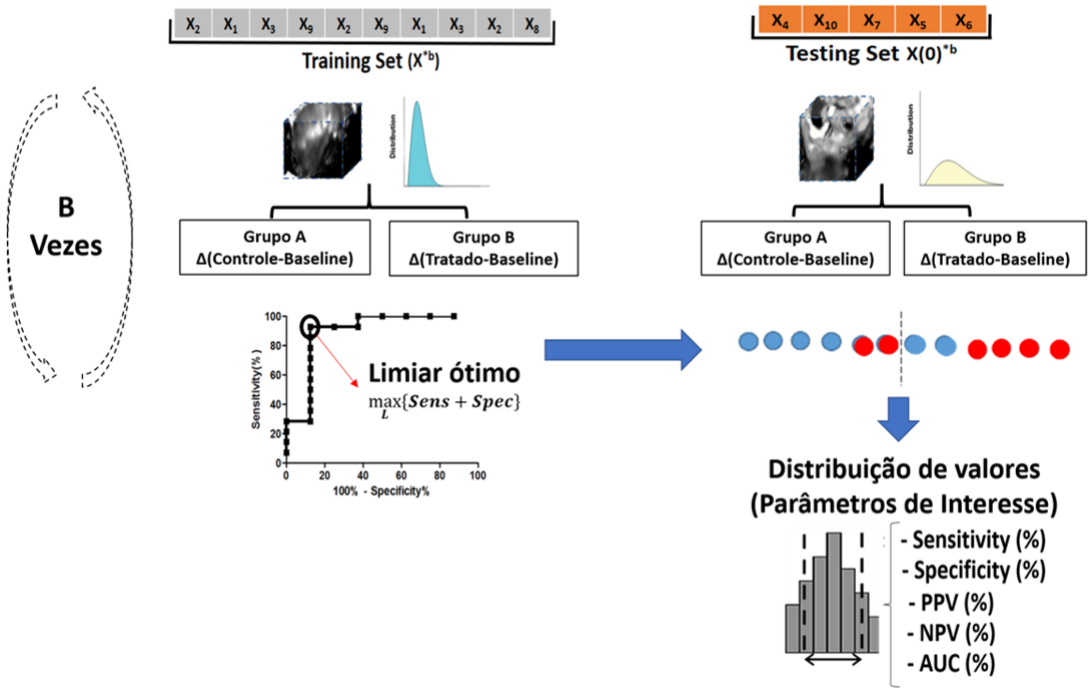


Tabela 3 – Descritores não-adaptativas usados nesse trabalho: (i) ADC histograma, (ii) Geometria, (iii) GLCM, (iv) GLSZM, (v)GLRLM, (vi) NGTDM e (vii)GLDM. Maiores informações algorítmicas sobre implementação de tais métodos podem ser encontrados em (ZWANENBURG et al., 2016).

Tipo	Sigla	Nome
(i)	ΔSkewness	Simetria
	ΔKurtosis	Curtose
	ΔMaximum	Máxima intensidade
	Δ90Perc	90th Percentil
	ΔMedian	Mediana
	ΔStddDeviation	Desvio Padrão
	ΔVariance	Variança
	ΔMinimum	Mínima intensidade
	ΔMean	Média
	ΔRootMeanSq	Raiz média quadrática
	ΔRange	Intervalo
	ΔEnergy	Energia
	ΔTotalEnergy	Energia Total

Continua na próxima página

Tabela 3 – continuação da página anterior

Tipo	Sigla	Nome
	ΔRobustMeanAbsDev	Média da desvio absoluto
	ΔEntropy	Entropis
	Δ10Percentile	10th Percentil
	ΔInterquartileRange	Intervalo interquartil
	ΔUniformity	Uniformidade
	ΔMeanAbsDeviation	Desvio absoluto médio
(ii)	ΔSurfArea	Área da superfície
	ΔMinorAxis	Eixo-menor
	ΔElongation	Razão dos eixos principais
	ΔCompactness1	Compacidade1
	ΔCompactness2	Compacidade1
	ΔSpherDisprop	Desvio da Esfera
	ΔSphericity	Esfericidade
	ΔLeastAxis	Menor eixo principal
	ΔMax2DDiamRow	Máx. diâmetro 2D
	ΔSurfVolRatio	Densidade superficial
	ΔVolume	Volume
	ΔMajorAxis	Eixo-maior
	ΔMax3DDiam	Máx. diâmetro 3D
	ΔFlatness	Planicidade
	ΔMax2DDiamColumn	Máx. distância entre os voxels da superfície (plano coronal)
	ΔMax2DDiamSlice	Máx. distância entre os voxels da superfície (plano sagital)
(iii)	ΔClusterProminence	
	ΔClusterShade	Simetria da matriz GLCM
	ΔJointEntropy	Entropia da GLCM
	ΔClusterTendency	Tendência de Variabilidade
	ΔDifferVariance	Variância da diferença
	ΔSumEntropy	Entropia da soma dos vizinhos
	ΔIdmn	Inverso do momento diferença normalizado
	ΔMaxProbability	Predominante par de intensidade
	ΔSumSquares	Variância
	ΔDifferEntropy	Entropia da diferença dos vizinhos
	ΔSumAverage	Média da soma dos vizinhos
	ΔJointEnergy	Energia
	ΔJointAverage	Média do GLCM

Continua na próxima página

Tabela 3 – continuação da página anterior

Tipo	Sigla	Nome
	Δ Imc2	Correlação exponencial
	ΔAutocorrelation	Fineza da textura
	ΔIdn	Inverso diferença normalizado
	ΔContrast	Contraste
	ΔImc1	Correlação
	ΔHomogeneity2	Uniformidade
	ΔHomogeneity1	Uniformidade
	ΔDissimilarity	Dissimilaridade
	Δ Diff.Average	Média da diferença entre vizinhos
	ΔIdm	Inverso momento diferença
	ΔId	Inverso diferença
	ΔInv.Variance	Inverso variança
	ΔCorrelation Correlação	
(iv)	ΔGLVar	Variância nível de cinza (nc) na zona
	ΔSALGLE	Áreas de Nível de Cinza Baixa
	ΔLAHGLE	Ênfase área de nível de cinza alto
	ΔGLNUN	Nível de cinza não uniformizado normalizado
	ΔZP	Porcentagem da Zona
	ΔSAE	Ênfase de Área Pequena
	ΔSZNU	Zone de tamanho não uniforme
	ΔSZNUN	Zone de tamanho não uniforme normalizado
	ΔZVar	Variância da Zona
	ΔHGLZE	Ênfase de Nível nível de cinza alto
	ΔSAHGLE	Ênfase de nível de cinza alto de área pequena
	ΔGLNU	Nível de Cinza não-uniforme
	ΔLAE	Ênfase de área grande
	ΔLALGLE	ênfase de nível de baixo nível de área baixa
	ΔLGLZE	Ênfase da Zona de Nível Baixo Grau
	ΔZEntrop	Zone Entropy
(v)	ΔGLVar	Variância de nível de cinza
	ΔRunEntropy	Entropia of GLRLM matriz
	ΔGLNUN	nível de cinza não uniformizado normalizado
	ΔLRLGLE	Ênfase de baixo nível de longa corrida
	ΔSRE:	Ênfase de corrida curta
	ΔLGLRE	Ênfase de baixo nível de longa corrida

Continua na próxima página

Tabela 3 – continuação da página anterior

Tipo	Sigla	Nome
	ΔHGLRE	Ênfase de auto nível de longa corrida
	ΔLRHGLE	Ênfase de alto nível de alto nível de cinza
	ΔRunPercentage	Rugosidade
	ΔGLNU	nível de cinza não uniforme
	ΔRunVariance	Variância da GLRLM
	ΔSRLGLE	corrida curta de ênfase de baixo nível de cinza
	ΔRLNUN	Não-uniformidade normalizada
	ΔSRHGLE	
	ΔLRE	Ênfase de longo prazo
	ΔRLNU	Uniformidade
(vi)	ΔStrength	
	ΔComplexity	Complexidade
	ΔCoarseness	Fineza
	ΔContrast	Contraste
	ΔBusyness	Aspereza
(vii)	ΔDVar	
	ΔGLVar	Variância de nível de cinza
	ΔSDHGLE	Baixa dependência de alto nível de cinza
	ΔDNUN	Dependência não-uniformizado normalizado
	ΔLDLGLE	Baixa dependência de baixo nível de cinza
	ΔSDE	Pequena ênfase de dependência
	ΔGLNU	nível de cinza não-uniforme
	ΔLDHGLE	Baixa dependência de alto nível de cinza
	ΔLGLE	Ênfase de nível de cinza baixo
	ΔSDLGLE	Pequena dependência de ênfase de baixo nível de cinza
	ΔDNU	Dependência não-uniforme normalizada
	ΔHGLE	Ênfase de nível de cinza alto
	ΔGLNUN	Nível de cinza não-uniforme normalizado
	ΔDEntrop	Dependência Entropia
	ΔLDE	Grande ênfase de dependência

Características Adaptativas: análise intratumoral

Por assumir que subregiões do VOI, em um sistema quase sempre heterogêneo, contribuem da mesma forma para discriminar diferentes grupos (tumor vs não-tumor, controle vs

tratado, etc), os métodos de extração de características tradicionais (não-adaptativos) descritos anteriormente podem conduzir a alto nível de interseção e, consequentemente, a baixo valor de discriminação.

Como uma alternativa a tal limite, um método adaptativo é desenvolvido aqui. Por essa abordagem, regiões discriminativas podem ser automaticamente extraídas usando a diferença e a distância entre a distribuição de probabilidade representando cada grupo.

Para tal, dado um conjunto de VOIs pertencentes a c_j grupos, computa-se inicialmente para cada VOI em S_{train} a DP condicional $P_n(i|c_j)$, $n = 1,2,...,N(c_j)$. Baseado nessa distribuição condicional, uma distribuição média para cada grupo, c_j, com um número total de casos $N(c_j)$, é obtida usando:

$$\bar{P}(i|c_j) = \frac{1}{N(c_j)} \sum_{n=1}^{N(c_j)} P_n(i|c_j)$$

Baseado nisso, a diferença entre as duas classes são computadas para cada grupo:

$$\Delta_P(i|c_1,c_2) = \bar{P}(i|c_2) - \bar{P}(i|c_1) \tag{4.5}$$

onde o grupo A é composto pelos pares *baseline*(c_1)-controle(c_2) e o grupo B por *baseline*($c_1^{'}$)-tratado($c_2^{'}$). Por simplicidade, daqui em diante, esses grupos serão chamados controle e tratado, uma vez elas são as classes de interesse.

Por fim, a variância $\sigma_P^2(i|c_j)$ associada aquela média é também estimada, assim como a distância $J_P(i|c_j)$ entre cada grupo c_1 e c_2:

$$\sigma_P^2(i|c_j) = \frac{1}{N(c_j)-1} \sum_{n=1}^{N(c_j)} \left(P_n(i|c_j) - \bar{P}(i|c_j)\right)^2$$

$$J_P(i|c_j) = \left[2\frac{[\Delta_P(i|c_1,c_2)]^2}{\sigma_P^2(i|c_1) + \sigma_P^2(i|c_2)}\right]^{\frac{1}{2}} \tag{4.6}$$

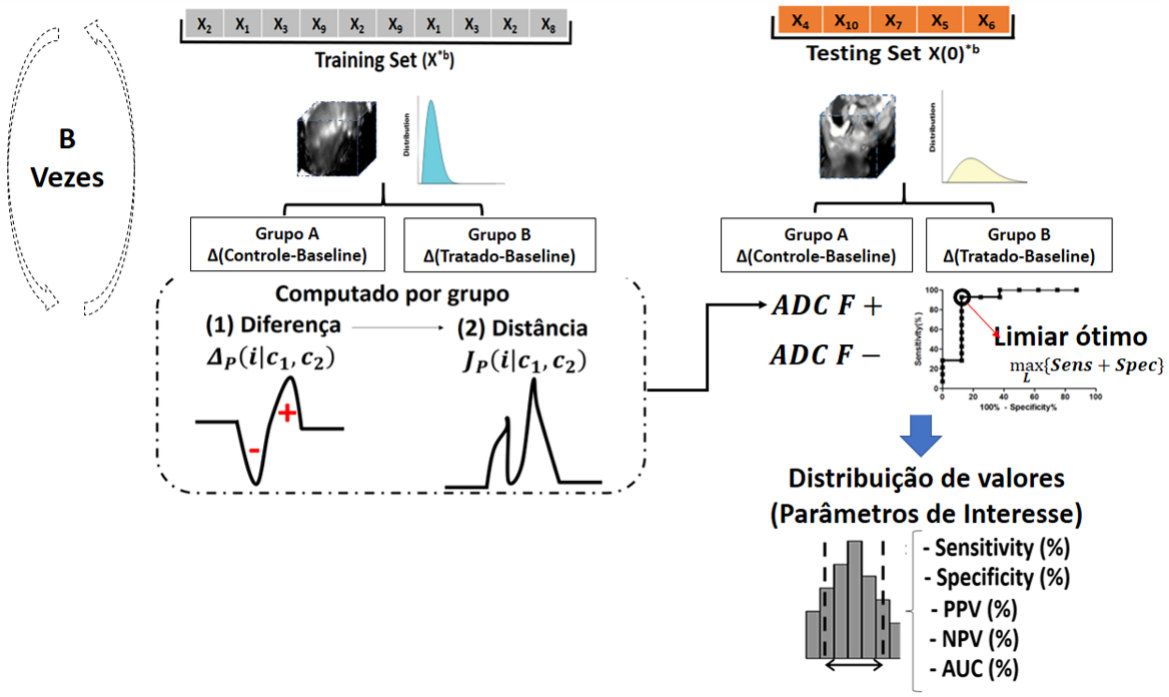
Seguindo esse raciocínio, ao invés de computar um alto número de características para cada VOI, os quais normalmente apresentam alto nível de redundância, o espaço de característica aqui passa a ser reduzido para duas variáveis por cada caso. Para isso, essas duas variáveis adaptativas usam as partes disjuntas, positiva e negativa, de $\Delta_P(i|c_1,c_2)$ como domínio das seguintes somas ponderadas:

$$ADCF_+ = \sum_{\Delta_P(i|c_1,c_2)\geq 0} P_j(i|c_j)\left[J_P(i|c_j)\right]^2 \tag{4.7}$$

$$ADCF_- = \sum_{\Delta_P(i|c_1,c_2)< 0} P_j(i|c_j)\left[J_P(i|c_j)\right]^2 \tag{4.8}$$

Contudo, note que, uma vez as diferenças (Equação 4.5) e distância (Equação 4.6) entre as classes são estimadas em S_{train}, as texturas adaptativas ADC F_+ e ADC F_- são computadas em S_{test}. A Figura 28 sumariza todos esses passos.

Figura 28 – Obtenção da diferença e distância entre grupos para estimativa de $ADCF+$ e $ADCF-$.



A principal diferença do método adaptativo aqui descrito para os métodos de análise de textura não-adaptativos é que, ao invés de usar $P_j(i|c_j)$ como fator de peso para computar indiretamente as texturas, $P_j(i|c_j)$ é usada como variável de entrada e o fator de peso passa a ser o quadrado da distância entre as distribuições de probabilidades que representam cada classe, $J_P(i|c_j)$. Nesse sentido, assume-se aqui que $J_P(i|c_j)$ será maior em regiões do VOI nas quais a diferenciação entre os grupos tratados e controle são proeminentes, comparado ao *baseline*. Isso significa poder fazer automaticamente o que os médicos já fazem visualmente, que é pesquisar sub-regiões da imagem onde um determinado tratamento se fez evidente, sobretudo no caso de desenvolvimento de novas drogas/tratamentos.

Além disso, enquanto ADC F_+ mede as regiões do mapa ADC discriminativas nas quais os valores de intensidade apresentam maior probabilidade de ser em média mais altos do que os valores do *baseline* (o que na prática indica redução da quantidade de células (Koh, DM, 2010), refletindo efeito do tratamento), ADC F_- mede o contrário, regiões com maior probabilidade desses valores de ADC serem menores que os observados no *baseline*.

Análise Estatística das Características

Para identificar nos estágios iniciais aqueles pacientes com maior verossimilhança de não responder ao tratamento, o poder discriminativo dos grupos de características adaptativa e não-adaptativa foram avaliados sobre mesmas condições. Para isso, as estatísticas de interesse foram estimadas nas amostras *bootstrap* (Figuras 27 e 28) usando parâmetros derivados da curva ROC, ou seja, baseados na sensitividade e especificidade, incluindo positivo e negativo valor preditivo

(PPV e NPV), assim como área sobre a curva ROC propriamente dita (AUC).

Após estimar os desempenhos de cada característica nas múltiplas amostras, obtêm-se uma distribuição de valores representando cada uma deles. Como método de comparação, utilizamos o método sugerido por (JANEZ, 2006), similarmente ao utilizado no caso 1 para comparar os classificadores no diagnóstico de linfonodos, uma vez que tal método permite comparar o desempenho de cada característica sobre as múltiplas amostras como um todo. Para isso, primeiramente uma matriz ordenada (*ranking matrix*) é criada contendo as amostras *bootstraps* como linhas e as características como colunas, para cada parâmetro de interesse (PPV, NPV, etc). Ou seja, em tal matriz, cada elemento r_{bf} representa a ordem de uma dada medida de desempenho de interesse (ex. PPV) fornecido por uma característica x^k na amostra b. Tal ordem para cada linha b é computada como se segue: o mais alto desempenho é atribuído o valor 1, ao segundo o valor 2 e assim sucessivamente. Em seguida, após cada coluna x^k ter recebido sua posição em cada uma das amostras b, um vetor média é estimado para cada característica, ao longo das amostras. Por fim, um teste estatístico para comparações múltiplas é realizado para verificar com 95% de confiança quais características apresentaram performance superior a suas contrapartes considerando seus *rank* médios. Tal teste estatístico é baseado em *Friedman* teste com *post-hoc* análise usando *Bonferroni-Dunn* teste. Dessa forma, cada características de interesse é comparada com ADC médio (referência usada na prática clínica), e considerada significativamente diferente se sua ordem média difere por pelo menos uma diferença crítica (CD) (JANEZ, 2006).

Por outra lado, além do desempenho, estabilidade também foi analisada, expressa em termos da CV(%) e definida como,

$$CV\% = 100x\frac{\sigma_{perf}}{\mu_{perf}} \tag{4.9}$$

onde μ_{perf} e σ_{perf} significam, respectivamente, média e desvio padrão dos valores da performance de interesse para cada característica.

Assim, em ordem de identificar a variável com maior performance e estabilidade, a mediana dos valores de cada performance (Θ) e estabilidade (θ) foram usadas como critérios de seleção de característica. Dessa forma, características, x^k, com performance (PPV, NPV, AUC, etc) $> \Theta$ e CV $\leq \theta$ são consideradas de alta performance e estabilidade.

