



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA



ESPECTROSCOPIA DE INFRAVERMELHO COM TRANSFORMADA DE
FOURIER COM REFLETÂNCIA TOTAL ATENUADA (ATR-FTIR) E
QUIMIOMETRIA COMO ADJUVANTES NO DIAGNÓSTICO DE CÂNCER DE
TIREOIDE

Ingrid Gabriela Bezerra de Lima Cruz





UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA



ESPECTROSCOPIA DE INFRAVERMELHO COM TRANSFORMADA DE
FOURIER COM REFLETÂNCIA TOTAL ATENUADA (ATR-FTIR) E
QUIMIOMETRIA COMO ADJUVANTES NO DIAGNÓSTICO DE CÂNCER DE
TIREOIDE

Ingrid Gabriela Bezerra de Lima Cruz*

Versão preliminar da Dissertação
apresentada como parte dos requisitos
para obtenção do título de Mestre em
Química pela Universidade Federal da

Orientador: Prof. Dr. Sherlan Guimarães Lemos

Coorientador: Prof. Dr. Wallace Duarte Fragoso

*Bolsista Capes

Catálogo na publicação
Seção de Catalogação e Classificação

C957e Cruz, Ingrid Gabriela Bezerra de Lima.

Espectroscopia de infravermelho com transformada de Fourier com refletância total atenuada (ATR-FTIR) e quimiometria como adjuvantes no diagnóstico de câncer de tireoide / Ingrid Gabriela Bezerra de Lima Cruz. - João Pessoa, 2022.

71 f. : il.

Orientação: Sherlan Guimarães Lemos.

Coorientação: Wallace Duarte Fragoso.

Dissertação (Mestrado) - UFPB/CCEN.

1. Glândula tireóide - Câncer. 2. Espectroscopia de infravermelho. 3. DD-SIMCA. 4. Classificação patológica. 5. SIMCA. I. Lemos, Sherlan Guimarães. II. Fragoso, Wallace Duarte. III. Título.

UFPB/BC

CDU 616.441(043)

Espectroscopia de infravermelho com transformada de Fourier com reflectância total atenuada (ATR-FTIR) e quimiometria como adjuvantes no diagnóstico de câncer de tireoide.

Dissertação de Mestrado apresentada pela aluna Ingrid Gabriela Bezerra de Lima Cruz e aprovada pela banca examinadora em 20 de maio de 2022.



Prof. Dr. Sherlan Guimarães Lemos
Departamento de Química – CCEN/UFPB
Orientador/Presidente



Prof. Dr. Wallace Duarte Fraagoso
Departamento de Química – CCEN/UFPB
2º. Orientador



Prof. Dr. Fabyan Esberard de Lima Beltrão
HU/UFPB-JPA-PB
Examinador Externo



Prof. Dr. Márcio José Coelho de Pontes
Departamento de Química – CCEN/UFPB
Examinador Interno

Assinaturas da Banca realizadas em modo Webconferência em 20/02/2022, digitalizadas e certificadas pelo Prof. Dr. Sherlan Guimarães Lemos (SIAPE 1650955) em 20/05/2022.

Dedicatória

*Dedico este trabalho de mestrado à Neves
minha mãe e ao meu esposo Luiz Eduardo.*

AGRADECIMENTOS

- Grande é o Senhor e mui digno de louvor. Agradeço a Ele por todos os seus feitos em minha vida, pela sabedoria e discernimento concedido por Ele. Confio no seu infinito AMOR, só Ele é o Deus eterno, sobre toda a terra e céus.
- Agradeço a minha mãe por todo esforço e dedicação em meu desenvolvimento físico, emocional, educacional. Por ser uma mulher guerreira que me inspira a cada dia ser melhor. Pelo enorme incentivo para que eu viesse a alcançar todos os meus sonhos.
- Agradeço ao meu esposo Eduardo, meu grande amor. Sempre me incentivando e mostrando que sou melhor. Alegria de poder compartilhar com você esses anos de namoro, noivado e agora casados. Como é bom ter você por perto, me apoiando e dizendo que tudo vai ficar bem, porque o Senhor não perde o controle sob nossas vidas.
- Agradeço ao meu companheiro de quatro patas, Duck pelo amor compartilhado durante esses anos. Quem conhece a história sabe o quanto ele sofreu até chegar em nosso lar.
- Agradeço a todos os meus familiares: minha avó materna Alzira e avó paterna Nalva. Meus tios e tias, Cristiane, Edson, Ezione, Francisco, Gilmar, Risonete, Severino, Severina, Sonia, Suely, Wilson. Aos meus primos, Elisângela, Geilza, Henrique, Jeilton, Jonathas, John Lucas, Lucas, Lyvia., aos meus cunhados Luiz Henrique, Willian Neto, Anna Luíza.
- Agradeço ao meu orientador Prof. Dr. Sherlan Guimarães Lemos que tive o prazer de conhecer durante o ensino médio através de um projeto de pesquisa que pude participar, foi um divisor de águas na escolha do curso de graduação. Depois por aceitar como aluna no grupo de pesquisa durante a graduação e agora no mestrado. Obrigada por compartilhar comigo um pouco do seu conhecimento. Obrigada pelos puxões de orelha nos momentos necessários, pela enorme paciência, incentivo e compreensão durante todos esses anos, assim ajudando em minha formação.
- Agradeço a todos meus amigos, por sempre acreditar no meu potencial. Alanny Firmino, Alex Mendonça, Amanda Letícia, Ângela Barros, Anne Karoline, Bruno Péricles, Carla Conceição, Isadora Vieira, Kell, Jefferson, João Gomes, Lauro,

Márcio Carvalho, Paula Braz, Rayana Eulina, Rosângela Conceição, Rossana Oliveira, Ruth Diniz, Suelly Fernandes.

- Agradeço a todos que fazem parte do GEAQA (*Grupo de Estudos Avançados em Química Analítica*), que foram de grande importância nessa jornada acadêmica. Alcides, Anne, Carol, Dayvison, Flávia, Jefferson, Josy, Kelvin, Levi, Matheus, Prof. Dr. Wallace.
- Agradeço ao Prof. Dr. Fabyan pela proposta do projeto e disponibilidade das amostras utilizadas neste estudo de mestrado.
- Agradeço a Laís Chantelle (LACOM), que realizou todas as análises ATR-FTIR das lâminas em seu laboratório. Sem essas análises, este trabalho estaria impossibilitado de ser concluído.
- Agradeço todas as contribuições e conselhos para melhoria deste trabalho dos avaliadores da banca de defesa de mestrado Prof. Dr. Ricardo Alexandre, Prof. Dr. Márcio Coelho, Prof. Dr. Fabyan Esberard, Prof. Dr. Sófacles.
- Agradeço aos professores da Pós-graduação em Química – UFPB, pela contribuição em minha formação.
- Agradeço a Universidade Federal da Paraíba.
- Agradeço a fundação vinculada ao Ministério da Educação do Brasil, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, pela bolsa concedida.

