

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-graduação em Física

Egleidson Frederik do Amaral Gomes

**Microscopia multifóton aplicada para a caracterização de câncer
de próstata**

Belo Horizonte
2023

Egleidson Frederik do Amaral Gomes

Microscopia multifóton aplicada para a caracterização de câncer de próstata

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Doutor em Ciências.

Orientador: Ana Maria de Paula

Belo Horizonte

2023

Dados Internacionais de Catalogação na Publicação (CIP)

G633m Gomes, Egleidson Frederik do Amaral.
Microscopia multifóton aplicada para a caracterização de câncer de próstata /
Egleidson Frederik do Amaral Gomes. – 2023.
102 f. : il.

Orientadora: Ana Maria de Paula.
Tese (doutorado) – Universidade Federal de Minas Gerais,
Departamento de Física.
Bibliografia: f. 67-76.

1. Microscopia. 2. Fótons. 3. Câncer. 4. Diagnóstico. I. Título. II. Paula, Ana
Maria de. III. Universidade Federal de Minas Gerais, Departamento de Física.

CDU – 543.456 (043)



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

FOLHA DE APROVAÇÃO

A presente tese, intitulada "**Microscopia multifóton aplicada para a caracterização de câncer de próstata**" de autoria de **EGLEIDSON FREDERIK DO AMARAL GOMES** submetida à Comissão Examinadora, abaixo-assinada, foi aprovada para obtenção do grau de **DOUTOR EM CIÊNCIAS**, em vinte e nove de junho de 2023.

Belo Horizonte, 29 de junho de 2023.

Profa. Ana Maria de Paula
Orientadora do estudante
Departamento de Física/UFMG

Prof. Leandro Malard Moreira
Departamento de Física/UFMG

Profa. Livia Siman Gomes
Departamento de Física/UFMG

Profa. Jaqueline dos Santos Soares
Departamento de Física/UFOP

Prof. Paulo Henrique Almeida Campos-Junior
Departamento de Ciências Naturais/UFSJ



Documento assinado eletronicamente por **Ana Maria de Paula, Professora do Magistério Superior**, em 29/06/2023, às 14:17, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Paulo Henrique de Almeida Campos Junior, Usuário Externo**, em 29/06/2023, às 14:43, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Jaqueline dos Santos Soares, Usuário Externo**, em 29/06/2023, às 15:08, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Livia Siman Gomes, Professora do Magistério Superior**, em 29/06/2023, às 17:05, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Leandro Malard Moreira, Professor do Magistério Superior**, em 06/07/2023, às 06:12, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2427976** e o código CRC **6E52D13C**.

Agradecimentos

Agradeço a meus pais e irmãos pelo suporte ao longo de todo esse período, em especial ao meu irmão Aruanã, você é um orgulho para mim. Esse trabalho não teria sido concluído sem apoio da minha amada namorada Polynne, que tem sido minha musa inspiradora, meu apoio e cuidado. Aos meus filhotes Derfel e Shuri, acho que eu não teria chegado até aqui sem amassar vocês todos os dias.

Aos meus amigos que tanto me ajudam tanto nos momentos de descanso quanto nas discussões a cerca deste trabalho: Paulin, Anne, Lucas, Vitória, Dan Bahia, Sol, Vivis, Gílson e Rada, vocês foram um apoio em muitos dos momentos difíceis que se apresentaram de 2020 para cá.

Agradeço também ao Joda, que me auxiliou com materiais de estatística e proporcionou mais conhecimento sobre esse território menos conhecido.

À Ana Maria, pela orientação durante todos esses anos e por estar sempre buscando meu aprimoramento.

Ao Marcelo Mamede e Eduardo Paulino, pelas discussões sobre a parte biológica e por fornecer as amostras que foram usadas nesse trabalho. Ao Frank Longford pelo desenvolvimento do software PyFibre que permitiu esse trabalho ser concluído.

Aos meus colegas de laboratório que contribuíram em discussões, sugestões e até mesmo nos experimentos no laboratório, alguns nomes em especial se fazem necessário citar: Maria Izabel, Luana, Mychel, Aloísio, Giovanna - vocês contribuíram muito para que esse trabalho fosse concluído.

À Secretaria do departamento de Física por nos orientar em assuntos mais burocráticos e às bibliotecárias do departamento sempre dispostas e solícitas. Aos professores e funcionários em áreas de manutenção que permitem que trabalhemos em um ambiente limpo e tranquilo.

Às agências de fomento, cujo financiamento proporcionou a realização desse trabalho: CAPES, CNPQ, FAPEMIG e INCT Nanomateriais de Carbono.

*“The saddest aspect of life right now is that science
gathers knowledge faster than society gathers wisdom.”*

— Isaac Asimov

Resumo

A cada dia tem crescido o interesse na utilização das técnicas de microscopia não linear, principalmente para análise e imageamento de tecidos biológicos. Dentre essas técnicas, a microscopia por geração de segundo harmônico (SHG) e fotoluminescência por absorção de dois fótons (PL), tem sido as mais importantes e de maiores utilizações quando se trata da análise e prognóstico de tumores. Essa crescente utilização vem do fato de ser uma técnica não invasiva, não necessitar da utilização de corantes ou qualquer outro tipo de preparação nas biópsias podendo ser aplicada em diversas situações como amostras in vivo, tecidos fixados, sistemas in vitro 3D, entre outros. O câncer de próstata tem uma evolução lenta e os principais marcadores prognósticos utilizados atualmente são a idade do paciente, o antígeno prostático específico e o sistema de graduação Gleason que quantifica o grau histológico. No entanto, esses marcadores não conseguem prever com grande acurácia o comportamento biológico do tumor, já que o tumor altera o ambiente a sua volta. Uma modificação relevante acontece no estroma, denominada de estroma reativo. Acredita-se que estudá-lo pode ajudar na graduação histológica e na determinação de prognósticos mais precisos. Neste trabalho utilizamos microscopia por SHG e PL, para estudo de biópsias de tecido de próstata fixadas em lâminas histológicas. As imagens de SHG nos permitem analisar as fibras de colágeno, que sofrem alterações na presença de um tumor, tanto na forma das fibras quanto em sua organização. E as imagens de PL permitem obter informações sobre a parte celular do tecido. Apresentamos resultados para amostras separadas em 3 grupos de análise: o padrão Gleason e outros dois grupos que foram propostos levando-se em consideração as alterações estruturais que ocorrem nos ácinos e no estroma do tecido de próstata na presença de um tumor. Apresentamos também novas metodologias de análises das imagens de SHG e PL pelo uso de um software desenvolvido para separar automaticamente as imagens em regiões de fibras e de células, o PyFibre. As métricas obtidas pelo software, foram propostas com o objetivo de quantificar as mudanças nas estruturas celulares e de estroma observadas pelos patologistas. Com o conjunto de métricas obtido realizamos análises estatísticas por LDA (Linear Discriminant Analysis) que permitiram distinguir entre tecido não neoplásico e áreas de carcinoma com uma precisão de $89 \pm 3\%$, mas entre grupos de Gleason com apenas $46 \pm 6\%$. A análise pelo estroma reativo melhorou a precisão para $65 \pm 5\%$, claramente demonstrando que os parâmetros estromais devem ser considerados como critérios adicionais para um diagnóstico mais preciso.

Palavras-chave: Microscopia; Multifóton; Câncer; Diagnóstico.

Abstract

The interest on the use of non-linear microscopy techniques has increased recently, mainly for the analysis and imaging of biological tissues. Among these techniques, second harmonic generation microscopy (SHG) and two-photon absorption photoluminescence (PL) have been the most important and most widely used when it comes to the analysis and prognosis of tumors. This increased interest is due to the fact that they are non-invasive technique, do not require the use of dyes or any other type of preparation in biopsies allowing the use in different situations such as in vivo samples, fixed tissues, in vitro 3D systems, among others. Prostate cancer has a slow evolution and the main prognostic markers currently used are the patient's age, the prostate-specific antigen and the Gleason grading system that quantifies the histological grade. However, these markers cannot accurately predict the biological behavior of the tumor, as the tumor alters the environment around it. A relevant modification takes place in the stroma which is called reactive stroma. Thus, studying the stroma could help in histological grading and in determining more accurate prognoses. In this work, we used SHG and PL microscopy to study prostate tissue biopsies fixed on histological slides. SHG images allow us to analyze collagen fibers, which undergo changes in the presence of a tumor, both in the shape of the fibers and in their organization. And the PL images allow to obtain information about the cellular part of the tissue. We present results for 3 different analysis groups: the Gleason System and two groups that were proposed taking into account the structural alterations that occur in the acini and stroma of the prostate tissue in the presence of a tumor. We present also new methodologies for analysis of SHG and PL images developing a software to automatically separate images into regions of fibers and cells: PyFibre. The metrics computed by the PyFibre were proposed with the aim of quantifying the changes in cellular and stroma structures observed by pathologists. The set of metrics was able to distinguish between non-neoplastic tissue and carcinoma areas by linear discriminant analysis (LDA) with an accuracy of $89 \pm 3\%$, but between Gleason groups of only $46 \pm 6\%$. The reactive stroma analysis improved the accuracy to $65 \pm 5\%$, clearly demonstrating that stromal parameters should be considered as additional criteria for a more accurate diagnosis.

Keywords: Microscopy; Multiphoton; Cancer; Diagnosis.

Sumário

1	INTRODUÇÃO	12
2	REVISÃO TEÓRICA	15
2.1	Geração de Segundo Harmônico	15
2.2	Fluorescência por Absorção de Dois Fótons	17
2.3	Próstata	18
2.3.1	Adenocarcinoma de Próstata	18
2.4	Colágeno	20
3	METODOLOGIA	22
3.1	Biópsias de Próstata	22
3.1.1	Novas Propostas de Classificação	22
3.1.1.1	Separação pelo tipo de Gleason	23
3.1.1.2	Separação pelo tipo de Ácinos	25
3.1.1.3	Separação pelo tipo de Estroma	25
3.1.2	Amostras	27
3.2	Microscopia de Fluorescência por Absorção de Dois Fótons	29
3.3	Microscopia de Geração de Segundo Harmônico	30
3.4	Análise de Imagens	32
3.4.1	PyFibre	32
3.4.2	Métodos Computacionais	32
3.4.2.1	FIRE	32
3.4.3	Segmentação por Cor e atribuição das regiões de Fibras e Parênquima	35
3.4.4	Parâmetros extraídos das imagens	36
3.5	Análise de Variância e Valor P:	40
3.6	Análise por Discriminante Linear	41
3.6.1	Performance do Modelo	43
4	RESULTADOS E DISCUSSÕES	45
4.1	Imagens PL e SHG	45
4.2	Análise dos resultados para o padrão Gleason	48
4.3	Análise dos resultados para o padrão de Ácinos	51
4.4	Análise dos resultados para o padrão de Estroma	55
4.5	Análises LDA: aprendizado de máquinas	57
4.5.1	LDA da Separação por Gleason	57
4.5.2	LDA da Separação por Ácinos	60

4.5.3	LDA da Separação por Estroma	62
4.5.4	Separação dos Gleasons 3 + 4 e 4 + 3 através do Estroma	64
5	CONCLUSÕES E PERSPECTIVAS	66
	REFERÊNCIAS	68
	APÊNDICE A – PUBLICAÇÕES	78
	APÊNDICE B – GRÁFICOS DE CAIXA	79
B.1	Boxplots do Padrão Gleason	79
B.2	Boxplots do Padrão de Ácinos	84
B.3	Boxplot do Padrão de Estroma	89
	ANEXO A – SCRIPT DE ANÁLISE DOS DADOS	94

1 Introdução

Câncer, atualmente a segunda maior causa de morte em seres humanos [1], designa uma multi-variedade de doenças que podem acometer diversos órgãos e tecidos, sejam de humanos ou animais. Os tumores são caracterizadas por um crescimento anormal e descontrolado de células, que pode se disseminar, por extensão direta a outros órgãos ou à distância com a formação de metástases. O carcinoma de próstata é o segundo tipo de câncer mais comum em homens, e a quinta causa de morte por neoplasia maligna. Segundo a Agência Internacional para Pesquisa em Câncer (International Agency for Research on Cancer, IARC) foram diagnosticados 1,41 milhões de casos de câncer de próstata em 2020 [2]. No Brasil, o Instituto Nacional de Câncer (INCA) estima 704 mil casos de câncer por ano até 2025. O tumor maligno mais incidente no Brasil é o de pele não melanoma (31,3% do total de casos), seguido pelos de mama feminina (10,5%), próstata (10,2%), cólon e reto (6,5%), pulmão (4,6%) e estômago (3,1%) [3]. Em homens, o câncer de próstata é predominante em todas as regiões, totalizando 72 mil casos novos estimados a cada ano do próximo triênio, atrás apenas do câncer de pele não melanoma.

O câncer de próstata tem uma evolução lenta e os principais marcadores prognósticos utilizados atualmente são a idade do paciente, o antígeno prostático específico e a análise visual de lâminas histológicas por patologistas através de microscopia óptica, segundo o sistema de graduação Gleason [4]. O Sistema Gleason, descrito em 1966 por Donald Gleason [4], é utilizado para se definir o grau histológico, que está associado à agressividade do tumor. Além de descrever o grau histológico, este sistema é utilizado para o estabelecimento de prognóstico [4]. No entanto, esses marcadores não conseguem prever com grande acurácia o comportamento biológico do tumor, já que o tumor altera o ambiente a sua volta. Uma modificação relevante acontece no estroma, denominada de estroma reativo [5]. Acredita-se que estudá-lo pode ajudar na graduação histológica e na determinação de prognósticos mais precisos. Estes estudos foram parte dos objetivos deste trabalho de tese.

Recentemente, métodos envolvendo análise quantitativa, que sejam mais rápidos e com menor subjetividade, têm sido estudados visando maior acuidade nos diagnósticos [6]. Dentre essas metodologias, as técnicas de microscopia não linear, a microscopia por geração de segundo harmônico (SHG) e fotoluminescência por absorção de dois fótons (PL), tem sido as mais importantes e de maiores utilizações quando se trata da análise e prognóstico de tumores [7–35]. As principais razões para essa crescente utilização são devido a serem técnicas não invasivas, não necessitar da utilização de corantes ou qualquer outro tipo de preparação nas biópsias, e que podem ser utilizadas em diversas situações como para amostras in vivo, tecidos fixados, sistemas in vitro 3D, entre outros. Dentre as técnicas de

análises, uma das áreas que mais tem despertado interesse, são as que envolvem aprendizado por máquinas, em inglês, Machine Learning (ML) e Deep Learning (DL) [36–38], que utilizam métodos computacionais para entender padrões de mudança estrutural, bem como de evolução do sistema estudado.

O SHG permite detectar a organização estrutural nas fibras de colágeno e possui alta resolução, o que o torna eficiente na análise de amostras biológicas, principalmente nos estudos de tumores em diferentes tecidos, como mama, pâncreas, ovário, próstata, entre outros [39–53]. Alguns dos benefícios que a técnica usufrui é a de ser uma técnica não-invasiva, não necessitar que a amostra seja processada e não necessitar de corantes, podendo ser feita em biópsias recém coletadas, amostras *in vivo* [54], sistemas *in vitro* tridimensionais [14, 55, 56], biópsias de tecidos fixados em lâminas [13, 57], entre outros [58].

No caso específico do adenocarcinoma de próstata, foi observado uma correlação da organização (ou anisotropia) e da linearidade das fibras de colágeno com o comportamento mais agressivo do tumor [42, 52]. E atualmente tem-se buscado uma automatização dos processos de análises [59]. Para isso várias ferramentas computacionais tem surgido, na busca para aumentar a acurácia dos prognósticos, algumas dessas ferramentas disponíveis atualmente são: Fiji plugins of OrientationJ [60], Fibriltools [61], Curve-Align [62–64] e CT-Fire [45, 63]. O que se tem feito até o momento, quando se fala de câncer de próstata, é estudar os padrões e alterações na estrutura do colágeno e correlacionar estes achados com os escores de Gleasons presentes na biópsias analisadas [42, 52, 59]. Ainda são poucos os estudos que mostrem as alterações que também ocorrem no estroma decorrentes da presença de um carcinoma.

Um bom sistema de prognóstico que deseja ser amplamente utilizado necessita de 3 critérios básicos [65]: 1) Habilidade prognóstica que exceda os parâmetros clínicos, 2) Ser reprodutível por outros patologistas e 3) Os resultados desse sistema em biópsias escolhidas aleatoriamente devem representar de forma suficiente todo o tumor.

O objetivo pretendido com esse trabalho era o desenvolvimento de uma ferramenta para o diagnóstico do adenocarcinoma de próstata com algumas destas características. A ferramenta deve considerar as mudanças estruturais que podem ocorrer em tecidos nos quais existe um tumor, levando a maior acurácia. Visamos obter padrões como forma dos ácinos, aumento ou diminuição na quantidade de células, aumento na quantidade das fibras de colágeno, bem como da linearidade e a anisotropia das fibras de colágeno, conforme citado anteriormente.

Utilizamos as microscopias por SHG e PL para estudar as mudanças estruturais em amostras de tecido de próstata. O objetivo geral do trabalho foi o desenvolvimento de metodologias de obtenção de imagens e novos métodos de análise para a caracterização automatizada de biópsias. As imagens foram analisadas por métodos computacionais para a obtenção de diversas métricas que levam em conta tanto a organização das fibras de

colágeno bem como do arranjo celular nas amostras. O estudo envolve o desenvolvimento de software para realizar análises automáticas das imagens de SHG e PL, de forma a otimizar todo o processo de extração de parâmetros dos tecidos estudados, visando obter dados sobre a organização das células neoplásicas e da matriz extracelular nos tecidos. As análises trouxeram parâmetros promissores para se avaliar os tecidos, demonstrando haver padrões de organização e distribuição do colágeno que se associam mais a determinados arranjos arquiteturais de células neoplásicas.

Este trabalho está organizado como a seguir. O capítulo 2 apresenta uma revisão teórica da geração de segundo harmônico, da fluorescência por absorção de dois fótons, da glândula prostática, do adenocarcinoma de próstata e do colágeno. No capítulo 3 descrevo os procedimentos clínicos e laboratoriais para a obtenção das biópsias e do imageamento por SHG e PL, os métodos computacionais e de análises que foram utilizados no trabalho, e a descrição dos parâmetros extraídos pelo software desenvolvido: PyFibre. Os resultados obtidos no estudo são apresentados e discutidos no capítulo 4 e ao fim do estudo apresento as conclusões e perspectivas do trabalho, no capítulo 5. No apêndice A estão listadas as publicações referentes ao trabalho e no apêndice B mostramos em gráficos de caixa todas as métricas obtidas através do PyFibre. Os detalhes mais técnicos sobre o script escrito para as análises está apresentado no anexo A.

2 Revisão Teórica

Neste capítulo apresentamos uma breve revisão da teoria envolvida na geração de segundo harmônico (SHG) e da fluorescência por absorção de dois fótons (PL). Mostramos o desenvolvimento das relações que permitem a ocorrência de segundo harmônico em materiais não centro-simétricos, como é o caso das fibras de colágeno encontradas nas biópsias de próstata, bem como da PL. Apresentamos também alguns detalhes do material a ser estudado, a glândula prostática e o colágeno.

2.1 Geração de Segundo Harmônico

Das equações de Maxwell, temos:

$$\nabla \cdot \vec{D} = 4\pi\rho. \quad (2.1)$$

$$\nabla \cdot \vec{B} = 0. \quad (2.2)$$

$$\nabla \times \vec{E} = -\frac{1}{c} \frac{\partial \vec{B}}{\partial t}. \quad (2.3)$$

$$\nabla \times \vec{H} = \frac{1}{c} \frac{\partial \vec{D}}{\partial t} + \frac{4\pi\vec{J}}{c}. \quad (2.4)$$

Supondo um material que possua permeabilidade unitária, ou seja, $\vec{B} = \vec{H}$ e que não existem correntes livres, $\rho = \vec{J} = 0$, para facilitarmos os cálculos. Os campos $\vec{D}$ e $\vec{E}$, estão relacionados pela seguinte equação:

$$\vec{D} = \vec{E} + 4\pi\vec{P}, \quad (2.5)$$

onde $\vec{P}$ é a polarização do meio não-linear. Das equações 2.1 - 2.4 com a 2.5, é possível escrever a equação de onda que irá nos fornecer a propagação do campo elétrico no meio material.

$$\nabla \times \nabla \times \vec{E} + \frac{1}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} = \frac{4\pi}{c^2} \frac{\partial^2 \vec{P}}{\partial t^2}. \quad (2.6)$$

Isso mostra que a resposta de polarização $\vec{P}$ em um material pode ser interpretada também como geradora do campo elétrico, da mesma forma que nas equações de Maxwell, o campo $\vec{B}$ gera o campo $\vec{E}$ e vice-versa. Para os casos não-lineares, ou seja, na presença de um campo elétrico que seja muito intenso é possível expandir a equação de polarização em potências do campo elétrico $\vec{E}$ [66]:

$$\vec{P} = \epsilon_0(\chi^{(1)}\vec{E} + \chi^{(2)}\vec{E} \cdot \vec{E} + \chi^{(3)}\vec{E} \cdot \vec{E} \cdot \vec{E} + \dots), \quad (2.7)$$

onde $\chi^{(n)}$ é o tensor susceptibilidade elétrica do material de ordem n e ϵ_0 é a permissividade elétrica no vácuo. Vamos considerar no presente caso apenas o termo de segunda ordem, responsável pela geração de segundo harmônico e para esse caso podemos escrever a polarização não-linear $\vec{P}^{(2)}(2\omega)$, como [66]:

$$\vec{P}^{(2)}(2\omega) = \epsilon_0 \chi_{ijk}^{(2)}(2\omega, \omega, \omega) \vec{E}(\omega) \cdot \vec{E}(\omega). \quad (2.8)$$

Para facilitar os cálculos, vamos escrever o campo elétrico como $\vec{E}_n(\omega) = \mathcal{E}_n(\vec{r})e^{-i\omega t} + c.c$, onde $n = (x, y, z)$, substituindo o campo elétrico na equação 2.8 e considerando uma propagação na direção z , temos:

$$P_i^{(2)}(2\omega) = \epsilon_0 \chi_{ijk}^{(2)}(2\omega, \omega, \omega) [\mathcal{E}_i(\vec{r})\mathcal{E}_j(\vec{r})e^{-i2\omega t} + \mathcal{E}_i(\vec{r})\mathcal{E}_j^*(\vec{r}) + c.c] \quad (2.9)$$

Podemos ver na equação 2.9 que a polarização tem um termo $(\mathcal{E}_i(\vec{r})\mathcal{E}_j(\vec{r})e^{-i2\omega t})$ que oscila com o dobro da frequência incidente, que chamamos de segundo harmônico, bem como um termo de ordem zero $(\mathcal{E}_i(\vec{r})\mathcal{E}_j^*(\vec{r}))$, que chamamos de retificação ótica.

Para que haja geração de segundo harmônico, é necessário que o material não tenha simetria de inversão, isso é necessário para que o tensor $\chi^{(2)}$ tenha termos não-nulos. Para materiais com centro de inversão, ao trocarmos o sentido do campo elétrico na eq. 2.7 para $-E$, teríamos que também alterar o sentido da polarização $P \rightarrow -P$, temos assim:

$$-\vec{P}^{(2)}(2\omega) = \epsilon_0 \chi_{ijk}^{(2)}(2\omega, \omega, \omega)(-\vec{E}(\omega) \cdot -\vec{E}(\omega)), \quad (2.10)$$

de onde temos:

$$-\vec{P}^{(2)}(2\omega) = \epsilon_0 \chi_{ijk}^{(2)}(2\omega, \omega, \omega)(\vec{E}(\omega) \cdot \vec{E}(\omega)). \quad (2.11)$$

Em materiais com simetrias de inversão precisamos, para que a expressão $\vec{P} = -\vec{P}$ seja válida, que o tensor χ^2 seja nulo: $\chi^2 = 0$. Como para que ocorra a geração de segundo harmônico é necessário que o termo de segunda ordem na eq. 2.9 seja diferente de zero, isso implica que materiais centrossimétricos não possuem SHG.

A figura 1 mostra de forma simplificada, como ocorre a interação do campo elétrico, com uma frequência ω que incide em um material com tensor de segunda ordem não-nulo. A resposta resultante da polarização é um campo elétrico com o dobro da frequência incidente e um campo elétrico resultante da interação de primeira ordem.

Esse é um resultado geral, que pode ser aplicado a qualquer material ao qual incidimos um campo elétrico intenso, que nos permita calcular a polarização e a intensidade do segundo harmônico. A primeira observação experimental do fenômeno de geração de

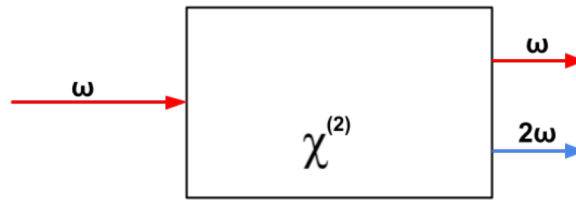


Figura 1 – Simplificação esquemática da interação de um campo elétrico com frequência ω , resultando em uma resposta de polarização de segunda ordem 2ω e uma resposta resultante da interação de primeira ordem.

segundo harmônico, foi obtida em 1961 por Franken et al. [67] ao incidir um laser de rubi em um cristal de quartzo.

Neste trabalho o SHG foi utilizado para estudar as fibras de colágeno em biópsias de próstata, já que as mesmas não possuem simetria de inversão.

2.2 Fluorescência por Absorção de Dois Fótons

Previsto inicialmente em 1931 de forma teórica por Maria Göppert-Mayer em seu trabalho de doutorado [68], no qual discorre sobre a teoria de transições quânticas de dois fótons, a excitação por absorção de dois fótons só foi observada experimentalmente 30 anos depois. Em seu trabalho denominado Two-Photon Excitation in $\text{CaF}_2:\text{Eu}^{2+}$ [69], Kaiser e Garret descrevem a emissão de luz azul por fluorescência, num comprimento de aproximadamente 425 nm excitando cristais de $\text{CaF}_2:\text{Eu}^{2+}$ com uma luz vermelha (694,3 nm) gerada por um laser de rubi. Mas foi apenas na década de 90 que a técnica de microscopia de fluorescência por absorção de dois fótons foi desenvolvida [70].

A excitação de um material por dois fótons é um processo ótico não-linear de terceira ordem, na qual o material absorve de forma (quasi) simultânea os dois fótons, que possuem energia combinadas que permitam, pela regra de seleção [41, 71, 72], uma transição do estado fundamental para um estado excitado. Os fótons podem ter energias iguais, ou seja, $2E_1$ ou energias diferentes, $E_1 + E_2$, oriundas, por exemplo, da combinação de dois lasers diferentes. A Figura 2 mostra de forma esquemática como ocorre esse tipo de transição.

O que ocorre nesse tipo de situação é que um fóton com uma determinada energia E_1 excita uma molécula para um estado virtual intermediário, a molécula pode permanecer nesse estado por período de tempo que varia entre 10^{-15} s e 10^{-16} s (tempo estimado pelo princípio da incerteza para fótons de energia na faixa do visível e infravermelho próximo) [71]. Se durante esse intervalo um outro fóton com uma energia E_2 interagir com a molécula, ela pode ser levada, novamente caso seja obedecida a regra de seleção, para um determinado estado excitado [41, 71, 72].

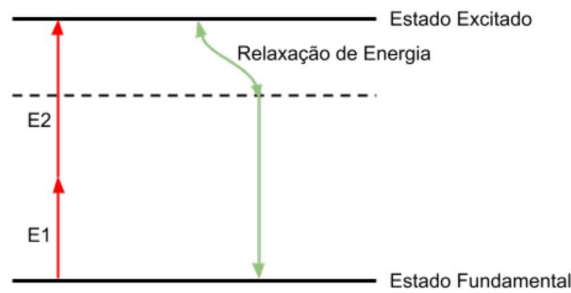


Figura 2 – Esquema simplificado de absorção por dois fótons e posterior decaimento com emissão de fluorescência.

2.3 Próstata

A próstata, mostrada na figura 3, é uma glândula exócrina do sistema reprodutor masculino, ou seja, responsável pela secreção de parte do líquido do sêmen [73]. Esta secreção é básica e ajuda a neutralizar a acidez do canal vaginal, além de possibilitar uma maior motilidade dos espermatozoides [73].

A próstata fica localizada logo abaixo da bexiga urinária e envolve parte do canal uretral. É formada por tecidos glandulares e conjuntivo, sendo este último composto de tecido fibroso e muscular [73]. Os elementos fundamentais para o funcionamento da próstata formam o parênquima prostático, constituído pelas glândulas tubuloacinares ou tubuloalveolares. O outro componente importante é o estroma, que é a estrutura de sustentação do parênquima, sendo composto por tecido conjuntivo (fibras elásticas e colágenas) e especialmente músculo liso [74], como mostrado na figura 4. As células epiteliais, que secretam os seus produtos no lúmen acinar, estão situadas junto à lâmina basal da membrana basal e contém células basais entre estes dois componentes. Também estão presentes células neuroendócrinas intercaladas que regulam a secreção da próstata. Externo aos ácinos tem-se o compartimento do estroma, constituído por vasos sanguíneos, matriz extracelular (MEC) e vários outros elementos celulares, incluindo-se células musculares lisas, fibroblastos, terminações de nervos do sistema nervoso autônomo, células do sistema imunológico, dentre outras [73].

2.3.1 Adenocarcinoma de Próstata

Usualmente os tumores malignos na próstata são classificados como adenocarcinomas, ou tumores glandulares. Estes ocorrem quando as células que produzem a secreção do esperma sofrem mutações e se tornam células tumorais [75]. Os tumores surgem das células epiteliais secretoras (perfil secretor) e na grande maioria das vezes representam adenocarcinomas de padrão acinar [75].