5 RESULTADOS E DISCUSSÃO

5.1 Caso 1

Análise e seleção das Características

A maioria dos estudos na literatura adotam apenas ADC médio como parâmetro quantitativo em DW-MRI (KUANG et al., 2013; PAYNE et al., 2010). Além disso, nos estudos nos quais usa-se toda a lesão nas estimativas das características do histograma do VOI, nota-se que tais análises consideram essas variáveis individual e independentemente (MCVEIGH et al., 2008; DOWNEY et al., 2013; XUE et al., 2014).

Nessa perspectiva, considerou-se aqui primeiramente uma abordagem univariada para a diferenciação de linfonodos e estendeu-se o conjunto de parâmetros padrões (F1-F9) com três variáveis extras baseadas no histograma acumulativo volume-intensidade (CAVH: F10,F11 e F12; Tabela 4) (NAQA et al., 2009).

Essa análise univariada está sintetizada nas Figuras 29 e 30. As variáveis F6 (assimetria), F2 (Média), F3 (Mediana), F5 (curtose) (Figura 29) e todas as variáveis baseadas no CAVH (F10, F11 e F12) (Figura 30) demostraram diferença significativa ($p < 0.05$, MW teste) entre linfonodos da pelve maligno (M) e benigno (B), o que não foi o caso para F1(volume), F7 (coeficiente de variação do ADC histograma), F8(entropia) e F9(energia) (Tabela 4). Contudo, entre as variáveis em destaque, sabe-se que F2 e F3 falham ao descrever regiões heterogêneas quando consideradas univariadamente (PADHANI et al., 2009), o que exige uma abordagem complementar.

Tabela 4 – Características utilizadas: sumário estatístico.

Característica	Benigno		Maligno	
	Mediana	95% IC	Mediana	95% IC
F1: Volume	810.24	[763.43;955.95]	878.91	[592.62;6593.94]
*F2: Média	116.44	[108.88;154.37]	96.36	[89.5;101.99]
*F3: Mediana	117.00	[113.92;124.81]	89.50	[84.9;97.33]
*F4: Desvio padrão	35.30	[36.52;43.08]	30.78	[27.8;34.56]
*F5: Curtose	2.63	[2.58;2.85]	3.36	[3.49;5.23]
*F6: Assimetria	-0.05	[-0.13;0.03]	0.63	[0.64;1.03]
F7: Coef. de Var.	0.31	[0.31;0.37]	0.32	[0.29;0.37]
F8: Entropia	6.49	[6.36;6.57]	6.33	[6.04;6.42]
F9: Energia	0.01	[0.01;0.02]	0.01	[0.01;0.02]
*F10: AUC-CAVH	86.42	[84.26;95.69]	63.00	[56.6;72.36]
*F11:I_{10-90}-CAVH	86.42	[84.26;95.69]	63.00	[56.6;72.36]
*F12:V_{10-90}-CAVH	0.93	[0.9;0.93]	0.97	[0.94;0.97]

Visão geral das diferentes características extraídas do histograma simples e acumulado (CAVH) do mapa de ADC. Média e 95% intervalo de confidência (IC). *Significantemente diferente entre linfonodos benigno e maligno da pelvis (MW, $p < 0.05$).

Figura 29 – (a) Comportamento geral das características normalizadas ($\mu = 0, \sigma = 1$). Análise univariada (box-plots). Variáveis com (*) apresentaram diferença significativa segundo MW teste ($p < 0.05$).

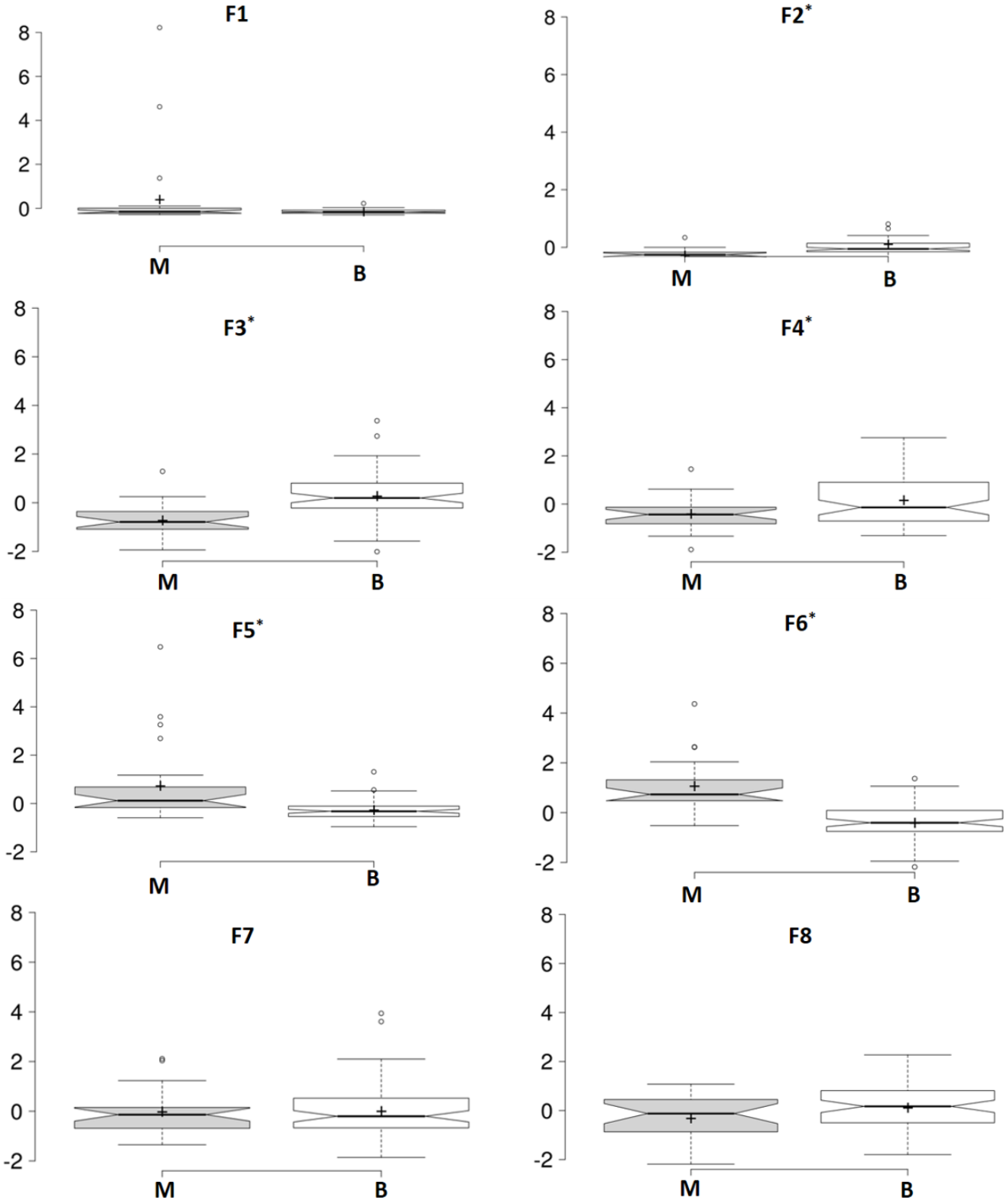
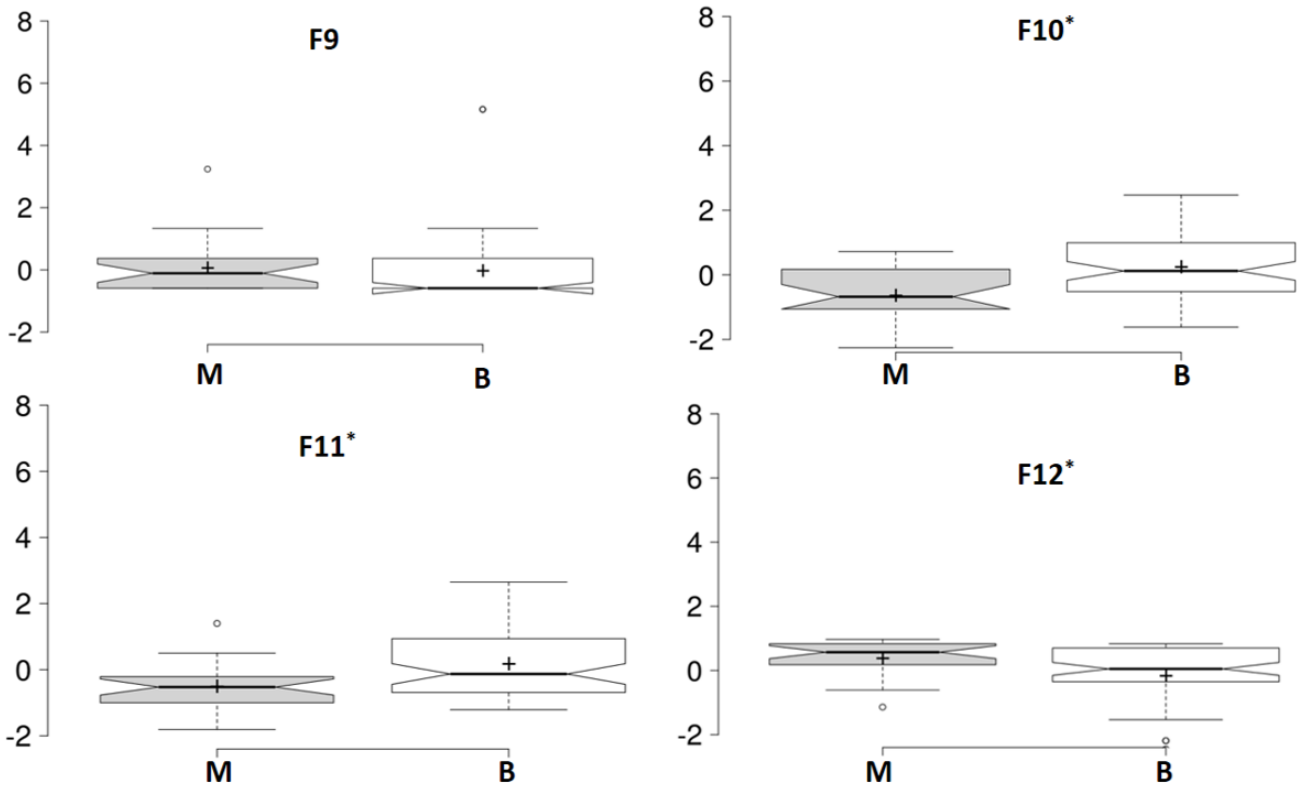


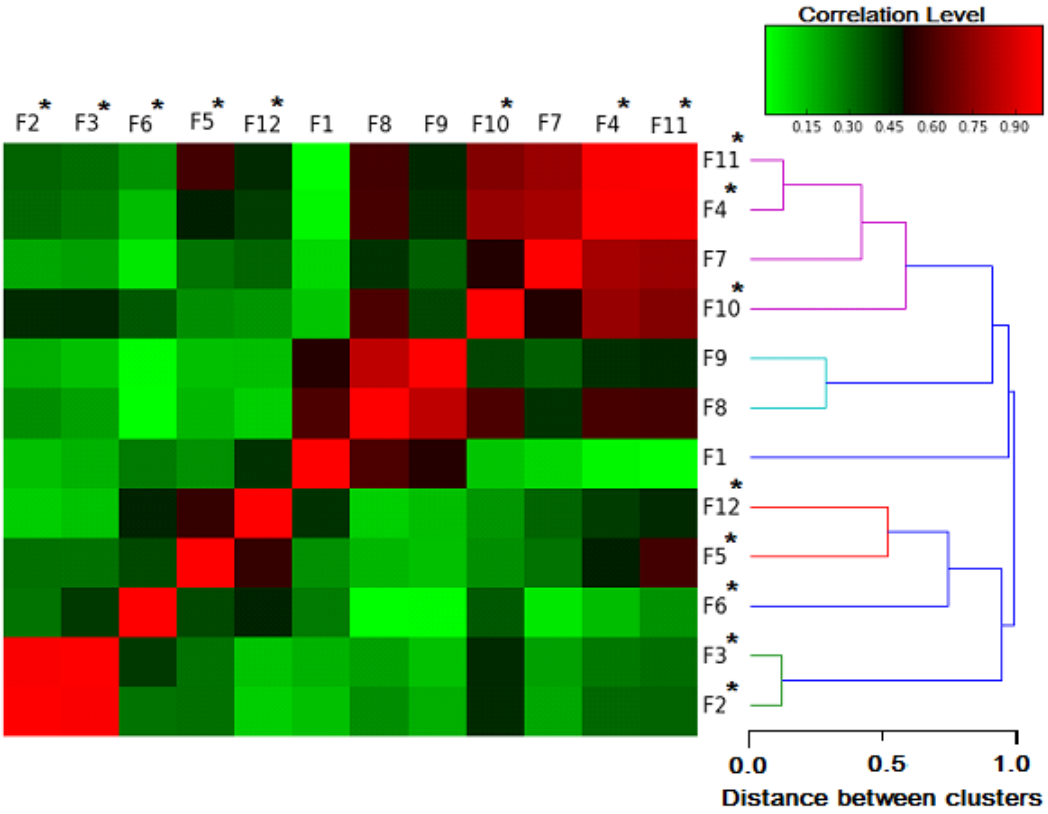
Gráfico de caixa mostram distribuição dos grupos maligno e benigno para cada característica (F1-F12). Os entalhes representam intervalo de confiança (95%) para a mediana (linha horizontal preta) e são definidos como Mediana $\pm$ 1.58*IQR/$\sqrt{N}$, onde IQR é o intervalo interquartílico e N o número de pontos. (*) Indica diferença significativa entre os grupos maligno (M) e benigno (B) segundo Mann Whitney (MW) teste ($p < 0.05$). A cruz em cada gráfico indica o valor médio da distribuição.

Figura 30 – (b) Continuação



Assim, considerando que esses achados univariados não fornecem informação suficiente para determinar o quanto essas características se comportam de modo similar (ou não) em um contexto multivariado, maiores análises se fizeram necessárias. Desse modo, em termos de correlação entre os pares de características (ρ), a matriz de correlação com os correspondentes *clusters* é exibida na Figura 31.

Figura 31 – Comportamento geral das características normalizadas ($\mu = 0, \sigma = 1$). Análise de *clusters* baseados no grau de correlação.



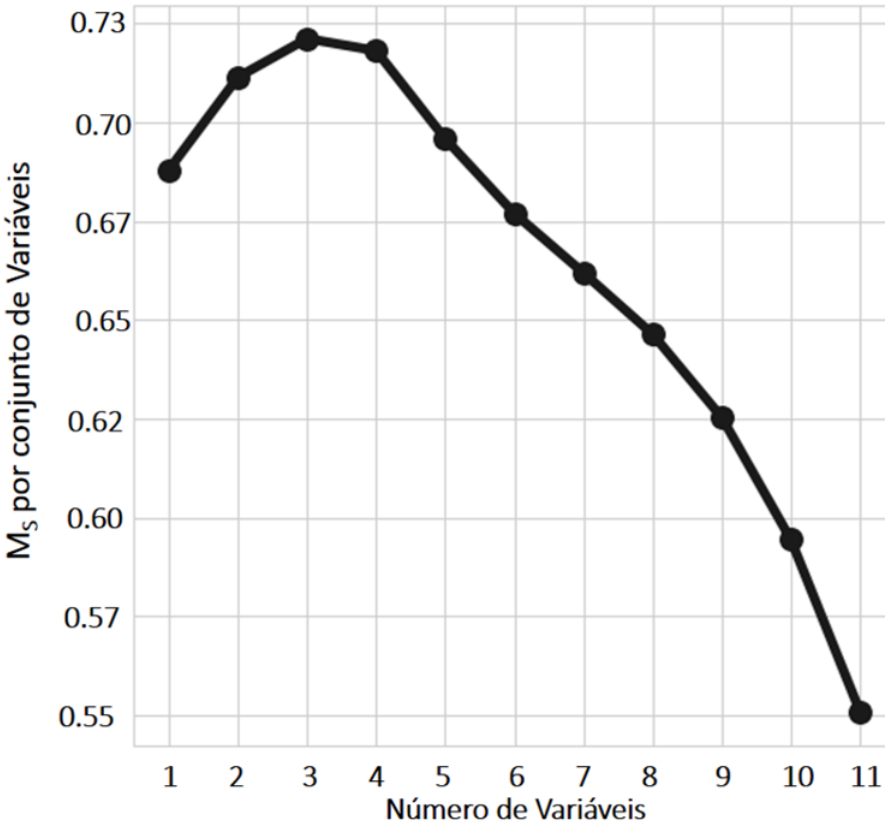
variáveis com (*) apresentaram diferença significativa segundo MW teste ($p < 0.05$).

Especialmente para as três extras variáveis baseadas na CAVH, F10 apresentou mais alta correlação com F4 ($\rho = 0.77$) e F11 ($\rho = 0.72$), enquanto F12 mostrou mais alta correlação com F5 ($\rho = 0.56$). Além disso, considerando o grau de correlação entre cada características, a princípio pode-se dizer que, o grau de correlação entre F2 e F3 (ou F4 e F11) indica comportamento similar e um deles poderia ser excluídos sem perda de informação. Todavia, permanece difícil estimar o grau de correlação necessário para exclusão de características correlacionadas.

Tendo esse limite em mente, em seguida, diferentes conjuntos de variáveis foram analisados levando em conta adicionalmente seus graus de correlação com as classes (alvo). A Figura 32 mostra o comportamento do conjunto de variáveis, medida em termos do mérito de um conjunto de variáveis M_s (Equação 4.2). Tal comportamento reflete a razão entre $\bar{\rho}_{Fc}$ (correlação variável-alvo) e $\bar{\rho}_{FF}$ (correlação variável-variável) ao longo de todas as possíveis dimensões do espaço de características. A curva apresentou um crescimento até atingir um platô em torno de $k = 3$ e $k = 4$ variáveis, compostos respectivamente por $S_3 = \{F6, F3, F2\}$ e $S_4 = \{F6, F3, F2, F5\}$, a partir da qual decresce drasticamente para valores $k > 6$, quando se observa que a correlação entre as variáveis ($\bar{\rho}_{FF}$) aumenta mais rapidamente que a correlação entre estas e o alvo ($\bar{\rho}_{Fc}$). Além disso, considerando que a diferença relativa de M_s entre S_2 e S_3 foi

menor que 2%, optou-se pelo modelo mais simples e maiores análises dos classificadores foram realizadas usando duas variáveis ($S_2 = \{F6, F3\}$) sem perda de informação. Esses conjuntos ótimos achados são compostos por características que também apresentaram univariadamente maior poder discriminativo em estudos anteriores, como pode ser visto em (MCVEIGH et al., 2008; PADHANI et al., 2009; JUST, 2014).

Figura 32 – Mérito por número de característica.



