



TESE DE DOUTORADO

**RADIÔMICA E APRENDIZADO DE MÁQUINA PARA
IDENTIFICAÇÃO DA CODELEÇÃO CROMOSSÔMICA 1p/19q EM
GLIOMAS DE BAIXO GRAU**

TONY ALEXANDRE MEDEIROS DA SILVA

Brasília, Fevereiro de 2023

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

**RADIÔMICA E APRENDIZADO DE MÁQUINA PARA
IDENTIFICAÇÃO DA CODELEÇÃO CROMOSSÔMICA
1p/19q EM GLIOMAS DE BAIXO GRAU**

Tony Alexandre Medeiros da Silva

ORIENTADOR: João Luiz Azevedo de Carvalho

COORIENTADOR: Francisco Assis de Oliveira Nascimento

TESE DE DOUTORADO

Publicação: PGEA.TD 195/23

Brasília/DF: Fevereiro – 2023

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA

TESTE DE DOUTORADO EM ENGENHARIA DE SISTEMAS
ELETRÔNICOS E AUTOMAÇÃO

**RADIÔMICA E APRENDIZADO DE MÁQUINA PARA
IDENTIFICAÇÃO DA CODELEÇÃO CROMOSSÔMICA 1p/19q EM
GLIOMAS DE BAIXO GRAU**

Tony Alexandre Medeiros da Silva

*Tese de doutorado submetida ao Departamento de Engenharia Elétrica da
Faculdade de Tecnologia da Universidade de Brasília, como parte dos
requisitos necessários para a obtenção do grau de doutor.*

Banca Examinadora

Prof. João Luiz Azevedo de Carvalho, Ph.D. ENE/UnB (Orientador)	_____
Prof. Francisco Assis de Oliveira Nascimento, Ph.D. ENE/UnB (Coorientador)	_____
Prof. Flávia Maria Guerra S. Aranha Oliveira, Ph.D. ENE/UnB (Examinador Interno)	_____
Prof. João Souza Neto, Dr. CCET/UFS (Examinador Externo)	_____
Prof. Túlio Augusto Alves Macedo, Dr. FAMED/UFU (Examinador Externo)	_____
Prof. Cristiano Jacques Miosso, Ph.D. ENE/UnB (Suplente)	_____

FICHA CATALOGRÁFICA

SILVA, TONY ALEXANDRE MEDEIROS DA

Radiômica e aprendizado de máquina para identificação da codeleção cromossômica 1p/19q em gliomas de baixo grau.

[Distrito Federal] 2023.

xviii, 123p., 210 x 297 mm (ENE/FT/UnB, Doutor, Tese de Doutorado).

Universidade de Brasília. Faculdade de Tecnologia.

Departamento de Engenharia Elétrica

- | | |
|--------------------------------|---------------------------|
| 1. Radiômica | 2. Aprendizado de máquina |
| 3. Extração de características | 4. Gliomas de baixo grau |
| 5. Biomarcador de imagem | 6. PyRadiomics |
| I. ENE/FT/UnB | II. Título (série) |

REFERÊNCIA BIBLIOGRÁFICA

SILVA, T. A. M. (2023). Radiômica e aprendizado de máquina para identificação da codeleção cromossômica 1p/19q em gliomas de baixo grau. Tese de Doutorado em Engenharia de Sistemas Eletrônicos de Automação, Publicação PGEA.TD-195/23, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 123p.

CESSÃO DE DIREITOS

AUTOR: Tony Alexandre Medeiros da Silva.

TÍTULO: Radiômica e aprendizado de máquina para identificação da codeleção cromossômica 1p/19q em gliomas de baixo grau.

GRAU / ANO: Doutor / 2023

É concedida à Universidade de Brasília permissão para reproduzir cópias desta tese de doutorado e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte dessa tese de doutorado pode ser reproduzida sem autorização por escrito do autor.

Tony Alexandre Medeiros da Silva

Rua Geraldino M. Rodrigues, 625, B. Granada.

38.410-134 Uberlândia - MG – Brasil.

Dedicatória

Esse trabalho é dedicado às pessoas que, por infelicidade, foram acometidas por algum tipo de tumor cerebral.

Tony Alexandre Medeiros da Silva

Agradecimentos

Primeiramente agradeço a Deus pela conclusão deste trabalho. Tenho certeza que em toda a caminhada ele esteve presente.

Em segundo lugar, gostaria de agradecer à minha amável esposa Kézia e aos meus filhos Gabryell, Adryell e Rafaell pelo carinho e imensa paciência.

Agradeço a 2 grandes amigos que esta etapa da vida me proporcionou, o professor e orientador João Luiz Azevedo de Carvalho e o professor e coorientador Francisco Assis de Oliveira Nascimento, presenças marcante neste trabalho, repassando conhecimentos e trocando experiências, e principalmente pelo constante exemplo de responsabilidade e respeito.

Também não posso deixar de agradecer o imenso apoio do neurorradiologista Dr. Guilherme Cássia, outro grande amigo que este trabalho me proporcionou. Simplesmente não media esforços e nem horário para me ajudar.

Aos professores do PGEA e PPGEE deixo aqui registrado minha profunda admiração pelo empenho com os quais ministram suas disciplinas e, principalmente, a forma respeitosa e educada com as quais tratam a comunidade acadêmica. Agradeço ao professor Francisco Assis de Oliveira Nascimento, por me permitir utilizar a Machine Learning Framework, e ao professor Marcelino Monteiro de Andrade por me apresentar a biblioteca PyRadiomics.

Ao GPDS (Grupo de Processamento Digital de Sinais) onde conheci pessoas fantásticas, mega inteligentes e um local super agradável para se trabalhar.

Ao grupo de pesquisa Machine Learning e todos os seus integrantes, por todos os artigos compartilhados, discussões, apresentações realizadas e construções de ideias que foram fundamentais neste trabalho.

Aos servidores das secretarias do PGEA e PPGEE pela eficiência. Nas várias vezes em que precisei de algo relacionado ao trabalho deles, foram 100% assertivos.

Aos técnicos em radiologia Marcello Lucio e Eduardo Batista, verdadeiros “feras” em exames de ressonância magnética, e dois grandes amigos que sempre me socorreram em momentos de dúvidas.

Ao Instituto Federal Goiano, por me permitir realizar parte deste trabalho com

dedicação total do tempo. Foi muito importante o afastamento de minhas atividades docentes em momento sensível da pesquisa. Obviamente, quero deixar registrado meu especial agradecimento aos colegas docentes, técnicos administrativos e demais prestadores de serviços do Campus Cristalina.

Gostaria de agradecer também aos membros que participaram da banca avaliadora, Prof. João Luiz Azevedo de Carvalho, Prof. Francisco Assis de Oliveira Nascimento, Profa. Flávia Maria Guerra de Sousa Aranha Oliveira, Prof. João Souza Neto, Prof. Túlio Augusto Alves Macedo, e Prof. Cristiano Jacques Miosso, pela cordialidade e pelos diversos comentários e sugestões que ajudaram a melhorar este trabalho e que, certamente, farão toda a diferença nas pesquisas futuras que esse projeto ainda irá proporcionar.

Finalmente, agradeço a todos que direta ou indiretamente contribuíram para a realização deste momento.

Tony Alexandre Medeiros da Silva

RESUMO

Os gliomas representam em torno de 40% dos tumores cerebrais. Destes, 50% são de baixo grau, sendo representados quase que em sua totalidade por oligodendrogliomas e astrocitomas, ambos de grau II ou III. Tendo em vista que a codeleção cromossômica 1p/19q é um importante marcador tumoral de fator prognóstico (favorável) positivo nesses gliomas, propõe-se a utilizar a radiômica na análise de imagens como um abrangente quantificador de fenótipos tumorais não invasivos para identificar a codeleção cromossômica 1p/19q em gliomas de baixo grau. Trata-se de uma abordagem potencialmente utilizada em oncologia, auxiliando na detecção, diagnóstico e prognóstico do câncer, previsão de resposta ao tratamento e monitoramento do estado da doença. Por meio da plataforma PyRadiomics, que extrai características radiômicas que quantificam a intensidade, a forma e a textura da imagem do tumor em exames de ressonância magnética, utilizou-se um dataset com 159 pacientes diagnosticados com gliomas de baixo grau e extraiu-se 120 características. Em seguida implementou-se 4 experimentos onde foram aplicadas técnicas de pré-processamento de dados, tais como seleção, escalonamento, redimensionamento e aumento de dados, e aprendizado de máquina supervisionado. Após avaliação de um conjunto de dados de teste em cada experimento, obteve-se resultados promissores para acurácia, sensibilidade, especificidade, precisão, area sob a curva (AUC) e pontuação F1. Os mesmos mostram-se favoráveis quando comparados ao estado da arte. Este estudo descreve uma pesquisa exploratória cujos resultados ressaltam que a radiômica em conjunto com aprendizado de máquina configura uma abordagem promissora na identificação da codeleção cromossômica 1p/19q em gliomas de baixo grau.

ABSTRACT

Gliomas represent approximately 40% of all brain tumors, and approximately 50% of all gliomas are low-grade, almost entirely represented by oligodendrogliomas and astrocytomas, both grade II or III. Since 1p/19q chromosome codeletion is an important tumor marker of positive prognostic (benign) factor in these gliomas, we propose the use of radiomics in image analysis as a comprehensive quantifier of non-invasive tumor phenotypes to identify 1p/19q chromosome codeletion in low-grade gliomas. This is an approach potentially used in oncology, helping in the detection, diagnosis and prognosis of cancer, prediction of response to treatment and monitoring of the state of the disease. Using the PyRadiomics platform, which extracts radiomic features that quantify the intensity, shape and texture of the tumor image in magnetic resonance images, a dataset with 159 patients diagnosed with low-grade gliomas was used and 120 features were extracted. Then, 4 experiments were implemented where data pre-processing techniques were applied, such as selection, scaling, resizing and data augmentation, and supervised machine learning. After evaluating a set of test data in each experiment, promising results were obtained for accuracy, sensitivity, specificity, precision, area under the curve (AUC) and F1 score. These results were competitive when compared to the state of the art. This study describes an exploratory research whose results underscore that radiomics in combination with machine learning constitutes a promising approach in identifying 1p/19q chromosome codeletion in low-grade gliomas.

Sumário

1	Introdução	19
1.1	Contextualização do problema	19
1.2	Motivação	21
1.3	Objetivos	21
1.4	Organização do trabalho	22
2	Trabalhos Correlatos	23
2.1	Análise radiômica	23
2.2	Extração de características quantitativas em tumores cerebrais	25
2.3	Inteligência artificial no auxílio ao diagnóstico por imagem	28
3	Conceitos Fundamentais	32
3.1	Tumores cerebrais	32
3.2	Processamento digital de imagens	34
3.3	Plataforma PyRadiomics	35
3.3.1	Segmentação do volume de interesse	36
3.3.2	Extração de características - PyRadiomics	37
3.4	Pré-processamento	42
3.4.1	Seleção das características	42
3.4.2	Escalonamento das características	43
3.4.3	Redimensionamento das características	48
3.4.4	Aumento das características	51
3.5	Algoritmos de classificação	55
3.5.1	Multilayer Perceptron (MLP)	57
3.5.2	Logistic Regression (RL)	59
3.5.3	Random Forest (RF)	59
3.5.4	K-Nearest Neighbor (KNN)	59
3.5.5	Support Vector Machine (SVM)	60
3.5.6	Bagging Classifier (BC)	60
3.5.7	Decision Tree (DT)	60
3.5.8	Gaussian Naive Bayes (GNB)	60

3.5.9	Adaptive Boosting (AB)	61
3.5.10	Stochastic Gradient Descent (SGD)	61
3.5.11	Histogram-based Gradient Boosting (HGB)	61
3.5.12	Extreme Gradient Boosting (XGB)	62
3.5.13	Extra Trees Classifier (ET)	62
3.5.14	Gradient Boosting (GB)	62
3.5.15	Linear Discriminant Analysis (LDA)	62
3.5.16	Grid Search CV	63
3.5.17	Método Ensemble	63
3.6	<i>Datasets</i> de imagens cerebrais	64
3.6.1	Protocolos T1, T2 e Flair	65
3.6.2	Bases de dados	67
4	Análise radiômica da codeleção cromossômica 1p/19q	70
4.1	Softwares e equipamentos	70
4.2	Dataset	70
4.3	Extração das características	72
4.4	Normalização das características	73
4.5	Métricas de avaliação	74
4.6	Compilado de técnicas	75
4.7	Experimento I	75
4.7.1	Objetivos	75
4.7.2	Métodos	75
4.7.3	Resultados	79
4.7.4	Discussão	81
4.8	Experimento II	82
4.8.1	Objetivos	82
4.8.2	Métodos	83
4.8.3	Resultados	86
4.8.4	Discussão	87
4.9	Experimento III	88
4.9.1	Objetivos	88
4.9.2	Métodos	88
4.9.3	Resultados	91
4.9.4	Discussão	93
4.10	Experimento IV	94
4.10.1	Objetivos	94
4.10.2	Métodos	94
4.10.3	Resultados	96

4.10.4 Discussão	100
4.11 Comparação com estado da arte	103
5 Conclusão	106
Referências Bibliográficas	108

Lista de Figuras

2.1	Etapas da análise radiômica	26
3.1	Exemplos de tumores gliais	33
3.2	Processo da codeleção 1p/19q	34
3.3	Exemplo de uma imagem crua <i>versus</i> processada	35
3.4	RM cerebral com sua respectiva segmentação	36
3.5	Etapas de um <i>pipeline</i> apresentado pela plataforma PyRadiomics.	38
3.6	Exemplo de distribuição dos dados escalonados com método de padronização	46
3.7	Exemplo de distribuição dos dados escalonados com método mínimo e máximo	47
3.8	Exemplo de distribuição dos dados escalonados com método de redução de outliers	47
3.9	Exemplo da técnica de aumento de dados	53
3.10	Gráficos com plotagem de exemplo da técnica SMOTE	54
3.11	Fluxo de um projeto em Machine Learning	55
3.12	Visão geral dos algoritmos de ML	56
3.13	Visão do aprendizado supervisionado	57
3.14	Estrutura do neurônio biológico <i>versus</i> neurônio artificial	58
3.15	Arquitetura de uma MLP	58
3.16	Diagrama simplificado do processo de classificação de subconjuntos para- lelos, como por exemplo, o <i>bagging</i>	63
3.17	Apresentação dos planos Sagital, Coronal e Axial	66
3.18	Exemplo de cortes no plano axial	66
3.19	Imagem de RM nas sequências T1, T2 e Flair	67
4.1	Divisão do dataset entre instâncias codeletados e não codeletados	71
4.2	Exemplo de LGG com cromossomo 1p/19q codeletado/não-codeletado . . .	71
4.3	Trecho do arquivo de entrada na plataforma PyRadiomics	72
4.4	Gráfico com resultado do redimensionamento	76
4.5	<i>Output</i> apresentando as classes do conjunto treinamento balanceadas . . .	77
4.6	Fluxo completo de execução do experimento 1	79
4.7	Curva ROC da MLP no experimento 1	80

4.8	Gráfico com resultado da seleção de características pela técnica de importância	83
4.9	Relação das características selecionadas pela biblioteca YellowBrick	84
4.10	Funcionamento do algoritmo RF	86
4.11	Matriz de correlação para o conjunto de dados, obtida após a redução da dimensionalidade NCA.	89
4.12	Boxplots dos valores de cada componente NCA (f1 a f4) para cada grupo de instâncias (azul: não deletado; laranja: codeletado).	89
4.13	Curva ROC para o resultado médio da assinatura completa	93
4.14	Curva PR para o resultado médio da assinatura completa	93
4.15	Curva ROC obtida pela média das 100 simulações do classificador <i>bagging</i> baseado em DT. A área sob a curva ROC foi de 0,842.	101

Lista de Tabelas

3.1	Exemplos de características radiômicas	41
3.2	Resultado comparativo do método <i>StandardScaler</i> nos algoritmos de classificação	44
3.3	Resultado comparativo entre dois métodos de normalização nos algoritmos de classificação	45
4.1	Quantitativo de características radiômicas extraídas	73
4.2	Compilado de técnicas e algoritmos usados nos métodos dos quatro experimentos.	75
4.3	Resumo das funções e parâmetros utilizados na MLP	78
4.4	Execução da MLP, SVM e RF no conjunto de dados de teste	79
4.5	Matriz de confusão do algoritmo MLP no conjunto de dados de testes . . .	80
4.6	Quantitativo de características radiômicas extraídas por classe no experimento 2	84
4.7	Lista de parâmetros de entrada usados na implementação dos classificadores RL e RF.	86
4.8	Desempenho dos algoritmos RL e RF no conjunto de dados de teste do experimento II	87
4.9	Desempenho de classificação alcançado após a incorporação de cada etapa na execução da assinatura. A assinatura completa (versão 4) está destacado em negrito. Os resultados em cada linha correspondem a uma média de 30 simulações.	91
4.10	Resultados da classificação individual obtidos em cada uma das 30 simulações da assinatura completo, valores médios e intervalos de confiança (ao nível de confiança de 95%.	92
4.11	Matriz de confusão obtida pela média das matrizes de confusão obtidas em cada uma das 30 simulações da assinatura completa.	92
4.12	Quantitativo de características radiômicas de textura	96
4.13	Lista de parâmetros de entrada usados na implementação de cada classificador.	97

4.14	Desempenho de classificação usando 68 características radiômicas baseadas em textura. Cada entrada corresponde a uma média de 100 simulações. O melhor resultado para cada métrica é destacado em negrito.	98
4.15	Desempenho de classificação usando 100 características radiômicas (textura + estatísticas de primeira ordem + forma). Cada entrada corresponde a uma média de 100 simulações. O melhor resultado para cada métrica é destacado em negrito.	99
4.16	Desempenho de classificação usando cada classificador como estimador base do classificador <i>bagging</i> . O subconjunto de 68 características radiômicas foi usado. Cada entrada corresponde a uma média de 100 simulações. O melhor resultado para cada métrica é destacado em negrito.	99
4.17	Resumo dos resultados obtidos com o melhor classificador geral: método <i>bagging</i> com um estimador de base de árvore de decisão (DT), usando um conjunto de 68 características radiômicas. Média, desvio padrão e intervalos de confiança de 95% são mostrados para os valores da matriz de confusão e para todas as métricas de classificação.	100
4.18	Comparação com estudos atualizados para classificação da codeleção 1p/19q baseados em radiômica e aprendizado de máquina.	103

Lista de Abreviações

2D	Bidimensional (duas dimensões)
3D	Tridimensional (três dimensões)
ACS	American Cancer Society
ADAM	Adaptive Moment Estimation
BraTS	Brain Tumor Segmentation Challenge
CAD	Computer Aided Diagnosis
CBICA	Center for Biomedical Image Computing and Analytics
CNN	Convolutional Neural Network
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DNA	Ácido desoxirribonucleico
FLAIR	Fluid Attenuated Inversion Recovery
GLCM	Gray Level Cooccurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
HGG	High-grade Gliomas
IA	Inteligência Artificial
IBSI	Imaging Biomarker Standardization Initiative
IDH	Isocitrate Dehydrogenase
INCA	Instituto Nacional de Câncer
KNN	k-Nearest Neighbor
LGG	Low-grade Gliomas
LoG	Laplacian of Gaussian
ML	Machine Learning
MLF	Machine Learning Framework
MLP	Multi Layer Perceptron
MSE	Mean Squared Error
NCA	Neighborhood Component Analysis
NCI	National Cancer Institute
NGTDM	Neighbouring Gray Tone Difference Matrix

NIFTI	Neuroimaging Informatics Technology Initiative
NITRC	NeuroImaging Tools and Resources Collaboratory
NRRD	Nearly Raw Raster Data
OMS	Organização Mundial da Saúde
PCA	Principal Component Analysis
PDI	Processamento Digital de Imagens
PET	Positron Emission Tomography
QIBA	Quantitative Imaging Biomarkers Alliance
QIN	Quantitative Imaging Network
ReLU	Rectified Linear Unit
RF	Random Forest
RFE	Recursive Feature Elimination
RL	Logistic Regression
RNSA	Radiological Society of North America
RNA	Ácido ribonucleico
RM	Ressonância Magnética
RNA	Redes Neurais Artificiais
ROI	Region of Interest
SMOTE	Synthetic Minority Over-sampling Technique
SNC	Sistema Nervoso Central
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TC	Tomografia Computadorizada
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive
UCI	University of California Irvine - Machine Learning Repository
US	Ultrassonografia
VOI	Volume of Interest

Capítulo 1

Introdução

O presente capítulo contextualiza a problematização, as principais motivações e os objetivos que impulsionaram o desenvolvimento deste trabalho. Em seguida, descreve as contribuições alcançadas e por último a organização do mesmo.

1.1 Contextualização do problema

O câncer é uma doença caracterizada pelo crescimento desordenado das células e se transformou em um problema de saúde mundial, sendo considerada uma das doenças crônicas não transmissíveis que constitui uma das principais causas de morte prematura, ou seja, antes dos 70 anos de idade.

O câncer é uma doença que desafia os sistemas de saúde do mundo todo. O Instituto Nacional de Câncer (INCA) previu para o último triênio (2020-2022) 625 mil casos por ano. O mesmo relata que nesse triênio, no Brasil, serão diagnosticados em torno de 11 mil novos casos de tumores do sistema nervoso central (SNC), sendo uma previsão de 5,6 casos masculinos e 4,8 casos femininos a cada 100 mil habitantes [1]. Quanto antes o diagnóstico for realizado melhor a chance de sobrevivência do paciente e menor os custos de tratamento, pois 48% dos diagnósticos são feitos em estágio avançado. Para viabilizar o diagnóstico precoce, uma alternativa é a utilização de dados de mutação genética, que possibilita o diagnóstico de forma não invasiva em pacientes com maior probabilidade de desenvolver a doença quando comparado com a população sem a mutação.

Basicamente, o tumor cerebral pode ser definido como um conjunto de células anormais que crescem de forma desordenada dentro do cérebro. Na formação do tumor, a divisão celular ocorre muito rapidamente, pois os mecanismos que controlam o crescimento das células são comprometidos. Dessa forma, esta célula poderá se transformar em um tumor.

Segundo a American Cancer Society [2], os tumores do SNC classificam-se em:

- Primários: originam-se a partir de células do cérebro ou de células que encontram-

se em seu entorno e normalmente permanecem dentro do cérebro. Estes tipos de tumores geralmente são chamados de benignos (por serem menos agressivos) ou malignos (por serem agressivos e se desenvolverem rápido).

- Metastáticos: também conhecidos como secundários, originam-se de outras partes do corpo, como, intestino ou mama, e se espalham pelo cérebro, principalmente através da corrente sanguínea. São tumores malignos.

Os tumores cerebrais podem variar conforme o grau de malignidade, localização e até mesmo em relação ao tipo de tecido. A OMS classificou os tumores cerebrais de acordo com características histológicas que refletem seu nível de agressividade. Os gliomas de baixo grau (LGG, do inglês *low grade gliomas*) são considerados tumores de menor agressividade. Dentre eles, os oligodendrogliomas e os astrocitomas, normalmente grau tipos I e II e com melhor prognóstico quando comparados aos gliomas de alto grau (HGG, do inglês *high grade gliomas*), como por exemplo o glioblastoma de grau tipo IV que apresenta crescimento acelerado [3]. Vale registrar que o glioblastoma, considerado um dos tumores mais agressivos dentre os HGGs, tem sido muito explorado por pesquisadores, médicos oncologistas, neuro-cirurgiões, e neurorradiologistas na última década [4, 5, 6].

Nas últimas duas décadas, soluções têm sido desenvolvidas para identificar padrões em dados genômicos, porém a maioria das abordagens já desenvolvidas só apresentam bons resultados quando realizam um grande esforço manual para extração de características. A extração manual de características requer tempo e domínio dos dados e limita inevitavelmente a eficiência, robustez e a capacidade de generalização. Além disso, os métodos baseados em inteligência artificial (IA) e aprendizado de máquina (do inglês, *machine learning* (ML)) em sua grande maioria utiliza dados de sequenciamento de genes, possibilitando a identificação das doenças com mutação gênica [7].

Por sua vez, os oligodendrogliomas apresentam uma mutação genética chamada codeleção cromossômica 1p/19q, a qual, quando identificada, determina que o tumor terá pequena chance de se tornar um HGG. Essa identificação da codeleção cromossômica 1p/19q não é uma ação inovadora, mas, seguramente uma proposta relevante de resultados promissores para o ser humano. Portanto, este trabalho de doutorado visa identificar essa alteração genética através do uso de características quantitativas extraídas de imagens de RM e com auxílio de técnicas de aprendizado de máquina.

Levando em consideração o que foi previamente descrito em parágrafos anteriores, foram consultados vários estudos científicos, pesquisas e trabalhos computacionais, os quais serão apresentados nesse trabalho, que propõem auxílio ao diagnóstico de tumores do SNC e análise da codeleção cromossômica 1p/19q. Uma das principais atividades que vem ganhando destaque nesta tentativa de melhor diagnosticar o glioma de baixo grau é o uso da radiômica, uma prática de extrair características quantitativas de imagens médicas para gerar dados que possam ser analisados e minerados. A radiômica tem como objetivo

encontrar padrões e correlações que melhorem a tomada de decisão clínica em pacientes com câncer [8].

Alguns trabalhos citados no capítulo 2, apresentam recursos de IA, técnicas, modelos e algoritmos de ML que proporcionaram a obtenção de resultados que contribuíram com uma melhor caracterização dos gliomas de baixo grau. Nesse contexto, os objetivos do presente trabalho são apresentados na seção a seguir.

1.2 Motivação

O reconhecimento de padrões em dados genômicos (DNA, RNA, proteínas) pode auxiliar no entendimento da progressão de algumas doenças, como na fisiopatologia endócrina e câncer. Na última década, os dados genômicos foram amplamente utilizados para classificar diferentes tipos de patologias [9]. Porém, os dados genômicos apresentam características que dificultam a qualidade do aprendizado e taxas de acerto dos algoritmos de predição, como o desbalanceamento das classes e a alta quantidade de características para um baixo número de amostras. Já existem alguns trabalhos que usam técnicas de inteligência artificial para obter avanços significativos na compreensão de bases biomoleculares de doenças severas como o câncer [10]. Motivado por estes trabalhos, serão desenvolvidos alguns experimentos com utilização das características radiômicas, técnicas de pré-processamento de dados e abordagem supervisionada de aprendizado de máquina para lidar com esses desafios (por exemplo, acesso a exames genéticos). A utilização dessas técnicas e abordagens permitem tornar explícito a contribuição de cada experimento realizado e a apresentação dos resultados que proporcionem uma interpretação promissora da classificação da codeleção cromossômica 1p/19q em gliomas de baixo grau.

1.3 Objetivos

Visando auxiliar os profissionais na área da saúde na caracterização dos gliomas de baixo grau através de uma análise radiômica utilizando imagens de ressonância magnética (RM) do cérebro, os objetivos deste trabalho incluem:

- (a) a apresentação do estado da arte em análise radiômica (marcadores tumorais) de gliomas de baixo grau;
- (b) o levantamento das principais técnicas de processamento de imagens, sendo aquelas com grande potencial para utilização neste trabalho;
- (c) utilização da biblioteca *PyRadiomics* para extração das características quantitativas utilizadas neste estudo para identificação da codeleção cromossômica 1p/19q.;

- (d) construção de experimentos com uso de diversos algoritmos de ML, técnicas de escalonamento, redução de dimensionalidade e aumento de dados, a fim de avaliar as características radiômicas extraídas;
- (e) analisar e discutir os resultados obtidos com os experimentos desenvolvidos, e verificar a possibilidade de considerá-los promissores na identificação da codeleção cromossômica 1p/19q.
- (f) sugerir novas melhorias através de trabalhos futuros;

No decorrer do texto deste trabalho, os objetivos acima apresentados serão demonstrados. Porém, o objetivo maior deste trabalho, é complementar um diagnóstico mais rápido, menos invasivo para o paciente e com melhor prognóstico de vida, sempre sob o olhar clínico do profissional de saúde com acompanhamento de resultados de outros exames e demais correlações clínicas.

1.4 Organização do trabalho

Este trabalho está organizado em 5 capítulos e segue a seguinte estrutura:

O Capítulo 1 contextualiza as considerações iniciais, a motivação, a definição dos objetivos e a organização deste trabalho.

O Capítulo 2 apresenta um levantamento bibliográfico do estado da arte sobre trabalhos correlatos. São publicações que envolvem análise radiômica, extração de características em tumores cerebrais e uso de inteligência artificial e aprendizado de máquina aplicados à classificação tumoral.

No Capítulo 3, informações necessárias ao leitor são descritas para um melhor entendimento deste trabalho. É introduzida uma fundamentação teórica sobre os principais conhecimentos abordados, tais como processamento digital de imagens (PDI), análise radiômica, a biblioteca *PyRadiomics*, técnicas de escalonamento dos dados, técnicas de redimensionamento de dados, técnicas de aumento de dados, algoritmos classificadores de aprendizado de máquina, *datasets* de imagens cerebrais e protocolos de imagens de RM.

Já no Capítulo 4, são descritos 4 experimentos implementados neste trabalho, a fim de classificar a codeleção cromossômica 1p/19q. Serão apresentados, a metodologia utilizada para o desenvolvimento, os resultados e a discussão para cada experimento. Em seguida, é apresentada uma discussão comparando ambos os experimentos com o estado da arte.

Por fim, o Capítulo 5 apresenta-se as conclusões finais deste trabalho. São descritas considerações sobre o mesmo e identificadas algumas sugestões para trabalhos posteriores a partir desta tese de doutorado.

Capítulo 2

Trabalhos Correlatos

Neste capítulo, será apresentada uma revisão bibliográfica acerca dos seguintes temas: análise radiômica; extração de características em tumores cerebrais; e uso da inteligência artificial no auxílio ao diagnóstico por imagem.

2.1 Análise radiômica

A radiômica é um processo de extração de características quantitativas que resulta na conversão de imagens em dados mineráveis. Nos últimos 10 anos, é notável o elevado crescimento da possibilidade de se analisar uma imagem médica. Tal afirmação poderia ser justificada devido ao desenvolvimento de softwares e ferramentas que permitem o reconhecimento de padrões em imagens, bem como o crescimento de bancos de dados de imagens. Nesse sentido, a radiômica vem ao encontro deste avanço, possibilitando ações ou processos capazes de desenvolver atividades de extração de características em imagens, permitindo que essas informações se transformem em dados que possam ser analisados e utilizados para tomada de decisões. Em relação à prática médica tradicional em radiodiagnóstico, constata-se que a avaliação das imagens ocorre de forma visual, enquanto na prática radiômica os dados são explorados a fim de aprimorar o diagnóstico e prognóstico dos pacientes com câncer [11].

Com a evolução do desempenho computacional, tornou-se possível extrair características quantitativas em imagens de RM, tomografia computadorizada (TC) e tomografia por emissão de pósitrons (PET) em prazos que atendam os profissionais da saúde, proporcionando à radiômica uma análise favorável após extração de grande dimensão de características. Com a combinação de outros dados do paciente, tais como dados clínicos e genômicos, a radiômica pode se tornar uma ferramenta de suporte à decisão médica. Vale ressaltar que, embora apresente certa similaridade com um *computer aided diagnosis* (CAD), a radiômica difere-se em ser um processo explicitamente projetado para extrair características e disponibilizá-las em um banco de dados para testes e análises posteriores [12]. Os CADs foram e ainda são muito utilizados, como no diagnóstico do câncer de

mama, por serem sistemas geralmente autônomos e direcionados para entregar uma única resposta [13].

Devido ao seu crescimento, a radiômica tem despertado o interesse não somente de pesquisadores, mas também de médicos que estão na prática clínica do dia-a-dia. De acordo com Wibmer et al. [14], os médicos radiologistas consideram a radiômica uma ferramenta de auxílio no diagnóstico de tumores agressivos e raros. Wibmer et al. relatam em seu trabalho sobre a capacidade da análise radiômica em distinguir tecidos benignos e o nível de agressividade do câncer de próstata [14]. Considerando a possibilidade de radiologistas contribuírem com a formação de banco de dados no apoio, no suporte e em decisões futuras, ferramentas radiômicas têm sido desenvolvidas para utilização no trabalho clínico diário. Coroller et al. [15] também apresentam a radiômica como uma ferramenta de avaliação do prognóstico do paciente nos casos de câncer de pulmão e cerebral do tipo glioblastoma.

Abordando um pouco mais os estudos em radiômica, Liu et al. [8] destacam em sua pesquisa diversas áreas do corpo humano tais como, mama, tireoide e carótida, onde ela tem sido amplamente utilizada. Ren et al. [16] construíram uma assinatura radiômica de cabeça e pescoço baseada em imagens pré-operatório de RM visando identificar o estágio do câncer, e verificaram que essa assinatura em imagens ponderadas em T1 pós-contraste e imagens ponderadas em T2 obtiveram bom desempenho. Zhou et al. [17] utilizaram um modelo de aprendizagem profunda tridimensional em câncer de nasofaringe para prever metástases linfonodais. A incidência do câncer de pulmão, órgão que exerce funções essenciais ao corpo humano, apresenta um elevado crescimento mundial, e representa um dos índices de maior mortalidade. Portanto, o pulmão tem sido objeto de grande estudo pela radiômica no diagnóstico, tratamento e prognóstico [18]. Aerts et al. [19] em um de seus trabalhos de maior destaque, extraíram 440 características de TC de 1.019 pacientes e descobriram que muitas das características radiômicas estavam correlacionadas com o estágio geral, o estágio T (triagem), o estágio E (estadiamento) e o estágio M (monitoramento) do câncer de pulmão. Além disso, Coroller et al. [15] e Zhou et al. [20] desenvolveram trabalhos semelhantes utilizando a radiômica à distância para prever metástase em pulmão, sendo que o primeiro analisou 182 pacientes nos Estados Unidos com diagnóstico patológico de adenocarcinoma, e o segundo avaliou 348 pacientes da China com fatores de risco clínico identificados.

Outro órgão extremamente importante no corpo humano, e com maior incidência de casos em mulheres no mundo, é a mama. Estudos renomados em radiômica aliados a informações clínicas do paciente têm se destacado na detecção precoce do câncer de mama, no auxílio ao tratamento de câncer de mama e no auxílio ao prognóstico do câncer de mama. O câncer de mama pode ser diagnosticado por várias modalidades de exames, como ultrassonografia (US), TC, mamografia e RM, sendo estes utilizados em pesquisas para a previsão do subtipo de câncer [21, 22]. Antropova et al. [22] propuseram unir as

características de mamografia, US e RM com o intuito de obter um melhor diagnóstico de câncer de mama. A ideia da proposta era proporcionar ao paciente uma abordagem não invasiva para a previsão do linfonodo sentinela. A maioria dos estudos radiômicos atuais em mama têm se concentrado na resposta à terapia do paciente [23]. Chan et al. [23] apresentaram uma técnica automatizada para prevenir falhas no tratamento em pacientes em estado inicial utilizando características extraídas de imagens de RM pré-operatório. Por fim, vale o registro de estudos radiômicos com foco no prognóstico do câncer de mama, como revelam Park et al. após o desenvolvimento de uma assinatura radiômica combinando dados da RM e dados clínicos do paciente para uma estimativa individualizada da sobrevida livre da doença [24]. Vários outros exemplos de aplicação de análise radiômica em outros órgãos suscetíveis ao câncer, como fígado, intestino, próstata e região colorretal, cuja doença acomete diariamente milhares de pessoas, foram publicados na literatura [8].

A palavra radiômica é originada do termo em inglês *radiomics*, cujo sufixo *omics* refere-se ao estudo detalhado de moléculas biológicas [11]. Porém, no âmbito das ciências, a radiômica tem se tornado popular por permitir o estudo de uma alta complexidade de dados. A análise radiômica, em conjunto com a medicina de precisão, biologia molecular e novos biomarcadores.

Além do que foi citado na seção 1.1 acerca das estimativas que representam o número de casos de câncer no Brasil, outro dado alarmante é mostrado por Siegel [25], o qual relata que os Estados Unidos apresentam em média, 1,6 milhão de novos casos de câncer por ano. Como a maioria desses pacientes são submetidos a exames como RM, TC e PET, que geram imagens digitais, esforços têm sido dirigidos para a criação de grandes bancos de dados. Embora a radiômica ainda necessite ser melhor explorada e elucidada, esse grande conjunto de dados poderá proporcionar um aperfeiçoamento nos tratamentos.

A radiômica sempre apresenta desafios e novas oportunidades [26, 27], mesmo sendo reconhecida por possuir um conceito simples (extração de características quantitativas) e ser uma etapa objetiva, todo um conjunto de ações são ilustradas na Figura 2.1 e serão melhor descritas no Capítulo 3.

