UNIVERSIDADE DE SÃO PAULO
FFCLRP - DEPARTAMENTO DE FÍSICA
PÓS-GRADUAÇÃO EM FÍSICA APLICADA À MEDICINA E BIOLOGIA

LEONARDO FERREIRA MACHADO

**Análise de Textura em Imagens de Ressonância
Magnética na Predição de Recorrência Tumoral em
Pacientes Portadores de Adenomas Hipofisários
Clinicamente Não Funcionantes.**

**Magnetic Resonance Imaging Texture Analysis in
the Prediction of Tumor Recurrence in Patients
with Non-functioning Pituitary Adenomas.**

Dissertation presented to Faculdade de
Filosofia, Ciências e Letras de Ribeirão
Preto of Universidade de São Paulo, as part
of requirements for acquirement the grade of
Master of Sciences, Area: Physics applied to
Medicine and Biology.

Ribeirão Preto - SP
2017

LEONARDO FERREIRA MACHADO

**Análise de Textura em Imagens de Ressonância
Magnética na Predição de Recorrência Tumoral em
Pacientes Portadores de Adenomas Hipofisários
Clinicamente Não Funcionantes.**

**Magnetic Resonance Imaging Texture Analysis in
the Prediction of Tumor Recurrence in Patients
with Non-functioning Pituitary Adenomas.**

Dissertation presented to Faculdade de
Filosofia, Ciências e Letras de Ribeirão
Preto of Universidade de São Paulo, as part
of requirements for acquirement the grade of
Master of Sciences.

Concentration area:

Física aplicada à Medicina e Biologia.

Advisor:

Luis Otavio Murta Junior.

Rectified version

Original version available at FFCLRP - USP online repository

Ribeirão Preto - SP

2017

I authorize partial and total reproduction of this work, by any conventional or electronic means, for the purpose of study and research, provided the source is cited.

FICHA CATALOGRÁFICA

Machado, Leonardo Ferreira

Análise de Textura em Imagens de Ressonância Magnética na Predição de Recorrência Tumoral em Pacientes Portadores de Adenomas Hipofisários Clinicamente Não Funcionantes. / Leonardo Ferreira Machado; orientador Luis Otavio Murta Junior. Ribeirão Preto - SP, 2017.

135 f.:il.

Dissertação (Mestrado - Programa de Pós-Graduação em Física Aplicada à Medicina e Biologia) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, 2017.

1. Análise de textura. 2. Imagens por ressonância magnética. 3. Predição de recorrência tumoral. 4. APNF. 5. Aprendizagem de máquina.

To my family.

ACKNOWLEDGEMENTS

Firstly, I want to give thanks to God, for he has done wonders, and for he has conserved my sanity during these last two years.

Also, to my mother and family who never stopped believing and supporting me.

To my professor, advisor, and friend, Dr. Luis Otavio Murta Junior, for his endless patience and for the trust he put in my work. Definitely it made a huge difference.

To the clinician who most helped me, Dr^a Paula C. Lamparelli Elias, for her countless and priceless contributions. This work would not be feasible without her.

To my CSIM lab colleagues, Bizu, Mehran, Fabrício, Gustavo, Juliano, and Paula, for their constant support and valuable contributions, and for the good time we had. I learned so much with them.

To my friends from Shalom Catholic Community who made me feel at home in Ribeirão Preto and gave me spiritual and emotional support to accomplish it all. Specially Ana, Rafael, Dorival, Taciana, Bia, André, Regiane, Daniel, and many others. I am deeply grateful to have met them all.

And to all the other friends whom I met here in Ribeirão Preto and who also supported me to continue my work and to do my best. Specially Leo , Yaser, and Hiles ("O Piauí"). Friends are heavenly treasures.

My sincere gratitude.

"Commit your way to the Lord;
trust in him
and he will do this."
Psalm 37,5.

RESUMO

MACHADO, L. F. **Análise de Textura em Imagens de Ressonância Magnética na Predição de Recorrência Tumoral em Pacientes Portadores de Adenomas Hipofisários Clinicamente Não Funcionantes..** 2017. 135 f. Dissertação (Mestrado - Programa de Pós-Graduação em Física Aplicada à Medicina e Biologia) - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto - SP, 2017.

O presente trabalho propõe o uso de parâmetros de textura extraídos computacionalmente de imagens por ressonância magnética (IRM) como biomarcadores de imagem na predição de recorrência tumoral em pacientes de adenomas pituitários clinicamente não funcionantes (APNF). Para isso, esse estudo analisou IRMs de 15 pacientes de APNF retrospectivamente separados em dois grupos: O grupo de pacientes recorrentes, definido por 7 pacientes que exibiram recorrência tumoral em um período de $4,640 \pm 0,653$ anos (média $\pm$ erro *padrão*) de acompanhamento clínico após a primeira abordagem cirúrgica; e o grupo de pacientes estáveis, formado por 8 pacientes com lesões consideradas estáveis em um período de $4,512 \pm 0,536$. Uma máscara de segmentação tridimensional da lesão tumoral foi construída manualmente por um especialista sobre a imagem 3D T1-W com contraste pré-operatória para cada paciente. Em seguida, essa segmentação e a própria imagem de ressonância foram usadas para extrair 48 características numéricas de textura. Adicionalmente, 4 características clínicas foram consideradas no estudo: a imuno-histoquímica, invasividade, idade na primeira cirurgia e sexo, totalizando 52 características. Cada uma destas 52 características foi testada através de testes estatísticos convencionais univariados para avaliar o poder discriminatório dessas características na diferenciação desses dois grupos de

pacientes. Mais adiante, diferentes subconjuntos dessas características foram usados para construir modelos de predição baseados na teoria de aprendizagem de máquinas (usando os algoritmos k-nearest neighbor (kNN), decision tree (DTC), e random forest (RFC)) para investigar um modelo de classificação capaz de identificar os pacientes que experimentariam recorrência tumoral após a primeira cirurgia. Nove características de textura foram consideradas individualmente significantes ($p < 0,05$) na diferenciação dos grupos de pacientes recorrentes e estáveis. Afirmando esses achados, a análise com a curva ROC para cada uma das 9 características exibiu medidas de AUC de 0,803 a 0,857, significando uma boa performance de classificação. A idade, imuno-histoquímica, invasividade e sexo não mostraram evidências de associação com recorrência tumoral. As melhores performances com algoritmos de classificação foram obtidas com kNN e RFC ambos atingiram uma especificidade de 1,000 conservando alta acurácia (0,933) e obtendo 0,991 na análise com a curva ROC, o que caracteriza uma performance de classificação quase perfeita. DTC não mostrou nenhuma melhora se comparado com os resultados das classificações univariadas. Esses resultados permitem concluir que parâmetros de textura são úteis na predição de recorrência tumoral após a primeira cirurgia em pacientes de APNF. E que o poder de predição dessas características pode ser observado por testes estatísticos univariados convencionais e por análises multivariadas através de algoritmos baseados em aprendizagem de máquinas.

Palavras-chave: 1. Análise de textura. 2. Imagens por ressonância magnética. 3. Predição de recorrência tumoral. 4. APNF. 5. Aprendizagem de máquina.

ABSTRACT

MACHADO, L. F. **Magnetic Resonance Imaging Texture Analysis in the Prediction of Tumor Recurrence in Patients with Non-functioning Pituitary Adenomas..** 2017. 135 f. Dissertation (M.Sc. - Postgraduate Program in Physics Applied to Medicine and Biology) - Faculty of Philosophy, Sciences and Literature, University of São Paulo, Ribeirão Preto - SP, 2017.

The present work proposes the usage of texture features computationally extracted from MRI as imaging biomarkers in the prediction of tumor recurrence in patients with non-functioning pituitary adenomas (NFPA). With this purpose, this study analyzed MR images from 15 patients of NFPA retrospectively separated in two groups: the recurrent patient group, formed by seven (7) patients who exhibited tumor recurrence in a period of 4.640 ± 0.653 years (mean $\pm$ standard error) of follow-up after the first surgical approach; and the stable patient group formed by eight (8) patients with lesions considered stable in a period of 4.512 ± 0.536 years. A three-dimensional segmentation mask of the tumor lesion was manually performed by a specialist over the preoperative 3D T1-W MR image post contrast for each patient. Next, this segmentation and the preoperative MRI itself were used to extract 48 numerical textural features. Additionally, 4 clinical features were considered in the study: immunohistochemistry, invasiveness, age at first surgery, and gender, totalizing 52 features. Each one of those 52 features were tested through conventional univariate statistical tests to see if there were evidence of their discrimination power to differentiate between these two patient groups. Moreover, different subsets of those features were used to build machine learning prediction models (using k-nearest neighbor (kNN), decision tree (DTC), and random forest (RFC) algorithms) to investigate a classification model capable of

identifying the patients that would experience tumor recurrence after the first surgery. Nine quantitative textural features were found to be individually significant ($p < 0.05$) in the differentiation of recurrent and stable patient group. Affirming these findings, the ROC curve analysis for each one of those 9 features exhibited an AUC score from 0.803 to 0.857 meaning a good classification performance. Age, immunohistochemistry, invasiveness status, and gender did not show evidence of association with tumor recurrence. The best performances with classification algorithms were obtained with kNN and RFC, both reached specificity of 1.000 conserving high accuracy (0.933) and scoring 0.991 in ROC curve analysis, which characterizes an almost perfect classification performance. DTC did not show any improvement compared to the univariate classification results. These findings allow to conclude that textural features are useful in the prediction of tumor recurrence after first surgery in NFPA patients. And that the prediction value of those features can be observed with both conventional univariate statistical tests and multivariate analyses through machine learning algorithms.

Key-words: 1.Texture analysis. 2.Magnetic resonance imaging. 3.Prediction of tumor recurrence. 4.NFPA. 5.Machine learning.

LIST OF FIGURES

1.1	Structure of pituitary gland. Sella turcica is a bone cavity that encapsulates pituitary gland. The optic chiasm is a very sensitive nerve that connects the brain to the eyes and is easily affected by large pituitary lesions. For the sake of readability, the gland was illustrated bigger. Adapted from [2].	5
1.2	Examples of three different size pituitary adenomas. (a),(b) and (c) bring a sagittal slice of a T1-weighted MR image. Adapted from [2]. .	6
1.3	Hardy-Knosp classification for invasiveness. Hardy grades indicate both invasiveness and symmetry level of tumor mass. Knosp only evaluates invasiveness. Both authors consider the proximity of the tumor mass from surrounding structures to evaluate invasiveness. Adapted from [5].	8
1.4	Decision-making process for FUNCTIONING pituitary adenomas. The option for medical treatment is considered for prolactinomas [2, 11]	10
1.5	Decision-making process for NON-FUNCTIONING pituitary adenomas.	11
2.1	Radiography of the left hand of Wilhelm Rontgen's wife on December 22, 1895, shortly after his discovery of X-rays. Considered to be the first medical image ever produced and the kick-start of a revolution in medicine	16
2.2	Classical illustration of magnetic dipole moment vector (μ) around $\vec{B}_0$ direction. Adapted from [31]	18

2.3	Sample excitation by an RFP. a), b) and c) illustrate what happens with the individual magnetic dipole vectors($\vec{\mu}$) in the sample after a $\pi/2$ RFP. d), e) and f) illustrate what happens with the total magnetization $\vec{M}$. In a) and d) the total magnetization $\vec{M}$ is Longitudinal. In b) and e), after the RFP, the total magnetization $\vec{M}$ is transversal. Adapted from [31].	20
2.4	NFPA imaged before (a) and after (b) gadolinium contrast. The difference in tumor delineation is easily noticed.	23
2.5	Visualization of the field inhomogeneities present in a T1 image. Adapted from [40].	24
3.1	Digital image pixel-wise analysis. It is an 8- <i>bits</i> image, meaning that pixel intensities vary from 0 to 255 ($2^{bits} - 1$). The ROI highlighted was arbitrarily chosen, but in medical imaging studies they are, in fact, physiological structures of interest. Adapted from [27].	27
3.2	Histogram. It is an 12- <i>bits</i> image, meaning that pixel intensities vary from 0 to 4095 ($2^{bits} - 1$).	29
3.3	GLCM for a simplistic image. The image has 6 possible grey-level intensities. The darkest value, 0. The brightest, 5. Red signs indicate the occurred frequency of pair (4, 5) equivalent to (I, I') consecutively appearing in the horizontal direction. $g_{(i,j)}$ is the $(i, j)th$ element on GLCM and represents the number of times the pair of intensities (i, j) appears on the image in a given direction separated by a given distance d	31
3.4	Possible offsets for $d = 1$ in two (a) and three-dimensional (b) image spaces. For every offset a specific GLCM can be generated, meaning that many co-occurrence matrices can be generated from the same image or ROI.	32

3.5	GLRM for a simplistic image. The image has 6 possible grey-level intensities. The darkest value, 0. The brightest, 5. Red signs indicate the run frequency of intensities (4, 4, 4) in the horizontal direction. $P_{(i,j)}$ is the frequency that a run of size j of pixels with intensity i appears over the ROI.	36
3.6	Linear masks for both horizontal ($\hat{x}$) and vertical ($\hat{y}$) directions. Analogously, there is one for the ($\hat{z}$) direction. By applying those masks over an image ([27]), a resultant "gradient image" is generated for each respective direction.	39
3.7	Gradient image obtained from a simple two-dimensional image. Gradient Component images (b) and (c) are obtained, then gradient magnitude image is generated (d). Adapted from [27].	40
4.1	Receiver operator characteristic graphic. The graphical representation is another way to represent the defectiveness of a binary classification task	45
4.2	Machine learning process. The ML scheme depicted here refers to a supervised learning task of two possible classifications, tissue A or B.	47
4.3	Supervised learning tasks illustration. In (a), the problem is to separate the elements in two classes, here represented by circles and crosses. x_1 and x_2 are the two describing numerical features. In (b) the algorithm is trying to fit a prediction for the variable y as x increases. Adapted from [68]	48
4.4	Machine learning process. The ML scheme depicted here refers to a supervised learning task of three possible classifications, i.e., tissue A, B, or C.	48
4.5	2D representation of KNN classification. It is a binary classification problem where square and triangle represent known label classes, and the circle is the unlabeled sample. Solid line represents how the classification would be taken if $K = 3$ nearest neighbors were considered, and dashed line for $K = 5$	50

4.6	Illustration of decision tree process for Iris dataset. It is a multiclass classification problem where labels correspond to three classes of flowers	52
4.7	Illustration of decision tree process for Iris dataset. It is a multiclass classification problem where labels correspond to three classes of flowers	56
5.1	Segmentation process. In GREEN is the segmentation obtained by a semiautomatic method. In BEIGE, the radiologist marking. Difference is highlighted when both markings are overlapped.	61
5.2	Screenshot of 3DSlicer welcome screen.	62
5.3	Screenshot of CLI module interface. Input fields and features methods check-boxes are visible.	63
5.4	CLI module input. (a) raw 3D T1-Weighted DCE, sagittal view. (b) radiologist final segmentation in green. Those are the inputs received into CLI processing pipeline.	64
5.5	CLI module pipeline.	65
5.6	Normalization step output. (a) 3D T1-Weighted DCE normalized and cropped ROI and (b) normalized label. Both images are presented in the coronal view.	66
5.7	Gradient features extraction. (a) Gradient magnitude image obtained from Figure 5.6(a). (b) and (c) correspond to the eroded label letting out borders.	68
6.1	Distribution of the three clinical qualitative information for groups recurrent and stable. <i>Pos</i> - Immunohistopositivity for one or more hormone type; <i>Neg</i> - Immunonegativity for all the tested hormone. <i>Inva</i> - Any level of invasiveness observed second to Hardy classification; <i>Non - Inva</i> - absence of invasiveness in Hardy classification. <i>M</i> - Male, <i>F</i> - Female.	78
6.2	Outcome groups comparison second 'Age-first Surgery'.	84
6.3	Outcome groups comparison second to each one of the relevant features in the univariate analysis.	85

6.4	ROC curves analysis for each one of the relevant features present in Table 6.9.	88
6.5	kNN confusion matrix and ROC curve analysis. ROC - Receiver operator characteristic.	90
6.6	kNN confusion matrix and ROC curve analysis.	91
6.7	RFC confusion matrix and ROC curve analysis.	92

LIST OF TABLES

5.1	Image protocol specifications.	60
5.2	Summary of 48 quantitative textural features extracted with the build CLI module.	69
5.3	Summary of statistical tests used in the univariate analysis.	71
6.1	Age at first surgery and average period follow-up for both patient group with standard error (SE).	77
6.2	Averaged histogram features extracted for both groups with standard error (SE).	79
6.3	Averaged GLCM features extracted for both groups with standard error (SE).	80
6.4	Averaged GLRM features extracted for both groups with standard error (SE).	81
6.5	Averaged shape label features extracted for both groups with standard error (SE).	82
6.6	Averaged gradient features extracted for both groups with standard error (SE).	82
6.7	Association between immunohistochemistry and invasiveness status with PSTR by Fisher's exact test.	83
6.8	Age-first Surgery significance. CI: Confidence interval. tt: Student's t-test. KW: Kruskal-Wallis test.	84
6.9	Relevant features found in the univariate analysis. CI: Confidence interval. tt: Student's t-test. KW: Kruskal-Wallis p-value test.	86

6.10 Summary of most relevant features found in the univariate analysis.
CI: Confidence interval. tt: Student’s t-test. KW: Kruskal-Wallis
p-value test. AUC - area under the receiver operator characteristic
curve. 89

6.11 Best subset for kNN and performance indexes. 90

6.12 Best subset for DTC and performance indexes. 91

6.13 Best subset for RFC and performance indexes. 92

LIST OF ABBREVIATIONS

ACC	Accuracy.
ACTH	Adrenocorticotrophic hormone.
ADH	Antidiuretic hormone.
AUC	Area under the ROC curve.
CAD	Computer aided diagnosis.
CLI	Command line interface.
CT	Computational tomography .
DCE-MRI	Dynamic contrast enhanced MRI.
DTC	Decision tree classifier.
F1-S	F1-Score.
FFCLRP	Faculdade de Filosofia Ciências e Letras de Ribeirão Preto.
FID	Free induction decay.
FN	False negative.
FP	False positive.
FPA	Clinically functioning pituitary adenoma.
FPR	false positive rate.
FS	Feature selection.

FSH	Follicle-stimulating hormone.
GH	Growth hormone.
GLCM	Grey-level co-occurrence matrix.
GLN	Grey-level nonuniformity.
GLRM	Grey-level run-length matrix.
HCFMRP	Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto.
HGRE	High grey-level run emphasis.
KFCV	K-fold cross-validation.
kNN	K-nearest neighbor.
LGRE	Low grey-level run emphasis.
LH	Luteinizing hormone.
LOOCV	Leave-one-out cross-validation.
LRE	Long run emphasis.
LRHGE	Long run high grey-level emphasis.
LRLGE	Long run low grey-level emphasis.
ML	Machine learning.
MRI	Magnetic resonance imaging.
Neg	Negative immunohistochemical studies.
NFPA	Non-functioning pituitary adenoma.
NMR	Nuclear magnetic resonance.
Pos	Positive immunohistochemical studies.
PRE	Precision.

PRL	Prolactin.
PSTR	Postoperative short-term recurrence.
RFC	Random forest classifier.
RFP	Radio-frequency pulse.
RGB	Red, green, blue.
RLN	Run length nonuniformity.
ROC	Receiver operator characteristic.
ROI	Region of interest.
RP	Run percentage.
RT	Radiotherapy.
SE	Standard error.
SEN	Sensitivity.
SPE	Specificity.
SRE	Short run emphasis.
SRHGE	Short run high grey-level emphasis.
SRLGE	Short run low grey-level emphasis.
T1	Longitudinal relaxation time.
T2	Transverse relaxation time.
TN	True negative.
TP	True positive.
TPR	True positive rate.
TSH	Thyroid-stimulating hormone.

US Ultrasound.