Porém, embora haja na literatura uma variedade de métodos para seleção de características (THEODORIDIS; KOUTROUMBAS, 2008), estudos têm mostrado que a escolha do algoritmo de classificação é ainda mais crítico que a maneira pela qual as características são selecionadas, uma vez que todos estes métodos de seleção tenderão a convergir para a seleção do conjunto de variáveis que melhor caracterizam as classes em estudo (PARMAR et al., 2015). Por esse motivo, um método de seleção de características, independente do modelo classificador, foi utilizado no presente trabalho afim de garantir a generalização de cada etapa (seleção e modelagem) e portanto as análises como um todo.

Vale salientar aqui ainda que, diferente da abordagem padrão em imagens por difusão (MCVEIGH et al., 2008; PAYNE et al., 2010; KUANG et al., 2013; DOWNEY et al., 2013; XUE et al., 2014), neste trabalho diferentes modelos de aprendizagem foram comparados considerando combinações das características de interesse. Como critério de comparação, utilizou-se desempenho e estabilidade de classificação em termos de AUC, conforme abordado a seguir.

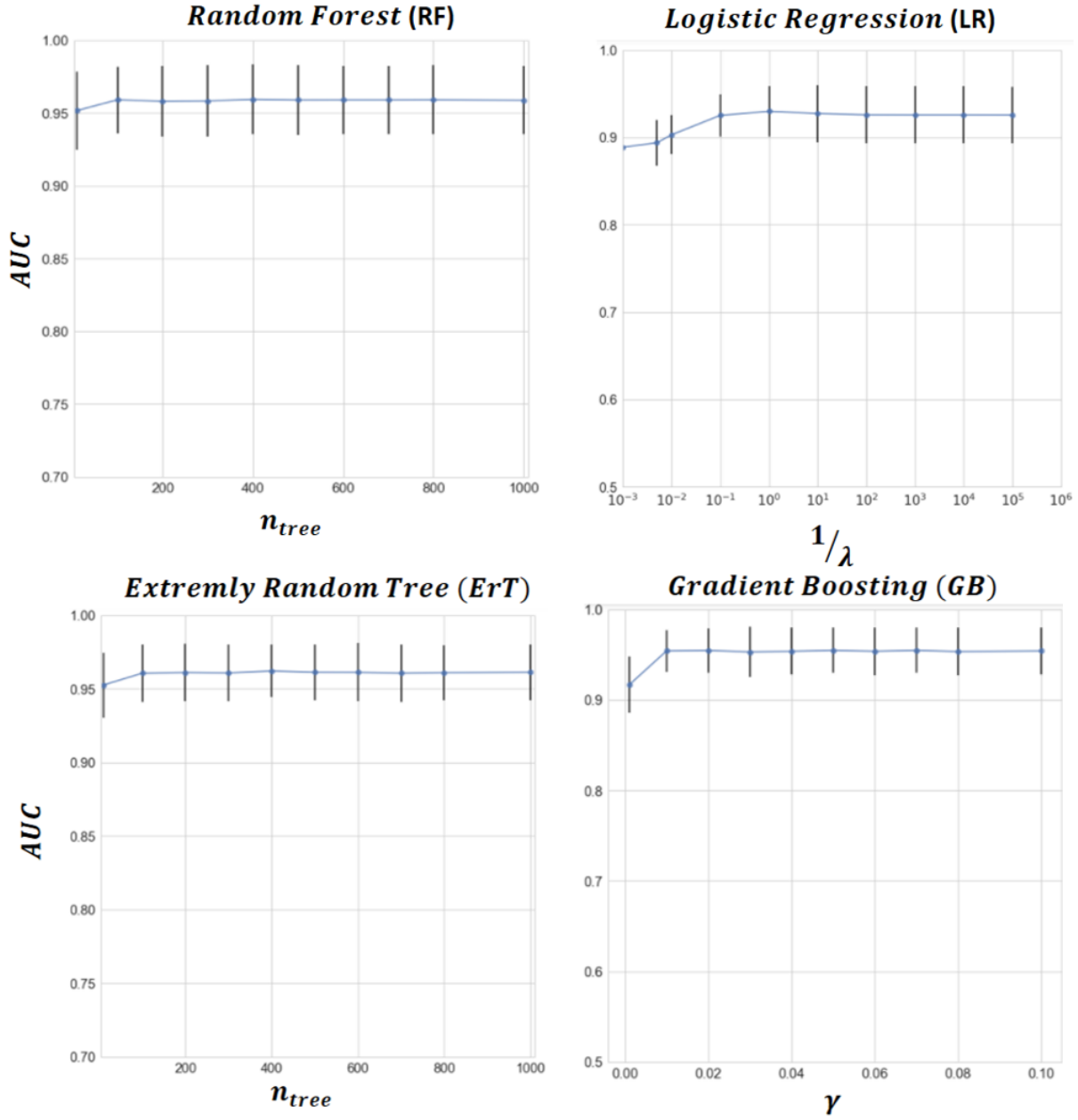
Performance e estabilidade dos classificadores

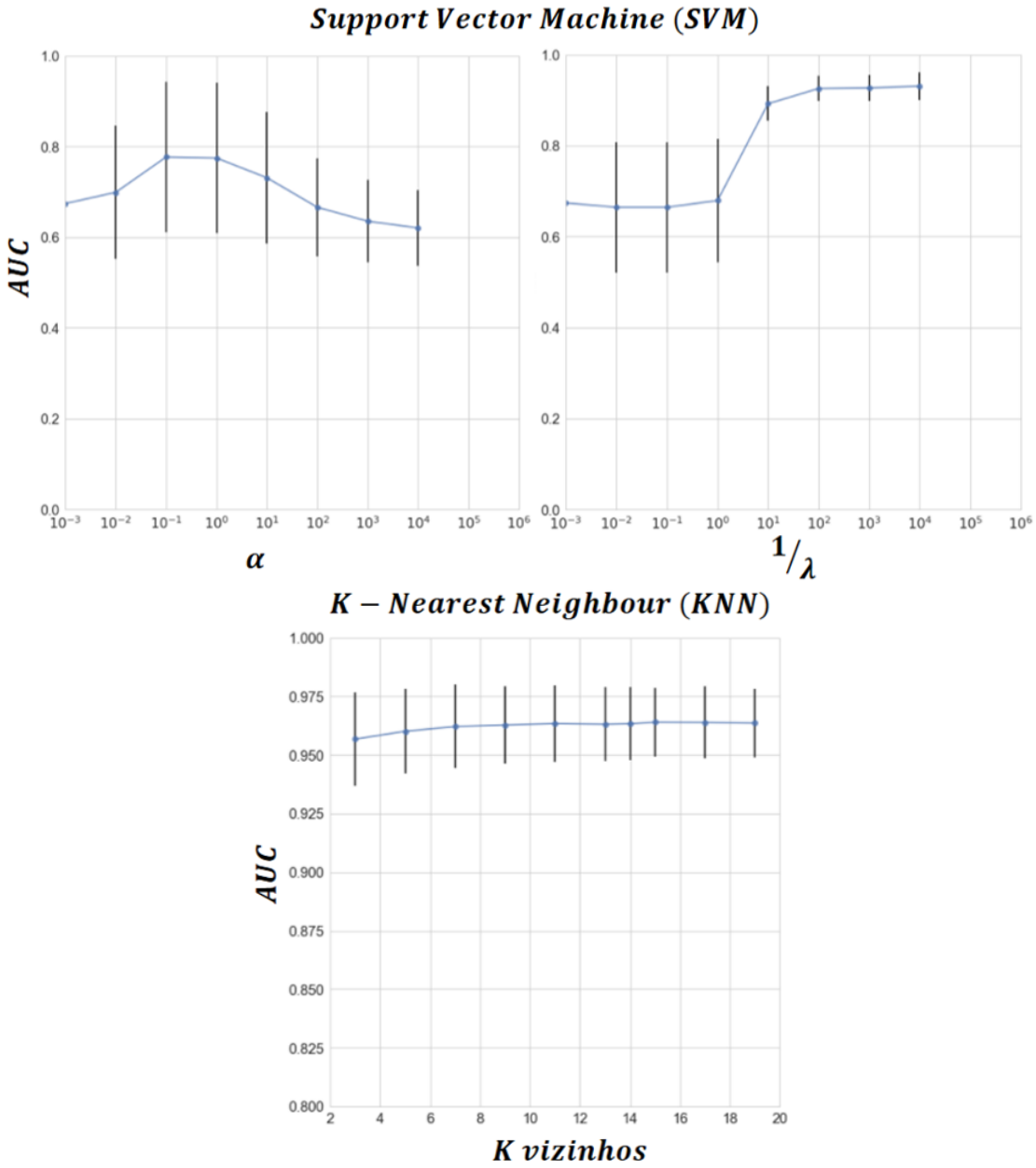
Os modelos de classificação usados seguem como princípio a aprendizagem supervisionada. Isso implica que uma função é inferida a partir de um conjunto de dados com categorias conhecidas *a priori*, o chamado conjunto treinamento. A partir dessa etapa, essa função (ou hipótese) pode ser usada para classificar novas categorias, não observadas até então.

Seguindo esse princípio, cada modelo de aprendizagem foi avaliado em termos de seu poder discriminativo, expresso pela AUC, assim como estabilidade, medido em termos do coeficiente de variação (CV%). Para isso, utilizou-se 100 amostras *bootstrap*, reamostrando os 129 casos (65 benignos e 64 malignos), conforme descrito anteriormente na Figura 24.

A Figura 33 mostra o comportamento do(s) parâmetro(s) de cada modelo durante a fase de treinamento. Considerando como critério a escolha dos parâmetros que fornecem maior desempenho de classificação em termos de AUC, os seguintes parâmetros foram utilizados na fase de teste dos modelos: RF (n_{tree}=300, AUC_{train}=0.959 $\pm$ 0.024), LR (λ=1, AUC_{train} =0.940 $\pm$ 0.029), ErT (n_{tree}=400, AUC_{train}=0.962 $\pm$ 0.018), GB (γ=0.01, AUC_{train}=0.961 $\pm$ 0.025), SVM (λ=0.01, $= 0.1$,AUC$_{train}$=0.943 $\pm$ 0.032), KNN ($K = 15$,AUC_{train}=0.964 $\pm$ 0.025).

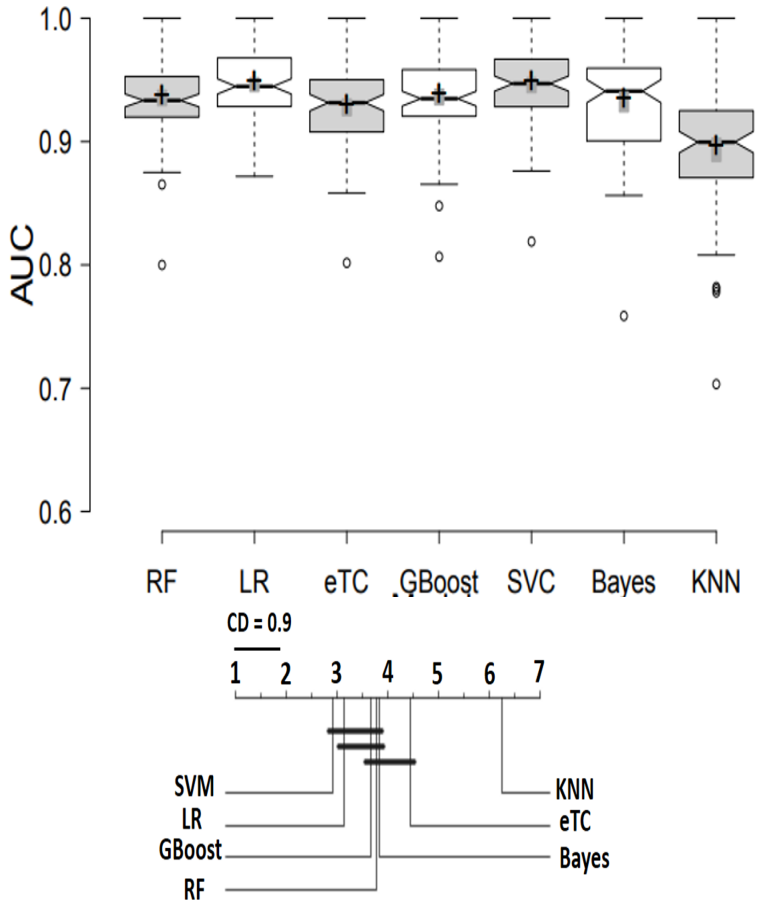
Figura 33 – Performance no conjunto treinamento para cada modelo ao variar os parâmetros destes. As barras representam o erro-padrão associado a cada parâmetro obtido nas amostras *bootstraps*.

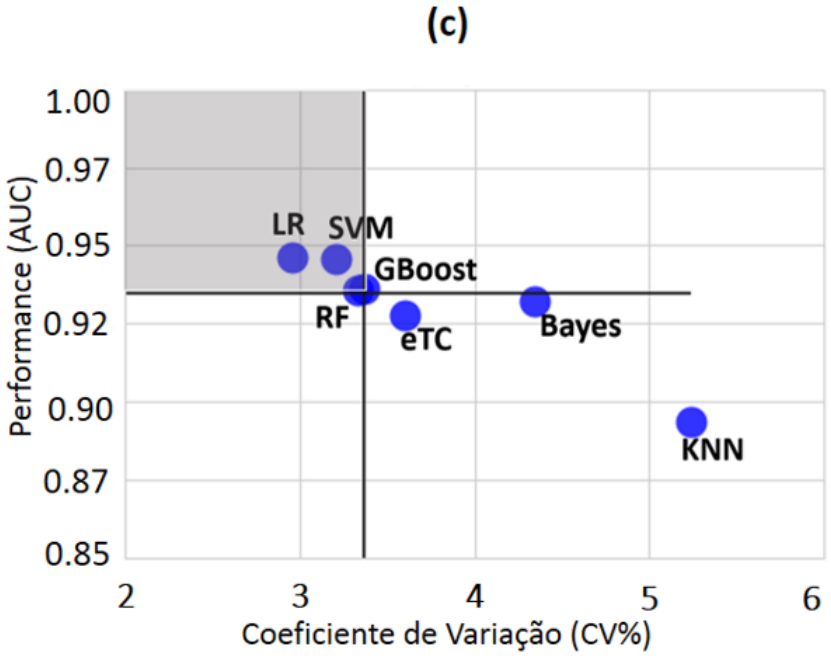




Por outro lado, o desempenho de classificação para cada modelo usando os dados de teste está sumarizada na Figura 34a. Ordenando a AUC (mediana ± std) do maior para o menor, observa-se a seguinte ordem SVM (AUC_{test}: 0.947 ± 0.031), LR (AUC_{test}: 0.945 ± 0.028), GB (AUC_{test}: 0.935 ± 0.032), seguido por NB (AUC_{test}: 0.941 ± 0.041), RF (AUC_{test}: 0.933 ± 0.031), ErT (AUC_{test}: 0.931 ± 0.033) e KNN (AUC_{test}: 0.899 ± 0.047).

Figura 34 – (a) Performance por modelo: (a) AUC e (b) ranking teste de cada modelo e (c) Modelos selecionados: performance *versus* estabilidade.





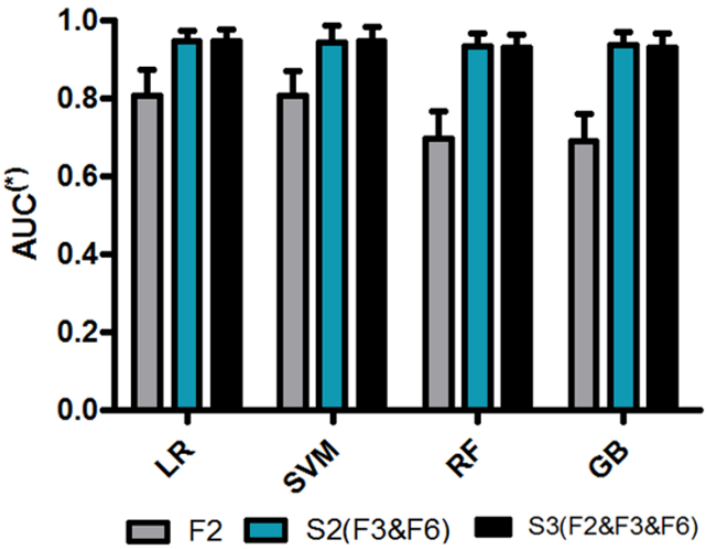
Esse comportamento foi confirmado por um teste dos *ranking* , seguido por *post – hoc* análise comparando todos os modelos entre sí (Figura 34b).

Todavia, essa mesma ordem muda levemente quando a estabilidade da predição desses modelos é o foco, e nesse sentido a seguinte sequencia foi observada: LR (CV: 2.956%), SVM (CV: 3.206%), RF (CV: 3.330%), GB (CV: 3.372%) como os mais estáveis modelos, seguidos por ErT (CV: 3.597%), NB (CV: 4.343%) e KNN (CV: 5.244%).

Assim sendo, em ordem de identificar o(s) modelo(s) mais acurado(s) e estáveis, os valores de medianas para AUC (Θ) e estabilidade (θ) foram utilizados como limiares. Métodos de classificação com $AUC \geq \Theta$ (AUC mediana de todos os modelos) e $CV \leq \theta$ (CV mediana de todos os modelos) foram considerados com alta performance e estabilidade na tarefa de discriminar linfonodos benignos de malignos. Esse comportamento está descrito no gráfico de dispersão mostrado na Figura 34c. Considerando como limiar os correspondentes valores medianos ($\Theta = 0.935$, $\theta = 3.33\%$), os modelos mais discriminativos e estáveis são aqueles cobertos pela região cinzenta na Figura 34c. Segundo esse critério, os modelos selecionados foram LR, SVM, GB e RF.