2.2 Extração de características quantitativas em tumores cerebrais

O uso de sofisticadas ferramentas computacionais tem possibilitado a extração de características quantitativas e a construção de modelos com potencial para o diagnóstico, prognóstico e estudos sobre o câncer [28]. Park et al. [28] registram ter encontrado 189 artigos publicados com o termo radiômica, em inglês *radiomics*, dentre os quais 51 apresentavam as palavras diagnóstico, prognóstico, predição e estudaram o tumor do tipo

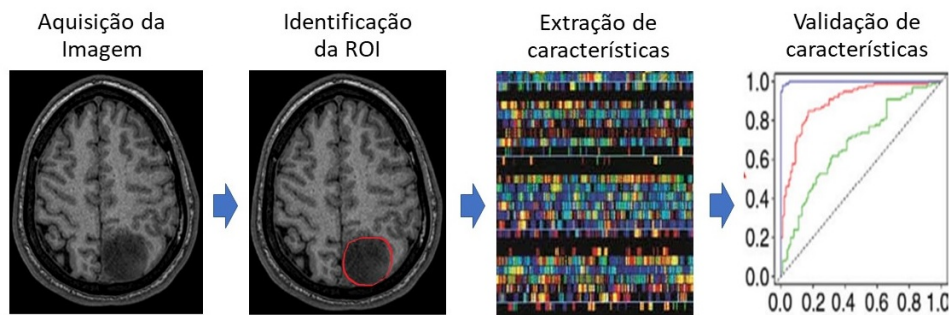


Figura 2.1: Inicialmente realiza-se a aquisição da imagem. No exemplo apresenta-se uma RM de cérebro com ponderação em T2 e corte axial. Em seguida identifica-se a região de interesse, que pode ocorrer de forma manual, semi-automática ou automática. Neste exemplo, identifica-se um LGG. Posteriormente aplicam-se algoritmos de extração das características radiômicas. Por fim ocorre a validação dessas características por meio de métodos matemáticos e estatísticos utilizando abordagens de aprendizado de máquina.

glioma.

Em neurooncologia diversos estudos têm sido conduzidos em relação à extração de características quantitativas e análises radiômicas. Suh et al. [4] apresentam vários estudos focados em diagnóstico, prognóstico e predição de diagnósticos como diferencial. Zhang et al. [29] focam seus estudos na classificação molecular e análise de sobrevivência do paciente. Park et al. [28] revelam que nos 51 principais artigos estudados em neurooncologia, 90,1% apresentam estudos sobre os gliomas. Destes, 47,1% são glioblastomas, 29,4% são LGG e os demais são tumores não gliais.

Zhou et al. [30] destacam estudos em análise de extração de características através de uma abordagem interdisciplinar, e revisam estudos radiológicos que interpretam imagens de tumores cerebrais através de modelos computacionais, como visão computacional e aprendizado de máquina, e apresentam novas percepções clínicas. Os autores registram as estratégias que mais têm sido utilizadas para extração de características em imagens para diferentes finalidades clínicas, além de apresentarem uma discussão sobre as contribuições de aprendizado de máquina.

Imagens radiológicas são capazes de gerar uma quantidade expressiva de informações, sendo elas relatadas no âmbito qualitativo e subjetivo. Gillies et al. [11] ressaltam uma clara demonstração de que o processo de extração de características quantitativas visa aperfeiçoar a compreensão acerca de tumores cerebrais. *Arrays* de características quantitativas em imagens podem ser utilizados para validar biomarcadores e caracterizar a ocorrência intra tumoral durante o tratamento do paciente [31]. Como modelos computacionais têm se tornado uma tecnologia capaz de identificar, analisar e validar recursos analíticos extraídos de imagens de câncer, novas percepções sobre fatores de risco, respostas precoce ao tratamento e adaptações de estratégias ao tratamento foram identificadas através de pesquisas [26].

Além de descreverem a importância da visão computacional na extração de ca-

racterísticas quantitativas de imagens, Lambin et al. [27] também destacam a importância de que praticamente 100% das características extraídas encontram-se na ROI, local que normalmente engloba todo o tumor ou regiões específicas dele. As características extraídas de imagens podem ser quantificadas e traduzidas radiologicamente quanto à forma volumétrica do tumor e à dinâmica da aparência visual. Lowe [32] utiliza a função de diferença da Gaussiana para calcular através da detecção de pontos-chave a transformação de característica invariante de escala, um cálculo da forma do tumor. Ou seja, durante o tratamento, o autor quantificou alterações sutis de quanto o tumor estava mais redondo ou mais elíptico.

Trabalhos como o de Zacharaki et al. [33] mostram como as técnicas computacionais têm sido largamente reproduzidas para extração de características de lesões tumorais em imagens de RM. De fato, devido à sua precisão, tais técnicas vêm sendo utilizadas na análise clínica de exames de imagens. Yun et al. [34] propõem um descritor de características biológicas, em que o conhecimento radiológico de características era transferido para hipóteses biológicas. Por exemplo, em um estudo, os autores levantaram a hipótese de que características farmacocinéticas oriundas da RM poderiam ser potenciais biomarcadores para diferenciar resultados obtidos entre tratamentos de radioterapia e quimioterapia. Ou seja, em seu descritor, sugere-se que características de uma doença inspira a sugerir características biológicas. Em termos comparativos, Zhou et al. [35] utilizam características extraídas das imagens de RM e inspiradas na biologia para definir a variação espacial, distribuição dos dados e evolução temporal de um tumor. De acordo com os autores, informações prognósticas distintas podem ser resultantes das variações espaciais entre sub-regiões tumorais. Além disso, o acompanhamento de alterações temporais pode revelar diferentes respostas fisiológicas durante a radioterapia em regiões de baixa e alta perfusão em gliomas. Os autores de Zhou et al. [36], propõem um novo espaço de imagens, sendo assim uma nova forma de quantificar as regiões e sub-regiões tumorais através do realce de contraste local, área de edema e vascularização em RM.

Outras aplicações importantes da radiômica são as pesquisas envolvendo o tratamento genômico em imagens, denominado radiogenômica, que representa uma associação entre biomarcadores de imagens e características genômicas de uma doença. O grande objetivo desta associação é permitir uma avaliação não invasiva através da extração de características em substituição aos resultados dos testes moleculares [37]. Alguns estudos mostram que o uso de características quantitativas extraídas através da RM de volumes inteiros de tumores podem distinguir atividades moleculares (fluxo sanguíneo e densidade molecular), por exemplo dos glioblastomas e seus subtipos [38].

Outros estudos realizaram a extração de características através de imagens para mostrar a codeleção cromossômica 1p/19q. Por exemplo, Akkus et al. [39] extraíram dados das imagens de TC a partir da varredura das regiões de interesse (já segmentadas por médicos radiologistas) utilizando uma rede neural convolucional (do inglês, *convolutional*

neural network (CNN)). Já no estudo conduzido por Zhang et al. [40], dois radiologistas realizaram a segmentação manual das regiões de interesse em cortes axiais, fatia a fatia, posteriormente validadas por outros dois radiologistas seniores, e então foi aplicada a ferramenta LIFEx Soft [41] para extração de 40 características radiômicas de primeira e segunda ordem. Em sua recente publicação do ano de 2021, Decuyper et al. desenvolveram um processo automatizado para segmentação de tumores cerebrais e previsão do grau do glioma, IDH e codeleção 1p/19q. Os autores reuniram alguns *datasets* e utilizaram CNN com arquitetura U-Net 3D para classificação das estruturas. É importante reforçar que métricas convencionais de ML (acurácia, sensibilidade, especificidade e precisão) foram utilizadas para amostrar os resultados [42].

2.3 Inteligência artificial no auxílio ao diagnóstico por imagem

Dentro das ciências exatas, temos as áreas de engenharias, ciências da computação, dentre outras que estudam a IA. Atualmente, através de técnicas, mecanismos, algoritmos, modelos matemáticos, e probabilidade e estatística, a IA busca conceder aos equipamentos a capacidade de aprender o conhecimento de um especialista. Apesar do mito de que o computador pode trabalhar como o cérebro humano, as avançadas pesquisas em neurociência ainda apontam que isto não é viável. Sem dúvida a IA está se relacionando com as ciências da saúde e suas inumeráveis pesquisas. Por ser capaz de processar grande quantidade de informação, principalmente no diagnóstico por imagem, esta tecnologia tem ganhado a atenção dos profissionais da saúde. A utilização da IA está sendo oportunamente utilizada em outras áreas médicas, tais como laboratorial, educacional, controle e alerta de doenças, medicina intensiva, telemedicina, dentre outras [43].

Basicamente, a IA em medicina atende duas frentes:

- Sistemas especialistas em apoio à decisão: são aqueles que apoiam o profissional da saúde em uma decisão. Por exemplo, a identificação de características que sugerem um tumor cerebral.
- Sistemas especialistas inteligentes: são aqueles que não necessitam do auxílio do ser humano para execução. Por exemplo, reconhecimento facial.

Topol [44] apresenta uma visão futurista do uso da IA na prática médica, citando que, em pouco tempo, as pessoas estarão consultando um sistema de IA antes de consultar um médico para procedimentos padrão. Segundo ele, a grande quantidade de informações clínicas e registros médicos que estão sendo armazenados eletronicamente possibilitarão aplicações de IA serem usadas cada vez mais na medicina. O autor ainda afirma que essas aplicações poderão mudar a forma em como os profissionais da saúde abordam a solução

de problemas clínicos. Porém, inúmeras indagações e preocupações precisam ser avaliadas e validadas por questões regulatórias.

Alguns trabalhos iniciais em IA aplicada à medicina apresentam comparativos inevitáveis entre a execução do algoritmo proposto como solução para determinado problema e a conclusão pelo profissional da saúde. Mukherjeedare [45] apresenta o relato de um pesquisador da Stanford University, que através da classificação de médicos dermatologistas, construiu um banco de dados de 130 mil imagens de lesões da pele. Com uma RNA, ele identificou padrões nas lesões através das características extraídas. Logo após o treinamento pela RNA, foram utilizadas 14 mil imagens para reconhecimento de 3 classes: lesões benignas, lesões malignas e lesões com crescimentos não cancerosos. A RNA acertou 72% contra 66% de acerto dos médicos dermatologistas.

O Google, subsidiário da *holding* Alphabet Inc., desenvolve projetos em IA com foco na medicina. Dentre esses projetos, vale destacar dois [46]: o primeiro, um algoritmo capaz de distinguir diversos tipos de câncer no ser humano, obtendo resultado de 89% de precisão contra 72% de médicos especialistas sem limite de tempo. Os testes têm sido realizados por uma equipe de pesquisadores da Universidade de Nova York e a ferramenta está sendo chamada de Inception V3. Os desenvolvedores do Google utilizaram *Deep Learning* (DL) para reconhecimento de mutações genéticas dentro do tumor através de imagens e, como fonte de treinamento, forneceram ao algoritmo milhares de imagens de lâminas histológicas obtidas do The Cancer Genome Atlas (TCGA). O segundo projeto é a utilização da IA para detecção de câncer de pulmão em estágio inicial de pacientes com histórico de tabagismo. Nos Estados Unidos a taxa de mortalidade de pessoas por ano com câncer de pulmão é elevada. Para este estudo, foram treinadas 42.290 imagens de TC de aproximadamente 15 mil pacientes. Na execução do algoritmo, foram fornecidas apenas informações das classes dos pacientes, ou seja, quem tinha câncer ou não. Para este teste, foi realizado um comparativo entre os resultados do algoritmo e a análise por 6 médicos radiologistas. Na conclusão do projeto, o Google cita que tanto na avaliação de uma única TC de um paciente quanto na avaliação de um conjunto de TC do paciente, o Inception V3 apresentou melhores resultados, sendo 5% melhor na detecção do câncer e 11% melhor no reconhecimento dos falsos positivos.

Através das técnicas de ML, a extração de características em imagens (mais especificamente a análise radiômica) fornece uma abordagem radiológica para desvendar características preditivas. O objetivo principal da análise radiômica é ajudar no diagnóstico, prognóstico e sobrevida do paciente com câncer. Sendo assim, as informações extraídas podem ser associadas a outras informações do paciente, tais como o desfecho clínico do paciente. Sendo assim, as técnicas de ML podem ser escolhidas para proceder à classificação dos dados do paciente considerando o histórico de informações clínicas, genômicas, biológicas ou imagens obtidas do indivíduo. Essas técnicas serão apresentadas no Capítulo 3. Em um estudo para previsão do tempo de sobrevivência do paciente com

glioblastoma [47], além dos dados clínicos e dados qualitativos (como lesões multifocais, forma do tumor e área de edema), foram utilizados biomarcadores quantitativos. Com a utilização de DL, os autores sugeriram, após o treinamento dos dados juntamente com características extraídas das imagens de RM com ponderação (protocolo de imagens que fornece informações complementares) em T2 e Flair, dois grupos de pacientes: aqueles com sobrevivência de longo prazo (maior que 400 dias) e com sobrevivência de curto prazo (menor que 400 dias).

Zacharaki et al. [33] desenvolveram um estudo com objetivo de identificar e diferenciar tumores cerebrais: gliomas do tipo II, tipo IV e metástases. Os autores apresentam o desenvolvimento do estudo desde a definição da ROI, extração das características quantitativas através de imagens de RM de perfusão, seleção das características e classificação. O algoritmo *support vector machine* (SVM), foi utilizado em um conjunto de 102 tumores cerebrais, ambos diagnosticados histologicamente. Destes, 34 eram glioblastoma. Após execução do algoritmo com validação cruzada, obtiveram os respectivos percentuais de acertos na diferenciação dos gliomas de grau II e grau IV: 88% de precisão, 85% de sensibilidade e 96% de especificidade.

Akai et al. [48] desenvolveram um estudo para prever a sobrevida global de pacientes com carcinoma hepatocelular ressecável, ou seja, um tipo de câncer de fígado maligno. Para esse estudo, foram utilizadas imagens de TC de 127 pacientes, sendo 97 homens e 30 mulheres, com idade média de 68 anos. Os respectivos tumores foram segmentados e, logo após, foram extraídas as características de textura com um filtro *Laplacian of Gaussian* (LoG). Em seguida, ocorreu o treinamento com o algoritmo *random forest* (RF) utilizando validação cruzada de tamanho 5. Dentre as 96 características de textura extraídas, o algoritmo foi treinado com 15 características para a determinação da sobrevida global do paciente e com 8 características para determinação da sobrevida do paciente livre da doença. Em seu resultado, os autores relatam ter identificado dois grupos de pacientes, um de baixo risco e outro de alto risco individual de previsão da doença. Registram ainda uma combinação satisfatória entre características radiômicas e o algoritmo RF para prever o prognóstico de pacientes com carcinoma hepatocelular ressecável.

Por fim, cita-se duas pesquisas utilizadas como base para o desenvolvimento dos experimentos apresentados no Capítulos 4, desta tese de doutorado. Ambas foram publicadas no ano de 2017, sendo que a primeira [49] propõe identificar a codeleção do cromossomo 1p/19q. Segundo os autores, foi desenvolvido um algoritmo para fazer um reconhecimento visual das imagens de RM ponderadas em T1 e T2 de 80 pacientes diagnosticados com gliomas e marcação genética. Em sequência o algoritmo SVM foi utilizado para classificação e os autores apresentam os resultados no nível de confiança com 95%. Em T2, (0.46 - 0.89) de precisão, (0.39 - 0.80) de sensibilidade e (0.57 - 0.94) de especificidade. Aos mesmos conjuntos de imagens em T2, neurorradiologistas acertaram 77%. Na segunda publicação [39], os autores propuseram uma CNN para prever a codeleção

do braço cromossômico 1p/19q, em gliomas LGG através de imagens de RM ponderadas em T1 pós-contraste e T2 de 159 pacientes, sendo 102 pacientes codeletados e 57 não-deletados, todos diagnosticados por biópsia. O conjunto de imagens foi dividido em treinamento, validação e teste. Ao conjunto de treinamento foram aplicadas técnicas de aumento de imagens através de deslocamento translacional, inversões horizontal e vertical e rotações. Após o treinamento e validação das imagens, o resultado dos testes apresentou 87,7% de acurácia, 93,3% de sensibilidade e 82,2% de especificidade.

Capítulo 3

Conceitos Fundamentais

Neste capítulo serão apresentados os conceitos de algumas técnicas de pré-processamento que foram utilizadas neste estudo para o desenvolvimento do trabalho proposto. Algumas pesquisas mais recentes e com crescente utilização na literatura abordando a utilização das mesmas serão apresentadas para demonstrar importância dessas técnicas.

3.1 Tumores cerebrais

Os tumores cerebrais referem-se ao crescimento anormal de células dentro do crânio, ou ainda podem ser identificados de acordo com seu local de surgimento, como por exemplo tumores que nascem de células do sistema nervoso ou das membranas que protegem o cérebro. Em sua grande maioria, são tumores benignos. Aqueles que nascem do tecido cerebral são conhecidos como tumores primários e os tumores que surgem das meninges (membranas que recobrem o cérebro) são também conhecidos como meningiomas. Além disso, existem os tumores que nascem em outros órgãos do corpo humano, como mama, pulmão, fígado, e se implantam dentro do SNC, caracterizando as chamadas metástases. Atualmente, 4% dos óbitos em decorrência do câncer estão relacionados aos tumores cerebrais, sendo catalogados mais de 100 tipos desses tumores, que recebem o nome de acordo com a célula na qual se desenvolvem [1].

Para um tratamento neuro-oncológico eficaz é necessária a identificação do tipo do tumor, e uma das formas mais eficientes de identificá-lo é realizar a retirada de parte desse tumor e encaminhá-la para avaliação patológica, denominada biópsia. De fato, a biópsia é considerada uma importante etapa neuro-cirúrgica, além de permitir a descompressão do cérebro e a realização de ressecção (o máximo possível).

Conforme classificação da OMS, os tumores que surgem no SNC apresentam características histológicas que refletem seu nível de agressividade [3] e são classificados como tumores de baixo grau (LGG) e tumores de alto grau (HGG). O LGG representa tumores que apresentam um crescimento relativamente lento (menor agressividade), enquanto os HGG são tumores altamente destrutivos e com rápido crescimento (maior agressividade).

Ambos se iniciam a partir de células gliais ou astrócitos, conhecidos por sua função de suporte, nutrição e proteção aos neurônios, e equivalem a 30% de todos os tumores do SNC. Dentre os LGGs, os mais comuns são os oligodendrogliomas e os astrocitomas, já entre os HGGs destaca-se o glioblastoma [3, 50]. A Figura 3.1 apresenta exemplos de gliomas do SNC.

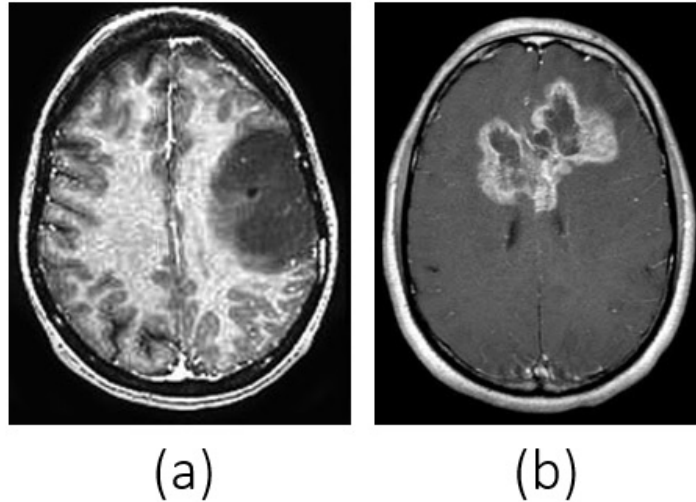


Figura 3.1: Imagens de RM exemplificando gliomas através de cortes axiais (a)LGG - oligodendroglioma e (b)HGG - glioblastoma, ambas as imagens em ponderação T1 pós-contraste [51].

Em sua atualização de 2016, a OMS recomenda uma classificação integrativa, acrescentando marcadores moleculares e genéticos às características histológicas dos tumores, permitindo uma melhor adequação da terapia de pacientes, melhor classificação para ensaios clínicos e estudos experimentais e uma categorização mais precisa [3]. Dessa forma, atualmente os tumores do tipo oligodendroglioma são caracterizados pela codeleção dos cromossomos 1p/19q, enquanto que os astrocitomas não apresentam essa codeleção cromossômica. Alguns estudos mostraram que a codeleção dos cromossomos 1p/19q deve ser considerada como importante marcador molecular tumoral preditivo de resposta ao tratamento quimioradioterápico e maior sobrevida em comparação aos tumores não-codeletados [52].

Em uma análise invasiva, após realização de exames de RM ou TC e obtêm-se o diagnóstico de câncer, muitas vezes torna-se necessário fazer uma biópsia para determinar com precisão que se trata de um glioma e o tecido da biópsia é usado para executar testes para determinar as características genéticas do glioma. Em uma classificação não invasiva, uma das alterações genéticas possíveis poderia ser a codeleção 1p/19q. Essa mutação genética consiste na perda do braço curto do cromossomo 1 (1p) e do braço longo do cromossomo 19 (19q). Se um paciente é positivo para essa codeleção 1p/19q, isso estabelece o diagnóstico de oligodendroglioma, o que indica um melhor prognóstico e

uma melhor resposta ao tratamento. A Figura 3.2 encena a codeleção 1p/19q.

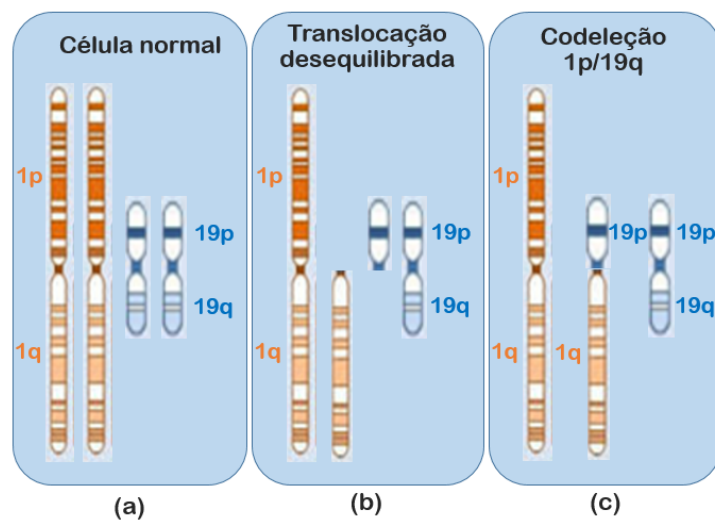


Figura 3.2: Processo de ocorrência da codeleção 1p/19q na célula. Em (a) os 2 pares de cromossomos 1p/1q e 19p/19q, em (b) a deleção do braço cromossômico 1p no par 1p/1q e a deleção do braço cromossômico 19q no par 19p/19q e em (c) a identificação da codeleção 1p/19q.

3.2 Processamento digital de imagens

A ciência médica tem sido impactada por um grande crescimento tecnológico nos últimos anos. Além disso, destaca-se o avanço conquistado a partir da análise e processamento de imagens médicas, haja visto seu rápido progresso em tratamentos guiados por imagens e o crescimento de técnicas de diagnóstico assistido por computador. O grande foco de PDI em diagnóstico por imagens está no processamento, análise, extração, quantificação e interpretação dos órgãos e estruturas do corpo humano. Enxerga-se um avanço do PDI na detecção, no diagnóstico, no prognóstico e na resposta ao tratamento do câncer [53].

Diferentemente de uma imagem convencional, que normalmente possui resolução de 8 bits com três canais de cores (vermelho, verde e azul), uma imagem médica apresenta uma resolução de contraste em tons de cinza. Atualmente, os equipamentos produzem imagens com 12 e 16 bits, ou seja, chegando a 65.536 tons de cinza, versus os 8 bits (256 níveis de cinza) de contraste nos equipamentos produzidos até o ano 2000. Geralmente, imagens obtidas diretamente dos aparelhos de diagnóstico por imagem são cruas (*raw*, em inglês), e após alguma técnica de processamento digital, se transformam em um arquivo digital e estarão prontas para serem apresentadas aos profissionais da saúde para diagnóstico. Dependendo do fabricante, modelo do equipamento e software de captura de imagens as características das imagens podem sofrer variação [54]. A Figura 3.3 apresenta

um exemplo de uma imagem de raios-x crua capturada após emissão do feixe de raios-x e a mesma imagem disponibilizada para avaliação médica após processamento pelo software do equipamento médico.

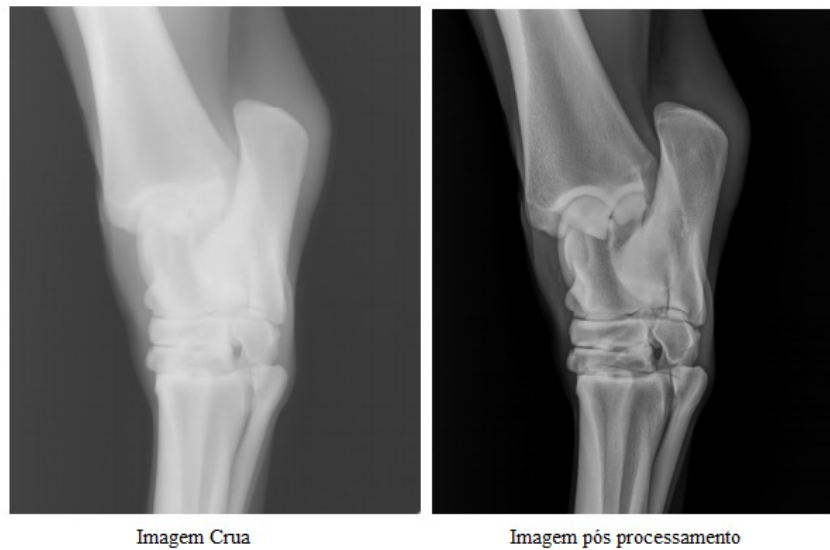


Figura 3.3: Exemplo de uma imagem de raios-x crua e seu processamento digital - imagem de domínio público disponível em: <http://www.agfahealthcare.com>.

O corpo humano permite a aquisição de dados e produz grande quantidade de informações acerca de suas propriedades, sendo caracterizado como um sistema complexo e intrincado. Dessa forma, torna-se um enorme desafio a aquisição, o processamento e a exibição das informações deste sistema. Tais informações permitem aos procedimentos médicos uma análise e interpretação mais apurada a fim de auxiliar no diagnóstico e prognóstico do paciente. Quando há interação entre os tecidos ou órgãos com algum tipo de técnica de radiodiagnóstico, por exemplo a RM, as imagens são geradas. Através de técnicas diferenciadas, pode-se visualizar a anatomia (estrutura) do corpo e suas funções (fisiologia).

Normalmente, imagens médicas digitais não são apropriadas para serem visualizadas diretamente, sem qualquer tipo de processamento. Mesmo após um procedimento de reconstrução, pode ser necessário realizar um pré-processamento da imagem para que a mesma passe por realce, correção, melhora no contraste, ajuste de pixels, redução de ruídos, dentre outras técnicas que, em situações específicas, podem demonstrar mais eficiência.

3.3 Plataforma PyRadiomics

Para o cérebro humano, reconhecer objetos por meio dos olhos é uma tarefa relativamente fácil. Entretanto, com os sistemas computacionais é diferente, pois há a

necessidade de um algoritmo para análise dos pixels de uma imagem.

3.3.1 Segmentação do volume de interesse

Ao avaliar uma imagem à procura de uma anormalidade, um médico analisa as características daquela região de interesse cuidadosamente para ser preciso em seu diagnóstico. No entanto, esta tarefa torna-se cansativa, e muitas vezes, demorada.

Em termos computacionais, esta tarefa de separação da imagem em regiões de interesse que possuem características semelhantes é denominada segmentação. Bordas, texturas e luminância do pixel são características básicas buscadas em ROI por algoritmos computacionais. Estes algoritmos são responsáveis pela separação das ROI nas imagens para que, posteriormente, outras tarefas tais como classificação e reconhecimento de padrões possam ser realizadas. A etapa de segmentação de imagens é reconhecida como uma das mais difíceis de se realizar em processamento de imagens. Ela pode determinar o sucesso ou o fracasso das etapas seguintes [55]. A Figura 3.4 ilustra uma demonstração de uma segmentação utilizada a partir do Capítulo 4 desta tese de doutorado na elaboração do projeto de análise radiômica da codeleção cromossômica 1p/19q.

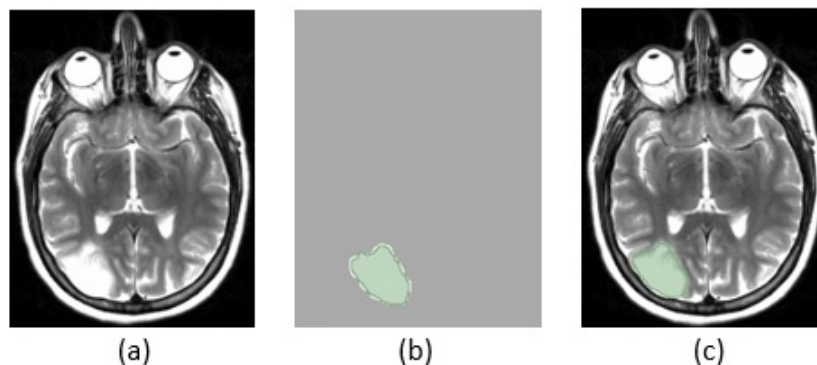


Figura 3.4: Imagem de RM do cérebro com tumor do tipo LGG com sua respectiva segmentação sendo a) corte axial com ponderação em T2; b) segmentação do tumor; c) imagem original com região segmentada em destaque.

Neste estudo, não foi necessário realizar a etapa de segmentação das imagens. As imagens adquiridas no TCIA já estavam acompanhadas de suas respectivas segmentações, também denominadas máscaras. Foi necessário apenas a conversão do tipo de imagem de DICOM para *nearly raw raster data* (NRRD) para que a biblioteca PyRadiomics pudesse realizar a extração das características. Maiores detalhes dessa extração serão abordados na seção seguinte e no Capítulo 4.

3.3.2 Extração de características - PyRadiomics

A extração de características é uma etapa normalmente subsequente à realização da segmentação de uma ROI, sendo responsável pela descrição e representação da região ou imagem analisada, que em seguida poderá passar por algum outro processamento ou execução de algum algoritmo de classificação. Computacionalmente, ela requer muito esforço, pois as imagens médicas armazenam informações complexas. Conforme estudo prévio [53], a extração automática de características que possa definir a essência da imagem é um desafio.

Duas formas de extrair informações de uma imagem podem ser consideradas, sendo a primeira delas a representação da região de interesse por meio das características externas, por exemplo, bordas, formato, contorno, dentre outras, e a segunda a representação interna da região de interesse, que tem como foco principal as propriedades dos pixels, por exemplo, textura, valores dos tons de cinza e outros [55].

Para descrever as características de um objeto, podem ser utilizados vários métodos. Haralick et al. relatam a utilização de sequências de coordenadas cartesianas, momentos angulares, invariantes de momento Hu e coeficientes gerados por transformadas *wavelet* [56].

A textura representa um dos principais métodos utilizados para a descrição das características de uma imagem médica, sendo ela a combinação entre magnitude e frequência da variação dos níveis de cinza de determinada imagem. Três propriedades são importantes para a caracterização da textura em uma imagem: i) o contraste local, que define a diferença entre os níveis de cinza mais claro e mais escuro entre os pixels vizinhos; ii) a orientação, que está associada ao direcionamento dessas regiões mais claras ou mais escuras; e iii) a escala, que traz dependência entre as regiões mais claras e mais escuras, ou seja, pode apresentar texturas mais finas ou mais grossas [57].

Para Gonzalez e Woods [55], a textura tem três funções em processamento de imagens, sendo elas: i) função espectral, a qual detecta periodicidade global e identifica os picos de alta de energia no espectro baseada no espectro de Fourier; ii) função estatística, onde um conjunto de métricas extraídas localmente define a textura; e iii) função estrutural, que traz a ideia de repetitividade e regularidade.

Nesse sentido, a radiômica pode quantificar um elevado número de características de forma e textura. Suas tecnologias têm como finalidade principal extrair dados quantitativos que possam ser mais facilmente usados para posterior classificação, mineração e visualização com métodos de IA, algoritmos de ML e, conseqüentemente, tornando-se importante no apoio à decisão clínica [58] e previsão de prognóstico de tumores intracranianos e classificação pré-cirúrgica [59]. O potencial da radiômica, em conjunto a essas metodologias, baseia-se na capacidade de reconhecer padrões através das características extraídas automaticamente. Nesse contexto, os dados quantitativos extraídos por aná-

lise radiômica podem refletir em características histopatológicas e genéticas de tumores. Exemplo disso está nos trabalhos apresentados no Capítulo 2, os quais avaliam tumores em diversas anatomias do corpo humano e fornecem informações complementares para apoio na decisão dos profissionais da saúde, em especial à área da oncologia [60].

Van Griethuysen et al. identificaram uma ausência de padronização nas características radiômicas extraídas das imagens; em função dessa situação, implementaram, por meio de conhecimento matemático e de engenharia, a biblioteca PyRadiomics [61], uma plataforma de código-fonte aberto para extração de características radiômicas, com principal aplicabilidade em imagens médicas. A PyRadiomics foi implementada em linguagem de programação Python, que pode ser instalada em qualquer sistema operacional e vem sendo intensamente utilizada para computação científica e ciência de dados. A PyRadiomics, além de ser uma plataforma flexível na análise de dados, possui uma interface *back-end* que permite manipulação das características radiômicas em lote e uma automação no processamento das mesmas, podendo ainda ser utilizada através da ferramenta 3DSlicer, uma interface *front-end* simples, de código aberto e gratuita para processamento de imagens médicas [62]. Ainda em [61], os autores demonstram a utilização da extração de atributos radiômicos na caracterização de lesões benignas e malignas de pulmão.

Toda a documentação da PyRadiomics pode ser acessada livremente¹, bem como sua arquitetura, fluxo de trabalho, versões, instruções de instalação, vídeos complementares, dentre outros. A Figura 3.5 ilustra um *pipeline* completo apresentado pela plataforma PyRadiomics, no qual a extração dos descritores radiômicos está incluída. As etapas são: aquisição da imagem, segmentação da lesão, extração das características e, por fim, a análise e interpretação dos dados.

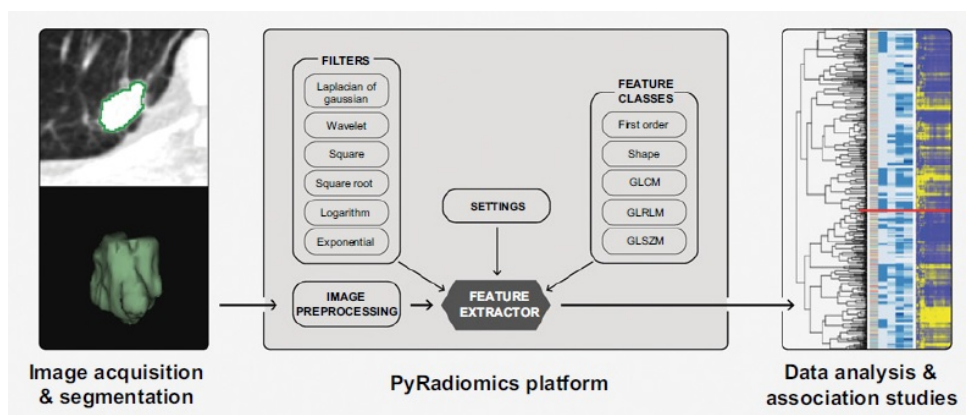


Figura 3.5: Etapas de um *pipeline* apresentado pela plataforma PyRadiomics [61].

Segundo Van Griethuysen et al. [61] a execução da PyRadiomics pode ser descrita em 4 etapas:

1) **Carregamento da imagem e da máscara de segmentação do tumor** - nesta

¹<https://pyradiomics.readthedocs.io/en/latest/>

etapa, imagens de RM, TC ou PET em formato DICOM, NRRD ou NIFTI (padrão de imagens que permite uma melhor operacionalização de imagens e dados médicos) do cérebro juntamente com sua respectiva ROI (denominada *mask* pela PyRadiomics) são carregadas e lidas pela plataforma. A PyRadiomics suporta a entrada de uma única fatia (2D) ou conjunto volumétrico (3D). Os cortes radiográficos podem ser nos planos axial, coronal ou sagital, enquanto os protocolos (as chamadas ponderações) podem ser T1, T1 pós-contraste, T2, Flair e outras. Ambas as informações relacionadas aos planos de corte da imagem e protocolos serão melhor detalhadas na seção 3.6.1. O plano axial e os protocolos T2 e Flair são os mais utilizados para extração de características radiômicas, pois permitem melhor visualização de líquidos e patologias (tumores).

2) Pré-processamento da ROI - a PyRadiomics utiliza a biblioteca SimpleITK, um conjunto de ferramentas de código aberto muito empregado para tratamento de imagens. Uma de suas principais funcionalidades consiste no suporte a diversos formatos de imagem, à grande variedade de recursos para forma e textura, além de várias opções de reamostragem do voxel (elemento de volume, é a contrapartida 3D do pixel, que é um elemento de imagem 2D) isotrópico (o qual indica que o objeto possui o mesmo tamanho em todas as dimensões), revelando distâncias iguais entres os voxels vizinhos em todas as direções [63].