Resumo

CRUZ, I. G. B. L. Espectroscopia de Infravermelho com Transformada de Fourier com Refletância Total Atenuada (ATR-FTIR) e Quimiometria como adjuvantes no diagnóstico de câncer de tireoide. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa, 2022.

O câncer de tireoide é a neoplasia maligna mais comum acometidos na região da cabeça e pescoço nos últimos anos. De acordo com projeções norte-americanas, será o quarto tipo câncer mais frequente na sociedade em 2030. A carência de profissionais especializados na distinção dos tecidos tiroidianos normais e cancerígenos em regiões mais distantes ou mais pobres dificulta a oferta do diagnóstico. Diante disto, a oferta de um método de diagnóstico ou triagem que pode ser executado por um profissional de maior disponibilidade pode reduzir o impacto deletério que essa carência causa. Nesse sentido, a espectroscopia de infravermelho em conjunção com técnicas quimiométricas pode ser uma alternativa viável para o desenvolvimento desse tipo de método de diagnóstico. Assim, este estudo teve como objetivo desenvolver um método empregando a espectroscopia no infravermelho com transformada de Fourier com reflexão total atenuada (ATR-FTIR) em conjunto com técnicas quimiométricas de classificação, a Modelagem Independente Flexível por Analogia de Classes (SIMCA) e Modelagem Independente Flexível por Analogia de Classes Direcionado pelos Dados (Data Driven SIMCA, DD-SIMCA), para auxiliar no diagnóstico de câncer de tireoide. O estudo empregou amostras de punção aspirativa por agulha fina (PAAF) de 49 pacientes (n = 49), sendo 27 amostras diagnosticadas como benignas, 11 amostras diagnosticadas como câncer de tireoide e 11 amostras diagnosticadas como indefinidas. Com os modelos desenvolvidos, foi possível avaliar a predição de 11 amostras indefinidas, para prever a similaridade da classificação patológica com o modelo proposto, foram obtidos valores de SENS global 61% e ESPEC global 89% para o conjunto de teste e indefinidas. A metodologia proposta representa uma ferramenta promissora no diagnóstico de câncer de tireoide, sendo possível concluir que a espectroscopia e quimiometria são capazes de distinguir tecidos tiroidianos saudáveis de tecidos tiroidianos com células cancerígenas.

Palavras-chave: Câncer de tireoide, espectroscopia de infravermelho, DD-SIMCA, classificação, SIMCA.

Abstract

CRUZ, I. G. B. L. **Attenuated Total Reflectance Fourier Transform Infrared Spectroscopy (ATR-FTIR) and Chemometrics as adjuvants in the diagnosis of thyroid cancer.** Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa, 2022.

Thyroid cancer is the most prevalent malignant neoplasm affected in the head and neck region in recent years. According to US projections, it will be the fourth most frequent type cancer in society by 2030. The lack of professionals specialized in distinguishing normal and cancerous thyroid tissues in more remote or poorer regions makes it difficult to offer diagnosis. Given this, offering a diagnostic or screening method that can be performed by a more readily available professional can reduce the deleterious impact that this shortage causes. In this sense, infrared spectroscopy in conjunction with chemometrics may be a viable alternative for the development of such a diagnostic method. Thus, this study aimed to develop a method employing attenuated total reflection Fourier transform infrared spectroscopy (ATR-FTIR) in conjunction with chemometric classification techniques, the soft independent modelling by class analogy (SIMCA) and data driven SIMCA (DD-SIMCA), to aid in the diagnosis of thyroid cancer. The study employed fine needle aspiration (FNA) samples from 49 patients ($n = 38$), with 27 samples diagnosed as benign, 11 samples diagnosed as thyroid cancer and 11 samples diagnosed as undefined. With the models developed, it was possible to evaluate the prediction of 11 undefined samples, to predict the similarity of the pathological classification with the proposed model, values of global SENS 61% and global ESPEC 89% were obtained for the test set and undefined. The proposed methodology represents a promising tool in the diagnosis of thyroid cancer, being possible to conclude that spectroscopy and chemometrics can distinguish between healthy and cancerous thyroid tissues.

Keywords: Thyroid cancer, infrared spectroscopy, DD-SIMCA, classification, SIMCA.

Lista de Figuras

Figura 1 – Ilustração anatômica da glândula tireoide. Fonte: “human thyroid gland anatomy illustration”, por purple carrot nutrition, licenciado por Creative Commons Attribution-Share Alike 4.0 International.....	21
Figura 2 - Matriz de dados a partir de espectro. as informações espectrais são organizadas em matriz de dados.	28
Figura 3 - Gráfico da área de aceitação com amostras regulares (verde) e amostra extrema (laranja).....	33
Figura 4 - Conjunto de treinamento, teste e indefinidas para os modelos SIMCA e DD-SIMCA.	40
Figura 5 - Espectros ATR-FTIR lâminas de pacientes com câncer de tireoide e pacientes saudáveis, na região de 4000 a 600 cm^{-1} sem pré-processamento.....	44
Figura 6 - Média dos Espectros ATR-FTIR para pacientes saudáveis (—) sobreposta a média dos espectros de pacientes doentes (—).	45
Figura 7 - Espectros com pré-processamentos. (a) Suavização Savitzky-Golay, (b) Suavização + MSC	46
Figura 8 - Gráficos dos escores referentes aos 38 pacientes da PC1 x PC2. Espectros 4000 – 600 cm^{-1} (a) sem pré-processamento; (b) Suavização (Savitzky-Golay) (c) Suavização (Savitzky-Golay) + MSC. Amostras de pacientes saudáveis (●) e amostras pacientes doentes (●)	47
Figura 9 - Gráfico de pesos para PC1, PC2, PC3 e PC4, obtendo uma variância explicada acima de 90%, para os espectros sem pré-processamento.	48
Figura 10 - Gráfico de área de aceitação para o conjunto de treinamento (a, b, c) e teste (d, e, f) para o algoritmo de classificação DD-SIMCA desenvolvidos no ATR-FTIR 4000-600 cm^{-1} (espectros inteiros). (a, d) espectro bruto; (b, e) espectro com suavização; (c, f) espectro com suavização mais MSC. (●) amostra regular. As linhas verdes representa a área de aceitação e as linhas vermelhas é a fronteira da área de aceitação.	51
Figura 11 - Gráficos dos escores referentes aos espectros (1400 – 650 cm^{-1}) dos 38 pacientes, PC1 x PC2. (a) sem pré-processamento (b) com suavização Savitzky-Golay; (c) com suavização Savitzky-Golay + MSC. Amostras de tecido saudáveis (●) e amostras de tecido com câncer de tireoide (●)	54