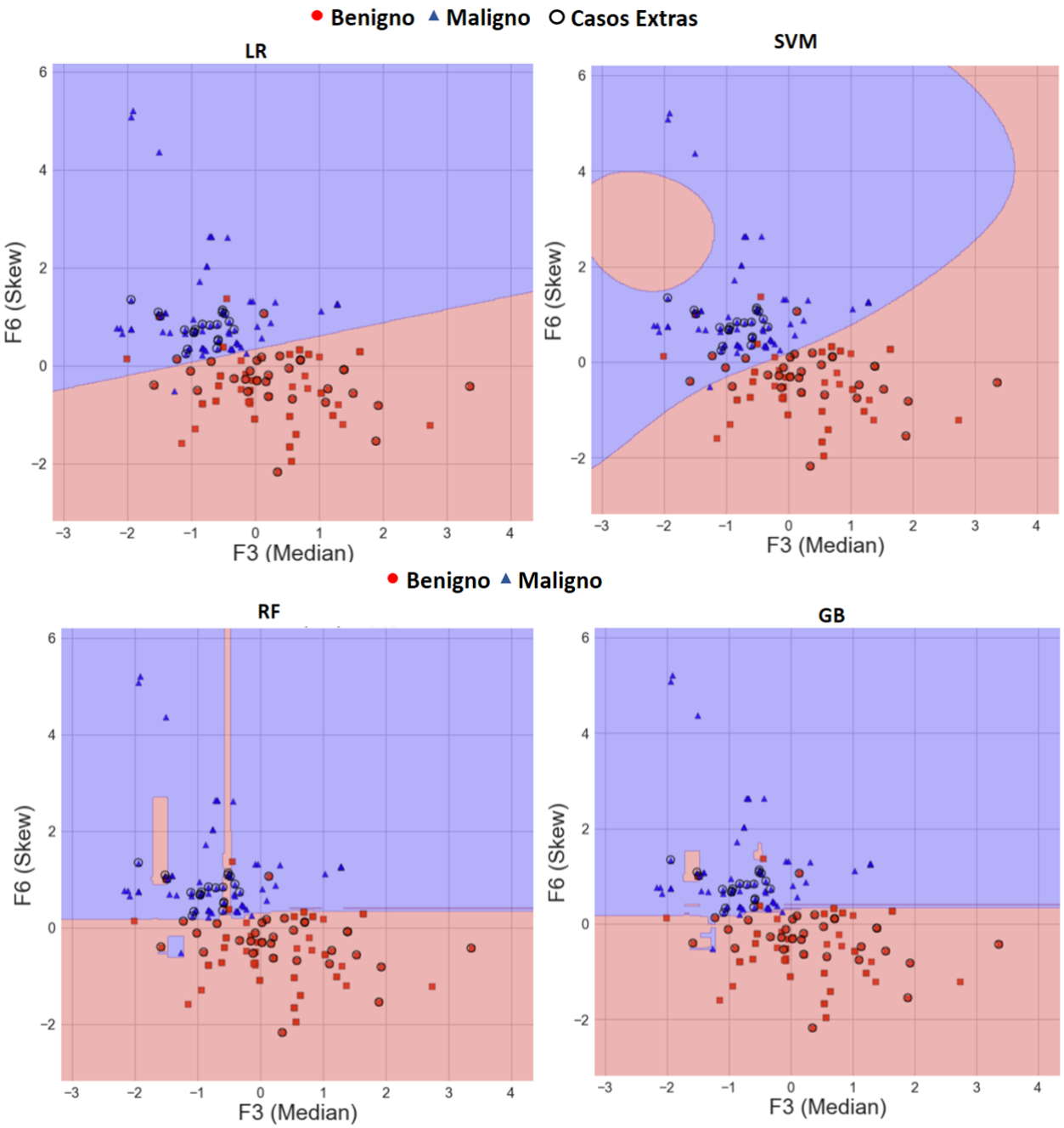
Em seguida, esses quatro modelos pré-selecionados foram avaliados usando os conjuntos $S_2 = \{F6, F3\}$ e $S_3 = \{F6, F3, F2\}$ (Figura 32) comparando-os a F2 (variável mais usada na prática clínica para esse tipo de caso). Os resultados são exibidos na Figura 35 em termos de AUC média e desvio padrão. Ao se usar quer seja S2 ou S3, a performance se mostrou praticamente a mesma independente do modelo utilizado. Além disso, comparado a utilização apenas de F2, o uso da combinação de duas ou mais características apresentou uma ganho de aproximadamente 14±0.10% para LR(F2:0.81±0.06; S2:0.95±0.03) e SVM (F2:0.81±0.06; S2:0.94±0.04) e de aproximadamente 24±0.11% para RF (F2: 0.69±0.07;S2:0.93±0.03) e GB (F2: 0.69±0.07;S2:0.93±0.04) ($p < 0.05$).

Figura 35 – AUC para cada modelo selecionado. Valores estimados nos conjuntos de teste nas amostras *bootstrap* (S_{test}).



F2- ADC médio F3- ADC mediana F6- ADC assimetria

Figura 36 – Fronteiras de decisão para os modelos selecionados usando $S_2 = \{F3, F6\}$.

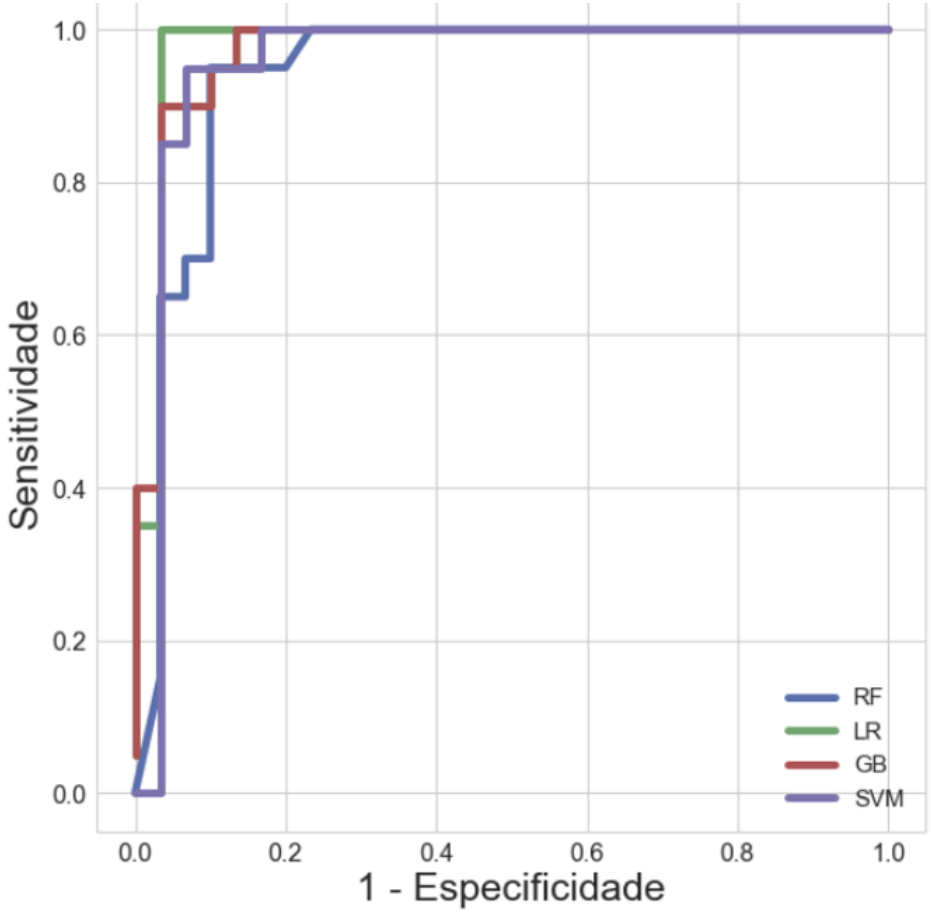


Os casos benignos e malignos marcados com círculo preto correspondem aos casos extras, não utilizados na fases de treinamento, nem na fase de teste.

Por fim, esses quatro modelos, com fronteiras de decisão mostradas na Figura 36, foram comparados usando dados extras para validação final. Para isso, utilizou-se um total de 50 casos (20 malignos e 30 benignos) que não fizeram parte nem da fase de treinamento, nem da fase de testes, marcados com círculos em preto na Figura 36. Baseadas nessas fronteiras de decisão, o desempenho de classificação dos casos extras em termos de AUCs foram equivalentes (Figura 37), com marginal ganho de sensitividade e especificidade para LR,

que apresentou (AUC_{extra}=0.978), seguido por SVM (AUC_{extra}=0.957), RF (AUC_{extra}=0.942) e GB(AUC_{extra}=0.972). Além disso, todos esses valores estão dentro do intervalo de confiança previsto nas amostras *bootstrap* e, portanto, refletem o grau de generalização de cada modelo.

Figura 37 – AUC para os casos extras usando $S_2 = F3, F6$.



5.2 Caso 2

A comparação dos métodos de extração de características tradicionais, ou não-adaptativos, e os adaptativos, tal como aqui proposto e descritos na Figura 25, foi dividida em duas etapas. Na primeira as 103 características não-adaptativas são comparadas e aquelas com diferença significativa entre os grupos controle e tratado ($p < 0.05$, MW teste) são selecionadas. Numa segunda etapa, aquelas características não-adaptativas selecionadas são comparadas com as características adaptativas, ADC F_+ e ADC F_-, em termos de sensitividade(%), especificidade(%), PPV(%), NPV(%) e AUC(%) usando amostras *bootstrap*.

5.2.1 Seleção das características não-adaptativas

Um total de 103 características divididas em 7 tipos foram inicialmente analisadas, tal como sumarizado na Tabela 5 (i-vii). Segundo uma análise não-paramétrica usando MW teste, as características baseadas em (ii) Forma, (iv) GLRLM (*Gray Level Size Zone Matrix*) e (vi) NGTDM (*Neigbouring Gray Tone Difference Matrix*) apresentaram uma alta interseção entre os grupos controle e tratados e, por isso, nenhuma diferença significativa foi observada usando como critério $p < 0.05$. Por outro lado, entre os tipos que apresentaram uma diferença significativa estão (i) ADC histograma, (iii) GLCM (*Gray Level Co-ocurence Matrix*), (v) GLRLM (*Gray Level Run Length Matrix*) e (vii) GLDM (*Gray Level Dependence Matrix*) (Tabela 5).

Tabela 5 – Características não-adaptativas: (i) ADC histograma, (ii) Geometria, (iii) GLCM, (iv) GLSZM, (v)GLRLM, (vi) NGTDM e (vii)GLDM. Características com (*) indicam significante diferença entre grupos controle e tratado ($p < 0.05$)

Tipo	Nome	Mediana[95% CI]	Mediana[95% CI]	p
		Controle	**Tratado**	
(i)	ΔSkewness*	199.01 [-543.74,941.76]	-68.79 [-140.29,2.71]	0.0012
	ΔKurtosis*	45.78 [-179.4,270.95]	-25.11 [-38.37 , -11.86]	0.0041
	ΔMaximum*	16.82 [-0.18 , 33.81]	6.01 [-1.05 , 13.06]	0.03
	Δ90Perc	0.39 [-5.03 , 5.8]	10.5 [6.67 , 14.33]	0.19
	ΔMedian	2.87 [-2.25 , 8]	10.98 [7.58,14.37]	0.22
	ΔStddDeviation	14.59 [9,20.18]	11.24 [-1.07 , 23.54]	0.22
	ΔVariance	31.31 [18.47,44.15]	23.74 [-2.84,50.31]	0.22
	ΔMinimum	33.02 [-681.83,747.87]	9.84 [-96.74,116.42]	0.31
	ΔMean	2.71 [-2.56,7.98]	10.93 [7.5,14.37]	0.39
	ΔRootMeanSq	2.78 [-2.33 , 7.89]	10.87 [7.54,14.2]	0.39
	ΔRange	7.9 [-20.29,36.09]	3.95 [-8.16 , 16.05]	0.43
	ΔEnergy	4.76 [-9.64,19.16]	15.7 [7.55 , 23.85]	0.43
	ΔTotalEnergy	4.76 [-9.64 , 19.16]	15.7 [7.55 , 23.85]	0.43
	ΔRobustMeanAbsDev	2.15 [-8.82 , 13.11]	14.57 [2.53,26.61]	0.52
	ΔEntropy	1.89 [-13.6,17.37]	7.5 [-4.40,19.40]	0.52
	Δ10Percentile	3.55 [-2.66,9.76]	8.85 [3.47 , 14.23]	0.57
	ΔInterquartileRange	6.62 [-4.13 , 17.37]	14.21 [2.35 , 26.08]	0.76
	ΔUniformity	3.9 [-8.09 , 15.89]	-14.69 [-26.73 , -2.66]	0.79
	ΔMeanAbsDeviation	9.78 [1.96 , 17.61]	15.1 [3.19,27.01]	0.85
(ii)	ΔSurfArea	3.54 [-6.66 , 13.73]	-5.05[-9.75 , -0.35]	0.19
	ΔMinorAxis	4.07 [-2.74,10.89]	-3.39 [-9.31,2.54]	0.22

Continua na próxima página

Tabela 5 – continuação da página anterior

Tipo	Nome	Mediana[95% CI]	Mediana[95% CI]	p
	ΔElongation	3.67 [-5.04,12.38]	-3.73 [-11.27,3.81]	0.25
	ΔCompactness1	-6.67 [-13.85,0.52]	2.20 [-1.84, 6.24]	0.31
	ΔCompactness2	-12.63[-26.26 , 1]	4.45 [-3.35,12.25]	0.31
	ΔSpherDisprop	4.88 [-0.4 , 10.17]	-1.44 [-4.33,1.46]	0.31
	ΔSphericity	-4.53 [-9.41,0.35]	1.46 [-1.27,4.19]	0.31
	ΔLeastAxis	-1.42 [-4.23,1.39]	-0.22 [-3.9,3.45]	0.48
	ΔMax2DDiamRow	-8.74 [-20.3,2.81]	-5.79 [-15.18,3.6]	0.63
	ΔSurfVolRatio	6.9 [0.27,13.53]	-0.57 [-4.49,3.35]	0.63
	ΔVolume	-4.09 [-15.89,7.7]	-3.29 [-8.85,2.27]	0.74
	ΔMajorAxis	-2.4 [-7.81,3.01]	0.88 [-2.84,4.59]	0.79
	ΔMax3DDiam	-0.2 [-4.51,4.12]	1.86 [-4.91,8.64]	0.85
	ΔFlatness	-0.16 [-6.31,6]	0.23 [-5.58,6.04]	0.91
	ΔMax2DDiamColumn	5.64 [-8.56,19.84]	6.4 [-0.94,13.8]	0.91
	ΔMax2DDiamSlice	-1.47[-5.69,2.75]	2.06[-2.36,6.48]	0.97
(iii)	ΔClusterProminence*	181.85 [43.78,319.92]	0.57 [-71.73,72.88]	0.01
	ΔClusterShade*	293.13 [-262.5,848.77]	-50.53 [-160.93,59.88]	0.01
	ΔJointEntropy	0.07 [-16.5,16.65]	7.97 [-4.41,20.36]	0.35
	ΔClusterTendency	22.01 [-0.98,44.99]	19.54 [-9.3,48.37]	0.43
	ΔDifferVariance	14.61 [-1.81,31.03]	10.75 [-1.5,23]	0.43
	ΔSumEntropy	2.34 [-13.59,18.27]	7.12 [-4.78,19.03]	0.43
	ΔIdmn	0.08 [-0.27,0.43]	-0.1 [-0.26 ,0.07]	0.43
	ΔMaxProbability	11.52 [-4.7,27.75]	-20.09 [-36.52,-3.66]	0.48
	ΔSumSquares	21.71 [2.81,40.61]	16.55 [-9.16,42.26]	0.53
	ΔDifferEntropy	4.89 [-6.74,16.52]	7.31 [-0.89,15.52]	0.53
	ΔSumAverage	-1.63 [-19.61,16.35]	11.38 [-6.37,29.14]	0.68
	ΔJointEnergy	8.99 [-10.79,28.78]	-24.82 [-43.89,-5.76]	0.68
	ΔJointAverage	-1.63 [-19.61,16.35]	11.38 [-6.37,29.14]	0.68
	Δ Imc2	8.48 [-8.95,25.91]	4.7 [-6.61,16.02]	0.68
	ΔAutocorrelation	-2.12 [-34.73, 30.49]	24.74 [-15.3,64.77]	0.74
	ΔIdn	0.16 [-0.9,1.22]	-0.37 [-0.83,0.09]	0.79
	ΔContrast	12.49 [-11.23,36.21]	18.28 [0.06,36.5]	0.79
	ΔImc1	15.75 [-8.56,40.05]	6.41 [-6.14,18.95]	0.85
	ΔHomogeneity2	-0.39 [-2.84,2.06]	-3.21 [-5.51,-0.9]	0.91
	ΔHomogeneity1	-0.3 [-2.73,2.13]	-3.06 [-5.28 , -0.83]	0.91
	ΔDissimilarity	4.71 [-18.67,28.08]	15.69 [-1.36 , 32.74]	0.91

Continua na próxima página

Tabela 5 – continuação da página anterior

Tipo	Nome	Mediana[95% CI]	Mediana[95% CI]	p
	Δ Diff.Average	4.71[-18.67,28.08]	15.69[-1.36,32.74]	0.91
	ΔIdm	-0.39 [-2.84,2.06]	-3.21 [-5.51,-0.9]	0.91
	ΔId	-0.3 [-2.73,2.13]	-3.06 [-5.28,-0.83]	0.91
	ΔInv.Variance	0.13 [-23.15 ,23.4]	13.23 [-3.49,29.96]	0.97
	ΔCorrelation	12.7 [-6.12 , 31.52]	8.23 [-3.54 , 20.01]	0.97
(iv)	ΔGLVar	15.45[-19.86,50.76]	1.13[-20.86,23.12]	0.057
	ΔSALGLE	51.2[-18.67,121.07]	-3.28[-28.44,21.87]	0.093
	ΔLAHGLE	-32.69[-52.64,-12.74]	-10.72[-42.02,20.57]	0.15
	ΔGLNUN	-6.42[-14.55,1.72]	5.41[-12.03,22.86]	0.19
	ΔZP	24.36[-32.41,81.14]	2.95[-13.39,19.3]	0.31
	ΔSAE	1 [-53.54,55.54]	3.29[-13.84,20.41]	0.48
	ΔSZNU	6.95[-101.27,115.16]	10.04[-14.57,34.66]	0.57
	ΔSZNUN	3.69[-31.47,38.85]	8.16[-13.12,29.45]	0.57
	ΔZVar	-12.36[-55.47,30.74]	-20.24[-45.84,5.37]	0.57
	ΔHGLZE	7.83[-16.08,31.75]	-1.32[-39.28,36.64]	0.63
	ΔSAHGLE	2.02[-203.19,207.23]	-2.54[-48.8,43.72]	0.68
	ΔGLNU	3.71[38.25,45.67]	-0.02[-19.77,19.73]	0.74
	ΔLAE	-10.1[-55.65,35.45]	-19.16[-45.3,6.99]	0.79
	ΔLALGLE	-16.31[-174.36,141.73]	-42.93[-123.6,37.73]	0.79
	ΔLGLZE	25.1[-18.39,68.58]	-19.7[-69.79,30.39]	0.97
	ΔZEntrop	1.23[-6.62,9.07]	-0.75[-4.74,3.24]	0.97
(v)	ΔGLVar*	35.15[17.86,52.43]	-2.16[-17.95,13.62]	0.006
	ΔRunEntropy	0.71[-2.33,3.76]	-0.91[-2.6,0.78]	0.13
	ΔGLNUN	-4.86[-9.27,-0.44]	-2.74[-8.78,3.3]	0.39
	ΔLRLGLE	-0.57[-83.18,82.03]	-31.67[-107.4,44.06]	0.48
	ΔSRE:	3.92[-5.96,13.79]	3.13[-1.31,7.56]	0.68
	ΔLGLRE	9.78[-57.27,76.83]	-16.91[-76.45,42.64]	0.74
	ΔHGLRE	0.85[-34.16,35.86]	24.48[-14.21,63.17]	0.79
	ΔLRHGLE	-25.56[-64.69,13.58]	-13.09[-41.24,15.05]	0.79
	ΔRunPercentage	2.57[-9.56,14.69]	8.48[0.7,16.26]	0.79
	ΔGLNU	2.19[-7.2,11.59]	-1.39[-9.51,6.72]	0.85
	ΔRunVariance	6.42[-14.61,27.45]	-26.08[-46.13,-6.03]	0.85
	ΔSRLGLE	25.03[-40.32,90.38]	-6.41[-64.74,51.92]	0.85
	ΔRLNUN	4.32[-11.9,20.53]	8.62[0.99,16.24]	0.91