3) Extração das características - Inicialmente, a plataforma PyRadiomics retorna 120 características radiômicas extraídas originalmente por meio de suas sete classes. Os filtros posteriormente citados podem ser aplicados ou não. Se aplicados, a plataforma PyRadiomics retorna além das características originais, as características de cada classe com seus respectivos filtros. A Tabela 3.1 exhibe alguns exemplos de características radiômicas extraídas.

As características extraídas das classes da plataforma PyRadiomics podem ser calculadas diretamente na imagem original e/ou em uma outra imagem derivada (por exemplo, em uma imagem obtida após aplicação de um filtro), com exceção dos descritores de forma 2D e 3D, pois são extraídos através do rótulo da máscara e independem do valor no nível de cinza. Quando ativados no arquivo de parâmetros, são mostrados como se ocorresse o cálculo pela imagem original, mas são calculados separadamente da imagem de entrada. Basicamente, todas as características extraídas através das classes da plataforma PyRadiomics foram definidas pelo Imaging Biomarker Standardization Initiative (IBSI) [64]. As classes da plataforma PyRadiomics são:

- (a) First Order Statistics - características que descrevem a ocorrência de distribuição da intensidade dos *voxels* na região da imagem usando métricas comuns e básicas. Podem ser extraídas 19 características nesta classe.
- (b) Shape-based 2D - características que incluem os descritores de forma e tamanho bidimensionais da ROI. Podem ser extraídas 10 características nessa classe.

- (c) Shape-based 3D - características que inclui os descritores de forma e tamanho tridimensionais da ROI. Podem ser extraídas 16 características nessa classe.
- (d) Gray Level Cooccurrence Matrix (GLCM) - A classe matriz de coocorrência de nível de cinza é muito utilizada na classificação de imagens, conhecida como uma tabulação da possibilidade de combinações que pode ocorrer entre pixels de uma imagem. Ou seja, ela representa a relação entre o pixel chamado de referência e seu vizinho, que pode ser o pixel acima, abaixo, da esquerda ou da direita, e ainda na diagonal. Podem ser extraídas 24 características nessa classe.
- (e) Gray Level Run Length Matrix (GLRLM) - A classe matriz de duração de nível de cinza realiza cálculos em execuções homogêneas (por exemplo, os tecidos humanos que se caracterizam como estruturas aleatórias, sem regularidade aparente) para cada nível de cinza. É uma quantificação das execuções para um determinado número de pixels que possui mesma intensidade. Molina et al. [65] afirmam que as características desta classe são as mais robustas, inclusive promovendo uma boa combinação com RNAs. Diferente de algumas ferramentas, a plataforma PyRadiomics calcula os índices de textura da matriz e, antes da exibição das características, realiza um corte entre os níveis de cinza mínimo e máximo, movendo os índices principais da matriz e, portanto, os valores dos índices texturais resultantes. Segundo a documentação, essa movimentação reduz o custo computacional no cálculo das características radiômicas. Podem ser extraídas 16 características nessa classe.
- (f) Gray Level Size Zone Matrix (GLSZM) - A classe matriz de tamanho de zona de nível de cinza irá quantificar o número de voxels que se conectam compartilhando a mesma intensidade de nível de cinza. Um voxel está conectado quando sua distância com outro voxel é de apenas 1 pixel. Ou seja, para imagens 2D, afirma-se que a região é 8-conectada. Diferentemente das classes anteriores, a GLSZM calcula uma única matriz para qualquer direção da ROI, sendo independente de rotação. Podem ser extraídas 16 características nessa classe.
- (g) Gray Level Dependence Matrix (GLDM) - A classe matriz de dependência de nível de cinza, como a própria descrição do nome informa, mensura a dependência entre o nível de cinza em uma imagem, sendo determinada como uma distância x entre os voxels conectados ao voxel central. Podem ser extraídas 14 características nessa classe.
- (h) Neighbouring Gray Tone Difference Matrix (NGTDM) - A classe matriz de diferença de nível de cinza de vizinhos quantifica a diferença entre o nível de cinza do voxel central e a média do nível de cinza dos vizinhos conectados a uma distância x . A soma absoluta das diferenças no nível de cinza para aquele voxel é armazenada na matriz. Podem ser extraídas 5 características nessa classe.

As matrizes de métricas de nível e tonalidade de cinza acima citadas são comumente utilizadas em diagnósticos médicos auxiliados por meio de imagens de RM, TC, US e outras para extração de características de textura com o objetivo de aprimorar a interpretação médica. Pode-se citar, por exemplo, a avaliação de células atingidas pelo câncer [66]. A Tabela 3.1 mostra, como exemplo, algumas características radiômicas extraídas pela PyRadiomics com e sem aplicação de filtro e como são nomeadas.

Tabela 3.1: Exemplos de características radiômicas extraídas por meio da plataforma PyRadiomics para as classes First Order e GLCM com e sem filtro.

Filtros	Classes	
	First Order	GLCM
Original	original_firstorder_Energy	original_glcml_Autocorrelation
Wavelet	wavelet-LLH_firstorder_Energy	wavelet-LLH_glcml_Autocorrelation
LoG	log-sigma-0-5-mm-3D_firstorder_Energy	log-sigma-0-5-mm-3D_glcml_Autocorrelation
Square	square_firstorder_Energy	square_glcml_Autocorrelation
SquareRoot	squareroot_firstorder_Energy	squareroot_glcml_Autocorrelation
Logarithm	logarithm_firstorder_Energy	logarithm_glcml_Autocorrelation
Exponential	exponential_firstorder_Energy	exponential_glcml_Autocorrelation
Gradient	gradient_firstorder_Energy	gradient_glcml_Autocorrelation
LBP2D	lbp-2D-m1_firstorder_Energy	lbp-2D-m1_glcml_Autocorrelation

4) **Aplicação de filtros** - além das características radiômicas originalmente extraídas, a plataforma permite aplicar filtros nas ROI pré-processadas. Os filtros disponíveis são²⁾:

- Wavelet: esse filtro extrai em cada nível de classe 8 decomposições combinando os filtros passa-alta e passa-baixa em cada uma de suas três dimensões.
- Laplacian of Gaussian (LoG): esse filtro produz aprimoramentos de bordas. Nele pode ser configurado um parâmetro *sigma*, que define texturas mais grossas ou mais finas, enfatizando áreas com alteração de nível de cinza.
- Square: esse filtro eleva as intensidades da imagem ao quadrado e redimensiona-as linearmente para o intervalo original.
- SquareRoot: esse filtro aplica a raiz quadrada nas intensidades absolutas da imagem e redimensiona-as linearmente para o intervalo original.
- Logarithm: esse filtro obtém o logaritmo da intensidade absoluta + 1 da imagem e redimensiona os valores para o intervalo original.
- Exponential: esse filtro calcula o exponencial da imagem original.
- Gradient: esse filtro calcula e retorna a magnitude do gradiente na imagem.

²disponível em: <https://pyradiomics.readthedocs.io/en/latest/>

- Local Binary Pattern 2D (LBP2D): esse filtro calcula e retorna o padrão binário da imagem em uma operação de fatia por fatia (2D).
- Local Binary Pattern 3D (LBP3D): esse filtro calcula e retorna o padrão binário da imagem em 3D usando harmônicas esféricas.

3.4 Pré-processamento

Esta é uma etapa muito importante da análise dos dados. É aqui que alguns ajustes e alterações serão realizados de forma a deixá-los com uma melhor qualidade de mineração.

3.4.1 Seleção das características

Em grande parte dos problemas reais de classificação, as características predictoras são de valor parcial ou irrelevantes para a predição ou ainda redundantes, o que não agrega valor ao resultado final. E além disso, com grandes *datasets*, fez-se necessário reduzir o número de características para melhorar o desempenho de execução dos algoritmos de classificação [67].

Utilizar características irrelevantes no conjunto de dados pode afetar negativamente o desempenho e a precisão dos modelos de classificação, fazendo com que os mesmos aprendam atributos insignificantes. Portanto a seleção de características visa selecionar os atributos mais descritivos para a variável de resposta. Com isso, ganha-se redução do tempo de treinamento, melhora na precisão, minimização de dados enganosos, menor quantidade de atributos redundantes, dentre outros a redução de *overfitting*, que geralmente ocorre quando o número de atributos é bem maior que o número de instância no *dataset* [68].

Quatro algoritmos de seleção de recursos são descritos abaixo, métodos de filtro (importância, correlação, chi-quadrado) e método *wrappers* (recursive feature elimination (RFE)). Esses métodos possuem implementação simples e podem modelar a dependência dos atributos.

1) Métodos de filtros - esses métodos examinam cada característica individualmente, sobretudo qual delas tem o relacionamento mais forte com a variável de resposta.

- Importância: essa técnica permite obter a importância de cada característica do conjunto de dados através da propriedade de importância do modelo de ML escolhido. Ele pontua cada característica do conjunto de dados, e por meio das maiores pontuações ele sugere quais características possuem maior relevância com a variável de saída.

1. YellowBrick: uma biblioteca desenvolvida para o pacote Python, que fornece visualização de características, visualização da classificação e visualização de modelos de seleção, dentre eles a importância das características. Ela seleciona os atributos mínimos necessários para produzir um desempenho válido.
- Matriz de correlação: essa técnica nos indica como as características estão relacionadas entre si e com a variável alvo. Uma correlação forte e positiva, significa que o valor se aproxima de 1. Quando o valor se aproxima de -1 , indica uma correlação forte, mas chamada negativa.
 - Qui-quadrado: é uma técnica estatística que mede a divergência de uma distribuição esperada a uma distribuição observada dos dados se assumirmos que a ocorrência do recurso é realmente independente do valor da classe. Calcula-se o qui-quadrado entre cada atributo e a variável alvo. Se a variável alvo for independente do atributo, ela obtém uma pontuação baixa, mas, se for dependente, recebe uma pontuação alta e o atributo é importante [69].

2) **Método *wrappers*** - esses métodos utilizam algum algoritmo de ML para encontrar os melhores subconjuntos de atributos.

- Eliminação recursiva de características (RFE): é uma técnica recursiva que treina um modelo de forma iterativa, e a cada iteração elimina o atributo menos significativo até avaliar todos os atributos. Ela utiliza o próprio sistema de pesos do modelo de ML como critério [70]. A RFE é uma técnica que sofre com tempo de execução, pois consome muito recurso computacional.

3.4.2 Escalonamento das características

A normalização, bem como a padronização ou escalonamento dos dados, assim chamados por diversos autores, são ações caracterizadas como recursos primordiais e projetadas visando atingir um objetivo similar, que é criar recursos que tenham intervalos semelhantes entre si, sendo amplamente utilizadas na análise de dados e muitas vezes entendido como a preparação de um novo *dataset*.

Em um de seus trabalhos, DeFellipi [71] apresenta o uso da normalização dos dados de entrada para avaliação de imóveis no distrito da Califórnia. As informações de entrada, o número de quartos e a renda média da família possuíam, cada um, uma escala de unidade diferente, e determinadas informações que, para o autor soavam estranhas, como o número de banheiros acima de 1000 e renda média familiar em um agrupamento de 500.000 dólares, faziam com que a classificação descartasse qualquer pessoa. Após aplicar a normalização dos dados chamada *Min-Max*, ou seja, colocar os valores entre 0 e 1, os *outliers* desapareceram e os dados passaram a ficar mais consistentes entre

si. Ao mesmo conjunto de dados, foi aplicada outra técnica de normalização chamada padronização (ou, do inglês *z-score*), inserindo os valores entre uma média 0 e desvio padrão 1. Assim como na normalização *Min-Max*, os mesmos resultados foram alcançados na normalização *z-score*, porém observou-se um melhor comportamento no processo de treinamento. Segundo o autor, com a padronização, a otimização do algoritmo torna-se mais aprimorada. Este trabalho faz parte de um desafio hospedado no Kaggle Inc³.

Geller [72] apresenta um conjunto de dados obtidos por meio de um sonar, instrumento utilizado para localização de submarinos, estudo e pesquisa dos oceanos, dentre outras finalidades, composto por 208 registros (linhas) e 60 entradas (colunas) de informações. Esses dados eram de sinais refletidos em um cilindro de metal e rochas. Apesar de parecer um conjunto de dados equilibrado, os mesmos não tinham a garantia de 1 ser o valor máximo e 0 o valor mínimo em cada recurso. O objetivo foi classificar corretamente qual sinal havia sido emitido. A Tabela 3.2 ilustra os resultados aplicados a 4 algoritmos de classificação. É possível verificar que a normalização através da padronização usando o método *StandardScaler()* da biblioteca Scikit-Learn aprimorou os resultados dos algoritmos MLP e SVM. Já o método RF não foi afetado e o método RL sofreu uma leve piora.

Tabela 3.2: Taxa de acerto na acurácia dos testes do *dataset* com e sem uso da normalização dos dados pelo método *StandardScaler()* na comparação dos algoritmos de classificação [72].

Algoritmos de Classificação	Sem padronização	Com padronização
MLP	78,0%	84,0%
RF	71,0%	71,0%
SVM	61,0%	84,0%
RL	75,0%	74,0%

Em outro estudo, também mencionado no trabalho de Geller [72], o autor utiliza um *dataset* com 56.420 registros (linhas) e 18 entradas (colunas) para prever dias de chuva. Ao realizar o treinamento, o autor comparou dois métodos de normalização. Conforme Tabela 3.3, foram obtidos melhores resultados para o método *RobustScaler()*. Esse método utiliza uma combinação entre valor médio da coluna e o segundo quartil, o que sugere um bom tratamento para problemas com outliers. Já o método *StandardScaler()* não apresentou melhoras significativas em relação ao treinamento sem o uso da padronização. O autor sugere que tal resultado seja devido ao grande número de registros (linhas) do *dataset*, diferentemente do teste anterior.

³www.kaggle.com/camnugent/california-housing-prices

Tabela 3.3: Taxa de acerto na acurácia dos testes do *dataset* com dois métodos de padronização dos dados na comparação dos algoritmos de classificação [72].

Algoritmos de Classificação	Sem padronização	Método <i>StandardScaler()</i>	Método <i>RobustScaler()</i>
MLP	99,0%	100,0%	100,0%
RF	97,0%	97,0%	97,0%
SVM	78,0%	97,0%	99,0%
RL	100,0%	99,0%	100,0%

Raschka e Mirjalili [73] em seu livro citam sobre a normalização e o redimensionamento dos dados após trazer explicações sobre o método de padronização (ou *Z-score normalization*), demonstram um exemplo do efeito da padronização para algoritmos de aprendizado de máquina. Eles utilizam o conjunto de dados “Wine” gratuito que encontra-se disponível no repositório de aprendizado de máquina da UCI⁴, o qual contém três classes diferentes para as “*n*” entradas de vinhos (registros das linhas). Neste exemplo, o pesquisador realiza a transformação dos dados usando o método de padronização *StandardScaler()* e aplica uma classificação utilizando o algoritmo Gaussian Naive Bayes. Para isso, ele divide os dados do *dataset* em treinamento (70%) e testes (30%). Desta forma, foram obtidos os seguintes resultados de acurácia: 96,77% no treinamento e 98,15% nos testes, versus 81,45% no treinamento e 64,81% nos testes sem a padronização.

Após demonstração de alguns trabalhos que utilizaram tal técnica e obtiveram resultados importantes, percebe-se que o escalonamento dos dados é muito utilizado por quem trabalha com classificação de atributos, seja em estatísticas, ciência de dados ou ML. A principal justificativa para esta tratativa nos dados é devido à variação nos quais os atributos a serem analisados e/ou tralhados podem possuir. Sendo assim, ao aplicar um escalonamento, os dados de entrada são trazidos para uma mesma ordem de grandeza, ou até mesmo para uma mesma unidade, não deixando que um atributo ou característica tenha maior peso que outro.

O escalonamento é geralmente uma abordagem escolhida pois, para modelos de ML como RL, SVM, MLP, RNA, dentre outros, os pesos estimados serão atualizados de maneira semelhante e não com taxas diferentes durante o processo de construção do treinamento, tendendo a tornar o treinamento um processo mais comportado. Isso fornecerá resultados mais precisos quando os dados forem escalonados inicialmente [73].

1. Standard Scaler: realiza uma distribuição normal dos dados de forma que a distribuição resultante seja centrada em 0 e um desvio padrão igual a 1. Também conhecida por *Z-score standardization*. Os atributos são escalonados com base na

⁴<http://archive.ics.uci.edu/ml/datasets/Wine>

equação:

$$z = \frac{x - \mu}{\sigma} \quad (3.1)$$

em que μ é a média e σ é o desvio padrão. A Figura 3.6 demonstra tal situação.

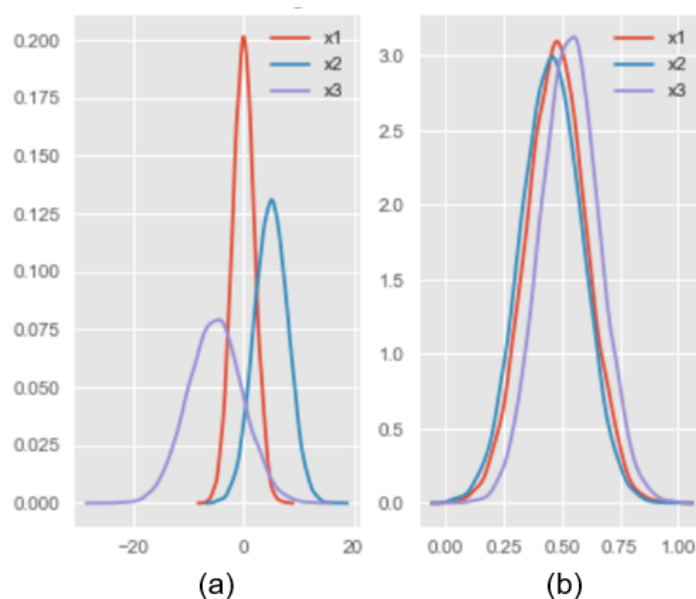


Figura 3.6: Exemplo da distribuição dos dados escalonados com método *Standard Scaler*, em (a) antes do escalonamento e (b) após o escalonamento - imagem gerada a partir do código disponível em: <https://benalexkeen.com/feature-scaling-with-scikit-learn/>.

2. Min-Max Scaler: realiza a divisão de um vetor por seu comprimento e transforma os dados em um intervalo entre 0 e 1. Os atributos são escalonados com base na equação:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.2)$$

em que $\min(x)$ é o menor valor da coluna e $\max(x)$ é o maior valor da coluna. A proposta da equação 3.2 é escalonar os dados para que tenham melhor resultado na classificação quando a distribuição dos dados não for Gaussiana ou o desvio padrão for muito pequeno. A Figura 3.7 demonstra tal situação.

3. Robust Scaler: semelhante ao Min-Max Scaler, ele usa o intervalo interquartil. O objetivo desse método é ser robusto na identificação de *outliers*. Consequentemente, ele utiliza-se de menos dados para realizar o escalonamento. Os atributos são escalonados com base na equação:

$$z = \frac{x - Q_1(x)}{Q_3(x) - Q_1(x)} \quad (3.3)$$

em que $Q_1(x)$ é o primeiro quartil e $Q_3(x)$ é o terceiro quartil. A proposta da

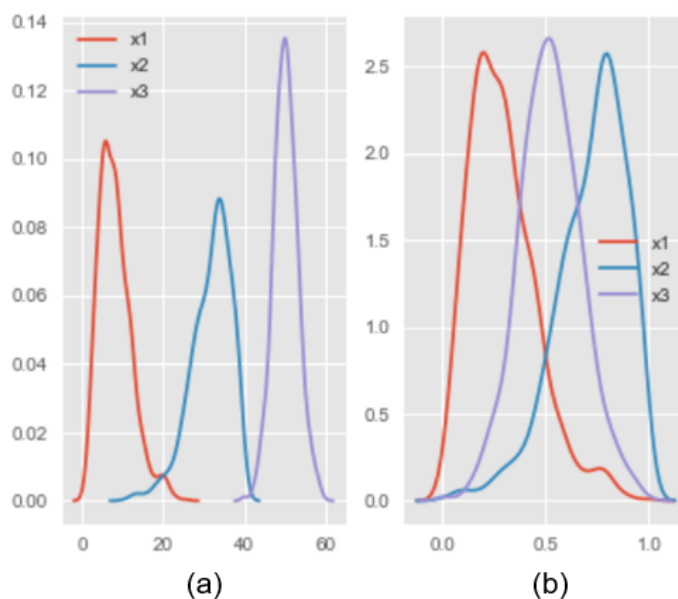


Figura 3.7: Exemplo da distribuição dos dados escalonados com *Min-Max Scaler*, em (a) antes do escalonamento e (b) após o escalonamento - imagem gerada a partir do código disponível em: <https://benalexkeen.com/feature-scaling-with-scikit-learn/>.

equação 3.3 é escalonar os dados para que tenham melhor resultado na classificação quando os dados apresentam valores discrepantes em relação a faixa de valores do conjunto em que se encontram. A Figura 3.8 demonstra tal situação.

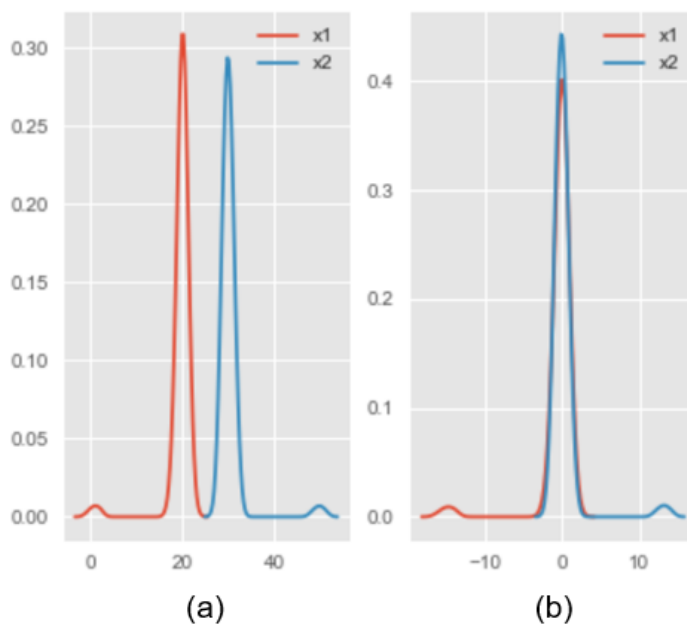


Figura 3.8: Exemplo da distribuição dos dados escalonados com *Robust Scaler*, em (a) antes do escalonamento e (b) após o escalonamento - imagem gerada a partir do código disponível em: <https://benalexkeen.com/feature-scaling-with-scikit-learn/>.

Todas as técnicas de escalonamento citadas anteriormente fazem parte da biblioteca Sklearn⁵, são utilizadas em diversos trabalhos por apresentarem bons resultados quando comparadas a resultados que não apresentam estas técnicas [72, 74].

Vale registrar que a biblioteca Sklearn possui outros métodos de escalonamento, sendo eles, *MaxAbsScaler()*, *QuantileTransformerNormal()*, *QuantileTransformerUniform()*, *PowerTransformerYeoJohnson()* e *Normalizer()*.

3.4.3 Redimensionamento das características

O redimensionamento de dados é uma técnica constantemente utilizada na tentativa de se realizar a redução do número de dimensões de um conjunto de dados, onde espera-se encontrar uma nova estrutura capaz de auxiliar na compreensão, bem como na visualização desses dados.

Pao e Meng [75] abordam a dificuldade de entendimento e possíveis problemas quanto ao multidimensionamento de dados. Uma das soluções que os autores apontam é a utilização de métodos e ferramentas para o dimensionamento dos conjuntos de dados, os quais possibilitam visualizar uma melhor correlação entre os dados, e inclusive, com projeções gráficas 2D. Também deixam registrado em seu estudo a importância da compreensão dos dados multidimensionais abordando três importantes aspectos: distribuição, categorização e relacionamento.

Outro estudo interessante foi apresentado por Kreuseler [76] no registro de observações marinhas por 25 anos. O autor reuniu vários dados, incluindo temperatura, concentração de oxigênio, salinidade, dentre outros e relatou dificuldade em analisar e encontrar uma correlação entre esses dados hidrológicos. Mesmo com apresentação de gráficos dos 10.000 registros coletados nos 25 anos, foi relatado pelo autor impossibilidade de análise dos mesmos. De acordo com o pesquisador, foi realizada redução da dimensionalidade, escolhendo algumas características mais importantes e dividindo os dados em subconjuntos. Isso possibilitou aos pesquisadores uma boa compreensão dos dados e a realização de uma análise.

Guo et al. [77] descreveram, a partir de dados multidimensionais, a criação de uma função de transferência na geração de imagens. Através do algoritmo *Multidimensional Scaling*, a função de transferência projetava um diagrama de coordenadas paralelas e dados em *clusters*.

Uma das técnicas clássicas, a PCA, foi utilizada por Yuan et al. para redimensionar dados [78]. Os autores desenvolveram uma aplicação que analisa dados multidimensionais, e, através da interação do usuário com a ferramenta, mapas de árvores de subespaços são construídos. O objetivo da aplicação era proporcionar aos usuários a projeção de gráficos 2D para visualização, identificação de relações e agrupamento dos

⁵documentação disponível em <https://scikit-learn.org/stable/>

dados.

Tsai [79] revela que tanto a PCA como o *Multidimensional Scaling* são excelentes técnicas para se trabalhar com dados lineares. Porém, quando o conjunto de dados apresenta uma não-linearidade, outras técnicas são sugeridas, tais como *isometric feature mapping* (Isomap), *kernel principal component analysis*, *hessian locally linear embedding* (HLLE), *locally linear embedding* (LLE), *local tangent space alignment* (LTSA), *diffusion maps* e *multilayer autoencoders* [79].

Balachander et al. [80] e De Backer et al. [81] realizaram estudos comparando alguns métodos de redimensionamento não supervisionados, dentre eles a PCA. O objetivo dos trabalhos foi classificar padrões de amostras citológicas, cujos resultados foram aferidos indiretamente através das taxas de acerto do classificador *k-Nearest Neighbor* (KNN) aplicado aos dados reduzidos.

Em aplicações de IA, Pao e Meng em seu trabalho citam alguns métodos importantes anteriormente utilizados para redução da dimensionalidade, como *ratio-conserving map*, transformada de Karhunen-Loeve, mapa de características de Kohonen, *generative topographic mapping*, *autoassociative mapping approach*, *nonlinear variance conserving approach* e *equalized orthogonal mapping* [75].

Os autores de Sorzano et al. [82] relatam sobre adaptações que surgiram com base nos métodos tradicionais, como por exemplo a técnica *Kernel PCA*, adaptada da PCA, para mapeamentos não lineares. Em sua publicação, os autores deixam registrada uma vasta relação de trabalhos sobre redução de dimensionalidade, mineração de dados, comparativo entre métodos de redução de dimensionalidade e, por fim, fazem um detalhamento do redimensionamento não linear.

Djerioui et al. [83] apresenta um sistema de previsão de doenças cardíacas utilizando *neighborhood component analysis* (NCA) e SVM. Os autores utilizam o NCA para selecionar as características mais relevantes em seu *dataset* obtido com dados reais de doenças cardíacas. Como resultado eles apontam uma redução no tempo, na utilização de materiais e mão de obra e aumento no desempenho da previsão. O sistema proposto superou em 1,99% a previsão de doenças cardíacas.

Yang, Wang and Zuo [84] consideram de extrema importância a seleção de características quando se trata sobre dados de alta dimensão. Os autores propõem um algoritmo de ponderação das características baseado nos vizinhos mais próximos por meio do cálculo da distância euclidiana. Experimentos realizados em *datasets* reais e artificiais mostraram que o algoritmo desenvolvido, *neighborhood component feature selection* (NCFS) é sensível a diminuição do número de características irrelevantes e apresentou desempenho superior aos algoritmos avaliados (LMFW, FSSun e Simba) em praticamente todos os casos.

No trabalho de Siddique, Sakib e Rahman [85] aplicou-se o NCA e algoritmos

autoencoders em 9 *datasets* do repositório UCI⁶. Os autores reduziram a dimensão desses *datasets* em 50%, selecionando as características consideradas mais relevantes. Posteriormente todos os novos *datasets* foram classificados através do KNN e o SVM, algoritmos de aprendizado de máquina com abordagem supervisionada. Dividiram o conjunto de dados em 90% e 10%, para treinamento e testes, respectivamente, e utilizaram a métrica precisão para avaliar o desempenho de cada classificador.

É importante o relato dos trabalhos utilizando técnicas de redimensionamento dos dados, para que se possa ter noção da importância de utilização dessa técnica. No capítulo 4 serão apresentados resultados de experimentos que usam técnicas de redimensionamento de dados.

A estratégia de redução da dimensionalidade consiste em identificar os elementos de um conjunto com n dimensões e representá-los com um número menor de dimensões m , sendo $m < n$, mantendo a relação entre os dados do conjunto. A função $f: R^n \times T \rightarrow R^m$, apresenta uma interpretação do comportamento dos algoritmos de redução de dimensionalidade, em que os elementos n de um conjunto de dados são transformados e representados em um novo conjunto em R^m . Porém cada algoritmo que realiza a redução de dimensionalidade dos dados apresenta sua peculiaridade em relação à disposição dos pontos vizinhos. Em geral, após execução do algoritmo, é essencial que ele mantenha relação de vizinhança entre os pontos do espaço m -dimensional assim como eram no espaço n -dimensional [86].

Dois métodos são descritos a seguir: o PCA e o NCA.

1. *principal component analysis* (PCA): Sendo um dos métodos mais antigos, inicialmente utilizado em um ambiente biológico e posteriormente em trabalhos psicométricos [87], o PCA é citado por Orlandic et al. [88] como um método de base para todas as técnicas de redução de dimensionalidade. Segundo ele, o PCA é um método que realiza uma operação linear sobre os dados, a fim de obter uma nova projeção com coordenadas ortogonais capaz de demonstrar maior variância dos dados numa sequência decrescente do percentual desta variância. Mesmo sendo um método clássico, o PCA ainda possui ampla aceitação em atividades de redimensionamento de dados e análise exploratória, sendo bastante útil quando o número de características é maior que o número de instâncias, tornando uma tarefa difícil representá-las graficamente.

O PCA tem como objetivo reduzir a dimensionalidade de um grande número de variáveis que se correlacionam entre si dentro de um conjunto de dados, e que após o alcance da redução através de uma transformação, obtêm-se as componentes principais, que são funções lineares das variáveis originais, em que a maior variância se estende sobre a primeira coordenada, a segunda maior variância se estende sobre a

⁶<http://archive.ics.uci.edu/ml/index.php>

segunda coordenada e assim por diante. Para tanto, calcula-se a matriz de covariância para o conjunto de características e, em seguida, calcula-se os autovetores e autovalores da matriz de covariância, e logo após classifica-os de acordo com seu autovalor [89]. A seguir, é descrito um pseudo-código do algoritmo do PCA:

- (a) subtrair a média dos dados para cada atributo;
- (b) calcular a matriz de covariâncias;
- (c) calcular os autovalores e autovetores da matriz de covariâncias;
- (d) ordenar os autovetores por ordem de importância;
- (e) realizar o mapeamento dos novos dados para o novo espaço dimensional.

2. *neighborhood component analysis* (NCA): Ao contrário de outros métodos, o NCA é não paramétrico, e baseia-se no modelo KNN. Ele aprende um encaixe linear de baixa dimensão das informações rotuladas para facilitar a visualização e categorização dos dados de altas dimensões [90, 91]. Seu objetivo é encontrar uma matriz de distância A que possa espelhar com fidelidade a realidade dos dados de teste. Para o critério de otimização, o NCA aproveita as atribuições de vizinhos estocásticos em vez de simplesmente usar k vizinhos mais próximos.

A proposta do NCA para aplicar a redução da dimensionalidade nos dados é semelhante ao PCA, que consiste em identificar os elementos de um conjunto com n dimensões e representá-los com um número menor de dimensões m , sendo $m < n$ mantendo a relação entre os dados do conjunto.

O algoritmo do NCA aproveita uma matriz A de dimensão $m \times n$ em que $Q = A^T A$ e esta matriz estima os vetores de treinamento x_i em um espaço dimensional m . Esta projeção da matriz A define a distância de Mahalanobis entre quaisquer dois vizinhos mais próximos no espaço projetado [85].

3.4.4 Aumento das características

Por definição, afirma-se que um conjunto de dados está desbalanceado quando suas classes não estão igualmente representadas. Provost e Fawcett [92] relatam que o desequilíbrio entre classes de dados no setor de detecção de fraudes financeiras chega a 100 por 1. Em outros conjuntos de dados, como chamadas telefônicas fraudulentas, gerenciamento de telecomunicações, classificação de texto e detecção de derramamentos de óleo em imagens de satélite, o desbalanceamento dos dados chega a 100.000 por 1.

Segundo Woods et al. [93], avaliar o desempenho dos algoritmos de aprendizado de máquina quando os dados encontram-se desbalanceados não é apropriado, já que os resultados podem se diferenciar acentuadamente, pois estes algoritmos são avaliados normalmente usando uma métrica de precisão preditiva. Como exemplo, os autores citam

em seu trabalho a classificação de pixels em imagens de mamografia como uma região de possível tumor. Segundo eles, um conjunto de dados pode conter 98% de pixels normais e 2% de pixels anormais, área de tumor e não área de tumor, respectivamente. Em uma classificação padrão, seria estipulada a classe majoritária com precisão preditiva de 98%, enquanto não haveria uma taxa razoavelmente alta para detectar a classe minoritária. Os autores concluem que uma precisão preditiva simples não é apropriada nessas situações.

Lertampaiporn et al. [94] desenvolveram um *webserver* para classificação de dados genéticos. Os autores extraíram 125 características em cada sequência genética, num total de 4000 sequências. Em seguida, aplicaram uma técnica de aumento sintético de dados, o SMOTE, o qual foi aplicado para resolver o problema de desbalanceamento dos dados, aumentando em 50% o número de amostras e, posteriormente, obter melhores resultados na execução dos algoritmos de classificação.

Marques [95] utilizou o SMOTE como técnica para gerar o balanceamento dos dados reguladores da expressão gênica em plantas e animais. O autor registra ter alcançado uma taxa de classificação de 97% no treinamento dos dados balanceados utilizando o algoritmo de RF, contra 81% de taxa de treinamento executado nos dados desbalanceados.

Akkus et al. [39] também utilizam-se de técnicas de aumento da base de dados. Como os autores implementaram uma CNN e extraíram as características do tumor diretamente das imagens de entrada, após dividir o conjunto de dados em treinamento, validação e testes, os mesmos utilizaram técnicas de deslocamento translacional aleatório, rotação, inversões horizontal e vertical para aumento dos dados no conjunto de treinamento para um balanceamento igual das classes. Um exemplo dessa técnica foi implementado e demonstrado conforme a Figura 3.9.

Outro trabalho recentemente publicado descreve a utilização de técnicas de aumento de dados em seu conjunto de imagens de treinamento [42]. Segundo os autores, foram realizadas inversão aleatória, rotações axiais, escala de intensidade, transformação elástica e ajuste nos canais de entrada. O trabalho consiste no treinamento de uma rede CNN com arquitetura 3D-Net para detecção do grau do tumor, mutação do gene isocitrato desidrogenase (IDH) e codeleção cromossômica 1p/19q.

Além de transformações que possam alterar o tamanho, a forma ou a posição de um pixel dentro de uma imagem, pode-se citar mudanças que alteram o valor de um pixel armazenado, como por exemplo, a adição de ruídos. Os ruídos em imagens são provocados por algum fator relacionado ao processo de aquisição da imagem, geralmente ele é um valor aleatório no pixel. A adição de ruídos em imagens digitais foi testado pela primeira vez no trabalho de Moreno et al. [96], em 9 *datasets*. Os autores citam que a adição de ruídos em imagens favorece no aprendizado de características mais robustas em algoritmos de ML.