Lista de Tabelas

Tabela 1 - Resumo dos estudos de pesquisa da espectroscopia e Quimiometria na caracterização de amostras de tecidos tireoidiano saudáveis e doentes, destacando os tipos de amostras, análises e condições.....	26
Tabela 2 - Característica clínica dos pacientes.	38
Tabela 3 - Característica clínica dos pacientes que apresentam diagnóstico indefinido.	41
Tabela 4 - Resultados dos valores da matriz de confusão para o modelo SIMCA nos conjuntos de treinamento (20 amostras) e teste (20 amostras) na região entre 4000 e 600 cm^{-1}	49
Tabela 5 - Valores de predição para amostras indefinidas, utilizando os modelos SIMCA para os espectros bruto e suavizado (Savitzky-Golay).	50
Tabela 6 - Resultados dos parâmetros de desempenho sensibilidade, especificidade e acurácia para o modelo DD-SIMCA nos conjuntos de treinamento e teste na região entre 4000 e 600 cm^{-1}	52
Tabela 7 - Valores de predição para amostras indefinidas, utilizando os modelos DD-SIMCA para os espectros bruto e suavizado (Savitzky-Golay)	53
Tabela 8 - Resultados da matriz de confusão para os modelos SIMCA e DD-SIMCA na região entre 1400 e 650 cm^{-1} , para os espectros bruto, com suavização Savitzky-Golay e com suavização Savitzky-Golay + MSC	55
Tabela 9 - Resultado de predição das amostras indefinidas, para o modelo DD-SIMCA para os dados brutos e os espectros suavizados Savitzky-Golay	57

Lista de abreviatura e siglas

AC – Acurácia

ATR-FTIR – Espectroscopia no infravermelho com transformada de Fourier com Reflexão Total Atenuada

CAT – Carcinoma Anaplásico de tireoide

CDA – Análise Discriminante Canônica

CEP – Comitê de Ética em Pesquisa

CFM – Conselho Federal de Medicina

CFT – Carcinoma Folicular de Tireoide

CMT – Carcinoma Medular de Tireoide

CPT – Carcinoma Papilífero de Tireoide

CREMESP – Conselho Regional de Medicina do Estado de São Paulo

DD-SIMCA – Modelagem Independente Flexível por Analogia de Classes Direcionado pelos Dados

ESPEC – Especificidade

FTIR – Espectroscopia de infravermelho com transformada de Fourier

FT-MIR – Espectroscopia de infravermelho médio com transformada de Fourier

GA – Algoritmo Genético

H&E – Hematoxilina & Eosina

HULW – Hospital Universitário Lauro Wanderley

KS – Kennard – Stone

LDA – Análise Discriminante Linear

MSC – Correção do Espalhamento Multiplicativo

NIH – National Institute of Health

PAAF – Punção Aspirativa por Agulha Fina

PCA – Análise de Componentes Principais

PLS – Mínimos Quadrados Parciais

QDA – Análise Discriminante Quadrática

RP – Reconhecimento de Padrões

SENS – Sensibilidade

SIMCA – Modelagem Independente Flexível por Analogia de Classes

USG – Ultrassonografia

Sumário

INTRODUÇÃO	14
OBJETIVOS.....	19
Objetivo Geral	19
Objetivo Específicos	19
FUNDAMENTAÇÃO TEÓRICA	21
Glândula tireoide.....	21
Câncer de tireoide	22
Diagnóstico de câncer de tireoide.....	23
Espectroscopia e Quimiometria como alternativa de diagnóstico de doenças.....	24
Quimiometria.....	27
PCA.....	29
SIMCA	30
DD-SIMCA	31
Validação dos modelos de classificação	34
Figuras de mérito	34
METODOLOGIA	37
- Aquisição das amostras PAAF	37
- Equipamento e materiais utilizados	37
- Tratamento de dados: Organização dos dados e algoritmo de tratamento	38
- Pré-processamentos	41
RESULTADOS E DISCUSSÃO	44
CONCLUSÕES	60
PESPECTIVAS	61
REFERÊNCIAS	62

Capítulo 1

INTRODUÇÃO

INTRODUÇÃO

O câncer de tireoide é a neoplasia maligna mais frequente na região da cabeça e pescoço observados nas últimas décadas. Este tumor acomete predominantemente mulheres, na maioria dos casos com idade superior a 30 anos (WARD et al., 2008; INCA, 2020; BORGES et al., 2020). Fatores de risco a saúde consolidados na literatura para o câncer de tireoide é a exposição à radiação ionizante, histórico familiar e ausência de iodo na dieta, porém é possível que outros fatores desconhecidos possam ser carcinogênicos (INCA, 2020; BORGES et al., 2020).

Os tipos histológicos de câncer na glândula tireoide mais frequentes são o carcinoma papilífero (CPT), carcinoma folicular (CFT), carcinoma medular (CMT) e carcinoma anaplásico (CAT), sendo este último o mais raro dentre eles e o CPT o de maior predominância (ENEWOLD et al., 2009; INCA, 2020; BORGES et al., 2020).

Segundo Rahib e colaboradores, o câncer de tireoide será o quarto tipo de tumor mais frequente nos Estados Unidos em 2030 e possivelmente com padrão similar em outros países (RAHIB et al., 2014). Porém, avanços na triagem e no tratamento precoce e preventivo podem alterar essa projeção nas taxas de incidências e/ou mortalidade por câncer de tireoide, mas isso exigirá um esforço da comunidade científica.

Atualmente o principal exame utilizado para diagnóstico de câncer de tireoide é a punção aspirativa por agulha fina (PAAF), seguido pela PAAF guiada por ultrassonografia (USG) de alta definição. O método PAAF convencional é pouco invasivo e de baixo custo. Porém, mesmo sendo considerado método padrão para todas as patologias, ainda apresenta limitações no diagnóstico, principalmente na associação de tumores malignos a nódulos benignos, resultando em falsos negativos (YOKOZAWA, 1998; PECCIN et al., 2003). Observa-se que apesar de todo desenvolvimento atual em termos de tecnologia e diagnóstico clínico, há uma carência de métodos de diagnósticos para tumores, a qual dificulta o tratamento precoce dos tumores malignos (GHAZY et al., 2020).