CONTENTS

List of Figures	ix
List of Tables	xiv
List of Abbreviations	xvi
1 Pituitary Adenoma: pituitary gland, tumor, and treatment.	4
1.1 The Gland	4
1.2 Pituitary tumors: the tumor spectrum	5
1.3 Hardy and Knosp Classification for Invasiveness	7
1.4 Aggressiveness and Malignancy assessment in pituitary tumors	9
1.5 Symptoms, treatment options, and possible outcomes	9
1.5.1 Medical treatment	11
1.5.2 Surgery	12
1.5.3 Radiotherapy	12
1.6 Our goal: Prediction of postoperative short-term recurrence in NFPA	13
2 Medical Imaging - Magnetic Resonance Imaging	15
2.1 Magnetic resonance imaging	17
2.1.1 MRI principle	17
2.1.2 Imaging	19
2.1.3 Relaxation Times	19
2.1.4 Contrast modalities	21
2.2 T1-weighted dynamic contrast enhanced	21
2.2.1 DCE-MRI for imaging pituitary tumors	22
2.3 Considerations in MRI intensity variability	22

2.3.0.1	Bias field correction	23
2.3.1	Intensity normalization for group analysis	24
3	A quantitative image analysis: Digital image texture analysis	26
3.1	Practical concepts on digital image texture analysis	27
3.2	Histogram features	28
3.3	Haralick's descriptors: the co-occurrence matrix	30
3.4	Run length matrix	35
3.5	Gradient features	38
3.6	Shape features	40
4	Statistical approach: Univariate and multivariate analysis	42
4.1	Univariate analysis	42
4.1.1	Student t-test and Kruskal-Wallis test	43
4.1.2	Receiver operator characteristic	44
4.2	Multivariate analysis: Machine learning	45
4.2.1	How do machine learning algorithms work?	46
4.2.2	Learning modalities and approaches	47
4.2.3	Some classification algorithms	49
4.2.3.1	k-nearest neighbor	49
4.2.3.2	Decision tree	50
4.2.3.3	Random forest	52
4.2.4	Preprocessing steps	53
4.2.4.1	Feature selection	53
4.2.4.2	Dataset stratification	54
4.2.5	Metrics	55
5	Study design, image and data analyses	58
5.1	Study design	58
5.1.1	Including and excluding criteria	59
5.1.2	Studied phenomena	60
5.2	Images used	60
5.2.1	Tumor segmentation	61

5.3	Image preprocessing and feature extraction software	62
5.3.1	Feature extraction plugin implementation	63
5.3.1.1	Clinical features	68
5.4	Data analysis	70
5.4.1	Univariate analysis	70
5.4.2	Multivariate analysis: machine learning	71
5.4.2.1	Standardization	74
5.4.2.2	Feature selection strategy	74
5.4.2.3	Metrics	75
6	Results and discussion	76
6.1	Clinical and textural features for both groups	76
6.2	Univariate analysis	83
6.2.1	Imaging texture feature analyses	84
6.3	Multivariate analysis	89
6.3.1	kNN	90
6.3.2	DTC	91
6.3.3	RFC	92
6.4	Closure discussion	92
7	Conclusions and final considerations	95
	Bibliography	97
	Annex A - Ethics committees' approval letters	106
A.1	HCFMRP ethics committee's approval letter	107
A.2	FFCLRP ethics committee's approval letter	111

INTRODUCTION

CANCER treatment is definitely one of the greatest challenges of medicine and correlated areas. The complexity is associated with the fact that it is not a single disease but a spectrum of diseases that affects practically every organ or system in the human body. It is majorly identified as a neoplasm that has roots on genetic inheritance, probabilistic cellular reproduction errors, and external agents.

Cancer management always focuses on cure. However, cure has different definitions depending on the cancer type. In some cases in which cure is not systematically achievable, the ultimate goal turns to be tumor control, overall survival, and symptoms relief. The difficulties faced when handling cancer are related to tumor resistance to treatments available, recurrence, aggressiveness, and malignancy.

Pituitary tumors correspond to 10% to 20% of all intracranial tumors documented. They are usually benign and are included in the group for which cure is not always systematically achieved. This implies that every patient needs to be maintained under surveillance before and after treatments. And some studies describe follow-up periods for those patients of up to 20 years after the first approach, mostly surgical.

For any cancer type, and this includes pituitary tumors, the effectiveness of the treatment is closely related to the quality of prognostic information available. It means, knowing which type of tumor, if it is resistant to any of the available treatments, if it is aggressive, and the chances of post treatment regrowth, gives to the specialists a step ahead when making diagnostic and treatment decisions.

In the treatment of pituitary tumors, the hormonal activity status plays an important role in the treatment decision-making process. The class of pituitary adenomas known as non-functioning pituitary adenomas correspond to one-third of

all pituitary adenomas. The main problems caused by this tumor type are hormonal hyposecretion and visual disturbances as a mass effect consequence. It is firstly treated with surgery, but almost half of the cases it exhibit tumor regrowth, what requires a secondary approach, and this process may happen over and over again.

Radiotherapy (RT) is used to tackle many types of cancer diseases, and for non-functioning pituitary adenomas it offers control rates around 90%, and for that reason, it was once used as a "must do" postoperative approach for non-functioning adenomas management. Nonetheless, secondary lesions were noticed to arise after RT procedures. Now, a more consistent evaluation needs to be considered in order to make the trade-off between assuming risks of secondary lesions and better chances of tumor control.

Many studies have been done willing to find good prognostic factors for predicting tumor regrowth for non-functioning pituitary adenomas. However, up to the moment, despite some evidence, no prediction factor (clinical or histopathological features) has convincingly achieved this goal.

Texture analysis (TA) is an ongoing field that intends to extract computationally quantitative features from digital images. This technique has been used in many cancer studies, such as malignancy assessment, tumor type classification, tissue characterization, and many others narrowly defined clinical questions based on magnetic resonance imaging analysis. The combination of texture analysis with machine learning prediction models led to the field known as computer aided diagnosis. In addition, many studies have provided reliable TA-based tools to be used in the clinical setting as a second "opinion" in the treatment decision process helping the specialists.

Considering that, to the best of our knowledge, no texture analysis study has been done to investigate the prediction value of texture analysis in pituitary tumors, the following question was proposed: Is it possible to use MRI texture features as image biomarkers for short term recurrence prediction in non-functioning pituitary adenomas?

The present study was carried out to investigate this question. This work is organized in seven chapters: The second chapter brings a brief review on pituitary tumor, its subtypes, treatment options, and possible outcomes; The

third chapter brings a basic introduction to medical images focusing on magnetic resonance imaging and the protocols used to assess pituitary adenomas; The fourth chapter details the main concepts and methodologies underlying digital image texture analysis; The fifth chapter introduces statistical approaches (univariate and multivariate) to investigate correlation, association, and prediction power of clinical and computationally extracted image features. The sixth chapter describes meticulously how the study was done, from patient including criteria up to the final statistical tests. The seventh chapter presents the results obtained from both univariate conventional statistical tests and multivariate machine learning prediction models. And finally, the eighth chapter brings the closure conclusions and considerations about the study here presented.

PITUITARY ADENOMA: PITUITARY GLAND, TUMOR, AND TREATMENT.

IN order to discuss how this study can improve the treatment of pituitary adenomas through imaging analysis and to comprehend the complexity of the phenomena studied here, an introduction to the the principal component of the endocrine system, the pituitary gland, the tumor types, the diseases related to them, and the treatment scenario will be given .

1.1 The Gland

Pituitary gland, or hypophysis gland, sizes approximately same as a pea and is localized in sella turcica, next to optic chiasm, and right bellow hypothalamus, which controls its activity. It is also known as the "master" gland for managing other important glands in our endocrine system such as thyroid gland [1]. Fig. 1.1 illustrates its form and surrounding structures.

Pituitary gland is subdivided in two structures (Fig. 1.1), each one with a specific secretory activity. Anterior lobe, called adenohypophysis, is responsible for secretion of prolactin (PRL), growth hormone (GH), follicle-stimulating hormone (FSH), and luteinizing hormone (LH). Posterior lobe, called neurohypophysis, produces oxytocin and antidiuretic hormone (ADH) [2].

Given its importance and functionality, many diseases such as amenorrhea, galactorrhea, decreased libido, sexual impotence, gigantism, hypogonadism, hypothyroidism, hypoadrenalism, and panhypopituitarism, may arise when this

Figure 1.1: *Structure of pituitary gland. Sella turcica is a bone cavity that encapsulates pituitary gland. The optic chiasm is a very sensitive nerve that connects the brain to the eyes and is easily affected by large pituitary lesions. For the sake of readability, the gland was illustrated bigger. Adapted from [2].*

gland is not functioning properly, affecting men and women. Those diseases are mostly related to pituitary gland hyper or hyposecretion of one or more hormone types [3].

1.2 Pituitary tumors: the tumor spectrum

There is a spectrum of tumors that may arise in the pituitary gland. But in general, they are primary *adenomas* (benign tumors) that develop on the anterior lobe of hypophysis gland. Other types of primary tumors or even metastases are very rare [4].

The classification of these tumors is multiparametric given the variety of criteria to specify the disease. This includes specifying the presence or absence of secretory activity, hormone type, lesion dimensions, and histological characteristics. It is important to state that a standard classification has not been yet established in the scientific community [5, 6].

Regarding the size, pituitary adenoma can be classified as *microadenoma*, when its biggest diameter is less than 10 mm; and *macroadenoma*, when its biggest diameter is equal or greater than 10 mm. Tumors with dimensions bigger than 30 mm, or volume greater than 10 mm³, are described as *giant adenomas*¹ [1, 3, 5].

¹Some studies, as [7], consider giant adenomas tumors with diameters bigger than 40 mm. There is no consensus on the definition of giant adenomas [5]

Macroadenomas are those responsible for mass effect, that is, the symptoms caused by the compression of surrounding areas, such as headache, visual field deficits, and endocrine deficiencies, that is, hormonal hyposecretion.

(a) Microadenoma.

(b) Macroadenoma.

(c) Giant adenoma.

Figure 1.2: *Examples of three different size pituitary adenomas. (a), (b) and (c) bring a sagittal slice of a T1-weighted MR image. Adapted from [2].*

Based on the presence or absence of hormonal activity, pituitary adenomas

are classified as *clinically functioning* (FPA) or *clinically non-functioning* (NFPA). The classification based in hypersecretion is better understood by presenting the disease, its hormonal activity, and some of its aspects:

- *Prolactinoma* - prolactin hypersecretion, presence of mass effect, chances of hypopituitarism, most common type of pituitary tumor;
- *Acromegaly* - GH hypersecretion, mass effect, gigantism (if occurs in childhood);
- *Cushing's Disease* - hypersecretion of adrenocorticotrophic hormone (ACTH), usually are microadenoma;
- *TSH-secreting adenoma* - Rarely diagnosed, secretion of one or more types of hormones including thyroid-stimulating hormone (TSH), GH, and some others like FSH, and LH;
- *Mixed growth hormone/prolactin* - Hypersecretion of both PRL and GH, mass effect;
- *Gonadotropin-secreting adenoma* - Hypersecretion of FSH, LH, and presence of non-functioning sellar mass;
- *Nelson's Syndrome* - similar to Cushing's disease, but with the presence of mass effect, aggressiveness.

Summarized from [1, 2].

1.3 Hardy and Knosp Classification for Invasiveness

Over 25 to 55% of pituitary adenomas are considered invasive tumors ² [5]. Regarding tumor mass and geometry, Hardy and Knosp proposed a classification to grade how invasive such a tumor might be considering some surrounding structures. Figure 1.3 describes Hardy-Knosp classification.

In Figure 1.3 Hardy specifies five grades or levels to describe tumor invasiveness. The first three grades (0, I and II) are considered noninvasive

²This rate varies across many studies.

Figure 1.3: *Hardy-Knosp classification for invasiveness. Hardy grades indicate both invasiveness and symmetry level of tumor mass. Knosp only evaluates invasiveness. Both authors consider the proximity of the tumor mass from surrounding structures to evaluate invasiveness. Adapted from [5].*

for they happen to be regularly placed in sella turcica. Grades III and IV are considered invasive for it is observed the presence of focal and extensive bone erosion, sellar destruction. The geometrical distribution of tumor mass appears in Hardy classification as symmetric (A, B, and C) and asymmetric (D and E). Knosp classification system treats cavernous sinus as a reference for invasiveness. Tumors in grade 3 or 4 are then considered invasive since they are found within cavernous sinus or crossing the hypothetical line right at the end of sphenoid sinus in Figure 1.3 (B). Besides this image based classification (CT or MRI), histological tests can gauge the invasiveness of the tumor by assessing the invasion of the mucosa of the sphenoid sinus [5, 8].

1.4 Aggressiveness and Malignancy assessment in pituitary tumors

Invasiveness is not a definitive criterion for aggressiveness. Aggressive pituitary adenomas are rare with no incidence percentage defined [6, 9, 10]. They may be invasive or not, microadenoma or macroadenoma, secreting or non-secreting. This diagnose takes into account a high rate of recurrence, lack of response to therapy, rapid growth. Additionally to those features, histological analyses play an important role identifying the levels of mitotic activity and immunoreactivity of gene p53. Atypical findings on that analyses, high mitotic activity or p53 immunopositivity, also point to tumor aggressiveness [5, 10].

Malignant pituitary tumors correspond to roughly 1% of all tumors developed in hypophysis gland ³ [2]. The strongest feature of those carcinomas is its spreading pattern to distant sites of the body, lymph nodes and liver, through the central nervous system, in other words, a metastatic behavior. The survival expectation when a malignant pituitary tumor is diagnosed is one year in over 66% of the cases [2, 10].

1.5 Symptoms, treatment options, and possible outcomes

As it has been described so far, pituitary tumors correspond to a wide variety of tumors with very specific features and behaviors concerning their hormonal activity, growth pattern, histological aspects, and metastasis presence. Diagnosis and treatment decisions for patients harboring those neoplasms is straightforwardly related to observed symptoms, clinical examinations, and tumor classification as it was seen in Section 1.2.

The main symptoms that occur for patients with this tumor arise from mass effect (headache, visual field loss, hemianopsia) and or hormonal dysfunction (hyper or hypo hormone secretion). Once there is some suspicion of hypophysis tumor a detailed and multidisciplinary clinical investigation is done ⁴. History and physical

³Some early studies like [9] show different percentages

⁴This multidisciplinary group is composed by general physician, endocrinologist, radiologist,

examination (including neuro-ophthalmic examination), imaging studies (enhanced MRI with gadolinium (footnote: high-resolution CT can be used when MRI is not available or contraindicated)), blood and urine tests, and endocrine tests compose the set of clinical information for diagnostic [2].

When dealing with those tumors, treatments intent primarily to diminish mass effect, for tumors appearing in overgrown stage, and to normalize hormonal overproduction, for clinically functioning tumors. To reach those goals, there are three treatment baseline options which are *medical treatment*, *surgery*, and *radiotherapy*, which might be used separately or in a combination of modalities. The schemes in Figure 1.4 and 1.5 illustrate the decision-making process for both functioning and non-functioning pituitary adenomas.

Figure 1.4: Decision-making process for FUNCTIONING pituitary adenomas. The option for medical treatment is considered for prolactinomas [2, 11]

Figure 1.5: *Decision-making process for NON-FUNCTIONING pituitary adenomas.*

1.5.1 Medical treatment

Medical Treatment is the other option for initial treatment. It is a common option (not always the first) for secretory tumors and the drug of choice depends on the related disease. Its goal is tumor size reduction, mass effect relief, and hormonal hypersecretion [2]. The most common functioning tumor type is prolactinoma. The second most common is growth-hormone secretory tumors that leads to acromegaly in adults and gigantism when occurs in children. Following comes ACTH-secreting tumor which causes Cushing's disease and sometimes leads to Nelson's Syndrome. [1].

Medical treatment is the first option for patients harboring prolactinoma with 90% of those subjects experiencing hormone normalization and size reduction. It is

commonly taken as salvage treatment (with 50% to 80% of control) in subjects with acromegaly who did not achieve biochemical cure through surgery. For Cushing's disease, clinical treatment may be an adjuvant therapy alongside surgery or radiotherapy [1].

1.5.2 Surgery

Surgery intends to immediately solve tumor mass effect (compression of surrounding structures and visual field limitations) and to provide rapid relief for hormone hypersecretion by removing as much neoplastic tissue as possible. The goal is to remove the tumor completely and, in some cases, the majority of pituitary gland itself is removed [1, 2]. *Transsphenoidal* approach is the most common surgical resection technique used for decompressing parasellar and suprasellar structures. For those subjects non-eligible for transsphenoidal surgery, *transcranial* approaches may reach the same goals.

Around 90% of subjects who undergo this procedure experience gains in visual field issues. Surgical intervention is the first choice for symptomatic and hormonally inactive tumors (NFPA) and the second option for those treated initially by medical or radiotherapeutic treatment with failure [1, 12]. Long-term tumor control by surgery alone may be achieved in 50% to 90% of cases ⁵. Roughly 16% of subjects can, up to 10 years after surgical intervention, experience recurrent tumor growth [1, 2].

1.5.3 Radiotherapy

Radiotherapy (RT) is a powerful treatment option pointing up to 85-95% of tumor growth control at 5-10 years of follow-up [1, 14, 15, 16]. The hypersecretion control rate can reach 80% for acromegaly and Cushing's disease, and more than 30% for subjects with prolactinoma [2, 15]. Although radiation therapy has high success rate, it offers some long-term side effect risks to patients, being hypopituitarism the most common one [12, 14, 17, 18, 19, 20, 21]. Up to 56% of patients develop this syndrome and, in some other rare cases, brain necrosis, deterioration of vision,

⁵Some studies, like the series analyzed in the southeast of Brazil [13], shows an inferior tumor control rate.

vascular changes, dementia, and even secondary malignancies were observed after RT treatment [14, 15, 22, 23, 24]. For this reason, this treatment is not routinely indicated as it formerly was [16].

RT is mainly recommended for surgically non-eligible subjects, for invasive tumors, and for recurrent tumors when there is a potential risk of visual damage, but yet asymptomatic [2, 12, 16]. It is also recommended for remaining lesions after incomplete resection observed in short-term radiological studies. For instance, when MRI at 6-12 months post operation demonstrates residual tumor [2, 16]. Finally, RT may be recommended as adjuvant treatment after complete resection of clinically functioning tumors with later detected persistent hormone hypersecretion, or when drug therapy is not recommended or has failed to solve hypersecretion [2, 15, 16].

Radiotherapy may be approached in a fractionated RT, stereotactic fractionated RT, or even radiosurgery for small tumors [2]. The efficiency of this treatment is based on overall survival, progression free survival (local tumor control), and subject's quality of life after treatment [16]. In this way, a long-term follow up of patients who undergo any modality of RT, alone or in combination with surgery, is necessary [24].

1.6 Our goal: Prediction of postoperative short-term recurrence in NFPA

NFPA are those tumors that show no hormonal activity. They are also said to be silent adenomas, and for that reason they are hard to be detected. This type is usually diagnosed as macroadenoma (15-30%). Diagnosis often happens incidentally through MRI exams or by investigating simpler symptoms like headache and visual problems that an overgrown tumors normally exhibit [1, 2, 25, 22].

Despite the name, it is known, since the early 80s, that most of those NFPA actually do secrete some hormone (e.g. gonadotropins). Some may even show positive immunolabeling for GH, PRL, TSH, and ACTH, but as this is not detected in vivo, they are said to be silent adenomas and included in the clinically non-functioning group [25].

One-third of pituitary adenomas are NFPA [22]. In some studies though, the

proportion of NFPA to FPA is higher than unity [23, 26].

As depicted in Figure 1.5, the mainstream first treatment option for this tumor type is transsphenoidal surgery focusing on reducing visual loss and compression effects. Nevertheless, 12-46% of patients have reported tumor progression after surgery [22]. As already stated (Section 1.5.3), fractionated RT is an extremely powerful treatment option for tumor control, but side effects make it a risky step to carry out in NFPA treatment. Therefore, postoperative RT needs to be carefully evaluated.

Another adjuvant or secondary treatment option is the stereotactic radiosurgery. It has shown tumor control rates higher than 90% and, despite long-term follow-up effects having not yet been fully detailed, lower side effect risks have been reported [22].

In fact, the decision for using stereotactic radiosurgery or even fractionated RT to achieve progression-free survival would be better taken considering the risk-benefit trade-off for the patient. However, NFPA natural history is not yet fully comprehended. In addition, prognostic factors for tumor regrowth are not well defined despite some studies having evidenced some correlation with invasiveness, gender, and histology results [11, 22].

The present work investigated the use of magnetic resonance image quantitative parameters (texture, intensity, and morphological features) as biomarkers to predict short-term recurrence in NFPA treatment. The following chapters will describe basic principles of magnetic resonance imaging and its advantage in the follow-up process of pre and postoperative NFPA, the founding concepts of digital image texture analysis, and statistical models to evaluate the prognostic power of such data.

MEDICAL IMAGING - MAGNETIC RESONANCE IMAGING

UR visual system, when seeing some object, detects its appearance and aspect subjectively, and instantly makes classifications about that object as a person, an animal, scene, or something else. Similarly, through our hands more information about such a being can be collected if physical contact is allowed. At the end, many features about this object are collected and based on those features the brain makes classifications, conclusions about the nature of that object and makes decisions regarding it. As testified by our evolutionary success, the eye-brain junction is the best system so far (and will likely always be) for classifying and making decisions based on images [27].