No trabalho de Zhang et al. [97], os autores desenvolveram uma CNN para classificação de tipos de frutas e utilizaram adição de ruído para aumentar os dados do conjunto

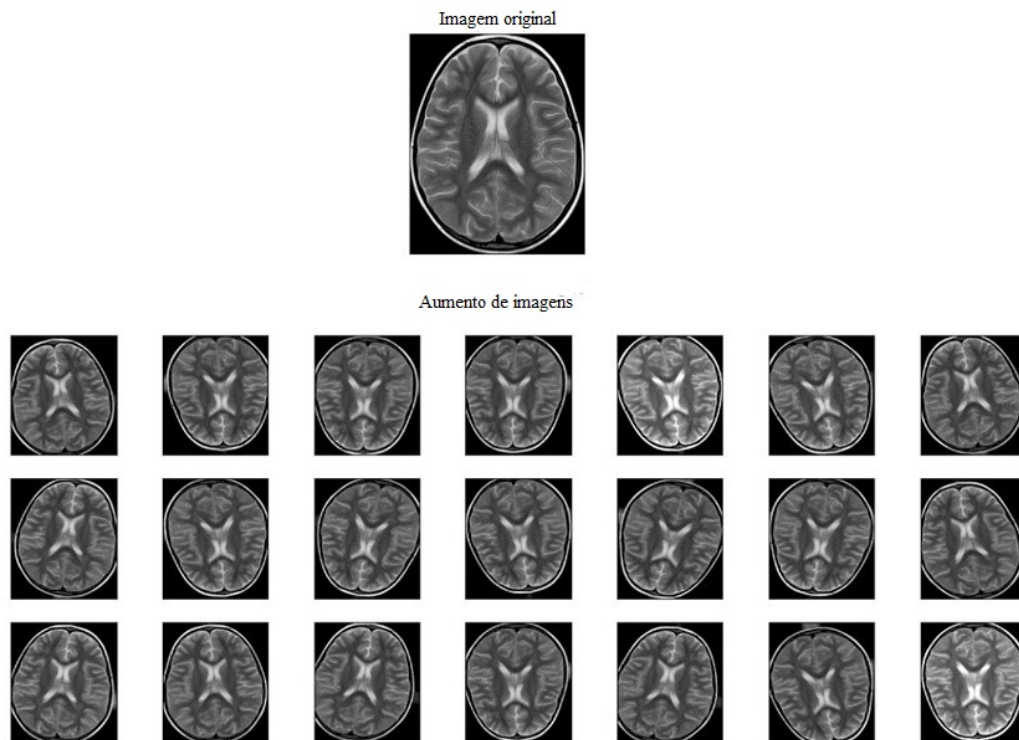


Figura 3.9: Exemplo de como ocorre o aumento de dados, com técnicas de translação, rotação e inversões.

de treinamento do *dataset*. A seção de resultados do trabalho aponta uma acurácia de 94,94%, sendo aproximadamente 5% melhor que os resultados encontrados na literatura.

Silva et al. [98] em seu trabalho para predição da codeleção cromossômica 1p/19q utilizaram adição de ruído gaussiano para aumento do conjunto de dados de treinamento e obter uma melhor performance dos modelos de ML, regressão logística e floresta aleatória aplicados na classificação dos dados do conjunto de teste. Os resultados obtidos tiveram acurácia de 83,24% e 84,67%, respectivamente, bastante promissores em comparação ao estado da arte.

Um dos propósitos das técnicas de aumento de características é minimizar o desafio de trabalhar com *datasets* pequenos ou desequilibrados (entende-se por desequilibrado um *dataset* que apresenta uma das classes sendo minoritária, com número de instâncias bem inferior às demais), e que a maioria das técnicas de aprendizado de máquina ignora e, por sua vez, apresenta baixo desempenho na classe minoritária, embora normalmente seja o desempenho na classe minoritária mais importante.

Uma possível maneira de resolver esse problema é aumentar o número de instâncias na classe minoritária. Sendo que, bastaria simplesmente duplicar os registros da classe minoritária no conjunto de dados de treinamento antes de realizar o ajuste do modelo, o que poderia equilibrar a distribuição de classes, mas não forneceria nenhuma informação adicional ao modelo [99]. Vejamos 2 abordagens que apresentam bons resultados descritos

pela literatura:

1) Aumento sintético de dados - um modo mais coerente para lidar com conjuntos de dados desequilibrados é sobreamostrar a classe minoritária. A abordagem mais simples envolve a duplicação de exemplos na classe minoritária, embora esses exemplos não adicionem novas informações ao modelo. Em vez disso, novos exemplos podem ser sintetizados a partir dos exemplos existentes. Esse é um tipo de aumento de dados para a classe minoritária que é chamado de técnica de super-amostragem por minoria sintética ou SMOTE [100]. Ele seleciona aleatoriamente uma característica, e logo após são encontrados os k vizinhos que estão mais próximos desta característica. Um dos vizinhos encontrados é selecionado aleatoriamente e gera-se uma nova característica sintética entre esses dois pontos [101]. Este é um procedimento que pode ser utilizado quantas vezes for necessário para gerar características sintéticas na classe minoritária. Conforme descrito em [101], os autores sugerem que primeiro seja utilizado a subamostragem aleatória para aparar o número de características na classe majoritária e, em seguida, usar o SMOTE para superexaminar a classe minoritária afim de equilibrar a distribuição da classe.

A Figura 3.10 apresentada a seguir, ilustra gráficos com um exemplo da implementação da biblioteca SMOTE. Na imagem 3.10(a) o gráfico de dispersão mostra uma distribuição aleatória de 990 pontos em uma classe majoritária e 10 pontos na classe minoritária, demonstrando um desequilíbrio entre as classes. Após aplicação da técnica SMOTE a esse conjunto de dados exemplo, pode-se ver por meio da imagem 3.10(b) que o gráfico de dispersão mostra a classe minoritária igualmente aumentada.

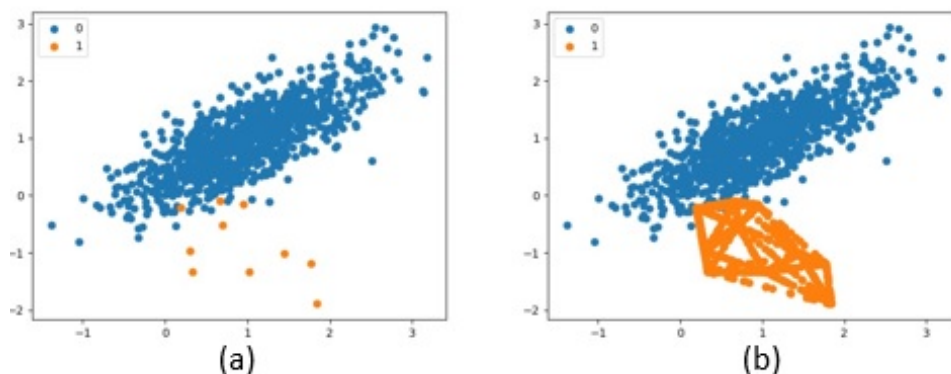


Figura 3.10: Figura com gráficos de dispersão exemplificando o balanceamento de pontos através da técnica SMOTE onde, os pontos azuis representam a classe majoritária e os pontos laranjas representam a classe minoritária. A imagem (a) mostra a plotagem inicial dos pontos e a imagem (b) mostra os pontos após aplicação do SMOTE.

2) Adição de ruído - outra abordagem possível para melhorar a classificação dos algoritmos em *datasets* pequenos ou desequilibrados é a geração de novas instâncias com adição de ruído nas características predictoras. Geralmente é um ruído com distribuição

Normal, que tem a sua potência distribuída uniformemente no espectro de frequência e valor médio nulo [102]. Um exemplo desse tipo de ruído é o Gaussiano e duas das principais aplicações do ruído Gaussiano constitui-se na síntese de sinais de fala e inserção de ruídos em imagens digitais. Segundo Moreno-Barea et al. [96], essa técnica favorece as CNNs no treinamentos de características mais robustas. Com esta técnica, os dados podem ser alterados pela adição de uma variável aleatória ou por multiplicação dos valores originais por uma variável aleatória [103].

3.5 Algoritmos de classificação

Após processamento da imagem, identificação da ROI e extração das características quantitativas por meio dos descritores, realiza-se a classificação ou associação de um padrão nas imagens através de um algoritmo. Tais procedimentos estão sendo fortemente utilizados através de abordagens supervisionadas, não supervisionadas e semi-supervisionadas. Kononenko [104] em seu trabalho publicado há quase 2 décadas já apontava o caminho do aprendizado de máquina, sendo hoje mundialmente difundido o ML. Esta representa uma das principais técnicas utilizadas na IA que realiza aprendizado de máquina baseado em matemática e estatística. Devido a grande quantidade de dados disponíveis para extração de informações, torna-se inviável o processamento manual imitando o comportamento do ser humano. Neste sentido, de posse dos dados brutos, utiliza-se ML para extração de características, representação e, posteriormente, por meio de um modelo matemático pode-se fazer uma inferência ou predição. Tal fato torna a utilização do ML interessante, pois trata-se de modelos matemáticos antigos, e que passaram a ter boa visibilidade a partir do crescimento das grandes bases de dados. É apresentada assim a utilização dos modelos matemáticos na descoberta de padrões que explique o relacionamento entre os dados extraídos, permitindo inclusive a automatização de tarefas e processos [104]. A Figura 3.11 exemplifica o fluxo de um projeto em ML.



Figura 3.11: Fluxo de um projeto em ML. Fonte: [105].

O ser humano tem a capacidade de adaptar-se, modificar seu comportamento e melhorar suas respostas, e este é o foco do ML. Quando um de seus algoritmos está construindo, treinando ou induzindo algum conhecimento, é comum afirmar que ele está aprendendo. Quando ele realiza alguma estimativa através de testes ou predição, afirma-se que o modelo está sendo aplicado. Conforme a necessidade do estudo e o *dataset* disponível, é possível utilizar vários algoritmos de ML, como apresenta a Figura 3.12. A grande maioria deles se encaixa em uma das abordagens a seguir [106]:

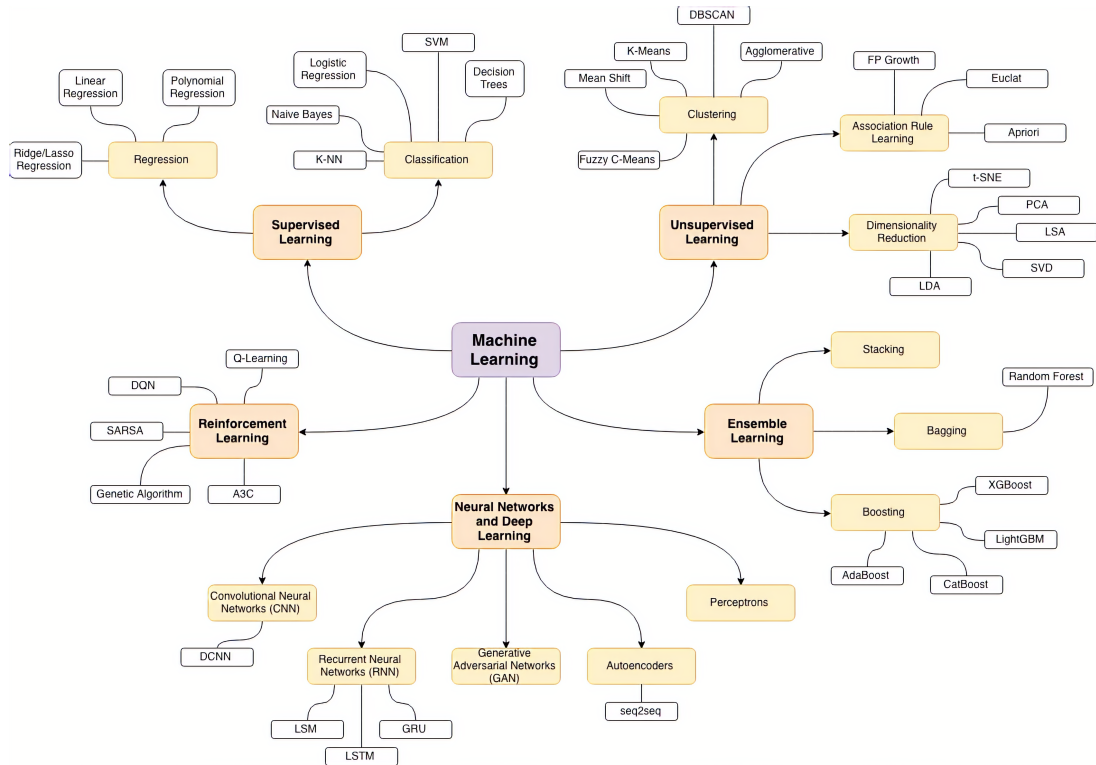


Figura 3.12: Visão geral dos algoritmos de ML. Fonte: disponível em: <https://medium.com/@andressasivolella/afinal-o-que-machine-learning-e-redes-neurais-fazem-7c89e1885064>.

- **Aprendizado supervisionado:** a partir de um conjunto de características previsoras e suas classes, os dados são rotulados, pois sabe-se qual será a saída esperada. O treinamento do modelo consiste em prever as saídas de algo desconhecido. Quando esse conjunto de características possui identificação discreta, de fato ocorrerá uma classificação, mas ao possuir valores contínuos, pode ocorrer uma regressão. A Figura 3.13 apresenta uma visão geral do aprendizado supervisionado. O conjunto de características geralmente é dividido em dois grupos de atributos, um grupo X que, como descrito no início, são os atributos previsores, utilizados na etapa de treinamento e teste, e é a partir do conjunto de treinamento que o modelo será construído, e um grupo Y com os rótulos que desejamos realizar a predição, o qual permite avaliar o comportamento do modelo em relação a dados que ele desconhece.

- **Aprendizado não supervisionado:** neste tipo de treinamento, as classes não são conhecidas, ou seja, não têm-se os atributos rotulados, portanto as saídas desejadas a serem estimadas são desconhecidas. Neste treinamento, o algoritmo explora os dados e tenta encontrar uma relação entre os atributos. Neste tipo de abordagem, um das estratégias principais é dividir os dados em agrupamentos (*clusters*), fazendo com que os atributos de um mesmo agrupamento possuam maior similaridade entre si.
- **Aprendizado semi-supervisionado:** neste caso, o conjunto de treinamento não possui todos os rótulos dos registros. Para alguns registros, a classe é desconhecida. É este um dos motivos da utilização deste tipo de aprendizado, pois grande parte dos dados disponíveis nos bancos de dados não estão rotulados, como por exemplo diagnósticos médicos, tornando-se uma tarefa dispendiosa.

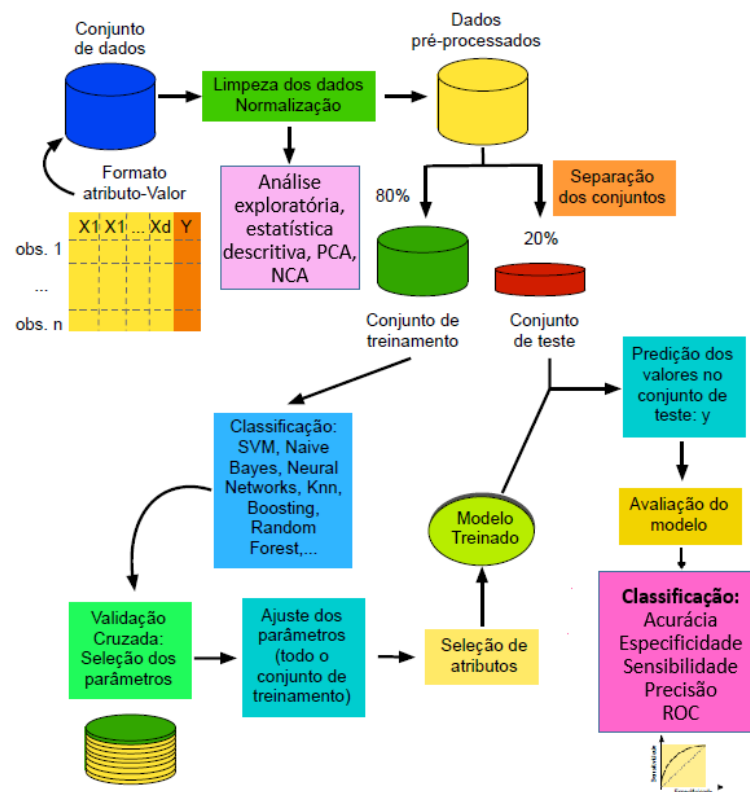


Figura 3.13: Visão geral do aprendizado supervisionado.

3.5.1 Multilayer Perceptron (MLP)

As redes neurais artificiais (RNA) foram desenvolvidas com inspiração nos neurônios, células nervosas que processam, transmitem e armazenam informações através dos axônios que produzem atividade elétrica, e nos proporcionam os sentidos como tato, olfato, paladar e visão. O cérebro humano possui bilhões de neurônios distribuídos em

rede e trabalhando em paralelo. Um modelo matemático foi desenvolvido para simular o neurônio humano [107]. A Figura 3.14 demonstra este funcionamento.

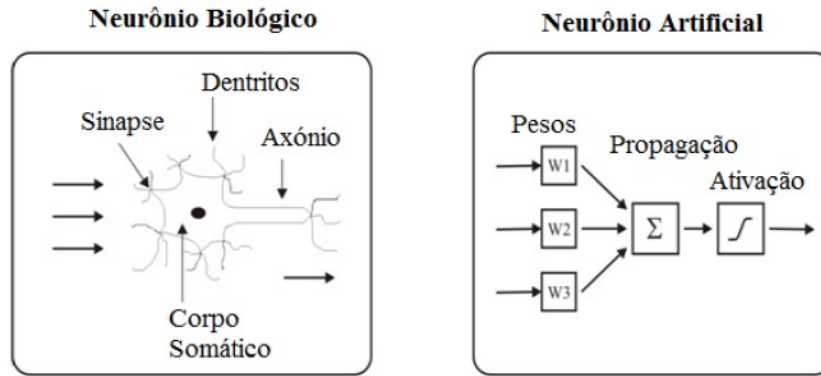


Figura 3.14: Estrutura do neurônio biológico *versus* neurônio artificial. Fonte: [107].

As RNAs com inúmeras arquiteturas têm sido utilizadas em diversas áreas, porém, arquiteturas multicamadas têm sido sem dúvida uma das mais utilizadas e constantemente encontradas na literatura. Devido a flexibilização que possui, ela consegue a partir de um único algoritmo propor soluções de qualidade para uma ampla variedade de problemas [108].

O *Perceptron* é uma arquitetura de rede com único neurônio, mas com capacidade para realizar classificação de padrões. No entanto, ele realiza o agrupamento de dados lineares que podem ser separáveis. Neste contexto, surgiram as redes neurais Perceptron Multicamadas (MLP) com o objetivo de classificação do grande número de dados que não são linearmente separáveis. Na MLP, os neurônios (chamados de nós) são distribuídos em camadas, geralmente três, denominadas camada de entrada, camadas intermediárias ou ocultas e camada de saída [109]. A Figura 3.15 traz uma representação desta descrição.

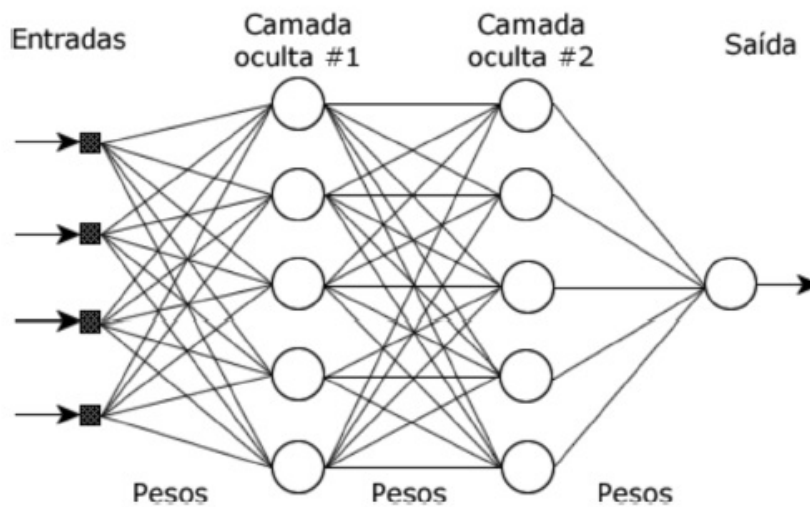


Figura 3.15: Arquitetura de uma MLP - Adaptado de [110].

O funcionamento básico de uma MLP consiste em propagação dos dados a partir da camada de entrada, camada a camada, até alcançar a camada de saída. O sinal obtido na camada de saída representa o resultado desejado (também chamado de resultado alvo), cuja diferença entre o resultado esperado e o resultado devolvido pela rede é chamado de erro. O erro é então retro-propagado, novamente camada a camada até alcançar a camada de entrada e os pesos são atualizados. Além desse funcionamento básico, a MLP se diferencia de outras arquiteturas de rede por caracterizar-se em: i) possuir para cada nó (neurônio) da rede uma função de ativação cuja curva apresenta suavidade, diferentemente do Perceptron que possui função de ativação com curvatura abrupta; ii) possuir uma ou mais camadas intermediárias (ocultas), os neurônios desta camada são os responsáveis pelo aprendizado de problemas complexos. Estes nós são diferentes dos nós das camadas de entrada e saída; e iii) possuir forte conectividade entre os nós, podendo um nó se conectar a qualquer outro nó da próxima camada e ser conectado por qualquer nó da camada anterior [110].

3.5.2 Logistic Regression (RL)

A RL, por ser uma técnica estatística, produzirá, a partir de um conjunto de atributos previsores, um modelo que possa ser capaz de prever uma saída categórica. Essa técnica mede a relação entre a variável dependente e variáveis independentes, estimando probabilidades [111], usando sua função logística subjacente, conforme abaixo :

$$p = \frac{1}{1 + e^{-y}} \quad (3.4)$$

3.5.3 Random Forest (RF)

Um algoritmo RF é um meta estimador que ajusta vários classificadores de árvore de decisão em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo. O tamanho da subamostra pode ser controlado com o parâmetro `max_samples`, caso contrário, todo o conjunto de dados é usado para construir cada árvore [112].

3.5.4 K-Nearest Neighbor (KNN)

K-vizinhos mais próximos (KNN) é um algoritmo de reconhecimento de padrões que armazena e aprende com os pontos de dados de treinamento, calculando como eles correspondem a outros dados no espaço n-dimensional. O KNN determina o rótulo de classificação de uma amostra, com base nas amostras de vizinhos, onde a variável k indica o número de vizinhos mais próximos aos quais uma medida de distância será aplicada, a fim de identificar a classe à qual a nova amostra pertence [113].

3.5.5 Support Vector Machine (SVM)

Uma máquina de vetores de suporte é um algoritmo simples, mas bastante usado pelos profissionais em atividades de classificação ou regressão. Ele funciona encontrando hiperplanos dentro de uma distribuição de dados, que pode-ser visualizado como uma linha que separa duas classes diferentes de dados. Muitas vezes existem muitos hiperplanos capazes de separar os dados, e o algoritmo selecionará a linha de separação ideal. No modelo SVM, o hiperplano ótimo é a linha divisória que oferece a maior margem entre as diferentes classes. SVMs são capazes de trabalhar em múltiplas dimensões se não conseguirem encontrar um hiperplano ideal para separar os dados em duas dimensões. Isso os torna extremamente eficazes para criar classificações de distribuições de dados complicadas. Quanto mais complexas as entradas de dados, mais preciso o SVM se torna, tornando-o excelentes ferramentas de aprendizado de máquina [114, 115].

3.5.6 Bagging Classifier (BC)

Bagging é um meta-estimador de conjunto no qual subconjuntos aleatórios do conjunto de dados de treinamento são usados para ajustar várias instâncias de um estimador de base. O estimador base padrão é um classificador de árvore de decisão, mas outros classificadores podem ser usados como estimador base (exploraremos isso mais adiante neste trabalho). A previsão final é obtida agregando as previsões individuais de cada instância do estimador [116].

3.5.7 Decision Tree (DT)

Uma árvore de decisão é um algoritmo de classificação de aprendizado de máquina supervisionado usado para construir modelos como a estrutura de uma árvore. Ele classifica os dados em categorias cada vez mais refinadas: de “tronco de árvore” a “galhos” a “folhas”. Ele usa a regra matemática **se-então** para criar subcategorias que se encaixam em categorias mais amplas e permite uma categorização precisa. Como as regras são aprendidas sequencialmente, do tronco à folha, uma árvore de decisão requer dados limpos e de alta qualidade desde o início do treinamento, ou as ramificações podem ficar superajustadas ou distorcidas [117, 118].

3.5.8 Gaussian Naive Bayes (GNB)

GNB é um algoritmo de classificação probabilístico baseado na aplicação do teorema de Bayes com fortes hipóteses de independência. No contexto da classificação, a independência refere-se à ideia de que a presença de um valor de uma característica não influencia a presença de outra (ao contrário da independência na teoria da probabilidade).

Naive refere-se ao uso de uma suposição de que as características de um objeto são independentes umas das outras. No contexto do aprendizado de máquina, os classificadores de Bayes são conhecidos por serem altamente expressivos, escaláveis e razoavelmente precisos, até porque, seu desempenho se deteriora rapidamente com o crescimento do conjunto de treinamento. Uma série de características contribuem para o sucesso dos classificadores de Bayes. Mais notavelmente, eles não exigem nenhum ajuste dos parâmetros do modelo de classificação, se ajustam bem a dimensão do conjunto de dados de treinamento e podem lidar facilmente com recursos contínuos [119].

3.5.9 Adaptive Boosting (AB)

Um classificador AdaBoost [120] é um meta-estimador que começa ajustando um classificador no conjunto de dados original e, em seguida, ajusta cópias adicionais do classificador no mesmo conjunto de dados, mas onde os pesos das instâncias classificadas incorretamente são ajustados de forma que os classificadores subsequentes se concentrem mais em casos difíceis.

3.5.10 Stochastic Gradient Descent (SGD)

Este estimador implementa modelos lineares regularizados com aprendizado estocástico de gradiente descendente (SGD): o gradiente da perda é estimado a cada amostra por vez (ao invés de usar o conjunto completo de dados) e o modelo é atualizado ao longo do caminho com um cronograma de força decrescente (também conhecido como taxa de aprendizado). Isso permite que o algoritmo seja mais rápido e possa lidar com grandes conjuntos de dados [121, 122].

3.5.11 Histogram-based Gradient Boosting (HGB)

Considerado uma árvore de decisão de aumento de gradiente [123], é um algoritmo de aprendizado de máquina amplamente utilizado, devido à sua eficiência, precisão e interpretabilidade. O HGB alcança desempenhos de ponta em muitas tarefas de aprendizado de máquina, como por exemplo em classificação multiclasse. Nos últimos anos, com o surgimento dos *big datas* (tanto em número de recursos quanto em número de instâncias), o HGB está enfrentando novos desafios, principalmente na escolha entre precisão e eficiência. Implementações convencionais de HGB precisam, para cada recurso, varrer todas as instâncias de dados para estimar o ganho de informação de todos os pontos de divisão possíveis. Portanto, suas complexidades computacionais serão proporcionais tanto ao número de recursos quanto ao número de instâncias. Isso torna essas implementações muito demoradas ao lidar com grandes bases de dados.

3.5.12 Extreme Gradient Boosting (XGB)

XGBoost (como é largamente conhecido) é um algoritmo de aprendizado de máquina baseado em árvores de decisão. Ele é uma variação do algoritmo GB, que foi criado para melhorar a performance e eficiência computacional. XGBoost usa uma técnica chamada regularização L1 e L2, que ajuda a controlar o sobreajuste. Ele também implementa um método de construção de árvores chamado “regressão quantílica” (regressão utilizada na análise estatística), que permite ajustar a distribuição dos dados de treinamento. Isso ajuda a melhorar a performance do modelo em conjuntos de dados desbalanceados. XGBoost também é amplamente utilizado em competições de ciência de dados e é conhecido por sua capacidade de lidar com grandes conjuntos de dados e sua alta precisão. Ele é muito popular para uso em problemas de classificação e regressão, e também tem aplicações em aprendizado não supervisionado, como por exemplo em detecção de anomalias [124].

3.5.13 Extra Trees Classifier (ET)

O ET implementa em suas classes um meta estimador que ajusta um número de árvores de decisão aleatórias (também conhecidas como árvores extras) em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo [125, 126].

3.5.14 Gradient Boosting (GB)

O Gradient Boosting constrói um modelo aditivo de uma maneira progressiva e permite a otimização de funções de perda diferenciáveis arbitrárias. O GB cria uma série de árvores de decisão fracas e adiciona-as ao modelo final para melhorar a precisão. A cada iteração, o algoritmo tenta corrigir os erros cometidos pelas árvores anteriores, adicionando uma nova árvore que se concentra nos exemplos que foram difíceis de serem previstos corretamente pelo modelo atual. Ele usa o gradiente da função de perda para guiar a construção de cada árvore de decisão, de forma que cada árvore tente corrigir os erros cometidos pelas árvores anteriores [127].

3.5.15 Linear Discriminant Analysis (LDA)

Essa técnica, não é apenas uma ferramenta de redução de dimensão, mas também um método de classificação robusto. Com ou sem suposição de normalidade dos dados, podemos chegar às mesmas características do LDA, o que explica sua robustez. O LDA é usada como uma ferramenta para classificação, redução de dimensão e visualização de dados. Apesar de sua simplicidade, o LDA geralmente produz resultados de classificação robustos, decentes e interpretáveis. Ao lidar com problemas de classificação do mundo

real, o LDA costuma ser o método de *benchmarking* antes que outros mais complicados e flexíveis sejam empregados [128].

3.5.16 Grid Search CV

GridSearchCV (assim como o comando é descrito) trata-se de uma técnica de seleção exaustiva para modelos de ML⁷. Ela avalia todas as combinações de diferentes valores de hiperparâmetros para obter a melhor performance do modelo. Em grandes *datasets* ou em grande número de parâmetros, ela pode ser muito lenta [73]. O ideal é utilizá-lo em pequenos domínios de hiperparâmetros.

3.5.17 Método Ensemble

O método ensemble é uma técnica de aprendizado de máquina que combina vários modelos individuais para produzir uma previsão mais precisa e robusta. Isso é feito combinando as previsões de vários modelos de maneira inteligente, geralmente através de votação ou média. É muito útil para lidar com problemas de aprendizado de máquina complexos, pois combina as forças de vários modelos e minimiza suas fraquezas. Ele também é útil para lidar com conjuntos de dados desbalanceados e ruidosos [129, 126]. A Figura 3.16 ilustra seu funcionamento.

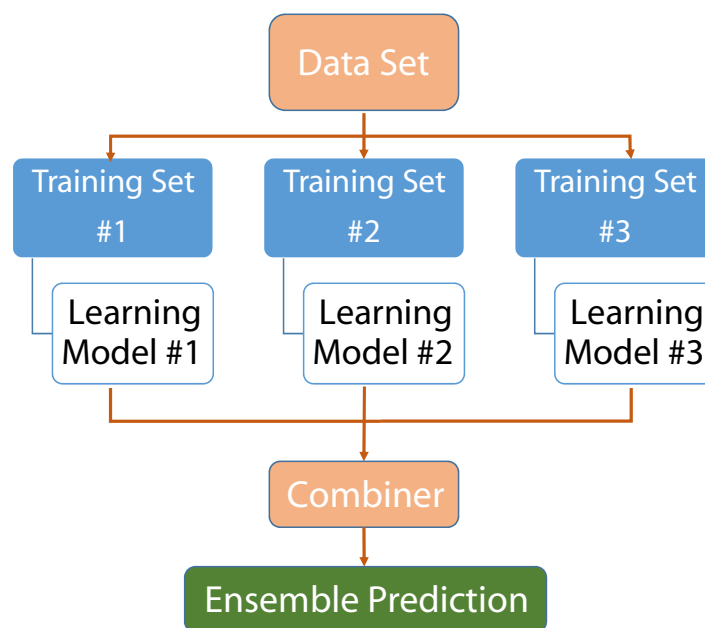


Figura 3.16: Diagrama simplificado do processo de classificação de subconjuntos paralelos, como por exemplo, o *bagging*.

Boosting e Bagging são dois métodos populares de aprendizado ensemble, cada um com sua própria abordagem para combinar vários modelos:

⁷https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

- Boosting é um método aditivo, onde vários modelos são treinados sequencialmente, com cada modelo tentando corrigir os erros cometidos pelos modelos anteriores. O AB é um exemplo de algoritmo de Boosting (descrito na década de 90), por sinal, o principal deles. O boosting ajusta a distribuição dos dados de treinamento a cada iteração, dando mais peso aos exemplos mal classificados pelos modelos anteriores.
- Bagging, é chamada de técnica em conjunto, onde várias versões do mesmo modelo são treinadas em subconjuntos aleatórios do conjunto de treinamento e suas previsões são combinadas. O algoritmo BC é um exemplo de algoritmo de Bagging.

3.6 *Datasets* de imagens cerebrais

Um *dataset* ou conjunto de dados é definido como uma coleção de dados organizados ou tabulados, em que cada instância desse conjunto pode possuir determinadas características. Um banco de dados se assemelha a uma tabela relacional, onde se encontram as linhas que referem-se aos registros e as colunas que referem-se aos atributos destes registros [130]. Em imagens, não é muito diferente: lê-se os pixels de determinada imagem e, através de técnicas diversas, são extraídas as características que vão constituir uma informação a respeito daquele objeto. Em relação à grande maioria dos trabalhos desenvolvidos na área do diagnóstico por imagem, são obtidas informações após identificação de uma região de interesse, etapa conhecida como segmentação [55]. De posse tanto das imagens médicas quanto das ROI, são construídos os repositórios, largamente utilizados por pesquisas, estudos e trabalhos. Alguns deles são descritos na sequência.

Akkus et al. [39] utilizaram um *dataset* construído e disponibilizado por Erickson et al. [131], com 159 casos de LGG diagnosticados com e sem codeleção cromossômica 1p/19q através de biópsias. Esse *dataset* propositalmente é o primeiro a ser citado, pois foi o *dataset* selecionado para ser utilizado nos experimentos desenvolvidos no Capítulo 4 deste trabalho. Ele será detalhado na seção 4.1.1.

Recentemente, Decuyper et al. [42] utilizaram um conjunto de imagens com 628 instâncias em sua pesquisa. Os mesmos não detalham quantas imagens usaram de cada instância, mas descrevem ter imagens do TCIA (tumores LGG e HGG), do BraST2019⁸ (repositório de imagens de câncer para competições) e do Hospital Universitário de Ghent (local onde a pesquisa foi desenvolvida).

Bakas et al. [132] em seu trabalho de avaliação do melhor algoritmo de ML para análise de tumores cerebrais utilizaram um *dataset* do BraST2018⁹. As imagens disponíveis em 2018 também foram utilizadas para medir o grau de progressão dos tumores em sub-regiões. Além da disponibilidade da ROI, o trabalho descreve que, para os algoritmos

⁸disponível em <https://www.nitrc.org/>

⁹disponível em <https://www.nitrc.org/>

testados, foram disponibilizadas características radiômicas.

Em um renomado estudo [19], Aerts et al. não descrevem a origem do seu *dataset*, mas registram a utilização de 1019 imagens de TC. A partir de imagens de pulmão, cabeça e pescoço (coluna cervical), são extraídos 440 atributos radiômicos de forma e textura que possam quantificar o tumor. Os dados extraídos neste estudo não sofreram normalização, sendo utilizados dados diretamente do padrão *Digital Imaging and Communications in Medicine* (DICOM).

A neuroimagem une pesquisadores do mundo inteiro com intenção de coletar, validar e utilizar dados que possam definir a progressão de uma doença. Por meio de imagens, dados clínicos e biomarcadores são construídos os *datasets* para análise cerebral. Além do foco desta tese de doutorado nas alterações tumorais, há diferentes estudos acerca de outras patologias, como doença de Alzheimer, esclerose múltipla, Parkinson, dentre outras.

3.6.1 Protocolos T1, T2 e Flair

Ao utilizar imagens radiológicas para extração de características quantitativas, busca-se conhecer um padrão mesmo estando sujeito a variações. Uma das formas de realizar este conhecimento é através da classificação, onde as classes que se deseja reconhecer se formam a partir das características a serem obtidas. É importante lembrar que os descritores utilizados para extração das características em uma imagem podem se adequar melhor conforme a característica extraída, por exemplo, forma, dimensão, cor, textura dentre outros [133].

Neste sentido, a RM é um método de imagem médica não invasivo, que fornece excelente contraste de tecidos moles, fluxo sanguíneo, difusão de tecidos, e reconhecidamente o padrão-ouro para o diagnóstico e controle de tratamento de tumores cerebrais [134].

As posições do corpo humano muitas vezes estão associadas a um plano imaginário que atravessam o mesmo e se encontra em uma posição anatômica. No estudo da RM, enfatiza-se a anatomia seccional, que compreende os seguintes cortes e planos: **plano sagital**, um plano longitudinal em que o corpo é dividido em lado direito e esquerdo; **plano coronal**, um plano longitudinal em que o corpo é dividido em parte anterior e posterior; e **plano axial**, um plano transversal que forma um ângulo reto com plano longitudinal e divide o corpo em parte posterior e inferior. A Figura 3.17 exemplifica tal visualização. No decorrer da execução dos exames de RM, esses planos passam por cortes ou fatias, sendo eles cortes longitudinais nos planos sagital e coronal, e cortes transversos no plano axial. A Figura 3.18 apresenta um exemplo de cortes realizados em uma imagem de RM do crânio no plano axial.