Outro problema associado é a obtenção do material aspirado adequado para análise e a utilização da USG. Tal análise depende minimamente da experiência do operador, necessitando de estudo e treinamento para a prática clínica, como um profissional médico

com especialização em patologia (YOKOZAWA, 1998; PECCIN et al., 2003). Segundo pesquisas do Conselho Federal de Medicina (CFM) e do Conselho Regional do Estado de São Paulo (CREMESP), em hospitais e clínicas no Brasil há uma carência de médicos patologistas. Assim como um declínio ou estagnação do número de novos residentes ao longo do tempo. As Regiões Norte e Centro-Oeste são as que apresentam menor número de médicos patologistas (DBM, 2020; CREMESP, 2014), tendo como consequência o retardo dos resultados de exames patológicos.

No que concerne ao método padrão de diagnóstico de câncer de tireoide, uma etapa importante do processo de diagnóstico é a classificação das lâminas de amostras retiradas de tecidos dos pacientes por PAAF. A classificação das lâminas PAAF emprega a escala Bethesda, baseada em seis níveis ou categorias de risco: inconclusivo (I); benigno (II); lesão folicular de significado indeterminado (III); neoplasia folicular (IV); suspeita de malignidade (V); maligno (VI) (LAYFIELD et al., 2010; LAYFIELD et al., 2008).

Essa classificação também nos dá informação sobre o risco de malignidade (%) nos nódulos caracterizados dentro desse padrão. Para o inconclusivo, não há célula suficiente para classificação e normalmente se repete a PAAF com USG. Para o segundo nível, benigno, o risco de malignidade é de 0-3% e normalmente o paciente recebe acompanhamento clínico. Para o terceiro nível, lesão folicular de significado indeterminado, o risco é entre 5-15%, repetindo-se a PAAF. Para o quarto nível, neoplasia folicular, o risco é de 15-30% e o lobo contendo células cancerígenas é retirado num procedimento chamado de lobectomia cirúrgica. Para o nível cinco, suspeita de malignidade, o risco está entre 60-75%, podendo ser indicado a lobectomia cirúrgica ou tireoidectomia quase total, que é a remoção quase total da glândula tireoide. Para o sexto nível, malignidade, o risco é de 97-99%, sendo recomendado a tireoidectomia quase total (VILLELA, 2017; LAYFIELD et al., 2010; LAYFIELD et al., 2008).

No cenário de diagnóstico clínico, na identificação de várias doenças podemos observar que são escassos testes econômicos, não invasivos e de boa acurácia (PARASKEVAIDI et al., 2018). Por exemplo, 25% dos diagnósticos recentes de tumores cerebrais ocorrem quando a doença está em fase de metástase, pois requerem um longo tempo para obtenção dos resultados clínicos (BURY et al., 2019). Diagnósticos como na doença de Alzheimer são utilizadas técnicas estruturais de neuroimagens ou baterias de

testes neuropsicológicos que são invasivos, demorados e de um custo financeiro elevado (PARASKEVAIDI et al., 2017).

Nesse contexto, a espectroscopia de infravermelho (IR) surge como uma alternativa promissora no diagnóstico de doenças, por ser uma técnica não invasiva, não destrutiva, que permite identificar a estrutura química e características químicas de uma determinada amostra biológica, diferenças moleculares em tecidos patológicos e saudáveis (LI et al., 2021; VILLELA, 2017; BAKER et al., 2014) como por exemplo, doença de Alzheimer (PARASKEVAIDI et al., 2017); câncer de ovário (LI et al., 2021); diabetes mellitus (CAIXETA, 2018); câncer de próstata (SIQUEIRA et al., 2017).

Novas alternativas no desenvolvimento de métodos para o diagnóstico de câncer de tireoide têm surgido. Segundo o Instituto Nacional de Câncer do Instituto Nacional de Saúde (NIH) norte-americano, as glândulas linfáticas apresentam grande importância no prognóstico do câncer de tireoide e eficiência no tratamento precoce do paciente (NIH, 2022). Liu e colaboradores desenvolveram uma análise preditiva do acometimento de linfonodos cervicais metastáticos e seu estudo consistiu em uma análise bioquímica dos tecidos pela técnica de espectroscopia de infravermelho com transformada de Fourier (FTIR). Por meio dos espectros de absorção diferentes tipos de histológicos são separados. Os espectros de absorção refletem as características químicas e vibracionais de grupos funcionais existentes na amostra alvo (LIU et al., 2011).

As ferramentas quimiométricas são constantemente explorados em combinação com técnicas espectroscópicas, pois conseguem gerar métodos rápidos, baratos e que permitem classificar amostras de interesse (alvo) ou diferenciá-las de outras amostras. Entre os exemplos de doenças com métodos de diagnóstico baseados no uso de técnicas espectroscópicas em associação com a quimiometria estão a asma (ULLAH et al., 2019); a disfunção tireoidiana e a insuficiência renal (WANG et al., 2021). Nesses casos foi utilizada a espectroscopia Raman e métodos quimiométricos. Doenças infecciosas como dengue, chikungunya e zika (SANTOS et al., 2018), além da fibromialgia (PASSOS et al., 2020), também tem métodos desenvolvidos empregando FTIR e quimiometria.

Para o diagnóstico de câncer de tireoide a literatura especializada registra um método que avaliou o uso das técnicas FTIR e Raman combinadas com análise multivariada para distinção entre adenomas de tireoide e carcinomas (DEPCIUCH et al.,

2019). Aplicando análise de componentes principais e análise discriminante linear (PCA-LDA), os autores obtiveram uma sensibilidade de 62% e especificidade de 70% usando espectroscopia Raman, e sensibilidade de 65% e especificidade de 75% para espectroscopia FTIR (DEPCIUCH et al., 2019). O objetivo desse estudo foi distinguir os tecidos tireoidianos normais dos tecidos tireoidianos cancerígenos, utilizando as duas técnicas espectroscópicas e quimiometria, investigando qual técnica espectroscópica era a melhor para detectar esses tecidos. O estudo empregou 17 pacientes com câncer de tireoide e 15 pacientes saudáveis e FTIR com resolução espectral de 2 cm^{-1} .