Biópsia transretal da próstata guiada por ultrassonografia é o método mais empre-

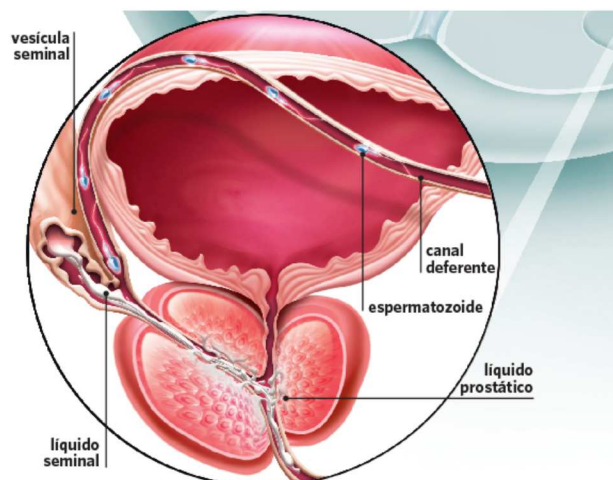


Figura 3 – Imagem mostrando a localização da próstata no corpo humano e a produção do líquido seminal (imagem retirada de <https://saude.abril.com.br/wp-content/uploads/2019/10/sqs-grafico1.png> no dia 14 de fevereiro de 2021).

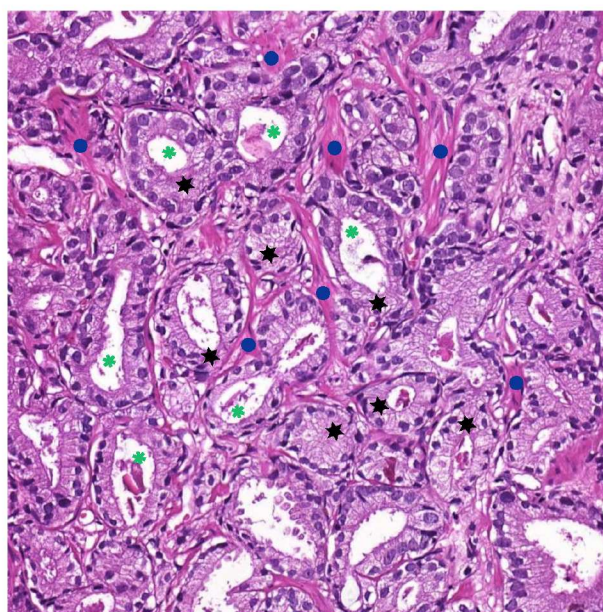


Figura 4 – Na figura temos a imagem de uma biópsia de próstata corada com hematoxilina e eosina (HE). As estrelas pretas mostram os ácinos que estão presentes no tecido, os asteriscos em verde mostram a região do lúmen, que compõem o parênquima prostático e os símbolos cheios azuis identificam a região do estroma, formado principalmente por fibras musculares e fibras de colágeno.

gado como procedimento padrão para o diagnóstico de câncer de próstata. Atualmente, envolve a coleta sistemática de 12 amostras de biópsia por agulha da zona periférica, sítio preferencial das neoplasias, 6 de cada lado da glândula, duas em cada região anatômica (ápice, terço médio e base), eventualmente colhendo-se amostras complementares de nódulos suspeitos ou da zona de transição [76]. O diagnóstico das biópsias após processamento histológico. O tipo mais comum de tumor maligno é o chamado adenocarcinoma de padrão acinar, que surge das glândulas externas da próstata. A agressividade do tumor está relacionada clinicamente com o seu grau de diferenciação ou grau histológico. Nesse aspecto, um tumor pode ser bem diferenciado, moderadamente diferenciado ou pouco diferenciado [76].

O tumor também altera o ambiente estromal, essa alteração é definida como Estroma Reativo (ER) [5]. O estroma se altera para tentar manter o equilíbrio na glândula prostática. Por conta do crescimento tumoral dá-se também o aumento de células inflamatórias que podem se espalhar pelo tecido. Isto ocorre devido a modificação da matriz extracelular ocasionada por esse novo ambiente estromal [5].

2.4 Colágeno

O colágeno é a proteína mais abundante do corpo humano, correspondendo a cerca de 6% do peso corporal e possui importância fundamental na constituição da matriz extracelular, sendo também um dos grandes responsáveis pelas propriedades físicas do tecido conjuntivo.

A figura 5 mostra uma representação das fibras de colágeno.

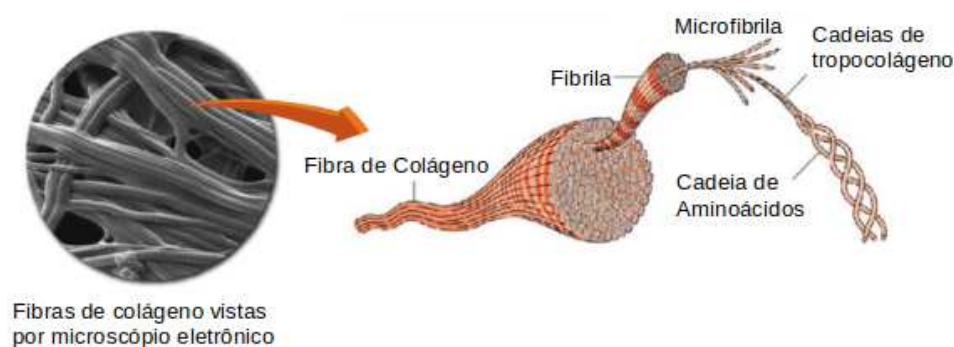


Figura 5 – Fibra de colágeno e as subdivisões estruturais mostrando desde a tripla hélice de aminoácidos até a fibra formada (imagem retirada e adaptada de <https://cnho.files.wordpress.com/2016/06/fibra-de-colageno.png> no dia 14 de fevereiro de 2021).

As cadeias de peptídeos formadas por aminoácidos como glicina, prolina e hidroxiprolina ligam-se entre si em uma estrutura de tripla hélice chamada de tropocolágeno, que possuem cerca de 300 nm de comprimento e 1,5 nm de diâmetro. O tropocolágeno por sua

vez, se organiza em agrupamentos de cerca de $1\ \mu\text{m}$ de comprimento, denominado fibrilas e estas, por fim, se agrupam e formam as fibras de colágeno.

Existem cerca de 28 agrupamentos diferentes das fibras de colágeno, cada um com uma função específica. Os tipos mais comuns de colágeno na próstata, são os colágenos tipo I e tipo III, que são fibrosos e o IV que é não fibrilar [5].

3 Metodologia

3.1 Biópsias de Próstata

As biópsias de próstata foram obtidas pelo grupo do Prof. Marcelo Mamede e Eduardo Paulino Junior da Faculdade de Medicina da Universidade Federal de Minas Gerais. E a utilização para estudos por imagens foi previamente aprovado pelo Comitê de Ética e Pesquisa da UFMG (COEP-UFMG) - CAAE 09288112.0.0000.5149.

Neste estudo utilizamos as lâminas histológicas com as preparações tradicionais para comparação dos resultados com o diagnóstico realizado pelo patologista. O diagnóstico tradicional é feito por lâminas histológicas, nas quais os cortes de tecido são fixadas em formol tamponado e processadas por técnicas histológicas, em seguida as amostras são colocadas em blocos de parafina. A partir dos blocos de parafina são feitos os cortes histológicos de 3 a 5 μm de espessura, e a marcação com corante Hematoxilina e Eosina (HE). Em seguida as lâminas de interesse foram digitalizadas, pelo patologista Eduardo Paulino Júnior, utilizando um escaneador de lâminas modelo Pannoramic MIDI – 3DHISTECH® (Hungary) no qual foi medida a extensão linear das amostras em milímetros e a área do tumor foi mapeada e medida em mm^2 .

As áreas a serem estudadas, por imagens de geração de segundo harmônico (SHG) e fotoluminescência por absorção de dois fótons (PL), foram marcadas pelo patologista. Quando na amostra haviam áreas com a presença de mais de um tipo de padrão carcinogênico ou uma área muito extensa de tumor, foram selecionadas um maior número de regiões de análise. Ao todo para esse trabalho foram analisadas 120 lâminas, cada uma correspondendo a um paciente. Nessas lâminas, foram separadas pelo patologista, um número total de 351 regiões, como mostrado na figura 6. Para cada área foram feitas dois conjuntos de imagens uma de SHG e uma de PL, durante a aquisição de dados.

3.1.1 Novas Propostas de Classificação

Nesta seção descrevo as separações propostas pelo patologista Eduardo Paulino Junior em sua tese de doutorado “Caracterização do Estroma Reativo Do Câncer de Próstata em Função Da Morfologia Acinar e Da Assinatura Do Colágeno por Imuno-histoquímica e À Microscopia Por Geração De Segundo Harmônico (SHG): A Busca De Novos Marcadores Prognósticos” [77]. Início pelo padrão ouro de classificação histológica para tumores de próstata, o padrão Gleason, já comentado previamente na seção 2.3.1, em seguida serão apresentadas as novas classificações por ácinos e por estroma reativo, respectivamente.

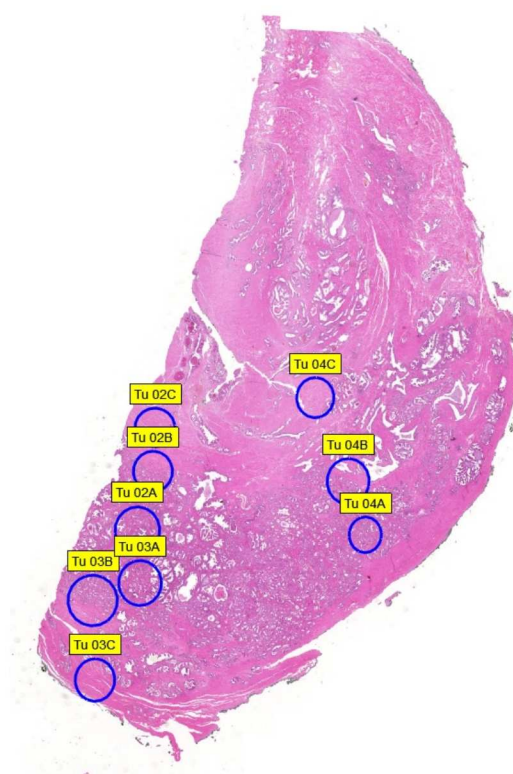


Figura 6 – Imagem de uma biópsia de próstata, utilizando o scanner de lâminas, a imagem mostra as áreas selecionadas pelo patologista para as análises de PL e SHG

3.1.1.1 Separação pelo tipo de Gleason

Proposto em 1966 por Donald F. Gleason [4], o sistema Gleason é atualmente o preditor utilizado para prognósticos de carcinomas de próstata. Foram determinados 5 padrões histológicos que podem ser observados através de biópsias coradas em HE, essa é a base do sistema Gleason.

Esses 5 padrões histológicos são convertidos em escores definidos de 1 a 5, que são dados para os diferentes arranjos dos ácinos neoplásicos observados num determinado tumor, sendo 1 e 2 aqueles tumores em que os arranjos celulares mais se assemelham aos ácinos normais e 5 aquelas lesões com perda completa da arquitetura acinar, havendo arranjos sólidos ou células isoladas. O escore final ou escore de Gleason representa a somatória dos dois padrões mais frequentes na amostra. Exemplo: tumor 3+4 (7), sendo 3 o padrão acinar mais frequente, ou Gleason primário, e 4 o segundo mais frequente, ou Gleason secundário. Vale ressaltar que tumores 3+4 (7) e 4+3 (7) devem apresentar prognósticos diferentes segundo a classificação, mesmo apresentando escores finais iguais, já que em 4+3 observamos maior frequência de lesões Gleason 4, teoricamente mais associadas a comportamento clínico desfavorável, com menor sobrevida livre de doença [78].

Desde a sua idealização em 1966, o sistema de Gleason passou por várias alterações. Em 2013, Epstein, J. I., et al. [79] propôs um sistema de classificação baseado nos padrões

Gleason, denominado de grupos prognósticos, onde temos:

- grupo prognóstico 1: Contempla todas as Somas Gleason, cujo resultado é ≤ 6 ;
- grupo prognóstico 2: Composto pela Soma Gleason 3+4 ;
- grupo prognóstico 3: Composto pela Soma Gleason 4+3;
- grupo prognóstico 4: Somatório dos escore Gleason cuja soma = 8.
- grupo prognóstico 5: Contempla as Somas Gleason, em que o resultado é ≥ 9

A figura 7, retirada de *Prostate cancer grading: a decade after the 2005 modified system* [80], mostra os padrões dos ácinos neoplásicos para os padrões Gleason 3 a 5.

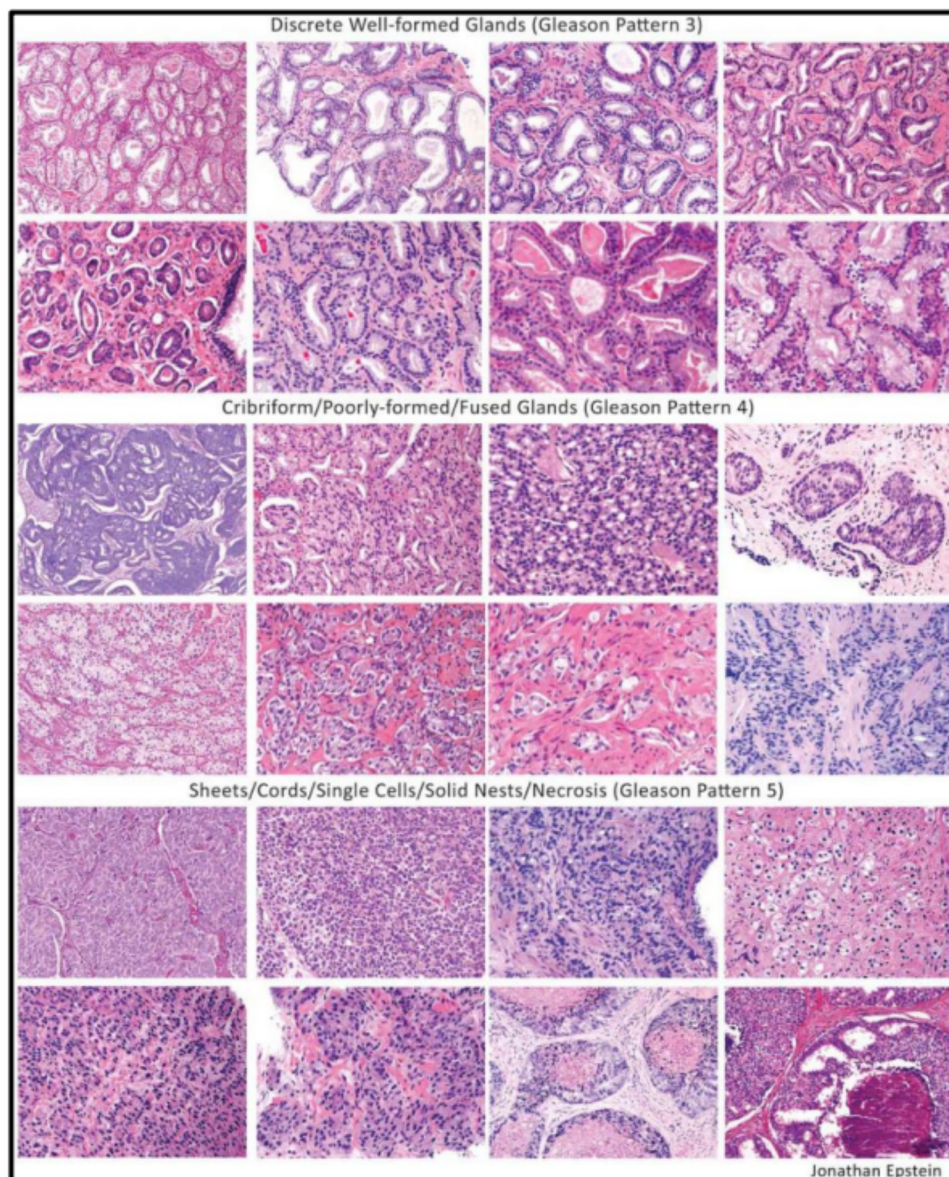


Figura 7 – Padrões Gleason 3, 4 e 5. Retirada de [80]

3.1.1.2 Separação pelo tipo de Ácinos

O padrão de separação por diferenciação acinar foi pensado levando-se em conta as mudanças na morfologia que são observadas na classificação Gleason. Aqui a proposta é a de categorizar de forma mais acurada essas mudanças e apresenta-las de forma mais simplificada e sem a necessidade de uma soma. Foram categorizados pelo patologista 13 padrões de ácinos neoplásicos, que são mostrados na figura 8 retirada de [77].

A descrição dos ácinos foi dividida em 5 tipos, de acordo com os arranjos descritos na figura 8 que são encontrados no corte histológico, denominados A, B, C, D, da seguinte forma:

Padrão A: Possui Ácinos bem formados.

Padrão B: Apresenta Ácinos Rudimentares.

Padrão C: Presença de Padrão cribriforme típico puro.

Padrão D: Composto por um Padrão Sólido e células isoladas.

Padrão E: Padrão misto, composto de mais de um dos padrões anteriores.

O padrão A consiste de ácinos neoplásicos bem diferenciados sem qualquer tipo de fusão, e de tamanhos variados (Arranjos 1 e 8). O Padrão B consiste de ácinos com estruturas rudimentares ou diferenciados, que podem estar fundidos ou não e separados por estroma (Arranjos 2, 6 e 12). O Padrão C possui um padrão cribriforme puro, que pode ser entendido como um aglomerado circular ou alongado de ácinos e é caracterizado pelo crescimento de células neoplásicas dentro do lúmen (Arranjo 4). O padrão D é constituído por tumores constituídos de células isoladas, arranjos compactos puros e/ou células isoladas e quase não forma ácinos (Arranjos 3, 5 e 13). O padrão E que corresponde aos arranjos 7, 9, 10, 11, não foi utilizado nas análises desse trabalho devido ao fato de ser um padrão misto, o que dificulta uma separação e melhor compreensão dos dados e com isso traz poucas informações relevantes.

3.1.1.3 Separação pelo tipo de Estroma

Como descrito brevemente na seção 2.3.1, o Estroma Reativo (ER) é uma resposta no sentido de manter a homeostase do sistema, o nome é dado pela forma análoga que essa resposta ocorre com a que acontece no processo de cicatrização de feridas.

As modificações causadas pela presença de um tumor no estroma, principalmente nas alterações das fibras colágeno, tem um papel muito importante na migração das células tumorais, facilitando a sua propagação e posterior metástase. A região de estroma é composta principalmente de fibras de colágeno e fibras musculares, assim nas análises de SHG pode-se observar as alterações que ocorrem nas fibras de colágeno derivadas da

TABELA 1 - ARRANJOS ARQUITETURAIS DOS ÁCINOS NEOPLÁSICOS.

ARRANJOS		CARACTERÍSTICAS
PADRÕES ÚNICOS	1	Ácinos neoplásicos bem diferenciados, pequenos, médios ou grandes, sem fusão, arredondados ou alongados
	2	Ácinos rudimentares, de lúmens virtuais, com ou sem fusão, isolados ou agrupados, separados por estroma
	3	Tumores constituídos de células isoladas ou dispostas em fileiras ou traves, sem ácidos desenvolvidos.
	4	Arranjos cribriformes puros
	5	Arranjos compactos/sólidos puros
PADRÕES MISTOS	6	Ácinos neoplásicos bem diferenciados e rudimentares, com ou sem fusão
	7	Ácinos neoplásicos bem diferenciados e rudimentares, com arranjos cribriformes
	8	Estruturas tubulares e traves ramificadas
	9	Ácinos neoplásicos rudimentares e arranjos cribriformes
	10	Ácinos neoplásicos bem diferenciados e arranjos cribriformes
	11	Ácinos neoplásicos rudimentares e arranjos cribriformes, com células isoladas
	12	Ácinos neoplásicos rudimentares e células isoladas
	13	Células isoladas e arranjos sólidos

Figura 8 – Tabela mostrando os 13 tipos acinares neoplásicos categorizados pelo patologista. Retirada de [77]

presença do carcinoma. Portanto, isso nos permite visualizar e quantificar a intensidade de reação estromal presente, devido aos fatores mencionados pode ser um bom marcador da atividade tumoral.

O ER foi caracterizado pelo patologista, em termos da reação tumoral presente em cada uma das regiões. Tem-se dessa forma 3 grupos distintos separados conforme mostrado abaixo:

Estroma 1 - Possui reação tumoral escassa ou ausente e tem uma relação de estroma para quantidade de células neoplásicas $< 5\%$.

Estroma 2 - Possui reação tumoral moderada, temos uma proporção de estroma reativo para células neoplásicas $\geq 5\%$ e $< 50\%$.

Estroma 3 - Possui reação tumoral acentuada, no qual a relação entre estroma reativo e as células neoplásicas é $\geq 50\%$.

A figura 9 retirada de [77], mostra os padrões de estroma prostático de acordo com o tipo de reação estromal. A imagem A mostra um padrão normal no qual a predominância do músculo liso(marcados por HE num tom rosa mais escuro), na imagem B temos um padrão de reação estromal do tipo 1 na qual já podemos notar um crescimento na quantidade das fibras de colágeno(tom rosa claro). A imagem C, mostra um estroma reativo do tipo 2 na qual podemos perceber um aumento na predominância das fibras de colágeno e diminuição drástica de músculo liso e, a imagem D que mostra o ER do tipo 3, na qual as fibras de colágeno compõem um valor maior ou igual a 50% da massa tumoral.

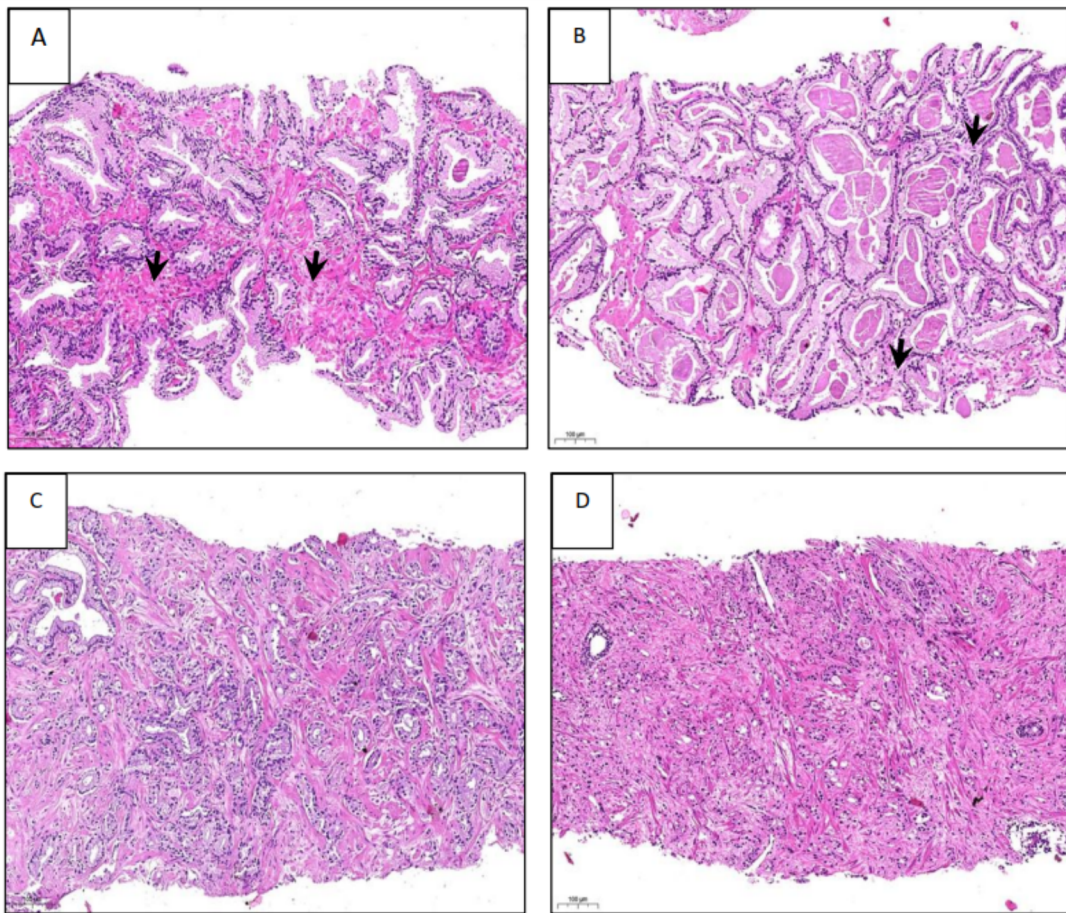


Figura 9 – Na imagem A temos um estroma prostático normal, no qual há uma predominância de músculo liso e poucas fibras de colágeno (marcadas com a seta). A imagem B mostra um tumor com reação estromal do tipo 1, na qual já podemos perceber um aumento na quantidade de colágeno(setas). Na imagem C um tumor com reação estromal moderada, ou seja do tipo 2, com um maior domínio de fibras de colágeno em relação ao músculo liso. E por fim, na imagem D, a presença de um tumor que causa uma reação estromal do tipo 3, no qual temos a massa tumoral composta de $> 50\%$ de colágeno. Retirada de [77]

3.1.2 Amostras

Para cada conjunto de separações foram realizadas medidas de microscopia multi-fótons, essas separações e o número de regiões totais medidas para cada subgrupo estão

listados na tabela 1. Na classificação pelo sistema Gleason, os somatório de escore Gleason com somas ≥ 8 , foram agrupados na coluna 4 + 4+.

Tabela 1 – Quantidade de áreas medidas para cada tipo de Gleason, Ácinos e Estroma reativo.

Normal	3+3	3+4	4+3	4+4+	Total
40	67	73	79	92	351
11,4%	19,1%	20,8%	22,5%	26,2%	100%
Ácinos A	Ácinos B	Ácinos C	Ácinos D	Ácinos E	Total
117	97	20	49	28	311
37,62%	31,2%	6,43%	15,75%	9%	100%
Estroma 1	Estroma 2	Estroma 3	Total		
93	192	26	311		
30,1%	61,5%	8,4%	100%		

3.2 Microscopia de Fluorescência por Absorção de Dois Fótons

O imageamento para as medidas de PL foi realizado utilizando um microscópio de varredura confocal (modelo FV-300 da Olympus) modificado de forma a ser possível o acoplamento de um laser pulsado para excitação dos processos multifótons. As imagens PL e SHG são obtidas pelo mesmo microscópio, em configurações específicas para cada caso. Para excitar a amostra utilizamos um laser de Ti:Safira pulsado (modelo Chameleon, da Coherent Inc.), sintonizável entre 650-1080 nm com pulsos de 140 fs.

A montagem experimental é feita alinhando-se o laser por um conjunto de espelhos tanto interna como externamente ao microscópio, de forma que seja possível fazer medidas em todas as fotomultiplicadoras (PMTs) presentes na montagem. Em seguida é ajustada a potência do laser por uma placa $\frac{\lambda}{2}$ e um polarizador externos ao microscópio antes de ser direcionado ao mesmo.

A Figura 10 mostra um esquema simplificado da montagem experimental e ilustra como é feita a captura das imagens de PL. Para a obtenção das imagens o laser foi sintonizado em 800 nm (que excita o corante eosina na amostra) com uma potência média na amostra de 8 mW. A baixa potência foi escolhida para não danificar as amostras e permitir excitar o corante por absorção de dois fótons. Está mostrado apenas o laser ao entrar no microscópio de varredura, após a montagem externa de direcionamento do laser.

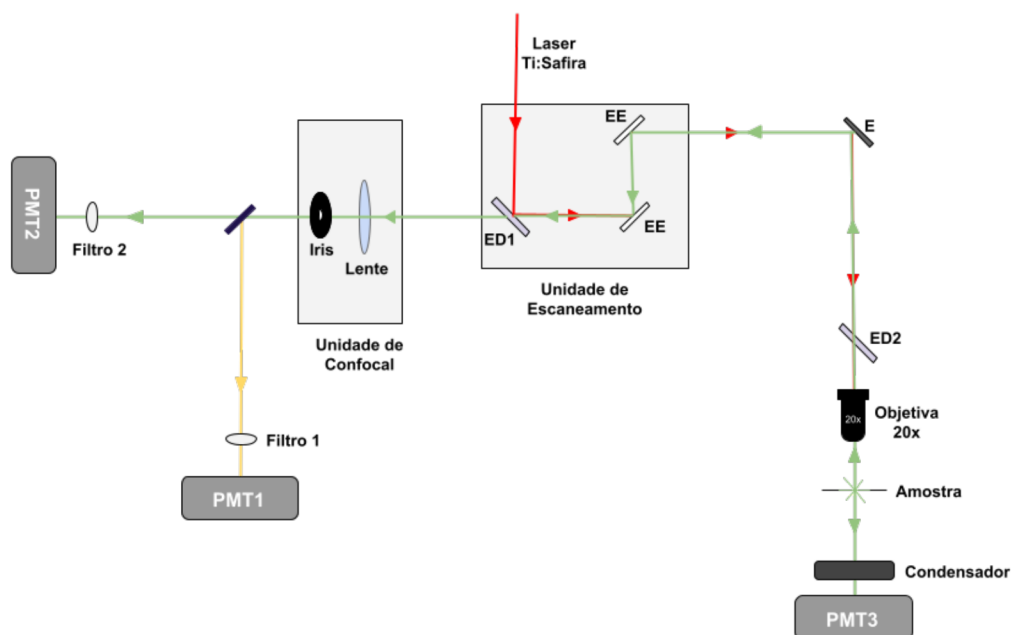


Figura 10 – Montagem esquemática para obtenção das imagens de PL.

Ao entrar no microscópio de varredura, o laser é direcionado por um espelho dicroico (ED1) e por um conjunto de dois espelhos móveis (EE) que são responsáveis por varrer o laser nas direções xy na amostra, o que nos permite obter uma imagem ponto a ponto do sinal na amostra. Após a unidade de escaneamento o laser é direcionado pelo espelho

(E), para ser focado na amostra pela objetiva. Nesse trabalho utilizamos uma objetiva com aumento de $20\times$ e abertura numérica de 0,95. A parte da luz que é transmitida pela amostra passa por uma lente no interior do condensador e é focalizada em uma fotomultiplicadora (PMT3), onde obtemos uma imagem de transmissão da amostra.