Continua na próxima página

Tabela 5 – continuação da página anterior

Tipo	Nome	Mediana[95% CI]	Mediana[95% CI]	p
	ΔSRHGLE	16.52[-19.86,52.9]	24.36[-16.52,65.24]	0.91
	ΔLRE	-3.44[-24.54,17.65]	-16.65[-35.01,1.71]	0.97
	ΔRLNU	5.7[-20.12,31.51]	11.11[-8.23,30.45]	0.97
(vi)	ΔStrength	76.57[-36.73,189.87]	0.41[-37.48,38.29]	0.08
	ΔComplexity	30.98[-57,118.96]	19.59[-20.13,59.31]	0.35
	ΔCoarseness	-7.07[-24.44,10.3]	-4.78[-21.74,12.17]	0.48
	ΔContrast	-5.92[-109.39,97.56]	50.65[21.37,79.94]	0.52
	ΔBusyness	-7.13[-88.02,73.76]	1.11[-31.99,34.22]	0.79
(vii)	ΔDVar*	-0.59[-6.2,5.01]	-7.9[-13.68,-2.12]	0.04
	ΔGLVar	26.44[13.15,39.74]	13.66[-10.54,37.85]	0.25
	ΔSDHGLE	35.78[-21.53,93.09]	13.12[-29.76,55.99]	0.48
	ΔDNUN	-1.17[-13.77,11.43]	0.31[-12.05,12.68]	0.52
	ΔLDLGLE	2.97[-68.48,74.42]	-29.65[-87.95,28.65]	0.57
	ΔSDE	6.5[-28.44,41.44]	9.19[-2.04,20.42]	0.57
	ΔGLNU	-5.4[-21.06,10.26]	-20.65[-32.73,-8.57]	0.63
	ΔLDHGLE	-18.49[-52.56,15.58]	6.02[-25.46,37.49]	0.63
	ΔLGLE	8.59[-59.46,76.64]	-16.74[-70.54,37.07]	0.68
	ΔSDLGLE	72.42[18.72,126.12]	-2.08[-40.87,36.71]	0.68
	ΔDNU	-3.27[-22.84,16.31	-5.6[-18.11,6.92]	0.74
	ΔHGLE	-1.38[-33.63,30.88]	24.1[-15.39,63.59]	0.74
	ΔGLNUN	3.9[-8.09,15.89]	-14.69[-26.73,-2.66]	0.79
	ΔDEntrop	-0.89[-6.19,4.41]	2.63[-2.04,7.29]	0.85
	ΔLDE	-1.78[-12.8,9.25]	-12.1[-20,-4.2]	0.91

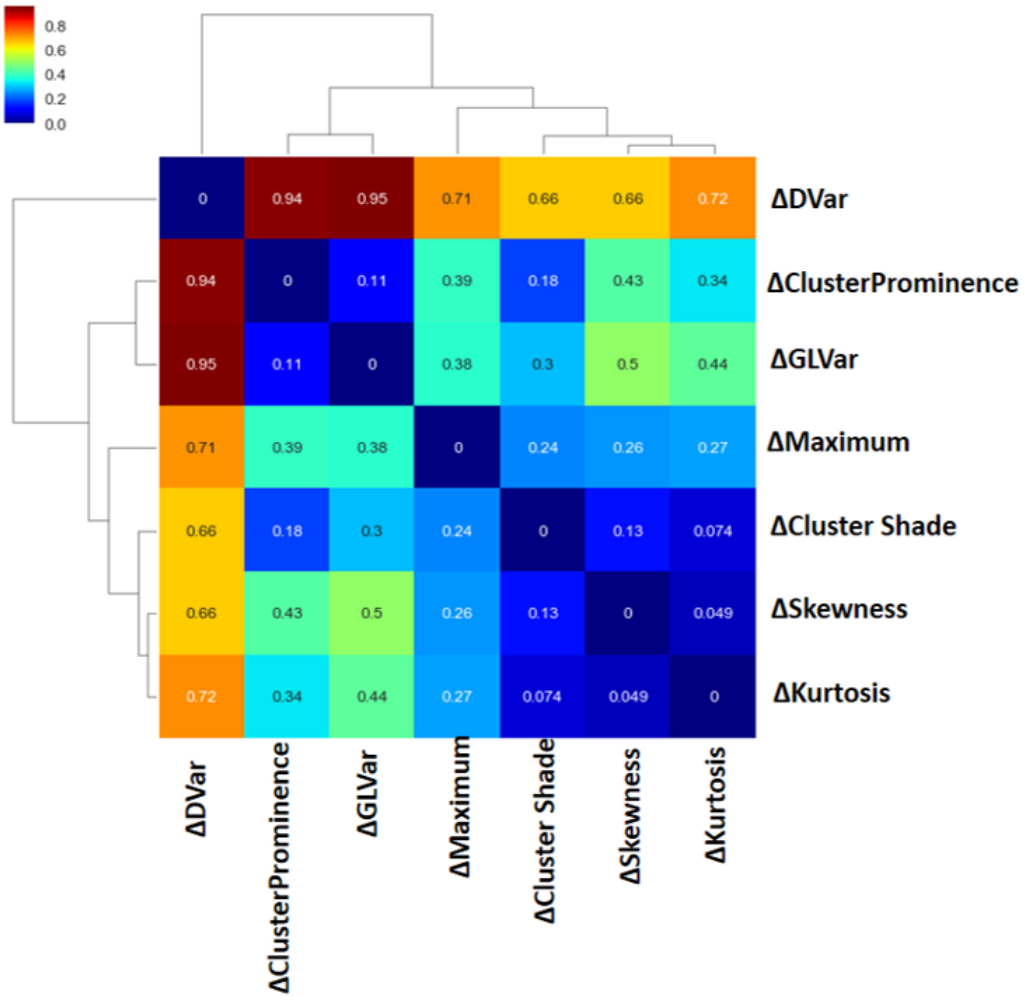
Contudo, das 19 variáveis analisadas em (i), apenas $\Delta Skewness$ ($p = 0.0012$), $\Delta Kurtosis$ ($p = 0.0041$) e $\Delta Maximum$ ($p = 0.03$) apresentaram diferença significativa. Essas três variáveis têm se mostrado promissoras no uso de mapa ADC na tarefa de distinguir diferente tipos de gliomas (TOZER et al., 2007; NOWOSIELSKI et al., 2011), câncer de cabeça e pescoço (KING et al., 2014), assim como osteosarcoma (FOROUTAN et al., 2013), nos contextos clínico e pré-clinico. Nos demais tipos, esse número foi ainda menor: duas variáveis com diferença significativa em (iii) $\Delta ClusterProminence$ ($p = 0.01$) e $\Delta ClusterShade$ ($p = 0.01$), e uma cada em (v) $\Delta GLVar$ ($p = 0.006$) e (vii) $\Delta DVar$ ($p = 0.04$) (Tabela 5).

Além disso, todas essas 7 variáveis pertencentes a esses quatro diferentes tipos apresentaram uma mesma tendência: significativo decrescimento nas mudanças relativas de ADC após o

tratamento. Essa alta correlação entre variáveis de diferentes ordens, tal como mostra Figura 38, pode ser explicada pela presença de regiões localizadas onde o tratamento surtiu efeito, de tal maneira que apenas uma fração dos *voxels* apresentaram diferença espacial, captada apenas por alguns métodos. Isso implica na presença de grande porção dos *voxels* apresentando intensidade parecida com a intensidade tumoral no grupo tratado, mascarando portanto a diferenciação dos grupos controle e tratado por métodos que exploram a relação espacial entre *voxels*, tal como nos tipos de iii a vii.

Assim sendo, considerando que as variáveis com $p < 0.05$ apresentaram similar comportamento univariado (Figura 38), com poder discriminativo de mesma ordem, optar-se-á nas análises a seguir pelo uso de variáveis derivadas do histograma (i) na comparação com as características adaptativas ADC $F+$ e ADC $F-$, tal como descrito a seguir.

Figura 38 – Mapa de clusteres mostrando o grau de redundância entra as características com $p <$ 0.05.



A distância utilizada como medida de similaridade foi $d_{ij} = 1 - \rho_{ij}$, onde ρ_{ij} é o grau de correlação entre as variáveis i e j. Quanto maior o grau de correlação entre i e j, menor a distância d_{ij} entre elas, o que implica similar poder discriminativo univariadamente falando. Os *clusters* expressos pelo dendogramas foram obtidos por meio da distância média entre os elementos de cada *cluster* (*linkage* médio) (ROKACH; MAIMON, 2005).

5.2.2 Características: Não-adaptativas versus Adaptativas

Considerando o poder discriminativo das características dos tipos (i), (iii), (v) e (vii) (Tabela 5), assim como o comportamento redundante entre elas (Figura 38), a diferença relativa entre os grupos tratados e controle será avaliada usando $\Delta Skewness$ ($p = 0.0012$) e $\Delta Kurtosis$ ($p = 0.0041$), assim como $\Delta Mean$, uma vez que, como discutido anteriormente, este é o parâmetro referência na prática clínica.

Essas variáveis foram então comparadas às características adaptativas, ADC $F+$ e ADC $F-$ sob a ótica de seus potenciais de discriminar um grupo do outro, assim como suas estabilidades, em termos de CV, nessa tarefa de discriminação.

Dada sensitividade(Sens %) e especificidade(Especif %) no ponto de ótimo limiar da curva ROC, aqui definido como o máximo $Sens\% + (1 - Sens\%)$ (VANDECAVEYE et al., 2009), o PPV (*Positive Predictive Value%*) e PPV (*Positive Predictive Value%*) foram calculados como,

$$PPV = \frac{Sens.Prev}{Sens.Prev + (1 - Especif).(1 - Prev)} \tag{5.1}$$

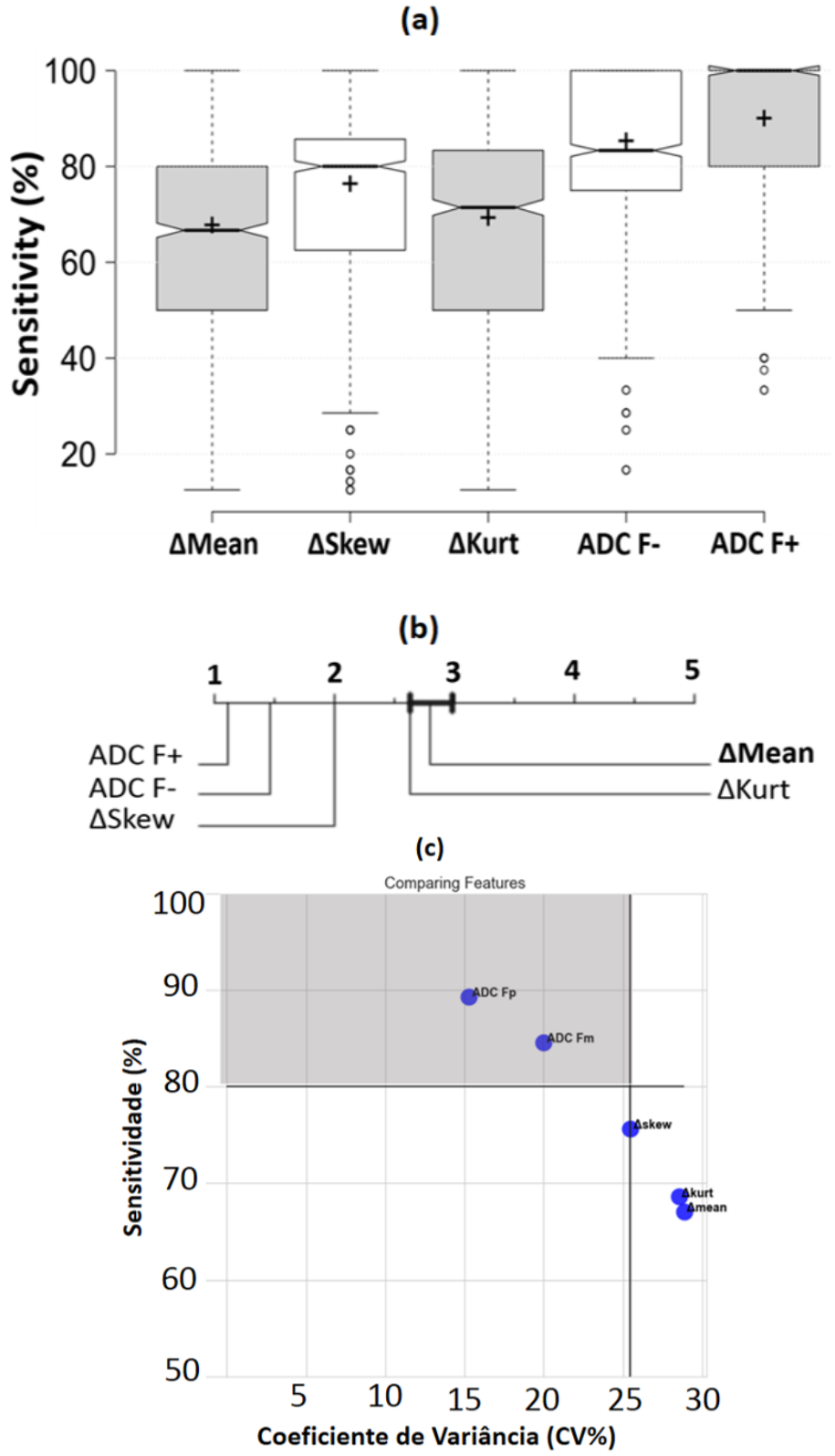
$$NPV = \frac{Especif.(1 - Prev)}{(1 - Sens).Prev + Especif.(1 - Prev)} \tag{5.2}$$

onde *Prev* é a fração do total de VOIs onde o tumor ainda está presente. Enquanto PPV mede a probabilidade de detectar um VOI que não respondeu ao tratamento, quando o mesmo é identificado como não respondendo; NPV vai informar a probabilidade do VOI ter respondido ao tratamento quando o mesmo é diagnosticado como tal. Quanto maior esses valores, melhor do ponto de vista clínico.

A Figura 39 agrupa as performances de cada característica nas amostras *bootstraps* na seguinte ordem: Sens(%), Especif(%), PPV(%), NPV(%) e AUC(%) . Para cada uma dessas 5 performances, há três gráficos com informações complementares. O primeiro gráfico (a) consiste de gráficos de caixa mostrando as distribuições das performances de cada característica. O segundo, (b), contém uma escala que ordena e compara cada variável com $\Delta Mean$ usando como teste estatístico o teste de *Bonferroni-Dunn*. Neste gráfico (b), as variáveis à esquerda apresentam melhor performance do que às posicionadas à direita. O intervalo demarcado corresponde ao intervalo de diferença crítica (CD) do teste: características com *rank* fora dessa área é significativamente diferente da $\Delta Mean$ com 95% de confiança de acordo com o teste de *Bonferroni-Dunn*. Isso permite ter uma ideia do valor clínico adicional de cada variável em estudo com relação ao que atualmente é utilizado na prática clínica. Por fim, em (c) um diagrama de pontos fornece a informação geral da estabilidade (CV%) *versus* performance média de cada característica.

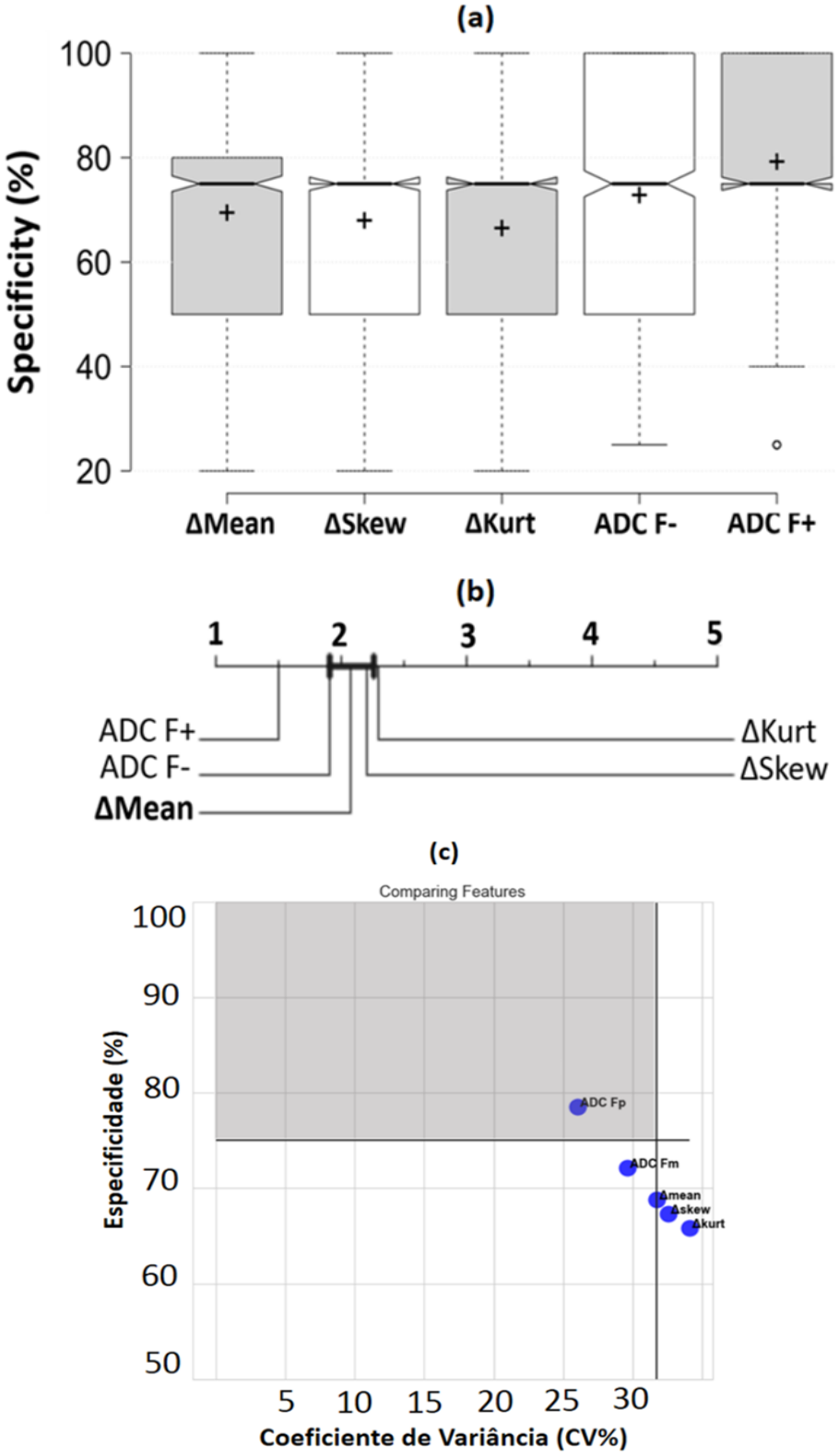
Figura 39 – Comparação das performances das características não-adaptativas *versus* adaptativas.