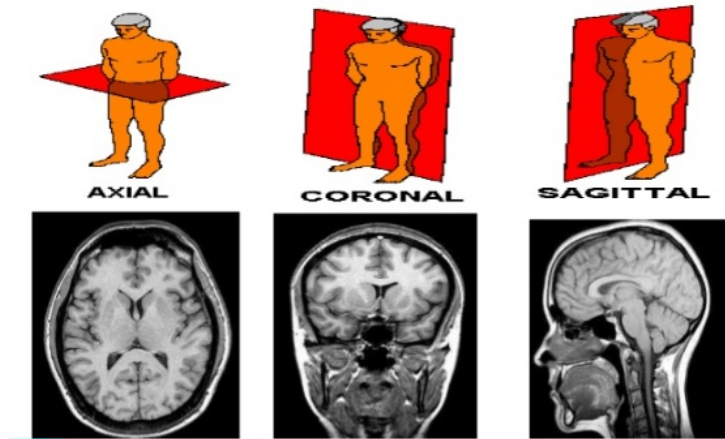


Figura 3.17: Exemplo dos planos axial, sagital e coronal - imagem de domínio público, disponível no Google Imagens em <https://www.google.com/imghp?hl=pt-br>.

PLANEJAMENTO AXIAL

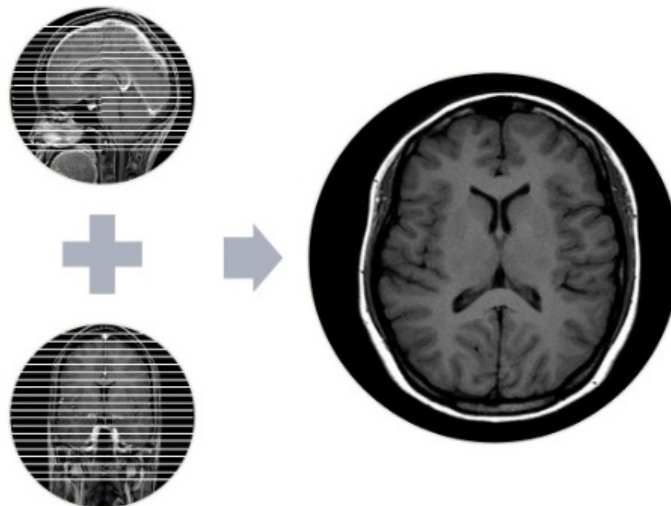


Figura 3.18: Exemplo de cortes em uma imagem de RM cerebral no plano axial - imagem de domínio público, disponível no Google Imagens em <https://www.google.com/imghp?hl=pt-br>.

A RM consiste na interação dos átomos de hidrogênio envolvidos por um fortíssimo campo magnético e submetidos a pulsos de radiofrequência. Sabe-se que esses átomos de hidrogênio estão sempre girando em torno de seu eixo. O momento angular do núcleo de hidrogênio é chamado de spin. Comumente usa-se o termo spin para se referir aos próprios núcleos de hidrogênio. As ponderações T1 e T2 são as sequências de contraste que vão medir as diferenças dos parâmetros nos tecidos. A ponderação em T1 basicamente representa a troca de energia dos núcleos de hidrogênio com o meio ambiente, enquanto a ponderação em T2 representa a troca de energia dos núcleos de hidrogênio vizinhos entre

si. Estas duas sequências são as principais utilizadas no diagnóstico de uma lesão. Por exemplo, caso seja preciso caracterizar melhor uma lesão em T1, usa-se um tempo de repetição (TR) longo e um tempo de eco (TE) curto, mas se é preciso caracterizar uma lesão com ponderação em T2, a situação é invertida, pois aplica-se um TR curto e um TE longo. Ou seja, TR controla a ponderação em T1 e TE em T2. As moléculas se comportam com maior mobilidade na água, gerando uma interação com os campos magnéticos. Como os tecidos patológicos apresentam-se em grande parte livres na água, por exemplo, edemas, tumores ou inflamações nestes tecidos, o T1 apresenta baixo sinal (hipointenso), enquanto o T2 apresenta alto sinal (hiperintenso). Sendo assim, torna-se melhor identificar lesões em T2, pois irá ocorrer a quebra de barreiras da massa encefálica e as características das células serão alteradas, e lesões e proteínas mesmo sem o uso do contraste gadolínio serão identificadas. O gadolínio é um tipo de contraste venoso utilizado em exames de imagens, e é utilizado como um marcador de atividades inflamatórias, permitindo a identificação de novas lesões inflamatórias no cérebro. Neste contexto pode-se ter ainda a ponderação T1 pós-contraste. Além desta, têm-se a sequência Flair conhecida como uma ponderação em T2, e como a própria sigla representa, é uma recuperação de inversão com fluido atenuado, ou seja, em T2 tudo que for líquido passará por inversão de sinal. Portanto, ela torna-se mais sensível na identificação de processos inflamatórios, processos tumorais, áreas de edema ou lesões, e numa sequência Flair ponderada em T2 essas regiões vão brilhar [135]. A Figura 3.19 exemplifica imagens de RM do cérebro extraída com sequências T1, T2 e Flair.

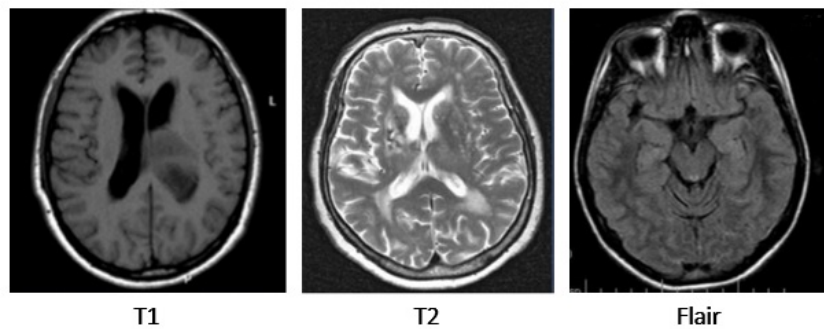


Figura 3.19: Imagens de RM do cérebro nas sequências T1, T2 e Flair - Imagens de domínio público disponível em [31].

3.6.2 Bases de dados

Os *datasets* (conjunto de dados) são considerados o insumo principal para extração e análise de informações. Alguns são formados por dados vindos diretamente da fonte, por exemplo, imagens médicas são fornecidas por clínicas, hospitais e instituições de pesquisa. Ou ainda, são formados por dados tabulados, sendo informações organizadas em registros, os quais possuem atributos extraídos de alguma base original.

Um dos repositórios mundiais com excelentes *datasets* e referência em imagens de câncer é o The Cancer Imaging Archive (TCIA) [51]. Sempre em constante atualização e com diversidade de coleção de imagens em relação à anatomia do corpo humano, o TCIA torna-se passagem obrigatória quando o assunto estudado é câncer versus imagiologia médica. O TCIA começa a apresentar forte integração com outro programa fundamental na análise de informações do tumor, o TCGA. O programa foi desenvolvido e está disponível pelo National Cancer Institute (NCI), possuindo informações genômicas extraídas de aproximadamente 20.000 cânceres em 33 tipos diferentes de câncer. Atualmente, já são 2,5 petabytes de dados genômicos disponíveis para a comunidade acadêmica utilizar em pesquisas [136].

Outra iniciativa muito interessante iniciada pelo NCI é o Quantitative Imaging Network (QIN), uma cooperativa com o objetivo de desenvolvimento de imagens quantitativas e métodos que melhorem a eficácia de novas terapias contra o câncer [137]. O QIN tem sido a “menina dos olhos” do NCI devido a grande possibilidade de ser a força propulsora de novas imagens e métodos, incluindo radiômica. Além disso, há um forte patrocínio da Radiological Society of North America (RSNA) e do National Institute for Biomedical Imaging and Bioengineering em patrocinarem o Quantitative Imaging Biomarkers Alliance (QIBA), um grande esforço em termos quantitativos radiológicos [138].

Um outro repositório bastante diversificado e muito utilizado nas diversas técnicas de ML é o Machine Learning Repository (UCI). Atualmente, com aproximadamente 500 *datasets* disponíveis, os dados disponibilizados são sempre em arquivos com extensão csv, txt, xls, dentre outros. Um dos *datasets* largamente utilizados é o Breast Cancer Wisconsin (Diagnostic) Data Set, o qual possui informações de 569 casos de câncer de mama, com 32 atributos extraídos de cada caso [139].

Bastante consultado também é o repositório Kaggle Datasets¹⁰. O mesmo disponibiliza alguns conjuntos de dados interessantes para competições de desempenho de processamento em técnicas diversas de ML. Dependendo da competição oferecida, são disponibilizadas as imagens ou dados já extraídos e organizados em arquivos tabulados.

E, por fim, é importante citar o NeuroImaging Tools and Resources Collaboratory (NITRC) definido como uma comunidade de pesquisas voltada para neuroimagem e câncer. O mesmo disponibiliza o repositório NeuroImaging Data Repository¹¹ (IR). O NITRC em parceria com o Center for Biomedical Image Computing and Analytics (CBICA) da Porelman School of Medicine na University of Pennsylvania, promove anualmente o Brain Tumor Segmentation Challenge (BraTS), um desafio para segmentação de tumores cerebrais. O MICCAI, nos anos seguintes ao das competições, disponibiliza as imagens cedidas aos competidores, para que possam ser utilizadas em pesquisas [140]. Ambos NITRC e CBICA têm demonstrado um forte interesse em competições e pesquisas

¹⁰disponível em <https://www.kaggle.com/datasets>

¹¹disponível em <https://www.nitrc.org/>

envolvendo LGG e HGG.

Capítulo 4

Análise radiômica da codeleção cromossômica 1p/19q

Este capítulo apresenta os quatro experimentos realizados com as características radiômicas através dos algoritmos de aprendizado supervisionado para classificar de forma não invasiva da codeleção cromossômica 1p/19q em gliomas de baixo grau.

Após apresentar os métodos e resultados de cada experimento, apresentaremos na seção 4.10 uma comparação dos nossos resultados com métodos da literatura.

4.1 Softwares e equipamentos

Todo o conjunto de softwares instalados, bem como todo o desenvolvimento do estudo e classificação dos conjuntos de dados de treinamento e teste foram realizados em um equipamento com sistema operacional Windows 10, processador Intel Core i7-4600M com arquitetura x86-64 (64 bits), 8 Gb de memória *random access memory* (RAM) e dispositivo de armazenamento *solid state disk* (SSD) com 256 GB utilizando a ferramenta de edição de código Jupyter Notebook versão 6.2.0 e Spyder versão 4.1.0, ambos do pacote Anaconda Navigator. Em algumas oportunidades, foram executados alguns dos algoritmos desenvolvidos no Google Colab¹ para comparativo de desempenho do equipamento e do modelo de classificação.

4.2 Dataset

Para este estudo, utilizou-se o *dataset* LGG-1P19QDeletion [131] do TCIA [51]. Esse conjunto de dados apresenta imagens de RM pré-operatória de 159 pacientes com LGG, mapeados em um período de 9 anos. Todos os pacientes tiveram biópsia comprovando o status dos cromossomos 1p/19q (codeletados ou não-codeletados). As imagens

¹disponível em: <https://colab.research.google.com/>

foram coletadas na Mayo Clinic, Minnesota, USA, administrado por Mayo Foundation for Medical Education and Research. Conforme mostrado na Figura 4.1, são 102 pacientes com codeleção e 57 pacientes com cromossomos 1p/19q não deletados. Destes LGGs, 17 são astrocitoma, 45 são oligodendroglioma e 97 oligoastrocitoma. As imagens foram adquiridas usando ponderação T2 e classificadas em tumores 1p/19q codeletados e 1p/19q não-codeletados, conforme exemplo na Figura 4.2. Segundo os autores de Akkus et al. [39], as imagens foram adquiridas utilizando um protocolo muito consistente, incluindo cortes de 1 mm e 3 mm em aparelhos de 1,5 T e 3 T. As imagens foram fornecidas em volume 3D e não cortes específico. As regiões de interesse foram segmentadas em 2D, utilizando a ferramenta semiautomática 3DSlicer [62] desenvolvida pelos autores.

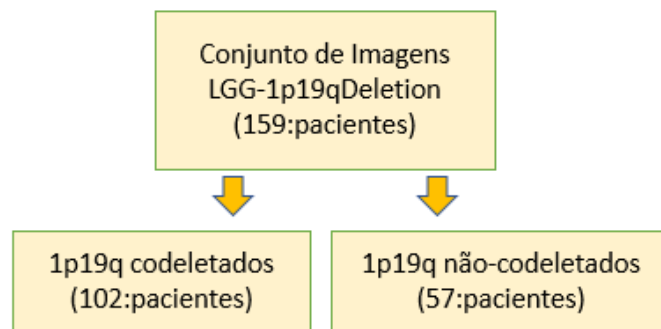


Figura 4.1: Divisão do dataset entre instâncias codeletados e não codeletados.

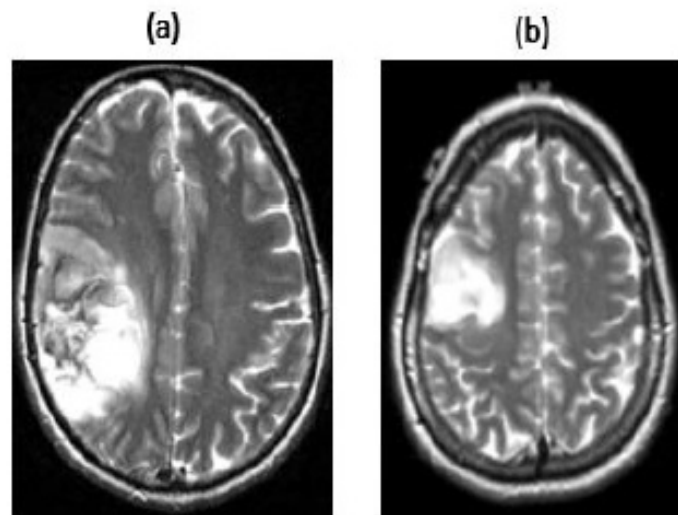


Figura 4.2: Exemplos de imagens de RM ponderadas em T2 de gliomas de baixo grau (LGG): (a) grande lesão infiltrativa heterogênea frontoparietal direita, com áreas de baixa intensidade de sinal, sugerindo hemorragia ou calcificação, com diagnóstico de oligodendroglioma (codeleção 1p/19q); e (b) lesão infiltrativa homogênea frontal direita com alta intensidade de sinal e diagnóstico de astrocitoma (ausência de codeleção 1p/19q) [51].

4.3 Extração das características

Após acesso ao TCIA e download do *dataset*, utilizou-se a ferramenta 3DSlicer [62] para realizar a conversão das imagens de RM que se encontravam no formato .nii (NIFTI) para o formato .nrrd (NRRD). Arquivos NIFTI vieram em substituição ao formato *Analyze*, o qual não possuía informações sobre o domínio espacial do objeto, provocando ambiguidades na interpretação dos dados extraídos (o mais visível era a falta de informação de orientação entre os lados esquerdo e direito das imagens do cérebro). O novo formato .nii foi definido pelo Data Format Working Group (DFWG) em 2003, quando os representantes de softwares de neuroimagem concordaram em incluir novas informações ao novo formato e inclusive o uso nativo, permitindo importação e exportação das imagens [141]. Os arquivos NRRD são aqueles em que os dados e o cabeçalho se encontram no mesmo arquivo. Foram projetados para visualização científica e processamento de imagens que envolvem múltiplas dimensões. Bastante utilizado nas aplicações de neuroimagem, é um arquivo que carrega simplicidade e flexibilidade, podendo ser aberto em um editor de texto simples. Outro recurso do formato de arquivo NRRD é o suporte de pares de chave/valor personalizado no cabeçalho para permitir o armazenamento de informações anonimizadas do usuário [142].

Em seguida, utilizou-se a biblioteca PyRadiomics para extração das características quantitativas. A plataforma PyRadiomics permite a extração de grande número de características quantitativas em imagens médicas com finalidade de diagnóstico do câncer. O conjunto volumétrico das imagens de RM cerebral e das máscaras correspondente a cada paciente do *dataset* foram processadas em lote conforme linha de comando `\footnote{pyradiomics entradas.csv -p c:/users/tony/params.yaml -o resultado1}` em que, `pyradiomics` representa o comando executável da plataforma para execução em lote, `entradas.csv` é o arquivo de entrada contendo a relação dos arquivos .nrrd dos pacientes e local de armazenamento, conforme ilustrado na Figura 4.3.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID,Image,Mask													
2	lgg_image_104,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-104_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-104-Segmentation.nrrd													
3	lgg_image_203,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-203_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-203-Segmentation.nrrd													
4	lgg_image_210,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-210_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-210-Segmentation.nrrd													
5	lgg_image_216,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-216_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-216-Segmentation.nrrd													
6	lgg_image_218,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-218_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-218-Segmentation.nrrd													
7	lgg_image_219,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-219_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-219-Segmentation.nrrd													
8	lgg_image_220,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-220_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-220-Segmentation.nrrd													
9	lgg_image_223,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-223_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-223-Segmentation.nrrd													
10	lgg_image_225,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-225_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-225-Segmentation.nrrd													
11	lgg_image_229,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-229_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-229-Segmentation.nrrd													
12	lgg_image_231,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-231_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-231-Segmentation.nrrd													
13	lgg_image_233,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-233_T2.nrrd,D:\UNB\Imagens_RM\Imagens_TRATADAS\LGG-233-Segmentation.nrrd													

Figura 4.3: Trecho do arquivo de entrada na plataforma PyRadiomics com local de armazenamento das imagens e sua respectiva segmentação.

Também é informado um arquivo de configuração, o `Params.yaml` que contém a configuração das classes e filtros que serão utilizados para extração das características radiômicas. Neste estudo, foram extraídas apenas as características da classe “original”, ou

seja, sem aplicação de filtros. A Tabela 4.1 demonstra os descritores utilizados, conforme apresentados na subseção 3.3.2.

Conforme nota de rodapé 2, após o parâmetro de configuração das classes radiômicas, é informado como parâmetro o arquivo de saída, `resultado1.csv`, com as características radiômicas extraídas do conjunto de imagens. Este arquivo de saída, assim como o arquivo de entrada, são do tipo `.csv` (*comma separated vales*), arquivos bastante utilizados na gestão de dados, podendo ser lidos por uma grande variedade de softwares. Particularmente, a linguagem de programação Python tem bibliotecas que trabalham muito bem com este tipo de extensão de arquivo, por exemplo, a biblioteca Pandas². O arquivo de saída possuirá tantos registros quanto o número de pacientes informado no arquivo de entrada, representando as linhas do arquivo. Já as características extraídas são representadas pelas colunas do arquivo, que, de acordo com a Tabela 4.1, são 120 características radiômicas (120 colunas). Em cada experimento relatado a seguir, será apresentado o sub-conjunto de características que foi utilizado.

Tabela 4.1: Quantitativo de características radiômicas extraídas.

Nome da classe	Número de Características	Descrição da classe
estatísticas de primeira ordem	19	Descreve a distribuição de intensidades do voxel dentro da ROI
forma	26	Descreve o tamanho tridimensional e a forma da ROI.
GLCM	24	Quantifica a coocorrência de intensidades do voxel dentro da ROI.
GLRLM	16	Quantifica as ocorrências de voxels consecutivos dentro da ROI que possuem o mesmo nível de cinza.
GLSZM	16	Quantifica o número de voxels conectados que compartilham a mesma intensidade de nível de cinza na ROI.
GLDM	14	Quantifica as dependências de nível de cinza na ROI.
NGTDM	5	Quantifica as diferenças de tons de cinza vizinhos na ROI.

4.4 Normalização das características

As características radiômicas apresentam-se em diferentes ordens de magnitude. Assim, a normalização foi utilizada para que todas apresentem intervalos de valores semelhantes. Isso ajuda o algoritmo de classificação, garantindo que todos os recursos sejam ponderados de forma semelhante. O método de padronização foi aplicado a cada um dos 120 conjunto de características, de modo que cada um deles tenha média zero e desvio padrão unitário [143, 144]. Os recursos normalizados são calculados como $z = (x - \mu)/\sigma$,

²documentação disponível em : <https://pandas.pydata.org/>

em que x são os recursos não normalizados e μ e σ são a média e o desvio padrão do conjunto de recursos correspondente, respectivamente. Observa-se que μ e σ são calculados apenas a partir das instâncias de treinamento, mas são usados para normalizar os conjuntos de recursos de treinamento e teste.

4.5 Métricas de avaliação

As seguintes métricas amplamente aceitas [145] foram usadas para avaliar o desempenho de cada classificador na previsão da codeleção 1p/19q: acurácia balanceada (BCA), acurácia (Acc), sensibilidade (Se), especificidade (Spe), precisão (Pre), F1-score (F1) e área sob a curva (AUC).

Os valores de AUC representam a área sob a curva ROC (*receiver operating characteristic*) [146], e são obtidos durante a execução da etapa de classificação, pois são calculados usando a função `roc_auc_score` implementação da biblioteca `sklearn`. A AUC mede a probabilidade de uma amostra positiva ser rotulada como positiva pelo classificador.

As demais métricas são calculadas após a etapa de classificação, a partir das matrizes de confusão obtidas de cada simulação, conforme:

$$\text{BCA} = \frac{\text{Se} + \text{Spe}}{2} \quad (4.1)$$

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (4.2)$$

$$\text{Se} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.3)$$

$$\text{Spe} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4.4)$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.5)$$

$$\text{F1} = 2 \cdot \frac{\text{Pre} \cdot \text{Se}}{\text{Pre} + \text{Se}} \quad (4.6)$$

em que TP, FP, TN e FN representam o número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, respectivamente, obtidos de cada simulação. Instâncias que apresentam a codeleção são tratadas como “positivas”, e instâncias que apresentam-se não deletadas são tratadas como “negativas”. Todas as métricas foram convertidas em valores percentuais, multiplicando cada uma por 100%.

BCA é a média aritmética de sensibilidade e especificidade. Essa métrica é particularmente útil ao lidar com dados desbalanceados [147], como é o nosso caso.

A métrica F1 é a média harmônica de precisão e sensibilidade. Assim, um valor muito baixo de precisão resulta em uma pontuação F1 muito baixa, mesmo que a sensi-

bilidade seja muito alta; da mesma forma, um valor muito baixo de sensibilidade produz uma pontuação F1 muito baixa, mesmo que a precisão seja muito alta.

4.6 Compilado de técnicas

Para melhor compreensão do leitor nas próximas seções onde serão detalhados os quatro experimentos aplicados neste estudo, apresenta-se através da Tabela 4.2 um compilado das técnicas e algoritmos usados nos métodos de todos os quatro experimentos.

Tabela 4.2: Compilado de técnicas e algoritmos usados nos métodos dos quatro experimentos.

técnicas	experimento I	experimento II	experimento III	experimento IV
Divisão do dataset (treino / teste)	75% = 119 25% = 40	80% = 120 20% = 30	90% = 143 10% = 16	90% = 143 10% = 16
Quantidade de características	120	100	100	100 68
Seleção de características	feature importance			
Escalonamento de características	sim		sim	sim
Redimensionamento de características	PCA = 8	PCA = 8	NCA = 4	
Aumento de características	SMOTE	adição de ruído Gaussiano $\mu = 0 / \sigma = 1 / p = 0,2$		
Classificadores	MLP	RL RF	KNN com GridSearchCV	Ensemble (bagging/boosting)
Número de simulações	100 épocas	10	30	100

4.7 Experimento I

4.7.1 Objetivos

Este experimento apresenta o primeiro desenvolvimento do projeto desta tese de doutorado. O objetivo aqui, foi implementar uma rede neural multilayer perceptron, utilizando as características radiômicas extraídas conforme Tabela 4.1, aplicando redimensionamento no número de características e aumento no número de instâncias minoritárias para balancear as classes alvo.

4.7.2 Métodos

Após atividades de extração das características (seção 4.2), realizou-se a normalização das características (seção 4.3) através do método `StandardScaler` da biblioteca `sklearn` da linguagem de programação Python.

Redimensionamento das características

Tendo em vista o grande número de características de entrada (vide Tabela 4.1) em relação ao tamanho do *dataset*, após aplicar a normalização nas características radiômicas, executou-se a PCA por meio da biblioteca `sklearn.decomposition`, uma técnica de aprendizado não-supervisionado para redução linear de dimensionalidade, que foi utilizada para extrair informações de um espaço de 120 dimensões, projetando-as em um subespaço de 8 dimensões. Em sua essência, a PCA dedica-se a preservar as partes essenciais (que têm mais variação dos dados) e remover as partes não essenciais (com menos variação), sendo essa uma característica da técnica para acelerar os algoritmos de aprendizado de máquina. As 8 novas componentes, apresentaram os seguintes percentuais de redução: PCA1=47%, PCA2=19%, PCA3=13%, PCA4=5%, PCA5=3%, PCA6=3%, PCA7=2% e PCA8=2%, totalizando 92% de redimensionamento (Figura 4.4).

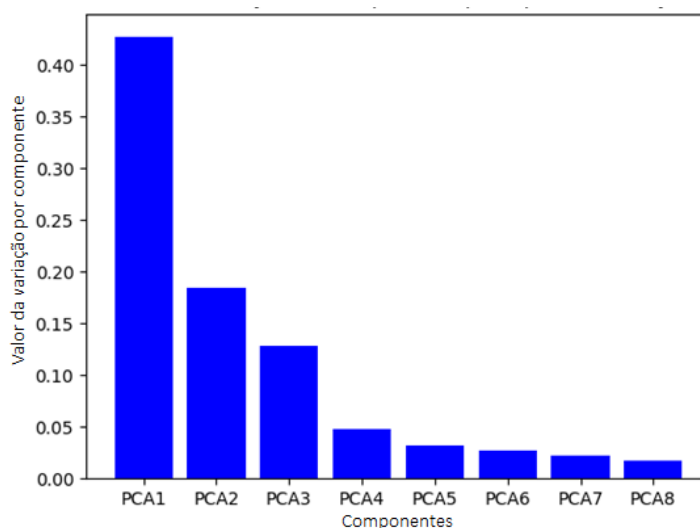


Figura 4.4: Gráfico com resultado do redimensionamento das características (normalizadas) pela técnica PCA.

Divisão do *dataset* em treino e teste

Após as etapas anteriores, o conjunto de dados com 159 instâncias (57 não deletadas e 102 codeletadas) foi dividido aleatoriamente em dois subconjuntos: 75% para treinamento (119 instâncias) e 25% para teste (40 instâncias).

Aumento das instâncias

Na sequência da divisão do *dataset*, realizou-se um aumento do número de instâncias a fim de se obter um balanceamento na quantidade de registros dos conjuntos de treinamento. Neste experimento, aplicou-se o aumento sintético através do método SMOTE, uma técnica de sobre amostragem, conforme descrito na seção 3.4.3. O método

SMOTE equilibra a distribuição de classes, aumentando aleatoriamente exemplos de classes minoritárias. De acordo com nosso *dataset* formado por 159 instâncias (pacientes), sendo 119 instâncias para treinamento; destas instâncias, 76 estavam classificados como “1” (oligodendrogliomas) e 43 como “0” (astrocitomas). A partir da classe minoritária, o método SMOTE gerou mais 33 instâncias sintéticas para o balanceamento das classes, obtendo assim 76 instâncias de cada rótulo. A Figura 4.5 demonstra o balanceamento realizado. Desta forma, obteve-se uma matriz de 152 linhas (instâncias) por 8 colunas (características redimensionadas pela PCA) para o conjunto de treinamento.

```
In [14]: print("After OverSampling, counts of label '1':  
{0}".format(sum(y_train_res==1)))  
...: print("After OverSampling, counts of label '0':  
{0}".format(sum(y_train_res==0)))  
After OverSampling, counts of label '1': 76  
After OverSampling, counts of label '0': 76
```

Figura 4.5: *Output* apresentando as classes do conjunto treinamento igualmente balanceadas.

Configuração e parametrização da MLP

As redes neurais artificiais (RNA) compostas por camadas e funções de ativação são consideradas uma das principais ferramentas para reconhecimento de padrões, já bastante consolidadas no meio científico e computacional. Dentre as RNAs mais difundidas, encontram-se a MLP, que têm sido utilizada em diversas aplicações. Contudo, é uma técnica de reconhecimento de padrões com uma fase de treinamento demorada e de grande esforço computacional. A Figura 3.14 ilustra a arquitetura de uma MLP. A rede MLP utilizada neste experimento caracteriza-se por ser uma rede sequencial, com uma camada de entrada, duas camadas ocultas e uma camada de saída. Foram utilizadas características de uma rede densa (conhecida como *fully-connected*), em que todos os neurônios se conectam aos neurônios da camada subsequente. A camada de entrada recebe os dados após a etapa de pré-processamento e nas camadas ocultas todos os neurônios possuem a mesma dimensão e aplicam uma função de ativação a cada neurônio da camada anterior. A função de ativação utilizada na implementação deste experimento foi a ReLU, conforme descrita na seção 3.5.1, a qual é amplamente descrita na literatura. Na camada de saída, utilizou-se a função de ativação sigmoide, também com excelente aceitação no meio acadêmico e bem descrita na literatura.

Inicialmente avaliou-se utilizar a biblioteca PyBrain, de código aberto, em linguagem Python, mas a mesma apresentava poucos recursos de otimização. Portanto, definiu-se por utilizar a biblioteca Keras (atualmente integrada a outra poderosa biblioteca chamada TensorFlow). Esta, por sua vez, aproveita muito bem as unidades de processamento gráfico e acelera os algoritmos de aprendizado de máquina, além de implementar uma infinidade de métodos para RNA [148]. Parametrizou-se na camada de

entrada 8 atributos previsores (ou seja, as características obtidas após redimensionamento pela PCA) e 152 instâncias (obtidas após etapa de geração de dados sintéticos através do SMOTE) para o conjunto de treinamento e 40 instâncias para o conjunto de teste. Nas camadas ocultas, foram utilizados 5 neurônios (metade do número de neurônios da camada de entrada mais 1, sugestão de exemplos de implementações da comunidade Python Brasil) e função de ativação ReLU. Na inicialização dos pesos, aplicou-se a técnica `random_uniform`, que inicializa os pesos através de uma distribuição uniforme entre $-0,05$ e $0,05$. Como o objetivo do experimento é identificar a deleção ou não do braço cromossômico $1p/19q$, a resposta será do tipo binária. Portanto, utilizou-se apenas um neurônio na camada de saída e a função de ativação aplicada foi a sigmoide.

Ao compilar a MLP, um dos parâmetros utilizados é o Adam com `batch_size = 10`, que através da descida do gradiente estocástico otimiza a função de ajuste dos pesos a cada dez registros. Outro parâmetro utilizado foi o `binary_crossentropy`, função que ativa o tratamento do cálculo do erro para problemas de classificação binária. Como o *dataset* é relativamente pequeno, foi utilizado o método `Cross_val_score = 10`. Com esse parâmetro, o algoritmo vai realizar validação cruzada, uma boa técnica para se evitar a ocorrência de *overfitting*. A validação cruzada divide o conjunto de treinamento em 10 dobras (chamadas de *k-folds*), e então treina em 9 dobras e valida em 1, fazendo isso até que todas as 10 dobras estejam validadas. Após a MLP treinada, realizou-se a avaliação no conjunto de instâncias separado para teste, por meio do parâmetro `binary_accuracy`, justamente por se tratar de um problema binário (somente duas classes de saída). A Tabela 4.3 apresenta um resumo das funções e parâmetros utilizados na implementação da MLP.

Tabela 4.3: Resumo das funções e parâmetros utilizados na MLP.

Função	Descrição
ReLU	função de ativação (camadas ocultas)
Sigmoid	função de ativação (camada de saída)
<code>random_uniform</code>	inicialização de pesos
<code>adam</code>	ajuste dos pesos
<code>binary_crossentropy</code>	tratamento de cálculo do erro
<code>binary_accuracy</code>	avaliação da acurácia
<code>cross_val_score</code>	treinamento dos dados em subconjuntos
<code>epochs</code>	quantidade de treinamentos executados

A MLP foi executada com 10 épocas e 100 épocas, por meio de configuração do parâmetro `epochs`. Como esse modelo de algoritmo representa uma abordagem de aprendizado supervisionado, na execução da rede neural, a MLP vai relacionar os atributos previsores (instâncias) e os atributos da classe de saída (rótulos).

A Figura 4.6 apresenta um fluxo do trabalho completo, desde a escolha do *dataset* até a avaliação dos resultados.

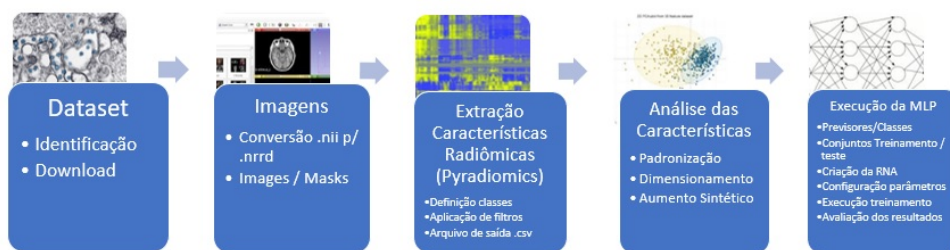


Figura 4.6: Fluxo completo de execução do experimento I.

4.7.3 Resultados

A Tabela 4.4 apresenta os resultados obtidos de sensibilidade, especificidade e precisão no conjunto de dados de teste na execução da MLP, em comparação com resultados obtidos com os classificadores SVM e RF. Os algoritmos clássicos de classificação, SVM e RF foram executados por meio da plataforma MLF (machine learning framework). A MLF é uma plataforma multiaplicativa que engloba o desenvolvimento de técnicas de pré-processamento de imagens, extração de características em imagens e diversos algoritmos de aprendizado de máquina. A mesma utiliza os algoritmos de classificação implementados pela biblioteca `sklearn` com os parâmetros default. A MLP foi testada com 100 épocas. A MLP apresentou melhor desempenho nas métricas sensibilidade e precisão em relação aos algoritmos SVM e RF; e em relação à métrica especificidade, a MLP manteve-se intermediária. Dentre os três resultados alcançados pela MLP, destaca-se o da precisão, apresentando-se com 85,0% de probabilidade de fornecer um resultado correto das instâncias classificadas corretamente como positivas dentre o total de instâncias positivas.

Tabela 4.4: Resultados da execução da MLP, SVM e RF no conjunto de dados de teste. O melhor resultado para cada métrica é destacado em negrito.

classificador	Se	Spe	Pre
MLP	88,0	73,5	85,0
SVM	85,0	70,0	77,5
RF	75,0	75,0	75,0

Nota: Os classificadores SVM e RF foram executados na MLF.

Além das métricas anteriormente utilizadas, que estima como foi o desempenho da MLP apresentando seu percentual de acerto, outra métrica utilizada em algoritmos de classificação é a matriz de confusão, que também oferece um detalhamento do desempenho da MLP. Ela exibe a distribuição dos registros em termos de suas classes atuais e de suas classes previstas. A Tabela 4.5 mostra o resultado obtido no conjunto de dados de teste após execução do algoritmo da MLP. Em um total de 40 casos, a rede errou apenas 3 casos de codeleção (falsos negativos) e 4 casos de não codeleção (falsos positivos). Sendo assim, acertou 22 casos que são positivos verdadeiros (casos de codeleção) e 11 casos que

são negativos verdadeiros (casos de não codeleção).

Tabela 4.5: Matriz de confusão do algoritmo MLP no conjunto de dados de teste.

casos de teste = 40	codeletado (real)	não codeletado (real)
codeletado (predito)	22	4
não codeletado (predito)	3	11

Utilizou-se também para avaliar o desempenho do modelo escolhido para classificação de dados, a curva ROC. A ROC é conhecida como uma ferramenta para medir e especificar problemas no desempenho do diagnóstico em medicina [149]. A curva ROC traça a taxa de positivos verdadeiros (isto é, sensibilidade) no eixo y versus a taxa de positivos falsos (isto é, $1 -$ especificidade) no eixo x , para diferentes valores do limiar de classificação. A Figura 4.7 exibe a curva ROC obtida com o método desenvolvido.

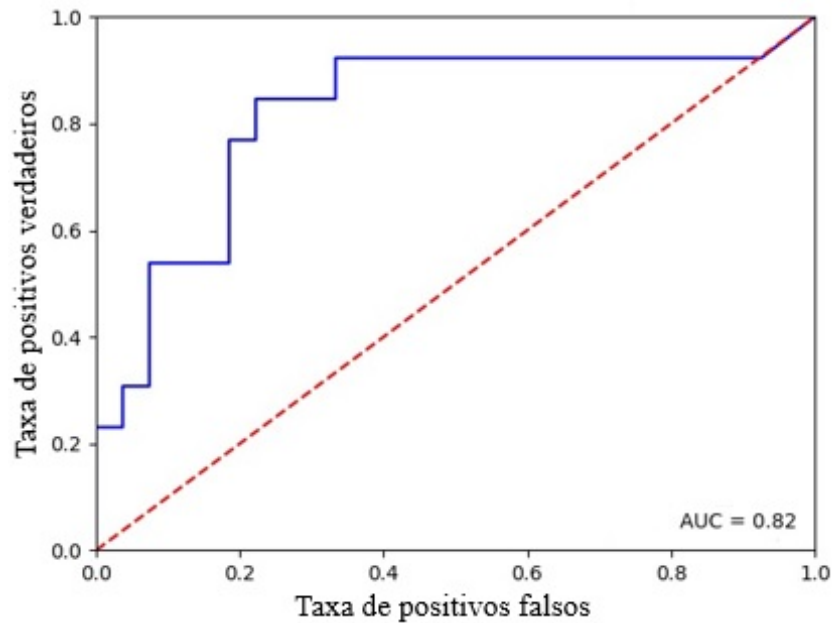


Figura 4.7: No gráfico, resultados acima e à esquerda da linha vermelha pontilhada são considerados melhores, enquanto que resultados abaixo e à direita da linha vermelha pontilhada são considerados piores. Um resultado no ponto $(x = 0, y = 1)$ indicaria classificação perfeita (especificidade 100% e sensibilidade 100%), enquanto que resultados ao longo da linha vermelha pontilhada indicariam resultados equivalentes a palpites aleatórios.