A fluorescência emitida pela amostra é coletada em retroespalhamento pela mesma objetiva que focaliza o laser de excitação. Em seguida, o sinal coletado passa pelo espelho dicróico ED1, que reflete o laser em 800 nm e transmite o sinal de PL para a unidade confocal. Este sinal passa por uma iris que é utilizada para barrar a luz proveniente de planos fora do plano focal. Em seguida, o feixe passa por um outro espelho dicróico (ED2), que reflete os comprimentos de onda menores que 505 nm, que é usado para separar o sinal da fluorescência que será medido nas fotomultiplicadoras PMT1 e PMT2. Na frente das fotomultiplicadoras, são utilizados filtros com diferentes faixas espectrais para selecionar as diferentes regiões de comprimento de onda da PL a serem medidas.

Nos experimentos feitos nesse trabalho utilizamos um filtro passa banda (560-600 nm) foi colocado antes da fotomultiplicadora de forma a eliminar algum possível laser espalhado no processo e selecionar a banda de fotoluminescência a ser coletada pela PMT2. A Eosina usada na coloração das biópsias apresenta a fluorescência nessa faixa de comprimento de onda. Foram coletadas imagens com 512×512 píxel², onde, para o caso da objetiva de $20\times$, cada píxel corresponde a $0,921 \mu\text{m}$, ou seja, obtemos uma imagem de $0,472 \times 0,472 \text{ mm}^2$. Vale aqui ressaltar, que em tecidos não corados, a autofluorescência da amostra pode ser utilizada para as análises de PL.

3.3 Microscopia de Geração de Segundo Harmônico

As medidas de SHG foram feitas nas mesmas condições e na mesma região das amostras onde foram coletadas as imagens de PL. O comprimento de onda do laser foi sintonizado também em 800 nm, de forma a termos o segundo harmônico centrado em 400 nm. A potência na amostra foi de 8 mW e com uma fluência de $7 \times 10^{-4} \text{ J/cm}^2$.

De forma semelhante ao que foi descrito na seção 3.2, o laser após passar pela unidade externa de alinhamento, é direcionado para ser focado na amostra pelo espelho ED1. Logo após passar pela unidade de escaneamento, o laser incide em um outro espelho dicróico (ED2), que é colocado de forma a transmitir o laser e refletir o SHG para a unidade fotomultiplicadora (PMT4), como mostrado na figura 11.

Após a interação com a amostra, coletamos a luz gerada por SHG por retroespalhamento pela mesma objetiva utilizada para focalização. A luz retro espalhada é refletida pelo dicróico ED2 para a fotomultiplicadora (PMT4). Na frente do detector PMT4 colocamos um filtro passa-banda centrado em 400 nm e largura de banda de 20 nm de forma a filtrar qualquer luz de laser espalhada pela amostra. Tanto as imagens de SHG quanto as

imagens de PL foram obtidas variando-se também a profundidade na amostra no plano z em $\pm 1 \mu m$ em relação ao plano focal da imagem.

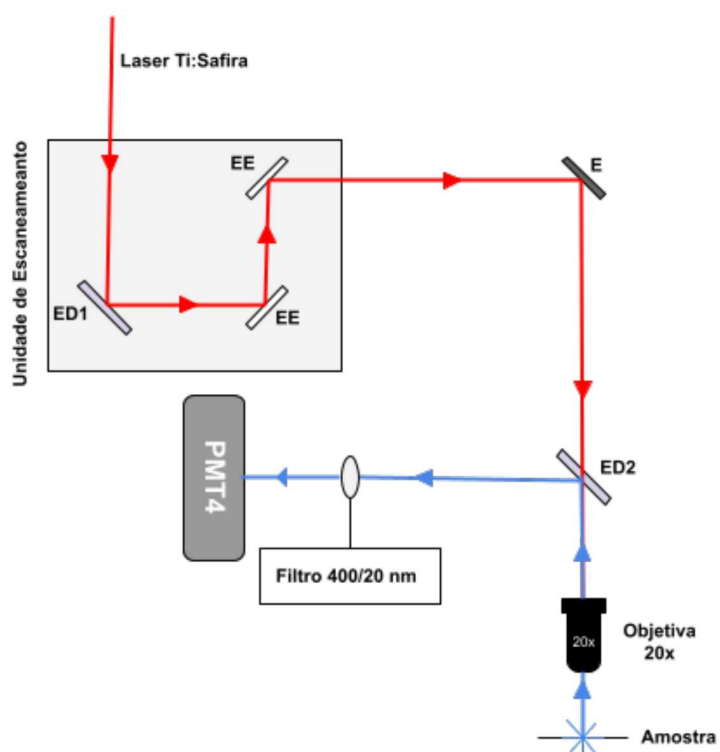


Figura 11 – Montagem esquemática para obtenção das imagens de Segundo Harmônico.

3.4 Análise de Imagens

Nesta seção apresentamos o software desenvolvido para a análise automática das imagens e descrevemos os métodos matemáticos utilizados para o cálculo dos parâmetros medidos para caracterizar os tecidos estudados.

3.4.1 PyFibre

O Software PyFibre (Python Fibrous Image Analysis Toolkit), escrito em python, foi desenvolvido em conjunto entre nosso grupo de pesquisa e pesquisadores da Universidade de Southampton (Inglaterra). O software combina novas propostas de análises dos tecidos tumorais e alguns métodos computacionais já previamente estabelecidos na literatura, que possibilita a extração de fibras de colágeno das imagens de SHG. Ele está disponível na plataforma GitHub [81].

O processo de desenvolvimento do software se deu da seguinte forma: baseado nas imagens de SHG e PL das biópsias obtidas, tanto de próstata quanto de tecidos de câncer de mama em cadelas (parte de outros estudos realizados por nosso grupo de pesquisa) foi discutido com os patologistas o que era preciso medir a partir dessas imagens (Fibras de colágeno, áreas dos ácinos, área de parênquima, número de fibras, número de segmentos de parênquima, entre outros). Desenvolvemos então a metodologia a ser utilizada para os cálculos, e a parte de programação e escrita do software para obter essas medidas foram feitas pelo Dr. Frank Longford, doutorando na Universidade de Southampton à época, durante uma visita de colaboração com o nosso grupo. A medida que o estudo e análise das amostras foram avançando, realizamos alguns refinamentos para obtermos um software mais eficiente para a extração dos dados.

3.4.2 Métodos Computacionais

3.4.2.1 FIRE

O FIRE [82] é um algoritmo desenvolvido para extrair redes de fibras de regiões fibrosa em uma imagem. As redes de fibras, sejam elas quais forem, são consideradas como grafos nos quais as pontas das fibras funcionam como limites entre os nós de diversas regiões que elas conectam. Um grafo pode ser considerado como uma matriz adjunta $n \times n$ que registra cada vez que um vértice que se interpõe entre um nó.

A figura 12 representa de forma simplificada como essa extração é feita, a rede completa é definida pelos seguintes parâmetros: vértices $\mathcal{N} = (1, \dots, p)$, coordenadas $\chi = (\vec{u}_1, \dots, \vec{u}_n)$ e pelo número de fibras conectadas em cada rede representadas pelos vetores $\vec{f}$, e dessa forma $\mathcal{F} = (\vec{f}_1, \dots, \vec{f}_m)$. Para o caso mostrado na figura temos vértices $\mathcal{N} = (1, \dots, 8)$,

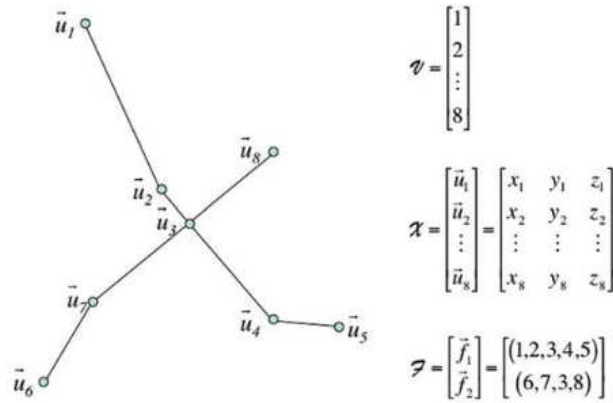


Figura 12 – Ilustração de como é escolhida a rede de fibras no FIRE, imagem retirada de [82]

coordenadas $\chi = (u_1, \dots, u_8)$ em que cada um dos vetores $\vec{u}$ tem suas coordenadas x , y e z e pelo número de fibras presentes $\vec{\mathcal{F}} = (\vec{f}_1, \vec{f}_2)$.

Na implementação do software, foi utilizada uma versão modificada do FIRE, de forma a melhorar resolução na extração da rede de fibras. Nossa versão do algoritmo utiliza um filtro SATO [83] simples na imagem a ser analisada de forma a melhorar a visualização de qualquer estrutura semelhante a um tubo, esse filtro se baseia nos autovalores Hessianos de cada píxel da imagem.

A versão desenvolvida por nós também possui a capacidade de criar um vértice entre qualquer conjunto de nós, o que é importante por nos permitir criar links cruzados dentro da própria rede determinada pelo algoritmo, ao invés de somente inferir que esses links estão presentes nos cruzamentos de vértices e por fim também definimos um comprimento mínimo r_0 no tamanho do vértice, que nos permite alterar a resolução de qualquer fibra que é extraída pelo software. A figura 13 mostra um exemplo de imagem extraída pelo software PyFibre para a rede fibras.

Para gerar a imagem que corresponde a figura 14, que mostra as fibras de forma individual e nos permite calcular as propriedades dessas fibras, como ondulação, comprimento, anisotropia, temos que definir o que é uma fibra. Para isto usamos uma designação parecida com a definida pelo algoritmo FIRE [82]. Primeiramente o software identifica todos os nós externos que contém apenas um vértice. Esses nós são denominados de “nós parentais” para as fibras. O primeiro nó a se conectar através de um vértice com um nó parental é chamado de “nó primogênito”, o algoritmo então analisa todos os nós ligados ao primogênito e traça as fibras tendo como base as regras impostas pelo algoritmo FIRE. Caso não seja possível traçar nenhum outro nó que se conecte ao nó propagante, seja por distância ou ângulo, o algoritmo considera o fim da fibra atual e uma nova fibra será traçada. Ao fim de todo o processo computacional, o software retorna uma imagem, como mostrada na figura 14, em que cada cor representa uma fibra presente na imagem.

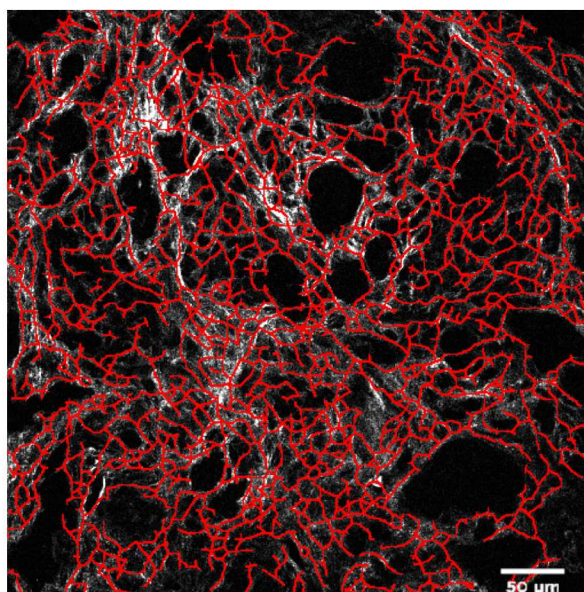


Figura 13 – Rede completa de fibras, mostrada em vermelho superpondo a imagem de SHG em tons de cinza, obtida pelo software PyFibre.

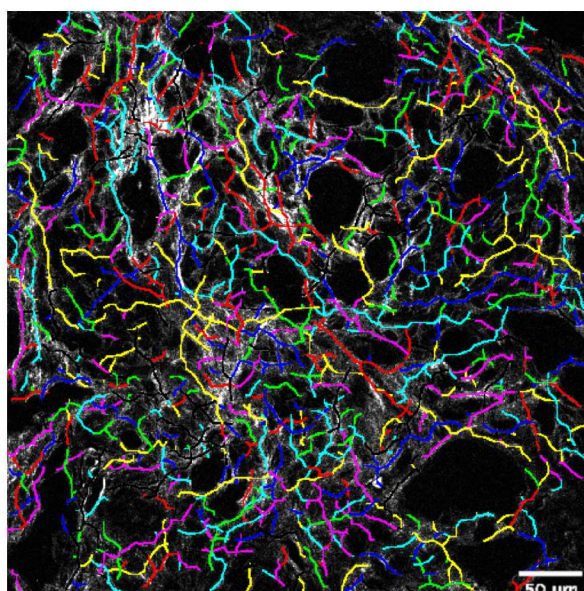


Figura 14 – Fibras que são extraídas utilizando uma versão modificada do FIRE [82] pelo software PyFibre, superpondo a imagem de SHG em tons de cinza. As cores das fibras na imagem são atribuídas de forma aleatória pelo programa.

3.4.3 Segmentação por Cor e atribuição das regiões de Fibras e Parênquima

Após a extração da rede de fibras, é possível obter uma aproximação estimada das bordas entre as regiões de fibras e células. Isto é feito criando-se uma máscara binária que representa a identificação inicial dos segmentos de fibras, baseado na localização de cada rede de colágeno. Para melhorar essa estimativa, utilizamos comparações entre as imagens de SHG, PL e de transmissão, que são feitas em uma escala de cinzas. Criamos um método para gerar uma imagem do tipo RGB com uma coloração padronizada, as fibras de colágeno aparecerão como vetores unitários com maior componente na direção da cor vermelha, enquanto os elementos do parênquima resultarão em vetores unitários com maiores componentes nas direções das cores verde e azul.

Um segundo método bastante comum de segmentação utilizado em imagens coradas com HE, utiliza o algoritmo “BDcreationHE” do software CurveAlign [62], que faz um agrupamento por semelhança de cor de cada píxel da imagem de RGB, não importando a localização desse píxel na imagem. Cada uma de nossas imagens possui 512×512 píxels, e temos 3 tipos diferentes de microscopia, de forma que, como o conjunto total dos dados possui $515^2 \times 3$ pontos, empregamos também um algoritmo de agrupamento do tipo “K-means”. Por questão de eficiência computacional, utilizamos o “MiniBatchKmeans” implementado pelo “sci-kit learn”, vale ressaltar que como o K-means é um algoritmo estocástico, ele não necessariamente fornecerá a melhor solução a cada iteração. A figura 15(a) mostra os segmentos de fibras e na figura 15(b) temos os segmentos de parênquima, ambos extraídos pelo software PyFibre, utilizando os métodos acima.

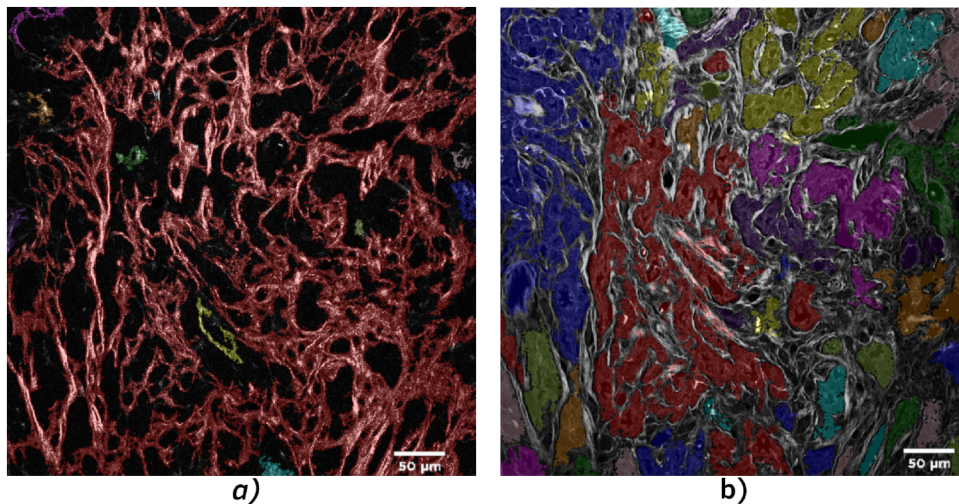


Figura 15 – Segmentos de fibras e parênquima extraídos pelo PyFibre: a) segmentos de fibras, superpondo a imagem de SHG em tons de cinza. b) segmentos de parênquima, superpondo a imagem de PL em tons de cinza. As cores nas imagens são atribuídas de forma aleatória.

3.4.4 Parâmetros extraídos das imagens

Após a segmentação da imagem em regiões de fibra e de células, o PyFibre nos permite obter 29 parâmetros que serão as bases para as análises presentes na seção 4. Apresentamos a seguir todos os parâmetros calculados. Para alguns dos parâmetros, serão explicados também os métodos computacionais que permitem o cálculo dos mesmos. Para os outros com cálculos mais diretos, a explicação é apresentada na tabela 2.

Tabela 2 – Métricas obtidas através do software e suas descrições.

Nome	Descrição	Grupo
No. Segments	Número de segmentos na imagem.	Global
No. Fibras	Número de fibras extraídas na imagem.	Global
Area	Área dos segmentos em pix^2 .	Segmento
Coverage	Razão entre área de pixels cobertos pelos segmentos pela área total da imagem.	Segmento
Intensity Mean	Intensidade média de todos os pixels que pertencem as regiões de segmentos na imagem.	Segmento
Intensity STD	Desvio Padrão da intensidade de todos os pixels que pertencem as regiões segmentadas.	Segmento
Intensity Entropy	Entropia de Shannon padrão da intensidade de Pixel de todos os pixels pertencentes as regiões segmentadas.	Segmento
Angle SDI	Razão da média dos valores máximos na distribuição angular de pixels.	Segmento
Coherence	Coerência do tensor de estrutura construído de todos os pixels das regiões segmentadas.	Segmento
Local Coherence	Coerência média dos tensores de estrutura para cada pixel das regiões segmentadas.	Segmento
Circularity	A circunferência de um círculo com área igual a área do segmento dividido pelo perímetro do segmento.	Segmento
Eccentricity	A distância focal de um elipsoide com a mesma matrix de covariância que o segmento dividido pelo eixo de maior comprimento.	Segmento
Fibre Length	Comprimento médio das fibras extraídas.	Fibra
Fibre Waviness	Ondulação média (comprimento por deslocamento) da fibra extraída.	Fibra
Fibre Network Degree	Número médio de conexões em cada nó da rede reduzida do FIRE.	Rede
Fibre Network Eigenvalue	Autovalor máximo da matriz adjacente da rede reduzida.	Rede
Fibre Network Connectivity	Conectividade da rede reduzida.	Rede
Fibre Network Cross-Link Density	Número médio de interseções cruzadas por fibra extraída na rede.	Rede

Coerência e Tensor Estrutural:

O cálculo da coerência, ou anisotropia, é baseado no tensor estrutural. Definimos a coerência como o grau de organização de um sistema, de forma que, um sistema altamente organizado terá um valor elevado de coerência e terá um valor baixo caso apresente um alto grau de desordem. No caso das análises de biópsias dos tecidos de próstata, a coerência está relacionada com a organização estrutural das fibras de colágeno. Usualmente as fibras de colágeno no tecido de próstata, apresentam baixa coerência não possuindo nenhuma direção preferencial, mas em algumas circunstancias, como na presença de um tumor, pode ocorrer um alinhamento das fibras de colágeno aumentando assim o valor da coerência [42].

Para calcularmos a coerência para o conjunto de imagens definimos primeiramente o tensor estrutural. Consideramos então a imagem de SHG ou PL, como uma matriz em duas dimensões $f(x,y)$ da intensidade do sinal medido. A direção de maior variação da intensidade será a direção ortogonal à direção preferencial de orientação das fibras de colágeno.

O tensor estrutural então para o caso de uma imagem em 2D, será uma matriz 2×2 , com as derivadas parciais ao longo de x e y , que contém as informações das orientações de todos os pixels da imagem [84]. O tensor estrutural é definido da seguinte forma [84]:

$$J = \begin{bmatrix} \langle f_x, f_x \rangle_w & \langle f_x, f_y \rangle_w \\ \langle f_x, f_y \rangle_w & \langle f_y, f_y \rangle_w \end{bmatrix} \quad (3.1)$$

onde f_x e f_y , são as derivadas espaciais da imagem $f(x,y)$, respectivamente ao longo de x e y . Da relação acima, definimos o produto interno do tensor estrutural, como:

$$\langle g, h \rangle_w = \iint_{R^2} w(x,y) g(x,y) h(x,y) dx dy$$

onde $w(x,y)$ é uma função que nos diz qual a resolução da área espacial de interesse da imagem $f(x,y)$. A derivada do tensor estrutural é maximizada na direção do autovalor máximo λ_{max} e minimizada na direção ortogonal que corresponde ao autovalor mínimo λ_{min} , é com essas informações que podemos determinar a direção do alinhamento das fibras de colágeno nas imagens SHG [84]. A direção das fibras, que chamaremos de θ , será dada nesse caso, por:

$$\theta = \frac{1}{2} \arctan \left(\frac{2\langle f_x, f_y \rangle_w}{\langle f_y, f_y \rangle_w - \langle f_x, f_x \rangle_w} \right) \quad (3.2)$$

E definimos a anisotropia ou coerência (n), como a razão entre a diferença dos valores máximo e mínimo do autovalor pela soma desses autovalores [60,85], sendo assim

temos:

$$n = \left(\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} \right)^2 = \frac{\sqrt{(\langle f_y, f_y \rangle_w - \langle f_x, f_x \rangle_w)^2 + 4\langle f_x, f_y \rangle_w}}{\langle f_y, f_y \rangle_w + \langle f_x, f_x \rangle_w} \quad (3.3)$$

Dessa forma os valores totais da coerência variam entre 0 e 1, sendo 0 o valor para um arranjo não alinhado, ou seja, fibras distribuídas aleatoriamente, e 1 pra um arranjo totalmente alinhado, ou seja, as fibras estão distribuídas em uma mesma direção.

Circularidade: O cálculo da circularidade tanto dos segmentos de fibra quanto dos segmentos de parênquima, mede o quão circular é o objeto da medida, e é definido da seguinte forma:

$$\text{circularidade} = \frac{\text{perímetro da área do segmento}}{\text{perímetro de círculo com a mesma área}} \quad (3.4)$$

Dessa forma, quanto mais próximo de 1, mais circular é o segmento de fibra ou de parênquima, e quanto mais próximo de zero, mais linear.

Entropia de Shannon: A entropia de Shannon é uma medida da quantidade de informação presente numa variável ou, em termos computacionais, da quantidade de bits necessários para armazenar essa informação [86], o cálculo da entropia de Shannon é dado por:

$$E = - \sum_n (p(n) \times \log_2(p(n))) \quad (3.5)$$

onde $p(n)$ é probabilidade que a informação n ocorre no sistema, por exemplo, ao jogar uma moeda não viciada, a informação “cara” (“coroa”) pode ocorrer metade das vezes, dessa forma a quantidade de bits necessários para armazenar a informação de uma moeda jogada seria, $E = 1$ bit.

Ondulação: A ondulação das fibras é calculada tomando-se a razão entre distância linear do ponto inicial ao ponto final da fibra, pelo comprimento total da fibra, como mostrado na figura 16, a distância L_0 (linha tracejada) representa o comprimento de ponta a ponta da fibra linearmente, enquanto a distância L_1 (linha contínua) é o comprimento total da fibra.

A ondulação das fibras então, será dada por:

$$P_s = \frac{L_0}{L_1} \quad (3.6)$$

O valor de P_s está confinado entre 0 e 1, se a fibra for totalmente linear a ondulação é igual a 1, e quanto mais ondulada mais próximo de zero será o valor de P_s [60]

Espectro Angular das Fibras: O PyFibre também fornece o espectro angular da distribuição de orientação das fibras, atribuindo uma cor a uma determinada direção.

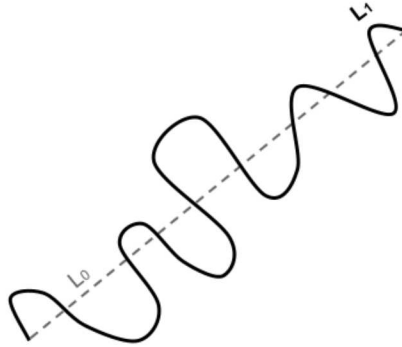
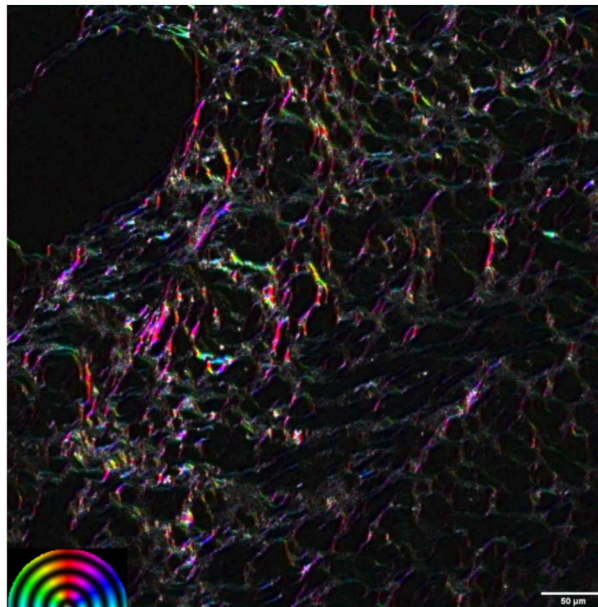


Figura 16 – O comprimento L_0 , mostrado pela linha tracejada, é a distância linear da ponta inicial ao ponto final da fibra, e L_1 , mostrado pela linha contínua, representa o comprimento total da fibra.

A imagem do espectro angular permite visualizar também a alteração e conformação das fibras em uma determinada direção a medida que o câncer aumenta em agressividade. A imagem 17 mostra um exemplo de distribuição angular das fibras de colágeno para uma das amostras estudadas nesse trabalho. A imagem está mostrada por uma superposição da distribuição angular das fibras sobre a imagem gerada pelo processo de SHG. O semicírculo no canto inferior esquerdo representa o mapa de cores para as direções das fibras.



Normal

Figura 17 – Distribuição angular das fibras de colágeno presentes em uma amostra analisada pelo software PyFibre. O semicírculo no canto inferior esquerdo indica o mapa de cores, a direção a qual cada cor corresponde. Imagem obtida por superposição da distribuição angular das fibras sobre imagem de SHG em escala de cinza

3.5 Análise de Variância e Valor P:

A Análise de Variância (ANOVA) é um método estatístico muito utilizado para comparação entre grupos independentes. A ANOVA nos permite verificar se as diferenças observadas nas médias entre esses grupos são estatisticamente significativas ou se podem ter ocorrido por um simples acaso.

A análise de variância tem como base a decomposição da variabilidade total dos dados em duas componentes principais, que são: a variabilidade encontrada entre cada um dos grupos e a variabilidade observada entre os indivíduos dentro de cada grupo. A ideia é que, se essa variabilidade entre os grupos tiver significância maior do que a variabilidade entre cada indivíduo do grupo, então pode-se concluir que essas médias são realmente diferentes.

Para utilizarmos a ANOVA, alguns pontos são necessários, supondo-se um espaço amostral com N grupos, temos os seguintes processos necessários para realizarmos a análise de variância:

1 - Formulação de hipóteses: A hipótese nula (H_0) afirma que as médias dos grupos são iguais, ou seja, $\mu_1 = \mu_2 = \dots = \mu_N$. A hipótese alternativa (H_1) afirma que pelo menos uma das médias é diferente, $\mu_i \neq \mu_j$ para algum $i \neq j$ quaisquer.

2 - Cálculo Estatístico: As estatísticas que medem a variabilidade entre cada grupo, são calculadas utilizando a ANOVA, da seguinte forma:

- A variabilidade entre os grupos, dada pela soma dos quadrados entre cada um dos grupos:

$$\Sigma_{Entre} = \Sigma_{i=1}^N n_i \cdot (\bar{x}_i) - (\bar{x}_N)^2, \quad (3.7)$$

- A variabilidade de cada um dos indivíduos dentro do grupo, dada pela soma dos quadrados dentro do grupo.

$$\Sigma_{Dentro} = \Sigma_{Entre} - \Sigma_T, \quad (3.8)$$

- Variabilidade Total, que é dada por:

$$\Sigma_T = \Sigma_{j=1}^k \Sigma_{i=1}^N (x_i^j - \bar{x})^2 \quad (3.9)$$

3 - Determinação das estatísticas de teste: A partir dos cálculos mostrados no passo 2, pode-se calcular a razão F, também chamada de função F de Fisher:

$$F = \frac{S_{Entre}^2}{S_{Dentro}^2}, \quad (3.10)$$

onde,

$$\begin{aligned} S_{Entre}^2 &= \frac{\Sigma_{Entre}}{k-1} \\ S_{Dentro}^2 &= \frac{\Sigma_{Dentro}}{\Sigma_{i=1}^n (x_i - k)}, \end{aligned} \quad (3.11)$$

sendo k o número de graus de liberdade da amostra. O valor-p para esse caso pode ser calculado como $\text{valor-p} = P(F \geq F_{\text{observado}})$. Para o caso apresentado nessa tese, se o valor-p for menor que 0.05 rejeita-se a hipótese nula. A análise de variância foi utilizada na construção das tabelas de valores p, discutidas no capítulo 4, e inicialmente fazemos uma análise métrica(i) x métrica(j). Para uma análise mais completa optamos pela análise por discriminante linear, que será discutida na próxima seção.

3.6 Análise por Discriminante Linear

Após a extração dos parâmetros, obtemos um conjunto de 29 métricas que serão utilizadas para a classificação automática das amostras. Para isso utilizamos a metodologia de Análise por Discriminante Linear, ou LDA, da sigla em inglês, Linear Discriminant Analysis. LDA é comumente utilizada como uma técnica de estatística supervisionada em aprendizado por máquinas para classificação e separação de dados [87]. A essência do método é utilizar os padrões que separam cada conjunto de dados, e atribuir a eles um rótulo específico para aumentar a separação entre cada conjunto.

Podemos pensar na LDA em termos de projeção de um conjunto de dados que está num espaço n -dimensional em um subespaço dimensional k , onde $k \leq n - 1$, para separar os grupos de dados, sem com isso perder informações que classificam cada conjunto [88]. A figura 18, mostra essa ideia de uma projeção de uma forma simplificada para um conjunto de dados quaisquer em duas dimensões. Note que, no eixo unidimensional obtido, os conjuntos de dados (vermelhos e azuis) estão completamente separados.