On the other hand, many other systems that produce images that can be saved into a file and interpreted by a computer have become very popular. They range from low-cost smartphones and web-cams to super expensive X-ray scans, magnetic resonance scanners, and satellites. The images produced by those systems are called digital images. The popularity and accessibility of those systems and tools (or at least some) allowed the production of an infinite set of digital images in a variety of fields. Industry, media, science and medicine, to say a few. Coupled with that phenomena, image processing and image analysis have grown widely to enhance the utility, manipulation, storage, and quality of those images [27].

This work places itself in the use of images in medicine. Nowadays, the use of images in the regular routine of diagnostic and treatment of patients with many diseases is a fact [28]. Radiology is the field of medicine that is concerned with

making diagnoses regarding inner parts of patients based on the interpretation of images. This field has its roots in Rontgen's discovery in 1895, when X-rays were first produced and detected, and, alongside it, what is considered the very first medical image from inside the body (Figure 2.1). Many other imaging techniques have been developed since then, *Ultrasound* (US), *Computed Tomography scanners* (CT), *Magnetic Resonance Imaging Scans* (MRI), and many others, which have become each time more available for clinical use, consolidating the power and clinical importance of imaging studies.

Figure 2.1: Radiography of the left hand of Wilhelm Rontgen's wife on December 22, 1895, shortly after his discovery of X-rays. Considered to be the first medical image ever produced and the kick-start of a revolution in medicine ¹.

Oncology, the branch of medicine dedicated to prevent, diagnose and treat cancer diseases, has also been positively affected with the development and growth

¹Source: <http://www.the-scientist.com/?articles.view/articleNo/30693/title/The-First-X-ray--1895/>

of (invasive and non-invasive) image modalities. But, of course, non-invasive and non-ionizing radiation based techniques stand out in that scenario.

Following these criteria, magnetic resonance imaging has increasingly grown in use considering that regular acquisition of MR images has become a mainstream step of treatment follow up for cancer disease patients. And for those reasons, studies that enhance and optimize the information extraction process from MRI by clinicians may improve the diagnose performance and precision of the treatments with low cost, since it is already there [29, 30].

2.1 Magnetic resonance imaging

Since its beginnings in 1973 with Paul C. Lauterbur and Peter Mansfield, magnetic resonance imaging (MRI) has greatly evolved becoming an extremely powerful non-invasive imaging technique, which uses non-ionizing radiation (magnetic field), with high flexibility and specificity for a wide variety of tissue properties. Its power goes from tissue discrimination capability to body structure and function understanding [31].

2.1.1 MRI principle

The contrasts found in MRI may be obtained checking different properties and aspects of the same essential principle: the nuclear magnetic resonance (NMR). This is the underlying phenomena in MRI. It consists in the interaction between an external radio-frequency pulse ($\vec{B}_1$) that temporally changes the spin orientation of the nucleus (mostly the hydrogen's ones) of a tissue molecule initially oriented by a constant magnetic field ($\vec{B}_0$).

To introduce this nuclear MR phenomenon a semi-classical model has been successfully used in the literature. Given its spin vector and the fact that it is charged, the hydrogen nucleus (which is a proton) has a magnetic dipole moment $\vec{\mu}$. The magnetic moment and spin of a proton point in the same direction, and in MRI jargon, the terms "spin" and "magnetic moment" are commonly interchanged. This proton spin, once immersed in a strong and constant magnetic field ($\vec{B}_0$), will "precess" around $\vec{B}_0$ direction (Figure 2.2), moved by the torque generated by the

interaction between $\vec{\mu}$ and $\vec{B}_0$ [31]

Figure 2.2: Classical illustration of magnetic dipole moment vector (μ) around $\vec{B}_0$ direction. Adapted from [31]

This precession frequency is called Larmor frequency and is given by

$$\omega_0 = \gamma B_0. \quad (2.1)$$

Where γ is a constant called gyromagnetic ratio. γ has different values for different nuclei. In particular, for the hydrogen nuclei, its value is $2,67.10^8 \text{ rad}/(s.T)$, or, $\gamma/2\pi$ is $42,7 \text{ MHz}/T$. Therefore, in clinical MRI systems, which have magnetic field strengths in the range of a few Teslas, the Larmor frequencies are in the range of MHz (that is, radiofrequency).

The described system (a magnetic dipole vector precessing around $\vec{B}_0$ axis) is "available" to interact with external fields. If an external field ($\vec{B}_1$) oscillating at the Larmor frequency is applied for a short period, a radio-frequency pulse (RFP), it can "change" the orientation of the precession axis of the magnetic dipole vector ($\vec{\mu}$). As soon as the transitory field ($\vec{B}_1$) is off, the magnetic dipole vector recovers back to the original position precessing around $\vec{B}_0$ axis.

This variation in the magnetic spin orientation generates a *magnetic field flux* that can be detected with a coil according to Faraday's law of induction.

The detected signal has a sinusoidal appearance and diminishes from its maximum intensity, right after the RFP is off, to zero when the magnetic spin vector is fully back to its original orientation. This detected response signal is called *Free Induction Decay* (FID).

2.1.2 Imaging

NMR is the underlying principle in the image formation in MR scans. But instead of considering a single proton spin as in the above explanation, the MR imaging process considers tissue samples of millimeters order. Those samples have, instead of a single dipole moment, a magnetization vector ($\vec{M}$).

Everything works pretty much the same if a whole object is imaged. Nevertheless, counting uniquely with the static field ($\vec{B}_0$) the whole object would have the same Larmor frequency, would answer similarly to the RFP and would respond as a single resultant FID. To make the imaging process feasible, a linearly variant magnetic component is added through a gradient component coil in $\vec{B}_0$ direction. This "slicing" process separates the object spatially in spectral components in $\vec{B}_0$ direction (z). Additional gradient coils are added in the other two directions perpendicular to $\vec{B}_0$ (x, y) to complete the spatial and spectral correlation [31].

The imaging construction is finished with the transformation of spectral space information separated with the gradient field configuration to the ordinary space. This is accomplished elegantly by using the Fourier transform [31].

2.1.3 Relaxation Times

Back to the idea of the sample total magnetization ($\vec{M}$) in the presence of the constant field $\vec{B}_0$, it is important to stress that $\vec{M}$ has two components (or can be decomposed into two components): one in $\vec{B}_0$ direction (z), the longitudinal magnetization, and one in the perpendicular direction (x, y plane), the transverse magnetization. The longitudinal magnetization has to do with spin-lattice effect, which is the interaction between a given spin and the lattice of atoms around it. Transverse magnetization has to do with spin-spin effect, which is the synchronization of the spin precession phase around $\vec{B}_0$ direction.

When the RFP is applied over a sample its total magnetization can be flipped (of a given angle, i.e. $\pi/2$ or π) depending on the pulse duration and strength (Figure 2.3). In a $\pi/2$ pulse, $\vec{M}_z$ goes to zero and $\vec{M}_{xy}$ to its maximum value. The rate that expresses how fast the total magnetization $\vec{M}$ of a sample grows back to its original orientation and intensity after RFP is off is a tissue property. The regrowth rate for $\vec{M}_z$ is called *Longitudinal Relaxation Time* ($1/T_1$). And the decay rate for $\vec{M}_{xy}$ is called *Transverse Relaxation Time* ($1/T_2$).

Figure 2.3: Sample excitation by an RFP. a), b) and c) illustrate what happens with the individual magnetic dipole vectors ($\vec{\mu}$) in the sample after a $\pi/2$ RFP. d), e) and f) illustrate what happens with the total magnetization $\vec{M}$. In a) and d) the total magnetization $\vec{M}$ is Longitudinal. In b) and e), after the RFP, the total magnetization $\vec{M}$ is transversal. Adapted from [31]

The following equations define the exponential recovery and decay process of $\vec{M}_z$ and $\vec{M}_{xy}$ components respectively:

$$M_z(t) = M_0(1 - e^{-t/T_1}), \quad (2.2)$$

$$M_{xy}(t) = M_0(e^{-t/T_2}). \quad (2.3)$$

Many variations of the Equations 2.2 and 2.3 are proposed as *pulse sequences*. Those sequences are defined as a combination of basic pulses (i.e. $\pi/2$ or π) in order to manipulate the signal and reach different contrasts. Two specific time intervals are regularly defined: time to echo (TE) and time to repetition (TR).

2.1.4 Contrast modalities

MRI is sensible to proton density, relaxation times, pulse sequences, temperature, proton motion, tissue heterogeneity, and to some biochemical phenomena that produce changes in the final image signal produced. The process of choosing one or other tissue property to be imaged is done by choosing not a single excitation RFP but rather a pulse sequence. In this way, different MRI contrast modalities may be accomplished by changing or adding some parameters in the pulse sequence or even by using some contrast enhancement agent [31, 29].

Given that versatility, MRI can capture extremely rich anatomical and functional information about the tissue or organ imaged. In addition to that, the advent of new physiology-based MR imaging modalities such as *diffusion tensor imaging*, *diffusion weighted imaging*, *perfusion imaging*, and *MR spectroscopy*, and the use of *gadolinium contrast agent*, in the modality called *Dynamic Contrast Enhanced MRI* (DCE-MRI), has brought MRI to the mainstream in the assessment and treatment follow up of brain lesions. The gains include precise lesion evaluation, more accurate tumor type assessment, surgical biopsy guidance, and treatment follow-up feasibility [29, 32].

For the purposes of this work, it will be given special attention to the T1-weighted DCE-MR image modality, focusing on pituitary tumor imaging gains, and no further explanation on other MRI contrast modalities will be provided despite their vast applicability to many other clinical settings.

2.2 T1-weighted dynamic contrast enhanced

The contrast observed in Dynamic Contrast enhanced MRI is obtained by the administration of pharmacokinetic tracers with bolus injection, mostly paramagnetic gadolinium-based, and relies pretty much on the tissue (e.g. tumor lesion) vasculature. Usually, the contrast agent dose administered is 0.1 mmol/kg , 4 ml/s . And the imaging protocol corresponds to an anatomical T1 weighted acquisition ² before contrast administration followed by a postcontrast acquisition [32].

²T2 or T2* weighted is used in susceptibility-based techniques [33]

The intensity of the resultant contrast is directly related to the regional blood flow, blood vessel density, hematocrit variations, vascularization of existing blood vessels, and surface area permeability of the endothelial membrane. The utility of DCE-MRI includes better tumor visualization, complex quantification of tumor microvasculature, analysis of tumor angiogenesis, "in vivo" tumor staging, and even malignancy assessment based on signal intensity curves (e.g. breast cancer [34, 35]) [32, 33].

2.2.1 DCE-MRI for imaging pituitary tumors

Given its power for separating soft tissues with good quality, MRI has become the mainstream imaging technique for evaluating sellar and parasellar lesions. More specifically, DCE-MRI is considered the best protocol for imaging pituitary adenomas [36, 37].

Gadolinium-based contrasts are the chosen ones and are administered in doses of 0.05 mmol/kg. The peak of enhancement occurs around 60 to 200 seconds after dose administration. This is, in fact, the best window to image the lesion. DCE-MRI acquisition protocol is composed of a regular T1-weighted image before contrast (Figure 2.4(a)) followed by a new acquisition post-contrast (Figure 2.4(b)) that allows physicians to clearly detect the lesion [36].

Looking at Figure 2.4(b) it is easy to detect the lesion and define its borders, even for a nonspecialist observer. For that reason, DCE-MRI is the protocol of choice for visual inspection of pituitary lesions, to detect both micro and macro adenomas, differentiate residual from recurrent lesions, and to evaluate invasion of the cavernous sinus [36, 37].

2.3 Considerations in MRI intensity variability

MR Imaging is a very complex imaging process and is unintentionally affected by many factors such as room temperature, scanner thermal noise, static field inhomogeneities, patient motion artifacts, analogical-digital/digital-analogical signal conversion, and coil calibration, causing it to suffer from intra and inter-patient variabilities, as well as across manufacturers [38, 39, 40].

(a) T1-weighted pre-contrast

(b) T1-weighted post-contrast

Figure 2.4: *NFPA* imaged before (a) and after (b) gadolinium contrast. The difference in tumor delineation is easily noticed.

With that in mind, any MRI data analyses must consider doing some procedures that correct or diminish such a variability. Some procedures are very well accepted and practiced in a wide variety of image analysis studies.

2.3.0.1 Bias field correction

N4 bias field correction algorithms, as well as its former version N3 (nonparametric nonuniform intensity normalization), is used to correct intensity nonuniformities (illustrated in Figure 2.5) associated with poor radio frequency coil uniformity, gradient-driven eddy currents, and patient anatomy [39, 40].

N3-based algorithms intend to restore frequencies distorted by performing

Figure 2.5: Visualization of the field inhomogeneities present in a T1 image. Adapted from [40].

a signal processing approach to the image. The performance of those methods is independent of the pulse sequence used and the presence of pathologies on the imaged tissue. Therefore, it fits pretty well as part of the preprocessing pipeline in an automated data analysis [39, 40].

2.3.1 Intensity normalization for group analysis

In a group study variabilities intra- or inter-patients, and intra or inter-scanners, play a considerable role. It means that grey pixel value in the obtained image will be slightly different if the subject takes the same protocol twice in the same scanner or in different scanners.

Other recommended procedures to diminish this variability are the intensity normalization methods. They work by excluding pixels of the image that lie outside of the main range. Pixels that have intensities lower than $A - 3\sigma$ and higher than $A + 3\sigma$ are excluded, where A is the pixel intensity *average* and σ is the *standard deviation* of the pixel intensity distribution across the image [38].

This preprocessing step is very important in studies such as the one presented here. It was shown to improve performance in machine learning classification tasks based on digital image texture analysis [38].

In the next two chapters, the foundations of digital image texture analysis and how it can be used to improve the decision making-process in the clinical scenario in general, and more specifically in oncology, will be introduced.

A QUANTITATIVE IMAGE ANALYSIS: DIGITAL IMAGE TEXTURE ANALYSIS

TISSUE characterization and abnormality detection are the main goals of any diagnostic imaging investigation. Radiologists are mainly concerned with distinguishing between normal and pathological tissue; grey matter, white matter, and cerebrospinal fluid; soft and solid tissue, and so on. This task is almost at no stage automatized and relies basically on radiologists' personal experience regarding the issue being investigated. Implicitly, the imaging diagnostic process implies in seeing and recognizing some specific patterns and features in the image and connecting such information with possible diagnoses. Despite the subjectivity involved in this process, specialists achieve expressive accuracy and precision in their diagnostic tasks, not easily reproduced by a computer [41, 42].

Those patterns and features sought by specialists in clinical imaging studies are implicitly or explicitly connected to intensity variations, morphological structures, or to image texture. Considering that human eyes have limited range sensibility for texture and image intensity discrimination, many other strategies for image texture classification and analysis have been investigated since the middle 1950s [28, 41].

The probable earliest work regarding image texture analysis dates from 1955 and treated quantification of aerial images. Since then, with image and digital computer popularization, many other techniques and computational approaches were developed for the extraction and analysis of imaging texture features for general purposes [41, 43]. Followed by the achievement of X-rays image digitalization, digital

image texture analyses were firstly carried out for medical applications in the early 1970s. Later on, a large amount of meaningful studies were prompted using other powerful image modalities like MRI and Ultrasound.

3.1 Practical concepts on digital image texture analysis

Computationally speaking, digital images (including MRI) are basically composed of a regular rectangular set (an array, a matrix) of small unities of image called pixels (or called voxels when dealing with three-dimensional images). Each pixel brings within a single intensity value, in the case of grey-level images, or a tuple of three numbers, as in an RGB (red, green, blue) color scheme. Also, such a pixel is identified by an index, a pair of coordinates (three for three-dimension images), and its physical position, in physical units [27, 28, 44]. Figure 3.1 brings an illustration of the representation of a grey-level image in its pixel intensities form.

Figure 3.1: Digital image pixel-wise analysis. It is an 8 – bits image, meaning that pixel intensities vary from 0 to 255 ($2^{bits} - 1$). The ROI highlighted was arbitrarily chosen, but in medical imaging studies they are, in fact, physiological structures of interest. Adapted from [27].

The organization or distribution of such pixels in an image is responsible for the patterns observed subjectively by human eyes. Those patterns are more specifically perceived as morphological structures, brightness or darkness, and

texture features. All those three types of features characterize the object, scene or pattern imaged and are passive of computerized quantification given the pixel distribution present in the image and some mathematical tools [28, 41]. In other words, a texture feature corresponds to a number that characterizes an image or *Region of Interest* (ROI) ¹ that is obtained by a mathematical procedure over this image's pixel distribution, e.g. the mean as in Figure 3.1. Therefore, there are many methods to calculate texture features from a given image that output different numbers.

Texture features are grouped or classified differently by different authors. [45] described texture extraction methods in four groups: *Structural*, *Statistical*, *Model-based*, and *Transform methods*. The structural ones are defined by primitive elements allocated according some placement rule. Statistical features are those acquired checking the statistical distribution that governs grey-level pixels allocation and relationship. They are yet subdivided into first and second-order statistical features. First-order ones are related to regular histogram measurements like mean, standard deviation, and percentiles. Second-order statistical features are those that consider the distribution of pairs of pixels or runs of pixels as those proposed by [43] and [46] respectively. Model-based methods imply in using complex mathematical models such as fractal or stochastic to interpret image texture. Transform methods refer to features obtained by sampling and analyzing the original image into an alternative space using space transforms like Fourier and Wavelet Transform [28, 41, 43, 45, 46, 47].

In the following, some of the most common texture analysis techniques and quantitative parameters extracted from medical images present in the literature, and that are relevant in the context of the present work, are presented.

3.2 Histogram features

Grey-level digital image pixel intensities range from 0 to $2^{bits} - 1$, where *bits* is the number of bits in the hard-drive dedicated to store each image pixel information

¹Region of interest may be defined only as a specific part of an image or as the whole image space. In this work, the term ROI will be used to relate to a whole image or a specific part of the image being analyzed.

². For instance, *bits* = 8 allows intensities from 0 to 255 grey-level tones. And each value is associated with an intensity in the grey-level range, being 0 the darkest and the 255 the brightest one.

MRI images mostly use 12 *bits* pixel-storage resolution. Given this amount of space, every pixel image may assume a grey-level value ranging from 0 to 4096 unities of intensity. Alternatively, a given image is liable of being represented in a shorter resolution using pixel-wise rescaling operations.

Histogram is a first-order statistical approach massively known and used to describe the pixel distribution across a ROI considering each single pixel value (*level* or *intensity*) and the number of times it appears in the image (*frequency*). Figure 3.2 brings an example. Needless to say that histograms are used to represent data distributions in various fields.

Figure 3.2: *Histogram.* It is an 12 – bits image, meaning that pixel intensities vary from 0 to 4095 ($2^{bits} - 1$). ³.

Many parameters may be computed from the histogram. The most popular ones are *mean*, *standard deviation*, *variance*, *minimum*, and *maximum*. Also, *kurtosis* and *skewness* might be measured. Kurtosis is defined as a measure of how (heavy or light) tailed is the distribution related to the normal one. Skewness is a measure of how asymmetric is the distribution (e.g. normal curves have skewness equal to zero). Lastly, *percentile* is the threshold intensity for which the given percentage of pixels has intensity smaller or equal to it. For example, having 25th

²Also known as *radiometric resolution* [27].

³Adapted from: https://matplotlib.org/examples/pylab_examples/mri_with_eeg.html

percentile equal to 47 means that 25% percent of pixels in the image have intensities smaller or equal to 47. 5th, 10th, 25th, 75th, and 90th percentiles are commonly present in the literature. Additionally, *interquartile interval* refers to the intensity range between the 75th and 25th percentiles, or in other words, the intensity range of 50% of the pixels in the image.

Entropy is another feature commonly computed directly from the image's grey-level distribution. It is Shannon's adaptation of the classical concept of entropy defined for physical systems to estimate the amount of information or uncertainty of a source, which in our case is an image [48]. Once obtained the histogram, entropy can be calculated by the following Equation:

$$H = -K \sum_{n=1}^n p_i \log p_i. \quad (3.1)$$

Where p_i is the probability of a random pixel assuming an intensity inside the bin i range. Being T_f the total amount of frequencies in the image, and f_i the frequency of bin i , $p_i = f_i/T_f$. And K is an arbitrary constant usually chosen to be $1/\log(2)$. Restating the Equation 3.1 for image's histogram it reads:

$$H = -\frac{1}{\log(2)} \sum_{n=1}^n \frac{f_i}{T_f} \log \frac{f_i}{T_f}. \quad (3.2)$$

3.3 Haralick's descriptors: the co-occurrence matrix

Those second-order features were introduced in 1973 by Haralick, Shanmugam and Dinstein for computational texture evaluation in micrographs of sandstones, panchromatic aerial photographs, and satellite images [43]). They became widely used for general image texture analysis, and certainly in medical image applications, for their proved discrimination power. Different than histogram features, Haralick's descriptors consider the statistical relationship and distribution of pairs of pixels, what gives them some spatial information.