Performance 1 (Sensitividade): (a) Gráfico de caixa, (b) Performance ordenada e (c) Desempenho (AUC) x Estabilidade (CV).



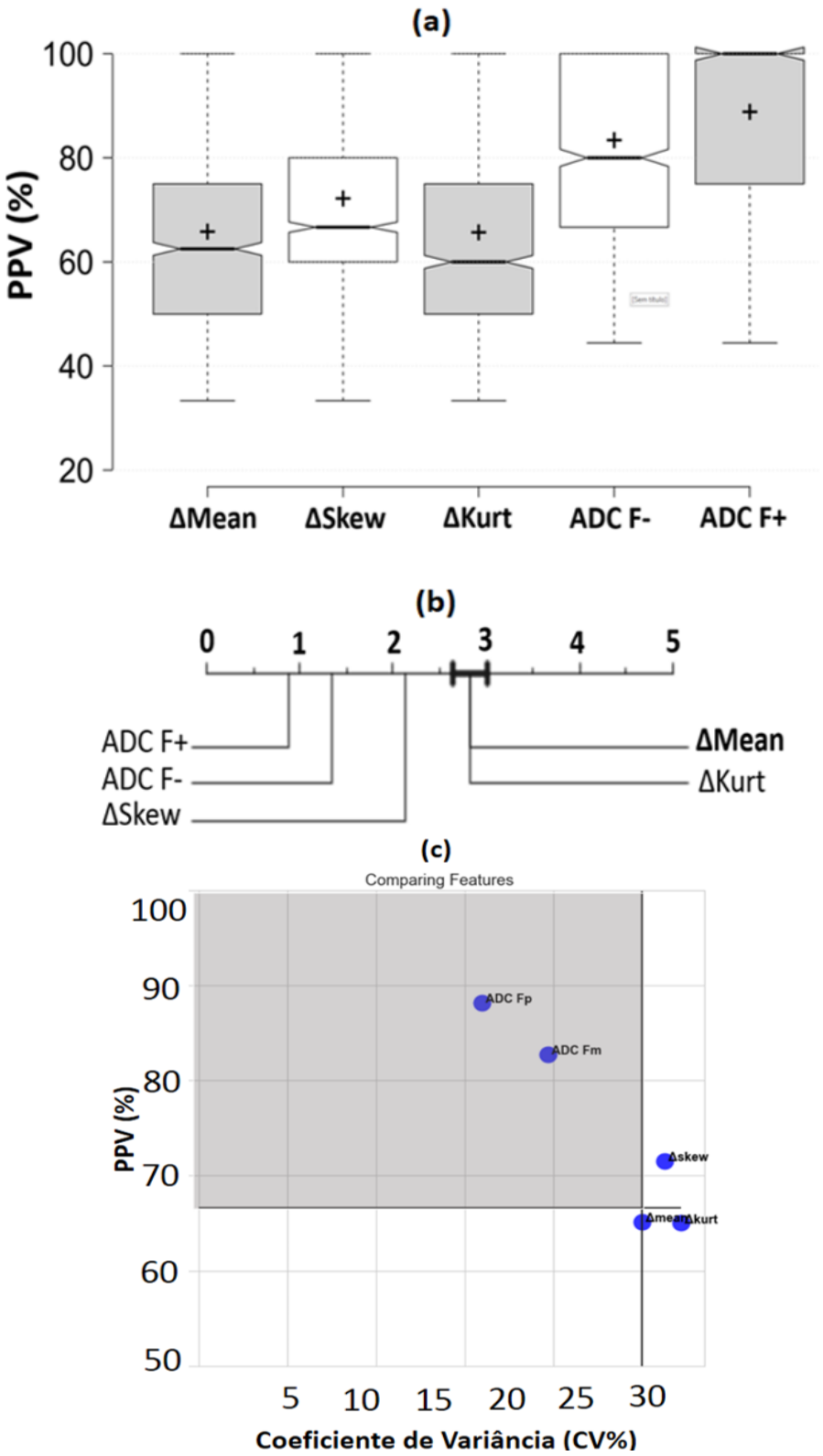
Gráficos de caixa mostram distribuição para cada característica. Os entalhes representam intervalo de confiança (95%) para a mediana (linha horizontal preta) e são definidos como Mediana $\pm$ 1.58*IQR/$\sqrt{N}$, onde IQR é o intervalo interquartílico e N o número de amostras *bootstrap*. A cruz em cada gráfico indica o valor médio da distribuição *bootstrap* das variáveis.

Performance 2 (Especificidade): (a) Gráfico de caixa, (b) Performance ordenada e (c) Desempenho (AUC) x Estabilidade (CV).



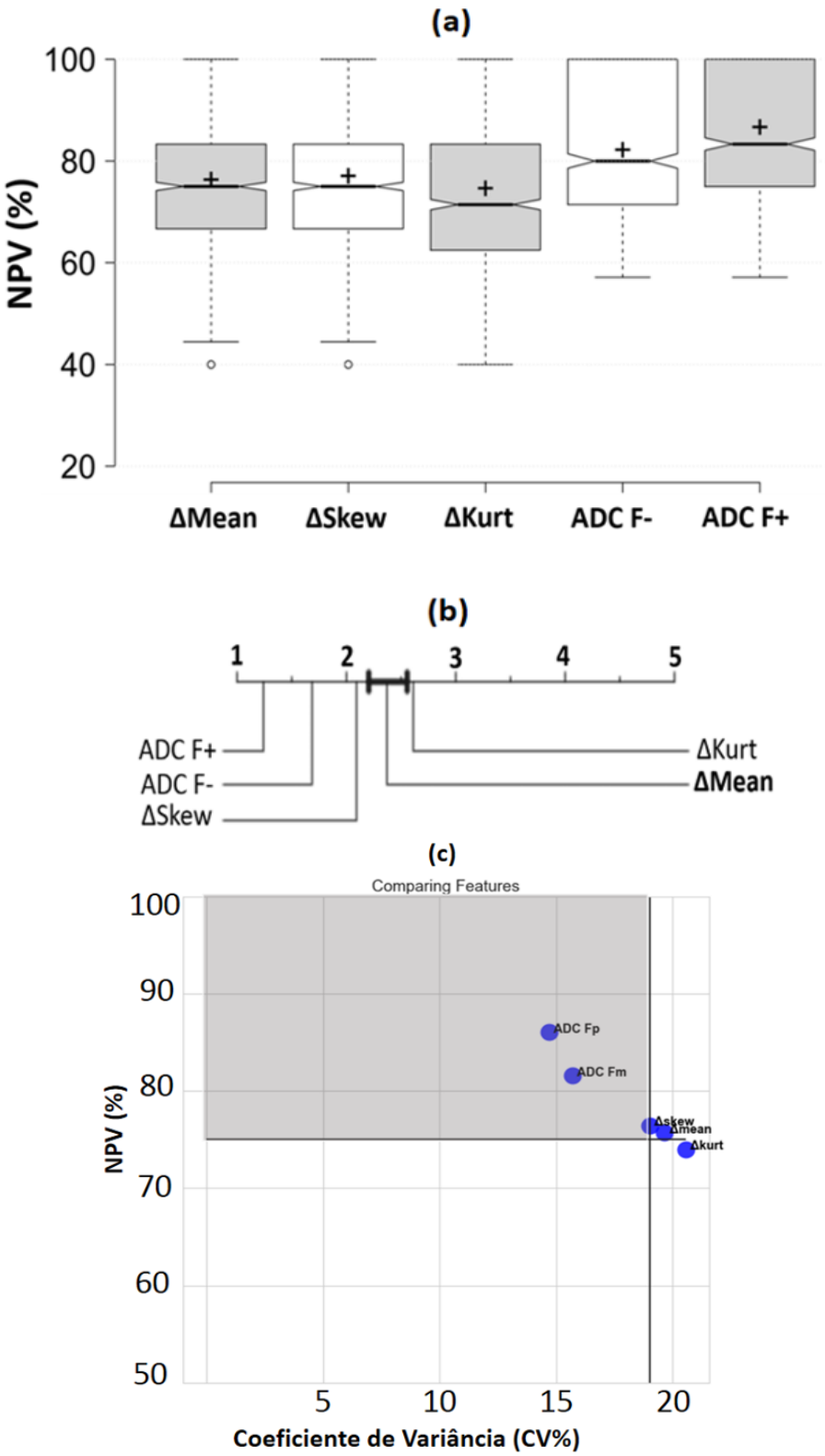
Gráficos de caixa mostram distribuição para cada característica. Os entalhes representam intervalo de confiança (95%) para a mediana (linha horizontal preta) e são definidos como Mediana $\pm$ 1.58*IQR/$\sqrt{N}$, onde IQR é o intervalo interquartílico e N o número de amostras *bootstrap*. A cruz em cada gráfico indica o valor médio da distribuição *bootstrap* das variáveis.

Performance 3 (PPV): (a) Gráfico de caixa, (b) Performance ordenada e (c) Desempenho (AUC) x Estabilidade (CV).



Gráficos de caixa mostram distribuição para cada característica. Os entalhes representam intervalo de confiança (95%) para a mediana (linha horizontal preta) e são definidos como Mediana $\pm$ 1.58*IQR/$\sqrt{N}$, onde IQR é o intervalo interquartílico e N o número de amostras *bootstrap*. A cruz em cada gráfico indica o valor médio da distribuição *bootstrap* das variáveis.

Performance 4 (NPV): (a) Gráfico de caixa, (b) Performance ordenada e (c) Desempenho (AUC) x Estabilidade (CV).



Gráficos de caixa mostram distribuição para cada característica. Os entalhes representam intervalo de confiança (95%) para a mediana (linha horizontal preta) e são definidos como Mediana $\pm$ 1.58*IQR/$\sqrt{N}$, onde IQR é o intervalo interquartílico e N o número de amostras *bootstrap*. A cruz em cada gráfico indica o valor médio da distribuição *bootstrap* das variáveis.

Performance 5 (AUC): (a) Gráfico de caixa, (b) Performance ordenada e (c) Desempenho (AUC) x Estabilidade (CV).

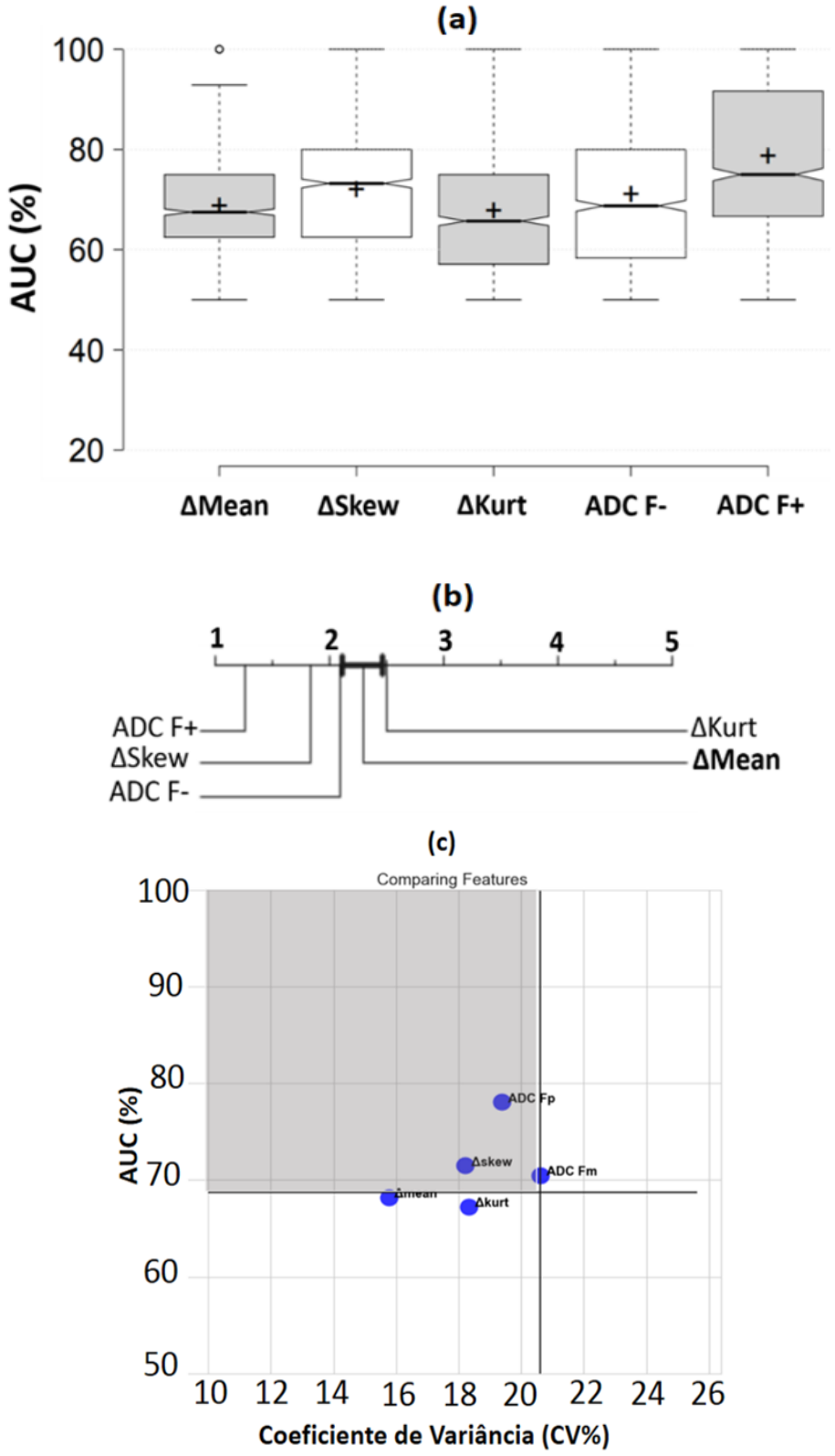


Figura 39

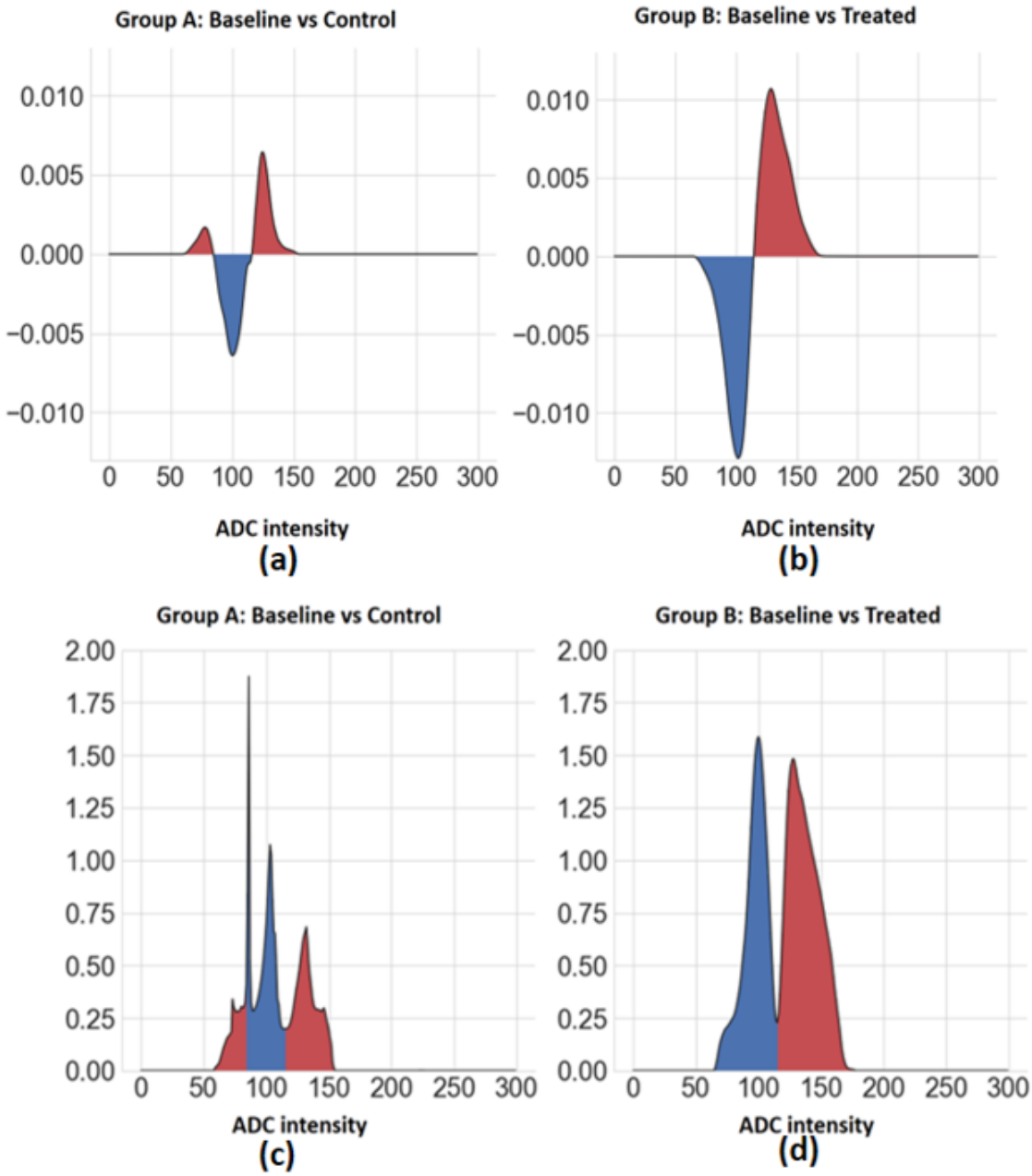
Gráficos de caixa mostram distribuição para cada característica. Os entalhes representam intervalo de confiança (95%) para a mediana (linha horizontal preta) e são definidos como Mediana $\pm$ 1.58*IQR/$\sqrt{N}$, onde IQR é o intervalo interquartílico e N o número de amostras *bootstrap*. A cruz em cada gráfico indica o valor médio da distribuição *bootstrap* das variáveis.