A LDA tende a diminuir a interferência de correlações muito fortes da análise computacional, porque ao projetar os conjuntos de dados em um espaço de dimensão menor, essas correlações tendem a desaparecer. A análise por discriminante linear é um dos muitos processos de estatística multivariada que são utilizados em casos nos quais uma grande quantidade de variáveis são medidas de forma simultânea em cada elemento de um espaço amostral.

A LDA tenta relacionar uma variável dependente como uma combinação linear das demais características presentes em um grupo. Para o caso em que temos $p > 1$ preditores, podemos então calcular a probabilidade de um objeto pertencer ao grupo, utilizando o Teorema de Bayes e supondo que seja uma distribuição normal ou gaussiana

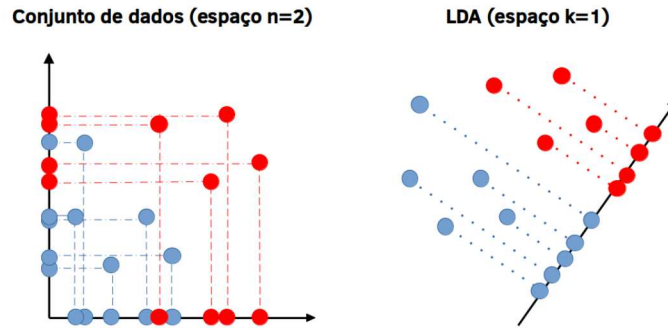


Figura 18 – Esquema simplificado de como é feita a separação de dados por LDA, o objetivo é encontrar o eixo cuja projeção dos dados possui maior separação para cada conjunto. Na figura um conjunto de dados qualquer (símbolo cheios azuis e vermelhas) em um espaço $n=2$, tem as projeções dos dados nesse espaço mescladas, encontramos então um espaço de dimensão reduzida, espaço $k=1$, que melhor separe esse conjunto, sem a perda das informações desses dados.

e considerando que uma variável qualquer tem uma distribuição gaussiana multivariada definida por $X \approx N(\mu, \Sigma)$, temos que [89]:

$$\mathbb{P}(Y = r|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^k \pi_l f_l(x)}, \quad (3.12)$$

sendo $f_k(x)$ a função densidade de probabilidade da distribuição normal multivariada, é definida por:

$$f_X = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right), \quad (3.13)$$

Em que $\mu \in \mathbb{R}^k$ é um vetor de médias e $\Sigma \geq 0$ é a matriz de covariância $p \times p$, que é comum a todas as classes, são dois parâmetros da distribuição normal. Para a k -ésima classe da equação 3.13, temos $f_k(X = x)$, levando esse resultado a equação 3.12, queremos o caso em que

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k, \quad (3.14)$$

é o maior possível. Para a utilização de LDA, em aprendizado de máquinas é necessário que o conjunto de dados tenham as classes bem definidas.

Nesse trabalho a LDA foi usada para a separação das diferentes classificações propostas para os tumores de próstata: padrão Gleason, padrão de ácinos e o padrão relacionado ao estroma reativo. Os parâmetros computacionais da função que calcula a LDA nesse trabalho são explicados no Anexo A, que contempla o script de análise utilizada nesta tese.

3.6.1 Performance do Modelo

Ao utilizarmos ferramentas de aprendizado de máquinas é imprescindível analisarmos a performance do modelo proposto. Nesse sentido algumas métricas relacionadas ao cálculo dessa performance aparecem com muita frequência, entre elas temos a Matriz de confusão, a acurácia do modelo, precisão, sensibilidade, eficiência e eficácia.

A matriz de confusão resume a performance do modelo em classificar as amostras separadas para teste em suas classes verdadeiras. A matriz de confusão possui quatro medidas importantes para calcularmos o rendimento do modelo: verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN) e falso negativo (FN). VP é o número de amostras classificadas como positivo e que realmente são positivos, FP é o número de amostras que foram preditas como positivo, mas que na verdade são negativo, VN é o número de amostras que são verdadeiro negativo e que o modelo predisse de forma correta e por fim, FN é o número de amostras que foram preditas como negativo pelo modelo e que realmente são positivo.

Valores Preditos	Positivo	VP	FN
	Negativo	FP	VN
		Positivo	Negativo
		Valores Reais	

Figura 19 – Exemplo de uma matriz de confusão 2x2, na parte superior temos os valores reais positivos e negativos e na lateral os valores que foram preditos pelo modelo. Valores positivos reais que são preditos de forma correta VP, os valores negativos reais que foram preditos como positivos FN, os valores positivos reais que foram preditos como negativos FP e por fim, os valores negativos reais que foram medidos corretamente como negativos VN.

A partir dos resultados obtidos pela matriz de confusão, podemos calcular outras métricas que servem para avaliação do modelo, incluindo a acurácia, precisão, sensibilidade e especificidade.

A acurácia mede o quanto o modelo realmente é bom em classificar as amostras de forma correta, é um cálculo do acerto em relação ao total de amostras. O cálculo da

acurácia é dado pela seguinte fórmula:

$$Acurácia = \frac{(VP + VN)}{(VP + FP + VN + FN)} \quad (3.15)$$

A precisão de um modelo pode ser entendida como uma relação entre todas as amostras que foram classificadas como positivas pelo modelo, quais delas são realmente positivas:

$$Precisão = \frac{VP}{(VP + FP)} \quad (3.16)$$

Já a sensibilidade é uma medida da proporção de amostras positivas classificadas pelo modelo, em relação a todos os valores positivos presente no espaço amostral. Podemos defini-la conforme a equação abaixo:

$$Sensibilidade = \frac{VP}{(VP + FN)} \quad (3.17)$$

E por fim, a especificidade que mede a taxa de valores negativos que foram medidos de forma correta, ela pode ser definida como:

$$Especificidade = \frac{VN}{(VN + FP)} \quad (3.18)$$

A eficácia do modelo é em sumo, a capacidade do mesmo de realizar uma tarefa, nesse caso a classificação dos grupos, com grande precisão e acurácia. Para que um modelo seja realmente eficaz é necessário que ele seja capaz de produzir resultados precisos e confiáveis, mesmo quando levamos em consideração dados com alto ruído ou desbalanceados (amostragens muito diferentes para cada tipo a ser classificado).

4 Resultados e Discussões

Neste capítulo apresento os resultados obtidos através das imagens multifótons, as análises de imagens através do software PyFibre e classificações dos tipos tumorais estudados.

4.1 Imagens PL e SHG

Apresento aqui os resultados para as biópsias de câncer de próstata selecionadas pelo patologista, que também selecionou em cada lâmina as regiões de interesse com diferentes prognósticos como descrito na seção 3.1. Para essas regiões foi feito o imageamento por transmissão, geração de segundo harmônico e fluorescência por dois fótons. Um exemplo é mostrado na figura 20, onde é possível ver alguns dos componentes presentes nas biópsias de tecido de próstata. Na imagem transmitida podemos identificar os componentes do parênquima prostático como, o lúmen que são as cavidades tubulares presentes no tecido, que na imagem correspondem aos espaços “vazios”. Também é possível observar as células que são as estruturas granulares que podem ser vistas na imagem transmitida. O conjunto de células fechadas em uma estrutura que se assemelham a bagos são os ácinos. E é possível também observar os componentes do estroma, que são as fibras de colágeno que são vistas na imagem de SHG. E ao sobrepor as imagens de SHG e PL, a diferença entre as duas nos permite ressaltar as fibras musculares, como podemos observar olhando para a imagem composta de PL e SHG.

Para as análises, as imagens foram separadas segundo os critérios de classificação em 3 grupos distintos para posterior obtenção dos parâmetros via software. Esses agrupamentos foram realizados em conformidade com o patologista e fazem parte da proposta de se analisar os tecidos carcinogênicos olhando também para as mudanças que ocorrem no estroma e nos ácinos. O primeiro dos grupos é o já padronizado e amplamente utilizado sistema Gleason, o segundo grupo é uma separação baseada no tipo de ácino presente no tecido, Padrão de Ácino e o terceiro grupo é a separação baseada no tipo de reação estromal, Padrão de Estroma. É importante ressaltar que os 3 grupos contém todas as imagens analisadas nessa tese.

As imagens obtidas pelo imageamento no microscópio e previamente separadas nos devidos grupos foram analisadas utilizando-se o nosso software PyFibre. O PyFibre extrai de cada imagem as 29 métricas informadas na seção 3.4 do capítulo 3. Esse conjunto de parâmetros formam a base de dados para as análises. Escrevemos um script de análise em linguagem Python (detalhes no apêndice A, para realizar as análises estatísticas e gráficos LDA, que serão apresentados e discutidos no decorrer desse capítulo. O objetivo é

verificar, através das técnicas de imageamento e análise das regiões tumorais marcadas pelo patologista, se existem outras alterações presentes nos tecidos que foram analisados que auxiliem no prognóstico do paciente. Essas alterações podem ser, por exemplo, nos ácinos presentes nos tecidos ou nas redes de fibras de colágeno ou no padrão de estroma. Foram feitas algumas tentativas de classificação e separação das amostras de acordo com as informações sobre essas alterações que nos foram fornecidas pelo patologista, essas separações serão discutidas a seguir.

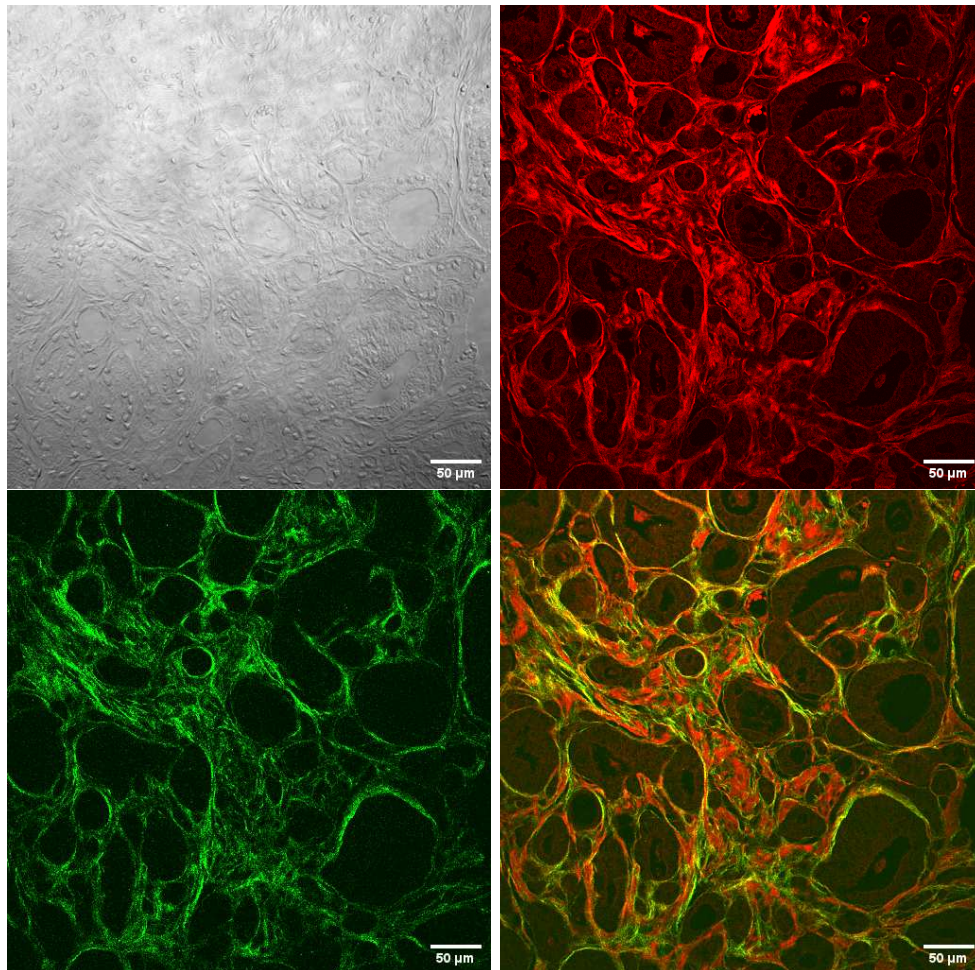


Figura 20 – Imagens de tecido de próstata: Na parte superior à esquerda temos a imagem por Transmissão, e à direita temos a imagem de PL, abaixo à esquerda temos a imagem de SHG e à direita a imagens de PL e SHG superpostas

Na figura 21 apresento exemplos de todas as imagens obtidas através do PyFibre, através das imagens multifótons obtidas de uma amostra de biópsia, para uma área medida de um tumor Gleason 3 + 3. Em (a) está apresentada a imagem de campo claro da região da biópsia corada com HE mostrando o componente acinar organizado aleatoriamente entre o estroma. As imagens do mapa de intensidade de PL, em vermelho, e SHG, em verde, são apresentadas em (b) e (c), respectivamente. As imagens destacam os detalhes dos componentes do tumor acinar e estromal, a imagem SHG mostra as fibras de colágeno e o PL mostra a emissão do corante eosina usado para corar o tecido.

A análise automatizada de imagens permite segmentar as imagens em conteúdo celular (d) e de fibras colágenas (e). Cada segmento conectado é mostrado em uma cor aleatória. A rede de fibras de colágeno extraída, em vermelho, e as fibras individuais extraídas, em cores aleatórias, são mostradas em (f) e (g), respectivamente. Um mapa de cores dos ângulos de alinhamento das fibras de colágeno é mostrado em (h). De acordo com o sistema de classificação de estroma reativo proposto, a região da biópsia mostrada em (a) foi classificada como estroma 2, mostrando mais de 5% de fibras colágenas (fios rosa claro) ao fundo.

Através destas análises para todas as áreas estudadas obtivemos um conjunto de dados, com as 29 métricas extraídas para cada imagem. Esse conjunto de dados será utilizado para as análises dos tipos tumorais, como descrito nas próximas seções.

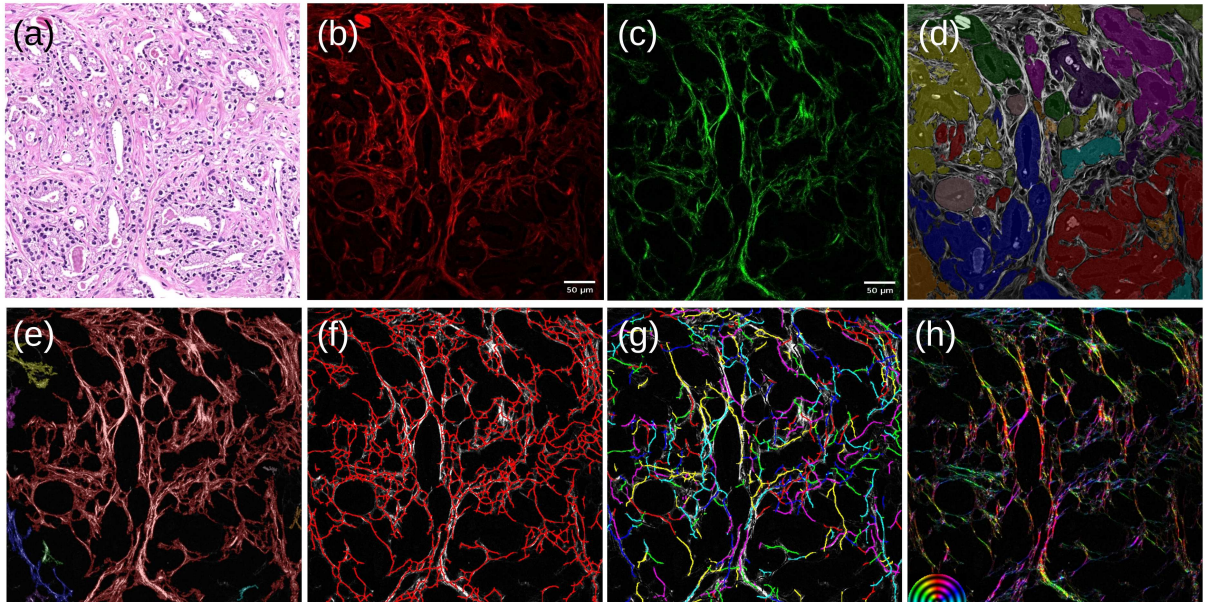


Figura 21 – Imagens obtidas: (a) imagem de microscopia de campo claro HE, (b) PL e (c) SHG. Imagens extraídas pelo software: (d) segmentos de células e (e) segmentos de fibra, cada segmento conectado é mostrado em uma cor aleatória, e os segmentos de célula e fibra estão superpostos nos mapas de intensidade PL e SHG em escala de cinza, respectivamente. (f) a rede de fibras de colágeno em vermelho, e (g) as fibras individuais extraídas (cada fibra tem uma cor aleatória), também sobrepostas no mapa de intensidade SHG em escala de cinza. (h) os ângulos de orientação das fibras de colágeno (escala do mapa de cores à esquerda). A barra de escala de 50 μm se aplica a todas as imagens extraídas e à imagem HE.

4.2 Análise dos resultados para o padrão Gleason

Por ser o padrão ouro nas classificações tumorais, começarei a apresentação dos resultados pela separação por Gleason, esse padrão serviu como referência para as tentativas de classificação pelo estroma e ácinos. Os valores das somas do padrão Gleason, da menor para a maior, servem como uma quantificação do grau de agressividade do câncer. Vale notar que essa soma varre os valores de 6 até 10, mas como os tipos mais agressivos são muito similares, os padrões de Gleason $4 + 4$, $4 + 5$, $5 + 4$ e $5 + 5$, por terem poucos casos documentados nessa tese, de forma que foram combinados e nas análises de Gleason, serão identificados como $4 + 4+$, essa junção é importante para evitarmos os efeitos de desbalanceamento de dados, o que pode trazer análises enviesadas.

Inicialmente, apresentarei algumas comparações estatísticas entre os tipos tumorais utilizando o valor- p obtido por ANOVA (do inglês: One-way analysis of variance). Serão discutidas somente as métricas que proporcionam um p significativo, dentre as 29 que são obtidas. Os gráficos de caixa (do inglês boxplot) para todas as métricas estão apresentados no apêndice B. Nas tabelas, utilizarei a nomenclatura $3 + 3$, $3 + 4$, etc, para significar os Gleason $3 + 3$, Gleason $3 + 4$, respectivamente, e a significância do valor p : $p < 0.05$ é representada por *, $p < 0.01$ é representado por ** e $p < 0.001$ é representado por ***.

A Tabela 3 mostra os valores p para o caso em que temos os tipos normais comparados com cada um dos padrões Gleason. É possível ver que o padrão Gleason $4 + 4+$ tem valores significativos para todas as métricas mostradas em relação ao padrão Normal, esse é o tipo tumoral que mais modifica o ambiente ao redor de si e implica também em uma maior agressividade do tumor presente no tecido, mas essa distinção entre um padrão histológico normal e tipos tumorais se faz presente em todos os casos mostrados.

Tabela 3 – Valores p para as comparações entre normal e as várias somas de Gleasons.

Métricas	Normal $\times 3 + 3$	Normal $\times 3 + 4$	Normal $\times 4 + 3$	Normal $\times 4 + 4+$
Circularidade Seg. de Parênquima		***	***	***
Angulo PL no Seg. Parênquima	*			**
Área Seg. de Parênquima	***	***		***
Cobertura Seg. de Parênquima	***	***		***
Desvio Padrão do SHG	***	***		***
Anis. do SHG Global	*	***	*	***
Anis. do SHG Local	***	***	***	***
No. de Seg. de Parênquima	***	***	**	***
No. de Fibras		***	*	*

Em seguida apresento os valores p para as métricas mais relevantes para as comparações de cada tipo de Gleason entre si, tabelas 4, 5 e 6. É interessante notar que as métricas com valor significativo se repetem em todas as tabelas. De forma geral podemos relacioná-las a dois grupos principais: métricas relacionadas ao parênquima celular e métricas que contemplam as fibras de colágeno. Estes resultados evidenciam a importância de se estudar as mudanças estruturais que ocorrem nas fibras nas regiões afetadas por algum tipo de tumor.

Vale destacar as comparações entre as somas de Gleason $3+4$ e $4+3$, que estruturalmente são muito parecidas, sendo somas que muitas vezes podem causar confusão mesmo quando analisadas por patologistas experientes. Ao observarmos os valores p mostrados na tabela 5, fica evidente que existem características presente nesses dois tipos que podem ser difíceis de se contemplar visualmente.

Apesar das tabelas de valores p mostradas na tabelas indicarem que existem diferenças entre os diversos padrões gleason para algumas métricas, analisar somente esses parâmetros e excluir da análise os demais não conseguiríamos de forma eficiente separar as diferentes categorias. Assim optamos por realizar a análise com todos os parâmetros por LDA, o que permite verificar melhor o quanto cada classe tumoral difere entre si.

Tabela 4 – Valores p para as comparações entre Gleason $3+3$ e as outras somas de Gleasons.

Métricas	$3+3$ $\times 3+4$	$3+3$ $\times 4+3$	$3+3$ $\times 4+4+$
Circularidade do Seg. de Parênquima	***		***
Angulo da PL no Seg. Parênquima	***	***	***
Área do Seg. de Parênquima		*	***
Cobertura Segmento de Parênquima	***		***
Desvio Padrão do SHG		***	
Anis. do SHG Global			
Anis. do SHG Local		**	*
No. de Segmento de Parênquima			***
No. de Fibras	***	**	**

Tabela 5 – Valores p para as comparações entre Gleason $3 + 4$ e as outras somas de Gleasons.

Métricas	$3 + 4$ $\times 4 + 3$	$3 + 4$ $\times 4 + 4 +$
Circularidade do Seg. de Parênquima	*	***
Angulo da PL no Seg. Parênquima	**	***
Área do Seg. de Parênquima	***	***
Cobertura Segmento de Parênquima	***	***
Desvio Padrão do SHG	***	**
Anis. do SHG Global	**	
Anis. do SHG Local	**	***
No. de Segmento de Parênquima	***	***
No. de Fibras		***

Tabela 6 – Valores p para as comparações entre Gleason $4 + 3$ e o Gleason $4 + 4 +$.

Métricas	$4 + 3$ $\times 4 + 4 +$
Circularidade do Seg. de Parênquima	***
Angulo da PL no Seg. Parênquima	*
Área do Seg. de Parênquima	***
Cobertura Segmento de Parênquima	***
Desvio Padrão do SHG	***
Anis. do SHG Global	***
Anis. do SHG Local	***
No. de Segmento de Parênquima	***
No. de Fibras	***

4.3 Análise dos resultados para o padrão de Ácinos

Como descrito anteriormente no Cap 3, essa classificação quanto ao tipo de padrão acinar presente foi dividida em 5 tipos, denominados A, B, C, D, E. Porém o padrão E não foi computado nas análises descritas nesse capítulo por se tratar de um padrão misto. Relembrando os 4 padrões estudados são definidos da seguinte forma:

Padrão A: Possui Ácinos bem formados

Padrão B: Apresenta Ácinos Rudimentares

Padrão C: Presença de Padrão cribriforme típico puro

Padrão D: Composto por um Padrão Sólido e células isoladas.

Vale ressaltar que alguns tipos de padrões de ácinos e também de Gleason, são mais comuns e abundantes. Os padrões A, B e D são os mais abundantes e por isso possuem um número maior de amostras. As tabelas a seguir mostram os valores p para cada tipo de padrão de ácino. De forma análoga ao que foi feito para o padrão Gleason, aqui primeiramente mostro a tabela 7, que relaciona o padrão normal com os demais tipos de ácinos. Novamente temos uma separação visível entre a classe Normal e as classes neoplásicas.

As tabelas 8, 9 e 10 mostram os valores p obtidos para as comparações entre os tipos de ácinos. Observe que apenas o padrão do tipo A, aparenta ser diferenciável de todos os demais. Já o padrão do tipo B apresenta uma boa separação do grupo C, mas temos uma dificuldade de percebermos uma separação entre os tipos B e D. O padrão C que consiste de ácinos cribriformes se difere bem dos demais, pois o que ocorre nesse caso é que as células carcinogênicas crescem dentro do lúmen. Vale ressaltar aqui que o PyFibre não categoriza cada componente celular individualmente, mas engloba dentro das métricas relacionadas ao parênquima, as estruturas de lúmen, células, ácinos, dentre outras.

Tabela 7 – Valores p para as comparações entre normal e os vários padrões de ácinos.

Métricas	Normal × Padrão A	Normal × Padrão B	Normal × Padrão C	Normal × Padrão D
Angulo da PL no Seg. Parênquima	***	*		**
Área do Seg. de Parênquima	***	***	*	
Cobertura Segmento de Parênquimas	***	***		***
Desvio Padrão do SHG	***	**	**	
Anis. do SHG Global	**	**	**	***
Anis. do SHG Local	*	***		***
No. de Segmento de Parênquima	***	***	***	**
No. de Fibras		***		***

Tabela 8 – Valores p para as comparações entre o padrão de ácino A e os outros padrões de ácinos.

Métricas	Padrão A × Padrão B	Padrão A × Padrão C	Padrão A × Padrão D
Angulo da PL no Seg. Parênquima	***	*	***
Área do Seg. de Parênquima		***	***
Cobertura Segmento de Parênquima	***	***	
Desvio Padrão do SHG	***		**
Anis. do SHG Global	*		
Anis. do SHG Local	***	***	***
No. de Segmento de Parênquima	***	***	
No. de Fibras	***		***

Tabela 9 – Valores p para as comparações entre o padrão de ácino B e os outros padrões de ácinos.

Métricas	Padrão B × Padrão C	Padrão B × Padrão D
Angulo da PL no Seg. Parênquimas		
Área do Seg. de Parênquima	***	***
Cobertura Segmento de Parênquima	***	
Desvio Padrão do SHG	*	
Anis. Global do SHG	*	
Anis. do SHG Local		
No. de Segmento de Parênquima	***	
No. de Fibras	***	

Tabela 10 – Valores p para as comparações entre o padrão de ácino C e os outros padrões de ácinos.

Métricas	Padrão C × Padrão D
Angulo da PL no Seg. Parênquima	*
Área do Seg. de Parênquima	
Cobertura Segmento de Parênquima	**
Desvio Padrão do SHG	
Anis. Global do SHG	
Anis. do SHG Local.	***
No. de Segmento de Parênquimas	***
No. de Fibras	***

Dentre as métricas extraídas pelo PyFibre, duas delas que se relacionam com parênquima, tanto na área total quanto no número de segmentos, chamam bastante a atenção. A figura 22 apresenta o gráfico em caixa (boxplot) para esses parâmetros, a área de parênquima e a quantidade dos segmentos de parênquima, respectivamente. Podemos ver que enquanto para os padrões Normal, A, B e D, temos uma diminuição da área de segmentos e um aumento do número desses segmentos, para o padrão C, temos a relação inversa. Isso significa que enquanto temos um aumento significativo da área do segmento devido ao crescimento anormal de células dentro do lúmen, o número desses segmentos diminui pelo mesmo motivo.

Já para os demais padrões como as células se multiplicam fora do lúmen, a área total coberta pelos segmentos diminui, porque as células estão mais espalhadas dentro do tecido, mas o número de segmentos aumenta de forma significativa. É importante salientar que a métrica não identifica as células especificamente, o que ela identifica é essa região de parênquima, que engloba as células, o lúmen, e até mesmo espaços vazios dentro do tecido. A figura 23 mostra a análise feita pelo software PyFibre para a métrica do segmento de parênquima para os padrões do tipo A e tipo C, respectivamente. Podemos observar que no casos em que temos o padrão do tipo C, temos um menor número de áreas de parênquima mas, um área maior observada do que no padrão A.

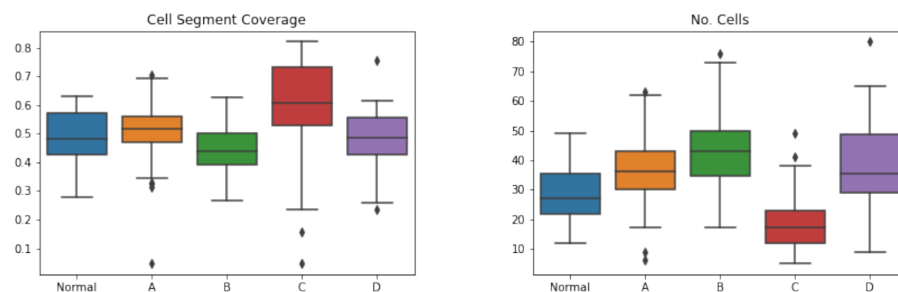


Figura 22 – Gráfico em Boxplot de duas das métricas analisadas, área coberta pelo segmento de parênquima (Cell Segment Coverage) número de segmentos de células (No. Cells), ambos para a separação pelos tipos de ácinos. No gráfico, as linhas centrais mostram as medianas; os limites da caixa indicam os percentuais de 25 e 75%; as barras indicam os limites 0 e 100% dos dados; e os pontos representam os dados discrepante (outliers).

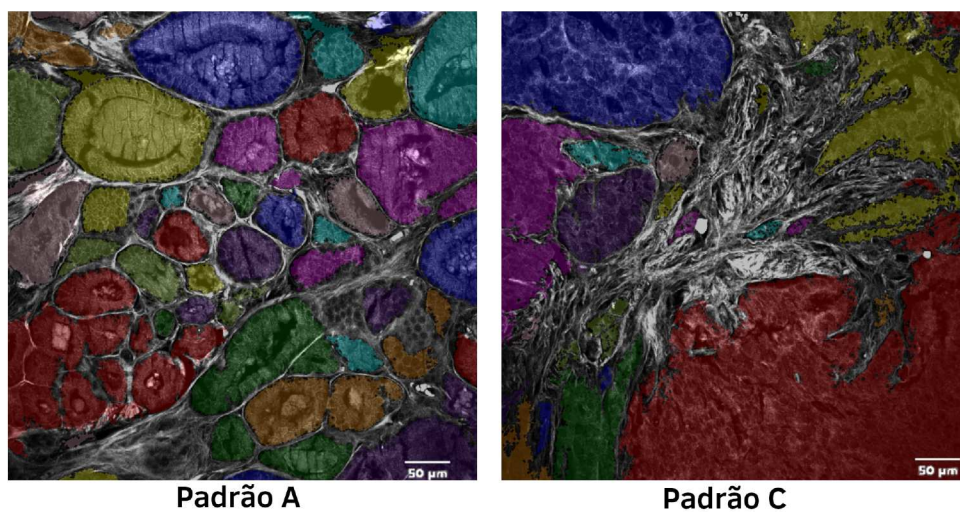


Figura 23 – Imagem obtida pelo software PyFibre, mostrando os segmentos de parênquima, para os padrões A (esquerda) e C (direita).