Assim, em vista da possibilidade de discriminação de tecidos tireoidianos normais dos tecidos tireoidianos cancerígenos por FTIR e quimiometria, neste estudo propõe-se explorar a técnica de espectroscopia de infravermelho com transformada de Fourier com acessório de reflectância atenuada (ATR-FTIR) e quimiometria para a criação de modelos de classificação de classe única focando na modelagem de amostras de tecidos normais retirados da tireoide por PAAF. A justificativa para o emprego de modelos de classe única está na maior disponibilidade de amostras de tecidos tireoidianos normais em relação à disponibilidade de amostras com tecidos cancerígenos. Adicionalmente, considerando a maior variedade de níveis de risco associadas a um diagnóstico de câncer de acordo com a classificação Bethesda, a modelagem das amostras de tecido normal pode produzir um método de triagem mais sensível e robusto para identificar quando o paciente deve receber acompanhamento especializado. O método proposto pode auxiliar no diagnóstico por meio de qualquer profissional da saúde minimamente treinado no manuseio e análises das amostras. Tendo em vista a existência de uma alta demanda de biópsia e a baixa disponibilidade de patologistas em hospitais, principalmente nos mais periféricos, dá-se como de grande importância essa metodologia, sendo uma ferramenta auxiliar aos médicos.

Capítulo 2

OBJETIVOS

OBJETIVOS

Objetivo Geral

Este trabalho tem como objetivo desenvolver um método para auxiliar no diagnóstico de câncer de tireoide, empregando espectroscopia de infravermelho com transformada de Fourier com refletância total atenuada (ATR-FTIR) e algoritmos de classificação multivariada de classe única.

Objetivo Específicos

- Avaliar o impacto de procedimentos de pré-tratamento dos espectros como a filtragem Savitzky-Golay para suavização e a correção de espalhamento multiplicativo de sinal (MSC) na eficiência de predição dos modelos;
- Avaliar a eficiência da técnica ATR-FTIR associada com os modelos de classificação SIMCA e DD-SIMCA na classificação dos tecidos tiroidianos saudáveis em comparação aos tecidos patologicamente alterados da glândula tireoide;
- Avaliar o modelo desenvolvido na classificação de amostras de categoria III, IV e V, que são consideradas como amostras indefinidas no diagnóstico patológico.

Capítulo 3

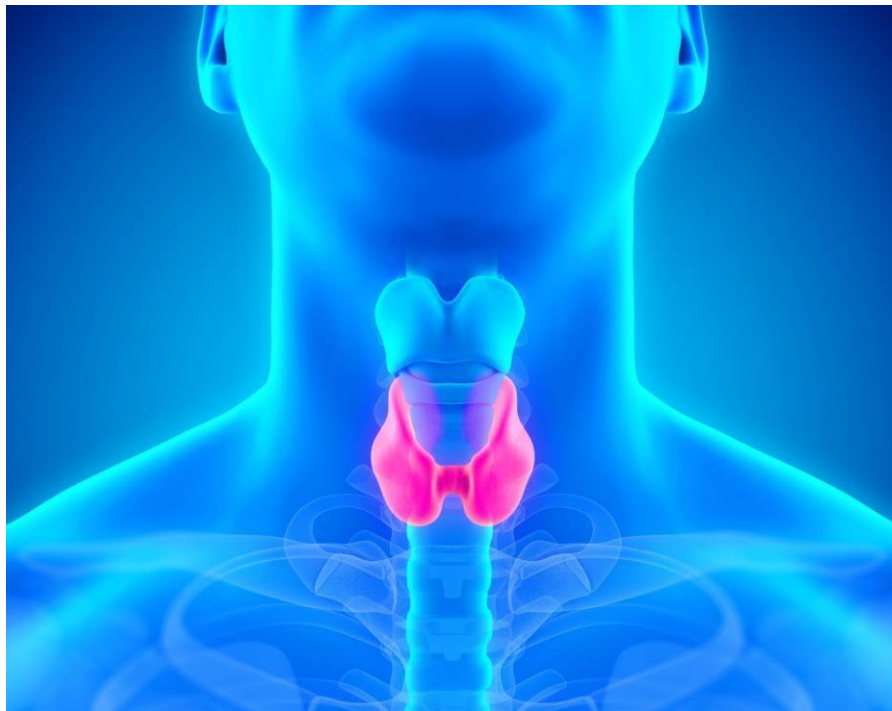
FUNDAMENTAÇÃO TEÓRICA

FUNDAMENTAÇÃO TEÓRICA

Glândula tireoide

A tireoide é a glândula em formato de borboleta, localizada entre a 5ª e 7ª vértebra cervical (figura 1), é constituída de dois lobos, direito e esquerdo, conectados por um istmo, podendo pesar entre 15 g e 25 g. A glândula tireoidiana é cercada por vasos linfáticos e sanguíneos que distribuem por todo o corpo os hormônios produzidos pela tireoide, atuando em todos os sistemas do nosso organismo, agindo no crescimento e desenvolvimento de crianças e adolescentes, humor, fertilidade, memória, peso, entre outras funções (KHAN et al., 2021; MINISTÉRIO DA SAÚDE, 2016; MALBURG et al., 1991).

Figura 1 – Ilustração anatômica da glândula tireoide. Fonte: “human thyroid gland anatomy illustration”, por purple carrot nutrition, licenciado por Creative Commons Attribution-Share Alike 4.0 International.



De acordo com aspectos histológicos, é na glândula de tireoide onde ocorre a síntese dos hormônios T3 (tri-iodotironina) e T4 (tiroxina), além do tecido interfolicular, produtor da calcitonina. Quando há uma produção em excesso de hormônios da tireoide,

esse estado é chamado de hipertireoidismo. Como consequência disso, o organismo se torna mais acelerado e em geral o paciente fica mais agitado, fala mais, anda mais rápido, dorme pouco, entre outras disfunções (KHAN et al., 2021; MINISTÉRIO DA SAÚDE, 2016). O oposto ocorre quando a tireoide não produz hormônios suficientes para o corpo humano, conhecido por hipotireoidismo. O metabolismo desacelera, o intestino não funciona bem, cabelos, unhas e pele ficam fracos, sem brilho, ressecados (GHAZY et al., 2021; KHAN et al., 2021; MINISTÉRIO DA SAÚDE, 2016).

Câncer de tireoide

O câncer de tireoide corresponde a 1% de todos os tumores malignos, atingindo uma faixa etária entre 30 e 74 anos, afetando três vezes mais indivíduo do sexo feminino do que masculino (BORGES et al., 2020; XAVIER et al., 2014). De acordo com uma projeção norte-americana, o câncer de tireoide ocupará uma posição entre os cinco cânceres mais comuns nos Estados Unidos no ano de 2030, com uma tendência similar sendo observada em outros países, como o Brasil (INCA, 2020; RAHIB et al., 2014; WARD et al., 2008). O diagnóstico de câncer de tireoide pode haver variações de acordo com: a região territorial, idade, gênero. Apesar desse crescimento na incidência de câncer de tireoide, vale ressaltar que há um declínio na taxa de mortalidade que está relacionada diretamente ao diagnóstico precoce e tratamento (DAVIES et al., 2006; KILFOY et al., 2011; LA VECCHIA et al., 2014; VACCARELLA et al., 2015; BORGES et al., 2020).