Um teste de *Friedman* com 5% de significância revelou diferença significativa entre o *rank* médio dos grupos de características para todos os parâmetros analisados derivados da curva ROC ($p < 0.0001$). Contudo, uma vez que o teste de *Friedman per si* não mostra quais características estariam originando tal diferença, análise *post-hoc* dos possíveis pares de texturas adaptativas *versus* não-adaptativas revelaram que $ADCF+$ forneceu melhores resultados comparado às outras características, as quais se diferenciaram de $\Delta Mean$ com uma pequena margem (ou mesmo não se diferenciaram, $p > 0.05$, vide diagramas Figura 39b) após abordagem mais detalhada usando as $N_{boot} = 1000$ amostras *bootstrap*. A Tabela 6 contém um sumário das distribuições dos desempenhos de interesse para cada característica, expresso em termos da mediana ($1.58IQR/\sqrt{N_{boot}}$,IQR), onde IQR é o intervalo interquartílico.

Essa diferença no desempenho fornecida pelos métodos de textura adaptativas, especialmente por ADC $F+$, seguida por ADC $F-$, pode ser explicada pelo fato de os métodos tradicionais não-adaptativos assumirem que toda a região no mapa de ADC do VOI contribui uniformemente na discriminação entre os grupos. Todavia, na abordagem adaptativa, ADC $F+$ e ADC $F-$ levarão em conta apenas sub-regiões discriminativas, usando como fator de peso a distância entre as classes, o que automaticamente tende a reduzir a contribuição de regiões responsáveis pela interseção entre os grupos. Tal interseção em geral é proveniente de mudanças biológicas após o tratamento que, assim como as células tumorais, tendem a reduzir a difusão e, portanto, os valores de ADC, mesmo em pacientes que responderam positivamente ao tratamento (PADHANI et al., 2009; JUST, 2011).

Nesse sentido, a Figura 40 exibe as regiões discriminativas baseadas na distância entre as classes. Nesta mesma figura, estas distâncias, $J_P(i|c_1,c_2)$, estão associadas às diferenças entre as classes, $\Delta_P(i|c_1,c_2)$, que podem ser maiores (região em vermelho) ou menores que zero (região em azul).

Figura 40 – (a),(b) Diferença $\Delta_P(i|c_1,c_2)$ e (c),(d) Distância, $J_P(i|c_1,c_2)$, entre os grupos controle e tratado com relação ao *baseline*.

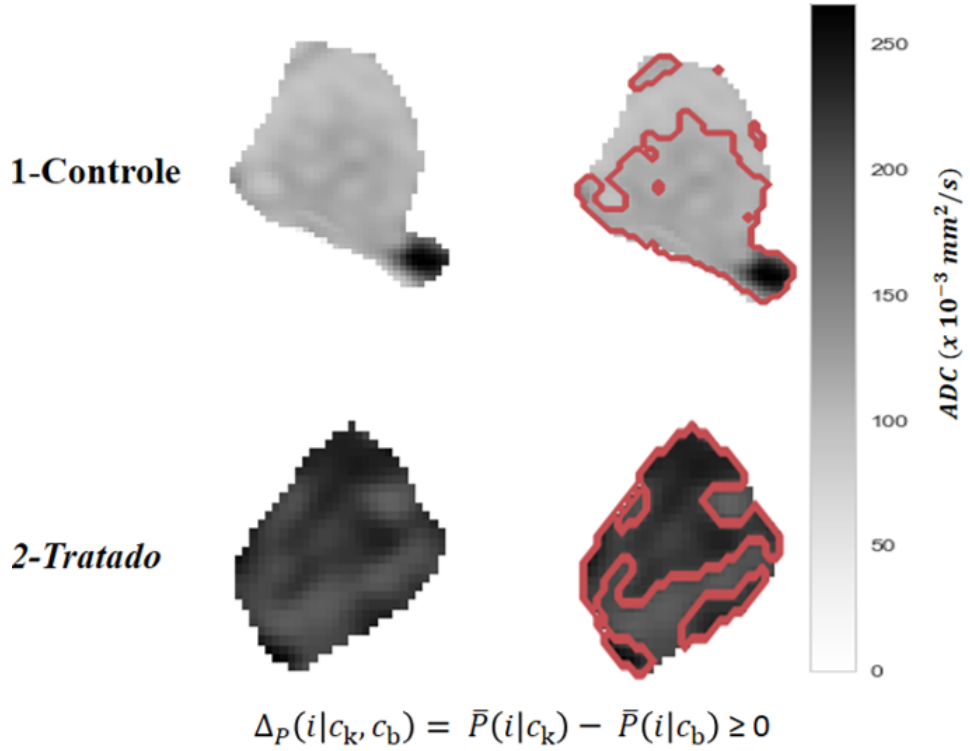


Regiões em azul mostram intensidade em ADC onde a diferença, $\Delta_P(i|c_1,c_2)$, entre os respectivos grupos são menor que zero. Em vermelho, regiões onde tal diferença é maior ou igual a zero.

Comparado ao grupo controle (Grupo A, Figura 40a e 40c), as curvas associadas ao grupo tratado (Grupo B, Figura 40b e 40d) apresentaram maiores diferença e distância com relação ao *baseline*, além de um comportamento mais homogêneo ao longo do mapa ADC de intensidades. Tais diferenças foram proeminentes em regiões onde $\Delta_P(i|c_1,c_2) > 0$, correspondendo, em média, a intensidades, i, entre $114x10^{-3} \leq i \leq 170x10^{-3} mm^2/s$ (Figura 40b e 40d). Isso não apenas mede o efeito do tratamento, assim como indica as regiões de intensidade nas quais estes efeitos foram mais proeminentes. A Figura 41 exibe por grupo, as correspondentes regiões mais

proeminentes, à nível da imagem, derivadas dos gráficos da Figura 40.

Figura 41 – Região discriminativa por grupo com relação ao *baseline*.



Os contornos vermelhos, nas ROIs à direita, evidenciam regiões das imagens originais (à esquerda) onde a diferença entre as classes c_k ($k = A$ e $k = B$ representam os grupos controle e tratado), com relação ao *baseline*, c_b, são maiores que zero. No grupo tratado, essa diferença $\Delta_P(i|c_k, c_b) > 0$, cobre áreas com maior intensidade, as quais apresentam maior verossimilhança de presença de efeito do tratamento, segundo os mecanismos biológicos conhecidos na mapeamento do ADC, como pode ser visto em (PADHANI et al., 2009).

Tabela 6 – Desempenho por característica: não-adaptativo *versus* adaptativo. *Friedman* teste mostrou diferença significativa ($p < 0.0001$) entre o desempenho dos grupos de características.

	$\Delta Mean$	$\Delta Skewness$	$\Delta Kurtosis$	ADC $F+$	ADC $F-$
Sens(%)	66.67	80.00	71.43	83.30	100.0
	(1.50,50.0-80.0)	(1.50,62.50-85.71)	(1.60,50.0-83.33)	(1.25,75.0-100.0)	(0.99,80.0-100.0)
Especif (%)	75.00	75.00	75.00	75.00	75.00
	(1.50,50.0-80.0)	(1.50,50.0-80.0)	(1.50,50.0-80.0)	(1.50,50.0-80.0)	(1.50,50.0-80.0)
PPV(%)	62.50	66.67	60.00	80.00	100.00
	(1.25,50.0-75.0)	(0.99,60.0-80.0)	(1.25,50.0-75.0)	(1.05,66.67-100.0)	(1.25,75.0-100.0)
NPV(%)	75.00	75.00	71.43	80.0	83.33
	(0.83,66.67-83.33)	(0.83,66.67-83.33)	(0.14,62.50-83.33)	(1.43,71.43-100.0)	(1.25,75.0-100.0)
AUC(%)	67.50	73.21	65.71	68.75	75.0
	(0.09,62.50-75.0)	(0.12,62.50-80.0)	(0.13,57.14-75.0)	(0.17,58.33-80.0)	(0.16,66.67-91.67)

6 CONCLUSÕES E PERSPECTIVAS FUTURAS

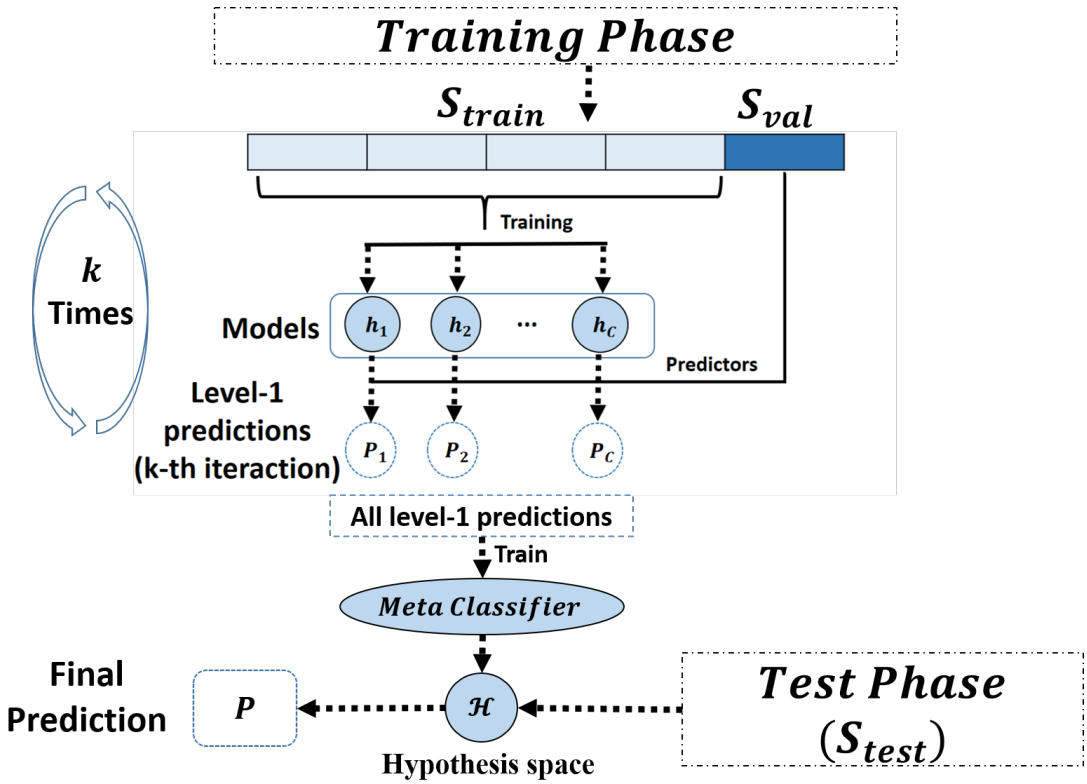
Para o caso 1, envolvendo discriminação entre linfonodos benignos e malignos de câncer de útero, demonstrou-se neste trabalho que, comparado à abordagem univariada, obtém-se melhor desempenho ao combinar variáveis, independente do modelo de aprendizagem de máquina utilizado (Figura 35).

Nesse sentido, observou-se que ao combinar entre duas (F3: ADC Mediana e F6: ADC assimetria), e até quatro características, os médicos ganham significativo poder de diagnóstico em termos de área sobre a curva ROC (AUC), se comparado somente ao que eles tendem a utilizar na rotina clínica, F2 (ADC Médio). Todavia, as vantagens de utilização de duas, com relação a quatro variáveis, são a possibilidade de visualização do espaço vetorial, fator importante para os médicos, assim como a redução do risco de *overfitting* (ou perda de generalização) por parte dos modelos de aprendizagem, o qual aumenta com o aumento do número de variáveis.

Investigações futuras, na próxima fase desse projeto, incluirão avaliação dos modelos construídos neste trabalho em diferentes tumores, assim como *scanners* MRI de diferentes fabricantes, em uma perspectiva multi-centro. Com isso, questões como influência do tipo de tumor, assim como padronização dos protocolos de aquisição de DWI e correspondentes padrões de qualidade de imagem em termos da razão sinal-ruído poderão ser amplamente analisadas, afim de verificar estabilidade e precisão dos modelos de aprendizagem sob essas óticas. Para isso, o grupo fechou algumas parcerias. Talvez a mais importante com a *Philippis image healthcare*, em seu projeto *Intelligent Porto*: uma interface web que permite visualização e mineração de imagens médicas, com uma série de recursos para segmentação automática já funcionais.

Além disso, dentro dessa perspectiva, um passo a mais na construção dos modelos de aprendizagem está em andamento, conforme ilustrado na Figura 42. Este passo consiste em combinar, dessa vez, não somente descritores, mas também classificadores, usando uma espécie de empilhamento de classificadores. Nessa abordagem, os resultados de predição de cada modelo (h_c) serve como entrada para um meta-classificador que aprenderá com os erros cometidos por cada modelo individualmente durante a fase de treinamento (Figura 42), potencialmente aumentando o desempenho em casos clínicos difíceis. Outra possibilidade que está sendo levada em conta é a utilização de métodos como *Deep Learning* (LECUN; BENGIO; HINTON, 2015), embora para isso precisemos de um banco de dados massivamente maior, o que tem implicações de ordem logística, impondo assim o uso de *Deep Learning* numa fase posterior, para a qual as ferramentas de segmentação do *Intelligent Porto* serão fundamental auxiliando os médicos.

Figura 42 – Arquitetura para combinação dos modelos de aprendizagem de máquina usados neste trabalho.



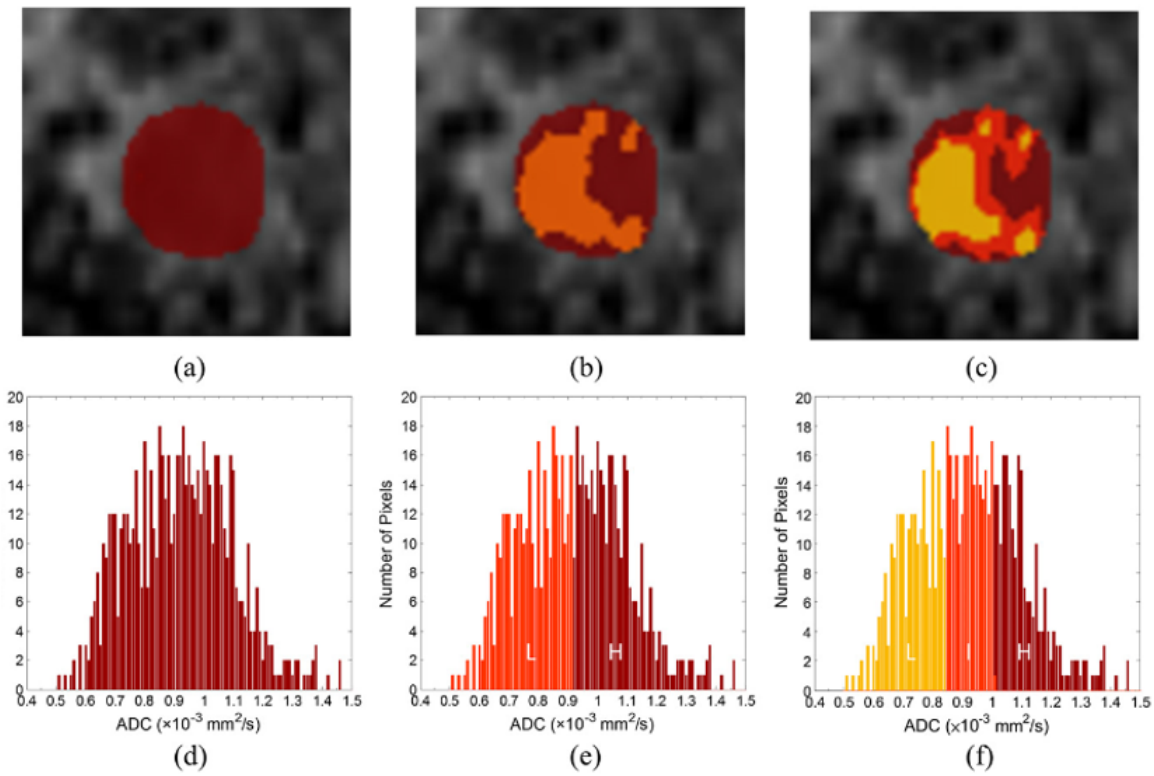
Por outro lado, no caso clínico 2, onde a meta era implementar e avaliar um número exaustivo de características para quantificar o efeito do tratamento em ratos com tumores gastrointestinais, observou-se que apenas algumas características apresentaram diferença estatística significativa ($p < 0.05$). Entre estas características discriminativas, observou-se equivalente desempenho entre descritores de primeira, segunda e terceira ordens estatística, em termos de diferenças relativas entre os grupos controle e tratados, com relação ao *baseline*. Aqui, novamente, ADC assimetria e curtose estiveram entre os descritores promissores (Tabela 5), embora com desempenho sub-ótimos considerando o grau de variabilidade apresentado durante a análise *bootstrap* (Figura 39).

Dessa feita, afim de lidar com os limites das métricas quantitativas *não-adaptativas* na avaliação da resposta ao tratamento, descritores adaptativos foram propostos. O raciocínio de base é que, embora haja regiões de resistência ao tratamento, refletidos nos resultados sub-ótimos dos métodos *não-adaptativas*, existem regiões específicas na distribuição do mapa ADC que contém, ao longo do tempo, as regiões onde o tratamento surtiu maior efeito (regiões discriminativas). A magnitude e a localização espacial dessas regiões discriminativas dependerão da extensão dos efeitos da droga, bem como de todos os processos biológicos associados. Contudo, tal magnitude e localização só foram possíveis de serem capturadas pelo método *adaptativo* proposto neste trabalho. Através deste método, todas as áreas ao nível da imagem, onde um determinado tratamento causa alterações, podem ser mapeadas

e visualizadas automaticamente, usando uma estratégia binária, tal como os médicos usam na análise de mapa ADC/DWI: observando regiões de aumento ou diminuição de atenuação de sinal. Seguindo esse raciocínio, o espaço de características foi naturalmente reduzido de d para apenas duas dimensões ($\mathbb{R}^d \rightarrow \mathbb{R}^2$), sem exigir procedimentos extras de redução de dimensionalidade, como PCA (*Principal Component Analysis*) ou LDA (*Linear Discriminant Analysis*) (THEODORIDIS; KOUTROUMBAS, 2008).

Numa próxima fase deste projeto, o método aqui desenvolvido será validado em humanos. Em adição, considerando que o filtro fornecido pelo método aqui proposto se assemelha de alguma forma a métodos de *clusters*, comparações serão realizadas com *Fuzzy C-Means Clusters*, o qual vem sendo utilizado na segmentação de imagens por difusão (LEE et al., 2016), como mostra a Figura 43.

Figura 43 – Cluster de mapa ADC baseado em lógica *Fuzzy* em processo de implementação.



Sobreposição de *clusters* de coeficientes de difusão aparente (ADC) assumindo (a)um, (b)dois e (c)três *clusters*. Seus correspondentes histogramas são mostrados nas figuras (e-f). Nesta Figura, *H* significa *cluster* com alto ADC; *I*, cluster com intermediário ADC; e *L*, cluster com baixo ADC.