4.4 Análise dos resultados para o padrão de Estroma

O Estroma Reativo (ER) foi caracterizado pelo patologista, em termos da reação tumoral presente em cada uma das regiões. Tem-se dessa forma 3 grupos distintos separados conforme mostrado abaixo:

Estroma 1 - Possui reação tumoral escassa ou ausente.

Estroma 2 - Possui reação tumoral moderada.

Estroma 3 - Possui reação tumoral acentuada.

A tabela 11 mostra os valores p para essas três classificações do estroma. O que podemos perceber, primeiramente, é que realmente existe uma maior diferenciação das amostras, dependendo do tipo de reação estromal que está presente na biópsia. Também vemos que esta diferença pode ser quantificada, principalmente quando a reação tumoral é acentuada (Estroma 3). Outro fato relevante que podemos observar é que a análise por SHG permite uma boa separação entre os tipos de estromas reativos, pois a anisotropia das fibras de colágeno realmente é um fator disruptivo entre os três tipos de estroma.

Tabela 11 – Valores p para as comparações entre os vários padrões de estroma.

Métricas	Estroma 1 × Estroma 2	Estroma 1 × Estroma 3	Estroma 2 × Estroma 3
Circularidade do Seg. de Parênquima		**	**
Angulo da PL no Seg. Parênquima		***	***
Área do Seg. Parênquima		***	***
Entropia da PL no Seg. Parênquima		**	***
Conectividade da Rede de Fibras		***	***
Densidade da Rede de Fibras		***	***
Comprimento Médio das Fibras		***	***
Anis. do SHG Global	***	***	**
Anis. do SHG Local	***	***	***
No. de Segmentos de Parênquimas	*	*	
No. de Fibras	***	***	***

De fato, se olharmos de forma comparada para os gráficos boxplots, mostrados nas figuras 24, 25 e 26 relacionadas à anisotropia, tanto local quanto global, podemos ver que das três classificações: **Gleason**, **Ácinos** e **Estroma**, a classificação por estroma é a que melhor demonstra o aumento da ordem de organização das fibras de colágeno a medida que o câncer evolui e se torna mais agressivo. Isso demonstra a importância de se olhar para as mudanças estruturais que decorrem da presença do tumor no estroma.

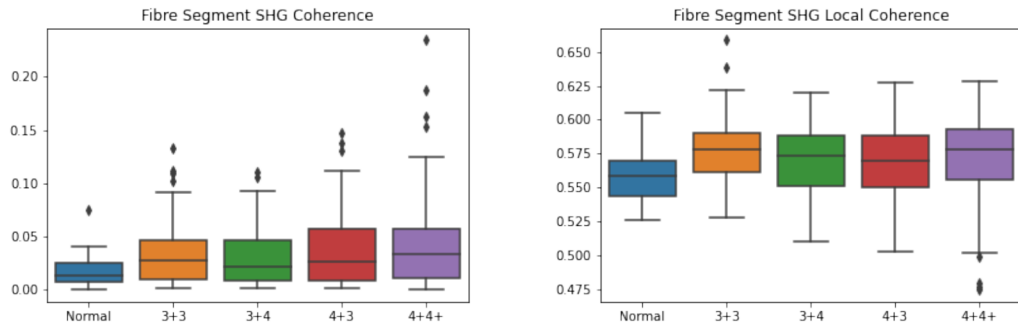


Figura 24 – Gráfico boxplot da anisotropia global (Fibre Segment SHG Coherence) e local (Fibre Segment SHG Local Coherence) para a separação pelo padrão Gleason.

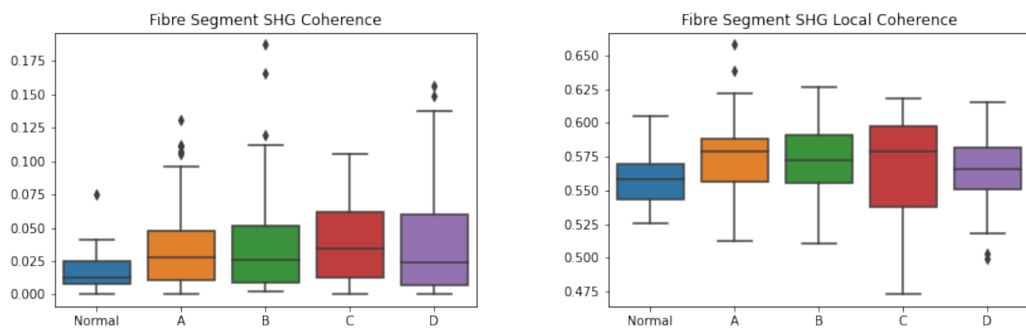


Figura 25 – Gráfico boxplot da anisotropia global (Fibre Segment SHG Coherence) e local (Fibre Segment SHG Local Coherence) para a separação pelo tipo de ácinos.

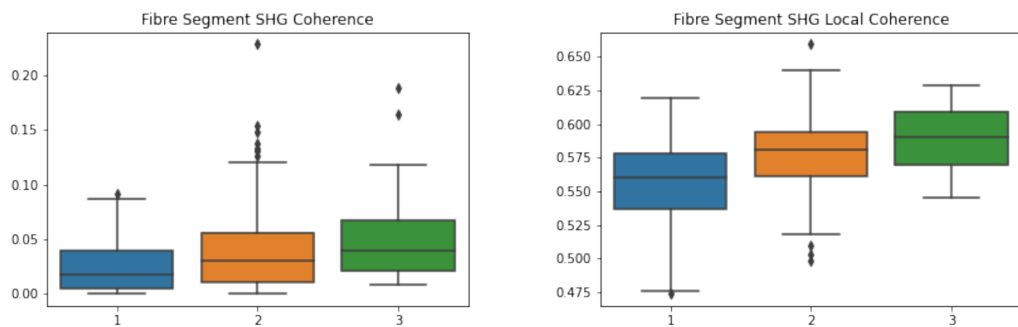


Figura 26 – Gráfico boxplot da anisotropia global (Fibre Segment SHG Coherence) e local (Fibre Segment SHG Local Coherence) para a separação pelo estroma.

4.5 Análises LDA: aprendizado de máquinas

Utilizamos as técnicas de aprendizado de máquinas para categorizar os dados de carcinomas de próstata, por análise de discriminante linear. Quando falamos de aprendizado de máquinas uma ação necessária para a modelagem dos dados é a separação desse conjunto de dados em Treino e Teste. As amostras separadas para Treinamento, contém todos os rótulos necessários para que modelo aprenda sobre os padrões e as relações que os dados possuem que o possibilite categorizar esses dados de acordo com os rótulos estabelecidos.

As amostras de teste, contém um conjunto de dados nos quais os rótulos são desconhecidos e serve para verificar o quanto o modelo aprendeu a partir dos dados de treino, a fazer ele mesmo esses rótulos. Um modelo que consegue separar bem os dados de treino, mas ao receber um conjunto de dados desconhecidos performa de forma ineficiente, provavelmente está fazendo o que é usualmente chamado de *superajuste* ("overfitting", em inglês). É necessário que as amostras de teste sejam uma representação boa dos dados reais e com isso termos um bom modelo.

Para as análises por LDA, o script em python que foi utilizado e que pode ser encontrado no anexo A, necessita de alguns argumentos previamente estabelecidos nas funções definidas dentro do script. Para explicar esses argumentos é necessário dizer que a função que computa a LDA, utiliza duas outras funções diferentes em Python, ambas do pacote de programação Scikit-Learn [90], que também são discutidas brevemente no anexo A.

A primeira dessas funções é a `RepeatedStratifiedKFold`, e os argumentos dessa função que foram configurados são: `n_splits=6`, `n_repeats=8`, `random_state=42`. O parâmetro `n_splits` se refere a quantos grupos de divisões em dados de treino e teste, serão feitos, nesse caso teremos 6 divisões sendo 5 para treinos e 1 para teste, é importante que essas divisões sejam feitas de formas aleatórias e contemplem todos os dados. O argumento `n_repeats` define quantas vezes será feita as divisões determinadas pelo parâmetro anterior, no caso temos 8 repetições dessa divisão em 6 grupos. E por fim, o argumento `random_state`, serve como uma semente para que ao repetirmos essas análises com os mesmos dados e mesmas configurações sejam obtido os mesmo resultados. Nas seções a seguir, apresentamos e discutimos os resultados da análise por LDA para as separações tumorais que foram estudadas neste trabalho.

4.5.1 LDA da Separação por Gleason

A figura 27 (A) mostra o gráfico LDA para o caso do Gleason. Os dados são apresentados nos dois eixos principais, correspondentes aos dois principais auto-valores dos dados, onde se obtém a melhor separação dos grupos. Os símbolos cheios representam os dados de treinamento do script de análise, as cruces os dados de teste, a estrela o valor

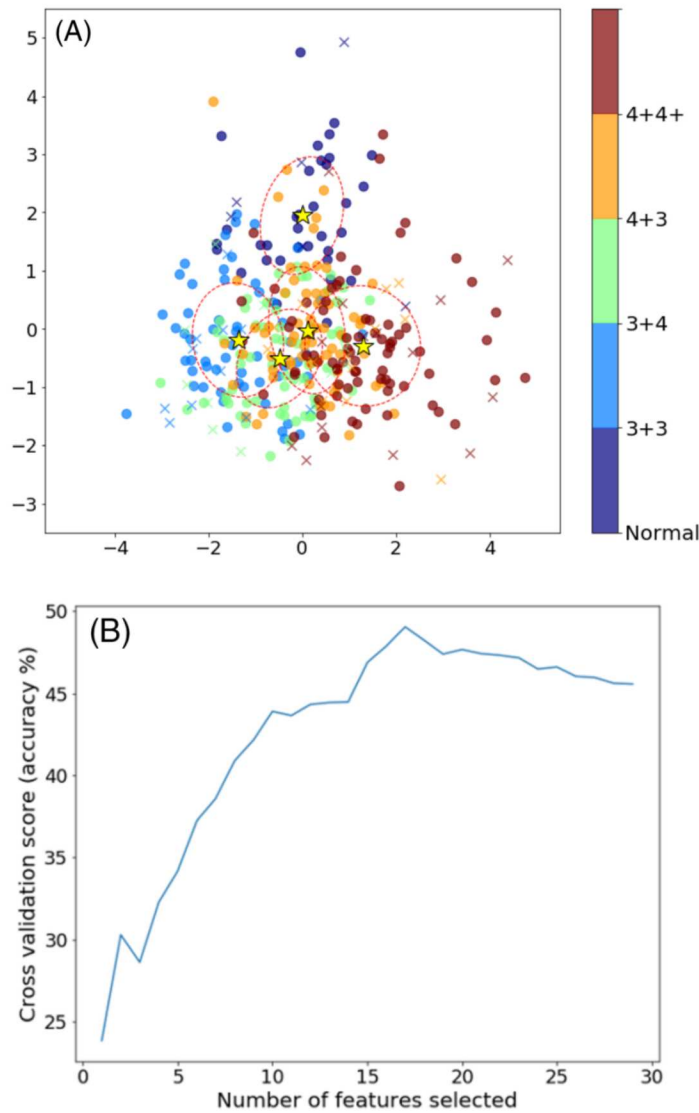


Figura 27 – Gráfico (A) é a LDA da separação pelo tipo de Gleason. Na imagem, o símbolo cheio representa os dados de treinamento do script de análise, a cruz os dados de teste, a estrela o valor médio para cada componente e a elipse envolve um desvio padrão da amostragem e mostra a direção preferencial de cada componente. A acurácia do LDA para a separação pelo tipo de Gleason é de $46.0\% \pm 6\%$, como mostrado no gráfico (B) para a validação cruzada em função do número de parâmetros.

médio para cada grupo e a elipse envolve um desvio padrão da amostragem e mostra a direção preferencial de cada componente. Podemos ver claramente que o padrão normal se separa dos demais padrões (média e desvio padrão bem separados dos outros grupos), em concordância com o já obtido pelas análises de valores p na Tabela 3. A acurácia da separação entre os tecidos normais e com tumores é de $89\% \pm 3\%$. Como todas as análises foram feitas para o mesmo conjunto de imagens, essa acurácia é a mesma para todas as separações. Observamos também uma separação entre os tipos de câncer mais agressivos e os demais. A acurácia do LDA para a separação pelo tipo de Gleason, o gráfico (B), mostra

a validação cruzada em função do número de parâmetros considerado na análise. Com aproximadamente 15 dos parâmetros obtemos o valor máximo. Como a curva continua aproximadamente constante, mantivemos todos os 29 parâmetros em todas as análises. A acurácia do LDA para a separação pelo tipo de escore de Gleason para todos os grupos é de $46.0\% \pm 6\%$

Para as somas de Gleasons do tipo $3 + 4$ e $4 + 3$, o que ocorre é que entre esses dois padrões as diferenças no tecido são difíceis de serem percebidas, principalmente quando atentamos para o fato de que o padrão Gleason desconsidera alterações no estroma. Verificamos, como será discutido posteriormente, que o estroma apresenta um importante papel na diferenciação desses dois tipos de tumor. Esses duas somas de Gleason são os tipos mais complicados de se separar mesmo sendo avaliados por patologistas experientes que podem divergir na classificação desses padrões [76,91].

A figura 28 e a tabela 12 mostram a qualidade dos resultados de classificação obtidos através da análise LDA para o diagnóstico por padrão Gleason. A figura 28 mostra os resultados da matriz de confusão com valores preditos para o conjunto de dados de teste, após a etapa de aprendizado. As linhas mostram as classificações preditas e as colunas as classificação reais (true values). Os valores de precisão, sensibilidade e especificidade calculados através da matriz de confusão são mostrados na tabela 12. É realmente interessante notar as separações para os padrões Normal, $3 + 3$ e $4 + 4+$. Mas como esperado para as somas $3 + 4$ e $4 + 3$ o modelo não os separa de forma eficiente, justamente por esses dois padrões serem muito semelhantes. Vale ressaltar que mesmo patologistas bem treinados e experientes tem um grau de acordo quanto a classificação desses tipos em torno de 21%, conforme mostrado em estudos recentes [76,91]. Logo temos que os resultados obtidos por LDA são compatíveis com o que foi encontrado entre os patologistas.

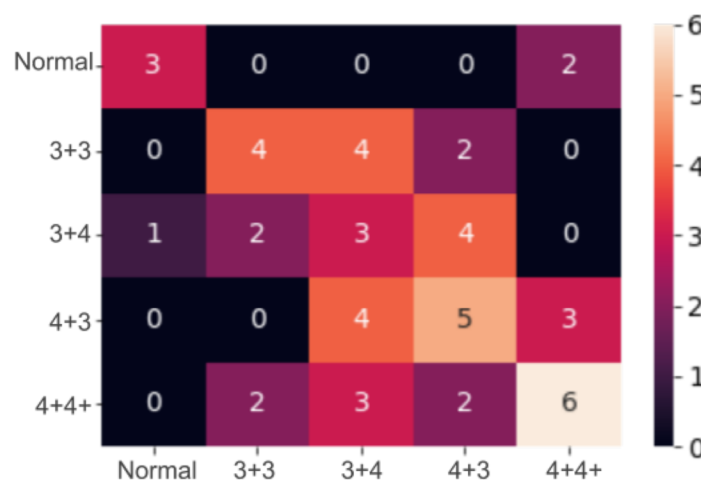


Figura 28 – Matriz de Confusão para a Separação por Gleason, na diagonal principal temos os verdadeiros positivos para essa separação.

Tabela 12 – Análises Estatísticas: Classificação pelo Padrão Gleason.

	Normal	3 + 3	3 + 4	4 + 3	4 + 4+
Precisão	0.75	0.50	0.21	0.38	0.54
Sensibilidade	0.60	0.40	0.30	0.42	0.46
Especificidade	0.95	0.81	0.62	0.67	0.75

4.5.2 LDA da Separação por Ácinos

A figura 29 (A) mostra o gráfico de LDA para a classificação de padrão de ácinos. Os símbolos cheios representam os dados de treinamento do script de análise, as cruzes os dados de teste, a estrela o valor médio para cada grupo e a elipse envolve um desvio padrão da amostragem e mostra a direção preferencial de cada componente. Ao olharmos para a figura fica evidente a separação das amostras não neoplásicas e o padrão do tipo A (média e desvio padrão bem separados dos outros grupos). Já para os tipos B e D, observa-se uma dificuldade de separação, provavelmente devido ao fato de termos estruturas semelhantes nas duas classes. Entretanto, para o padrão C por representar apenas padrões cribriformes puros, tende a se separar dos demais tipos. A acurácia do LDA para a separação pelo tipo de ácino, o gráfico (B), mostra a validação cruzada em função do número de parâmetros considerado na análise. Com proximadamente 12 dos parâmetros obtemos o valor máximo, mas como a curva continua aproximadamente constante, mantivemos todos os 29 parâmetros em todas as análises. A acurácia do LDA para a separação entre as amostras pelo tipo de ácino é de $55\% \pm 4\%$.

A matriz de confusão para a separação por tipos de ácinos é mostrada na figura 30, e os cálculos de sensibilidade, precisão e especificidade para a separação por ácinos é mostrada na tabela 13. Como podemos ver para esse caso o modelo tem um comportamento estranho ao não conseguir identificar nenhuma das amostras como Normal, todas as amostras classificadas como normais foram categorizadas como A, B e D. Isso levaria a resultados inconclusivos nos cálculos da tabela 13, pois no caso da sensibilidade do padrão normal teríamos uma divisão do tipo $0/0$, o valor então foi considerado como zero. Os demais valores fazem sentido quando comparados aos que foram mostrados na tabela 12. Como observado no gráfico de validação cruzada 29 (B), temos uma separação que assim como a feita pelo tipo de Gleason, onde tem-se alguns problemas para identificar e corretamente classificar os tipos tumorais.

Pode-se supor que apesar dos ácinos terem diferenças estruturais claras, como as métricas computadas pelo PyFibre leva em consideração todas as estruturas parênquimais isso interfira na classificação do modelo, mas também não podemos descartar que em ambos os casos, Gleason e ácinos, desconsideram o estroma reativo. Sabe-se hoje em dia que essas alterações no estroma, causadas pela presença de um tumor se tornam mais significativas com a evolução do câncer, como veremos na próxima seção.

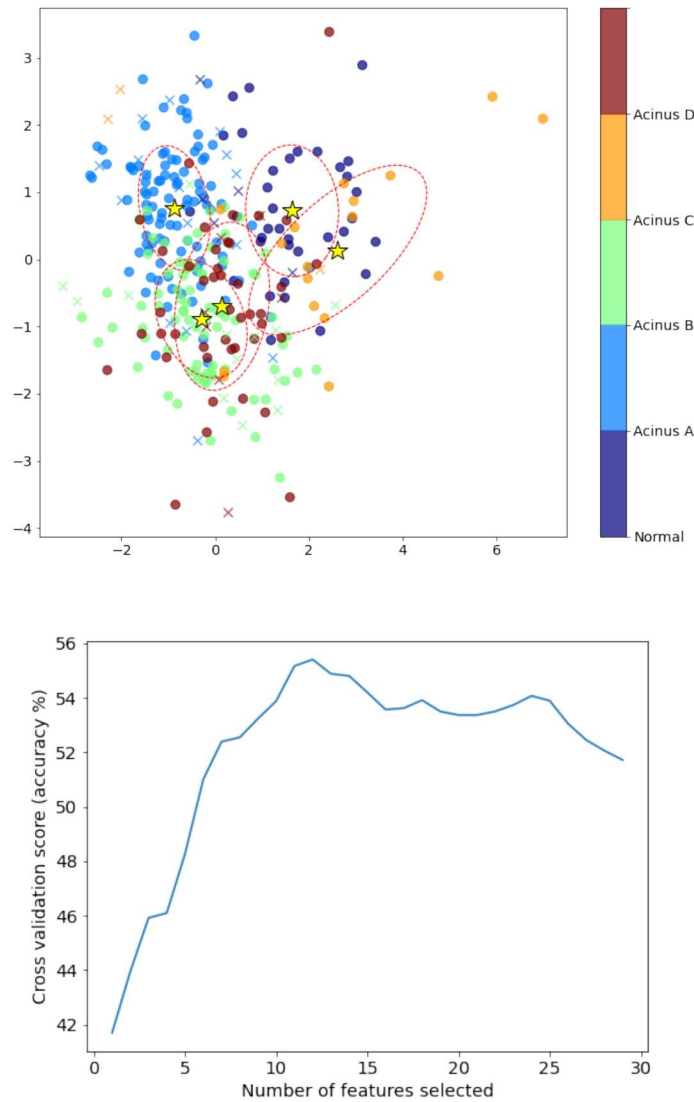


Figura 29 – Gráfico LDA (A) para a separação pelo tipos de Ácinos. Na imagem, o símbolo cheio representa os dados de treinamento do script de análise, a cruz os dados de teste, a estrela representa o valor médio para cada componente e a elipse envolve um desvio padrão da amostragem e mostra a direção preferencial de cada componente. A acurácia do LDA para a separação pelo tipo de ácino é de $55\% \pm 4\%$, como mostrado no gráfico (B) para a validação cruzada em função do número de parâmetros.



Figura 30 – Matriz de Confusão para a Separação por ácidos, na diagonal principal temos os verdadeiros positivos para essa separação.

Tabela 13 – Análises Estatísticas: Classificação pelo Padrão de Ácidos.

	Normal	A	B	C	D
Precisão	0.00	0.50	0.46	0.33	0.43
Sensibilidade	0.00	0.69	0.43	0.33	0.43
Especificidade	1	0.48	0.68	0.91	0.81

4.5.3 LDA da Separação por Estroma

A separação entre os tipos de estroma está apresentada no gráfico de LDA na figura 31 (A). É importante notar que o número de casos para os tumores estromogênicos (estroma 3), cuja massa tumoral é composta principalmente de estroma reativo, é bem menor do que para os demais padrões, correspondendo a apenas 8% dos casos analisados. Isso se deve ao fato dele estar associado aos tumores de maior agressividade, que são mais raros.

A acurácia do LDA para a separação pelo tipo de estroma reativo, excluindo-se as normais, é de $75\% \pm 5\%$, um valor consideravelmente maior do que os obtidos pelas duas classificações anteriores. Isso deixa bem evidente que as alterações que estão presentes quando o tumor altera o ambiente se tornam muito significativas, essa reação do estroma no sentido de manter a homeostase do sistema, pode ser um bom indicativo da gravidade do tumor presente.

Para a separação pelo tipo de estroma também apresento a figura 32 e a tabela 14 que mostram os resultados que são obtidos através da análise por LDA. A figura 32 mostra os resultados da matriz de confusão com valores preditos para o conjunto de dados de teste. Os valores de precisão, sensibilidade e especificidade são mostrados na tabela 14 e para as três classes temos resultados que são bem interessantes.

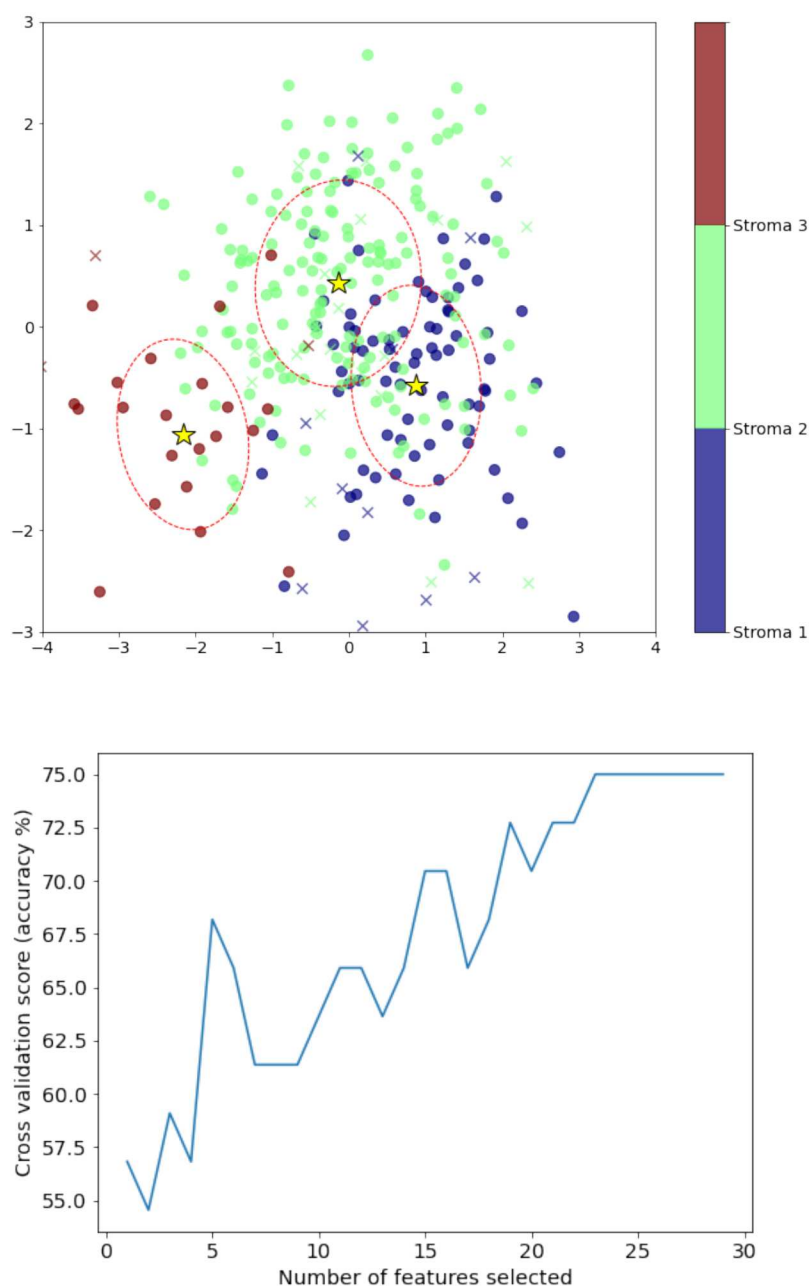


Figura 31 – Gráfico LDA para a separação entre os três tipos de reação estromal. De forma análoga aos demais casos, os símbolos cheios representam os dados de treinamento, a cruz os dados de teste, a estrela representa o valor médio para cada componente e a elipse envolve um desvio padrão da amostragem e mostra a direção preferencial de cada componente. A acurácia do LDA entre as amostras com câncer e as normais, para a separação pelo tipo de estroma é de $75\% \pm 5\%$, como mostrado em (B).

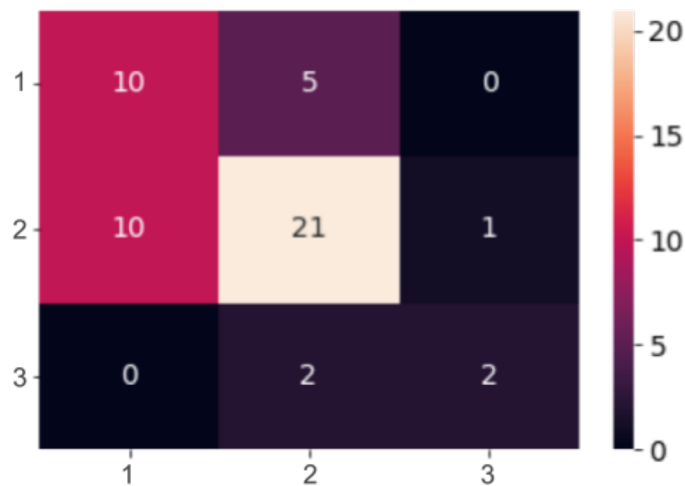


Figura 32 – Matriz de Confusão para a Separação por estroma, na diagonal principal temos os verdadeiros positivos para essa separação.

Tabela 14 – Análises Estatísticas: Classificação por Estroma.

	Estroma 1	Estroma 2	Estroma 3
Precisão	0.50	0.75	0.67
Sensibilidade	0.67	0.66	0.50
Especificidade	0.70	0.71	0.97

4.5.4 Separação dos Gleasons 3 + 4 e 4 + 3 através do Estroma

Vimos que os padrões Gleason 3 + 4 e 4 + 3 são difíceis de se separar quando olhamos essas duas somas por si só, mas se combinarmos o padrão Gleason e o padrão de Estroma, conseguimos observar uma melhor separação entre esses dois tipos, conforme apresentado na figura 33. Os grupos 3 + 4 e 4 + 3 foram separados pelo tipo de estroma 1 (RS1) e 2 (RS2), como 3 + 4 RS1, 3 + 4 RS2, 4 + 3 RS1 e 4 + 3 RS2. Podemos observar claramente os quatro grupos distintos. No entanto, o número de amostras, especialmente para 3 + 4 RS1 (14 imagens) é muito pequeno para realizar os cálculos estatísticos. A ideia é apenas mostrar a importância de se olhar todos os fatores presentes na região tumoral e computar todos para classificarmos um carcinoma de forma mais eficiente, o que melhora os prognósticos dados ao se fazerem as biópsias de próstata. Mas é necessário chamar a atenção que a classificação feita dessa forma é limitada pela pouca quantidade de amostras.

Em resumo, estes resultados indicam que um sistema de classificação composto que inclua informações sobre a intensidade do estroma reativo ao sistema de Gleason baseado na morfologia dos ácinos poderia levar a abordagens mais precisas de diagnóstico e prognóstico. Sabe-se que os tumores estromogênicos apresentam comportamento mais agressivo. No entanto, seriam necessários mais estudos e dados clínicos para melhor compreender e fortalecer estes resultados.