Outro fato interessante, mostrado por Nagar e colaboradores, indica o quão importante é o uso de novas tecnologias facilitadoras do diagnóstico. Eles observaram que médicos com idades mais avançadas e que não usavam de novas tecnologias produziam geralmente um diagnóstico incorreto devido à falta de uso da adaptação do ultrassom na região do pescoço (NAGAR et al., 2014).

Diagnóstico de câncer de tireoide

Desde a década de 70 tem-se observado uma evolução das práticas de diagnósticos clínicos. Os nódulos maiores eram avaliados por palpação e em casos duvidosos era realizado a biópsia. Este cenário no diagnóstico clínico, acarretava numa série de diagnósticos imprecisos e incertos (Na et al., 2017; YOKOZAWA 1998; MALBURG et al., 1991).

A técnica de diagnóstico mais utilizada é a PAAF, que é uma técnica citológica. É um processo de aspiração por agulha fina do nódulo a ser examinado, com objetivo de montar uma lâmina citológica do tecido coletado. Essa técnica é segura ao paciente e pouco invasiva (Na et al., 2017; BRITO et al., 2013; LEENHARDT et al., 2004; YOKOZAWA 1998; MALBURG et al., 1991).

No final dos anos 80, com o advento USG, foi implementado a PAAF guiada por USG, o qual permitia a detecção de nódulos menores até 2 mm, o que tornou o diagnóstico mais preciso nesta época. Segundo Marqusee e colaboradores a implementação da USG mudava a conduta clínica em aproximadamente 63% dos casos de pacientes com diagnóstico de câncer de tireoide (MARQUSEE et al., 2000). O estudo empregou 223 pacientes que foram avaliados na clínica de nódulos da tireoide entre os anos de 1995 e 1997 (MARQUSEE et al., 2000).

Após o ano de 2000, novas tecnologia de exames por imagem, como tomografia computadorizada e ressonância magnética, foram inseridas no diagnóstico de câncer de tireoide, elevando a confiabilidade nos diagnósticos clínicos (YOKOZAWA 1998; LEENHARDT et al., 2004; BRITO et al., 2013). Porém, é importante ressaltar as desvantagens das técnicas de diagnóstico. Para a PAAF, 18% dos resultados no diagnóstico são falsos negativos e em aproximadamente 50% dos diagnósticos dos pacientes são realizadas repetições da técnica (MARQUSEE et al., 2000). Adicionalmente, as novas tecnologias inseridas para diagnóstico clínico são invasivas, não tem custo acessível a todos os usuários, além da obtenção dos resultados requerer mais tempo (PARASKEVAIDI et al., 2018; Na et al., 2017; MARQUSEE et al., 2000).

Uma alternativa para o diagnóstico de câncer de tireoide surgiu por meio da biologia molecular, a partir de análises de genes, que na maioria dos casos apresentam mutação.

O uso da biologia molecular é importante para o descobrimento de marcadores de câncer de tireoide, que podem ser usados em uma triagem precisa de pacientes em relação ao tipo de tratamento a ser administrado (SILVA, 2019; PENNA et al., 2017).

Espectroscopia e Quimiometria como alternativa de diagnóstico de doenças

Alguns estudos demonstram que as medições de absorção na região do IR produzem espectros com picos que representam ligações químicas existentes nas amostras em estudo (MOVASAGHI et al., 2008; WALSH et al., 2008; BAKER. et al., 2014). É conhecido que as biomoléculas absorvem na região do IR, por conta dos movimentos vibracionais das ligações químicas, sendo essa região subdividida em IR próximo, IR médio e IR distante. A região do IR médio ($4000\text{--}400\text{ cm}^{-1}$) é conhecida por absorver uma série de moléculas bioquímicas, contendo regiões de impressão digital da célula bioquímica (KELLY et al., 2011).

A espectroscopia na região do IR é uma técnica que envolve basicamente a interação de luz com a matéria. O IR é dividido em radiação de IR próximo ($12800\text{--}4000\text{ cm}^{-1}$), médio ($4000\text{--}400\text{ cm}^{-1}$) e distante ($400\text{--}10\text{ cm}^{-1}$), sendo a região espectral mais utilizada entre 4000 cm^{-1} a 670 cm^{-1} . Os primeiros equipamentos de IR eram do tipo dispersivo, baseado em rede de difração. Tempo depois a maior parte desses equipamentos foram substituídos por FTIR devido a velocidade, confiabilidade e conveniência (MEDEIROS, 2009; SKOOG et al., 2002).

A técnica de reflexão total atenuada (ATR) é utilizada para fins qualitativos e quantitativos. Esse método se baseia quando um feixe de luz passa por um meio de maior densidade para um de menor, ocorrendo a reflexão. Quando o meio menos denso absorve a radiação, ocorre atenuação do feixe nos comprimentos de onda, sendo esse fenômeno conhecido por ATR. Essa técnica está sendo bastante utilizada por seus espectros serem obtidos de forma rápida (SKOOG et al., 2002).

A técnica Raman com uso de modelagem matemática também permitiu desenvolver um método de diagnóstico de câncer de tireoide que diferencia a partir dos espectros, as características de determinados tecidos humanos, sendo possível diferenciar tecidos

saudáveis e tecidos que apresentam células tumorais (DEPCIUCH et al., 2019; RAU et al., 2016; HARRIS et al., 2009).

Técnicas de espectroscopia vibracional, tal como infravermelho com transformada de Fourier (FTIR) têm sido bastante empregadas como método indicado para estudar a interação da luz com materiais biológicos (MORAIS et al, 2020). Pesquisas demonstram diferenças de intensidades nas regiões de absorção de DNA, RNA, lipídio, amidas e carboidratos para amostras saudáveis e cancerígenas, apresentando maior intensidade de absorção nas amostras malignas (LIU et al., 2011; ZHANG et al., 2017). Villela detalha minuciosamente sobre a caracterização dos tecidos tireoidianos por espectroscopia no infravermelho (VILLELA, 2017), trazendo informações sobre as regiões do espectro com maior informação relacionada ao DNA (900 - 1300 cm^{-1}), proteínas (1450 - 1700 cm^{-1}) e lipídeos (2800 - 3100 cm^{-1}), relacionadas ao aumento da razão CH_3/CH_2 (ANDRUS, 2006; VILLELA, 2017).

Devido ao fato destas técnicas não serem destrutivas para as amostras, elas estão sendo cada vez mais utilizadas, pois também possibilitam a rápida obtenção dos resultados, além da possibilidade de serem associadas as ferramentas quimiométricas, apresentando-se como uma alternativa eficiente para o diagnóstico de doenças (LI et al., 2021; MOVASAGHI et al., 2008).