To calculate Haralick's texture features for a ROI it is necessary to firstly compute a second-order histogram called *grey-level co-occurrence matrix* (GLCM).

This matrix is built considering a given angular direction θ ($= 0^\circ, 45^\circ, 90^\circ$, or 135° , for a two-dimensional image) and a distance d ($= 1, 2, \dots$). The GLCM has dimensions $Lx \times Ly$, where $Lx = Ly = \text{number of grey levels}$ (e.g. 256 grey levels, from 0 to 255). The elements $g_{(i,j)}$ in the matrix correspond to the number of times (frequency) the pair of intensities I and I' appears in the positions $P_{(i,j)}$ and $P_{(i+d,j+d)}$ (in this exact order) across the image, along a specific direction θ . Figure 3.3 illustrates how to generate a co-occurrence matrix for a very simple image of 4×4 pixels and 6 grey-level intensities (0 to 5), for the horizontal direction ($\theta = 0^\circ$) and $d = 1$.

Figure 3.3: GLCM for a simplistic image. The image has 6 possible grey-level intensities. The darkest value, 0. The brightest, 5. Red signs indicate the occurred frequency of pair (4,5) equivalent to (I, I') consecutively appearing in the horizontal direction. $g_{(i,j)}$ is the (i,j) th element on GLCM and represents the number of times the pair of intensities (i,j) appears on the image in a given direction separated by a given distance d .

It is easy to check visually the frequency of all possible pair of intensities across the image on Figure 3.3 considering the simplicity of the depicted image. As highlighted, the frequency of the pair of intensities $I = 4$ and $I' = 5$ is equal to 1. It means that the combination of intensities 4 followed by 5 at its right in the horizontal direction occurs a single time across the image. Symmetrically, the pair of intensities 5 followed by 4 at its left is found only once in the whole image. The occurrence of a specific pair of intensities on the image is counted in both ways,

left-to-right and *right-to-left*. This explains the even frequency for symmetric pair of intensities $(0, 0; 1, 1; 2, 2; \dots)$.

The direction of calculation θ and the distance d can be expressed simultaneously by an *offset*. An offset indicates the value to be summed in each coordinate of an image pixel $P_{(i,j)}$ to find its pair given the direction and distance. For example, in Figure 3.3, the offset $[0, 1]$ indicates that the neighbor pixel of pixel $P_{(2,0)}$ is the pixel $P_{(2,1)}$ ($\theta = 0^\circ$ and $d = 1$). This occurrence $(P_{(2,0)}, P_{(2,1)})$ would add one unity to the value of the element $g_{(P_{(2,0)}, P_{(2,1)})}$ (or $g_{(2,3)}$) of the respective $GLCM_{(d,\theta)}$. The possible offsets for *two* and *three-dimensional* images are illustrated below considering $d = 1$.

Figure 3.4: Possible offsets for $d = 1$ in two (a) and three-dimensional (b) image spaces. For every offset a specific GLCM can be generated, meaning that many co-occurrence matrices can be generated from the same image or ROI.

For two-dimensional neighborhoods 4 offset directions are usually used, meaning 4 offsets for every d value. In three-dimension images 13 regular directions may be traced for every desired d value. For each offset (a specific pair (d, θ)) a GLCM can be generated.

GLCM can yet be normalized. This is achieved by dividing each element $g_{(i,j)}$ on the matrix by the total number of frequencies found in the image for the given offset. With the GLCM at hand, 14 texture features, as initially defined by [43], can be calculated.

Considering

- $p_{(i,j)}$ being the (i,j) th element of the normalized GLCM;

- N_g the number of distinct grey-levels across the image or ROI;

- $p_{x,(i)} = \sum_{j=1}^{N_g} p_{(i,j)}$, and $p_{y,(j)} = \sum_{i=1}^{N_g} p_{(i,j)}$;

- $\sum_i = \sum_{i=1}^{N_g}$, and $\sum_j = \sum_{j=1}^{N_g}$;

- $p_{x+y,(k)} = \sum_{i=1}^{N_g} \sum_{\substack{j=1 \\ |i+j|=k}}^{N_g} p_{(i,j)}$, for $k = 2, 3, \dots, N_g$;

and

- $p_{x-y,(k)} = \sum_{i=1}^{N_g} \sum_{\substack{j=1 \\ |i-j|=k}}^{N_g} p_{(i,j)}$, for $k = 0, 1, \dots, N_g - 1$.

All the 14 features in [43] are defined as follows:

1. *Angular second moment, uniformity or energy:*

$$f_1 = \sum_i \sum_j p_{(i,j)}^2. \quad (3.3)$$

2. *Contrast or inertia:*

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{\substack{i=1 \\ |i-j|=n}}^{N_g} \sum_{j=1}^{N_g} p_{(i,j)} \right\}. \quad (3.4)$$

3. *Haralick's Correlation:*

$$f_3 = \frac{\sum_i \sum_j (ij) p_{(i,j)} - \mu_x \mu_y}{\sigma_x \sigma_y}, \quad (3.5)$$

where μ_x , μ_y , σ_x , and σ_y are the means and standard deviation of p_x and p_y .

4. *Sum of squares: Variance:*

$$f_4 = \sum_i \sum_j (i - \mu)^2 p_{(i,j)}. \quad (3.6)$$

5. *Inverse difference moment or homogeneity:*

$$f_5 = \sum_i \sum_j \frac{1}{1 + (i - j)^2} p_{(i,j)}. \quad (3.7)$$

6. *Sum average:*

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y,(i)}. \quad (3.8)$$

7. *Sum variance:*

$$f_7 = \sum_{i=2}^{2N_g} (i - f_6)^2 p_{x+y,(i)}. \quad (3.9)$$

8. *Sum entropy:*

$$f_8 = - \sum_{i=2}^{2N_g} p_{x+y,(i)} \log p_{x+y,(i)}. \quad (3.10)$$

9. *Entropy:*

$$f_9 = - \sum_i \sum_j p_{(i,j)} \log p_{(i,j)}. \quad (3.11)$$

10. *Difference variance:*

$$f_{10} = \text{variance of } p_{x-y}. \quad (3.12)$$

11. *Difference entropy:*

$$f_{11} = - \sum_{i=0}^{N_g-1} p_{x-y,(i)} \log p_{x-y,(i)}. \quad (3.13)$$

12. f_{12} and f_{13} are *Information measures of correlation:*

$$f_{12} = \frac{f_9 - HXY1}{\max\{HX, HY\}} \quad (3.14)$$

13. and

$$f_{13} = \sqrt{1 - \exp[-2(HXY2 - f_9)]}, \quad (3.15)$$

where

$$HXY1 = - \sum_i \sum_j p_{(i,j)} \log\{p_{x,(i)} p_{y,(j)}\} \quad (3.16)$$

and

$$HXY2 = - \sum_i \sum_j p_{x,(i)} p_{y,(j)} \log\{p_{x,(i)} p_{y,(j)}\}. \quad (3.17)$$

14. *Maximal correlation coefficient:*

$$f_{14} = (\text{Second largest eigenvalue } Q)^{1/2}, \quad (3.18)$$

where

$$Q_{(i,j)} = \sum_k \frac{p_{(i,k)}p_{(j,k)}}{p_{x,(i)}p_{y,(k)}}. \quad (3.19)$$

Some studies have proposed different subsets (of 6 to 8 features) to be computed instead of computing all the above coefficients [49, 50, 51]. The set implemented in the *Insight Toolkit* (ITK, [51]) based on [43, 49, 50] is composed by 8 features: f_1 , f_2 , f_3 , f_5 , and f_9 , exactly as defined in [43], and three of those 14 features differently defined [49, 50]:

- *Correlation:*

$$C = \frac{1}{\sigma_x \sigma_y} \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i - \mu_x)(j - \mu_y)p_{(i,j)}; \quad (3.20)$$

- *Cluster shade:*

$$CS = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i + j - \mu_x - \mu_y)^3 p_{(i,j)}; \quad (3.21)$$

- *Cluster shade prominence:*

$$CSP = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i + j - \mu_x - \mu_y)^4 p_{(i,j)}. \quad (3.22)$$

3.4 Run length matrix

With the same aim that Haralick's coefficients were introduced, Galloway, in 1974, proposed a group of five image features with confirmed discrimination power [46]. Galloway's coefficients are calculated out of a *grey-level run-length matrix* (GLRM), a second order histogram analogous to GLCM described in the previous section. Additional studies have defined another six features to be extracted along Galloway's traditional coefficients [52].

Figure 3.5: GLRM for a simplistic image. The image has 6 possible grey-level intensities. The darkest value, 0. The brightest, 5. Red signs indicate the run frequency of intensities (4,4,4) in the horizontal direction. $P_{(i,j)}$ is the frequency that a run of size j of pixels with intensity i appears over the ROI.

Given an image or a region of interest, a GLRM is built counting the number of runs of length j of pixels with intensity i along the direction θ . Figure 3.5 brings an example computing the GLRM analogous to the one presented in Figure 3.3.

A GLRM may be generated for each direction (θ) illustrated in Figure 3.4 for two and three-dimensional image space, since the GLRM is defined as a function of *direction* (θ). It implies in 4 matrices for two-dimensional images and 13 matrices for three-dimensional images. Moreover, for each matrix 11 features may be calculated. Next, all the 11 run-length features are defined.

For $P_{(i,j)}$ being the frequency that a run of size j of pixels with intensity i appears in the ROI, with n_r being the total number of runs in the ROI, and n_p , the number of pixels in the image. All the 11 features generated for each GLRM are defined as follows [52, 53]:

1. *Short run emphasis (SRE)*:

$$SRE = \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_{(i,j)}}{j^2}; \quad (3.23)$$

2. *Long run emphasis (LRE):*

$$LRE = \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N P_{(i,j)} j^2; \quad (3.24)$$

3. *Grey-level nonuniformity (GLN):*

$$GLN = \frac{1}{n_r} \sum_{i=1}^M \left(\sum_{j=1}^N P_{(i,j)} \right)^2; \quad (3.25)$$

4. *Run length nonuniformity (RLN):*

$$RLN = \frac{1}{n_r} \sum_{j=1}^N \left(\sum_{i=1}^M P_{(i,j)} \right)^2; \quad (3.26)$$

5. *Low grey-level run emphasis (LGRE):*

$$LGRE = \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_{(i,j)}}{i^2}; \quad (3.27)$$

6. *High grey-level run emphasis (HGRE):*

$$HGRE = \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N P_{(i,j)} i^2; \quad (3.28)$$

7. *Short run low grey-level emphasis (SRLGE):*

$$SRLGE = \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_{(i,j)}}{i^2 j^2}; \quad (3.29)$$

8. *Short run high grey-level emphasis (SRHGE):*

$$SRHGE = \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_{(i,j)} i^2}{j^2}; \quad (3.30)$$

9. *Long run low grey-level emphasis (LRLGE):*

$$LRLGE = \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N \frac{P_{(i,j)} j^2}{i^2}; \quad (3.31)$$

10. *Long run high grey-level emphasis (LRHGE)*:

$$LRHGE = \frac{1}{n_r} \sum_{i=1}^M \sum_{j=1}^N P_{(i,j)} j^2 i^2; \quad (3.32)$$

and

11. *Run percentage (RP)*:

$$RP = \frac{n_r}{n_p}. \quad (3.33)$$

3.5 Gradient features

Another feature group of image texture descriptors regularly cited by some authors is the *Gradient features* [28]. The classical definition of gradient ($\vec{\nabla}$) is a vector whose components are the partial derivatives of a function f in each direction ($\hat{i}$, $\hat{j}$, and $\hat{k}$) as in the following expression [27]:

$$\vec{\nabla} f = \frac{\partial f}{\partial x} \hat{x} + \frac{\partial f}{\partial y} \hat{y} + \frac{\partial f}{\partial z} \hat{z}. \quad (3.34)$$

The *magnitude* of the gradient vector for an image $I(x, y, z)$ is given by:

$$|\vec{\nabla} I_{(x,y,z)}| = \sqrt{\left(\frac{\partial I_{(x,y,z)}}{\partial x}\right)^2 + \left(\frac{\partial I_{(x,y,z)}}{\partial y}\right)^2 + \left(\frac{\partial I_{(x,y,z)}}{\partial z}\right)^2}. \quad (3.35)$$

On images, gradient vector's components express the variation of intensity for each specific direction. One way to obtain the one-directional components of gradient vector is by convoluting a *linear mask* (Figure 3.6) with a ROI. The result of that convolution is an *image of gradient* in which each pixel $P_{(i,j)}$ is the gradient component in $\hat{x}$ direction, for example [27].

Mathematically, applying the linear mask presented in Figure 3.6(a), for example, is equivalent to calculate the mean of the difference between the intensity value of pixel $P_{(i,j+1)}$ and pixel $P_{(i,j-1)}$.

$$\nabla_{P_{(i,j)}} = \frac{P_{(i,j+1)} - P_{(i,j-1)}}{2}. \quad (3.36)$$

Where $\nabla_{P_{(i,j)}}$ is the gradient value in the $\hat{x}$ direction at the pixel $P_{(i,j)}$, on the resultant gradient image. In other words, the component ∇_x of $\vec{\nabla}$ at the position (i, j) for the grey-level image.

Figure 3.6: Linear masks for both horizontal ($\hat{x}$) and vertical ($\hat{y}$) directions. Analogously, there is one for the ($\hat{z}$) direction. By applying those masks over an image ([27]), a resultant "gradient image" is generated for each respective direction.

The pixel $P_{(i,j)}$ does not contribute to the gradient component value, that is why its weighting value is 0 in the highlighted cell in Figure 3.6. The result of the application of the linear masks on Figure 3.6 to an image or ROI is a new image, a *gradient image*, which has same size and dimensions of the original ROI. In this way, an image for each image direction ($\hat{x}, \hat{y}$, and maybe $\hat{z}$) can be calculated.

After computing *gradient component images* with one-dimensional masks, it is possible and useful to calculate the *gradient magnitude image*. It is accomplished by using the components calculated with linear filter masks (Figure 3.6) into the Equation 3.35. In the resulting image, every pixel $P_{(i,j)}$ is equivalent to the magnitude of the gradient vector in that position in the original ROI. Figure 3.7 brings an illustration of how gradient magnitude and component images look like for a simple two-dimensional image.

Figure 3.7(a) is the original image. (b) is the $\hat{x}$ -component gradient image obtained with the linear mask over the original image. (c) is the $\hat{y}$ -component gradient image obtained with the linear mask over the original image. Moreover, (d) brings the magnitude gradient image computed with the two image components (b) and (c), by using $P_{(i,j)}$ in both Figure 3.7(b) and (c) in the Equation 3.35.

Once the gradient images are computed, conventional statistics parameters, described in Section 3.2, such as *mean*, *standard deviation*, *variance* and *maximum intensity*, may be used as features to characterize an image or ROI.

Figure 3.7: *Gradient image obtained from a simple two-dimensional image. Gradient Component images (b) and (c) are obtained, then gradient magnitude image is generated (d). Adapted from [27].*

3.6 Shape features

When considering three-dimensional image objects, some shape characteristics may be very useful. The most known one is the volume, which is extensively used in the radio-oncology field to describe a tumor stage. A considerable set of morphological features for characterization of regular and irregular ROIs (e.g. image objects or labels) can be found in [54].

The author defines:

1. *Size*: is the size of the object in number of pixels;
2. *Physical size*: is the size of the object in physical units;
3. *Centroid*: is the position of the center of the shape in physical coordinates;
4. *Region*: is the bounding box of the object given in pixel coordinates;
5. *Region elongation*: is the ratio of the largest physical size of the region on one dimension and its smallest physical size;
6. *Size region ratio*: is the ratio of the size of the object region (the bounding box) and the real size of the object;
7. *Size on border*: is the number of pixels in the objects which are on the border of the image;
8. *Physical size on border*: is the physical size of the objects which are on the border of the image;
9. *Feret diameter*: is the diameter in physical units of the sphere which include all the object;
10. *Binary elongation*: is the elongation of the shape, computed as the ratio of the largest principal moment by the smallest principal moment;
11. *Equivalent radius*: is the equivalent radius of the hypersphere of the same size than the label object. The value depends on the image spacing;
12. *Equivalent perimeter*: is the equivalent perimeter of the hypersphere of the same size than the label object;
13. *Equivalent ellipsoid perimeter*: is the size of the ellipsoid of the same size and the same ratio on all the axes than the label object;

Those features do not consider the grey-level intensity distribution across the ROI, only its geometrical organization.

STATISTICAL APPROACH: UNIVARIATE AND MULTIVARIATE ANALYSIS

As it was explained in Chapter 4, texture analysis consists of a set of techniques to extract numerical information from the pixel intensity distribution across the image. However, to make any judgment out of those extracted features or, better saying, to engage in any classification task it is necessary to make use of some statistical tools. Computers are the best servants for those tasks given their proper capacity to handle a vast amount of data and the variety of artificial intelligence algorithms already existent for that purpose.

In the context of the present work, which is the investigation of the utility of MRI quantitative features for outcome prediction in the cancer treatment, two statistical frameworks are considered: *Univariate* and *Multivariate analysis*. Univariate analysis is used to evaluate the classification performance, or in other words, the discrimination power of a *single* feature at a time. The multivariate analysis uses instead a *set*, or a *subset*, of quantitative features extracted through texture analysis to perform the classification task. The present chapter will describe both statistical approaches in details.

4.1 Univariate analysis

Univariate analysis implies comparing two or more groups of elements considering only one single feature to describe the groups involved. For instance, compare the blood pressure (feature) of sick (group 1) and healthy people (group 2) [55].

4.1.1 Student t-test and Kruskal-Wallis test

Student t-test is a conventional two-sample population mean comparison test that may be used for the comparison of two image objects (e.g. two tissue samples, tumor types) given a single quantitative feature describing the elements of each group. In this type of analysis, one single feature (e.g. *mean* or *variance* of pixel intensity across image or ROI) is used to describe each group of image objects studied.

In this test, the magnitude of the difference between each group feature mean is considered relevant to affirm difference between two groups depending on the p-value test output and the significance level, α , adopted. Traditionally $p < 0.05$ is arbitrarily chosen as significance level ¹ [55].

Some examples of this approach for medical image quantitative features for clinical purposes are: [56], to predict chemotherapeutic response in cancer treatment; [57], to predict overall survival and progression-free survival in glioblastoma patients; and [58], to evaluate malignancy in orbital tumors.

Student t-test makes use of the *normality assumption*. It means, to apply this test, one is considering that the distribution of subject features in the compared groups is reasonably close to the standard normal. This assumption is often correct when the number of subjects in each group is close to or greater than 30 [55]. If the normality assumption principle is not applicable, *non-parametric* tests are considered [59]. *Kruskal-Wallis* test is a suitable non-parametric alternative to Student's t-test for two-sample mean population comparison [55].

If any statistical evidence of difference between the studied groups is found, one could say that this feature is capable of differentiating or classifying the two objects studied. This type of analysis is carried out for every numerical feature extracted. It is important to stress that, as the name states, the relationship among features is not considered at any stage of this analysis.

¹Lately after a wide discussion on the topic, some authors have chosen $\alpha = 0,0125$ for considering any statistical evidence (e.g. [56])

4.1.2 Receiver operator characteristic

Coupled with two-sample Student t-test, the *receiver operator characteristic* (ROC) is used in univariate analyses to measure the effectiveness of the classification for each feature considered [55, 60].

Consider the following example. Suppose that 100 samples of two tissue types A and B (grey matter and white matter, for example) have been imaged, being 50 of each type. Consider $A_{variance} = 10.8$ and $B_{variance} = 13.8$ being the average of the variance for each group. How good would be a classification if one looks only to the grey-level variance value of the image for that classification task?

The situation described above is an example of a *binary classification* problem (it means only two possible outcomes, tissue A or tissue B , false or true, 0 or 1) where variance is the texture feature used to differentiate these two tissue types. To evaluate how accurate this classification is, it is necessary to draw a ROC curve. This is accomplished by arbitrarily choosing a set of different *cutoff value* for variance and making a simple procedure of classification: if *variance* > *cutoff values* it is a tissue type A , otherwise it is a tissue type B . For every cutoff value the ratio of true positive classifications (known as *true positive rate* (TPR) or *Sensitivity*) and the ratio of false negative classifications (known as *false positive rate* (FPR) or *1-Specificity*) is calculated.