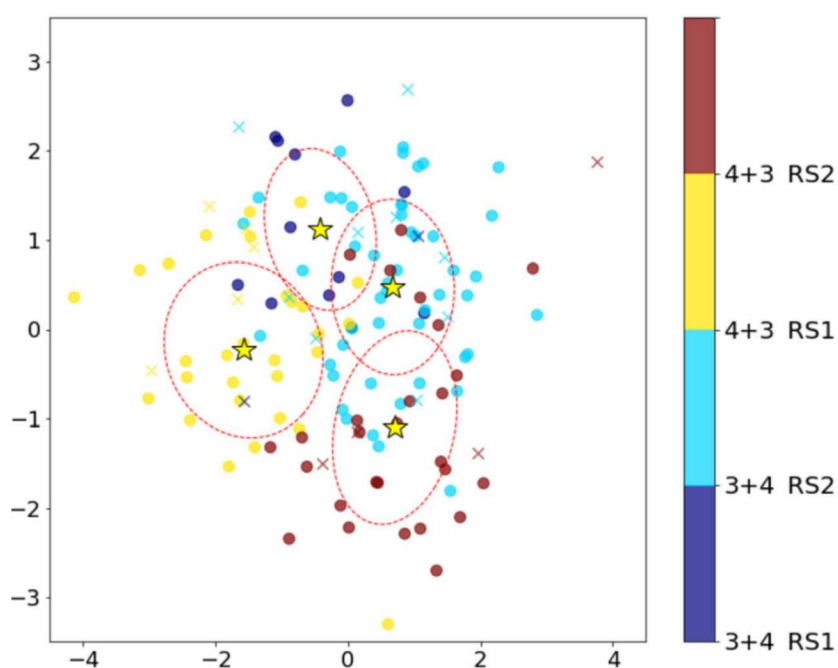


Figura 33 – Gráfico LDA combinando os padrões de soma Gleason $3 + 4$ e $4 + 3$, com o padrão de estroma, nesse caso estão presentes apenas os tipos 1 e 2. Símbolo cheio representa os dados de treinamento, a cruz os dados de teste, a estrela representa o valor médio para cada componente e a elipse envolve um desvio padrão da amostragem e mostra a direção preferencial de cada componente.

5 Conclusões e Perspectivas

Neste trabalho utilizei medidas de microscopia não linear por geração de segundo harmônico e também a fotoluminescência por absorção de dois fótons para analisar biópsias de tecido de próstata, tanto apresentando tecido normal quanto com a presença de carcinoma. As amostras com câncer foram separadas em 3 grupos distintos para comparação e quantificação das mudanças estruturais que ocorrem na presença de um tumor. As mudanças, que são identificadas por um patologista, incluem as alterações na organização das fibras de colágeno, mudanças nas formas e características dos ácinos e alterações no estroma.

Também foi apresentado e elaborado em conjunto com outro grupo de pesquisa, um software de análises, baseado na linguagem python, que tendo como base o tensor estrutural e a extração de redes de fibras, me permitiu identificar e separar com boa eficiência as áreas de fibras e parênquima, presentes na amostra. Isso possibilita a comparação entre os grupos de classificação propostos pelo patologista e o padrão Gleason, padrão ouro de classificação tumoral para a próstata. As imagens por mapa de cores, obtidas pelo software, permitem que sejam analisadas tanto a organização quanto a orientação das fibras de colágeno. Foi possível observar alterações que ocorrem na região que envolve o parênquima, como o crescimento do número desses segmentos que de forma implícita demonstram o uma progressão do número de células tumorais. Foi possível perceber a diferença entre alguns tipos de tumorais onde analisamos a situação em relação as características dos ácinos e como essas alterações podem afetar a estrutura em torno do tumor. Todos estes detalhes se mostraram eficientes nas separações entre as amostras de tumores de próstata normais e malignos.

As separações que foram feitas pelo padrão Gleason, pelo tipo de ácino e pelas alterações no estroma, mostram que as alterações presentes no estroma reativo bem como as mudanças nos tipos acinares podem ser fatores importantes para a definição de prognósticos mais eficientes e com isso proporcionar uma melhor qualidade de vida para o paciente. Dentre as separações, quando é feita uma análise por discriminante linear, que pode ser observada pelos gráficos de LDA, as três separações tem a mesma acurácia para a comparação entre tecidos normais $\times$ câncer, em torno de 90%, o que demonstra a eficiência dos modelos propostos na interpretação entre esses dois tipos. Além desse fato, quando comparamos essas separações levando-se em conta apenas a presença de um carcinoma, temos que a análise do estroma reativo permite uma melhor separação entre os tipos tumorais, o que infere que as alterações no estroma tem carácter significativo na progressão tumoral e nos prognósticos. Também é valido notar que quando olhamos para o estroma reativo juntamente com o padrão de Gleason, conseguimos separar de forma

mais eficiente os tipos de tumor Gleason $3 + 4$ e $4 + 3$, que de outras formas se confundem bastante por terem estruturas muito semelhantes, essa é uma análise que seria interessante de se refazer obtendo um número maior de biópsias contendo essas duas situações.

O que podemos concluir, até o presente momento é que em se tratando do câncer de próstata, é importante se levar em consideração alterações em todo o tecido da próstata devido a presença de um tumor, ou seja, olharmos para as modificações acinares e no estroma, e que essas alterações muitas vezes podem não ser visualmente identificadas de forma adequada. Por esse motivo tem crescido o interesse em utilizar um método que permita prognóstico que possa ser quantificado e que tenha pouca ou nenhuma influência de um viés confirmatório. É importante apontar aqui um fator que é um diferencial nesse trabalho, essa colaboração entre um grupo de Física e um grupo de Medicina, que permite que a pesquisa avance de forma eficiente tanto na parte da análise computacional e de imagens, quanto na identificação de padrões biológicos que podem sofrer alteração por conta do câncer e dessa forma poder prever de forma mais eficiente a evolução dos diversos tipos de carcinoma, bem como a discussão rápida e eficiente que pude ter com os patologistas nas discussões sobre os resultados apresentados aqui nessa tese.

Os próximos passos nesta pesquisa será a avaliação das predições em relação à evolução de casos clínicos para uma possível validação da metodologia de diagnóstico.

Referências

- [1] Organization, World Health: *Cancer*. Em *Key facts*. (Data de acesso 13/01/2021), 2020. <https://www.who.int/news-room/fact-sheets/detail/cancer>. Citado na página 12.
- [2] Cancer, International Agency for Research on: *Cancer Charts*. Em *Prostate cancer facts*. (Data de acesso 29/05/2023), 2020. <https://bit.ly/3mLzyKz>. Citado na página 12.
- [3] Câncer, INCA Instituto Nacional de: *INCA estima 704 mil casos de câncer por ano no Brasil até 2025*. Em *Estimativa*. (Data de acesso 06/07/2023), 2022. <https://www.gov.br/inca/pt-br/assuntos/noticias/2022/inca-estima-704-mil-casos-de-cancer-por-ano-no-brasil-ate-2025>. Citado na página 12.
- [4] Gleason, Donald F: *Classification of prostatic carcinomas*. Cancer Chemother. Rep., 50:125–128, 1966. Citado 2 vezes nas páginas 12 e 23.
- [5] Ayala, Gustavo, Jennifer A. Tuxhorn, Thomas M. Wheeler, Anna Frolov, Peter T. Scardino, Makoto Otori, Marcus Wheeler, Jeffrey Spitler e David R. Rowley: *Reactive Stroma as a Predictor of Biochemical-Free Recurrence in Prostate Cancer*. Clinical Cancer Research, 9(13):4792–4801, 2003. Citado 3 vezes nas páginas 12, 20 e 21.
- [6] Ouellette, Jonathan N., Cole R. Drifka, Kelli B. Pointer, Yuming Liu, Tyler J Lieberthal, W John Kao, John S. Kuo, Agnes G. Loeffler e Kevin W. Eliceiri: *Navigating the Collagen Jungle: The Biomedical Potential of Fiber Organization in Cancer*. Bioengineering, 8(2):17, 2021, ISSN 2306-5354. <https://www.mdpi.com/2306-5354/8/2/17>. Citado na página 12.
- [7] Falzon, G, S Pearson e R Murison: *Analysis of collagen fibre shape changes in breast cancer*. Physics in Medicine & Biology, 53(23):6641, 2008. Citado na página 12.
- [8] Provenzano, Paolo P, Kevin W Eliceiri, Jay M Campbell, David R Inman, John G White e Patricia J Keely: *Collagen reorganization at the tumor-stromal interface facilitates local invasion*. BMC Medicine, 4(1):38, 2006. Citado na página 12.
- [9] Han, Xiaoxing, Ryan M Burke, Martha L Zettel, Ping Tang e Edward B Brown: *Second harmonic properties of tumor collagen: determining the structural relationship between reactive stroma and healthy stroma*. Optics Express, 16(3):1846–1859, 2008. Citado na página 12.

- [10] Provenzano, Paolo P, David R Inman, Kevin W Eliceiri, Justin G Knittel, Long Yan, Curtis T Rueden, John G White e Patricia J Keely: *Collagen density promotes mammary tumor initiation and progression*. BMC Medicine, 6(1):11, 2008. Citado na página 12.
- [11] Tuer, Adam E, Serguei Krouglov, Nicole Prent, Richard Cisek, Daaf Sandkuijl, Kazuhiro Yasufuku, Brian C Wilson e Virginijus Barzda: *Nonlinear optical properties of type I collagen fibers studied by polarization dependent second harmonic generation microscopy*. The Journal of Physical Chemistry B, 115(44):12759–12769, 2011. Citado na página 12.
- [12] Campagnola, Paul: *Second harmonic generation imaging microscopy: applications to diseases diagnostics*. Analytical Chemistry, 83(9):3224–3231, 2011. Citado na página 12.
- [13] Conklin, Matthew W, Jens C Eickhoff, Kristin M Riching, Carolyn A Pehlke, Kevin W Eliceiri, Paolo P Provenzano, Andreas Friedl e Patricia J Keely: *Aligned collagen is a prognostic signature for survival in human breast carcinoma*. The American Journal of Pathology, 178(3):1221–1232, 2011. Citado 2 vezes nas páginas 12 e 13.
- [14] Ajeti, Visar, Oleg Nadiarnykh, Suzanne M Ponik, Patricia J Keely, Kevin W Eliceiri e Paul J Campagnola: *Structural changes in mixed Col I/Col V collagen gels probed by SHG microscopy: implications for probing stromal alterations in human breast cancer*. Biomedical Optics Express, 2(8):2307–2316, 2011. Citado 2 vezes nas páginas 12 e 13.
- [15] Burke, Kathleen, Ping Tang e Edward B Brown: *Second harmonic generation reveals matrix alterations during breast tumor progression*. Journal of Biomedical Optics, 18(3):031106, 2012. Citado na página 12.
- [16] Ghazaryan, Ara, Halley F Tsai, Gor Hayrapetyan, Wei Liang Chen, Yang Fang Chen, Myung Yung Jeong, Chang Seok Kim, Shean Jen Chen e Chen Yuan Dong: *Analysis of collagen fiber domain organization by Fourier second harmonic generation microscopy*. Journal of Biomedical Optics, 18(3):031105, 2012. Citado na página 12.
- [17] Zhang, Kun, Callie A Corsa, Suzanne M Ponik, Julie L Prior, David Piwnica-Worms, Kevin W Eliceiri, Patricia J Keely e Gregory D Longmore: *The collagen receptor discoidin domain receptor 2 stabilizes SNAIL1 to facilitate breast cancer metastasis*. Nature Cell Biology, 15(6):677–687, 2013. Citado na página 12.
- [18] Bredfeldt, Jeremy S, Yuming Liu, Carolyn A Pehlke, Matthew W Conklin, Joseph M Szulczewski, David R Inman, Patricia J Keely, Robert D Nowak, Thomas R Mackie e Kevin W Eliceiri: *Computational segmentation of collagen fibers from second-harmonic*

- generation images of breast cancer*. Journal of Biomedical Optics, 19(1):016007, 2014. Citado na página 12.
- [19] Adur, Javier, Vitor B. Pelegati, Andre A. de Thomaz, Mariana O. Baratti, Diogo B. Almeida, L. A. L. A. Andrade, Fátima Bottcher-Luiz, Hernandes F. Carvalho e Carlos L. Cesar: *Optical Biomarkers of Serous and Mucinous Human Ovarian Tumor Assessed with Nonlinear Optics Microscopies*. Plos One, 7(10):e47007, outubro 2012. Citado na página 12.
- [20] Adur, Javier, Vitor B. Pelegati, Andre A. de Thomaz, Mariana O. Baratti, Liliana A. L. A. Andrade, Hernandes F. Carvalho, Fátima Bottcher-Luiz e Carlos Lenz Cesar: *Second harmonic generation microscopy as a powerful diagnostic imaging modality for human ovarian cancer*. Journal of Biophotonics, 7(1-2):37–48, 2014, ISSN 1864-0648. Citado na página 12.
- [21] Wu, Xiufeng, Gang Chen, Jianping Lu, Weifeng Zhu, Jingting Qiu, Jianxin Chen, Shusen Xie, Shuangmu Zhuo e Jun Yan: *Label-free detection of breast masses using multiphoton microscopy*. PLOS One, 8(6):e65933, 2013. Citado na página 12.
- [22] Bizzarri, Mariano e Alessandra Cucina: *Tumor and the microenvironment: a chance to reframe the paradigm of carcinogenesis?* BioMed Research International, 2014:934038. Citado na página 12.
- [23] Tao, Yuankai K, Dejun Shen, Yuri Sheikine, Osman O. Ahsen, Helen H. Wang, Daniel B. Schmolze, Nicole B. Johnson, Jeffrey S. Brooker, Alex E. Cable, James L. Connolly et al.: *Assessment of breast pathologies using nonlinear microscopy*. Proceedings of the National Academy of Sciences, 111(43):15304–15309, 2014. Citado na página 12.
- [24] Brabrand, Anders, Ian I. Kariuki, Monica J. Engstrøm, Olav A. Haugen, Linda A. Dyrnes, Bjørn O. Åsvold, Magnus B. Lilledahl e Anna M. Bofin: *Alterations in collagen fibre patterns in breast cancer. A premise for tumour invasiveness?* APIMIS, 123(1):1–8, 2015. Citado na página 12.
- [25] Tan, Wai Jin, Jie Yan, Shuoyu Xu, Aye Aye Thike, Boon Huat Bay, Hanry Yu, Min Han Tan e Puay Hoon Tan: *Second harmonic generation microscopy is a novel technique for differential diagnosis of breast fibroepithelial lesions*. Journal of Clinical Pathology, 68(12):1033–1035, 2015. Citado na página 12.
- [26] Tilbury, Karissa e Paul J. Campagnola: *Applications of Second-Harmonic Generation Imaging Microscopy in Ovarian and Breast Cancer*. Perspectives in Medicinal Chemistry, 7:21–32, 2015. Citado na página 12.
- [27] Golaraei, Ahmad, Lukas Kontenis, Richard Cisek, Danielle Tokarz, Susan J. Done, Brian C. Wilson e Virginijus Barzda: *Changes of collagen ultrastructure in breast cancer*

- tissue determined by second-harmonic generation double Stokes-Mueller polarimetric microscopy*. Biomed. Opt. Express, 7(10):4054–4068, Oct 2016. Citado na página 12.
- [28] Hall, Matthew S, Farid Alisafaei, Ehsan Ban, Xinzeng Feng, Chung Yuen Hui, Vivek B Shenoy e Mingming Wu: *Fibrous nonlinear elasticity enables positive mechanical feedback between cells and ECMs*. Proceedings of the National Academy of Sciences, 113(49):14043–14048, 2016. Citado na página 12.
- [29] Wang, Karin, Fei Wu, Bo Ri Seo, Claudia Fischbach, Weisi Chen, Lauren Hsu e Delphine Gourdon: *Breast cancer cells alter the dynamics of stromal fibronectin-collagen interactions*. Matrix Biology, 60–61:86–95, 2017. Citado na página 12.
- [30] Case, Ashley, Becky K. Brisson, Amy C. Durham, Suzanne Rosen, James Monslow, Elizabeth Buza, Pascale Salah, Julie Gillem, Gordon Ruthel, Sridhar Veluvolu, Veronica Kristiansen, Ellen Puré, Dorothy C. Brown, Karin U. Sørenmo e Susan W. Volk: *Identification of prognostic collagen signatures and potential therapeutic stromal targets in canine mammary gland carcinoma*. PLOS ONE, 12(7):e0180448, julho 2017. Citado na página 12.
- [31] Conklin, Matthew W, Ronald E Gangnon, Brian L Sprague, Lisa Van Gemert, John M Hampton, Kevin W Eliceiri, Jeremy S Bredfeldt, Yuming Liu, Nuntida Surachaicharn, Polly A Newcomb, Andreas Friedl, Patricia J Keely e Amy Trentham-Dietz: *Collagen Alignment as a Predictor of Recurrence after Ductal Carcinoma In Situ*. Cancer Epidemiol Biomarkers Prev., 27(2):138–145, 2018. Citado na página 12.
- [32] Natal, Rodrigo A, José Vassallo, Geisilene R Paiva, Vitor B Pelegati, Guilherme O Barbosa, Guilherme R Mendonça, Caroline Bondarik, Sophie F Derchain, Hernandes F Carvalho, Carmen S Lima *et al.*: *Collagen analysis by second-harmonic generation microscopy predicts outcome of luminal breast cancer*. Tumor Biology, 40(4):1010428318770953, 2018. Citado na página 12.
- [33] Natal, Rodrigo Andrade, Geisilene R Paiva, Vitor B Pelegati, Ludwing Marenco, César A Alvarenga, Renato F Vargas, Sophie F Derchain, Luis O Sarian, Camille Franchet, Carlos L Cesar *et al.*: *Exploring collagen parameters in pure special types of Invasive Breast Cancer*. Scientific Reports, 9(1):7715, 2019. Citado na página 12.
- [34] Best, Sara L., Yuming Liu, Adib Keikhosravi, Cole R. Drifka, Kaitlin M. Woo, Guneet S. Mehta, Marie Altwegg, Terra N. Thimm, Matthew Houlihan, Jeremy S. Bredfeldt, E. Jason Abel, Wei Huang e Kevin W. Eliceiri: *Collagen organization of renal cell carcinoma differs between low and high grade tumors*. BMC Cancer, 19(1):490, May 2019. Citado na página 12.
- [35] Gole, L, J Yeong, J C T Lim, K H Ong, H Han, A A Thike, Y C Poh, S Yee, J Iqbal, W Hong, B Lee, W Yu e P H Tan: *Quantitative stain-free imaging and digital profiling*

- of collagen structure reveal diverse survival of triple negative breast cancer patients.* Breast Cancer Research, 22(1):42, 2020. Citado na página 12.
- [36] Nguyen, Tan Huu, Shamira Sridharan, Virgilia Macias, Andre Kajdacsy-Balla, Jonathan Melamed, Minh N Do e Gabriel Popescu: *Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning.* Journal of Biomedical Optics, 22(3):036015, 2017. Citado na página 13.
- [37] Quiros, Adalberto Claudio, Roderick Murray-Smith e Ke Yuan: *Pathology GAN: learning deep representations of cancer tissue.* arXiv preprint arXiv:1907.02644, 2019. Citado na página 13.
- [38] Goldenberg, S. Larry, Guy Nir e Septimiu E. Salcudean: *A new era: artificial intelligence and machine learning in prostate cancer.* Nature Reviews Urology, 16(7):391–403, 2019. Citado na página 13.
- [39] Campagnola, PJ, W Mohler e AE Millard: *3-dimensional high-resolution second harmonic generation imaging of endogenous structural proteins in biological tissues.* Biophysical Journal, 82(1):175A–175A, 2002. Citado na página 13.
- [40] Campagnola, Paul: *Second harmonic generation imaging microscopy: applications to diseases diagnostics.* Analytical Chemistry, 83(9):3224–3231, 2011. Citado na página 13.
- [41] Garcia, Aloisio Miguel: *Fluorescência por absorção de dois fótons em corantes com aplicações biológicas.* Tese de Mestrado, Universidade Federal de Minas Gerais, 2013. Citado 2 vezes nas páginas 13 e 17.
- [42] Garcia, Aloísio M, Felipe L Magalhes, Jaqueline S Soares, Eduardo Paulino Junior, Mário FR de Lima, Marcelo Mamede e Ana M de Paula: *Second harmonic generation imaging of the collagen architecture in prostate cancer tissue.* Biomedical Physics & Engineering Express, 4(2):025026, 2018. Citado 3 vezes nas páginas 13, 37 e 79.
- [43] Hu, Wenyan, Hui Li, Ling Fu, Chunyou Wang e Shanmiao Gou: *Characterization of collagen fibers by means of texture analysis of second harmonic generation images using orientation-dependent gray level co-occurrence matrix method.* Journal of Biomedical Optics, 17(2):026007, 2012. Citado na página 13.
- [44] Wen, Bruce L, Molly A Brewer, Oleg Nadiarnykh, James D Hocker, Vikas Singh, Thomas R Mackie e Paul J Campagnola: *Texture analysis applied to second harmonic generation image data for ovarian cancer classification.* Journal of Biomedical Optics, 19(9):096007, 2014. Citado na página 13.

- [45] Bredfeldt, Jeremy S, Yuming Liu, Carolyn A Pehlke, Matthew W Conklin, Joseph M Szulczewski, David R Inman, Patricia J Keely, Robert D Nowak, Thomas R Mackie e Kevin W Eliceiri: *Computational segmentation of collagen fibers from second-harmonic generation images of breast cancer*. Journal of Biomedical Optics, 19(1):016007, 2014. Citado na página 13.
- [46] Reis, Luana A, Ana PV Garcia, Egleidson FA Gomes, Francis GJ Longford, Jeremy G Frey, Geovanni D Cassali e Ana M de Paula: *Canine mammary cancer diagnosis from quantitative properties of nonlinear optical images*. Biomedical Optics Express, 11(11):6413–6427, 2020. Citado na página 13.
- [47] Birk, John W, Micheal Tadros, Koorosh Moezardalan, Oleg Nadyarnykh, Faripour Forouhar, Joseph Anderson e Paul Campagnola: *Second harmonic generation imaging distinguishes both high-grade dysplasia and cancer from normal colonic mucosa*. Digestive Diseases and Sciences, 59(7):1529–1534, 2014. Citado na página 13.
- [48] Ambekar, Raghu, Tung Yuen Lau, Michael Walsh, Rohit Bhargava e Kimani C Toussaint: *Quantifying collagen structure in breast biopsies using second-harmonic generation imaging*. Biomedical Optics Express, 3(9):2021–2035, 2012. Citado na página 13.
- [49] Golaraei, Ahmad, Richard Cisek, Serguei Krouglov, Roya Navab, Carolyn Niu, Shingo Sakashita, Kazuhiro Yasufuku, Ming Sound Tsao, Brian C Wilson e Virginijus Barzda: *Characterization of collagen in non-small cell lung carcinoma with second harmonic polarization microscopy*. Biomedical Optics Express, 5(10):3562–3567, 2014. Citado na página 13.
- [50] Hompland, Tord, Arne Erikson, Mikael Lindgren, Tore Lindmo e Catharina de Lange Davies: *Second-harmonic generation in collagen as a potential cancer diagnostic parameter*. Journal of Biomedical Optics, 13(5):054050, 2008. Citado na página 13.
- [51] Adur, Javier, Vitor B Pelegati, Andre A de Thomaz, Mariana O Baratti, Liliana ALA Andrade, Hernandes F Carvalho, Fátima Bottcher-Luiz e Carlos Lenz Cesar: *Second harmonic generation microscopy as a powerful diagnostic imaging modality for human ovarian cancer*. Journal of Biophotonics, 7(1-2):37–48, 2014. Citado na página 13.
- [52] Ling, Yuting, Chunhui Li, Kairui Feng, Scott Palmer, Paul L. Appleton, Stephen Lang, David McGloin, Zhihong Huang e Ghulam Nabi: *Second harmonic generation (SHG) imaging of cancer heterogeneity in ultrasound guided biopsies of prostate in men suspected with prostate cancer*. Journal of Biophotonics, 10(6):911–918, 2017. Citado na página 13.

- [53] Esquibel, Corinne R, Kristy D Wendt, Heui C Lee, Janak Gaire, Andrew Shoffstall, Morgan E Urdaneta, Jenu V Chacko, Sarah K Brodnick, Kevin J Otto, Jeffrey R Capadona *et al.*: *Second harmonic generation imaging of collagen in chronically implantable electrodes in brain tissue*. *Frontiers in Neuroscience*, 14:95, 2020. Citado na página 13.
- [54] Condeelis, John e Jeffrey E Segall: *Intravital imaging of cell movement in tumours*. *Nature Reviews Cancer*, 3(12):921–930, 2003. Citado na página 13.
- [55] Drifka, Cole R, Kevin W Eliceiri, Sharon M Weber e W John Kao: *A bioengineered heterotypic stroma–cancer microenvironment model to study pancreatic ductal adenocarcinoma*. *Lab on a Chip*, 13(19):3965–3975, 2013. Citado na página 13.
- [56] Sung, Kyung Eun, Gui Su, Carolyn Pehlke, Steven M Trier, Kevin W Eliceiri, Patricia J Keely, Andreas Friedl e David J Beebe: *Control of 3-dimensional collagen matrix polymerization for reproducible human mammary fibroblast cell culture in microfluidic devices*. *Biomaterials*, 30(27):4833–4841, 2009. Citado na página 13.
- [57] Hanley, Christopher J, Fergus Noble, Matthew Ward, Marc Bullock, Cole Drifka, Massimiliano Mellone, Antigoni Manousopoulou, Harvey E Johnston, Annette Hayden, Steve Thirdborough *et al.*: *A subset of myofibroblastic cancer-associated fibroblasts regulate collagen fiber elongation, which is prognostic in multiple cancers*. *Oncotarget*, 7(5):6159, 2016. Citado na página 13.
- [58] Brown, Edward, Trevor McKee, Alain Pluen, Brian Seed, Yves Boucher, Rakesh K Jain *et al.*: *Dynamic imaging of collagen and its modulation in tumors in vivo using second-harmonic generation*. *Nature Medicine*, 9(6):796–800, 2003. Citado na página 13.
- [59] Yang, Qinqin, Zhexin Xu, Chenxi Liao, Jianyong Cai, Ying Huang, Hong Chen, Xuan Tao, Zheng Huang, Jianxin Chen, Jiyang Dong *et al.*: *Epithelium segmentation and automated Gleason grading of prostate cancer via deep learning in label-free multiphoton microscopic images*. *Journal of Biophotonics*, 13(2):e201900203, 2020. Citado na página 13.
- [60] Rezakhaniha, Rana, Aristotelis Agianniotis, Jelle Tymen Christiaan Schrauwen, Alessandra Griffa, Daniel Sage, CVC vd Bouten, FN Van De Vosse, Michaël Unser e Nikolaos Stergiopoulos: *Experimental investigation of collagen waviness and orientation in the arterial adventitia using confocal laser scanning microscopy*. *Biomechanics and Modeling in Mechanobiology*, 11(3):461–473, 2012. Citado 3 vezes nas páginas 13, 37 e 38.
- [61] Boudaoud, Arezki, Agata Burian, Dorota Borowska-Wykręt, Magalie Uyttewaal, Roman Wrzalik, Dorota Kwiatkowska e Olivier Hamant: *FibrilTool, an ImageJ*

- plug-in to quantify fibrillar structures in raw microscopy images*. Nature protocols, 9(2):457–463, 2014. Citado na página 13.
- [62] Liu, Yuming, Adib Keikhosravi, Guneet S Mehta, Cole R Drifka e Kevin W Eliceiri: *Methods for quantifying fibrillar collagen alignment*. Em *Fibrosis*, páginas 429–451. Springer, 2017. Citado 2 vezes nas páginas 13 e 35.
- [63] Liu, Yuming, Adib Keikhosravi, Carolyn A Pehlke, Jeremy S Bredfeldt, Matthew Dutson, Haixiang Liu, Guneet S Mehta, Robert Claus, Akhil J Patel, Matthew W Conklin *et al.*: *Fibrillar collagen quantification with curvelet transform based computational methods*. Frontiers in Bioengineering and Biotechnology, 8:198, 2020. Citado na página 13.
- [64] Bredfeldt, Jeremy S, Yuming Liu, Matthew W Conklin, Patricia J Keely, Thomas R Mackie e Kevin W Eliceiri: *Automated quantification of aligned collagen for human breast carcinoma prognosis*. Journal of Pathology Informatics, 5, 2014. Citado na página 13.
- [65] Kweldam, CF, GJ van Leenders e T van der Kwast: *Grading of prostate cancer: a work in progress*. Histopathology, 74(1):146–160, 2019. Citado na página 13.
- [66] Shen, Yuen Ron: *Principles of nonlinear optics*. Wiley-Interscience, New York, NY, USA, 1984. Citado 2 vezes nas páginas 15 e 16.
- [67] Franken, P A, A E Hill, C W Peters e G Wienreich: *Generation of Optical Harmonics*. Physical Review Letters, 7(4):118–119, 1961. Citado na página 17.
- [68] Göppert-Mayer, Maria: *Über elementarakte mit zwei quantensprüngen*. Annalen der Physik, 401(3):273–294, 1931. Citado na página 17.
- [69] Kaiser, W e CGB Garrett: *Two-photon excitation in $\text{CaF}_2: \text{Eu}^{2+}$* . Physical Review Letters, 7(6):229, 1961. Citado na página 17.
- [70] Zipfel, Warren R., Rebecca M. Williams e Watt W. Webb: *Nonlinear magic: multiphoton microscopy in the biosciences*. Nature Biotechnology, 21(11):1369–1377, 2003. Citado na página 17.
- [71] Rumi, Mariacristina e Joseph W Perry: *Two-photon absorption: an overview of measurements and principles*. Advances in Optics and Photonics, 2(4):451–518, 2010. Citado na página 17.
- [72] So, Peter TC, Chen Y Dong, Barry R Masters e Keith M Berland: *Two-photon excitation fluorescence microscopy*. Annual Review of Biomedical Engineering, 2(1):399–429, 2000. Citado na página 17.