A técnica ATR-FTIR combinada com técnicas quimiométricas foram usadas para analisar amostras de plasma sanguíneo para diagnóstico de doenças neurodegenerativa, a exemplo do Alzheimer (PARASKEVAIDI et al., 2017). A utilização desta técnica permite o diagnóstico rápido e robusto de doença neurodegenerativa e segregação entre diferentes demências, a exemplo de Bury e colaboradores conseguiram distinguir tumores cerebrais normais ou metastáticos primários, esta metodologia demonstrou um grande potencial (BURY et al., 2019; PARASKEVAIDI et al., 2017).

A técnica de espectroscopia e quimiometria são exploradas na classificação de câncer de próstata com base em dados de espectroscopia de infravermelho médio com transformada de Fourier (FT-MIR) associado com a análise de componente principal, os modelos desenvolvidos oferecem aumento na qualidade do diagnóstico e classificação do câncer de próstata. Identificando variações na estrutura de proteínas secundárias e

alterações de DNA/RNA que são considerados os principais marcadores biomoleculares para os graus de câncer de próstata (SINQUEIRA et al., 2017).

Na Tabela 1, será mostrado um resumo de alguns artigos encontrados nos últimos anos utilizando Espectroscopia e Quimiometria no diagnóstico de câncer de tireoide. Como mencionado, estudos com amostras citológicas são desafiadoras, devido a coleta de amostra adequada para análise, por sua natureza invasiva do procedimento PAAF, pequena quantidade de amostras aspirada. No entanto, o uso de amostras PAAF, em nossa opinião, é mais vantajoso, pois sua análise, realizada antes da cirurgia, pode contribuir para evitar a remoção da tireoide parcial ou quase total em pacientes com diagnóstico impreciso.

Tabela 1 - Resumo dos estudos de pesquisa da espectroscopia e Quimiometria na caracterização de amostras de tecidos tireoidiano saudáveis e doentes, destacando os tipos de amostras, análises e condições.

TIPO DE AMOSTRA	CONDIÇÕES	SENS(%)	ESPEC (%)	AUTORES
n = 119, 67 amostras de câncer anaplásico, 52 amostras saudáveis	Raman / PCA Laser 783 nm 1800 – 600 cm ⁻¹	92	95	Harris et al., 2009
n = 184, 61 metástase, 123 não metátese, 22 amostras câncer papilar	FTIR / LDA Varredura: 32 Resolução: 8 cm ⁻¹ 4000 – 1000 cm ⁻¹	80	91	Liu et al., 2011
n = 60, 17 amostras de câncer papilar; 43 amostras de bócio nodular	ATR / CDA Varredura: 32 Resolução: 4 cm ⁻¹ 4000–800 cm ⁻¹	83	95	Zhang et al., 2011
n = 38, 20 amostras de câncer papilar, 18 amostras saudáveis	Raman / LDA Laser 532 nm Resolução: 5 cm ⁻¹ 3400 – 200 cm ⁻¹	93	100	Rau et al., 2016
n = 32, 17 adenoma folicular e 15 câncer folicular	FTIR / PCA-LDA Varredura: 32 Resolução: 2 cm ⁻¹ 1800 – 800 cm ⁻¹	65	75	Depciuch et al., 2019

O ambiente químico computacional tem sido amplamente utilizado devido à diversidade de programação de rotinas que podem ser utilizadas viabilizando o uso de métodos quimiométricos, como por exemplo Unscrambler, Pirouette, Matlab, Octave (VALDERRAMA et al., 2016; SOUZA et al., 2012). Por meio dos espectros de IR é possível conhecer o perfil espectral das amostras de tecidos normais e amostras de tecidos patologicamente alterados (LI et al., 2021; BAKER, et al., 2014; MOVASAGHI et al., 2008).

Quimiometria

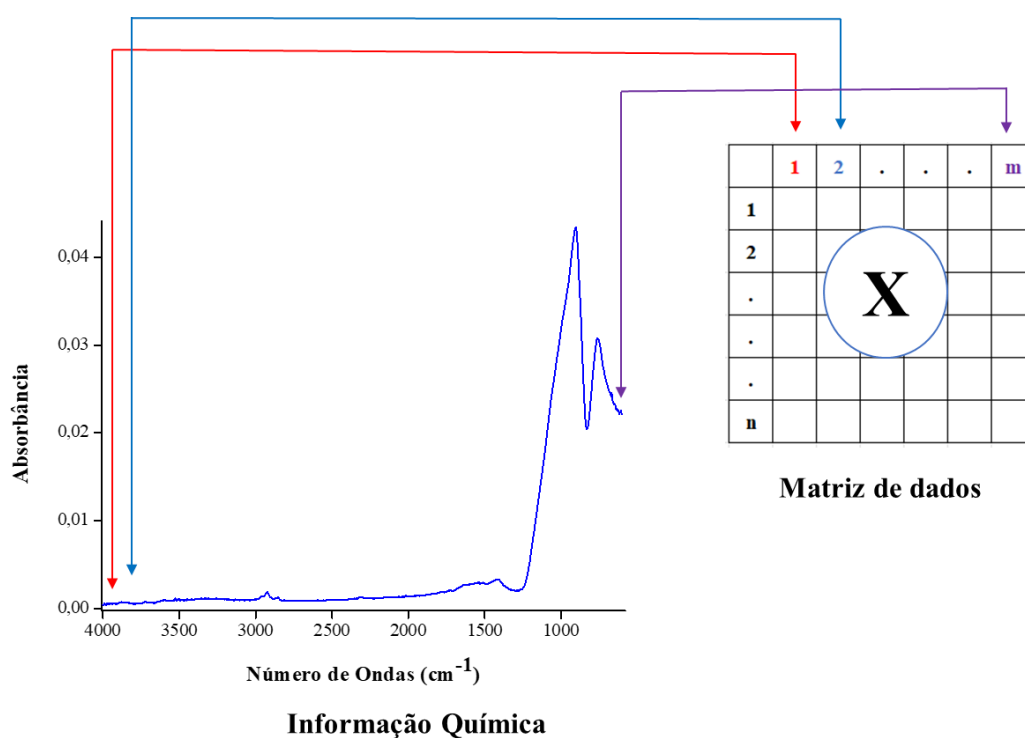
A quimiometria surgiu da necessidade de extrair informações químicas de dados produzidos pela instrumentação moderna. Em outras palavras, na década de 60 havia uma grande quantidade de informação química disponível e faltavam recursos para extrair informações relevantes a partir de um determinado conjunto de dados (NETO et al., 2006; KOWALSKI, 1975).