Therefore, each cut-off value gives an ordered pair (1-specificity, sensitivity) and with those pairs it is possible to draw a ROC curve and even select the best cut-off value of variance for differentiating those tissue types. Figure 4.1 brings some illustration of how ROC curves look like.

ROC curve itself gives a qualitative idea of how good is the chosen feature in a classification task. Nevertheless, a more objective measurement may be obtained by calculating the *area under the ROC curve* (AUC). It is, as the name reads, the total area under the ROC curve, as highlighted in Figure 4.1(d). AUC ranges from 0 to 1, where 1 means a perfect classifier (Figure 4.1(c)), 0.5 is equivalent to a random classification (Figure 4.1(a)), and anything less than 0.5 is worse than guessing [55, 60]. ROC curve is also widely used to evaluate the performance of

²Source: <http://blog.yhat.com/posts/roc-curves.html>

Figure 4.1: *Receiver operator characteristic graphic. The graphical representation is another way to represent the defectiveness of a binary classification task*².

classification algorithms in multivariate analysis, as discussed in the next section.

4.2 Multivariate analysis: Machine learning

Instead of using each extracted feature individually to classify or differentiate image objects across subjects, a stronger analysis can be approached using *Machine Learning* (ML) algorithms for compiling a classification task (as the example early given of tissue type discrimination A and B), considering all the quantitative features extracted simultaneously.

Many definitions of machine learning can be found in literature. ML is "the study that gives computers the ability to learn without being explicitly programmed.", by the pioneer in the study of artificial intelligence, Arthur Samuel.

"Machine learning is the design and study of software artifacts that use past experience to make future decisions; it is the study of programs that learn from data." [61]. The computer scientist Thom Michel gives a more formal definition of such "learning", he said that "a program can be said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E ." [61, 62].

Machine learning is a transdisciplinary field that conjugates artificial intelligence, computational complexity, probability, statistics, information theory, neurobiology and even philosophy [62]. Moreover, it has been used to tackle a wide variety of problems, other than the clinical ones just mentioned, such as house pricing, pattern recognition, computer game playing, spam detection, drug response, house pricing, image segmentation, customer grouping and many others [62, 30].

One can find some very fascinating demonstrations of the power of this type of analyses in medical image studies in the literature. It was used for tissue characterization of the human brain [41]; for childhood tumor type classification [63]; to differentiate primary lung cancer and granulomatous lesions [64]; for diagnosing glaucoma [65]; to visualize histological heterogeneity and identify rich tumor areas in glioblastoma [66]; and for primary breast cancer classification [67].

The junction of an automated process of texture features extraction from medical images and sophisticated algorithms to help in a clinical decision, classification or diagnosis has provided enough ground to raise a field known as *computer aided diagnosis* (CAD) [42].

4.2.1 How do machine learning algorithms work?

The simplest way to illustrate how those powerful algorithms work is to think of a binary classification task, for instance, the classification problem of tissue characterization earlier mentioned. That example is said to be a binary problem because it involves only two labels or classes (Tissue A or B). In that example, the whole dataset contains 100 samples (S), 50 of tissue A and 50 of tissue B .

A dataset is best represented by a table. A sample (S_i) in the dataset is composed by its texture features ($f_1, f_2, f_3, \dots, f_n$) extracted from the image and is associated with a label (L_i) that identify it as tissue A or tissue B .

Initially, the dataset is split into two subsets, the *training* and *testing set*. Say, 80 samples (40 of tissue A, 40 of tissue B) for training, and 20 samples (10 of tissue A, 10 of tissue B) for testing. Afterwards, the algorithm receives the training set and *learns* what is a tissue sample of type A and B. The testing set samples are given to the algorithm without their *already* known classification labels. The algorithm, then, classifies the testing set and its outcome is compared with the already known label classes for those samples. Finally, a *metric* routine computes how good was the performance of the algorithm. This broad example of machine learning application is translated into a flowchart in Figure 4.2.

Figure 4.2: Machine learning process. The ML scheme depicted here refers to a supervised learning task of two possible classifications, tissue A or B.

4.2.2 Learning modalities and approaches

Two major modalities of learning processes characterize machine learning applications: *supervised* and *unsupervised learning*.

The supervised learning modality consists of a learning process where samples that compose the dataset are *already* separated into categories or classes. The algorithm "knows what it is learning" in the learning step and "knows what it is looking for" in the testing step.

Supervised processes are used to attack classification or regression problems. In classification ones, the algorithm acts by identifying which category or class an object belongs to, it makes predictions about classes. The performance result

of a binary classification problem (two classes) is illustrated in Figure 4.3(a) as a two dimensional graph. Differently, regression algorithms function predicting a continuous-valued attribute variation, they fit a function and predict further values (Figure 4.3(b)) [62, 30].

Figure 4.3: *Supervised learning tasks illustration. In (a), the problem is to separate the elements in two classes, here represented by circles and crosses. x_1 and x_2 are the two describing numerical features. In (b) the algorithm is trying to fit a prediction for the variable y as x increases. Adapted from [68]*

Figure 4.4: *Machine learning process. The ML scheme depicted here refers to a supervised learning task of three possible classifications, i.e., tissue A, B, or C.*

In the other hand, in unsupervised learning methods samples in the training set are not labeled. Some algorithms in this modality look for patterns in the data

not yet known and that may be relevant for a well defined purpose. This strategy is called *clustering* (Figure 4.4) [30].

4.2.3 Some classification algorithms

There are several types of machine learning algorithms [30, 61, 62]. They differ mostly in the strategy and mathematics used in the learning and prediction steps. For the sake of the present work's interest, which is to model prediction of tumor recurrence based on imaging features, a detailed explanation will be given only for the three algorithms used here: *k-nearest neighbor*, *decision tree*, and random forest classifiers.

4.2.3.1 k-nearest neighbor

k-Nearest Neighbor (kNN) is a non-parametric algorithm used for both supervised and unsupervised learning frameworks. It is defined by very simple modeling principles. Given the $N - dimensional$ feature space where samples are "localized", consider an unlabeled sample X_0 . The algorithm finds the k ($1, 2, 3, \dots$) nearest neighbors of that X_0 . The k parameter is defined by the user, is commonly set to 5 ([30]), and can influence very much the model's performance and flexibility. The distance calculated between samples maybe Euclidian or other estimation of distance. Based on neighbors' labels y_i ($(0, 1)$ or $(a, b, c, \dots)$), the model assign a probability P of sample X_0 being classified with label y_j . P is given by:

$$P(y = y_j | X = X_0) = \frac{1}{K} \sum_{i=0}^{K-1} I(y_i = y_j). \quad (4.1)$$

Where I is a conditional variable that assumes the value 1 when the given condition is *true* and 0 otherwise. It means that the more neighbor samples X_i with label y_j are close to sample X_0 , the higher the probability P is to assign label y_j to sample X_0 [30, 69]. Figure 4.5 image illustrates how kNN works and how the number of neighbors (K) affects the final classification.

In Figure 4.5, the green sample may be classified as a red triangle, if its two nearest neighbors are considered, and as a blue square, if its three nearest neighbors

³Source: https://www.researchgate.net/publication/267953942_Video_-based_Fall_Detection_in_Elderly%27s_Houses/figures

Figure 4.5: 2D representation of KNN classification. It is a binary classification problem where square and triangle represent known label classes, and the circle is the unlabeled sample. Solid line represents how the classification would be taken if $K = 3$ nearest neighbors were considered, and dashed line for $K = 5$ ³.

are taken into account. After calculating the labeling probabilities of classification (P) based on nearest neighbors (X_k), the algorithm decides the final classification (y_j) for X_0 sample [30, 69].

4.2.3.2 Decision tree

Similarly to kNN, decision tree classifier (DTC) is a very intuitive algorithm and can be easily translated to a comprehensive logic for human understanding making it an attractive choice for decision-making problems in the clinical setting. It is also an algorithm versatile enough to be used in both classification and regression tasks with applications ranging from celestial objects classification to credit risk of loan application assessment [62].

DTC can be explained as a hierarchical If-Else set of splitting rules (a decision tree) that is produced as binary decisions are done out of a single initial node (the top of the tree). The classification trees are built from the given data. DTC starts at a root node that is represented by a parameter or feature in the dataset that describes the samples. The algorithm makes a binary decision (False or True) that partitions the original dataset (of dimensions *number of samples* x *number of features*) in two smaller sets that form two different *branches* (left and right). Each smaller sample will be checked by another *node* (feature, parameter) that will judge and separate

the samples again in two even smaller subsets, and this process continues until it reaches the final labeling classification, the leaves. This process of tree generation is called recursive binary splitting [69].

Decision trees, once generated, can be visualized and intuitively comprehended. Figure 4.6 illustrates the decision tree learned from the public dataset *Iris*⁴. It brings 150 samples of three different types of iris flower (*Setosa*, *Versicolour*, and *Virginica*). Samples are described by 4 numerical features (*sepal length*, *sepal width*, *petal length*, and *petal width*). The tree was generated using the whole dataset (150, 4).

In this tree, the root node is defined with the petal length parameter, and the binary rule is *petal length* ≤ 2.5 . This node receives initially all the 150 samples, 50 of each type indicated in the label value. This rule eliminates or isolates all 50 samples of iris flower of type setosa in one branch, and then finishes this branch with that label. The other 100 samples are left to the right branch that originates many other branches, and this is recursively repeated until each branch ends in a *leaf*.

In this recursive process, one of the possible criteria used to find the best binary split is the *Gini's index* (G) (brought in each node of the tree in Figure 4.6) that measures variance across all the classes (K) [69]. Consider that the tree is divided in regions (branches, m), and inside those regions different proportions (p) of the classes inside each region are observed. Gini's index is then defined as

$$G = \sum_{k=0}^{K-1} p_{mk}(1-p_{mk}). \quad (4.2)$$

This index indicates the 'purity' of the node. It means, if it is close to zero, that node has predominantly one single class. Otherwise, it indicates the variance of classes out of that node. Another strategy also used to evaluate the quality of split is the *cross-entropy* (Shannon entropy, Section 3.2) that is equally sensible to node's purity [69].

⁴Available at: <https://archive.ics.uci.edu/ml/datasets/iris>

⁵Source: <http://scikit-learn.org/stable/modules/tree.html>

Figure 4.6: Illustration of decision tree process for Iris dataset. It is a multiclass classification problem where labels correspond to three classes of flowers ⁵.

4.2.3.3 Random forest

Random forest classifier (RFC) is an example of ensemble classification model. The ensemble technique mainly intends to create models more generalized and robust by gathering a set of different base models. This ensemble model may use different strategies to reach the final prediction. The final goal is to improve prediction performance [30].

RFC uses a set of independently constructed DTC as its base. Final classification for a tested sample is achieved by averaging the individual base decision tree predictions. The random aspect of the classifier is present in the tree generation

process, more specifically in the splitting node step [30, 69].

4.2.4 Preprocessing steps

In order to improve the performance of ML classification models some *preprocessing* steps are highly recommended, and even required by some algorithms. Preprocessing steps are operations that may be done over the *raw dataset* (X) that will output a *new dataset* (X'), sometimes conserving original dimensions, sometimes with smaller dimensions. In other words, a *new dataset* with same or less amount of parameters or features. Those processes include *standardization*, *scaling*, *normalization* and *feature selection* [30].

Some classification algorithms (e.g. kNN) may function incorrectly if the dataset given does not bring features with distributions close to the standard normal (mean equals to 0, and variance equals to 1). The process to achieve this is called standardization. Scaling is the process through which every feature in the dataset is resampled inside a specific range, e.g. (0,1). Normalization is a process similar to standardization, but it is done over the samples not over the features as in the standardization process [30].

4.2.4.1 Feature selection

Cautions related to the number of features (N_f) and number of samples (N_s) in a dataset must also be considered when building a classification model. *Overfitting* problems are likely to be observed when there is a huge difference between the number of features and the number of samples ($N_f \gg N_s$) in the dataset. It means that the model is finding relations that do not really exist leading it to misclassifications. *Feature selection* (FS) processes are extensively used as a reasonable way out of this situation. Those operations intend to remove either irrelevant or redundant features [30, 70].

There are a considerable variety of strategies to accomplish a dimensionality reduction ranging from very intuitive ones to more sophisticated ones. A straightforward approach for this is the *variance thresholding*. Features with variance lower than a given threshold, consequently all features with zero variance across samples, are removed from the dataset. Another alternative is the *univariate* FS

where features are checked individually with statistical tests, e.g. *chi-square* and *F-test* [30].

Mutual information feature selection is a more sophisticated FS method developed in the basis of *probability and information theory* [71]. This method intends to calculate the amount of information obtained from a random variable X about another variable Y . With that it measures the dependence between those two variables. The mutual information (I) for discret (Equation 4.3) and continuous (Equation 4.4) variables is defined as follows: For discret variables:

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p_{(x,y)} \log \left(\frac{p_{(x,y)}}{p_{(x)}p_{(y)}} \right), \quad (4.3)$$

and for continuous ones:

$$I(X, Y) = \int_{y \in Y} \int_{x \in X} p_{(x,y)} \log \left(\frac{p_{(x,y)}}{p_{(x)}p_{(y)}} \right) dx dy. \quad (4.4)$$

Where, $p_{(x,y)}$ is the joint probability function, and $p_{(x)}$ and $p_{(y)}$ are the marginal probability functions. I is the Boltzmann-Gibbs entropy translated into the information theory setting. In the feature selection context, it is calculated for a number k (usually 3) of nearest neighbors. And it assumes zero when there is no dependency at all among the random variables and increases as dependency is observed.[30, 71].

Some FS strategies consider using the classifier itself to performe this dimensionality reduction. *Recursive* feature elimination is an approach where the classification algorithm itself assigns some weights and ranks the features in the dataset, and removes recursively less important ones until it reaches an optimum set or a desired feature set size [30]. Those mentioned methods work by simply removing certain features from the original data matrix. A different approach widely used is *Principal Component Analysis* (PCA) that mathematically transforms the dataset to a new space and identifies the components with the highest variation, which turn to be the principal components on the data [30].

4.2.4.2 Dataset stratification

In order to carry out a classification task the dataset needs to be separated in two subsets, the training set and the testing set, as illustrated in Figure 4.2. There

are some stratification strategies to accomplish that.

The simplest way is to separate k folds out of the total amount of samples N_s to use them as the testing set while $N_s - k$ samples are used as the training set. There is the *k-fold cross-validation* (KFCV) where many classifications are executed, different k samples are removed from N_s at a time and used as the testing set, the rest is used as training set. The classification process is repeated $N_s - (k - 1)$ times and the final result is the average of all the classifications made. And finally, *the leave-one-out cross-validation* (LOOCV) which is similar to KFCV, but for $k = 1$. LOOCV is highly recommended for small datasets where separating the original dataset is a quite interesting option [30].

4.2.5 Metrics

Once the classification model is built it needs to be validated and for that, the performance it exhibited over the testing and validation step needs to be quantified. This is achieved using some very well-known metrics. The metric type chosen to score a classification model performance may vary with the characteristics of the model itself. Some commonly used metrics for both binary and multiclass classification tasks are *accuracy*, *sensitivity*, *specificity*, *AUC*, *precision*, and *F1-score* [30].

Before defining those scores, four more basic outcome numbers on the classification scenario must be stated:

- *True Positive* (TP) number: amount of correct positively classified samples;
- *True Negative* (TN) number: amount of correct negatively classified samples;
- *False Positive* (FP) number: amount of wrong positively classified samples;
- *False Negative* (FN) number: amount of wrong negatively classified samples;

Sometimes those numbers come arranged in a matrix called *confusion matrix*. Figure 4.7 illustrates how those quantities come organized in such a matrix. This confusion matrix was generated for a multiclass classification task over the public dataset Iris already introduced in Section 4.2.3.2. The tested sample consisted of 13 Setosa, 16 Versicolour, and 9 Virginica.

Figure 4.7: Illustration of decision tree process for Iris dataset. It is a multiclass classification problem where labels correspond to three classes of flowers ⁶.

All the four quantities are computed for each of the three classes in the problem (same presented in section 4.2.3.2). For Versicolour, for instance, Figure 4.7 shows that 10 samples out of 16 samples were correctly classified (TP). No sample out of 22 was wrongly said to be Versicolour (FP). 6 samples out of 16 were wrongly said not to be Versicolour (FN). And 22 samples out of 22 were correctly said not to be Versicolour by the classifier.

Following, the scorers and rates computed out of TN, FN, TP, FP numbers:

- *Accuracy* (ACC): Measures how well a model rightly classifies a sample, whether positive or negative:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}; \quad (4.5)$$

- *Sensitivity* (SEN): Measures how well a model can correctly classify a sample as positive. Also known as recall, and true positive rate:

$$SEN = \frac{TP}{TP + FN}; \quad (4.6)$$

- *Specificity* (SPE): Measures how well a model can correctly classify a sample as negative. Also known as true negative rate:

$$SPE = \frac{TN}{TN + FP}; \quad (4.7)$$

- *Precision* (PRE): Measures the ability of a model not to misclassify a sample that is negative:

$$PRE = \frac{TP}{TP + FP}; \quad (4.8)$$

- *F1-Score* (F1-S): Computes the harmonic mean of sensitivity and precision:

$$F1 - S = \frac{2(Spe \cdot Pre)}{Spe + Pre}; \quad (4.9)$$

- *AUC*: Is the same index introduced in Section 4.1 and measures how good a binary classification performance is by integrating the area under ROC curve.

All those metrics listed above intend to evaluate how much prediction errors happened in a classification task. All of them intend to give a piece of information about how good the designed model is in some aspect of the classification process. The metrics defined above achieve their best value at 1 and worst at 0 [30, 61].

With those tools in hand for univariate and multivariate analysis and metrics explicitly defined, models can then be fed and the classification task performed. In the following chapter, a meticulous explanation on how the specific goals of this work were tackled with theoretical background explained in the previous chapters is given. It will be discussed in detail the subjects including and excluding criteria, MRI preprocessing steps, implementation of an algorithm for feature extraction, univariate statistical tests and machine learning approaches used to investigate the proposed hypothesis: predict tumor recurrence after first surgery in patients harboring non-functioning pituitary adenoma using MRI texture analyses as image biomarkers.

STUDY DESIGN, IMAGE AND DATA ANALYSES

EACH of the methods for quantitative digital image analysis mentioned in Chapters 4 and 5 have been introduced at different moments and have shown usefulness for many *well-defined classification tasks*.

Nevertheless, it is important to stress that for every new narrowly defined classification problem those texture features must be tested in order to evaluate if they have any discrimination power in a given scenario.

In this way, this work investigated the prognostic power of the texture features introduced in Chapter 3 in the following clinical task: Prediction of NFPA recurrence or progression after first resection surgery through analysis of patients preoperative MRI.

5.1 Study design

To carry out this investigation a *retrospective study* was proposed with patients harboring non-functioning pituitary adenomas with at least two and a half years of documented follow-up after first surgical approach.

Two-hundred and five NFPA patients treated in Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, localized in São Paulo, Brazil (HCFMRP) participated in this study. The study was approved by both HCFMRP ethics committee and the Faculdade de Filosofia Ciências e Letras de Ribeirão Preto (FFCLRP) ethics committee, and all patients signed an inform consent form. Those proofing documents are available in the annex section.

5.1.1 Including and excluding criteria

As described in Section 2.3, MR images carry a built-in variability. With that in mind and considering the proper variability involved in this tumor type treatment (Section 1.5), including and excluding criteria were proposed in order to eliminate or at least reduce such variability, and to make the study question more narrow and well-defined.

The *including* criteria were:

- Adult patients harboring non-functioning pituitary macroadenoma;
- Patients with at least two and a half years of follow-up properly documented. This includes patient clinical history and MRI;

The *excluding* criteria were:

- Patients without follow-up documents (clinical history or MR images);
- Patients with observed apoplexy in preoperative MRI. The image is consistently different and the imaging pattern (of tumor) associated with that phenomena may introduce a variability that affects the data of interest;
- Patients without preoperative images in the desired MRI protocol (DCE 3D T1-weighted, in 3T scanner). The patient group considered for the study must had been imaged in the same MRI protocol;
- Patients with preoperative MR image obtained in scanners with different field intensity (here, 1.5T);
- Patients who were treated with RT in less than two and a half years after the first surgery for reasons different than tumor recurrence;
- Patients who underwent a secondary surgical approach in less than two and a half years after the first surgery for reasons different than tumor recurrence.

Considering only the including criteria, 83 out of the 205 patients were kept in the study. The other 122 patients were removed mostly for not having preoperative images available in the HCFMRP image database.

Checking the excluding criteria, another 65 patient were removed by not having images in the desired protocol or by being treated with RT in less than two and a half years after the first surgical approach. In the end, only 15 patients were considered appropriate for the study.

Those 15 patients were divided in two groups of interest: *stable lesion group* (8 patients) and *recurrent lesion group* (7 patients).