- [73] Nascimento-Júnior, Braz José do: *Anatomia humana sistemática básica*. UNIVASF, 2020. Citado na página 18.
- [74] Duarte, Antonio H, Sicilia Colli, Jorge L Alves-Pereira, Max P Martins, Francisco JB Sampaio e Cristiane F Ramos: *Collagen I and III and metalloproteinase gene and protein expression in prostate cancer in relation to Gleason score*. International Braz. J. Urol., 38:341–355, 2012. Citado na página 18.
- [75] Society, America Cancer: *Prostate Cancer*. Em *cancer*. (Data de acesso 13/04/2021), 2019. <https://www.cancer.org/cancer/prostate-cancer/about/what-is-prostate-cancer.html>. Citado na página 18.
- [76] Lomas, Derek J e Hashim U Ahmed: *All change in the prostate cancer diagnostic pathway*. Nature reviews Clinical oncology, 17(6):372–381, 2020. Citado 2 vezes nas páginas 20 e 59.
- [77] Paulino Júnior, Eduardo: *Caracterização do estroma reativo do câncer de próstata em função da morfologia acinar e da assinatura do colágeno por imuno-histoquímica e à microscopia por geração de segundo harmônico (SHG): a busca de novos marcadores prognósticos*. Tese de Doutorado, Universidade Federal de Minas Gerais - Faculdade de Medicina, 2022. Citado 4 vezes nas páginas 22, 25, 26 e 27.
- [78] Loeb, Stacy, Sophie M Bruinsma, Joseph Nicholson, Alberto Briganti, Tom Pickles, Yoshiyuki Kakehi, Sigrid V Carlsson e Monique J Roobol: *Active Surveillance for Prostate Cancer: A Systematic Review of Clinicopathologic Variables and Biomarkers for Risk Stratification*. European Urology, 67(4):619 – 626, 2015. Citado na página 23.
- [79] Epstein, Jonathan I, Michael J Zelefsky, Daniel D Sjoberg, Joel B Nelson, Lars Egevad, Cristina Magi-Galluzzi, Andrew J Vickers, Anil V Parwani, Victor E Reuter, Samson W Fine *et al.*: *A contemporary prostate cancer grading system: a validated alternative to the Gleason score*. European urology, 69(3):428–435, 2016. Citado na página 23.
- [80] Epstein, Jonathan I: *Prostate cancer grading: a decade after the 2005 modified system*. Modern Pathology, 31:47–63, 2018. Citado na página 24.
- [81] G., Longford Frank: *Pyfibre: Python fibrous image analysis toolkit*. <https://github.com/franklongford/PyFibre>, 2022. Citado na página 32.
- [82] Stein, Andrew M, David A Vader, Louise M Jawerth, David A Weitz e Leonard M Sander: *An algorithm for extracting the network geometry of three-dimensional collagen gels*. Journal of Microscopy, 232(3):463–475, 2008. Citado 3 vezes nas páginas 32, 33 e 34.

- [83] Sato, Yoshinobu, Shin Nakajima, Nobuyuki Shiraga, Hideki Atsumi, Shigeyuki Yoshida, Thomas Koller, Guido Gerig e Ron Kikinis: *Three-dimensional multi-scale line filter for segmentation and visualization of curvilinear structures in medical images*. Medical Image Analysis, 2(2):143–168, 1998. Citado na página 33.
- [84] Ávila, Francisco J e Juan M Bueno: *Analysis and quantification of collagen organization with the structure tensor in second harmonic microscopy images of ocular tissues*. Applied Optics, 54(33):9848–9854, 2015. Citado na página 37.
- [85] Natal, Rodrigo A, José Vassallo, Geisilene R Paiva, Vitor B Pelegati, Guilherme O Barbosa, Guilherme R Mendonça, Caroline Bondarik, Sophie F Derchain, Hernandes F Carvalho, Carmen S Lima *et al.*: *Collagen analysis by second-harmonic generation microscopy predicts outcome of luminal breast cancer*. Tumor Biology, 40(4):1010428318770953, 2018. Citado na página 37.
- [86] Vajapeyam, Sriram: *Understanding shannon’s entropy metric for information*. arXiv preprint arXiv:1405.2061, 2014. Citado na página 38.
- [87] RESEARCH, NIRPY: *Classification of NIR spectra by Linear Discriminant Analysis in Python*. Em LDA. (Data de acesso 28/01/2021), 2018. <https://nirpyresearch.com/classification-nir-spectra-linear-discriminant-analysis-python/>. Citado na página 41.
- [88] Raschka, Sebastian: *Linear Discriminant Analysis*. Em LDA. (Data de acesso 28/01/2021), 2014. https://sebastianraschka.com/Articles/2014_python_lda.html. Citado na página 41.
- [89] James, Gareth, Daniela Witten, Trevor Hastie e Robert Tibshirani: *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. <https://faculty.marshall.usc.edu/gareth-james/ISL/>. Citado na página 42.
- [90] Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt e Gaël Varoquaux: *API design for machine learning software: experiences from the scikit-learn project*. Em ECML PKDD Workshop: Languages for Data Mining and Machine Learning, páginas 108–122, 2013. Citado 2 vezes nas páginas 57 e 94.
- [91] Egevad, Lars, Brett Delahunt, Daniel M Berney, David G Bostwick, John Cheville, Eva Comperat, Andrew J Evans, Samson W Fine, David J Grignon, Peter A Humphrey *et al.*: *Utility of Pathology Imagebase for standardisation of prostate cancer grading*. Histopathology, 73(1):8–18, 2018. Citado na página 59.

APÊNDICE A – Publicações

- **Egleidson F. A. Gomes**, Eduardo Paulino Junior, Mario F. R. de Lima, Luana A. Reis, Giovanna Paranhos, Marcelo Mamede, Francis G. J. Longford, Jeremy G. Frey, Ana Maria de Paula. *Prostate cancer tissue classification by multiphotonimaging, automated image analysis and machine learning*. J. Biophotonics 2023, e202200382. <https://doi.org/10.1002/jbio.202200382>
- Luana A. Reis, Ana P. V. Garcia, **Egleidson F. A. Gomes**, Francis G. J. Longford, Jeremy G. Frey, Geovanni D. Cassali, and Ana M. de Paula. *Canine mammary cancer diagnosis from quantitative properties of nonlinear optical images*. Biomed. Opt. Express 11, 6413-6427 (2020). <https://doi.org/10.1364/B0E.400871>

APÊNDICE B – Gráficos de Caixa

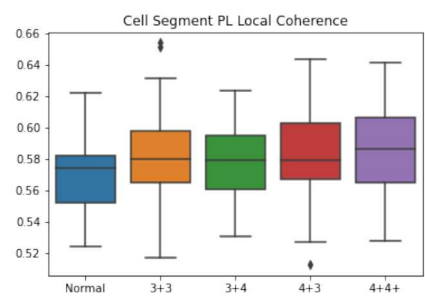
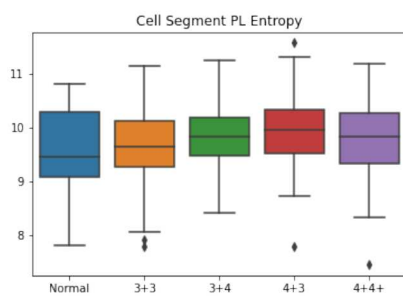
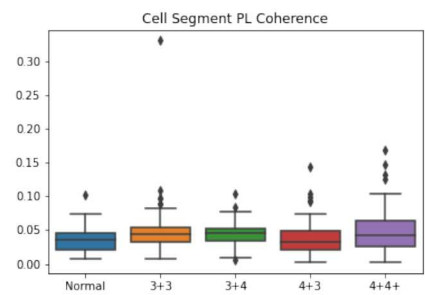
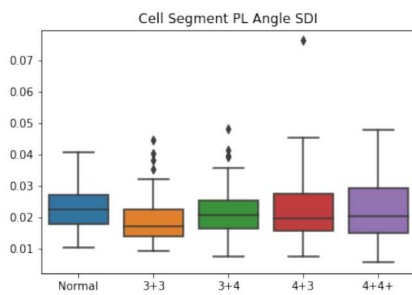
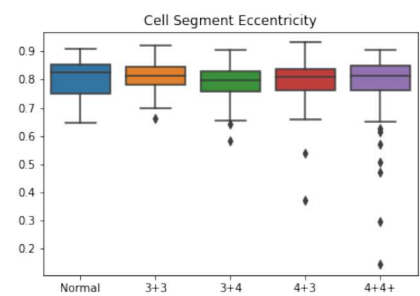
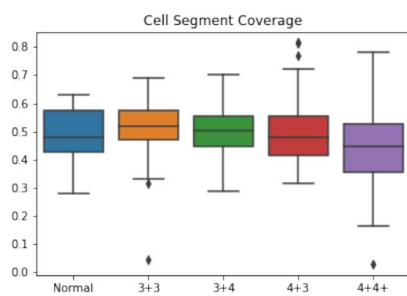
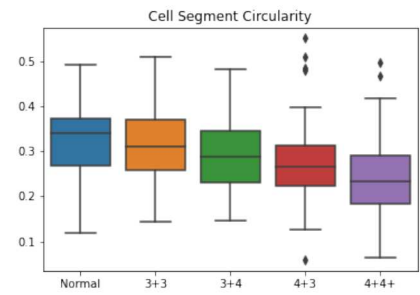
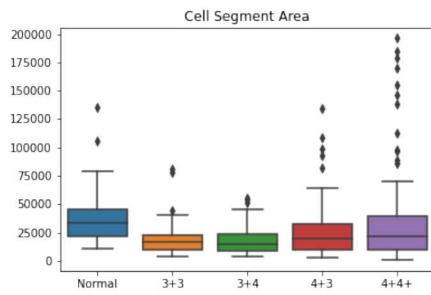
Neste apêndice apresento os boxplots (ou gráficos de caixa) mostrando todas as métricas computadas pelo PyFibre. No total temos um conjunto de 29 métricas. Primeiramente serão apresentados os gráficos relacionados ao Padrão Gleason, em seguida o de Ácinos e por fim, o de Estroma.

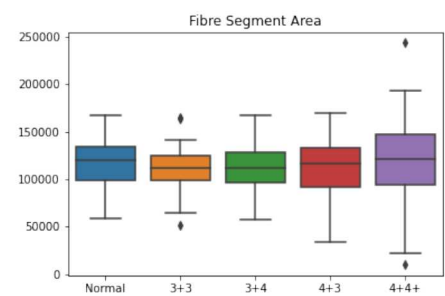
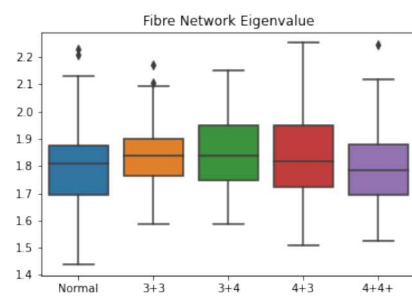
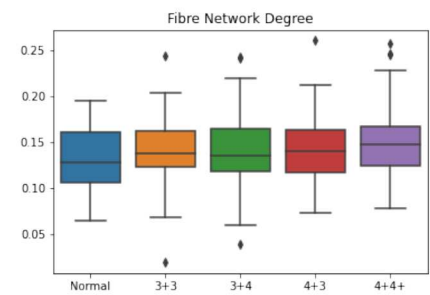
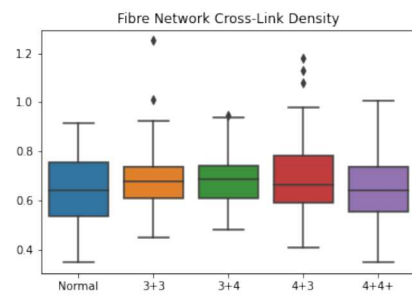
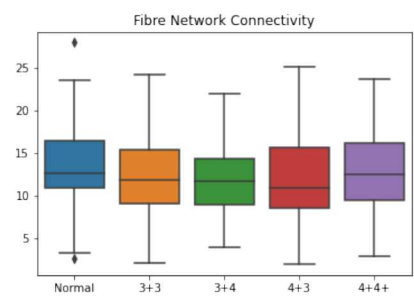
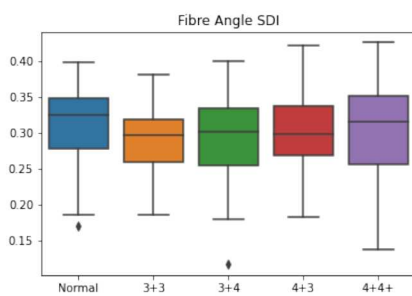
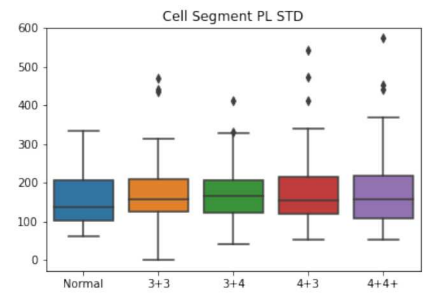
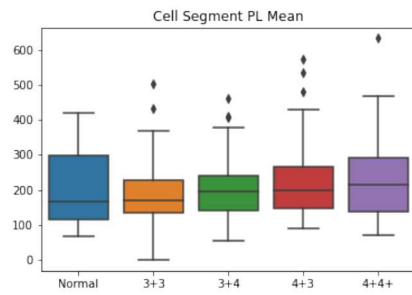
O principal intuito ao apresentarmos esses gráficos boxplots neste apêndice é o de mostrar que individualmente as métricas não são suficientes para separar significativamente os tipos tumorais. O que nos levou à análise geral via LDA.

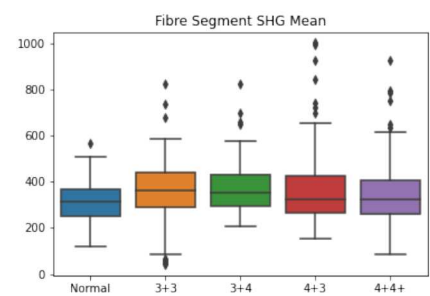
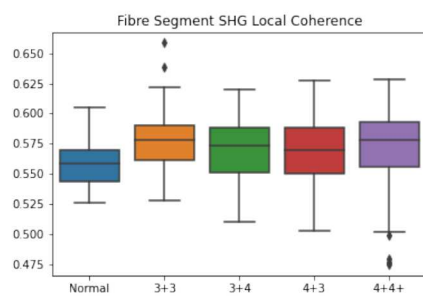
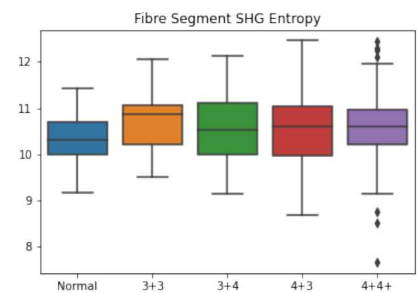
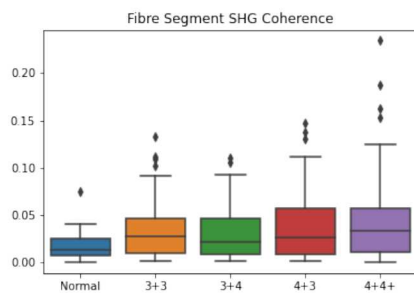
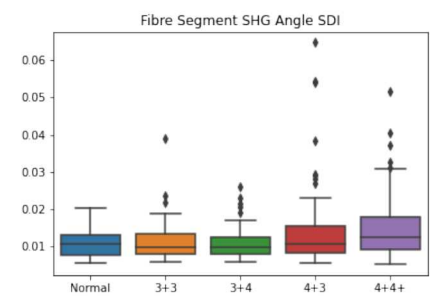
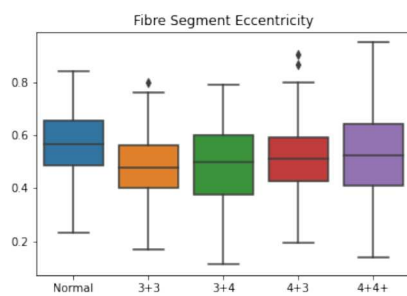
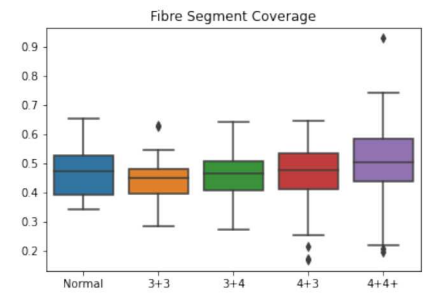
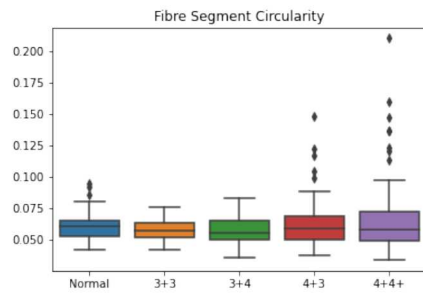
B.1 Boxplots do Padrão Gleason

Ao observar os gráficos relacionados ao padrão de Gleason, principalmente aqueles que dizem respeito às métricas da região de parênquima, quase não é possível tirar conclusões à respeito das diferentes classes. Apenas algumas métricas mostram diferenças principalmente para os tumores mais agressivos, como por exemplo a medida de circularidade.

Mesmo as métricas relacionadas às fibras de colágeno, fornecem pouca informação. A anisotropia das fibras, que já sabemos alterar e aumentar de acordo com a agressividade do tumor [42], parece não variar muito se olharmos globalmente, apesar de existir uma separação maior quando observamos a anisotropia local.







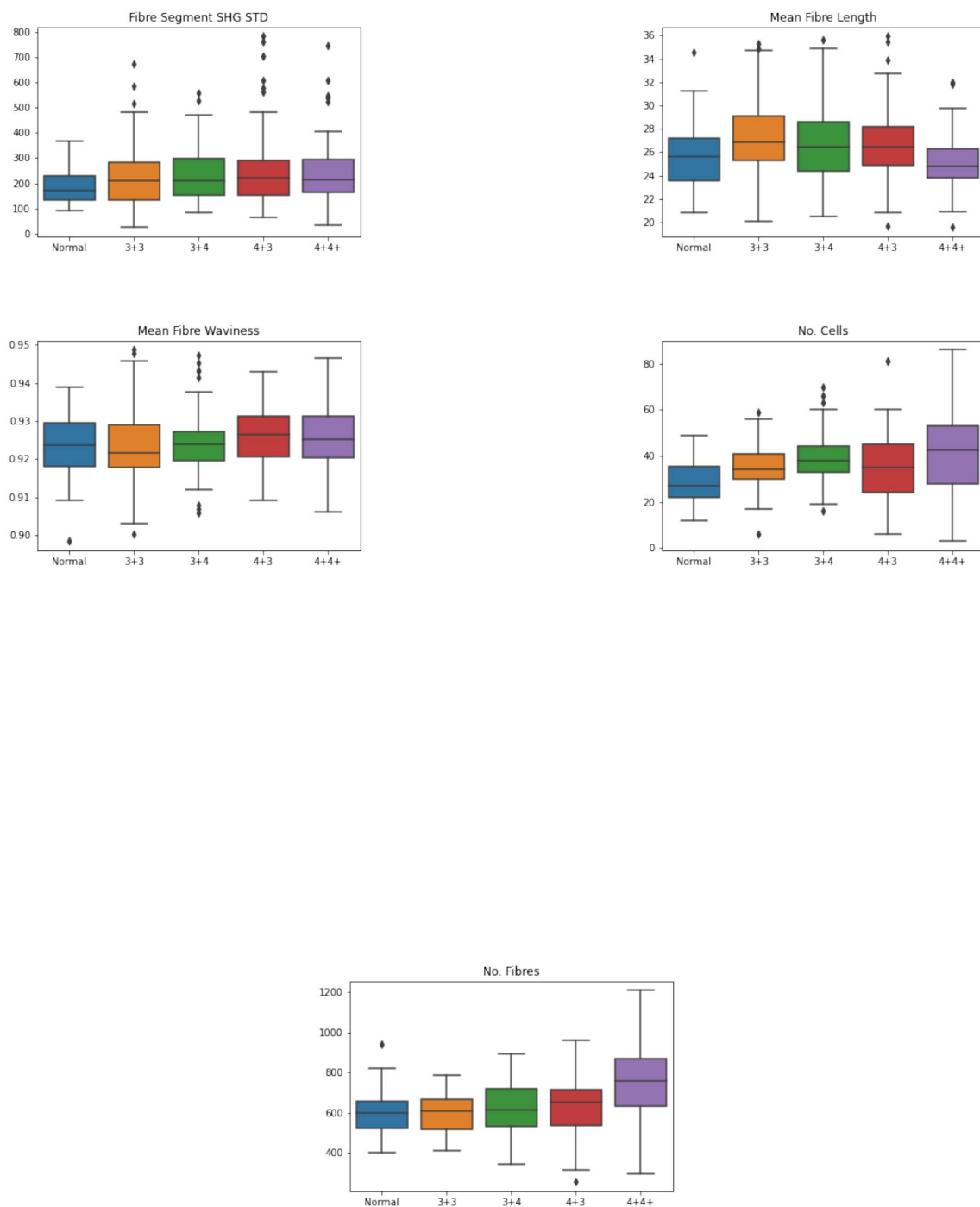
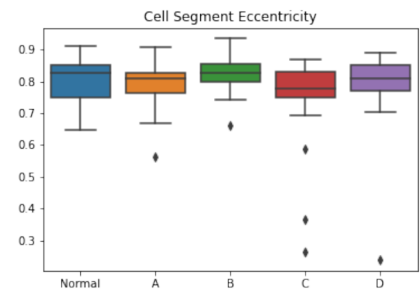
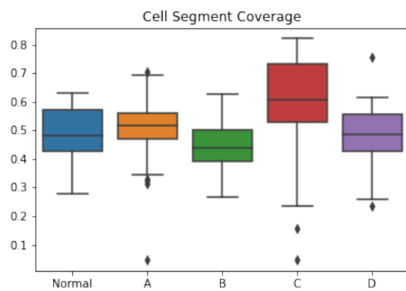
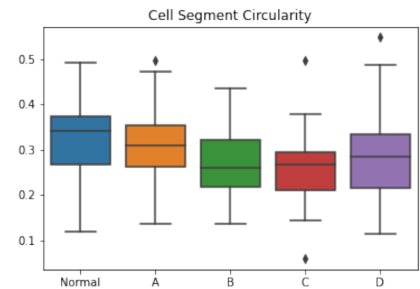
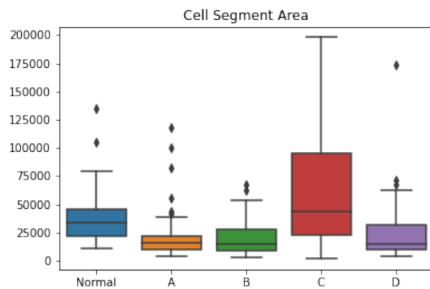


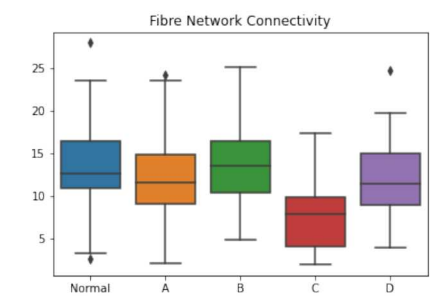
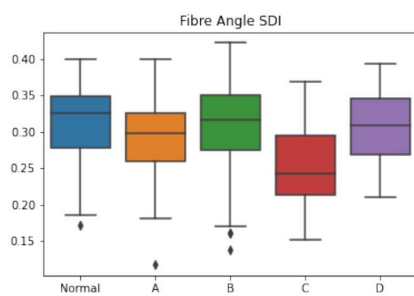
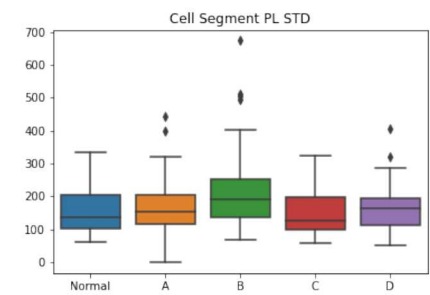
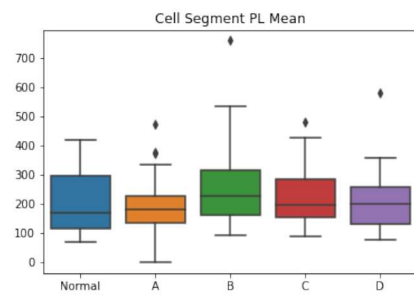
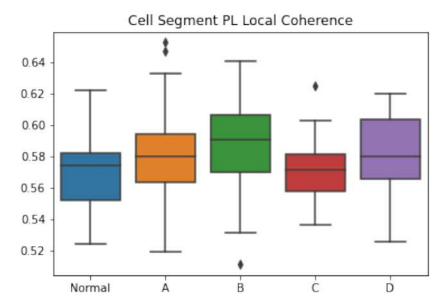
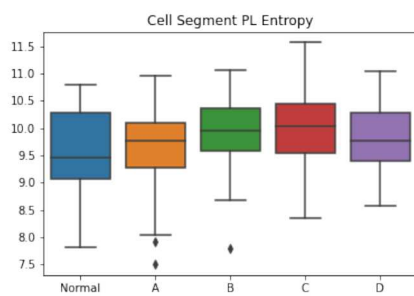
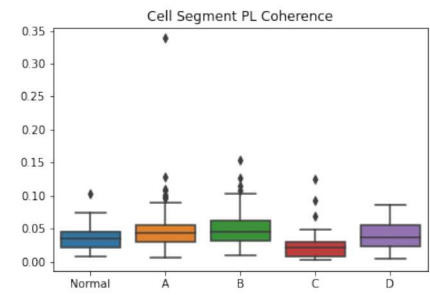
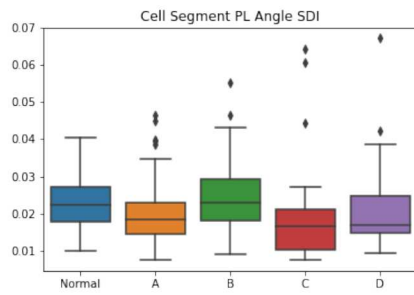
Figura 41 – Gráficos boxplots das métricas para a separação pelo padrão Gleason

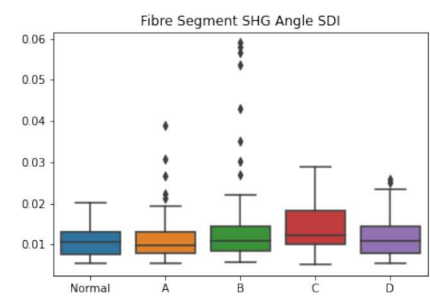
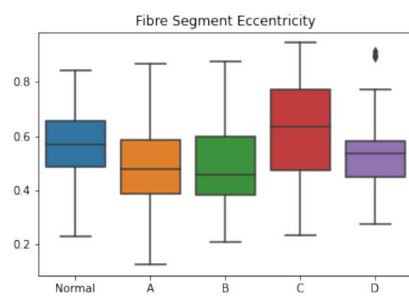
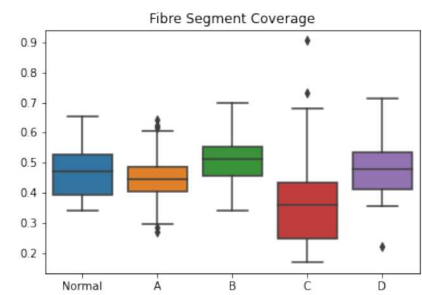
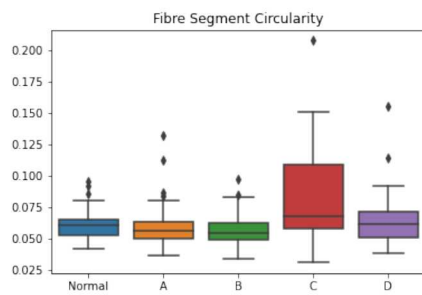
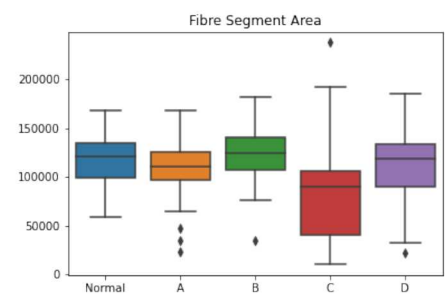
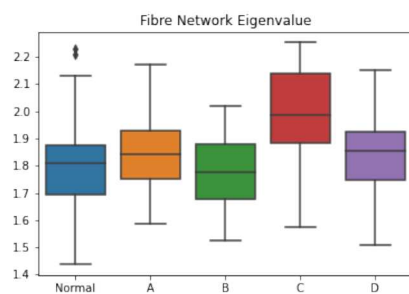
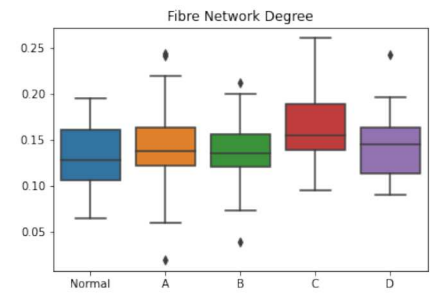
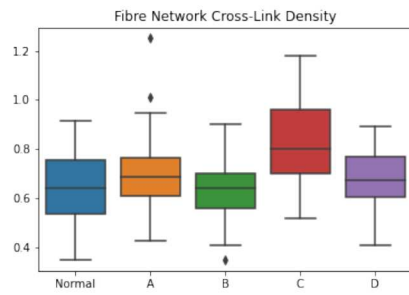
B.2 Boxplots do Padrão de Ácinos

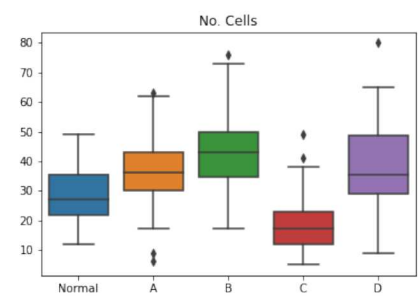
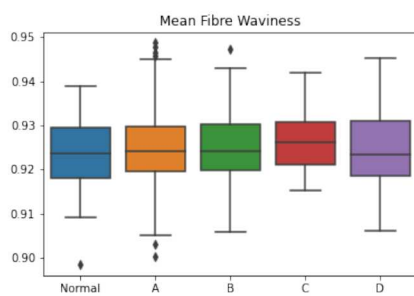
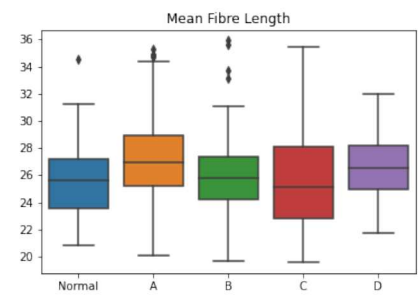
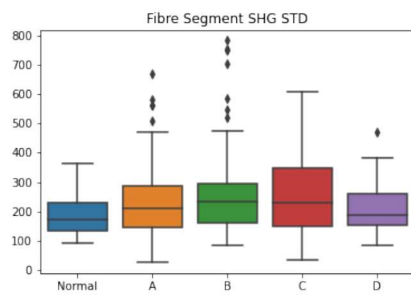
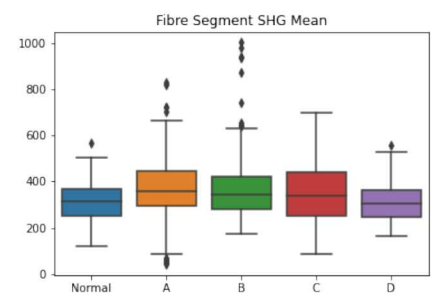
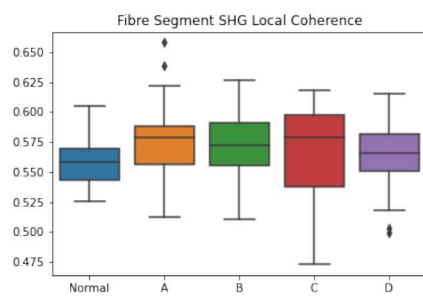
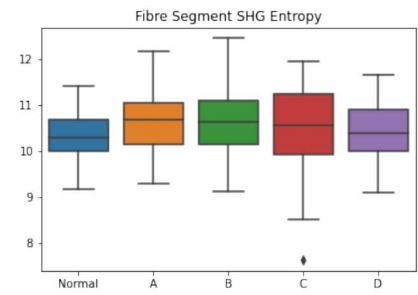
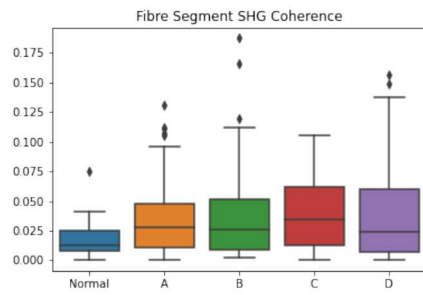
De um modo geral, os gráficos do padrão de ácinos também fornece poucas informações para classificar os diferentes tipos de carcinomas. Entretanto, aqui tem-se um pouco mais de informação sobre a região de parênquima, justamente por essa classificação focar nas diferente separações de ácinos.