A área sob a curva ROC (AUC, do inglês *area under curve*) nada mais é que uma maneira de resumir a curva ROC em um único valor, agregando todos os limiares da curva ROC, calculando a “área sob a curva”. O valor do AUC varia de 0,0 até 1,0 e o limiar entre a classe é 0,5. Ou seja, acima desse limite, o algoritmo classifica em uma

classe e abaixo em outra classe. Quanto maior o AUC, melhor. A curva ROC está entre as métricas mais utilizadas para a avaliação de um modelo de *Machine Learning* [149].

4.7.4 Discussão

Um grande desafio para execução de algoritmos de aprendizado de máquina é a obtenção de um número adequado de imagens para treinamento e testes. Entretanto, o *dataset* continha as imagens necessárias para a avaliação da codeleção 1p/19q. As imagens disponíveis permitiram a correta conversão para o padrão de leitura PyRadiomics. As imagens com ponderação em T2 foram escolhidas por apresentar melhor condição de visualização do tumor devido as suas características.

O escalonamento das características radiômicas através da padronização (ou, normalização *z-score*) se fez necessário para eliminar a discrepância entre atributos de uma mesma classe. Após escalonamento, a execução do algoritmo juntamente com as demais etapas do pré-processamento ganhou mais desempenho na etapa de treinamento.

Mesmo sendo uma técnica não supervisionada, o redimensionamento do número de atributos pela PCA demonstrou ser eficiente em um modelo de aprendizado supervisionado. As 120 características radiômicas extraídas e padronizadas foram redimensionadas em 8 novas componentes, obtendo 92% do redimensionamento. Isto mostra que o algoritmo conseguiu manter a máxima proximidade entre as características.

O SMOTE alterou o número de instâncias sintéticas do conjunto de treinamento de forma que ambas as classes (0 - não codeletados e 1 - codeletados) apresentassem o mesmo número de casos. No desenvolvimento do projeto, o número de casos na classe minoritária (classe 0 - não codeletados) foi aumentado de 43 para 76. O SMOTE se baseia nas características dos casos existentes para criar seus vizinhos próximos sem perda de informação.

A execução da MLP mostrou capacidade de aprender as características extraídas da imagem, obtendo melhor acurácia no conjunto de dados de treinamento do que no conjunto de dados de teste, o que era esperado, sobretudo pela quantidade de casos ser pequena nos dados de teste. Conforme resultados da Tabela 4.4, a MLP apresentou precisão e sensibilidade melhor que os demais modelos comparados. O SVM e a RF foram utilizadas para comparação após vários outros algoritmos serem testados por meio da MLF e terem apresentados os melhores resultados, pois a MLP implementada na MLF utiliza-se de parâmetros padrão de implementação da biblioteca `sklearn`.

Vale registrar aqui que as 120 características, quando utilizadas sem nenhum pré-processamento para treinamento da MLP, resultaram em um valor de precisão de 65% no conjunto de dados de teste. O escalonamento, o redimensionamento e o aumento sintético dos dados de treinamento foram importantes para o aprendizado do modelo de classificação e obtenção de um resultado mais robusto da MLP sobre os demais mode-

los (algoritmos) avaliados. Os valores de especificidade e sensibilidade obtidos indicam que a MLP foi capaz de identificar com bom percentual, as características de codeleção cromossômica em gliomas de baixo grau. O resultado da matriz de confusão apresentado na Tabela 4.5 demonstra que o modelo escolhido se adaptou bem às características radiômicas, tendo uma boa taxa de acerto dos valores preditos em relação aos valores reais.

A curva ROC apresenta AUC acima de 0,70, percentual em que a literatura considera desempenho satisfatório. A Figura 4.7 mostra que a curva movimentou-se para cima e para a direita, e isto significa um critério para identificar os pontos positivos (maior sensibilidade e menor especificidade).

Este experimento apresentou um algoritmo que demonstra uma técnica não invasiva, baseada em abordagens clássicas como MLP, PCA e padronização de dados, que se mostrou promissor para prever a codeleção cromossômica 1p/19q através de análise de características radiômicas extraídas de imagens de RM com ponderação em T2.

Na próxima seção, será apresentado um novo experimento explorando novas técnicas de pré-processamento e novos algoritmos de classificação.

4.8 Experimento II

4.8.1 Objetivos

Com o constante crescimento da radiômica, o segundo experimento desta tese de doutorado continua tendo como foco a análise das características radiômicas em tumores de baixo grau para classificação da codeleção cromossômica 1p/19q através de métodos e técnicas computacionais, extraindo atributos quantitativos das imagens de RM. Assim como no primeiro experimento, a proposta desse desenvolvimento é que seus resultados possam ser bem definidos e cuidadosamente avaliados para fornecer aos profissionais da saúde, auxílio ao diagnóstico por imagem, prognóstico e respostas ao tratamento de terapias.

O objetivo desse experimento foi avaliar 2 algoritmos bastante conhecidos na literatura, a regressão logística (do inglês, *logistic regression* ou RL) e a floresta aleatória (do inglês, *random forest*, ou RF) que tem apresentado resultados significantes. Ambos classificadores possuem abordagem supervisionada. Nesse experimento foi avaliado a técnica de seleção de características por importância, redimensionamento dos dados, utilização de nova técnica de aumento de dados e um subconjunto das classes predictoras com 100 características. A seleção de características e o redimensionamento das mesmas por PCA foram executadas por ferramentas implementadas na biblioteca `sklearn`, enquanto que o aumento de instâncias e a classificação foram executadas na ferramenta MLF.

4.8.2 Métodos

Após a extração de características (seção 4.2), realizou-se as atividades a seguir para alcançar os resultados esperados.

Seleção de características

Nesse experimento, optou-se por seguir as orientações dos autores de [150]. Segundo eles, imagens de 9 instâncias do *dataset* LGG-1p19qDeletion poderiam ser pós-operatório e pareciam não ter utilizado o mesmo protocolo de aquisição das imagens utilizado nas demais 150 instâncias. Por isso, essas 9 instâncias foram excluídas neste experimento.

Neste experimento, a primeira ação realizada foi de utilizar uma técnica de seleção de características para obter as melhores previsoras candidatas a um resultado de classificação promissor. Dentre as técnicas testadas, nesse experimento utilizou-se a importância de características (*Feature Importances*). Trata-se de uma técnica recorrente no dia-a-dia de cientistas de dados, que é de execução rápida e tem boa performance em modelos de ML baseados em árvores [151]. Para a respectiva seleção foi utilizado a biblioteca YellowBrick, de código aberto em linguagem de programação Python, permitindo uma análise visual. As Figuras 4.8 e 4.9, mostram a identificação das características utilizadas na próxima etapa. Enquanto a Figura 4.8 apresenta os valores de correlação positiva identificada pelo algoritmo YellowBrick, a Figura 4.9 demonstra a ordem das características que corresponde ao eixo x no gráfico da Figura 4.8.

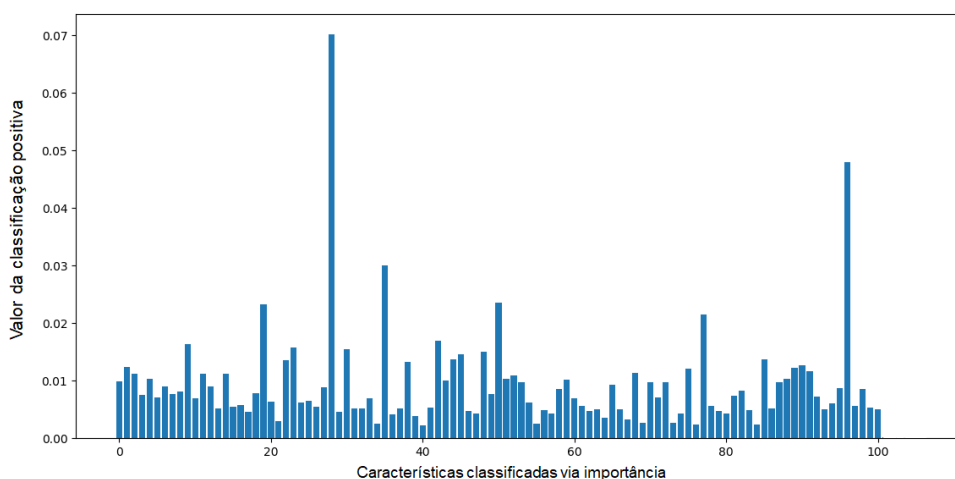


Figura 4.8: Gráfico com resultado da seleção de características com correlação positiva por meio da técnica de importância de características usando a biblioteca YellowBrick.

A Tabela 4.6 sumariza a quantidade de características extraídas com correlação positiva, conforme Figura 4.9.

Características - Original	Ordem	Características - Original	Ordem
original_shape_Elongation	1	original_gldm_MaximumProbability	51
original_shape_Fitness	2	original_gldm_SumAverage	52
original_shape_LeastAxisLength	3	original_gldm_SumEntropy	53
original_shape_MajorAxisLength	4	original_gldm_SumSquares	54
original_shape_Maximum2DDiameterColumn	5	original_gldm_GrayLevelNonUniformity	55
original_shape_Maximum2DDiameterRow	6	original_gldm_GrayLevelNonUniformityNormalized	56
original_shape_Maximum2DDiameterSlice	7	original_gldm_GrayLevelVariance	57
original_shape_Maximum3DDiameter	8	original_gldm_HighGrayLevelRunEmphasis	58
original_shape_MeshVolume	9	original_gldm_LongRunEmphasis	59
original_shape_Sphericity	10	original_gldm_LongRunHighGrayLevelEmphasis	60
original_shape_SurfaceArea	11	original_gldm_LongRunLowGrayLevelEmphasis	61
original_shape_VoxelVolume	12	original_gldm_LowGrayLevelRunEmphasis	62
original_firstorder_10Percentile	13	original_gldm_RunEntropy	63
original_firstorder_30Percentile	14	original_gldm_RunLengthNonUniformity	64
original_firstorder_Energy	15	original_gldm_RunLengthNonUniformityNormalized	65
original_firstorder_Entropy	16	original_gldm_RunPercentage	66
original_firstorder_InterquartileRange	17	original_gldm_RunVariance	67
original_firstorder_Kurtosis	18	original_gldm_ShortRunEmphasis	68
original_firstorder_Maximum	19	original_gldm_ShortRunHighGrayLevelEmphasis	69
original_firstorder_MeanAbsoluteDeviation	20	original_gldm_ShortRunLowGrayLevelEmphasis	70
original_firstorder_Mean	21	original_glszm_GrayLevelNonUniformity	71
original_firstorder_Median	22	original_glszm_GrayLevelNonUniformityNormalized	72
original_firstorder_Minimum	23	original_glszm_GrayLevelVariance	73
original_firstorder_Range	24	original_glszm_HighGrayLevelZoneEmphasis	74
original_firstorder_RobustMeanAbsoluteDeviation	25	original_glszm_LargeAreaEmphasis	75
original_firstorder_RootMeanSquared	26	original_glszm_LargeAreaHighGrayLevelEmphasis	76
original_firstorder_Skewness	27	original_glszm_LargeAreaLowGrayLevelEmphasis	77
original_firstorder_TotalEnergy	28	original_glszm_LowGrayLevelZoneEmphasis	78
original_firstorder_Uniformity	29	original_glszm_SizeZoneNonUniformity	79
original_firstorder_Variance	30	original_glszm_SizeZoneNonUniformityNormalized	80
original_gldm_Autocorrelation	31	original_glszm_SmallAreaEmphasis	81
original_gldm_JointAverage	32	original_glszm_SmallAreaHighGrayLevelEmphasis	82
original_gldm_ClusterProminence	33	original_glszm_SmallAreaLowGrayLevelEmphasis	83
original_gldm_ClusterShade	34	original_glszm_ZoneEntropy	84
original_gldm_ClusterTendency	35	original_glszm_ZonePercentage	85
original_gldm_Contrast	36	original_glszm_ZoneVariance	86
original_gldm_Correlation	37	original_gldm_DependenceEntropy	87
original_gldm_DifferenceAverage	38	original_gldm_DependenceNonUniformity	88
original_gldm_DifferenceEntropy	39	original_gldm_DependenceNonUniformityNormalized	89
original_gldm_DifferenceVariance	40	original_gldm_DependenceVariance	90
original_gldm_JointEnergy	41	original_gldm_GrayLevelNonUniformity	91
original_gldm_JointEntropy	42	original_gldm_GrayLevelVariance	92
original_gldm_MCC	43	original_gldm_HighGrayLevelEmphasis	93
original_gldm_lmcl	44	original_gldm_LargeDependenceEmphasis	94
original_gldm_lmcl2	45	original_gldm_LargeDependenceHighGrayLevelEmphasis	95
original_gldm_ldm	46	original_gldm_LargeDependenceLowGrayLevelEmphasis	96
original_gldm_ldmn	47	original_gldm_LowGrayLevelEmphasis	97
original_gldm_ld	48	original_gldm_SmallDependenceEmphasis	98
original_gldm_ldn	49	original_gldm_SmallDependenceHighGrayLevelEmphasis	99
original_gldm_InverseVariance	50	original_gldm_SmallDependenceLowGrayLevelEmphasis	100

Figura 4.9: Relação das características selecionadas na Figura 4.8 e sua ordem.

Tabela 4.6: Quantitativo de características radiômicas extraídas por classe após técnica de seleção de características por meio da biblioteca YellowBrick.

Classe de características	Quantidade de características
estatísticas de primeira ordem	18
forma	14
GLCM	22
GLRLM	16
GLSZM	16
GLDM	14

Redimensionamento das características

Por ter-se mostrado uma técnica promissora no experimento 1, e ter sido avaliado novamente nesse experimento, aplicou-se novamente a técnica PCA por meio da biblioteca `sklearn.decomposition`. Porém, neste experimento temos inicialmente 100 dimensões (vide Tabela 4.6, projetando-as em um espaço de 8 dimensões (componentes)). Ao aplicar a PCA neste conjunto de 100 características, o percentual total de redimensionamento foi

de 92%, exatamente o mesmo valor alcançado no experimento apresentado anteriormente.

Divisão do *dataset* em treino e teste

Inicialmente, o conjunto de dados com 150 instâncias (54 não deletadas e 96 codeletadas) foi dividido aleatoriamente em dois subconjuntos: 80% para treinamento (120 instâncias) e 20% para teste (30 instâncias).

Aumento das instâncias

Várias técnicas de aumento de instâncias são comumente utilizadas para proporcionar maior eficiência aos algoritmos de classificação [152], tais como: transformações geométricas (*flipping*, rotação, translação, corte, zoom, dentre outras) e transformações fotométricas, que atuam na mudança do valor do pixel, como por exemplo adição de ruído, filtragem com kernel e *random erasing*.

Neste experimento foi implementado a técnica de aumento do número de instâncias (casos) através da adição de um ruído Gaussiano com média igual a 0, desvio padrão igual a 1 e potência igual a 0,2, sendo que a potência do ruído deve ser baixa para que os dados não se encontrem fora da faixa de grandeza dos demais atributos. Zhang et al. [96] utilizaram adição de ruído em 9 bases de dados e comprovaram que essa técnica deixa o aprendizado de máquina mais eficaz no aprendizado de características robustas.

Após execução dessa etapa em 3 vezes, o novo conjunto de dados para treinamento ficou com 480 instâncias.

Configuração e implementação dos algoritmos RL e RF

As técnicas de ML (traduzido e popularmente chamado de “aprendizado de máquina” no Brasil) tem evoluído e alcançado resultados importantes com relação ao reconhecimento de padrões. Nesse experimento, para avaliação dos dados e predição da codeleção 1p/19q, foram selecionados 2 modelos de algoritmos [14]: RL e RF, ambos classificadores com abordagem supervisionada.

A RL, por ser uma técnica estatística, produzirá, a partir de um conjunto de atributos previsores, um modelo que possa ser capaz de prever uma saída categórica. Essa técnica mede a relação entre a variável dependente e variáveis independentes, estimando probabilidades, usando sua função logística subjacente, conforme abaixo:

$$p = \frac{1}{1 + e^{-y}} \quad (4.7)$$

O método RF é baseado no conceito de *ensemble learning*, que é um processo de combinação de vários classificadores para resolver um problema complexo e melhorar o desempenho do modelo. Ele constrói um conjunto de árvores de decisão em vários

subconjuntos a partir dos dados fornecidos e, ao invés de depender de uma única árvore de decisão, mescla os resultados dessas árvores com base na maioria dos votos para prever um resultado de saída. A Figura 4.10 exemplifica seu funcionamento.

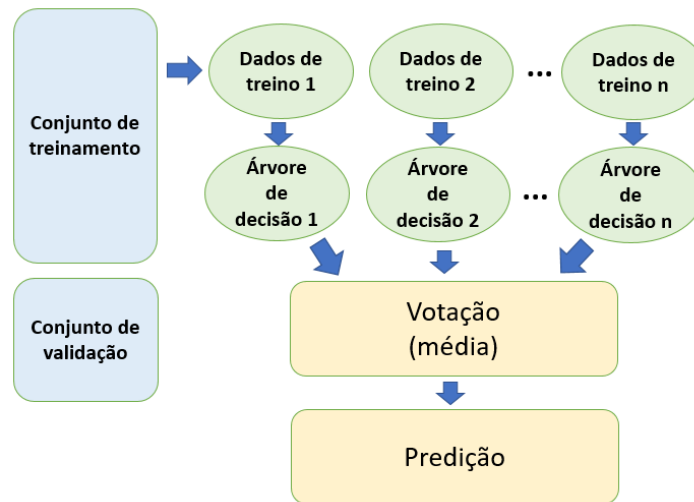


Figura 4.10: Funcionamento do algoritmo RF.

Em resumo, o conjunto de dados de treinamento passou a ter 480 instâncias, o conjunto de dados de teste 30 instâncias e ambos com 8 características (redimensionadas).

A Tabela 4.7 apresenta os parâmetros de entrada usados em cada classificador (RL e RF) implementados na MLF por meio da biblioteca `scikit-learn 0.24.2`.

Tabela 4.7: Lista de parâmetros de entrada usados na implementação dos classificadores RL e RF.

classificador	parâmetros de entrada
LR	<code>LogisticRegression(class_weight='balanced', solver='newton-cg')</code>
RF	<code>RandomForestClassifier(n_estimators=200)</code>

Cada classificador foi simulado 10 vezes, com cada nova simulação começando com uma nova divisão de dados de treinamento e teste, e mesma parametrização. Para cada classificador, as métricas acurácia, sensibilidade, especificidade e precisão foram avaliadas para cada simulação e, em seguida, calculadas as médias.

Por se tratar de um aprendizado supervisionado, ao executar o treinamento dos dados pelos algoritmos RL e RF, os mesmos vão relacionar os atributos previsoires (entrada) e os atributos categóricos (saída), conforme a divisão do *dataset* em treinamento e teste.

4.8.3 Resultados

A Tabela 4.8 apresenta o resultado do desempenho obtido pelos modelos RL e RF no conjunto de dados de teste, conforme as métricas acurácia, especificidade, sensibilidade

e precisão.

Tabela 4.8: Desempenho dos algoritmos RL e RF no conjunto de dados de teste do experimento II. O melhor resultado para cada métrica é destacado em negrito.

classificador	Acc	Se	Spe	Pre
RL	83,24	87,52	81,43	94,91
RF	84,67	88,28	83,45	95,94

Ambos os algoritmos utilizados para a classificação da codeleção cromossômica 1p19q neste experimento apresentaram bons resultados. Porém, o algoritmo RF apresentou melhor resultado que o RL em todas as métricas avaliadas.

4.8.4 Discussão

Como já citado, é um desafio a obtenção de um *dataset* com número suficiente de casos (instâncias) para treinamento e teste. Os autores de [153] indicam uma relação de 10 a 15 instâncias para cada característica radiômica extraída. Devido a proposta de trabalhar com 100 características radiômicas em uma base de 150 casos, a utilização da PCA para redução da dimensionalidade do número de atributos se mostrou interessante para obtenção de bons resultados com boa taxa de redução para as 8 novas componentes.

Os modelos RL e RF com 10 simulações foram capazes de reconhecer as características radiômicas extraídas das imagens de RM permitindo o alcance de bons resultados. Isto foi possível neste experimento usando uma técnica baseada em adição de ruído Gaussiano para aumento do conjunto de dados para treinamento. As métricas de desempenho utilizadas — acurácia, sensibilidade, especificidade e precisão — foram escolhidas por serem amplamente utilizadas em publicações de trabalhos semelhantes, permitindo, assim, uma comparação mais adequada.

As 9 instâncias excluídas do conjunto de dados deste experimento foram avaliadas por um neurorradiologista, que recomendou sua reintegração ao *dataset* original. Assim, essas 9 instâncias não foram excluídas nos experimentos III e IV, a seguir.

Novamente, o experimento desenvolvido apresenta técnicas não invasiva — baseada em análise de características radiômicas extraídas de imagens de RM com ponderação em T2, análise de importância para seleção de atributos, PCA para redução de dimensionalidade, adição de ruído Gaussiano para aumento do conjunto de dados de treinamento, e abordagens RL e RF para classificação —, que se mostraram eficazes para prever a codeleção cromossômica 1p/19q em LGGs.

Na próxima seção, será apresentado um novo experimento avaliado, explorando novas técnicas de pré-processamento e novos algoritmos de classificação.

4.9 Experimento III

4.9.1 Objetivos

Neste terceiro experimento, propõe-se uma assinatura para avaliação da codeleção cromossômica 1p/19q através de técnicas não invasivas baseadas em análise de características radiômicas extraídas de imagens de RM com ponderação em T2. Para construção da assinatura utilizou-se da técnica de hiperparâmetros, redução de dimensionalidade através da abordagem NCA e classificador KNN.

4.9.2 Métodos

Após atividades de extração das características (seção 4.2), realizou-se a normalização das características (seção 4.3) através do método `StandardScaler` da biblioteca `sklearn` da linguagem de programação Python.

Divisão do *dataset* em treino e teste

O conjunto de dados com 159 instâncias (57 não deletadas e 102 codeletadas) foi dividido aleatoriamente em dois subconjuntos: 90% para treinamento (143 instâncias) e 10% para teste (16 instâncias).

Redimensionamento das características

A análise de componentes de vizinhança (NCA) foi utilizada para a redução de dimensionalidade, que consiste em projetar os dados de características de um espaço de alta dimensionalidade para um subespaço de menor dimensionalidade. O NCA é um método para aprender uma métrica de distância quadrática que otimiza o erro de classificação no algoritmo KNN, usando uma abordagem de gradiente descendente, conforme já apresentando na seção 3.4.3 e referências [90, 84]. O NCA é construído com base exclusivamente no conjunto de treino, mas o mesmo modelo é aplicado também ao conjunto de teste.

Usando o NCA, o conjunto de recursos de 100 dimensões foi projetado em um subespaço de 4 dimensões. A Figura 4.11 mostra um mapa de calor da matriz de correlação para o conjunto de recursos de 4 dimensões obtido após a redução da dimensionalidade NCA, onde valores próximos a +1 indicam forte correlação positiva, valores próximos a -1 indicam forte correlação negativa e valores próximos a zero indicam correlação fraca. Os valores de correlação entre diferentes componentes do NCA variaram de -0,31 a +0,41. A Figura 4.12 mostra um gráfico do tipo *boxplot* dos valores de cada componente NCA para cada grupo de instâncias (azul: não deletado; laranja: codeletado). Esses *boxplots* permitem apreciar o intervalo de valores e a distribuição estatística (mediana, quartis, outliers, etc.) de cada componente, para cada grupo. Eles também nos permitem observar

que alguns componentes parecem ser mais propensos a contribuir para discriminar com sucesso entre codeletados e não deletados do que outros (compare f3 a f1, por exemplo).

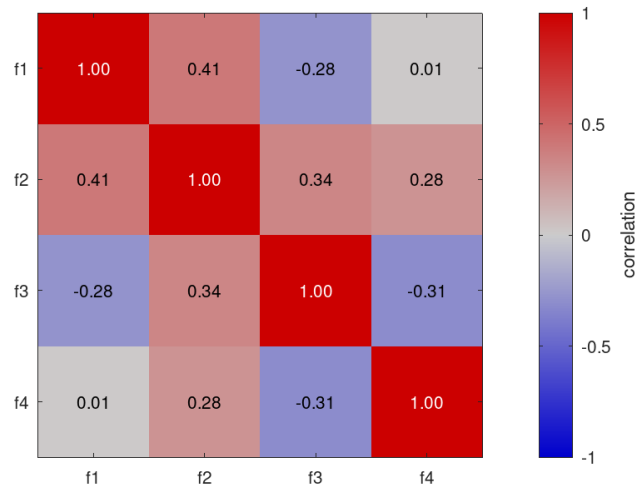


Figura 4.11: Matriz de correlação para o conjunto de dados, obtida após a redução da dimensionalidade NCA.

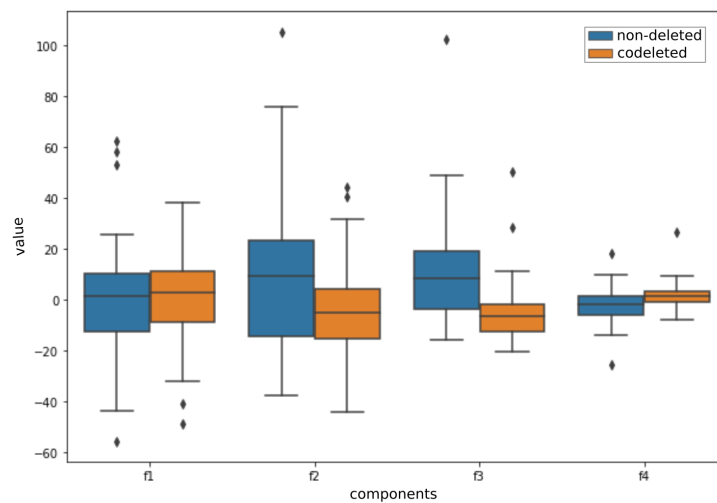


Figura 4.12: Boxplots dos valores de cada componente NCA (f1 a f4) para cada grupo de instâncias (azul: não deletado; laranja: codeletado).

Classificação

A assinatura proposta para classificação da codeleção 1p/19q usa o algoritmo KNN, um classificador supervisionado que determina o rótulo de classificação de uma amostra com base nas amostras vizinhas, em que a variável k indica o número de vizinhos mais próximos aos quais uma medida de distância será aplicada a fim de identificar a classe à qual a nova amostra pertence.

As instâncias de treinamento foram usadas para encontrar os parâmetros ótimos para o classificador KNN, ou seja, número de vizinhos mais próximos e abordagem de ponderação. Sendo assim, o processo de ajuste de hiperparâmetros foi realizado usando `grid-search` sobre uma grade de parâmetros. Esse processo avalia sistematicamente várias combinações dos parâmetros do modelo e retorna o ajuste que resulta na melhor acurácia. Para uma abordagem ponderada, avaliou-se os parâmetros `weights='uniform'` e `weights='distance'`. Para o número de vizinhos mais próximos, avaliamos k variando de 1 a 30.

Durante o ajuste dos hiperparâmetros, a avaliação de cada configuração de parâmetro foi feita usando validação cruzada estratificada com repetição de k -folds. Isso envolve particionar aleatoriamente o conjunto de dados em k subconjuntos de tamanho igual (não confundir com o k do classificador KNN, que representa o número de vizinhos mais próximos que esse classificador considera), em que cada subconjunto contém aproximadamente as mesmas proporções de cada classe de instâncias (codeletadas e não codeletadas). Um subconjunto é retido para validação, enquanto os subconjuntos restantes são usados para treinamento. Esse processo é repetido k vezes (para que cada subconjunto seja usado uma vez para validação) e os k resultados são calculados para produzir uma estimativa da precisão da classificação. Todo esse processo (partição aleatória seguida por k repetições das etapas de treinamento e validação) é repetido n vezes e a média da acurácia da classificação é calculada nessas n execuções. Usamos $k = 5$ dobras (*folds*) e $n = 10$ repetições. Observe que esta etapa faz parte do treinamento KNN e, portanto, apenas o conjunto de dados de treinamento (143 instâncias) é usado. As 16 instâncias de teste não são vistas pelo validador cruzado. Assim, o validador cruzado usou dois subconjuntos de 28 instâncias e três subconjuntos de 29 instâncias.

Finalmente, as 16 instâncias de teste são aplicadas ao classificador KNN, usando o número ótimo de vizinhos mais próximos e a abordagem de ponderação encontrada durante o ajuste de hiperparâmetros. O número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos é calculado e o desempenho é avaliado de acordo com várias métricas (seção 4.5).

Todo o processo descrito anteriormente foi simulado 30 vezes, com cada nova simulação começando com uma nova divisão aleatória do conjunto de dados (90% para treinamento e 10% para teste), seguida de ajuste de hiperparâmetros (usando repetição por validação cruzada estratificada k -fold), seguida pela classificação KNN (usando o número ideal de vizinhos mais próximos e a abordagem de ponderação encontrada durante o ajuste de hiperparâmetros). O desempenho alcançado pela assinatura em cada simulação foi avaliada e a média foi calculada. Também usou-se os resultados dessas 30 simulações para calcular intervalos de confiança para cada métrica, no nível de confiança de 95%, com base em uma distribuição t -Student.

Para avaliar a importância de cada etapa da assinatura, extraiu-se também os

resultados obtidos com versões “incompletas” da assinatura. No total, foram apreciadas 4 versões da assinatura:

1. classificador KNN com $k = 1$ vizinho mais próximo e ponderação baseada na distância;
2. normalização dos dados (padronização), seguida de classificador KNN com $k = 1$ vizinho mais próximo e ponderação baseada na distância;
3. normalização de dados (padronização), seguida de classificador KNN com ajuste de hiperparâmetros; e
4. assinatura completa, ou seja, normalização de dados (padronização), seguida de redução de dimensionalidade (NCA), seguida de classificador KNN com ajuste de hiperparâmetros.

Observe que cada versão da assinatura foi apreciada usando 30 simulações, cada simulação usando uma nova divisão aleatória do conjunto de dados (90% para treinamento e 10% para teste). No entanto, as mesmas 30 simulações foram usadas para avaliar cada versão da assinatura.

4.9.3 Resultados

A Tabela 4.9 apresenta o desempenho de classificação alcançado após a incorporação de cada etapa na execução da assinatura (média das 30 simulações). Os valores mostram que os resultados melhoram à medida que incorporamos cada etapa adicional a assinatura. Os melhores resultados são obtidos ao executar a assinatura completa (linha 4).

Tabela 4.9: Desempenho de classificação alcançado após a incorporação de cada etapa na execução da assinatura. A assinatura completa (versão 4) está destacado em negrito. Os resultados em cada linha correspondem a uma média de 30 simulações.

etapas da assinatura	Acc	Se	Spe	Pre	AUC	F1
1.KNN	56.25	67.92	35.17	67.80	56.70	66.35
2.Padronização + KNN	73.33	85.15	47.36	77.80	73.80	80.79
3.Padronização + KNN + Hiperparâmetros	76.04	80.76	66.51	83.23	73.60	81.31
4.Padronização + NCA + KNN + Hiperparâmetros	81.88	88.48	69.25	86.04	82.70	86.38

A Tabela 4.10 apresenta os resultados obtidos após execução da assinatura completa; valores médios e intervalos de confiança (IC) para cada métrica também são apresentados. Esses resultados demonstram a consistência da assinatura na obtenção de bons resultados de classificação. A Tabela 4.11 mostra uma matriz de confusão obtida pela média das matrizes de confusão obtidas em cada uma das 30 simulações da assinatura completa, o que mostra que o número médio de erros de classificação (FP e FN) é pequeno, comparado ao número de classificações corretas (TP e TN).

Tabela 4.10: Resultados da classificação individual obtidos em cada uma das 30 simulações da assinatura completo, valores médios e intervalos de confiança (ao nível de confiança de 95%).

métrica	média	intervalo de confiança
TP	9,4	9,0 – 10,9
FN	1,3	0,9 – 2,1
TN	3,7	3,2 – 5,1
FP	1,6	0,9 – 1,9
Acc	81,88	74,4 – 89,3
Se	88,48	79,7 – 97,2
Spe	69,25	55,2 – 83,3
Pre	86,04	77,9 – 94,1
AUC	82,70	77,5 – 87,9
F1	86,38	80,4 – 92,3

Tabela 4.11: Matriz de confusão obtida pela média das matrizes de confusão obtidas em cada uma das 30 simulações da assinatura completa.

casos de teste = 16	codeletado (predito)	não deletado (predito)
codeletado (real)	9,4 (<i>TP</i>)	1,3 (<i>FN</i>)
não deletado (real)	1,6 (<i>FP</i>)	3,7 (<i>TN</i>)

Métricas como sensibilidade, especificidade e precisão dependem do limiar de discriminação do classificador. Portanto, muitas vezes é útil analisar gráficos como a curva ROC (Figura 4.13) e a curva de precisão-sensibilidade (PR) (Figura 4.14). Essas ferramentas ilustram como os resultados da classificação mudam à medida que o limite de discriminação do classificador varia. A curva ROC é um gráfico da taxa de verdadeiros positivos (Se) contra a taxa de falsos positivos (1-Spe), enquanto a curva PR é um gráfico de precisão (Pre) contra a sensibilidade (Se). Essas curvas são calculadas com base em estimativas de probabilidade da classe positiva (com codeleção 1p/19q). No entanto, ao usar a assinatura completa, todas as estimativas de probabilidade em todas as 30 simulações foram 0,0 ou 1,0, de modo que a variação do limite de discriminação não alteraria os valores na matriz de confusão de cada simulação. Isso resultou em apenas dois pontos disponíveis para a construção da curva ROC e um ponto da curva PR para cada simulação [154]. Como resultado ilustrativo, apresentamos as duas curvas construídas para o resultado médio da assinatura completa, ou seja, usando $Se = 0,88$, $Spe = 0,69$ e $Pre = 0,86$. A área sob a curva ROC (Figura 4.13) foi de 0,827, que corresponde ao valor AUC de 82,7%. Esta é uma estimativa da probabilidade de rotular com precisão uma amostra positiva (com codeleção 1p/19q). Embora a curva ROC seja mais comumente usada, ela pode não ser o indicador de desempenho mais apropriado ao lidar com conjuntos de dados altamente desbalanceados; nesses casos, pode ser benéfico usar a curva PR. A área sob a curva PR (Figura 4.14) foi de 0,864, o que corresponde ao score F1 (média

harmônica entre precisão e sensibilidade) de 86,4%.

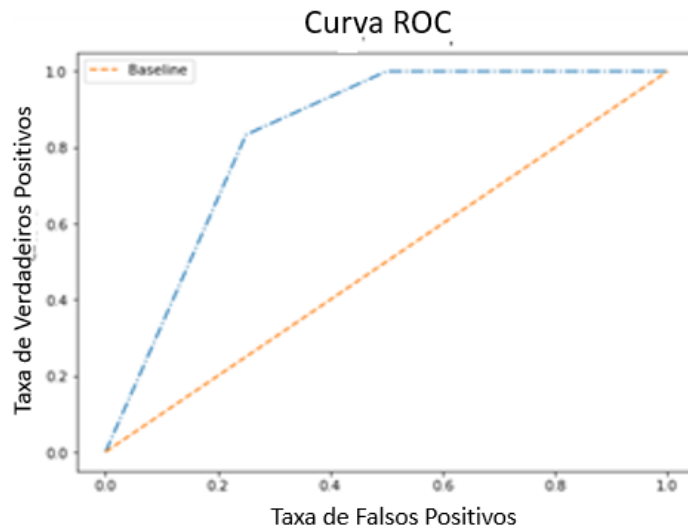


Figura 4.13: Curva ROC para o resultado médio da assinatura completa. A área sob a curva ROC foi de 0,827.