5.1.2 Studied phenomena

The phenomena studied here was the *postoperative short-term recurrence (PSTR) in NFPA treatment*.

Recurrence was defined as an observed and measured (through MRI studies) tumor volume increase after the first postoperative image.

Tumor recurrence was evaluated and diagnosed by a multiprofessional healthcare team (endocrinologists, neurosurgeons, and radiologists) that followed those patients during the specified period.

5.2 Images used

To carry out the proposed quantitative MR image study, a specific MRI protocol available for all included patient was chosen, which is:

Image protocol:	3D T1-Weighted Dynamic Contrast Enhanced with Gadolinium (Section 2.2)
Image size:	240 px X 240 px X 160 px
Slice thickness:	1 mm
Voxel resolution:	1.0 mm X 1.0 mm X 1.0 mm *
Pulse RT:	7.2 s
Pulse ET:	3.35 s
Scanner field:	3.0 T
Scanner manufacturer:	Philips Medical Systems, model: Achieva

Table 5.1: *Image protocol specifications.*

* *Four* patients included in the study had same image protocol of choice but in a different voxel resolution, they were imaged with $0.6\text{ mm} \times 0.6\text{ mm} \times 1.0\text{ mm}$. Those patient images were registered (rigid registration) into the desired voxel resolution to ensure uniformity of the data.

5.2.1 Tumor segmentation

Additionally to patients' preoperative image, an image *segmentation* that defines the ROI to be used in feature extraction step was necessary to draw.

The segmentation was obtained in two steps: firstly with a *semi-automatic grown-cut tumor segmentation method* [72]. Then, the obtained segmentation was appropriately corrected by a radiology specialist.

(a) Axial slice

(b) Sagittal slice

(c) Coronal slice

Figure 5.1: *Segmentation process. In GREEN is the segmentation obtained by a semiautomatic method. In BEIGE, the radiologist marking. Difference is highlighted when both markings are overlapped.*

The radiologist in charge of the tumor segmentation chose a *conservative*

criteria. That means, marking only what he was absolutely certain that was tumor tissue, and letting out what he considered to be *partial volume*. This criteria ensured that a "good" sample of the imaged tumor lesion was analyzed.

5.3 Image preprocessing and feature extraction software

All image preprocessing (including the segmentation process just described) and data extraction were carried out in the medical image processing and analysis platform *3DSlicer* [73]. It is an *open source* environment based on C++ that supports Python extensions and Matlab bridges, which makes it a very versatile tool in the image processing and analysis scenario (Figure 5.2).

Figure 5.2: Screenshot of *3DSlicer* welcome screen.

Additional to 3D Slicer, the *Insight Segmentation and Registration Toolkit* (ITK) was also used [44, 51]. It is an open source, cross-platform system that provides an extensive library of C++ algorithms for image processing and analysis.

Preprocessing includes image registration of the five patients described in the earlier section, and both semi-automatic segmentation and manual specialist's

correction of tumor delimitation on the included patient images.

5.3.1 Feature extraction plugin implementation

3D Slicer supports an extension development mode (Extension Wizard) where one can create a pipeline process that can be executed inside and outside the 3D Slicer environment.

To carry out image texture feature extraction in 3D Slicer a *command line interface* (CLI) module (Figure 5.3) was constructed. It was built using some already implemented ITK classes for image preprocessing and texture analysis. The steps of the CLI module created for the present work are explained in details further.

Figure 5.3: Screenshot of CLI module interface. Input fields and features methods check-boxes are visible.

Figure 5.4: *CLI module input. (a) raw 3D T1-Weighted DCE, sagittal view. (b) radiologist final segmentation in green. Those are the inputs received into CLI processing pipeline.*

The segmentation `label` (green) in Figure 5.4(b) is a mask that holds pixel value 1 in the 3D ROI and 0 everywhere else.

The following numbered items combined with Figure 5.5 describe the CLI module pipeline and details of the implementation process. The 3 preprocessing steps (items 1 to 3) and the 5 quantitative image feature extraction methods (items 4 to 8) were used to extract 48 numerical features in the proposed study. Item 9 is the report output step.

1. Bayes correction:

Bayes correction (Section 2.3) was the first step applied in the image preprocessing. This routine receives as `input`: 3D MRI (Figure 5.4(a)) and tumor label (segmentation that specifies ROI, Figure 5.4(b)). It uses the class `itk::BinaryDilateImageFilter` to dilate the given label.

MRI is given as parameter for method `SetInput()`, and the dilated label passed to method `SetMaskImage` in the `CorrectionFilter`. Then it applies the N4 algorithm with class `itk::N4BiasFieldCorrectionImageFilter` only over the dilated ROI. This outputs a corrected MRI.

2. Image type conversion:

Figure 5.5: CLI module pipeline.

The MRI outputted by the previous routine is then cast to a known data type (unsigned short) with class `itk::CastImageFilter`.

3. Intensity normalization:

Pixel intensity normalization (Section 2.3) is taken to standardize image group into the desired range to make averages and appropriate comparisons.

It is done by analyzing only the defined ROI. It is done in two steps: It firstly *removes* from ROI pixels with intensities ($I_{(i,j,k)}$) *out* of the image pixel intensity range of interest. It removes the following pixels:

$$I_{(i,j,k)} < \mu - 3\sigma \text{ and}$$

$$I_{(i,j,k)} > \mu + 3\sigma.$$

Where μ and σ are ROI pixel intensity average and standard deviation respectively. Next, it rescales the resulting ROI to a predefined range: 1 to 128 grey-level intensity. This step ensures that every patient will have

images inside the same grey-level intensity range. In [38] and [63] 1 to 64 grey-level intensity range was used, but here a bit more conservative approach was used, since reducing image pixel intensity scale also reduces the amount of information present in the image.

This routine outputs a cropped image with only the normalized ROI and zero elsewhere, and the normalized label.

Figure 5.6: *Normalization step output. (a) 3D T1-Weighted DCE normalized and cropped ROI and (b) normalized label. Both images are presented in the coronal view.*

4. Histogram features:

Histogram features extraction was performed with the class `itk::LabelImageToStatisticsLabelMapFilter`. The filter created with that class received as input the normalized 3D grey-level image (Figure 5.6(a)) and its respective ROI label (Figure 5.6(b)). The following features were calculated:

Mean, Std Deviation, Median, Skewness, Kurtosis, Variance, Histogram Elongation, Flatness, Minimum, and Maximum.

Moreover, the following features were manually calculated:

Percent 5th, Percent 10th, Percent 25th, Percent 75th, Percent 90th, Quartile, and GL Entropy.

Totalizing 15 features with that method.

5. GLCM features:

To calculate GLCM features the class `itkScalarImageToCooccurrenceMatrixFilter` was used. Not all the 14 features described in Section 3.3, only the subset implemented in ITK ([51]), which are:

Entropy, Energy, Correlation, Haralick-Correlation, Cluster Prominence, Cluster Shade, Inertia, and Inverse Difference Moment.

Totalizing 8 features. This feature set was calculated for the 13 offsets described in Figure 3.4(b) for $d = 1$ and then averaged. It resulted in 8 features invariant to rotation and uncorrelated among themselves.

6. GLRM features:

Similarly to GLCM features, GLRM features (Section 3.4) were calculated using the normalized grey-level image and ROI. Furthermore, the same 13 offsets describe in Figure 3.4(b) was used to obtain features invariant to rotation with stronger discrimination power.

The class `itkScalarImageToRunLengthMatrixFilter` was used to calculate the 11 features described in Section 3.4:

SRE, LRE, GLN, RLN, LGRE, HGRE, SRLGE, SRHGE, LRLGE, LRHGE, and RP.

7. Gradient features:

The process to calculate gradient features is a bit different than the ones described earlier since it was not already implemented in ITK.

Firstly, the class `itkGradientMagnitudeImageFilter` was used to calculate the gradient intensity image (Section 3.5) of the normalized grey-level image (Figure 5.6(a)) , resulting in Figure 5.7(a). Then the normalized label (Figure 5.6(b)) was eroded (`itk::BinaryErodeImageFilter`) in order to mark a

smaller ROI letting out the "contaminated borders" (Figure 5.7(b) and (c)). Finally the class `itk::LabelImageToStatisticsLabelMapFilter` was used with the eroded label (new ROI) to calculate the following 4 features:

Gradient Mean, Gradient Std, Gradient Variance, and Gradient Maximum.

Figure 5.7: *Gradient features extraction. (a) Gradient magnitude image obtained from Figure 5.6(a). (b) and (c) correspond to the eroded label letting out borders.*

8. Shape features:

To calculate shape features (Section 3.6) the label image (Figure 5.6(b)) and the class `itk::LabelImageToShapeLabelMapFilter` were used. The following 8 features were obtained:

Volume In Physical Units, Size Ratio, Elongation, Feret Diameter, Equiv Spher Perimeter, Equiv Spher Radius, Perimeter, and Roundness.

All those methods of feature extraction and their respective extracted features are summarized in Table 5.2.

5.3.1.1 Clinical features

Based on the clinical documentation, four clinical parameters were added to the analysis. These are:

- *Invasiveness status*: Hardy invasiveness classification (Section 1.3) was used to evaluate the presence of any level of tumor invasion. This feature was

Method	Features
Histogram features	Mean, Std Deviation, Median, Skewness, Kurtosis, Variance, Histogram Elongation, Flatness, Minimum, Maximum, Percent 5th, Percent 10th, Percent 25th, Percent 75th, Percent 90th, Quartile, and GL Entropy.
GLCM features	Entropy, Energy, Correlation, Haralick-Correlation, Cluster Prominence, Cluster Shade, Inertia, and Inverse Difference Moment
GLRM features	SRE, LRE, GLN, RLN, LGRE, HGRE, SRLGE, SRHGE, LRLGE, LRHGE, and RP
Gradient features	Gradient Mean, Gradient Std, Gradient Variance, and Gradient Maximum
Shape features	Volume In Physical Units, Size Ratio, Elongation, Feret Diameter, Equiv Spher Perimeter, Equiv Spher Radius, Perimeter, and Roundness

Table 5.2: Summary of 48 quantitative textural features extracted with the build CLI module.

binarized: 0, or *false*, when no invasive level was observed by the specialist (Grade 0 I, or II, Figure 1.3); and 1, *true*, when invasion into surrounding structures was detected (Grade III and IV, Figure 1.3);

- *Immunohistochemistry*: Immunonegativity and immunopositivity were considered as qualitative binary information. 0, or *false*, if no sign of immunopositivity was observed; and 1, *true*, if the tumor showed positive result for any of the five hormones tested (LH, ACTH, GH, PRL, and FSH);
- *Age at first surgery*: It is the time in years counted between date of birth and date of first surgical approach.
- and *gender*.

The number of features (clinical and computationally extracted) considered in the study totaled 52 features.

5.4 Data analysis

Initially, a *descriptive data analysis* was done to evaluate if the collected data looked alike for both sample groups. Histograms and boxplots were used to visualize data distribution and check the presence of outliers.

The *statistical data analysis* was carried out in two approaches: univariate analysis (Section 4.1) to evaluate the classification and characterization performance of each extracted feature individually; and multivariate analysis (Section 4.2) where ML classification algorithms tested the characterization performances of sets of features.

5.4.1 Univariate analysis

Each feature was checked individually to assess its capability to separate those two patient groups (*stable* and *recurrent group*). The following tests were used in that univariate analysis:

- *Shapiro-Wilks test* was used to evaluate the normality of both groups and see if it was acceptable to proceed with parametric or non-parametric mean group comparison. Evidence of an absent normality was considered for $p - values < 0.05$, and *normality* otherwise.
- *Levine's test* was used to evaluate variance in both groups. This is important for using some parametrical tests that are applied differently if groups' variances are different or equal. Evidence of *variance difference* between *stable* and *recurrent* groups was considered for $p - values < 0.05$, and *equal variance* otherwise.
- *Student's t-test* for independent samples was used to compare means of *stable* and *recurrent* groups for each one of the numerical features. Evidence of *differences* between compared groups were considered for $p - values < 0.05$, and *similarity* otherwise.
- *Kruskal-Wallis test* was used as a non-parametric alternative for independent samples comparisons. Evidence of *differences* between compared groups was considered for $p - values < 0.05$, and *similarity* otherwise.

- *Exact Fisher's test* was used to evaluate the correlation between the qualitative clinical features (invasiveness, immunohistochemistry, and gender). Evidence of *correlation* between compared groups was considered for $p - \text{values} < 0.05$, and *similarity* otherwise.

Table 5.3 summarizes statistical tests used in the univariate analysis.

Test	Description	Evidence considered
Shapiro-Wilk's test	Assess sample distributions' normality;	$p < 0.05$
Levine's test	Evaluate if two groups have equal or different variance;	$p < 0.05$
Student's t-test	Compare two independent population mean;	$p < 0.05$
Kruskal-Wallis' test	Nonparametric comparison of two independent sample;	$p < 0.05$
Exact Fisher's test	Evaluation of two qualitative phenomena (binary clinical information and tumor recurrence);	$p < 0.05$

Table 5.3: Summary of statistical tests used in the univariate analysis.

The tests described above are useful to inform if a feature is good or not to characterize the studied phenomena on both patient groups. The area under the ROC curve (Section 4.1(d)) was used to assess *how useful* each feature is in classifying a patient as belonging to *stable* or *recurrent* group.

5.4.2 Multivariate analysis: machine learning

In addition to the univariate analysis just described, a multivariate analysis with different sets of features was carried out. Three ML algorithms were used: kNN, DTC, and RFC (introduced in Section 4.2.3).

LOOCV was used as the stratification method in all machine learning tests since the dataset available for the present study was very short.

Those three algorithms are already implemented and available in package Sci-Kit Learn for Python users. They were set as follows:

- kNN:

It is implemented in the

`class sklearn.neighbors.KNeighborsClassifier.` The standard configuration from Sci-Kit Learn was used:

```
- n_neighbors = 5
- weights = 'uniform'
- algorithm = 'auto'
- leaf_size = 30
- p=2
- metric = 'minkowski'
- metric_params = None
- n_jobs = 1
```

- DTC:

It is implemented in the

`class sklearn.tree.DecisionTreeClassifier.` The standard configuration from Sci-Kit Learn was used:

```
- criterion = 'gini'
- splitter = 'best'
- max_depth = None
- min_samples_split = 2
- min_samples_leaf = 1
- min_weight_fraction_leaf = 0.0
- max_features = None
- random_state = None
- max_leaf_nodes = None
```

- `min_impurity_decrease = 0.0`
- `min_impurity_split = None`
- `class_weight = None`
- `presort = False`

- RFC:

It is implemented in the

class

`sklearn.ensemble.RandomForestClassifier`. The standard configuration from Sci-Kit Learn was used:

- `n_estimators = 10`
- `criterion = 'gini'`
- `max_depth = None`
- `min_samples_split = 2`
- `min_samples_leaf = 1`
- `min_weight_fraction_leaf = 0.0`
- `max_features = 'auto'`
- `max_leaf_nodes = None`
- `min_impurity_decrease = 0.0`
- `min_impurity_split = None`
- `bootstrap = True`
- `oob_score = False`
- `n_jobs = 1`
- `random_state = None`
- `verbose = 0`
- `warm_start = False`
- `class_weight = None`

5.4.2.1 Standardization

As explained in Section 4.2.4, to improve the classification performance of ML algorithms, a recommended step is standardization of the data. Here, the class `sklearn.preprocessing.QuantileTransformer` was used. This method works by transforming the feature set to exhibit a *uniform* or a *normal* distribution. This transformation reduces the effects of outliers over the data, which makes it a robust preprocessing approach [30].

The standard configuration present in `Sci-Kit Learn` was used:

- `n_quantiles = 1000`
- `output_distribution = 'uniform'`
- `ignore_implicit_zeros = False`
- `subsample = 100000`
- `random_state = None`
- `copy = True`

5.4.2.2 Feature selection strategy

A method for feature selection was tested in order to find the best algorithm to reach the best feature set for characterization of both clinical outcome groups in the study, for each of the three algorithms used. The mutual information classifier (Section 4.2.4.1) was used as it is implemented in the class `sklearn.feature_selection.mutual_info_classif`.

The following standard parameters were used:

- `discrete_features = 'auto'`
- `n_neighbors = 3`
- `copy=True`
- `random_state = None`
- `n_features = n`

The feature selection step was carried out simultaneously with the classification step itself. This was done inside a `loop`. For `n` (number of desired features, `n_features = n`) varying from 1 to 15, and for each `n`, `mutual_info_classif` filter ran 100 times searching for the best `n` features in the raw dataset (52 features). And for every subset selected, evaluation metrics were computed. In this way, the best feature set was chosen for each classification algorithm individually.

5.4.2.3 Metrics

The metrics *ACC*, *SPE*, *SEN*, *F1-S*, and *AUC* were used to evaluate how good was the classification performance of each algorithm tested. This was used to inform both the classification performance and the most relevant feature set in the prediction of NFPA recurrence. In other words, the best features to help to identify tumors with high regrowth chances after the first resection surgery.

In the next chapter the results obtained from univariate and multivariate analyses will be fully described and discussed.

RESULTS AND DISCUSSION

THE data collection that includes MRI texture analyses for all the 15 patients was done as described in the previous chapter. The patient population, as previously stated, was divided in two groups: the *stable* group (8 patients) defined by those patients whose NFPA lesion did not exhibit any sign of regrowth in the period considered in the study (no signs of PSTR); and, the *recurrent* group (7 patients) defined by the patients who experienced tumor regrowth in the follow-up period (PSTR observed).

In the following sections, these two groups will be described by the measured features, and the results obtained in both univariate and multivariate analyses will be presented and discussed.

6.1 Clinical and textural features for both groups

All the analyses performed were proposed to investigate the existence of differences between these two groups concerning the 52 features (4 clinical and 48 radiological) described in Chapter 6.

Table 6.1 summarizes *age at first surgery* in which patients underwent their first surgical approach and the average *period of clinical follow-up* for both groups.

Feature	Stable: Mean $\pm$ SE	Recurrent: Mean $\pm$ SE
Age at first surgery	55.870 $\pm$ 4.836	44.051 $\pm$ 4.642
Follow-up period	4.512 $\pm$ 0.536	4.640 $\pm$ 0.653

Table 6.1: Age at first surgery and average period follow-up for both patient group with standard error (SE).

The average age in which patients went through the first resection procedure is different between the groups. The average follow-up periods of both patient groups of almost 5 years were similar in both groups. This characterizes the work as an investigation for postoperative short-term recurrence (PSTR).

Figure 6.1 shows patients in the recurrent and stable groups regarding hormonal positive (Pos) or negative (Neg) immunohistochemical studies (LH, ACTH, GH, PRL, and FSH) in (a); according to signs of invasiveness (based on Hardy classification, Figure 1.3) or without any invasiveness observed. Also, in (c), the distribution of gender for both outcome groups is shown.

(a)

(b)

(c)

Figure 6.1: Distribution of the three clinical qualitative information for groups recurrent and stable. *Pos* - Immunohistopositivity for one or more hormone type; *Neg* - Immunonegativity for all the tested hormone. *Inva* - Any level of invasiveness observed second to Hardy classification; *Non - Inva* - absence of invasiveness in Hardy classification. *M* - Male, *F* - Female.

Tables 6.2 to 6.6 describe the 48 quantitative parameters computationally extracted from tumor ROI (Figure 5.1) of both patient groups.

Feature	Stable: Mean $\pm$ SE	Recurrent: Mean $\pm$ SE
Percent 5th	45.125 $\pm$ 5.115	33.571 $\pm$ 5.826
Percent 10th	50.375 $\pm$ 5.425	38.429 $\pm$ 6.003
Percent 25th	58.500 $\pm$ 5.735	47.429 $\pm$ 6.241
Percent 75th	76.625 $\pm$ 6.617	80.571 $\pm$ 3.644
Percent 90th	87.125 $\pm$ 5.097	90.571 $\pm$ 3.605
Quartile	18.125 $\pm$ 1.968	33.143 $\pm$ 7.015
GL Entropy	5.736 $\pm$ 0.146	6.107 $\pm$ 0.153
Mean	67.901 $\pm$ 5.696	65.271 $\pm$ 3.458
Std Deviation	14.706 $\pm$ 0.877	20.349 $\pm$ 2.727
Median	67.547 $\pm$ 6.263	67.556 $\pm$ 4.005
Skewness	0.131 $\pm$ 0.273	0.013 $\pm$ 0.128
Kurtosis	0.596 $\pm$ 0.422	-0.483 $\pm$ 0.258
Variance	221.655 $\pm$ 27.825	458.676 $\pm$ 123.957
Histogram Elongation	1.454 $\pm$ 0.091	1.541 $\pm$ 0.160
Flatness	1.182 $\pm$ 0.029	1.182 $\pm$ 0.036
Minimum	22.750 $\pm$ 5.561	11.143 $\pm$ 5.922
Maximum	114.000 $\pm$ 4.675	120.143 $\pm$ 4.160

Table 6.2: Averaged histogram features extracted for both groups with standard error (SE).