Fonte: (LEE et al., 2016)

REFERÊNCIAS

AERTS, H.; VELAZQUEZ, E.; LEIJENAAR, R.; PARMAR, C.; GROSSMANN, P.; CARVALHO, S.; BUSSINK, J.; MONSHOUWER, R.; HAIBE-KAINS, B.; RIETVELD, D.; HOEBERS, F.; RIETBERGEN, M.; LEEMANS, C.; DEKKER, A.; QUACKENBUSH, J.; GILLIES, R.; LAMBIN, P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*, v. 5, 2014. ISSN 2041-1723.

ALESSANDRO, A. M. Ressonância magnética: princípios de formação da imagem e aplicações em imagem funcional. *Revista Brasileira de Física Médica*, v. 3, n. 3, p. 117–29, 2009.

AMADASUN, M.; KING, R. Textural features corresponding to textural properties. *Systems, Man and Cybernetics, IEEE Transactions on*, v. 19, p. 1264–1274, 1989.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. 1th. ed. London: Chapman and Hall, 1993.

BIHAN, D. L.; BRETON, E. Imagerie de diffusion in-vivo par résonance magnétique nucléaire. *Comptes-Rendus de l'Académie des Sciences*, v. 93, n. 5, p. 27–34, dez. 1985. Disponível em: ⟨https://hal.archives-ouvertes.fr/hal-00350090⟩.

BIHAN, D. L.; BRETON, E.; LALLEMAND, D.; GRENIER, P.; CABANIS, E.; LAVAL-JEANTET, M. Mr imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology*, v. 161, n. 2, p. 401–407, 1986. PMID: 3763909.

BOELLAARD, R. Standards for pet image acquisition and quantitative data analysis. *J. Nucl. Med.*, v. 500, p. 11S–20S, 2009.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: ⟨https://doi.org/10.1023/A:1010933404324⟩.

CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, ACM, v. 2, n. 3, p. 27:1–27:27, maio 2011. ISSN 2157-6904. Disponível em: ⟨http://doi.acm.org/10.1145/1961189.1961199⟩.

CHEN, F.; KEYZER, F. D.; FENG, Y.-B.; CONA, M. M.; YU, J.; MARCHAL, G.; OYEN, R.; NI, Y.-C. Separate calculation of dw-mri in assessing therapeutic effect in liver tumors in rats. *World J. Gastroenterol.*, v. 47, n. 19, p. 9092–9103, Dec 2013.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565. Disponível em: ⟨https://doi.org/10.1023/A:1022627411411⟩.

DOWNEY, K.; RICHES, S. F.; MORGAN, V. A.; GILES, S. L.; ATTYGALLE, A. D.; IND, T. E.; BARTON, D. P. J.; SHEPHERD, J. H.; DESOUZA, N. M. Relationship Between Imaging Biomarkers of Stage I Cervical Cancer and Poor-Prognosis Histologic Features: Quantitative Histogram Analysis of Diffusion-Weighted MR Images. *American Journal of Roentgenology*, v. 200, n. 2, p. 314–320, feb 2013. ISSN 0361-803X. Disponível em: ⟨http://www.ajronline.org/doi/abs/10.2214/AJR.12.9545⟩.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification (2Nd Edition)*. [S.l.]: Wiley-Interscience, 2000. ISBN 0471056693.

FOROUTAN, P.; KREAHLING, J.; MORSE, D.; GROVE, O.; LLOYD, M.; REED, D. Diffusion MRI and Novel Texture Analysis in Osteosarcoma Xenotransplants Predicts Response to Anti-Checkpoint Therapy. *PLoS ONE*, v. 8, n. 12, 2013.

GALLOWAY, M. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, v. 4(2), p. 172–179, 1975.

GEORGES, D.; MASATO, K. Bootstrap resampling for unbalanced data in supervised learning. *European Journal of Operational Research*, v. 134, n. 1, p. 141–156, 2001.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine Learning*, v. 36, n. 1, p. 3–42, 2006. Disponível em: ⟨http://www.montefiore.ulg.ac.be/services/stochastic/pubs/2006/GEW06a⟩.

GONZALEZ, R. C.; WOODS, R. *Digital Image Processing*. 3 edição. ed. [S.l.]: Wiley, 2007.

GREGOIRE, V.; CHITI, A. Pet in radiotherapy planning: Particularly exquisite test or pending and experimental tool? *Radiotherapy and Oncology*, v. 96, p. 275–276, 2010.

GROVE, O.; ANDERS, E. B.; MATTHEW, B.; SCHABATH, H. J. W.; AERTS, L.; DEKKER, L.; WANG, H.; VELAZQUEZ, E. R.; LAMBIN, P.; GU, Y.; BALAGURUNATHAN, Y.; EIKMAN, E.; GATENBY, R. A.; ESCHRICH, S.; GILLIES, R. J. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PLoS ONE10*, v. 3, 2015. Disponível em: ⟨https://doi.org/10.1371/journal.pone.0118261⟩.

GUO, Y.; CAI, Y.; CAI, Z. Differentiation of clinically benign and malignant breast lesions using diffusion-weighted imaging. *J. Magn. Reson. Imaging*, v. 16, p. 172–178, 2002.

HARALICK, R. Statistical and structural approaches to texture. *Proceedings of the IEEE*, v. 67, p. 786–804, 1973.

JANEZ, D. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, v. 7, p. 1–30, 2006. Disponível em: ⟨http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf⟩.

JAPKOWICZ, N.; SHAH, M. *Evaluating learning algorithms: a classifiation perspective*. 1th. ed. Cambridge/New York: Cambridge University Press, 2011. ISBN 0521196000,9780521196000.

JONES, D. *Diffusion MRI: Theory, Methods, and Application*. Oxford University Press, 2010. ISBN 9780199828678. Disponível em: ⟨https://books.google.be/books?id=7X9DAQAACAAJ⟩.

JUST, N. Histogram analysis of the microvasculature of intracerebralhuman and murine glioma xenografts. *Magn Reson Medr*, v. 65, n. 12, p. 778–789, dec 2011.

JUST, N. Improving tumour heterogeneity MRI assessment with histograms. *British Journal of Cancer*, v. 111, n. 12, p. 2205–2213, dec 2014. ISSN 0007-0920. Disponível em: ⟨http://www.nature.com/doifinder/10.1038/bjc.2014.512⟩.

KING, A.; CHOW, K.; YU, K.; MO, F.; YEUNG, D.; YUAN, J.; BHATIA, K.; VLANTIS, A.; AHUJA, A. Head and neck squamous cell carcinoma: diagnostic performance of diffusion-weighted MR imaging for the prediction of treatment response. v. 266, n. 2, p. 531–8, Feb 2014.

Koh, DM. Qualitative and Quantitative Analyses: Image Evaluation and Interpretation. In: Koh D.M., T. H. (Ed.). *Diffusion-Weighted MR Imaging*. 1. ed. [S.l.: s.n.], 2010. cap. 3.

KOUROU, K.; EXARCHOS, T. P.; EXARCHOS, K. P.; KARAMOUZIS, M. V.; FOTIADIS, D. I. *Machine learning applications in cancer prognosis and prediction*. 2015.

KUANG, F.; REN, J.; ZHONG, Q.; LIYUAN, F.; HUAN, Y.; CHEN, Z. The value of apparent diffusion coefficient in the assessment of cervical cancer. *Eur Radiol.*, v. 4, n. 23, p. 1050–8, Apr 2013.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, May 28 2015.

LEE, M.-C.; CHUANG, K.-S.; CHEN, M.-K.; LIU, C.-K.; LEE, K.-W.; TSAI, H.-Y.; LIN, H.-H. Fuzzy C-means clustering of magnetic resonance imaging on apparent diffusion coefficient maps for predicting nodal metastasis in head and neck cancer. *The British Journal of Radiology*, The British Institute of Radiology., v. 89, n. 1063, p. 20150059, jul 2016. ISSN 0007-1285. Disponível em: ⟨http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5257296/⟩.

LORENSEN, W. E.; CLINE, H. E. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, ACM, New York, NY, USA, v. 21, n. 4, p. 163–169, ago. 1987. ISSN 0097-8930. Disponível em: ⟨http://doi.acm.org/10.1145/37402.37422⟩.

MANN, H.; WHITNEY, D. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, v. 18, p. 50–60, 1947.

MCVEIGH, P. Z.; SYED, A. M.; MILOSEVIC, M.; FYLES, A.; HAIDER, M. A. Diffusion-weighted MRI in cervical cancer. *European Radiology*, v. 18, n. 5, p. 1058–1064, may 2008. ISSN 0938-7994. Disponível em: ⟨http://link.springer.com/10.1007/s00330-007-0843-3⟩.

MERBOLDT, K.; HANICKE, W.; FRAHM, J. Self-diffusion nmr imaging using stimulated echoes. *Journal of Magnetic Resonance (1969)*, v. 64, n. 3, p. 479 – 486, 1985. ISSN 0022-2364. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/0022236485901118⟩.

MYERS, J. L.; WELL, A. D. *Research Design & Statistical Analysis*. 1. ed. [S.l.]: Routledge, 1995. ISBN 0805820671.

NAQA, E.; ISSAM, L.; RUIJIANG, M.; J., M. *Machine Learning in Radiation Oncology*. 1. ed. [S.l.]: Springer International Publishing, 2015. 336 p. ISBN 978-3-319-18305-3.

NAQA, I. E.; GRIGSBY, P.; APTE, A.; KIDD, E.; DONNELLY, E.; KHULLAR, D.; CHAUDHARI, S.; YANG, D.; SCHMITT, M.; LAFOREST, R.; THORSTAD, W.; DEASY, J. O. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.*, v. 42(6):, p. 1162–1171, 2009.

NEIL, J. Measurement of water motion (apparent diffusion) in biological systems. *Concepts Magn Reson*, v. 9, p. 385–401, 1997.

NOWOSIELSKI, M.; RECHEIS, W.; GOEBEL, G.; GÜLER, Ö.; TINKHAUSER, G.; KOSTRON, H.; SCHOCKE, M.; GOTWALD, T.; STOCKHAMMER, G.; HUTTERER, M. Adc histograms predict response to anti-angiogenic therapy in patients with recurrent high-grade glioma. *Neuroradiology*, v. 53, n. 4, p. 291–302, Apr 2011. ISSN 1432-1920. Disponível em: ⟨https://doi.org/10.1007/s00234-010-0808-0⟩.

ORLHAC, F.; SOUSSAN, M.; MAISONOBE, J.; GARCIA, C.; VANDERLINDEN, B.; BUVAT, I. Tumor texture analysis in 18f-fdg pet: Relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J. Nucl. Med.*, v. 55, p. 414–422, 2014.

PADHANI, A. R.; LIU, G.; KOH, D. M.; CHENEVERT, T. L.; THOENY, H. C.; TAKAHARA, T.; DZIK-JURASZ, A.; ROSS, B. D.; Van Cauteren, M.; COLLINS, D.; HAMMOUD, D. A.; RUSTIN, G. J. S.; TAOULI, B.; CHOYKE, P. L. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. *Neoplasia (New York, N.Y.)*, v. 11, n. 2, p. 102–25, feb 2009. ISSN 1476-5586. Disponível em: ⟨http://www.ncbi.nlm.nih.gov/pubmed/19186405http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2631136⟩.

PALHARES, D.; AZEVEDO, T.; VENEZIANI, A.; GADIA, R.; AFFONSO, R.; CANTON, H. P.; SPADIM, M. D.; MARCONI, D. The predictive impact of tumor volume assessment using magnetic resonance imaging before and during chemorradiation in patients with locally advanced cervical cancer. *International Journal of Radiation Oncology*Biology*Physics*, v. 99, n. 2, Supplement, p. E301 – E302, 2017. ISSN 0360-3016. Proceedings of the American Society for Radiation Oncology. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S0360301617323775⟩.

PAQUET, N.; ALBERT, A.; FOIDART, J.; HUSTINX, R. Within-patient variability of 18f-fdg: Standardized uptake values in normal tissues. *J. Nucl. Med.*, v. 45, p. 784–788, 2004.

PARMAR, C.; GROSSMANN, P.; BUSSINK, J.; LAMBIN, P.; AERTS, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports*, v. 5, n. 1, p. 13087, oct 2015. ISSN 2045-2322. Disponível em: ⟨http://www.nature.com/articles/srep13087⟩.

PAUL, E. K.; JAMES, W. F. Pet/ct standardized uptake values (suvs) in clinical practice and assessing response to therapy. *Semin Ultrasound CT MR*, v. 6, p. 496–505, 2010.

PAYNE, G.; SCHMIDT, M.; MORGAN, V.; GILES, S.; BRIDGES, J.; IND, T.; DESOUZA, N. Evaluation of magnetic resonance diffusion and spectroscopy measurements as predictive biomarkers in stage 1 cervical cancer. *Gynecol Oncol.*, v. 2, n. 116, p. 246–52, Feb 2010.

REBEKKA, K.; GASPAR, D.; SIBYLLE, I. Z. Simulation study of tissue-specific positron range correction for the new biograph mmr whole-body pet/mr system. *IEEE Trans. Nucl. Sci.*, v. 59, p. 1900, 2012. Disponível em: ⟨http://dx.doi.org/10.1109/TNS.2012.2207436⟩.

ROKACH, L.; MAIMON, O. *Clustering Methods in Data Mining and Knowledge Discovery Handbook*. 1. ed. New York: Springer-Verlag, 2005. 321-352 p. Disponível em: ⟨http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471494631.html⟩.

SATTLER, B.; LEE, J. A.; LONSDALE, M.; COCHE, E. Pet/ct (and ct) instrumentation, image reconstruction and data transfer for radiotherapy planning. *Radiotherapy and Oncology*, v. 96, p. 288–297, 2010.

SCHWARTZ, W. R.; PEDRINI, H. Evaluation of Feature Descriptors for Texture Classification. *Journal of Electronic Imaging*, v. 21, p. 1–17, 2012.

SCOTT, D. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York, Chicester,: John Wiley & Sons, 1992.

STEJSKAL, E. O.; TANNER, J. E. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *The Journal of Chemical Physics*, v. 42, n. 1, p. 288–292, 1965.

SULLIVAN, F.; ROY, S.; EARY, J. A statistical measure of tissue heterogeneity with application to 3d pet sarcoma data. *Biostatistics*, v. 4, p. 433–448, 2003.

TAKAHARA, T.; IMAI, Y.; YAMASHITA, T.; YASUDA, S.; NASU, S.; CAUTEREN, M. V. Diffusion weighted whole body imaging with background body signal suppression (dwibs): technical improvement using free breathing, stir and high resolution 3d display. *Radiat Med.*, v. 4, n. 22, p. 275–82, Jul-Aug 2004.

TAN, S.; KLIGERMAN, S.; CHEN, W.; LU, M.; KIM, G.; FEIGENBERG, S.; D'SOUZA, W.; SUNTHARALINGAM, M.; LU, W. Spatial-temporal [^{18}f]fdg-pet features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy. *Int J Radiat Oncol Biol Phys.*, v. 85, n. 5, p. 1375–82, 2012.

TAYLOR, D.; BUSHELL, M. The spatial mapping of translational diffusion coefficients by the nmr imaging technique. *Physics in Medicine and Biology*, v. 30, n. 4, p. 345–349, 1985. Disponível em: ⟨https://hal.archives-ouvertes.fr/hal-00350090⟩.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition, Fourth Edition*. 4th. ed. [S.l.]: Academic Press, 2008. ISBN 1597492728, 9781597492720.

THORWARTH, D.; GEETS, X.; PAIUSCO, M. Physical radiotherapy treatament planning based on functional pet/ct data. *Radiotherapy and Oncology*, v. 96, p. 317–324, 2010.

TIXIER, F.; REST, C. L.; HATT, M.; ALBARGHACH, N.; PRADIER, O.; METGES, J.; CORCOS, L.; VISVIKIS, D. Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer. *J. Nucl. Med.*, v. 52(3), p. 369–78, 2011.

TOZER, D. J.; JäGER, H. R.; DANCHAIVIJITR, N.; BENTON, C. E.; TOFTS, P. S.; REES, J. H.; WALDMAN, A. D. Apparent diffusion coefficient histograms may predict low-grade glioma subtype. *NMR in Biomedicine*, John Wiley & Sons, Ltd., v. 20, n. 1, p. 49–57, 2007. ISSN 1099-1492. Disponível em: ⟨http://dx.doi.org/10.1002/nbm.1091⟩.

VANDECAVEYE, V.; KEYZER, F. D.; VANDER, P. V.; DIRIX, P.; VERBEKEN, E.; NUYTS, S.; R., H. Head and neck squamous cell carcinoma: Value of diffusion-weighted mr imaging for nodal staging 1. *Radiology*, v. 251, 2009.

WANG, J.; TAKASHIMA, S.; TAKAYAMA, F.; KAWAKAMI, S.; SAITO, A.; MATSUSHITA, T.; MOMOSE, M.; ISHIYAMA, T. Head and neck lesions: Characterization with diffusion-weighted echo-planar mr imaging. *Radiology*, v. 220, n. 3, p. 621–630, 2001. PMID: 11526259.

WEBER, W. Use of pet for monitoring cancer therapy and predicting outcome. *J. Nucl. Med.*, v. 46(6), p. 983–95, 2005a.

WEBER, W. Quantitative analysis of pet studies. *Radiotherapy and Oncology*, v. 96, p. 308–310, 2010.

XUE, H.; REN, C.; YANG, J.; SUN, Z.; LI, S.; JIN, Z.; SHEN, K.; ZHOU, W. Histogram analysis of apparent diffusion coefficient for the assessment of local aggressiveness of cervical cancer. *Archives of Gynecology and Obstetrics*, v. 290, n. 2, p. 341–348, aug 2014. ISSN 0932-0067. Disponível em: ⟨http://link.springer.com/10.1007/s00404-014-3221-9⟩.

YANG, F.; THOMAS, M.; DEHDASHTI, F.; GRIGSBY, P. Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer. *Eur J Nucl Med Mol Imaging*, v. 40, n. 5, p. 716–27, 2013.

YOUDEN, W. J. Index for rating diagnostic tests. *Cancer*, Wiley Subscription Services, Inc., A Wiley Company, v. 3, n. 1, p. 32–35, 1950. ISSN 1097-0142. Disponível em: ⟨http://dx.doi.org/10.1002/1097-0142(1950)3:1⟨32::AID-CNCR2820030106⟩3.0.CO;2-3⟩.

YU, H. *Automated Segmentation of Head and Neck Cancer Using Texture Analysis with Co-Registered PET/CT Images*. Tese (Doutorado) — University of Toronto, 2010.

ZHANG, X.; XU, X.; TIAN, Q.; LI, B.; WU, Y.; YANG, Z.; LIANG, Z.; LIU, Y.; CUI, G.; LU, H. Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging. *J. Magn. Reson. Imaging.*, 2017. Disponível em: ⟨http://dx.doi.org/10.1002/jmri.25669⟩.

ZWANENBURG, A.; LEGER, S.; VALLIÈRES, M.; LÖCK, S. Image biomarker standardisation initiative - feature definitions. *CoRR*, abs/1612.07003, 2016. Disponível em: ⟨http://arxiv.org/abs/1612.07003⟩.