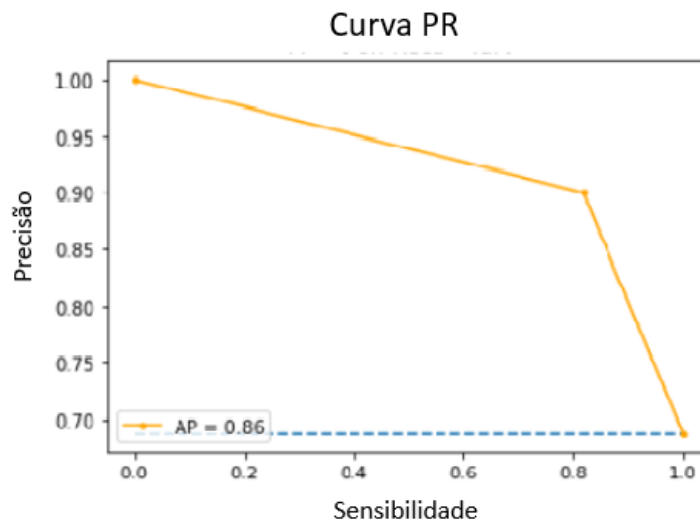


Figura 4.14: Curva PR para o resultado médio da assinatura completa. A área sob a curva PR foi de 0,864.

4.9.4 Discussão

Este experimento demonstrou mais uma vez que, por meio do uso de técnicas de aprendizado de máquina, a radiômica pode ser uma ferramenta valiosa para a identificação não invasiva da codeleção 1p/19q em gliomas de baixo grau. Resultados se mostraram promissores aos demais resultados obtidos nos experimentos anteriores.

Conforme sugestão de Li et al. [153], é importante ter uma relação de 10 a 15 instâncias por característica radiômica utilizada na etapa de classificação. Portanto, a

etapa de redução de dimensionalidade foi importante para o bom desempenho da assinatura proposta nesse experimento. De fato, os resultados da Tabela 4.9 demonstram que cada etapa contribuiu para melhorar o desempenho do classificador.

Sendo assim, a assinatura proposta neste experimento usou a técnica de redução de dimensionalidade NCA, que se mostrou eficiente na redução de dimensionalidade quando aplicada em conjunto com o algoritmo KNN. Ambas trabalham olhando para o vizinho mais próximo. Em conjunto com essas duas abordagens, a busca pelos melhores parâmetros do KNN em cada simulação, para obtenção de um modelo eficaz quando executado sobre o conjunto de dados de teste, foi determinante para os bons resultados obtidos. Há uma percepção aqui de que, com um número maior de instâncias no treinamento e teste, os resultados poderiam ser melhores.

A etapa completa da assinatura apresentou acurácia e sensibilidade compatível com os demais resultados alcançados nos experimentos anteriores, 81,8% e 88,4%, respectivamente. A Tabela 4.11 mostrou que o número médio de erros de classificação (FP e FN) é pequeno, comparado ao número de classificações corretas (TP e TN).

Em conclusão, apresentou-se uma assinatura para classificação não invasiva da codeleção 1p/19q em gliomas de baixo grau com base em radiômica e aprendizado de máquina. O método proposto obteve resultados promissores se comparado aos métodos de ponta da literatura e resultados consideravelmente melhores do que os obtidos por neurocirurgiões e neurorradiologistas analisando as mesmas imagens.

4.10 Experimento IV

4.10.1 Objetivos

Nesse último experimento, continua-se a avaliação da codeleção cromossômica 1p/19q através de técnicas não invasiva baseada em análise de características radiômicas extraídas de imagens de RM com ponderação em T2. Avaliou-se 2 subconjuntos de características radiômicas: com 68 (Tabela 4.12) e 100 (Tabela 4.6) atributos. Após normalização das características, experimentou-se vários classificadores, incluindo SVM, RF, KNN, ET, GB, AB, SGD, HGB, XGB, LR, DT, GNB e LDA. Também experimentou-se métodos *ensemble*, incluindo classificadores baseados em *boosting* e *bagging*.

4.10.2 Métodos

Divisão do *dataset* em treino e teste

Em ambos os subconjuntos de 68 e 100 características, manteve-se as 159 instâncias (57 não deletadas e 102 codeletadas) e dividiu-se aleatoriamente em dois subconjuntos: 90% para treinamento (143 instâncias) e 10% para teste (16 instâncias).

Classificadores

Experimentou-se vários classificadores [155] para a classificação da coleção 1p/19q:

- support vector machine (SVM) classifier [114, 115];
- random forest (RF) classifier [112];
- k-nearest neighbors (KNN) classifier [113];
- extra-trees (ET) classifier [125, 126];
- gradient boosting (GB) classifier [127];
- adaptive boosting (AB) classifier [120];
- stochastic gradient descent (SGD) classifier [121, 122];
- histogram-based gradient boosting (HGB) classification tree classifier [123];
- extreme gradient boosting (XGB) classifier [124];
- logistic regression (LR) classifier [111];
- decision tree (DT) classifier [117, 118];
- Gaussian naive Bayes (GNB) classifier [119]; and
- linear discriminant analysis (LDA) classifier [128].

Uma lista de parâmetros de entrada usados para implementar cada classificador é mostrada na Tabela 4.13. Os experimentos com os classificadores acima foram realizados usando implementações da biblioteca `scikit-learn` 0.24.2 (`sklearn`) [144], exceto para o classificador XGB, que foi implementado usando a biblioteca `XGBoost` 1.7.1 (`xgboost`) [124].

Também foi avaliado o *bagging* [116], um meta-estimador de *ensemble* no qual subconjuntos aleatórios do conjunto de dados de treinamento são usados para ajustar várias instâncias de um estimador de base e, em seguida, a previsão final é obtida agregando as previsões individuais de cada instância do estimador (Figura 3.16). O estimador de base padrão na implementação `scikit-learn` do *bagging* é um classificador de árvore de decisão (DT), mas, assim como avaliado nesse experimento, o usuário pode especificar um classificador diferente como estimador de base. Neste experimento utilizou-se o *bagging* usando cada um dos outros métodos listados na Tabela 4.13 como estimador de base através da classe `BaggingClassifier` (BC) da biblioteca `sklearn` com parâmetros de entrada padrão, exceto por `n_estimators=200`.

Comparando características radiômicas diferentes

Dois tumores diferentes podem ter tamanho, forma e/ou distribuição de nível de cinza de pixel semelhantes (estes seriam quantificados por características estatísticas baseados em forma e de primeira ordem), mas podem diferir na interrelação espacial de intensidades de pixel (que seriam quantificados por características baseadas em textura). A análise de textura pode ser especialmente útil quando se trata de tumores marcadamente heterogêneos [156]. Recursos quantitativos de textura extraídos de imagens com ponderação em T2 demonstraram prever a codeleção 1p/19q com alta sensibilidade [157].

Assim, experimentou-se conjuntos de 68 (Tabela 4.12) e 100 (Tabela 4.6) características radiômicas. Estatísticas de primeira ordem e recursos de forma não foram usados ao trabalhar com apenas 68 características, ou seja, apenas recursos baseados em textura foram usados. A média de cada métrica para cada classificador, obtida com 68 atributos, foi comparada com a obtida com 100 atributos e os resultados serão apresentadas a seguir na seção 4.9.3.

Tabela 4.12: Quantitativo de características radiômicas de textura.

Classe de características	Quantidade de características
GLCM	22
GLRLM	16
GLSZM	16
GLDM	14

Também experimentou-se o *bagging* usando cada um dos classificadores listados na Tabela 4.13 como estimador de base e, em seguida, comparamos cada métrica para cada classificador com aquela obtida sem *bagging*. Isso foi feito porque o *bagging* padrão (baseado em DT) mostrou resultados muito promissores nos experimentos iniciais, então decidiu-se investigar se essa abordagem *ensemble* poderia melhorar o desempenho de outros classificadores. Em cada uma dessas comparações, realizou-se o um t-test de duas amostras usando o GNU Octave 7.2.0. Quando $p < 0,05$, a hipótese nula de médias iguais foi rejeitada. Testes de normalidade não foram realizados, pois nosso tamanho amostral ($n = 100$) é grande [158].

4.10.3 Resultados

Cada classificador foi simulado 100 vezes, com cada nova simulação começando com uma nova divisão de dados de treino e teste, seguida pela normalização das características. Para cada classificador, as métricas listadas na seção 4.5 foram avaliadas para cada simulação e, em seguida, calculou-se as médias. Intervalos de confiança para cada métrica e para cada classificador também foram calculados — no nível de confiança de 95%, com base em uma distribuição t de Student — a partir dos resultados dessas 100 simulações.

Tabela 4.13: Lista de parâmetros de entrada usados na implementação de cada classificador.

classificador	parâmetros de entrada
SVM	<code>SVC(kernel='linear', probability=True)</code>
RF	<code>RandomForestClassifier(n_estimators=200)</code>
KNN	<code>KNeighborsClassifier(n_neighbors=3, weights='uniform')</code>
ET	<code>ExtraTreesClassifier(n_estimators=200)</code>
GB	<code>GradientBoostingClassifier(n_estimators=200, max_depth=1)</code>
AB	<code>AdaBoostClassifier(n_estimators=200)</code>
SGD	<code>CalibratedClassifierCV(base_estimator=SGDClassifier(penalty='elasticnet', loss='log', learning_rate='optimal', eta0=1, alpha= 0.1))</code>
HGB	<code>HistGradientBoostingClassifier(max_bins=255, max_iter=100)</code>
XGB	<code>XGBClassifier(n_estimators=200, eval_metric='logloss')</code>
LR	<code>LogisticRegression(class_weight='balanced', solver='newton-cg')</code>
DT	<code>DecisionTreeClassifier(max_features=30)</code>
GNB	<code>GaussianNB()</code>
LDA	<code>LinearDiscriminantAnalysis()</code>

Nota: O classificador `XGBClassifier` é da biblioteca `xgboost`; todos os outros classificadores fazem parte da biblioteca `sklearn`; parâmetros de entrada padrão foram usados onde não especificado acima.

A Tabela 4.14 apresenta os resultados de desempenho obtidos com cada um dos classificadores listados na Tabela 4.13, usando 68 características radiômicas baseadas em textura (características associadas com estatísticas de primeira ordem e forma não foram usado neste experimento; vide Tabela 4.12). Cada entrada corresponde a uma média de 100 simulações e o melhor resultado para cada métrica é destacado em negrito. Nesta primeira avaliação, SVM e RF apresentaram os melhores resultados gerais, apresentando desempenhos similares. O SVM apresentou o maior BCA (77,3%), enquanto o RF apresentou a maior acurácia (80,3%) e AUC (84,6%). Esses classificadores também apresentaram sensibilidade, especificidade, precisão e pontuação F1 competitivas. O classificador SGD apresentou maior sensibilidade (95,7%) do que SVM e RF, mas forneceu especificidade muito baixa (51,5%). Isso porque o SGD apresentou poucos falsos negativos (instâncias codeletadas classificadas como não deletadas), mas muitos falsos positivos (instâncias não deletadas classificadas como codeletadas). O classificador LR apresentou a maior especificidade (66,5%) e a menor sensibilidade (75,7%), pois apresentou menos falsos positivos, porém mais falsos negativos.

A Tabela 4.15 apresenta os resultados obtidos com os mesmos classificadores, mas usando 100 características radiômicas (ou seja, características associadas com estatísticas de primeira ordem e forma não foram excluídas). Diferenças estatisticamente significativas foram encontradas para algumas métricas e alguns classificadores. O uso de características radiômicas adicionais resultou em pior desempenho (média significativamente menor para pelo menos uma métrica, conforme indicado com *) para a maioria dos classificadores: SVM, RF, KNN, ET, GB, AB, SGD e LDA. Não observou-se nenhuma diferença signi-

Tabela 4.14: Desempenho de classificação usando 68 características radiômicas baseadas em textura. Cada entrada corresponde a uma média de 100 simulações. O melhor resultado para cada métrica é destacado em negrito.

classificador	BCA	Acc	Se	Spe	Pre	AUC	F1
SVM	77,3	79,8	91,0	63,5	79,5	80,0	84,2
RF	76,0	80,3	90,9	61,1	81,2	84,6	85,2
KNN	75,3	79,3	91,4	59,1	78,7	77,8	84,0
ET	74,8	79,1	90,1	59,6	79,5	84,2	83,9
GB	74,8	79,2	90,8	58,8	79,9	82,9	84,4
AB	74,5	79,4	88,4	60,5	81,4	82,0	84,2
SGD	73,6	80,3	95,7	51,5	78,9	79,3	86,0
HGB	72,9	76,4	84,6	61,1	79,8	79,9	81,4
XGB	72,5	76,0	85,9	59,0	78,1	77,7	81,1
LR	71,1	71,6	75,7	66,5	78,3	78,8	76,1
DT	68,8	71,4	77,5	60,0	78,8	68,8	77,2
GNB	67,9	72,3	85,8	49,9	74,5	70,9	79,0
LDA	61,9	66,2	77,3	46,5	72,6	64,3	73,8

ficativa no desempenho destes classificadores: HGB, XGB e DT. Observou-se resultados melhores (indicados com **) para apenas dois classificadores: LR, para o qual a pontuação F1 melhorou significativamente de 76,1% para 79,5%, e GNB, para o qual a AUC melhorou significativamente de 70,9% para 75,7%. Nota-se que o GNB ainda apresenta desempenho geral muito ruim em comparação com os outros classificadores da Tabela 4.14 e 4.15, enquanto LR ainda apresenta alta especificidade, mas baixa sensibilidade. Com base nesses resultados, prosseguiu-se as análises com o conjunto de 68 características, em vez do conjunto de 100 características.

A Tabela 4.16 apresenta os resultados obtidos com os mesmos classificadores, mas desta vez usando cada classificador como estimador base do classificador *bagging*. O conjunto de 68 características foi usado. O uso do *bagging* melhorou significativamente todas as métricas (indicadas com **) para os classificadores XGB e DT, enquanto piorou significativamente todas as métricas (indicadas com *) para o classificador KNN. *Bagging* também melhorou significativamente a sensibilidade do SVM (de 91,0% para 96,3%), mas ao custo de uma especificidade significativamente menor (de 63,5% para 39,1%). A HGB apresentou melhora importante na sensibilidade (de 84,6% para 90,2%), o que resultou em melhora significativa no BCA, na acurácia e na pontuação F1. Não observa-se nenhuma diferença significativa no desempenho destes classificadores: RF, ET, GB e GNB. Mais notavelmente, *bagging* com DT apresentou o maior BCA (79,3%), acurácia (81,9%) e especificidade (67,8%) de todos os nossos testes; nossos segundos maiores valores foram de precisão (82,6%), AUC (84,2%) e pontuação F1 (85,9%); e a sensibilidade (90,9%) altamente promissor. Considerou-se o *bagging* baseado em DT o melhor classificador geral. Pode-se notar que o *bagging* com XGB e o *bagging* com HGB também apresentaram

Tabela 4.15: Desempenho de classificação usando 100 características radiômicas (textura + estatísticas de primeira ordem + forma). Cada entrada corresponde a uma média de 100 simulações. O melhor resultado para cada métrica é destacado em negrito.

classificador	BCA	Acc	Se	Spe	Pre	AUC	F1
SVM	71,8*	75,6*	85,3*	58,2	79,2	73,5*	81,3*
RF	74,6	78,1	88,8	60,5	79,9	81,0*	83,4
KNN	64,2*	69,4*	82,9*	45,6*	74,3*	70,3*	77,2*
ET	73,5	77,7	89,9	57,1	78,4	80,3*	83,1
GB	74,7	77,6	85,7*	63,6	80,3	80,9	82,2
AB	70,4*	74,1*	82,7*	58,1	78,5	74,7*	79,6*
SGD	72,8	78,5	93,3*	52,3	77,7	79,3	84,2
HGB	71,4	75,8	85,8	57,1	79,3	77,5	81,7
XGB	73,2	76,7	86,8	59,6	79,8	81,1	82,4
LR	72,4	74,6	78,8	66,0	81,3	77,9	79,5**
DT	67,6	69,9	76,0	59,2	78,1	67,3	76,0
GNB	69,0	73,6	84,9	53,0	76,4	75,7**	79,7
LDA	61,8	64,5	71,2*	52,4	73,0	61,5	71,0

* $p < 0,05$ e a média é menor do que com 68 características.

** $p < 0,05$ e a média é maior do que com 68 características.

excelentes resultados.

Tabela 4.16: Desempenho de classificação usando cada classificador como estimador base do classificador *bagging*. O subconjunto de 68 características radiômicas foi usado. Cada entrada corresponde a uma média de 100 simulações. O melhor resultado para cada métrica é destacado em negrito.

estimador base	BCA	Acc	Se	Spe	Pre	AUC	F1
SVM	67,7*	76,1*	96,3**	39,1*	74,4*	78,0	83,4
RF	75,9	79,2	90,1	61,8	79,8	81,7	83,9
KNN	61,4*	66,1*	76,0*	46,7*	71,9*	68,6*	73,2*
ET	72,4	78,3	89,6	55,2	80,8	82,5	84,4
GB	76,9	81,2	91,0	62,8	81,8	83,3	85,6
AB	76,8	80,3	92,5**	61,1	79,8	82,9	85,0
SGD	76,1	79,4	87,8*	64,3**	81,5	80,7	83,8
HGB	77,9**	80,9**	90,2**	65,6	81,9	83,1	85,3**
XGB	78,5**	81,6**	90,3**	66,7**	83,0**	82,8**	85,8**
LR	69,5	71,9	79,9**	59,2*	76,7	76,2	77,4
DT	79,3**	81,9**	90,9**	67,8**	82,6**	84,2**	85,9**
GNB	68,4	73,8	84,0	52,8	77,4	72,5	79,9
LDA	61,7	65,6	78,3	45,2	71,2	71,5**	73,4

* $p < 0,05$ e a média é menor do que sem *bagging*.

** $p < 0,05$ e a média é maior do que sem *bagging*.

A Tabela 4.17 resume os resultados que obteve-se com o melhor classificador geral: método *bagging* com um estimador de base de árvore de decisão (DT), usando um conjunto de 68 características radiômicas. Média, desvio padrão e intervalos de confiança

de 95% são mostrados para os valores da matriz de confusão e para todas as métricas de classificação. Os valores da matriz de confusão mostram que o número médio de erros de classificação (FN e FP) é pequeno, comparado ao número de classificações corretas (TP e TN). Particularmente, o número de falsos negativos é muito pequeno em comparação com o número de verdadeiros positivos. Isso garante alta sensibilidade (90,9%). O número de falsos positivos não é tão baixo quando comparado com o número de verdadeiros negativos, o que deixa baixa a especificidade. Ainda assim, o melhor resultado de especificidade (67,8%) foi obtido usando *bagging* baseado em DT.

Tabela 4.17: Resumo dos resultados obtidos com o melhor classificador geral: método *bagging* com um estimador de base de árvore de decisão (DT), usando um conjunto de 68 características radiômicas. Média, desvio padrão e intervalos de confiança de 95% são mostrados para os valores da matriz de confusão e para todas as métricas de classificação.

	média ± desvio padrão	intervalo de confiança
TP	9,32 ± 1,86	8,96 – 9,68
FN	0,94 ± 0,93	0,76 – 1,12
TN	3,78 ± 1,35	3,51 – 4,05
FP	1,96 ± 1,36	1,69 – 2,23
BCA	79,3 ± 11,3	77,1 – 81,5
Acc	81,9 ± 9,9	79,9 – 83,8
Se	90,9 ± 8,9	89,2 – 92,7
Spe	67,8 ± 20,0	63,8 – 71,7
Pre	82,6 ± 12,0	80,2 – 84,9
AUC	84,2 ± 10,0	82,2 – 86,2
F1	85,9 ± 8,3	84,3 – 87,6

A Figura 4.15 apresenta a curva ROC obtida pela média das 100 simulações do classificador *bagging* baseado em DT. Métricas como sensibilidade, especificidade e precisão dependem do limite de discriminação do classificador, portanto, muitas vezes é útil analisar a curva ROC, que ilustra como os resultados da classificação mudam à medida que o limite de discriminação do classificador varia. A curva ROC é um gráfico da taxa de verdadeiros positivos (Se) contra a taxa de falsos positivos ($1 - \text{Spe}$). A curva ROC é calculada com base nas estimativas de probabilidade da classe positiva (identificação de codeleção 1p/19q). A área sob a curva ROC média foi de 0,842, o que corresponde ao valor médio de AUC de 84,2%. Esta é uma estimativa da probabilidade de rotular com precisão uma amostra positiva.

4.10.4 Discussão

Este experimento mais uma vez avaliou, através do aprendizado de máquina, o potencial das características radiômicas baseadas em estatísticas de primeira ordem, forma e textura na classificação da codeleção cromossômica 1p/19q em gliomas de baixo grau. As

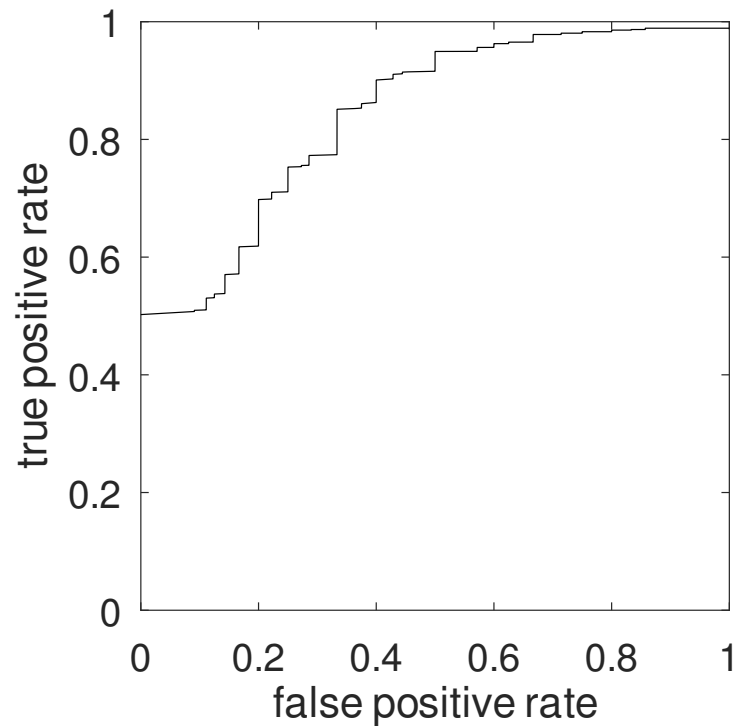


Figura 4.15: Curva ROC obtida pela média das 100 simulações do classificador *bagging* baseado em DT. A área sob a curva ROC foi de 0,842.

características baseadas na forma quantificam a forma e o tamanho do tumor; características estatísticas de primeira ordem quantificam a distribuição de intensidade do pixel do tumor (ou histograma); e características baseadas em textura (associadas a estatísticas de segunda e ordem superior) quantificam interrelações de pixel, como relações espaciais e co-ocorrência dos valores dos pixels [156]. A biblioteca PyRadiomics novamente, se mostrou adequada. Demonstrou-se resultados de classificação promissores usando exclusivamente características radiômicas baseadas em textura extraídas de imagens de ressonância magnética com ponderação em T2. A textura correlaciona-se potencialmente com a densidade celular, angiogênese e necrose; portanto, a análise de textura pode ser usada para medir a heterogeneidade do tumor e prever a biologia do tumor [156]. Por exemplo, os oligodendrogliomas geralmente apresentam calcificações e sangramento, enquanto os astrocitomas tendem a ser mais homogêneos. Assim, a heterogeneidade é um aspecto visual importante que os especialistas avaliam nas imagens de ressonância magnética ao tentar diferenciar esses tipos de tumores. A análise de textura demonstrou prever a codeleção 1p/19q com alta sensibilidade e especificidade [157]. Este experimento mostrou que, não apenas características baseadas em textura são suficientes para a classificação da codeleção 1p/19q, mas incluir características radiômicas associadas a estatísticas de primeira ordem e forma não melhoram o desempenho dos classificadores e, em muitos casos, o piora.

Métodos *ensemble*, como *bagging* e *boosting*, demonstraram fornecer melhor acurácia do que classificadores de base única em várias aplicações [159, 160, 161, 162]. Os

estimadores de base única geralmente têm um viés alto ou muita variância. Os métodos de ensemble tentam reduzir o viés e/ou a variância combinando vários algoritmos “aprendizes” [163, 164, 129]. Na abordagem *bagging*, vários algoritmos aprendizes homogêneos são treinados independentemente, em paralelo (conforme ilustrado na Figura 3.16); enquanto que, no *boosting*, os algoritmos aprendizes homogêneos são treinados sequencialmente. Experimentou-se vários classificadores baseados em *boosting* (GB, AB, HGB e XGB), mas os melhores resultados foram obtidos usando classificadores baseados em *bagging*. Curiosamente, o uso da abordagem *bagging* melhorou o desempenho de alguns dos classificadores *boosting* (ver Tabela 4.16). O melhor resultado, no entanto, foi obtido com um classificador *bagging* usando o algoritmo DT como estimador base.

Ao avaliar a codeleção 1p/19q, é mais importante evitar falsos negativos (diagnosticar incorretamente um tumor codeletado) do que evitar falsos positivos (diagnosticar incorretamente um tumor não deletado); ou seja, alta sensibilidade é mais crítica do que alta especificidade. O diagnóstico de codeleção positiva é importante porque os pacientes com tumores codificados têm melhor prognóstico e maior sobrevida, pois esses tumores respondem bem à quimioterapia dedicada (à base de procarbaniza, lomustina e vincristina) e à radioterapia [165]. O classificador *bagging* baseado em DT alcançou maior sensibilidade (90,9%) do que a maioria dos métodos da literatura. Alguns dos classificadores testados alcançaram sensibilidade ainda maior (até 96,3%), ao custo de menor especificidade. Embora nenhum dos classificadores avaliados tenha alcançado especificidade superior a 67,8%, acredita-se que isso pode ser devido ao pequeno tamanho do conjunto de dados LGG-1p19qDeletion, particularmente ao pequeno número de instâncias não excluídas. Usar um conjunto de dados desequilibrado costuma ser um problema, pois os padrões de dados da classe majoritária podem se destacar daqueles da classe minoritária. Um modelo treinado com um conjunto de dados desequilibrado pode projetar maior acurácia na classe majoritária e, possivelmente, exibir desempenho de classificação ruim para a classe minoritária [101]. Com o objetivo de resolver esse problema, experimentou-se técnicas de sobreamostragem de minorias, como o algoritmo sintético adaptativo [166] e a técnica de sobreamostragem de minorias sintéticas [100], mas isso não melhorou os resultados da classificação. Também tentou-se igualar o número de instâncias de cada classe no conjunto de treinamento reduzindo o número de instâncias da classe majoritária [167], mas isso também se mostrou infrutífero. Experimentou-se com técnicas de redução de dimensionalidade [168], também não apresentou melhora no desempenho da classificação. Li et al. [153] sugerem uma relação de 10 a 15 instâncias por característica radiômica utilizada na etapa de classificação. Assim, acredita-se fortemente que os classificadores testados poderiam ter obtido melhor desempenho se um conjunto de dados maior estivesse disponível.

4.11 Comparação com estado da arte

Todos as referências da Tabela 4.18 apresentam estudos com no máximo 5 anos de publicação, período de tempo importante quando se pretende realizar uma comparação e citação de trabalhos correlacionados. Dentre as referências citadas na Tabela 4.18, os autores de [169, 170] utilizaram dataset próprio e os autores de [171] utilizaram outro dataset público. Os demais utilizaram o dataset LGG-1p19qDeletion.

A Tabela 4.18 apresenta uma comparação entre os experimentos implementados nesta tese e vários métodos da literatura. Eles representam o estado da arte na classificação da codeleção 1p/19q com base em radiômica e aprendizado de máquina.

Tabela 4.18: Comparação com estudos atualizados para classificação da codeleção 1p/19q baseados em radiômica e aprendizado de máquina.

referência	dataset	algoritmo	BCA	Acc	Se	Spe	Pre	AUC	F1
[172]	LGG-1p19qDeletion	neurosurgeons*		48,9	41,5	63,8		51,5	
		neuroradiologists*		63,2	53,5	81,8		81,1	
	Own - EMC/HMC (train) LGG-1p19qDeletion (test)	SVM		69,3	73,2	61,7	78,7	72,3	69,7
[173]	LGG-1p19qDeletion	quadratic SVM		80,0	66,7	100,0			
[169]	Own	RF		70,0	68,3	71,2		76,0	
[150]	LGG-1p19qDeletion	multilayer perceptron		83,8	87,5	75,8	90,5	86,9	88,3
[170]	Own - PUMCH	RF		83,3	100,0	75,0		88,9	
[171]	TCGA-LGG	XGB			75,0	85,0		80,0	
[42]	LGG-1p19qDeletion TCGA-LGG BraTS 2019	CNN		83,3	86,7	79,2		82,1	
	Own - GUH			75,0	58,3	82,1		86,6	
experimento I	LGG-1p19qDeletion	MLP			88,0	73,5	85,0		
experimento II		RL		83,2	87,5	81,4	94,9		
		RF		84,6	88,2	83,4	95,9		
experimento III		KNN + NCA		81,8	88,4	69,2	86,0	82,7	86,3
experimento IV		BC com DT	79,3	81,9	90,9	67,8	82,6	84,2	85,9

* A ref. [172] apresenta um estudo comparativo no qual dois neurocirurgiões e dois neuroradiologistas classificaram as imagens do conjunto de dados em relação à sua capacidade de determinar com segurança a codeleção 1p/19q do tumor, usando uma escala de 1 a 5 (em que 1 correspondia a “inseguro” e 5 correspondia a “confiante”). Essa pontuação foi convertida em um marcador de previsão de codeleção, a partir do qual essas métricas foram calculadas.

Apesar do grande desafio para execução de algoritmos de aprendizado de máquina ser a obtenção de um número adequado de instâncias para treinamento e testes, no experimento I, a execução da MLP mostrou capacidade de aprender as características radiômicas extraídas da imagem, obtendo melhor precisão, sensibilidade e especificidade na base de treinamento sobre a base de testes, sobretudo pela quantidade de instâncias ser pequena na base de testes e pelo aumento sintético das instâncias no conjunto de treinamento. Conforme resultados da tabela 4.18, a MLP implementada no experimento I apresentou uma alta sensibilidade (88,0%) se comparada aos demais resultados e um resultado pouco acima do valor obtido em [150]. A especificidade e precisão obtidas ficaram abaixo, mas próximas, do valor obtido em [150], trabalho que também implementou uma MLP.

Importante registrar que as características utilizadas no experimento I (Tabela 4.1), quando utilizadas sem nenhum pré-processamento para treinamento da MLP, resultaram

em um valor de precisão de 65,0% no conjunto de dados de teste. A normalização, o redimensionamento e o aumento na base de dados foram importantes para um resultado mais robusto da MLP sobre as demais técnicas comparadas (veja Tabela 4.4). Os valores de especificidade e sensibilidade obtidos indicam que a MLP foi capaz de identificar com bom percentual de acerto a codeleção cromossômica em gliomas de baixo grau.

Pode-se notar que os resultados obtidos no experimento II são bem favoráveis quando comparados com as referências citadas na Tabela 4.18. Em relação à precisão, os algoritmos RL e RF demonstraram resultados bem superiores aos apresentados. Quanto à acurácia, o algoritmo RL apresenta pequena diferença a menor do que o valor obtido nas referências [170, 150], porém, o algoritmo RF obteve um resultado superior a todos os demais valores obtidos pelas referências apresentadas na Tabela 4.18. Com relação a sensibilidade, que representa no contexto do diagnóstico médico o percentual de acerto dos casos verdadeiros positivos, os resultados obtidos pelo experimento foram os segundos melhores valores, sendo inferior, apenas ao valor apresentado na referência [170] e no experimento III. Em relação à especificidade, que apresenta o percentual de acerto de casos verdadeiros negativos, ambos os algoritmos RL e RF apresentaram excelentes resultados, mas o algoritmo RF se destaca com o segundo melhor resultado dentre todos da Tabela 4.18.

A assinatura proposta no experimento III, alcançou resultados compatíveis com os demais experimentos implementados nesse estudo, bem como, resultados promissores com o estado da arte. Apresentou uma alta sensibilidade (88,5%) e em conjunto com o experimento IV, uma alta pontuação F1 (86,4%). Dentre os estudos apresentados na Tabela 4.18 apenas os autores de [150] e [172] com o classificador SVM apresentaram o valor obtido com a pontuação F1, portanto, o resultado obtido no experimento III, representa o segundo melhor resultado da Tabela 4.18. Os valores obtidos com as métricas AUC e Pre, 82,7% e 86,0%, respectivamente, demonstram a promissora capacidade da assinatura em identificar os casos positivos (aqueles em que o algoritmo classificou como codeletado e realmente eram codeletados).

A Tabela 4.18 inclui métricas de um estudo no qual dois neurocirurgiões e dois neurorradiologistas avaliaram as imagens ponderadas em T2 do conjunto de dados LGG-1p19qDeletion [172]. Os avaliadores classificaram cada imagem em relação à capacidade de determinar com segurança a codeleção 1p/19q do tumor, usando uma escala de 1 a 5 (em que 1 correspondia a “inseguro” e 5 correspondia a “confiante”). Essa pontuação foi convertida em um marcador de previsão de codeleção, a partir do qual foram calculadas as métricas apresentadas. A comparação na Tabela 4.18 mostra que o experimento IV com abordagem *ensemble* usando um classificador *bagging* baseado em DT fornece precisão e valores de AUC comparáveis aos melhores métodos da literatura. Trazendo o foco para trabalhos que utilizaram exclusivamente a mesma base de dados (Refs. [173] e [150]), o experimento IV apresenta melhor sensibilidade, mas menor especificidade. Conforme

descrito na seção 4.8.4, alta sensibilidade é mais crítica do que alta especificidade, pois o diagnóstico de codeleção positiva é importante porque os pacientes com tumores codeletados têm melhor prognóstico e maior sobrevida, pois esses tumores respondem bem à quimioterapia e à radioterapia [165]. O método utilizado no experimento IV também obteve melhor sensibilidade e AUC do que os fornecidos por neurocirurgiões e neurorradiologistas analisando as mesmas imagens.

Akkus et al. [39] relatou 87,7% de acurácia usando um método baseado em CNN aplicado ao conjunto de dados LGG-1p19qDeletion. Decidiu-se não incluir este resultado na Tabela 4.18 porque o mesmo foi obtido usando três cortes de cada paciente e, em seguida, dividindo os cortes em conjuntos de treinamento e teste, de modo que os cortes do mesmo paciente podem ter sido usados para as etapas de treinamento e teste. Percebeu-se que isso pode configurar vazamento de dados.

Capítulo 5

Conclusão

Este trabalho abordou a extração de características radiômicas em imagens de ressonância magnética com ponderação em T2 por meio da plataforma PyRadiomics, utilizando-as para classificação da codeleção cromossômica 1p/19q em gliomas de baixo grau através de técnicas clássicas de pré-processamento e aprendizado de máquina.

A identificação da codeleção cromossômica 1p/19q em pacientes com LGG representa uma oportunidade positiva de se determinar melhores decisões terapêuticas. Essa tomada de decisão é importante, pois tumores codeletados (oligodendrogliomas) tem melhor resposta quimioterápica e, portanto esses pacientes tem uma maior sobrevida global, em comparação aos pacientes com tumores não codeletados (astrocitomas).

A busca pelo *dataset* ideal foi uma atividade importante desse trabalho, ou seja, vários conjuntos de dados de imagens de RM eram compostos por imagens sem alterações tumorais, com alterações tumorais (gliomas de baixo e alto grau) e alterações diversas, como outros tumores do SNC, hemorragias, fissuras cirúrgicas, dentre outros. Foi realizada uma busca minuciosa em bancos de dados públicos, visando encontrar conjuntos de imagens adequados ao problema do trabalho. Ainda assim, sabendo que um tumor pode ter áreas com codeleção e sem deleção, não foi disponibilizado informação sobre como a ferramenta que interpretou esta análise interpretou tal situação (casos em que o tumor tem as duas características). Confiou-se na informação disponível junto com o dataset sobre os rótulos dos tumores dos 159 pacientes.

A extração de características de uma imagem é um processo desafiador. Dessa forma, em se tratando de imagens médicas para o auxílio no diagnóstico de uma doença tão considerável como o câncer, torna-se maior a responsabilidade do estudo. Entretanto, a etapa de execução foi concluída com sucesso, e a plataforma PyRadiomics se mostrou eficiente na extração de características radiômicas.

As etapas seguintes de análise das características extraídas foram importantes para melhorar o desempenho do modelo de classificação escolhido. A padronização, o dimensionamento e o aumento dos dados foram bastante importantes para a obtenção de resultados promissores.

Os algoritmos utilizados nos experimentos desse trabalho foram desenvolvidos no ambiente Anaconda Navigator 6.0 com Jupyter Notebook 6.2.0 e Spider 4.1.0 (The Scientific Python Development Environment) e concomitantemente executados na plataforma do Google Colab a fim de verificar resultados e performance. Apesar de algoritmos e *dataset* relativamente pequenos, tanto os resultados apurados pelas métricas como o tempo de execução foram os mesmos em ambos os ambientes. Houve um pequeno ganho de tempo (menor que 1 segundo) na etapa de treinamento em 100 épocas pelo Google Colab no experimento I. Portanto, assim como constatado para o equipamento utilizado para o desenvolvimento do estudo e para os algoritmos desenvolvidos, ambos ambientes de desenvolvimento demonstraram ser satisfatórios.