Feature	Stable: Mean $\pm$ SE	Recurrent: Mean $\pm$ SE
Entropy	9.303 $\pm$ 0.320	9.430 $\pm$ 0.233
Energy	0.001 $\pm$ 0.001	0.001 $\pm$ 0.000
Correlation	0.001 $\pm$ 0.000	0.002 $\pm$ 0.000
Haralick Correlation	39397138.750 $\pm$ 6443706.296	67173885.714 $\pm$ 9278598.427
Cluster Prominence	5618006.250 $\pm$ 991916.431	9862800.000 $\pm$ 3522112.118
Cluster Shade	8298.116 $\pm$ 14159.508	-13285.545 $\pm$ 19067.224
Inertia	336.385 $\pm$ 38.103	238.006 $\pm$ 17.593
Inverse Difference Moment	0.096 $\pm$ 0.019	0.103 $\pm$ 0.010

Table 6.3: Averaged GLCM features extracted for both groups with standard error (SE).

Feature	Stable: Mean $\pm$ SE	Recurrent: Mean $\pm$ SE
SRE	0.963 $\pm$ 0.009	0.966 $\pm$ 0.004
LRE	2.146 $\pm$ 0.173	1.630 $\pm$ 0.098
GLN	2068.909 $\pm$ 995.100	3356.736 $\pm$ 1589.421
RLN	59475.775 $\pm$ 14952.967	142445.429 $\pm$ 42618.354
LGRE	0.001 $\pm$ 0.000	0.001 $\pm$ 0.000
HGRE	4548.604 $\pm$ 197.852	4698.240 $\pm$ 148.623
SRLGE	0.001 $\pm$ 0.000	0.001 $\pm$ 0.000
SRHGE	4407.474 $\pm$ 207.599	4547.033 $\pm$ 147.186
LRLGE	0.002 $\pm$ 0.000	0.002 $\pm$ 0.000
LRHGE	9183.068 $\pm$ 638.641	7392.104 $\pm$ 326.570
RP	0.960 $\pm$ 0.011	0.962 $\pm$ 0.005

Table 6.4: Averaged GLRM features extracted for both groups with standard error (SE).

Feature	Stable: Mean $\pm$ SE	Recurrent: Mean $\pm$ SE
Volume In Physical Units	5461.748 $\pm$ 1632.922	12632.523 $\pm$ 3978.771
Size Ratio	2.538 $\pm$ 0.052	4.655 $\pm$ 2.463
Elongation	4.041 $\pm$ 0.309	3.639 $\pm$ 0.346
Feret Diameter	29.466 $\pm$ 3.079	42.309 $\pm$ 5.900
Equiv Spher Perimeter	1412.190 $\pm$ 275.704	2457.477 $\pm$ 532.510
Equiv Spher Radius	10.286 $\pm$ 0.969	13.505 $\pm$ 1.482
Perimeter	1841.118 $\pm$ 349.417	3107.106 $\pm$ 691.313
Roundness	0.769 $\pm$ 0.016	0.798 $\pm$ 0.025

Table 6.5: Averaged shape label features extracted for both groups with standard error (SE).

Feature	Stable: Mean $\pm$ SE	Recurrent: Mean $\pm$ SE
Gradient Mean	10.273 $\pm$ 1.064	10.905 $\pm$ 0.918
Gradient Std	5.772 $\pm$ 0.405	6.874 $\pm$ 1.159
Gradient Variance	34.465 $\pm$ 4.818	55.306 $\pm$ 19.327
Gradient Maximum	40.625 $\pm$ 2.556	44.857 $\pm$ 4.426

Table 6.6: Averaged gradient features extracted for both groups with standard error (SE).

The relevance of presented features to characterize and classify *recurrent* and *stable* patient groups was evaluated with univariate and multivariate statistical approaches and the results will be presented and discussed in the next sections.

6.2 Univariate analysis

The association of the binary clinical features presented in Figure 6.1 were evaluated with Fisher's exact test and the results are present in the following table. Evidences of association were considered for $p - values < 0.05$.

Feature	Fisher's p-value
Immunohistochemistry	0.610
Invasiveness	1.000
Gender	1.000

Table 6.7: Association between immunohistochemistry and invasiveness status with PSTR by Fisher's exact test.

As described in table 6.7, no evidence of association was found between invasiveness and immunohistochemistry with short term recurrence. Also, gender was not found to be associated with PSTR in the data here analyzed, despite some studies had reported it [74].

The relevance of *age at surgery* and the other 48 image descriptors extracted was assessed with Student's t-test and Kruskal-Wallis test. Before those tests, variance (Levine's test) and normality (Shapiro-Wilk's test) of each sample group (*recurrent* and *stable*) were evaluated in order to check the reliability of such tests. Finally, evidence of differences between these two groups according each one of those features was considered for $p - values < 0.05$. The results are as follows.

Figure 6.2: Outcome groups comparison second 'Age-first Surgery'.

Feature	Variance	Normality	CI 95%	tt p-value	KW p-value
Age-first Surgery	Equal	Normal	[-1.775, 25.412]	0.104	0.105

Table 6.8: Age-first Surgery significance. CI: Confidence interval. tt: Student's *t*-test. KW: Kruskal-Wallis test.

Despite some distinction observed between the average age at surgery for both recurrent and stable group, and the graphical difference observed in the boxplot of Figure 6.2, in the data here analyzed no statistical evidence was found to support this difference ($p > 0.05$, Table 6.8). Neither the p-value observed (for both parametric and non-parametric tests) nor CI point association between these two facts.

6.2.1 Imaging texture feature analyses

Out of the 48 features extracted, 9 were found to show statistical evidence of significance in at least one of the tests taken. Following, boxplot distributions for those features are presented and the tests results summarized in a table.

Figure 6.3: Outcome groups comparison second to each one of the relevant features in the univariate analysis.

Feature	Variance	Normality	CI 95%	tt p-value	KW p-value
Quartile	Equal	Normal	[-28.752, -1.284]	0.047	0.032
GL Entropy	Equal	Not			0.049
Kurtosis	Equal	Not			0.021
Haralick Correlation	Equal	Normal	[-49996044.968, -5557448.961]	0.026	0.028
Cluster Prominence	Equal	Not			0.028
Inertia	Equal	Normal	[9.425, 187.333]	0.044	0.028
LRE	Equal	Normal	[0.099, 0.933]	0.027	0.028
LRHGE	Equal	Normal	[277.331, 3304.595]	0.033	0.021
Size Ratio	Unequal	Not			0.028

Table 6.9: *Relevant features found in the univariate analysis. CI: Confidence interval. tt: Student's t-test. KW: Kruskal-Wallis p-value test.*

For all those tests $\alpha = 0.05$ was considered. Student t-test has many ways to be carried out: for dependent or independent samples, and for samples with equal or unequal variance. The sample groups considered here are independent and Levine's test was considered to evaluate if these two groups of samples had equal or different variance, and then the appropriate version of t-test was carried out. All the modalities of t-test are applicable only for samples with distribution close to the normal one.

Considering both groups' size (recurrent group, 7; stable group 8), the assumption of normality cannot be adopted right away. Shapiro-Wilkson test was carried out over both patient groups for every feature. In Table 6.9, the value *Normal* means that both outcome groups showed evidence of normality for the given significance level. The value *Not* means that at least one group did not offer evidence of normality.

To fully understand t-test results it is important to check both p-value and CI . If p-value is smaller than the chosen α it is said that statistical evidence of differences between the two groups concerning that specific feature was found. However, for that to be meaningful, CI must not contain the value 0 ($0 \notin CI$), since CI is the interval that estimates how much is the difference between the two sample means (here is $mean_{stable} - mean_{recurrent}$) with a precision of 95%.

The features *Quartile*, *Haralick Correlation*, *Inertia*, *LRE*, *LRHGE* showed statistical evidence of differences between the two outcome groups, considering the Student's t-test p-value ($< \alpha$) and the CI ($0 \notin CI$). The other four features, *GL Entropy*, *Kurtosis*, *Cluster prominence*, and *Size Ratio*, did not show evidence for Student's t-test, but on the other hand they were not found to present normal distributions by Shapiro-Wilk test, what makes t-test results not trustworthy. A non-parametric two-independent-sample comparison, Kruskal-Wallis test, was carried out as an alternative to the parametric Student's t-test. For this non-parametric test, all those 9 features were found to be significant in the differentiation between *recurrent* and *stable* group.

Those tests for population mean comparison give us a good first step in finding relevant image features related to the studied phenomena. Nevertheless, they are not the endpoint. For that reason, ROC curve analysis was used to evaluate the classification performance of those features found to be relevant (Table 6.9). The found results are described next.

Figure 6.4 brings the plotted ROC curve for each one the features found to be relevant.

Figure 6.4: ROC curves analysis for each one of the relevant features present in Table 6.9.

Moreover, Table 6.10 summarizes all the relevant results for those nine features. In this table t-test p-value and confidence interval for the features that did not show relevance through t-test were omitted since they are not trustworthy. Rather, only the non-parametric p-value is described again.

Feature	CI 95%	tt p-value	KW	AUC
Quartile	[-28.752, -1.284]	0.047	0.032	0.830
GL Entropy	–	–	0.049	0.803
Kurtosis	–	–	0.021	0.857
Haralick Correlation	[-49996044.968, -5557448.961]	0.026	0.028	0.839
Cluster Prominence	–	–	0.028	0.839
Inertia	[9.425, 187.333]	0.044	0.028	0.839
LRE	[0.099, 0.933]	0.027	0.028	0.839
LRHGE	[277.331, 3304.595]	0.033	0.021	0.857
Size Ratio	–	–	0.028	0.839

Table 6.10: Summary of most relevant features found in the univariate analysis. CI: Confidence interval. tt: Student's t-test. KW: Kruskal-Wallis p-value test. AUC - area under the receiver operator characteristic curve.

All of those nine features achieved ROC curve classification above 0.800, which emphasizes the evidence of their prognostic value to predict short-term tumor recurrence in NFPA treatment, considering the available data.

6.3 Multivariate analysis

The whole set of 51 features plus *gender* information was given to three machine learning algorithms (kNN, DTC, RFC) to evaluate the prediction value of those features. Their performances were checked in the task of correctly identifying *recurrent* and *stable* patients based on those features, the best feature subset for such a task, and which one of those classifiers was the best one.

Accuracy (ACC), sensitivity (SEN), specificity (SPE), F1-Score (F1-S), and the area under ROC curve (AUC) was used to assess their performance. The search for the best results and performances in the task here considered was carried by maximizing ACC, SPE and AUC. The best results for each classifier are presented

ahead.

6.3.1 kNN

Figure 6.5: *kNN confusion matrix and ROC curve analysis. ROC - Receiver operator characteristic.*

N. of features	Features	ACC	SEN	SPE	F1-S	AUC
5	GL Entropy, Std Deviation, Inertia, LRE, and Size Ratio	0.933	0.857	1.000	0.923	0.991

Table 6.11: *Best subset for kNN and performance indexes.*

kNN algorithm showed an excellent performance, achieving perfect SPE and conserving very high ACC (0.933) and almost perfect AUC (0.991). Looking at the confusion matrix in Figure 6.5(a) it is possible to see that out of those 15 patients to be classified only 1 was misclassified, and this one belonged to the *recurrent* group and was classified as *stable*. This type of error (false negative) has less consequences since this is the baseline decision in the current clinical practice.

6.3.2 DTC

Figure 6.6: *kNN* confusion matrix and ROC curve analysis.

N. of features	Features	ACC	SEN	SPE	F1-S	AUC
1	GL Entropy	0.800	0.857	0.750	0.800	0.804

Table 6.12: *Best subset for DTC and performance indexes.*

DTC showed an average performance and its best performance considered only one feature. It is a negative point since 52 features were given and that those algorithms are especially suited for multivariate analysis. Yet, the chosen feature, *GL Entropy*, had the worst performance in the univariate analysis with an AUC of 0.803. DTC achieved the same result and uses the same classification strategy, which does not justify its usage for the studied task. And finally, DTC made a dangerous type of error (*false positive*) for two patients (Figure 6.6(a)). It classified as *recurrent* a patient from *stable* group. This type of error is considered dangerous because, in the clinical practice, it would cause the patient to undergo unnecessary risky treatment approaches (e.g. radiotherapy).

6.3.3 RFC

Figure 6.7: *RFC confusion matrix and ROC curve analysis.*

N. of features	Features			ACC	SEN	SPE	F1-S	AUC
10	Quartile,	GL	Entropy,	0.933	0.857	1.000	0.923	0.991
	Std Deviation,		Variance,					
	Haralick		Correlation,					
	Cluster		Prominence,					
	Inertia,	LRE,	LRHGE,					
	Size Ratio							

Table 6.13: *Best subset for RFC and performance indexes.*

RFC achieved a performance similar to kNN, reaching perfect SPE, conserving higher ACC and AUC, and making a single misclassification. In fact, RFC used the same five features used by kNN and 5 additional features to achieve this result, which implies a more complex classification process.

6.4 Closure discussion

Prediction of NFPA regrowth after surgery is in fact very difficult to achieve with high precision, mostly due to the scarce knowledge of NFPA natural history and prognostic factors [22, 75]. Nevertheless, some evidences were found about the

association between NFPA postoperative regrowth and high ki-67 staining indices through hypothesis tests ($p < 0.01$) considering 5 years of period follow-up [76]. In other study, postoperative regrowth for pituitary adenomas in general (mostly functioning) was found to be associated with age ($p = 0.015$) and gender ($P = 0.017$) [74]. Pattern and growth rate were observed to be very similar for patients with different residual tumor volumes [75] which enforces no certainty of recurrence based on the residual lesion.

Although the definition of aggressiveness for NFPA is not clearly defined by WHO, some characteristics are considered as signs of aggressiveness such as PSTR and conventional therapy resistance. In spite of that, no biomarker has consistently been proven to distinguish aggressive from nonaggressive pituitary adenomas [5]. This distinction is essential and extremely valuable since it may improve outcome quality and allow a more precise treatment decision process for each patient.

It is known that radiotherapy offers high tumor control rates, but given its risk of secondary lesions it is not the mainstream adjuvant therapy of choice in NFPA treatment, or even for pituitary adenomas in general. Yet, radiosurgery offers similar tumor control rates with less risks of secondary lesions [5, 22]. Those risks should not preclude the use of those therapy modalities if the choices become more precise and reliable in pituitary adenoma management.

The present computational image analysis study was carried out to investigate the use of MRI texture features and some clinical features (age, gender, invasiveness, and immunohistochemistry) to assess the chances of PSTR for NFPA patients. All of those features were investigated individually with conventional statistics, and machine learning algorithms were used to perform a multivariate analysis considering different subsets of features to predict tumor recurrence in the patient groups included in the study.

Some agreement was found between univariate and multivariate analysis. Out of those 9 features identified as relevant in the univariate analysis, only *Kurtosis* was not considered in the multivariate subsets of kNN and RFC. From the other hand, *Std Deviation* ($tt = 0.057$, $KW = 0.064$) and *Variance* ($tt = 0.068$, $KW = 0.064$) were not detected as relevant, but were found useful in the multivariate analysis. None of the clinical features (age at first surgery, immunohistochemistry,

invasiveness status, and gender) were found to be relevant neither in univariate classification tests nor for machine learning algorithm's performances.

Multivariate entirely overcame univariate tests performances reaching almost perfect classification results ($SPE = 1.000$ and $AUC = 0.991$) for two algorithms (kNN and RFC). That validates and supports the use of machine learning algorithms to compile computationally extracted features into a prediction model for PSTR in patients harboring NFPA.

The statistical analyses presented in these two last sections gave enough evidence that MRI texture features extracted semi-automatically through computer tools may help making predictions of PSTR in NFPA treatment given the available data. Despite those MRI texture feature having no clinical or histological interpretation established, this approach offers the advantage of being extracted computationally out of patient MR images acquired routinely in their treatment follow-up. It does reduce bias present in visual image analysis by specialists and requires no extra expenses with patient management.

The next chapter presents the main conclusions, limitations, and perspectives for this study.

CONCLUSIONS AND FINAL CONSIDERATIONS

ANY cancer disease offers a considerable level of complexity that requires right tools, strategies, and precise information to treat it well. Management of non-functioning pituitary adenomas is as well a challenging task with delicate decisions to be taken. Although its clinical follow-up relies on a multiprofessional team, precise biomarkers are necessary to achieve the desired outcomes.

Given the results obtained in this study, it is possible to conclude that: i) DCE-MRI texture features showed to be highly effective in predicting short-term NFPA recurrence. ii) The prediction value of those features can be observed with both conventional univariate statistical tests and multivariate analyses through machine learning algorithms. iii) Machine learning models showed more effective and convincing classification performances for the clinical task here studied. iv) And, the computerized methods described and evaluated in this study can be useful in pituitary adenomas management supporting the clinical decision process toward a postoperative therapy, i.e. fractionated RT or radiosurgery.

Despite the excellent results found here, limitations surround the achieved knowledge. The number of patients included, and the retrospective study design limitations are the main drawbacks. However, the selection criteria and controlled conditions make the presented results reliable and conclusive. Further studies (e.g. cohorts and multicenter) are extremely encouraged and necessary to confirm this knowledge and make it reach clinical settings.

Nonetheless, this study contributes to the efforts of the scientific community

to improve tumor treatment and to develop a precision medicine. Also, for the 3DSlicer platform users, this work has provided a compact tool for automatic 2D and 3D texture features extraction for medical image with free access and open-source.

In conclusion, medical image texture analysis has shown to be of great value in the clinical decision process in many studies. In fact, it has become a research field itself, the radiomics. In addition, computer aided diagnosis applications, the combination of radiomics with machine learning models, have been shown to be a powerful tool to tackle clinical problems that demand a significant amount of data processing. Hence, the study here developed is relevant since it is consistently aligned with those proposals.

And finally, these methodologies represent an improvement in the usage of MRI that are already there, routinely acquired, non-invasive, and non-ionizing radiation based.

BIBLIOGRAPHY*

- [1] PHILLIPS, T. L. et al. *Leibel and Phillips textbook of radiation oncology*. 3rd. ed. [S.l.]: Elsevier/Saunders, 2010. 477–486 p. ISBN 9781437737752.
- [2] LU, J. J.; BRADY, L. W. *Decision Making in Radiation Oncology*. 2nd. ed. Berlin, Heidelberg: Springer, 2011. 923–940 p. (Medical Radiology). ISBN 978-3-642-16332-6.
- [3] SALVAJOLI, J. V.; SOUHAMI, L.; FARIA, S. L. *Radioterapia Em Oncologia*. 2nd. ed. São Paulo: Atheneu, 2013. ISBN 9788538803812.
- [4] SYMONDS, P. R. P. et al. *Walter and Miller's textbook of radiotherapy : radiation physics, therapy and oncology*. 7th. ed. [S.l.]: Elsevier Churchill Livingstone, 2012. 547–549 p. ISBN 9780702046629.
- [5] Di Ieva, A. et al. Aggressive pituitary adenomas-diagnosis and emerging treatments. *Nat. Rev. Endocrinol.*, Nature Publishing Group, v. 10, n. 7, p. 423–435, 2014. ISSN 1759-5029.
- [6] SAEGER, W. et al. Pathohistological classification of pituitary tumors: 10 years of experience with the German Pituitary Tumor Registry. *Eur. J. Endocrinol.*, v. 156, n. 2, p. 203–216, feb 2007. ISSN 0804-4643.
- [7] LINSLER, S. et al. Prognosis of pituitary adenomas in the early 1970s and today-Is there a benefit of modern surgical techniques and treatment modalities? *Clin. Neurol. Neurosurg.*, v. 156, p. 4–10, may 2017. ISSN 03038467.
- [8] MELMED, S. et al. *Williams textbook of endocrinology*. 13th. ed. [S.l.: s.n.], 2016. 232–299 p. ISBN 9780323341578.