Por não focar em alterações que ocorrem no estroma, as métricas relacionadas as fibras quase não fornecem informações relevantes para este caso. Como explicado no capítulo 4, o que conseguimos ver com bastante clareza está relacionado à métrica que quantifica os número de regiões de parênquima. Podemos ver que o número aumenta do Normal para o padrão B, cai de forma abrupta no padrão C e aumenta novamente no D. Isso pode ser explicado justamente pela existência de padrão cribriforme puro, que leva ao aumento da área mas faz diminuir o número desses segmentos.









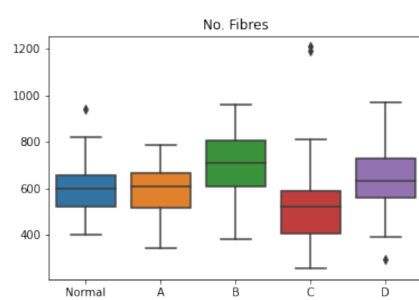


Figura 49 – Gráficos boxplots das métricas para a separação pelo tipo de Ácino

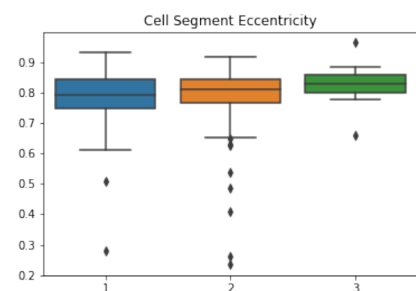
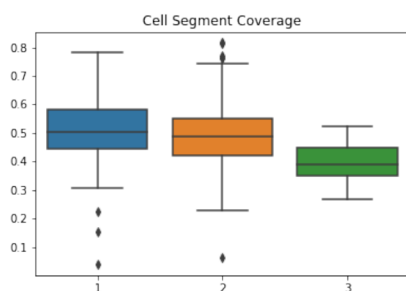
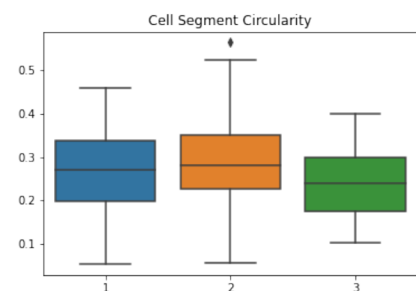
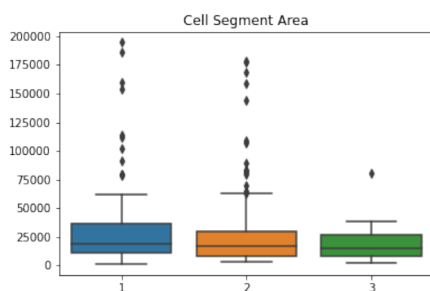
B.3 Boxplot do Padrão de Estroma

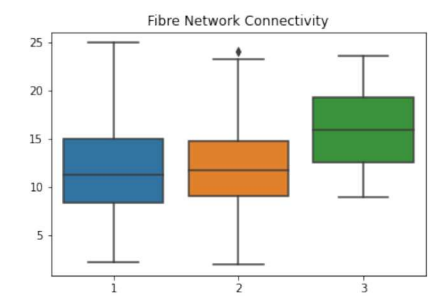
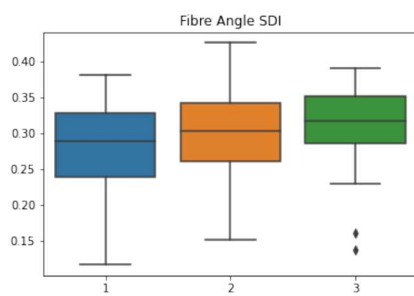
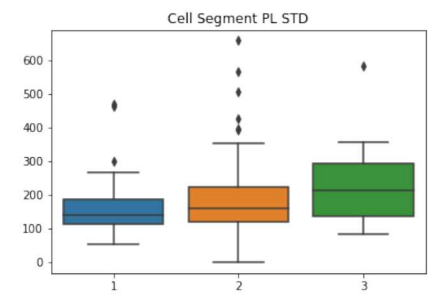
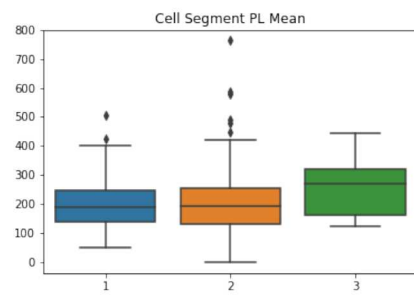
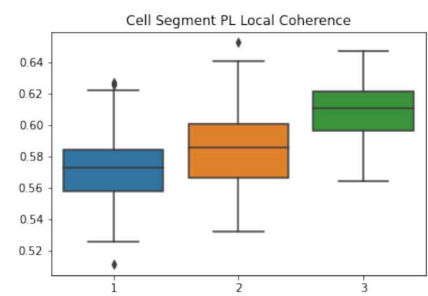
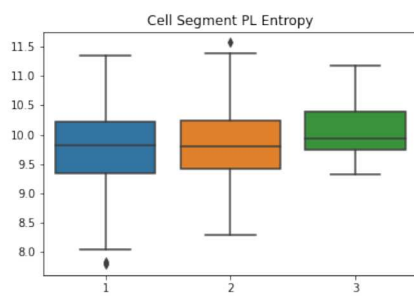
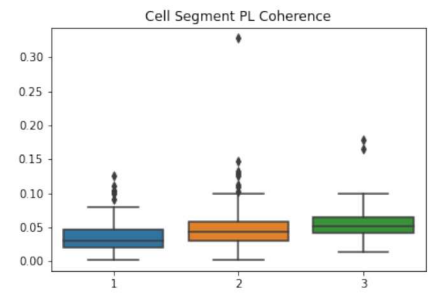
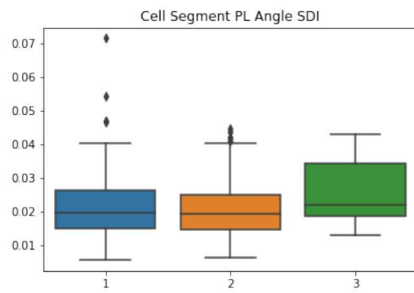
Os gráficos para a separação pelo estroma reativo, fornece informações que são interessantes à respeito das alterações que ocorrem por conta da reação estromal. Aqui ao contrário do que acontece com o padrão de ácinos estamos olhando mais para as informações contidas em relação as fibras de colágeno.

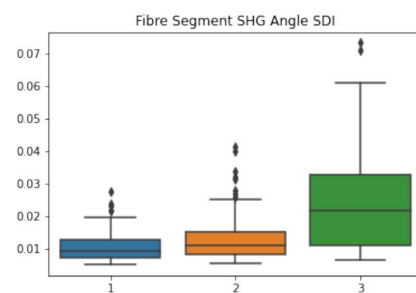
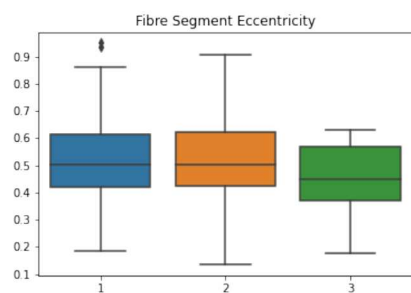
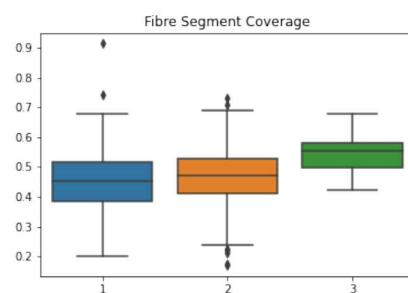
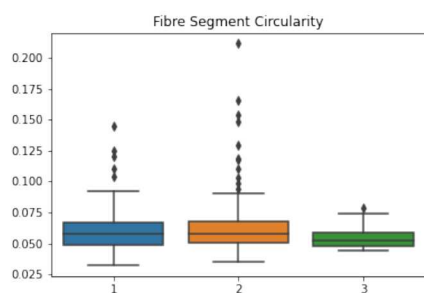
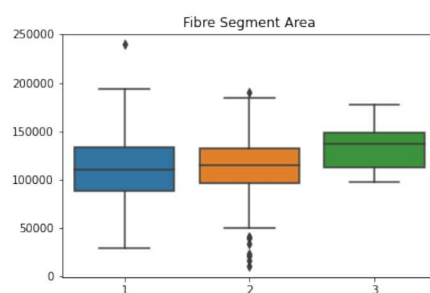
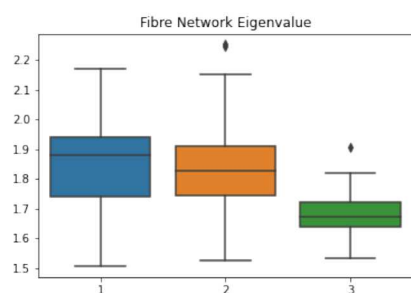
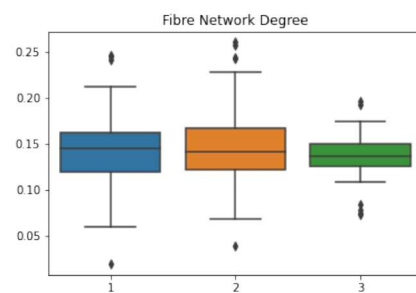
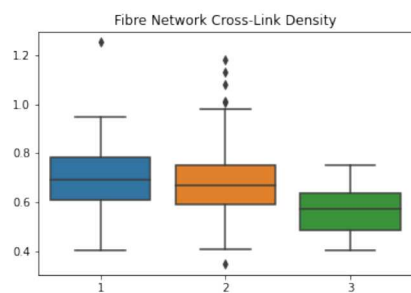
A anisotropia do SHG, tanto local quanto globalmente tem um aumento tímido mas visível e que se relaciona com o esperado de acordo com as pesquisas mais recentes. Bem como podemos ver um aumento significativo na quantidade de fibras com o crescimento da atividade tumoral.

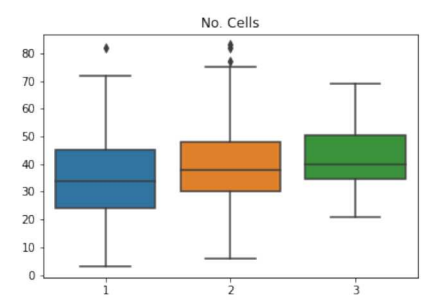
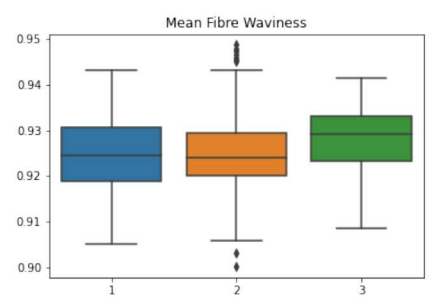
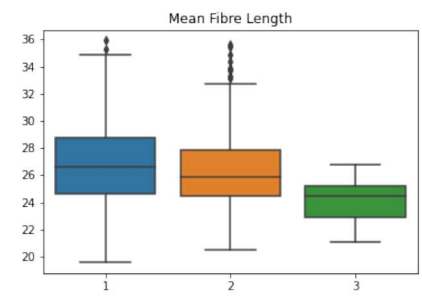
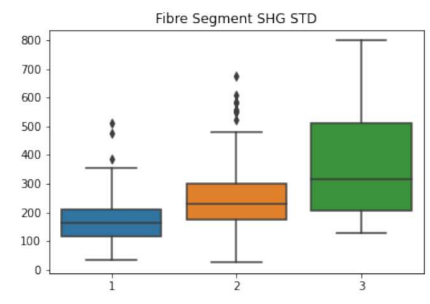
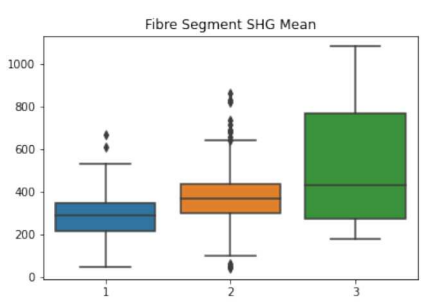
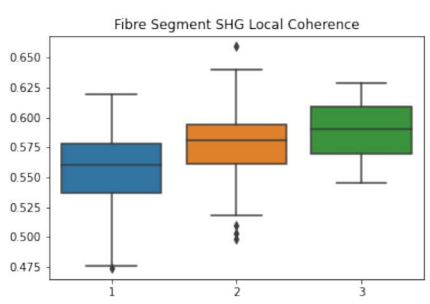
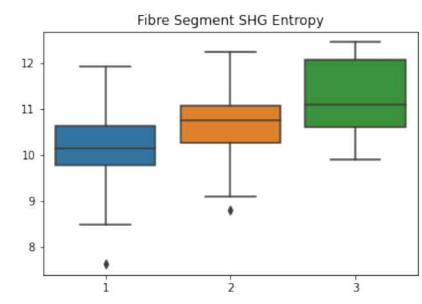
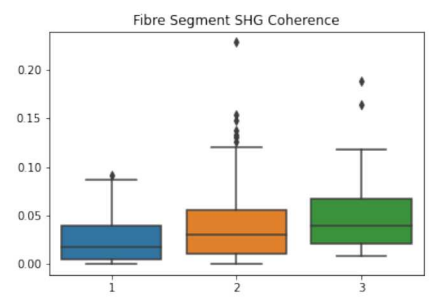
Porém, apesar de conseguirmos visualizar algumas alterações específicas como, por exemplo, o aumento do número de fibras nos estromas 2 e 3, ou na quantidade de regiões de parênquima entre os diferentes ácinos, elas não nos permitem realmente separar as classes em cada uma das separações propostas.

Devido ao fato de cada uma delas se concentrar em alterações em regiões específicas, fica claro que olhar para toda a região carcinogênica de forma completa, olhar as alterações nos ácinos, no estroma reativo, etc., contribui para um prognóstico mais eficiente.









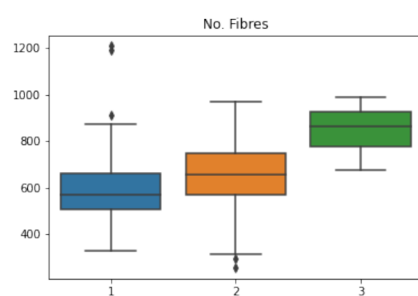


Figura 57 – Gráficos boxplots das métricas para a separação pelo tipo de Estroma.

ANEXO A – Script de Análise dos Dados

Os arquivos com a base de dados das métricas criados pelo software PyFibre, são analisados por Análise de Discriminante Linear (LDA). Utilizamos a linguagem Python para compilar esses dados e gerar os gráficos e imagens relacionados às técnicas mencionadas previamente. Abaixo está apresentado o script como utilizado para as análises.

Vale evidenciar e explicar em mais detalhes alguns dos parâmetros utilizados na função que faz o LDA. Veja as linhas 192 a 228. A função `RepeatedStratifiedKFold`, presente no pacote `Sklearn` [90] em Python, é utilizada para dividir os dados em k grupos de iguais tamanhos (caso seja possível) e isso é feito n vezes e em cada iteração é feita uma separação diferente. De forma que $k - 1$ grupos serão usados para treinar o modelo e grupo restante usado pra teste.

Para o caso apresentado no trabalho, o número de divisões é igual a 6, ou seja, os dados serão separados em 6 grupos, definida em `n_splits=6`, 5 usados para treino do modelo e 1 para teste e teremos 8 repetições, que vão implicar em 8 separações diferentes, definido em `n_repeats=8`. Além disso utilizamos o parâmetro `random_state=42`, como uma semente, de forma que os valores encontrados possam ser reproduzidos em momentos posteriores.

Para o caso da LDA, a função utilizada é a `LinearDiscriminantAnalysis`, também presente no pacote `Sklearn` [90], que será usada para a classificação dos dados. O parâmetro `solver='svd'` (svd = Singular Value Decomposition), por ser preferida em casos nos quais temos um grande número de características. Devido ao fato do svd, não computar a matriz de covariância, utilizamos o parâmetro `store_covariance=True` de forma a explicitamente computá-la.

```
# %%
import sys, os

import pandas as pd
from pandas.plotting import scatter_matrix
import seaborn as sn
import numpy as np
from scipy import linalg

# Set plotting options for the notebook
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
text_size = 14
mpl.rc('xtick', labelsizes=text_size)
```

```

mpl.rcParams['ytick', labelsizes=text_size)
mpl.rcParams.update({'font.size': text_size})

from sklearn.pipeline import Pipeline
from sklearn.model_selection import RepeatedStratifiedKFold,
    cross_val_score, train_test_split
from sklearn.preprocessing import StandardScaler, QuantileTransformer
from sklearn.decomposition import PCA
from sklearn.feature_selection import RFECV, f_classif
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.metrics import (
    auc, roc_curve, confusion_matrix, accuracy_score,
    classification_report, roc_auc_score)

# %%
from pyfibre_analysis_tools import scatter, load_databases

# %%
current_dir = os.getcwd()

# Extension options to load each type of PyFibre database
extensions = ['_global', '_fibre', '_cell', '_network']

# Enter in the names of directories here. By default these are also
    chosen as the
# grading labels for each image
directories = ['1', '2', '3']
data_directories = [
    os.path.join(current_dir, group) for group in directories
]

filename = 'pyfibre_database'
database = load_databases(filename + extensions[0], data_directories)

# %%
database_cut = database.copy()
groups = np.unique(database['Label'])

# Only select the largest segments if loading in Cell / Fibre segment
    databases
n_segments = 4
try:
    database_cut = database_cut.where(database_cut['ID'] <= n_segments
        -1)
    database_cut = database_cut.drop(['ID'], axis=1)
except:
    pass

```

```

# Remove any 'File' entries for simplicity
database_cut = database_cut.drop(['File'], axis=1)

# Remove any Nan / zero / infinite values from the database
database_cut = database_cut.replace(0, np.nan)
database_cut = database_cut.replace(np.inf, np.nan)
database_cut = database_cut.dropna()

# Map each seperate group name to the new database
database_cut['Group'] = database_cut['Label'].map(
    {float(index + 1): value for index, value in enumerate(groups)})
database_cut['Group'] = pd.Categorical(
    database_cut['Group'], categories=groups, ordered=True)

# Display an excerpt from the new database
database_cut.head()
database.head ()

# %%
columns = list(database_cut.columns)
columns.remove('Label')
columns.remove('Group')

# Extract any numerical values
df_norm = database_cut.copy()
x = df_norm[columns].values

# StandardScaler scales all values between -1 to 1
x_scaled = StandardScaler().fit_transform(x)
# QuantileTransformer enforces a Gaussian distribution of variables
x_scaled = QuantileTransformer(output_distribution='normal', n_quantiles
    =308).fit_transform(x_scaled)

# Replace the scaled numerical values back in
df_norm[columns] = x_scaled
df_norm.head()

# %%
df_norm.boxplot(column=columns, by=['Group'], figsize=(25, 25))
plt.savefig('full_metric_boxplots.png')
plt.show()

# %%
def feature_selection(database, labels, n_splits=6, n_repeats=8):
    """Retain significant features based on cross validation score of

```

```

LDA"""
104
    print("Total number of raw metrics = {}".format(database.shape[1])) 105
    X = database.values 106
    y = labels.values 107
    108
    kfold = RepeatedStratifiedKFold(n_splits=n_splits, n_repeats= 109
n_repeats)
    lda = LinearDiscriminantAnalysis(solver='svd', store_covariance=True 110
)
    rfecv = RFECV(estimator=lda, step=1, cv=kfold, scoring='f1') 111
    112
    rfecv.fit(X, y) 113
    114
    print("Optimal number of features : %d" % rfecv.n_features_) 115
    print("Features selected: ", database.loc[:, rfecv.support_].columns 116
)
    print("Features dropped: ", database.loc[:, ~rfecv.support_].columns 117
)
    118
    # Drops features with low significance 119
    features = database.loc[:, rfecv.support_] 120
    cv_results = pd.DataFrame(rfecv.cv_results_) 121
    122
    # Plot number of features VS. cross-validation scores 123
    plt.figure(0, figsize=(8,6)) 124
    plt.xlabel("Number of features selected") 125
    plt.ylabel("Cross validation score (accuracy %)") 126
    plt.plot(range(1, len(cv_results) + 1), 100 * cv_results[' 127
split1_test_score'])
    plt.savefig(f'cv_scores_lda.png') 128
    plt.show() 129
    130
    return features 131
    132
# %% 133
labels = df_norm['Label'].copy() 134
    135
df_full = df_norm.drop(['Group', 'Label'], axis=1) 136
    137
df_red = feature_selection(df_full, labels) 138
    139
# %% 140
def plot_lda_analysis(lda, training, test, columns, tick_labels, tag='') 141
:
    142
    # Plotting the training and test sets for an example LDA 143

```

```

X_plot_train = lda.transform(training[0])
X_plot_test = lda.transform(test[0])

plt.figure(0)
fig, ax, cb = scatter(X_plot_train, training[1], np.ones(training
[1].shape))
fig, ax, cb = scatter(X_plot_test, test[1], np.ones(test[1].shape),
                      ellipse=False, marker='x', alpha=0.6, fig=fig,
ax=ax, cb=cb)
cb.set_ticklabels(tick_labels)
#plt.axis('off')
plt.axis([-4, 4, -3, 3])
plt.tight_layout()
plt.savefig(f'lda_{tag}.png')
plt.show()

# Plot LDA coefficients
fig = plt.figure(figsize=(10, 8))
ax = plt.gca()

im = ax.matshow(lda.coef_, interpolation='none', cmap='coolwarm')
fig.colorbar(im)

ax.set_xticks(np.arange(len(columns)))
ax.set_xticklabels(list(columns))
ax.set_yticks(np.arange(len(tick_labels)))
ax.set_yticklabels(tick_labels)

plt.setp([tick.label2 for tick in ax.xaxis.get_major_ticks()],
rotation=45,
         ha="left", va="center", rotation_mode="anchor")
plt.show()

# Plotting metric covariance
fig = plt.figure(figsize=(10, 8))
ax = plt.gca()

im = ax.matshow(lda.covariance_, interpolation='none', cmap='
coolwarm')
fig.colorbar(im)

ax.set_xticks(np.arange(len(columns)))
ax.set_xticklabels(list(columns))
ax.set_yticks(np.arange(len(columns)))
ax.set_yticklabels(list(columns))

```

```

plt.setp([tick.label2 for tick in ax.xaxis.get_major_ticks()], 187
rotation=45,
            ha="left", va="center",rotation_mode="anchor") 188
plt.show() 189
190
191
def lda_analysis(database, labels, tick_labels=None, n_splits=6, 192
n_repeats=8, tag=''): 193
194
195
196
X = database.values 194
y = labels.values 195
196
197
kfold = RepeatedStratifiedKFold(n_splits=n_splits, n_repeats= 197
n_repeats, random_state=42)
lda = LinearDiscriminantAnalysis(solver='svd', store_covariance=True 198
) 199
200
201
202
scores = [] 200
binary_scores = [] 201
202
203
for train_index, test_index in kfold.split(X, y): 203
204
205
206
X_train, y_train = X[train_index], y[train_index] 204
X_test, y_test = X[test_index], y[test_index] 205
206
207
lda.fit(X_train, y_train) 207
y_predict = lda.predict(X_test) 208
209
210
211
print("Average cross validation score for all classes: ", np.mean( 211
scores), "+=", np.std(scores))
print("Average cross validation score for Normal vs other classes: " 212
, np.mean(binary_scores), "+=", np.std(binary_scores))
213
214
215
216
217
218
plot_lda_analysis( 213
lda, (X_train, y_train), (X_test, y_test), 214
columns=database.columns, 215
tick_labels=tick_labels, 216
tag=tag 217
) 218
219
220
# %% 220
221
222
# Change this if you want to use different labels for the figures (i.e. 222
use English spellings)
tick_labels = ['Stroma 1', 'Stroma 2', 'Stroma 3'] 223
print('Using full set of PyFibre Metrics') 224
lda_analysis(df_full, labels, tick_labels=tick_labels, tag='full') 225
226

```

```
print('Using reduced set of PyFibre Metrics')
```

227

```
lda_analysis(df_red, labels, tick_labels=tick_labels, tag='red')
```

228

E por fim apresento o código usado para obtenção dos gráficos de caixa, dos quais também foram extraídos os valores p , presentes nas tabelas mostradas na seção de Resultados e Discussões 4.

```

import pandas as pd
import seaborn as sb
import numpy as np
import os, sys
import matplotlib.pyplot as plt
from scipy import stats
curr_dir = os.getcwd()
dir_raiz = os.path.join(curr_dir+'/databases')
df_bases = []
labels = ['1', '2', '3']
label_diretorios = [
    os.path.join(curr_dir, group) for group in labels
]
#print(label_diretorios)
for i, diretorio in enumerate(label_diretorios):
    label = os.path.split(diretorio)[-1]
    #print(label)

for dirpath, dirnames, filenames in os.walk(dir_raiz):

    for filename in filenames:
        if filename == 'pyfibre_database_global.h5':
            file_path = os.path.join(dirpath, filename)
            dirname = os.path.basename(dirpath)
            # Le os arquivos h5 como dataframe
            df = pd.read_hdf(file_path)
            df['Label'] = dirname
            # Une os dfs em um unico df
            df_bases.append(df)

df_tipo = pd.concat(df_bases, ignore_index=True)
df_tipo['Label_2'] = df_tipo['Label'].map(lambda x: labels.index(x))

df_tipo = df_tipo.sort_values('Label_2')

df_tipo = df_tipo.drop('Label_2', axis = 1)

df_tipo

columns = list(df_tipo.columns)
columns.remove('Label')
columns.remove('File')
#xtickpositions = [0,1,2,3,4]
```

```

for col in columns:
    ax = df_tipo.boxplot(column=col,by=['Label'], grid=False)
    ax.set_xlabel('')
    ax.set_xticklabels(labels)
    plt.suptitle('')
    plt.savefig(f'estroma_boxplot_{col}.png')
    plt.close()
df_tipo = df_tipo.drop('File',axis =1)
label = ['1','2','3']
label
grouped = df_tipo.groupby('Label')

# Iterate over each group
for label, group in grouped:
    group_label = f"df_{label}"
    globals()[group_label] = group.copy()
df_name = [df_1,df_2,df_3]
with open('p_values.txt', 'w') as file:
    file.write('')
p_values = open('p_values.txt', 'w')

for k in range(0,len(df_name)):

    for n in range(k+1, len(df_name)):
        for col1 in df_name[k].columns:
            for col2 in df_name[n].columns:
                if (col1 == col2) and (col1 != 'Label') and (col2 != '
Label'):
                    print(df_name[k]['Label'].iloc[0] + 'X' + df_name[n
] ['Label'].iloc[0])

                    p_values.write(df_name[k]['Label'].iloc[0] + 'X' +
df_name[n]['Label'].iloc[0]+ '\n')

                    fscore, pvalor = stats.f_oneway(df_name[k][col1],
df_name[n][col2])
                    print(f'{col1} F-score: {fscore} pvalor: {pvalor}\n'
)
                    p_values.write(f'{col1} F-score: {fscore} pvalor: {
pvalor}\n')
                    p_values.write('
-----\n')

                else:
                    pass

```

```
        p_values.write('
*****\n')
p_values.close()
```