Quatro formas de apresentar os resultados da codeleção cromossômica 1p/19q foram descritas nesse estudo em seus experimentos: as métricas (acurácia balanceada, acurácia, sensibilidade, especificidade, precisão, area sob a curva ROC e pontuação F1), a matriz de confusão e a exibição da curva ROC. Os recursos e técnicas utilizados, bem como os resultados deste trabalho, foram acompanhados por um médico neurorradiologista com conhecimento e experiência em gliomas de baixo grau, podendo proporcionar à pesquisa um olhar clínico.

A avaliação da codeleção cromossômica 1p/19q no glioma de baixo grau pode fornecer ao paciente um prognóstico mais favorável, melhor estimativa de sobrevida e melhor resposta aos tratamentos neoadjuvantes. Os resultados apresentados nesse trabalho demonstram o potencial que as ferramentas aqui desenvolvidas têm para contribuir nesse processo de avaliação.

Trabalhos Futuros

Conhecendo o glioma de alto grau (do inglês, *high-grade glioma*, HGG), um exemplo dele o glioblastoma tipicamente conhecido como “selvagem”, onde o paciente após diagnóstico normalmente terá uma sobrevida estimada entre 12 a 15 meses de vida, deixa-se, como proposta futura, a continuidade deste trabalho, para avaliação de um conjunto maior de classes (rótulos), juntamente com as características radiômicas extraídas, e utilizar os algoritmos implementados nos experimentos deste trabalho para realizar uma classificação das mutações genômicas disponíveis em outros *datasets* e uma classificação da previsão de sobrevivência do paciente [174] em HGG. *Datasets* como o BraTS HGG, o BTP HGG e o Ivy HGG possuem conjuntos de imagens que necessitam da averiguação da possibilidade de incorporar as características radiômicas à classificação, pois os mesmos não possuem os rótulos genômicos. Sugere-se ainda, para novos trabalhos, avaliar novas técnicas de escalonamento, redimensionamento, aumento de dados, implementação e testes das diversas CNNs pré-treinadas já disponíveis na literatura e construção de um *dataset* próprio.

Referências Bibliográficas

- [1] I. N. de Câncer, “Estatísticas de câncer.” Disponível em: <https://www.inca.gov.br/numeros-de-cancer>. Acesso em: 05 de Maio de 2020.
- [2] A. C. Society, “Types of brain and spinal cord tumors in adults.” Disponível em: <https://www.cancer.org/cancer/brain-spinal-cord-tumors-adults/about/types-of-brain-tumors.html>. Acesso em: 07 de Maio de 2020.
- [3] D. N. Louis, A. Perry, G. Reifenberger, A. Von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison, “The 2016 world health organization classification of tumors of the central nervous system: a summary,” *Acta neuropathologica*, vol. 131, no. 6, pp. 803–820, 2016.
- [4] H. B. Suh, Y. S. Choi, S. Bae, S. S. Ahn, J. H. Chang, S.-G. Kang, E. H. Kim, S. H. Kim, and S.-K. Lee, “Primary central nervous system lymphoma and atypical glioblastoma: differentiation using radiomics approach,” *European radiology*, vol. 28, no. 9, pp. 3832–3839, 2018.
- [5] Y. Fan, C. Chen, F. Zhao, Z. Tian, J. Wang, X. Ma, and J. Xu, “Radiomics-based machine learning technology enables better differentiation between glioblastoma and anaplastic oligodendroglioma,” *Frontiers in oncology*, vol. 9, 2019. doi: 10.3389/fonc.2019.01164.
- [6] D. Krex, B. Klink, C. Hartmann, A. von Deimling, T. Pietsch, M. Simon, M. Sabel, J. P. Steinbach, O. Heese, G. Reifenberger, *et al.*, “Long-term survival with glioblastoma multiforme,” *Brain*, vol. 130, no. 10, pp. 2596–2606, 2017. doi: 10.1093/brain/awm204.
- [7] S. Park, S. Min, H.-S. Choi, and S. Yoon, “Deep recurrent neural network-based identification of precursor micrnas,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [8] Z. Liu, S. Wang, J. W. Di Dong, C. Fang, X. Zhou, K. Sun, L. Li, B. Li, M. Wang, and J. Tian, “The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges,” *Theranostics*, vol. 9, no. 5, p. 1303, 2019.

- [9] D. Zapparolli, “Inovação no diagnóstico de câncer.” Disponível em: <https://revistapesquisa.fapesp.br/wp-content/uploads/2018/02/077-079>. Acesso em: 01 de Março de 2018.
- [10] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, “Deep learning: new computational modelling techniques for genomics,” *Nature Reviews Genetics*, vol. 20, no. 7, pp. 389–403, 2019.
- [11] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: images are more than pictures, they are data,” *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.
- [12] G. S. Omenn, S. J. Nass, C. M. Micheel, *et al.*, *Evolution of translational omics: lessons learned and the path forward*. National Academies Press, 2012.
- [13] M. L. Giger, “Update on the potential of computer-aided diagnosis for breast cancer,” *Future Oncology*, vol. 6, no. 1, pp. 1–4, 2010.
- [14] A. Wibmer, H. Hricak, T. Gondo, K. Matsumoto, H. Veeraraghavan, D. Fehr, J. Zheng, D. Goldman, C. Moskowitz, S. W. Fine, *et al.*, “Haralick texture analysis of prostate mri: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different gleason scores,” *European radiology*, vol. 25, no. 10, pp. 2840–2850, 2015.
- [15] T. P. Coroller, P. Grossmann, Y. Hou, E. R. Velazquez, R. T. Leijenaar, G. Hermann, P. Lambin, B. Haibe-Kains, R. H. Mak, and H. J. Aerts, “Ct-based radiomic signature predicts distant metastasis in lung adenocarcinoma,” *Radiotherapy and Oncology*, vol. 114, no. 3, pp. 345–350, 2015.
- [16] J. Ren, J. Tian, Y. Yuan, D. Dong, X. Li, Y. Shi, and X. Tao, “Magnetic resonance imaging based radiomics signature for the preoperative discrimination of stage i-ii and iii-iv head and neck squamous cell carcinoma,” *European journal of radiology*, vol. 106, pp. 1–6, 2018.
- [17] Z. Zhou, L. Chen, D. Sher, Q. Zhang, J. Shah, N.-L. Pham, S. Jiang, and J. Wang, “Predicting lymph node metastasis in head and neck cancer by combining many-objective radiomics and 3-dimensional convolutional neural network through evidential reasoning,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–4, IEEE, 2018.
- [18] R. Siegel, K. Miller, and A. Jemal, “Cancer statistics, 2017,” *CA Cancer J Clin*, vol. 67, pp. 7–30, 2017.
- [19] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, *et al.*, “Decoding tumour

- phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nature communications*, vol. 5, no. 1, pp. 1–9, 2014. doi: 10.1038/ncomms5006.
- [20] H. Zhou, D. Dong, B. Chen, M. Fang, Y. Cheng, Y. Gan, R. Zhang, L. Zhang, Y. Zang, Z. Liu, *et al.*, “Diagnosis of distant metastasis of lung cancer: based on clinical and radiomic features,” *Translational oncology*, vol. 11, no. 1, pp. 31–36, 2018.
- [21] Y. Guo, Y. Hu, M. Qiao, Y. Wang, J. Yu, J. Li, and C. Chang, “Radiomics analysis on ultrasound for prediction of biologic behavior in breast invasive ductal carcinoma,” *Clinical breast cancer*, vol. 18, no. 3, pp. e335–e344, 2018.
- [22] N. Antropova, B. Q. Huynh, and M. L. Giger, “A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets,” *Medical physics*, vol. 44, no. 10, pp. 5162–5171, 2017.
- [23] H. Chan, B. H. van der Velden, C. E. Loo, and K. G. Gilhuijs, “Eigentumors for prediction of treatment failure in patients with early-stage breast cancer using dynamic contrast-enhanced mri: a feasibility study,” *Physics in Medicine & Biology*, vol. 62, no. 16, p. 6467, 2017.
- [24] H. Park, Y. Lim, E. S. Ko, H.-h. Cho, J. E. Lee, B.-K. Han, E. Y. Ko, J. S. Choi, and K. W. Park, “Radiomics signature on magnetic resonance imaging: association with disease-free survival in patients with invasive breast cancer,” *Clinical Cancer Research*, vol. 24, no. 19, pp. 4705–4714, 2018.
- [25] R. Siegel, “Cancer statistics, 2014,” *CA Cancer j Clin*, vol. 64, pp. 9–29, 2014.
- [26] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, *et al.*, “Radiomics: the process and the challenges,” *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.
- [27] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, *et al.*, “Radiomics: extracting more information from medical images using advanced feature analysis,” *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.
- [28] J. E. Park, H. S. Kim, D. Kim, S. Y. Park, J. Y. Kim, S. J. Cho, and J. H. Kim, “A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features,” *BMC cancer*, vol. 20, no. 1, pp. 1–11, 2020.
- [29] X. Zhang, Q. Tian, L. Wang, Y. Liu, B. Li, Z. Liang, P. Gao, K. Zheng, B. Zhao, and H. Lu, “Radiomics strategy for molecular subtype stratification of lower-grade

- glioma: detecting idh and tp53 mutations based on multimodal mri,” *Journal of Magnetic Resonance Imaging*, vol. 48, no. 4, pp. 916–926, 2018.
- [30] M. Zhou, J. Scott, B. Chaudhury, L. Hall, D. Goldgof, K. W. Yeom, M. Iv, Y. Ou, J. Kalpathy-Cramer, S. Napel, *et al.*, “Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches,” *American Journal of Neuroradiology*, vol. 39, no. 2, pp. 208–216, 2018.
- [31] F. W. Prior, K. Clark, P. Commean, J. Freymann, C. Jaffe, J. Kirby, S. Moore, K. Smith, L. Tarbox, B. Vendt, *et al.*, “Tcia: an information resource to enable open science,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1282–1285, IEEE, 2013.
- [32] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.
- [33] E. I. Zacharaki, S. Wang, S. Chawla, D. Soo Yoo, R. Wolf, E. R. Melhem, and C. Davatzikos, “Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 62, no. 6, pp. 1609–1618, 2009.
- [34] T. J. Yun, C.-K. Park, T. M. Kim, S.-H. Lee, J.-H. Kim, C.-H. Sohn, S.-H. Park, I. H. Kim, and S. H. Choi, “Glioblastoma treated with concurrent radiation therapy and temozolomide chemotherapy: differentiation of true progression from pseudo-progression with quantitative dynamic contrast-enhanced mr imaging,” *Radiology*, vol. 274, no. 3, pp. 830–840, 2015.
- [35] M. Zhou, L. O. Hall, D. B. Goldgof, R. J. Gillies, and R. A. Gatenby, “Survival time prediction of patients with glioblastoma multiforme tumors using spatial distance measurement,” in *Medical Imaging 2013: Computer-Aided Diagnosis*, vol. 8670, p. 86702O, International Society for Optics and Photonics, 2013.
- [36] M. Zhou, L. Hall, D. Goldgof, R. Russo, Y. Balagurunathan, R. Gillies, and R. Gatenby, “Radiologically defined ecological dynamics and clinical outcomes in glioblastoma multiforme: preliminary results,” *Translational oncology*, vol. 7, no. 1, p. 5, 2014.
- [37] P. O. Zinn, B. Majadan, P. Sathyan, S. K. Singh, S. Majumder, F. A. Jolesz, and R. R. Colen, “Radiogenomic mapping of edema/cellular invasion mri-phenotypes in glioblastoma multiforme,” *PloS one*, vol. 6, no. 10, 2011.

- [38] H. Itakura, A. S. Achrol, L. A. Mitchell, J. J. Loya, T. Liu, E. M. Westbroek, A. H. Feroze, S. Rodriguez, S. Echegaray, T. D. Azad, *et al.*, “Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities,” *Science translational medicine*, vol. 7, no. 303, pp. 303ra138–303ra138, 2015.
- [39] Z. Akkus, I. Ali, J. Sedlář, J. P. Agrawal, I. F. Parney, C. Giannini, and B. J. Erickson, “Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from mr images using machine intelligence,” *Journal of digital imaging*, vol. 30, no. 4, pp. 469–476, 2017.
- [40] Y. Zhang, C. Chen, W. Guo, H. Xu, X. Ou, J. Wang, X. Ma, and J. Xu, “Ability of radiomics in differentiation anaplastic oligodendroglioma from atypical low-grade oligodendroglioma using machine learning approach,” *Frontiers in oncology*, vol. 9, p. 1371, 2019. doi: 10.3389/fonc.2019.01371.
- [41] C. Nioche, F. Orhac, S. Boughdad, S. Reuzé, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. Frouin, and I. Buvat, “Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity,” *Cancer research*, vol. 78, no. 16, pp. 4786–4789, 2018. doi: 10.1158/0008-5472.CAN-18-0125.
- [42] M. Decuyper, S. Bonte, K. Deblaere, and R. Van Holen, “Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma,” *Computerized Medical Imaging and Graphics*, vol. 88, p. 101831, 2021.
- [43] S. O. Rezende, *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda, 2003.
- [44] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.
- [45] S. Mukherjeedare, “Ai versus md: what happens when diagnosis is automated,” *The New Yorker*, 2017.
- [46] I. Google, “Research that helps push care forward.” Disponível em: <https://health.google/health-research/>. Acesso em: 11 de Maio de 2020.
- [47] M. Zhou, B. Chaudhury, L. O. Hall, D. B. Goldgof, R. J. Gillies, and R. A. Gatenby, “Identifying spatial imaging biomarkers of glioblastoma multiforme for survival group prediction,” *Journal of Magnetic Resonance Imaging*, vol. 46, no. 1, pp. 115–123, 2017.

- [48] H. Akai, K. Yasaka, A. Kunimatsu, M. Nojima, T. Kokudo, N. Kokudo, K. Hasegawa, O. Abe, K. Ohtomo, and S. Kiryu, “Predicting prognosis of resected hepatocellular carcinoma by radiomics analysis with random survival forest,” *Diagnostic and interventional imaging*, vol. 99, no. 10, pp. 643–651, 2018.
- [49] S. R. van der Voort, R. Gahrman, M. J. van den Bent, A. J. Vincent, W. J. Niessen, M. Smits, and S. Klein, “Radiogenomic classification of the 1p/19q status in presumed low-grade gliomas,” in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 638–641, IEEE, 2017.
- [50] A. Malmström, *Studies for Better Treatment of Patients with Glioma*. PhD thesis, Linköping University Electronic Press, 2019.
- [51] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, *et al.*, “The cancer imaging archive (tcia): maintaining and operating a public information repository,” *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013. doi: 10.1007 / s10278-013-9622-7.
- [52] S. Fellah, D. Caudal, A. M. De Paula, P. Dory-Lautrec, D. Figarella-Branger, O. Chinot, P. Metellus, P. J. Cozzone, S. Confort-Gouny, B. Ghattas, *et al.*, “Multimodal mr imaging (diffusion, perfusion, and spectroscopy): is it possible to distinguish oligodendroglial tumor grade and 1p/19q codeletion in the pretherapeutic diagnosis?,” *American Journal of Neuroradiology*, vol. 34, no. 7, pp. 1326–1333, 2013.
- [53] A. Marques Da Silva, A. Patrocínio, and H. Schiabel, “Processamento e análise de imagens médicas,” *Revista Brasileira de Física Médica*, vol. 13, p. 34, 09 2019. doi: 10.29384/rbfm.2019.v13.n1.p34-48.
- [54] E. D. Pisano, E. B. Cole, B. M. Hemminger, M. J. Yaffe, S. R. Aylward, A. D. Maidment, R. E. Johnston, M. B. Williams, L. T. Niklason, E. F. Conant, *et al.*, “Image processing algorithms for digital mammography: a pictorial essay,” *Radiographics*, vol. 20, no. 5, pp. 1479–1491, 2000.
- [55] R. C. Gonzalez and R. E. Woods, *Processamento de imagens digitais*. Pearson, 3 ed., 2010.
- [56] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973. doi: <https://doi.org/10.1109/TSMC.1973.4309314>.
- [57] R. M. Haralick, “Statistical and structural approaches to texture,” *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, 1979. doi: <https://doi.org/10.1109/PROC.1979.11328>.

- [58] P. Lambin, R. T. Leijenaar, T. M. Deist, J. Peerlings, E. E. De Jong, J. Van Timmeren, S. Sanduleanu, R. T. Larue, A. J. Even, A. Jochems, *et al.*, “Radiomics: the bridge between medical imaging and personalized medicine,” *Nature reviews Clinical oncology*, vol. 14, no. 12, pp. 749–762, 2017. doi: 10.1038/nrclinonc.2017.141.
- [59] L. Papp, N. Pötsch, M. Grahovac, V. Schmidbauer, A. Woehrer, M. Preusser, M. Mitterhauser, B. Kiesel, W. Wadsak, T. Beyer, *et al.*, “Glioma survival prediction with combined analysis of in vivo 11c-met pet features, ex vivo features, and patient features by supervised machine learning,” *Journal of Nuclear Medicine*, vol. 59, no. 6, pp. 892–899, 2018. doi: 10.2967/jnumed.117.202267.
- [60] H. J. Aerts, “The potential of radiomic-based phenotyping in precision medicine: a review,” *JAMA oncology*, vol. 2, no. 12, pp. 1636–1642, 2016.
- [61] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, “Computational radiomics system to decode the radiographic phenotype,” *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017. doi: <https://doi.org/10.1158/0008-5472.CAN-17-0339>.
- [62] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, *et al.*, “3d slicer as an image computing platform for the quantitative imaging network,” *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1323–1341, 2012.
- [63] H. J. Johnson, M. M. McCormick, and L. Ibanez, “The itk software guide book 2: Design and functionality fourth edition updated for itk version 4.7,” *Kitware, Inc. (January 2015)*, 2016.
- [64] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, *et al.*, “Image biomarker standardisation initiative-feature definitions,” *arXiv preprint arXiv:1612.07003*, 2016.
- [65] D. Molina, J. Pérez-Beteta, A. Martínez-González, J. Martino, C. Velásquez, E. Arana, and V. M. Pérez-García, “Influence of gray level and space discretization on brain tumor heterogeneity measures obtained from magnetic resonance images,” *Computers in biology and medicine*, vol. 78, pp. 49–57, 2016. doi: <https://doi.org/10.1016/j.compbiomed.2016.09.011>.
- [66] R. F. Walker and P. T. Jackway, “Statistical geometric features-extensions for cytological texture analysis,” in *Proceedings of 13th international conference on pattern recognition*, vol. 2, pp. 790–794, IEEE, 1996.
- [67] M. Dash and H. Liu, “Feature selection for classification,” *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

- [68] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, “Optimal number of features as a function of sample size for various classification rules,” *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005.
- [69] X. Jin, A. Xu, R. Bie, and P. Guo, “Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles,” in *International workshop on data mining for biomedical applications*, pp. 106–115, Springer, 2006.
- [70] B. Seijo-Pardo, V. Bolón-Canedo, and A. Alonso-Betanzos, “On developing an automatic threshold applied to feature selection ensembles,” *Information Fusion*, vol. 45, pp. 227–245, 2019.
- [71] R. R. F. DeFellipi, “Standardize or normalize? — examples in python.” Disponível em: <https://medium.com/@rrfd>. Acesso em: 03 de Março de 2020.
- [72] S. Geller, “Normalization vs standardization — quantitative analysis.” Disponível em: <https://towardsdatascience.com/@shayzm1>. Acesso em: 05 de Março de 2020.
- [73] S. Raschka and V. Mirjalili, *Python Machine Learning, 3rd Ed.* Birmingham, UK: Packt Publishing, 2019.
- [74] J. Han, M. Kamber, and J. Pei, “Data mining concepts and techniques,” *Morgan Kaufmann*, 2011.
- [75] Y.-H. Pao and Z. Meng, “Visualization and the understanding of multidimensional data,” *Engineering Applications of Artificial Intelligence*, vol. 11, no. 5, pp. 659–667, 1998.
- [76] M. Kreuzeler, “Visualization of geographically related multidimensional data in virtual 3d scenes,” *Computers & Geosciences*, vol. 26, no. 1, pp. 101–108, 2000.
- [77] H. Guo, H. Xiao, and X. Yuan, “Multi-dimensional transfer function design based on flexible dimension projection embedded in parallel coordinates,” in *2011 IEEE Pacific Visualization Symposium*, pp. 19–26, IEEE, 2011.
- [78] X. Yuan, D. Ren, Z. Wang, and C. Guo, “Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2625–2633, 2013.
- [79] F. S. Tsai, “A visualization metric for dimensionality reduction,” *Expert Systems with Applications*, vol. 39, no. 2, pp. 1747–1752, 2012.

- [80] T. Balachander, R. Kothari, and H. Cualing, “An empirical comparison of dimensionality reduction techniques for pattern classification,” in *International Conference on Artificial Neural Networks*, pp. 589–594, Springer, 1997.
- [81] S. De Backer, A. Naud, and P. Scheunders, “Non-linear dimensionality reduction techniques for unsupervised feature extraction,” *Pattern Recognition Letters*, vol. 19, no. 8, pp. 711–720, 1998.
- [82] C. O. S. Sorzano, J. Vargas, and A. P. Montano, “A survey of dimensionality reduction techniques,” *arXiv preprint arXiv:1403.2877*, 2014.
- [83] M. Djerioui, Y. Brik, M. Ladjal, and B. Attallah, “Neighborhood component analysis and support vector machines for heart disease prediction,” *Ingénierie des Systèmes d’Information*, 2019.
- [84] W. Yang, K. Wang, and W. Zuo, “Neighborhood component feature selection for high-dimensional data.,” *J. Comput.*, vol. 7, no. 1, pp. 161–168, 2012.
- [85] M. A. B. Siddique, S. Sakib, and M. A. Rahman, “Performance analysis of deep autoencoder and nca dimensionality reduction techniques with knn, enn and svm classifiers,” in *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, pp. 1–6, IEEE, 2019.
- [86] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea, “Visual analysis of dimensionality reduction quality for parameterized projections,” *Computers & Graphics*, vol. 41, pp. 26–42, 2014. doi: <http://dx.doi.org/10.1016/j.cag.2014.01.006>.
- [87] H. Hotelling, “Analysis of a complex of statistical variables into principal components.,” *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.
- [88] R. Orlandic, Y. Lai, and W. G. Yee, “Clustering high-dimensional data using an efficient and effective data space reduction,” in *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 201–208, 2005.
- [89] I. T. Jolliffe, “Principal components in regression analysis,” in *Principal component analysis*, pp. 129–155, Springer, 1986.
- [90] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, “Neighbourhood components analysis,” *Advances in neural information processing systems*, vol. 17, 2004.

- [91] N. Singh-Miller, M. Collins, and T. J. Hazen, “Dimensionality reduction for speech recognition using neighborhood components analysis,” in *Eighth Annual Conference of the International Speech Communication Association*, 2007.
- [92] F. Provost and T. Fawcett, “Robust classification for imprecise environments,” *Machine learning*, vol. 42, no. 3, pp. 203–231, 2001.
- [93] K. S. Woods, J. L. Solka, C. E. Priebe, W. P. Kegelmeyer Jr, C. C. Doss, and K. W. Bowyer, “Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography,” in *State of The Art in Digital Mammographic Image Analysis*, pp. 213–231, World Scientific, 1994.
- [94] S. Lertampaiporn, C. Thammamongtham, C. Nukoolkit, B. Kaewkamnerdpong, and M. Ruengjitchatchawalya, “Heterogeneous ensemble approach with discriminative features and modified-smotebagging for pre-mirna classification,” *Nucleic acids research*, vol. 41, no. 1, pp. e21–e21, 2013.
- [95] Y. B. Marques, “Mirnacle: aprendizagem de máquina utilizando smote e random forest para prover aumento da seletividade na predição ab initio de pre-mirnas,” dissertação de mestrado, Universidade Federal de Viçosa, 2015.
- [96] F. J. Moreno-Barea, F. Strazzera, J. M. Jerez, D. Urda, and L. Franco, “Forward noise adjustment scheme for data augmentation,” in *2018 IEEE symposium series on computational intelligence (SSCI)*, pp. 728–734, IEEE, 2018.
- [97] Y.-D. Zhang, Z. Dong, X. Chen, W. Jia, S. Du, K. Muhammad, and S.-H. Wang, “Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation,” *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3613–3632, 2019.
- [98] T. Silva, F. Cassia, GS Nascimento, and J. Carvalho, “Predição da codeleção 1p/19q utilizando atributos radiômicos e aprendizado de máquina,” XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 2021.
- [99] J. Brownlee, “Imbalanced classification with python: Better metrics, balance skewed classes, cost-sensitive learning,” *Machine Learning Mastery*, 2020.
- [100] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953.
- [101] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

- [102] S. N. Gerges, “Ruído: fundamentos e controle,” in *Ruído: fundamentos e controle*, pp. 600–600, 1992.
- [103] J. Domingo-Ferrer, J. Drechsler, and S. Poletini, “Essnet-sdc deliverable report on synthetic data files,” *Universitat Rovira i Virgili & Institute for Employment Research*, 2009.
- [104] I. Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective,” *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001.
- [105] T. Escovedo and A. S. Koshiyama, *Introdução a Data Science - Algoritmos de Machine Learning e métodos de análise*. Casa do Código, 2020.
- [106] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larrañaga, and J. A. Lozano, “Machine learning: an indispensable tool in bioinformatics,” in *Bioinformatics methods in clinical research*, pp. 25–48, Springer, 2010.
- [107] G. Dougherty, *Pattern recognition and classification: an introduction*. Springer Science & Business Media, 2012.
- [108] F. J. Von Zuben, *Modelos parametricos e não-parametricos de redes neurais artificiais e aplicações*. PhD thesis, University of Campinas, Brazil, 1996.
- [109] L. G. Hafemann, “An analysis of deep neural networks for texture classification,” master’s degree dissertation, Universidade Federal do Paraná, 2014.
- [110] S. Haykin, *Redes neurais: princípios e prática*. Bookman Editora, 2008.
- [111] M. P. LaValley, “Logistic regression,” *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.
- [112] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [113] Y. Wu, K. Ianakiev, and V. Govindaraju, “Improved k-nearest neighbor classification,” *Pattern recognition*, vol. 35, no. 10, pp. 2311–2318, 2002.
- [114] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, 1992.
- [115] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [116] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

- [117] Y. Freund and L. Mason, “The alternating decision tree learning algorithm,” in *icml*, vol. 99, pp. 124–133, 1999.
- [118] J. Su and H. Zhang, “A fast decision tree learning algorithm,” in *Aaai*, vol. 6, pp. 500–505, 2006.
- [119] I. Rish *et al.*, “An empirical study of the naive bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46, 2001.
- [120] R. E. Schapire, “Explaining adaboost,” in *Empirical inference*, pp. 37–52, United States: Springer, 2013.
- [121] S.-i. Amari, “Backpropagation and stochastic gradient descent method,” *Neurocomputing*, vol. 5, no. 4-5, pp. 185–196, 1993.
- [122] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT’2010*, pp. 177–186, United States: Springer, 2010.
- [123] J. Shafer, R. Agrawal, M. Mehta, *et al.*, “SPRINT: A scalable parallel classifier for data mining,” in *Vldb*, vol. 96, pp. 544–555, Citeseer, 1996.
- [124] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.
- [125] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [126] B. S. Bhati and C. Rai, “Ensemble based approach for intrusion detection using extra tree classifier,” in *Intelligent computing in engineering*, pp. 213–220, United States: Springer, 2020.
- [127] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.
- [128] S. Balakrishnama and A. Ganapathiraju, “Linear discriminant analysis-a brief tutorial,” *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.
- [129] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.
- [130] C. J. Date, *Introdução a sistemas de bancos de dados*. Elsevier Brasil, 2004.
- [131] B. Erickson, Z. Akkus, J. Sedlar, and P. Korfiatis, “Data from lgg-1p19qdeletion,” 2017. doi: 10.7937/K9/TCIA.2017.DWEHTZ9V.

- [132] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, *et al.*, “Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge,” *arXiv preprint arXiv:1811.02629*, 2018.
- [133] J. West, *Quantitative Magnetic Resonance Imaging of the Brain: Applications for Tissue Segmentation and Multiple Sclerosis*. PhD thesis, Linköping University Electronic Press, 2014.
- [134] S. Cha, “Update on brain tumor imaging: from anatomy to physiology,” *American Journal of Neuroradiology*, vol. 27, no. 3, pp. 475–487, 2006.
- [135] W. E. Brant and C. A. Helms, *Fundamentals of diagnostic radiology*. Lippincott Williams & Wilkins, 2012.
- [136] M. R. Grever, S. A. Schepartz, and B. A. Chabner, “The national cancer institute: cancer drug discovery and development program.,” in *Seminars in oncology*, vol. 19, pp. 622–638, 1992.
- [137] L. P. Clarke, R. J. Nordstrom, H. Zhang, P. Tandon, Y. Zhang, G. Redmond, K. Farahani, G. Kelloff, L. Henderson, L. Shankar, *et al.*, “The quantitative imaging network: Nci’s historical perspective and planned goals,” *Translational oncology*, vol. 7, no. 1, p. 1, 2014.
- [138] A. J. Buckler, L. Bresolin, N. R. Dunnick, D. C. Sullivan, and Group, “A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging,” *Radiology*, vol. 258, no. 3, pp. 906–914, 2011.
- [139] D. Dua and C. Graff, “UCI machine learning repository,” 2017.
- [140] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, “The multimodal brain tumor image segmentation benchmark (brats),” *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014. doi: 10.1109/TMI.2014.2377694.
- [141] N. I. of Health, “Neuroimaging informatics technology initiative.” Disponível em: <https://nifti.nimh.nih.gov/>. Acesso em: 05 de Julho de 2020.
- [142] M. Everts, “Pynrrd’s documentation.” Disponível em: <https://pynrrd.readthedocs.io/en/latest/index.html>. Acesso em: 05 de Julho de 2020.
- [143] K. Erwin, *Advanced engineering mathematics*. United States: John Wiley, Inc. 4 th edition, 1979, 1979.

- [144] A. Géron, *Hands-on machine learning with scikit-learn and tensorflow: Concepts Tools and Techniques to Build Intelligent Systems*. United States: O'Reilly Media, 2 nd edition, 2019, 2019.
- [145] H. Dalianis, "Evaluation metrics and evaluation," in *Clinical text mining*, pp. 45–53, United States: Springer, 2018.
- [146] A. Saxena, J. Celaya, E. Balaban, K. Goebel, B. Saha, S. Saha, and M. Schwabacher, "Metrics for evaluating performance of prognostic techniques," in *2008 international conference on prognostics and health management*, pp. 1–17, IEEE, 2008.
- [147] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*, pp. 3121–3124, IEEE, 2010.
- [148] O. ended Neuro-Eletronic Intelligent Robot Operating System, "Keras and tensorflow 2.0." Disponível em: <https://www.keras.io>. Acesso em: 10 de Junho de 2020.
- [149] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
- [150] B. Kocak, E. S. Durmaz, E. Ates, I. Sel, S. Turgut Gunes, O. K. Kaya, A. Zeynalova, and O. Kilickesmez, "Radiogenomics of lower-grade gliomas: Machine learning-based mri texture analysis for predicting 1p/19q codeletion status," *European radiology*, vol. 30, no. 2, pp. 877–886, 2020.
- [151] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification: Algorithms and Applications*, 2014.
- [152] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [153] R. Li, L. Xing, S. Napel, and D. L. Rubin, *Radiomics and radiogenomics: technical basis and clinical applications*. CRC Press, 2019.
- [154] T. Katostaras and N. Katostara, "Area of the roc curve when one point is available.," in *ICIMTH*, pp. 219–221, 2013.
- [155] G. Bonaccorso, *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. United States: Packt Publishing Ltd, 2018.
- [156] N. Soni, S. Priya, and G. Bathla, "Texture analysis in cerebral gliomas: a review of the literature," *American Journal of Neuroradiology*, vol. 40, no. 6, pp. 928–934, 2019.

- [157] R. Brown, M. Zlatescu, A. Sijben, G. Roldan, J. Easaw, P. Forsyth, I. Parney, R. Sevick, E. Yan, D. Demetrick, *et al.*, “The use of magnetic resonance imaging to noninvasively detect genetic signatures in oligodendroglioma,” *Clinical Cancer Research*, vol. 14, no. 8, pp. 2357–2362, 2008.
- [158] M. W. Fagerland, “t-tests, non-parametric tests, and large studies—a paradox of statistical practice?,” *BMC medical research methodology*, vol. 12, no. 1, pp. 1–7, 2012.
- [159] O. El Aissaoui, Y. E. A. EL MADANI, L. Oughdir, and Y. EL ALLIOUI, “Combining supervised and unsupervised machine learning algorithms to predict the learners’ learning styles,” *Procedia computer science*, vol. 148, pp. 87–96, 2019.
- [160] S. E. Viswanath, P. V. Chirra, M. C. Yim, N. M. Rofsky, A. S. Purysko, M. A. Rosen, B. N. Bloch, and A. Madabhushi, “Comparing radiomic classifiers and classifier ensembles for detection of peripheral zone prostate tumors on t2-weighted mri: A multi-site study,” *BMC Medical Imaging*, vol. 19, no. 1, pp. 1–12, 2019.
- [161] D. Lavanya and K. U. Rani, “Ensemble decision tree classifier for breast cancer data,” *International Journal of Information Technology Convergence and Services*, vol. 2, no. 1, pp. 17–24, 2012.
- [162] M. N. KP and P. Thiyagarajan, “Alzheimer’s classification using dynamic ensemble of classifiers selection algorithms: A performance analysis,” *Biomedical Signal Processing and Control*, vol. 68, p. 102729, 2021.
- [163] A. Idri, M. Hosni, and A. Abran, “Systematic mapping study of ensemble effort estimation.,” *ENASE*, pp. 132–139, 2016.
- [164] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. United States: CRC press, 2012.
- [165] J. E. Eckel-Passow, D. H. Lachance, A. M. Molinaro, K. M. Walsh, P. A. Decker, H. Sicotte, M. Pekmezci, T. Rice, M. L. Kosel, I. V. Smirnov, *et al.*, “Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors,” *New England Journal of Medicine*, vol. 372, no. 26, pp. 2499–2508, 2015.
- [166] H. He, Y. Bai, E. A. Garcia, and S. Li, “Adasyn: Adaptive synthetic sampling approach for imbalanced learning,” in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328, IEEE, 2008.

- [167] J. Zhang, L. Chen, and F. Abid, “Prediction of breast cancer from imbalance respect using cluster-based undersampling method,” *Journal of healthcare engineering*, vol. 2019, 2019.
- [168] S. Ray and B. Analytics, “Beginners guide to learn dimension reduction techniques,” *Analytics Vidhya*, vol. 19, 2015.
- [169] Y. Han, Z. Xie, Y. Zang, S. Zhang, D. Gu, M. Zhou, O. Gevaert, J. Wei, C. Li, H. Chen, *et al.*, “Non-invasive genotype prediction of chromosome 1p/19q co-deletion by development and validation of an mri-based radiomics signature in lower-grade gliomas,” *Journal of neuro-oncology*, vol. 140, no. 2, pp. 297–306, 2018.
- [170] Z. Kong, C. Jiang, Y. Zhang, S. Liu, D. Liu, Z. Liu, W. Chen, P. Liu, T. Yang, Y. Lyu, *et al.*, “Thin-slice magnetic resonance imaging-based radiomics signature predicts chromosomal 1p/19q co-deletion status in grade ii and iii gliomas,” *Frontiers in neurology*, vol. 11, p. 551771, 2020.
- [171] Z. A. Shboul, J. Chen, and K. M. Iftekharuddin, “Prediction of molecular mutations in diffuse low-grade gliomas using mr imaging features,” *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.
- [172] S. R. van der Voort, F. Incekara, M. M. Wijnenga, G. Kapas, M. Gardeniers, J. W. Schouten, M. Starmans, R. Nandoe Tewarie, G. J. Lycklama, P. J. French, *et al.*, “Predicting the 1p/19q codeletion status of presumed low-grade glioma with an externally validated machine learning algorithm prediction on 1p/19 codeletion status of glioma,” *Clinical Cancer Research*, vol. 25, no. 24, pp. 7455–7462, 2019.
- [173] C.-F. Lu, F.-T. Hsu, K. L.-C. Hsieh, Y.-C. J. Kao, S.-J. Cheng, J. B.-K. Hsu, P.-H. Tsai, R.-J. Chen, C.-C. Huang, Y. Yen, *et al.*, “Machine learning-based radiomics for molecular subtyping of gliomas machine learning for molecular subtyping of gliomas,” *Clinical Cancer Research*, vol. 24, no. 18, pp. 4429–4436, 2018.
- [174] H. Ishwaran and M. Lu, “Random survival forests,” *Wiley StatsRef: Statistics Reference Online*, pp. 1–13, 2014.