*De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

- [9] RAGEL, B. T.; COULDWELL, W. T. Pituitary carcinoma: a review of the literature. *Neurosurg. Focus*, American Association of Neurological Surgeons, v. 16, n. 4, p. 1–9, apr 2004. ISSN 1092-0684.
- [10] EZZAT, S.; ASA, S. L. Mechanisms of Disease: the pathogenesis of pituitary tumors. *Nat. Clin. Pract. Endocrinol. Metab.*, Nature Publishing Group, v. 2, n. 4, p. 220–230, apr 2006. ISSN 1745-8366.
- [11] ROELFSEMA, F.; BIERMASZ, N. R.; PEREIRA, A. M. Clinical factors involved in the recurrence of pituitary adenomas after surgical remission: a structured review and meta-analysis. *Pituitary*, Springer US, v. 15, n. 1, p. 71–83, 2012.
- [12] TSANG, R. W. et al. Radiation therapy for pituitary adenoma: treatment outcome and prognostic factors. *Int. J. Radiat. Oncol. Biol. Phys.*, Elsevier, v. 30, n. 3, p. 557–65, 1994.
- [13] CURY, M. L. C. d. A. R. et al. Non-functioning pituitary adenomas: clinical feature, laboratorial and imaging assessment, therapeutic management and outcome. *Arq. Bras. Endocrinol. Metabol.*, ABE&M, v. 53, n. 1, p. 31–39, feb 2009. ISSN 0004-2730.
- [14] MINNITI, G. et al. Stereotactic radiotherapy and radiosurgery for non-functioning and secreting pituitary adenomas. *Reports Pract. Oncol. Radiother.*, v. 21, n. 4, p. 370–378, jul 2016. ISSN 15071367.
- [15] BECKER, G. et al. Radiation Therapy in the Multimodal Treatment Approach of Pituitary Adenoma. *Strahlentherapie und Onkol.*, v. 178, n. 4, p. 173–186, apr 2002. ISSN 0179-7158.
- [16] WELSH, L.; AJITHKUMAR, T.; BRADA, M. *Radiotherapy for Pituitary and Hypothalamic Tumours*. [S.l.]: MDText.com, Inc., 2013.
- [17] LITTLE, M. D. et al. Hypopituitarism following external radiotherapy for pituitary tumours in adults. *Q. J. Med.*, v. 70, n. 262, p. 145–60, feb 1989.

- [18] TSANG, R. W. et al. Role of radiation therapy in clinical hormonally-active pituitary adenomas. *Radiother. Oncol.*, Elsevier, v. 41, n. 1, p. 45–53, 1996.
- [19] PUATAWEEPONG, P. et al. Outcomes for Pituitary Adenoma Patients Treated with Linac-Based Stereotactic Radiosurgery and Radiotherapy: a Long Term Experience in Thailand. *Asian Pacific J. Cancer Prev.*, v. 16, n. 13, p. 5279–5284, 2015. ISSN 1513-7368.
- [20] SNYDER, P. J. et al. Hypopituitarism following radiation therapy of pituitary adenomas. *Am. J. Med.*, Elsevier, v. 81, n. 3, p. 457–462, 1986.
- [21] DIALLO, A. M. et al. Long-term results of fractionated stereotactic radiotherapy as third-line treatment in acromegaly. *Endocrine*, v. 50, n. 3, p. 741–748, dec 2015. ISSN 1355-008X.
- [22] CHEN, Y. et al. Natural History of Postoperative Nonfunctioning Pituitary Adenomas: A Systematic Review and Meta-Analysis. *Neuroendocrinology*, v. 96, p. 333–342, 2012.
- [23] BRADA, M. et al. The long-term efficacy of conservative surgery and radiotherapy in the control of pituitary adenomas. *Clin. Endocrinol. (Oxf)*., v. 38, n. 6, p. 571–578, jun 1993. ISSN 0300-0664.
- [24] VARSSEVELD, N. C. van et al. Cerebrovascular Events, Secondary Intracranial Tumors, and Mortality After Radiotherapy for Nonfunctioning Pituitary Adenomas: A Subanalysis From the Dutch National Registry of Growth Hormone Treatment in Adults. *J. Clin. Endocrinol. Metab.*, Oxford University Press, v. 100, n. 3, p. 1104–1112, 2015.
- [25] CHANSON, P. et al. Management of clinically non-functioning pituitary adenoma. *Ann. Endocrinol. (Paris)*., v. 76, n. 3, p. 239–247, jul 2015. ISSN 00034266.
- [26] MILKER-ZABEL, S. et al. Fractionated stereotactically guided radiotherapy and radiosurgery in the treatment of functional and nonfunctional adenomas of the pituitary gland. *Int. J. Radiat. Oncol. Biol. Phys.*, Elsevier, v. 50, n. 5, p. 1279–86, 2001.

- [27] BURGER, W.; BURGE, M. J. *Digital Image Processing*. 1st. ed. London: Springer London, 2008. (Texts in Computer Science). ISBN 978-1-84628-379-6. Disponível em: <<http://link.springer.com/10.1007/978-1-84628-968-2>>.
- [28] CASTELLANO, G. et al. Texture analysis of medical images. *Clin. Radiol.*, W.B. Saunders, v. 59, n. 12, p. 1061–1069, 2004.
- [29] KIMURA, M.; CRUZ, L. C. H. da. Multiparametric MR Imaging in the Assessment of Brain Tumors. *Magn. Reson. Imaging Clin. N. Am.*, v. 24, n. 1, p. 87–122, feb 2016. ISSN 10649689.
- [30] PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, v. 12, p. 2825–2830, 2011.
- [31] BROWN, R. W. et al. *Magnetic Resonance Imaging*. 2nd. ed. Chichester, UK: John Wiley & Sons Ltd, 2014. ISBN 9781118633953.
- [32] JACKSON, A.; BUCKLEY, D. L.; PARKER, G. J. M. *Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Oncology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. (Medical Radiology). ISBN 978-3-540-42322-5.
- [33] SOURBRON, S. P.; BUCKLEY, D. L. Classic models for dynamic contrast-enhanced MRI. *NMR Biomed.*, v. 26, n. 8, p. 1004–1027, aug 2013. ISSN 09523480.
- [34] FLICKINGER, F. W. et al. Differentiation of benign from malignant breast masses by time-intensity evaluation of contrast enhanced MRI. *Magn. Reson. Imaging*, v. 11, n. 5, p. 617–620, jan 1993. ISSN 0730725X.
- [35] STACK, J. P. et al. Breast disease: tissue characterization with Gd-DTPA enhancement profiles. *Radiology*, v. 174, n. 2, p. 491–494, feb 1990. ISSN 0033-8419.
- [36] CHAUDHARY, V.; BANO, S. Imaging of the pituitary: Recent advances. *Indian J. Endocrinol. Metab.*, v. 15, n. 7, p. 216, 2011. ISSN 2230-8210.

- [37] INDRAJIT, I. K. et al. Value of dynamic MRI imaging in pituitary adenomas. *Indian J. Radiol. Imaging*, Indian journal of radiology & imaging, v. 11, n. 4, p. 185–190, 2001.
- [38] COLLEWET, G.; STRZELECKI, M.; MARIETTE, F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. *Magn. Reson. Imaging*, v. 22, n. 1, p. 81–91, jan 2004. ISSN 0730725X.
- [39] SLED, J.; ZIJDENBOS, A.; EVANS, A. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging*, v. 17, n. 1, p. 87–97, feb 1998. ISSN 02780062.
- [40] TUSTISON, N. J. et al. N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging*, v. 29, n. 6, p. 1310–1320, jun 2010. ISSN 0278-0062.
- [41] LERSKI, R. A. et al. MR image texture analysis—an approach to tissue characterization. *Magn. Reson. Imaging*, v. 11, n. 6, p. 873–87, 1993.
- [42] TOURASSI, G. D. Journey toward Computer-aided Diagnosis: Role of Image Texture Analysis. *Radiology*, v. 213, n. 2, p. 317–320, 1999.
- [43] HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man. Cybern.*, SMC-3, n. 6, p. 610–621, nov 1973. ISSN 0018-9472.
- [44] JOHNSON, H. J.; MCCORMICK, M. M.; IBANEZ, L. *The ITK Software Guide Book 1: Introduction and Development Guidelines*. 1. ed. [S.l.]: Kitware Inc, 2015. ISBN 9781-930934-27-6.
- [45] MATERKA, A.; STRZELECKI, M. Texture analysis methods – a review. *Inst. Electron. Tech. Univ. LODZ*, 1998. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.4968>>.
- [46] GALLOWAY, M. M. Texture analysis using gray level run lengths. *Comput. Graph. Image Process.*, v. 4, n. 2, p. 172–179, jun 1975. ISSN 0146664X.

- [47] CHANG, T.; KUO, C.-c. J. Texture analysis and classification with tree-structured wavelet transform. *IEEE TRANS. IMAGE Process.*, v. 2, n. 4, p. 432—435, 1993.
- [48] SHANNON, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, v. 27, p. 379—423, 1948.
- [49] CONNERS, R. W.; HARLOW, C. A. A Theoretical Comparison of Texture Algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-2, n. 3, p. 204–222, 1980.
- [50] CONNERS, R. W.; TRIVEDI, M. M.; HARLOW, C. A. Segmentation of a high-resolution urban scene using texture operators. *Comput. Vision, Graph. Image Process.*, Academic Press, v. 25, n. 3, p. 273–310, 1984.
- [51] YOO, T. et al. Engineering and Algorithm Design for an Image Processing API: A Technical Report on ITK - The Insight Toolkit. *Proc. Med. Meets Virtual Real.*, p. 586—592, 2002.
- [52] TANG, X. Texture information in run-length matrices. *IEEE Trans. Image Process.*, v. 7, n. 11, p. 1602–1609, 1998. ISSN 10577149.
- [53] TUSTISON, N. J.; GEE, J. C. Run Length Matrices For Texture Analysis. *Insight J.*, 2008.
- [54] LEHMANN, G. Label object representation and manipulation with ITK. *Insight J.*, 2008.
- [55] PAGANO, M.; GAUVREAU, K. *Princípios de Bioestatística*. 3rd. ed. São Paulo: Cenage Learning, 2004. 522 p.
- [56] KLOTH, C. et al. Evaluation of Texture Analysis Parameter for Response Prediction in Patients with Hepatocellular Carcinoma Undergoing Drug-eluting Bead Transarterial Chemoembolization (DEB-TACE) Using Biphasic Contrast-enhanced CT Image Data. *Acad. Radiol.*, jun 2017. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/28652049>>.

- [57] CHOI, Y. S. et al. The Added Prognostic Value of Preoperative Dynamic Contrast-Enhanced MRI Histogram Analysis in Patients with Glioblastoma: Analysis of Overall and Progression-Free Survival. *Am. J. Neuroradiol.*, v. 36, n. 12, p. 2235–2241, dec 2015. ISSN 0195-6108.
- [58] XU, X.-Q. et al. Utility of histogram analysis of ADC maps for differentiating orbital tumors. *Diagnostic Interv. Radiol.*, v. 22, n. 2, p. 161–167, jan 2016. ISSN 13053825.
- [59] GHASEMI, A.; ZAHEDIASL, S. Normality tests for statistical analysis: A guide for non-statisticians. *Int. J. Endocrinol. Metab.*, v. 10, n. 2, p. 486–489, 2012. ISSN 1726913X.
- [60] HAJIAN-TILAKI, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J Intern Med*, v. 4, n. 2, p. 627–635, 2013.
- [61] HACKELING, G. *Mastering Machine Learning with scikit-learn*. Birmingham: Packt Publishing, 2014. 238 p. ISBN 9781783988365.
- [62] MITCHELL, T. *Machine Learning*. McGraw-Hill, 1997. ISBN 0070428077. Disponível em: <<http://www.cs.cmu.edu/~tom/mlbook.html>>.
- [63] FETIT, A. E. et al. Three-dimensional textural features of conventional MRI improve diagnostic classification of childhood brain tumours. *NMR Biomed.*, v. 28, n. 9, p. 1174–1184, sep 2015. ISSN 09523480.
- [64] DENNIE, C. et al. Role of quantitative computed tomography texture analysis in the differentiation of primary lung cancer and granulomatous nodules. *Quant. Imaging Med. Surg.*, AME Publications, v. 6, n. 1, p. 6–15, 2016.
- [65] SILVA, F. R. et al. Sensitivity and specificity of machine learning classifiers for glaucoma diagnosis using Spectral Domain OCT and standard automated perimetry. *Arq. Bras. Oftalmol.*, Conselho Brasileiro de Oftalmologia, v. 76, n. 3, p. 170–174, 2013.

- [66] HU, L. S. et al. Multi-Parametric MRI and Texture Analysis to Visualize Spatial Histologic Heterogeneity and Tumor Extent in Glioblastoma. *PLoS One*, v. 10, n. 11, p. e0141506, 2015. ISSN 1932-6203.
- [67] WAUGH, S. A. et al. Magnetic resonance imaging texture analysis classification of primary breast cancer. *Eur. Radiol.*, v. 26, n. 2, p. 322–330, feb 2016. ISSN 0938-7994.
- [68] RASCHKA, S. *Python machine learning*. Birmingham, UK: Packt Publishing, 2015. ISBN 9781783555130. Disponível em: <<https://www.packtpub.com/big-data-and-business-intelligence/python-machine-learning>>.
- [69] JAMES, G. et al. *An Introduction to Statistical Learning*. New York: Springer New York, 2013. (Springer Texts in Statistics, v. 103). ISBN 978-1-4614-7137-0.
- [70] YU, L.; LIU, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res.*, v. 5, p. 1205–1224, 2004.
- [71] KRASKOV, A.; STÖGBAUER, H.; GRASSBERGER, P. Estimating mutual information. *Phys. Rev. E*, American Physical Society, v. 69, n. 6, p. 066138, 2004.
- [72] EGGER, J. et al. Pituitary Adenoma Volumetry with 3D Slicer. *PLoS One*, Public Library of Science, v. 7, n. 12, p. e51788, 2012.
- [73] FEDOROV, A. et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging*, v. 30, n. 9, p. 1323–1341, nov 2012. ISSN 0730725X.
- [74] MONSALVES, E. et al. Growth Patterns of Pituitary Adenomas and Histopathological Correlates. *J. Clin. Endocrinol. Metab.*, v. 99, n. 4, p. 1330–1338, apr 2014. ISSN 0021-972X.
- [75] TANAKA, Y. et al. Growth pattern and rate in residual nonfunctioning pituitary adenomas: correlations among tumor volume doubling time, patient age, and MIB-1 index. *J. Neurosurg.*, v. 98, n. 2, p. 359–365, 2003.

- [76] EKRAMULLAH, S. M. et al. The correlation of Ki-67 staining indices with tumour doubling times in regrowing non-functioning pituitary adenomas. *Acta Neurochir. (Wien)*., v. 138, n. 12, p. 1449–55, 1996.

ETHICS COMMITTEES' APPROVAL LETTERS

The HCFMRP and FFCLRP ethics committee's authorization letter supporting and acknowledging the use of patients clinical information in the present study are attached next.

A.1 HCFMRP ethics committee's approval letter

PARECER CONSUBSTANCIADO DO CEP

Elaborado pela Instituição Coparticipante

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Análise Quantitativa de Imagens de Ressonância Magnética no Tratamento de Macroadenomas de Hipófise

Pesquisador: Leonardo Ferreira Machado

Área Temática:

Versão: 1

CAAE: 62569916.0.3001.5440

Instituição Proponente: Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto- USP

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 1.996.309

Apresentação do Projeto:

Trata-se de projeto de pesquisa da instituição proponente da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto- USP- USP e o HCFMRP_USP será uma coparticipante da pesquisa.

Objetivo da Pesquisa:

De acordo com o parecer da Instituição Proponente.

Avaliação dos Riscos e Benefícios:

De acordo com o parecer da Instituição Proponente.

Comentários e Considerações sobre a Pesquisa:

De acordo com o parecer da Instituição Proponente.

Considerações sobre os Termos de apresentação obrigatória:

Todos os termos de apresentação obrigatória foram anexados na Plataforma Brasil e estão de acordo com as normas éticas vigentes. Apresentou carta de anuência do responsável pelo local de coleta no HCFMRP-USP.

Recomendações:

Não há.

Endereço: CAMPUS UNIVERSITÁRIO
Bairro: MONTE ALEGRE **CEP:** 14.048-900
UF: SP **Município:** RIBEIRÃO PRETO
Telefone: (16)3602-2228 **Fax:** (16)3633-1144 **E-mail:** cep@hcrp.usp.br

Continuação do Parecer: 1.996.309

Conclusões ou Pendências e Lista de Inadequações:

O CEP concorda com o Parecer da instituição proponente e aprova o HCFMRP-USP como coparticipante da pesquisa.

Considerações Finais a critério do CEP:

"O CEP do HC e da FMRP-USP concorda com o parecer ético emitido pelo CEP da Instituição Proponente, que cumpre as Resoluções Éticas Brasileiras, em especial a Resolução CNS 466/12. Diante disso, o HCFMRP-USP, como instituição co-participante do referido projeto de pesquisa, está ciente de suas coresponsabilidades e de seu compromisso no resguardo da segurança e bem-estar dos sujeitos desta pesquisa, dispondo de infra-estrutura necessária para a garantia de tal segurança e bem-estar".

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_831206.pdf	20/02/2017 16:16:08		Aceito
Declaração de Instituição e Infraestrutura	Termo.pdf	20/02/2017 16:03:39	Leonardo Ferreira Machado	Aceito
Orçamento	custos.pdf	20/02/2017 13:35:18	Leonardo Ferreira Machado	Aceito
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_831206.pdf	24/11/2016 18:11:19		Aceito
Projeto Detalhado / Brochura Investigador	ProjetoCEP_Brochura.pdf	24/11/2016 18:10:05	Leonardo Ferreira Machado	Aceito
Declaração de Instituição e Infraestrutura	Termo_de_Concordancia_HCFMRP.pdf	24/11/2016 18:09:02	Leonardo Ferreira Machado	Aceito
Folha de Rosto	FOLHA_DE_ROSTO.pdf	24/11/2016 18:04:39	Leonardo Ferreira Machado	Aceito

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Endereço: CAMPUS UNIVERSITÁRIO
 Bairro: MONTE ALEGRE CEP: 14.048-900
 UF: SP Município: RIBEIRÃO PRETO
 Telefone: (16)3602-2228 Fax: (16)3633-1144 E-mail: cep@hcrp.usp.br

Continuação do Parecer: 1.996.309

RIBEIRAO PRETO, 03 de Abril de 2017

Assinado por:
MARCIA GUIMARÃES VILLANOVA
(Coordenador)

Endereço: CAMPUS UNIVERSITÁRIO
Bairro: MONTE ALEGRE **CEP:** 14.048-900
UF: SP **Município:** RIBEIRAO PRETO
Telefone: (16)3602-2228 **Fax:** (16)3633-1144 **E-mail:** cep@hcrp.usp.br

A.2 FFCLRP ethics committee's approval letter

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Análise Quantitativa de Imagens de Ressonância Magnética no Tratamento de Macroadenomas de Hipófise

Pesquisador: Leonardo Ferreira Machado

Área Temática:

Versão: 2

CAAE: 62569916.0.0000.5407

Instituição Proponente: Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto- USP

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 1.990.990

Apresentação do Projeto:

Ver relatório inicial.

Objetivo da Pesquisa:

Ver relatório inicial.

Avaliação dos Riscos e Benefícios:

Ver relatório inicial.

Comentários e Considerações sobre a Pesquisa:

Ver relatório inicial.

Considerações sobre os Termos de apresentação obrigatória:

Ver relatório inicial.

Recomendações:

Ver relatório inicial.

Conclusões ou Pendências e Lista de Inadequações:

Atendidas as solicitações do CEP.

Considerações Finais a critério do CEP:

Endereço: Av. Bandeirantes 3.900
Bairro: Monte Alegre **CEP:** 14.040-901
UF: SP **Município:** RIBEIRÃO PRETO
Telefone: (16)3315-4811 **Fax:** (16)3315-9101 **E-mail:** coetp@ffclrp.usp.br

Continuação do Parecer: 1.990.990

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_831206.pdf	20/02/2017 16:16:08		Aceito
Declaração de Instituição e Infraestrutura	Termo.pdf	20/02/2017 16:03:39	Leonardo Ferreira Machado	Aceito
Orçamento	custos.pdf	20/02/2017 13:35:18	Leonardo Ferreira Machado	Aceito
Projeto Detalhado / Brochura Investigador	ProjetoCEP_Brochura.pdf	24/11/2016 18:10:05	Leonardo Ferreira Machado	Aceito
Folha de Rosto	FOLHA_DE_ROSTO.pdf	24/11/2016 18:04:39	Leonardo Ferreira Machado	Aceito

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

RIBEIRAO PRETO, 30 de Março de 2017

Assinado por:
Marina Rezende Bazon
(Coordenador)

Endereço: Av. Bandeirantes 3.900
Bairro: Monte Alegre **CEP:** 14.040-901
UF: SP **Município:** RIBEIRAO PRETO
Telefone: (16)3315-4811 **Fax:** (16)3315-9101 **E-mail:** coetp@ffclrp.usp.br