O termo quimiometria foi introduzido por Svante Wold, em 1971. Alguns anos após em parceria com Bruce Kowalski, fundaram a *International Chemometric Society* (Sociedade internacional da Quimiometria) o objetivo seria fortalecer a ligação entre as áreas de Química, Estatística e Matemática (NETO et al., 2006; KOWALSKI, 1975). A partir de então, a quimiometria tem sido desenvolvida e explorada em diversas áreas para fins qualitativos e quantitativos (FERREIRA, 2015; NETO et al., 2006). Kowalski juntamente com seu grupo de pesquisa, desenvolveram o primeiro pacote computacional quimiométrico chamado ARTHUR, programa de reconhecimento de padrões (NETO et al., 2006; KOWALSKI, 1975).

A quimiometria apesar de ser atual e essencial na química analítica, não está limitada apenas a essa área. A quimiometria tem se mostrado útil nas indústrias farmacêutica (SILVA, 2020; SABIN et al., 2004), química (MEDEIROS, 2009) e de alimentos (NÓBREGA, 2021; HANSEN, 2019), que utiliza equipamentos analíticos para controle de qualidade, por exemplo. A quimiometria também está presente na área médica, auxiliando no diagnóstico, na identificação e classificação de doenças (SILVA, 2019; PARASKEVAIDI et al., 2017; SINQUEIRA et al., 2017).

Medir um sinal nada mais é que registrar a magnitude da resposta de um instrumento em função da sua variável. Por exemplo, em espectroscopia as informações obtidas em espectros são organizadas em uma matriz de dados como podemos observar na figura 2. Em outras palavras, a posição de um pico é determinada pelo comprimento de onda, que podemos chamar de variável (FERREIRA, 2015).

Figura 2 - Matriz de dados a partir de espectro. as informações espectrais são organizadas em matriz de dados.



Os trabalhos em quimiometria podem ser divididos em três áreas principais: planejamento e otimização de experimentos, reconhecimento de padrões e a calibração multivariada (FERREIRA, 2015; NETO, et al., 2006).

No reconhecimento de padrões, pode-se aplicar com diferentes finalidades, podendo destacar análise exploratória de dados, a classificação de amostras. Por exemplo: medidas espectrais, busca-se encontrar agrupamentos de amostras similares entre si, dessa maneira, detectar tendências nos dados (FERREIRA, 2015; NETO, et al., 2006).

Em geral as técnicas de reconhecimento de padrões (RP) são classificadas como não-supervisionados e supervisionada. RP não-supervisionada procura por objetos (amostras) similares, como por exemplo: *Análise de Componente Principal* (PCA), *Análise hierárquica do cluster* (HCA, do inglês: *Hierarchical Cluster Analysis*) foram desenvolvidos para facilitar a interpretação dos resultados, extraís maior informação do conjunto de dados (NETO et al., 2006).

As técnicas de RP supervisionada utiliza de um conjunto de treinamento para classificar os dados obtidos de acordo com as categorias já existentes, como por exemplo *K-Vizinho Mais Próximo* (KNN, do inglês: *K-Nearest Neighbor*), *Análise discriminante linear* (LDA, do inglês: *Linear Discriminant Analysis*), *Modelagem independente e flexível por Analogia de classe* (SIMCA, do inglês: *Soft Independent Modeling of Class*) (NETO et al., 2006).

Para este trabalho, será dado enfoque a técnica PCA, SIMCA e Data Driven - Modelagem Independente e Flexível por Analogia de Classes (DD-SIMCA).

PCA

A análise de PCA é uma técnica frequentemente utilizada para reduzir as dimensões de um conjunto de dados, preservando ao mesmo tempo o máximo de informação. A componente principal, PC_1 é a combinação linear de maior informação. A segunda componente, PC_2 apresenta maior informação na ortogonalidade da PC_1 . A terceira componente principal, PC_3 apresenta máxima informação na ortogonalidade as duas primeiras componentes, PC_1 e PC_2 .

Normalmente as informações de maior importância ficam concentradas nas duas ou três primeiras componentes principais, sendo assim podem ser examinadas à procura de padrões. Ou seja, nas amostras analisadas procura-se observar agrupamentos e padrões (SOUZA; POPPI, 2012; NETO et al., 2006).

Matematicamente, na PCA a matriz (X) é decomposta em um produto de duas matrizes, denominadas de escores (T) e pesos (P), mais a matriz (E) contém os resíduos (WOLD et al., 1987), como mostrado na equação 1.

$$X = TP^T + E \quad (1)$$

Na PCA os escores são as coordenadas das amostras projetadas em um novo sistema de eixos, formados pelas componentes principais, que podem ser usadas para revelar similaridades e diferenças entre as amostras (VALDERRAMA et al., 2016; OLIVIERI et al., 2012; RIBEIRO et al., 2013).

SIMCA

O método SIMCA, é construído a partir de modelos baseados em componentes principais, para construir separadamente um modelo para cada classe correspondente no conjunto de treinamento, ou seja, uma região no espaço multidimensional é delimitada através de um modelo de PCA, uma vez estabelecido, o modelo pode ser usados para classificar qualquer conjunto de amostras.

Para realizar o cálculo da distância da amostra ao modelo SIMCA, utilizam-se o desvio-padrão residual de cada objeto i , que é designado como S_i^c (equação 2) e o outro desvio-padrão residual médio dessa classe, S_0^c (equação 3).

$$S_i^c = \sqrt{\frac{\sum_{j=1}^p (res_{ij}^c)^2}{p - A_c}} \quad (2)$$

$$S_0^c = \sqrt{\frac{\sum_{i=1}^{N_c} \sum_{j=1}^p (res_{ij}^c)^2}{(N_c - A_c - 1) \cdot (p - A_c)}} \quad (3)$$

Onde N_c é o número de amostras pertencentes ao conjunto de treinamento da classe c ; A_c é o número de componentes principais utilizadas pela classe c ; p representa o número de variáveis, i e j representam os índices das amostras e variáveis, respectivamente.

Um teste F é utilizado para determinar se $(S_i^c)^2$ é significativamente maior do que $(s_0^c)^2$. A razão entre as variâncias é calculada para cada uma das amostras da classe c, de acordo com a equação 4.

$$F_{Ca_i} = \frac{(S_i^c)^2}{(s_0^c)^2} \cdot \frac{N_c}{N_c - A_c - 1} \quad (4)$$

O valor calculado de F_{Ca_i} pode ser comparado a um valor F_{crit} que pode ser obtido empiricamente ou tabelado para um determinado nível de confiança e grau de liberdade. Se a amostra sob investigação apresentar um valor de F_{Ca_i} menor do que o obtido pelo F_{crit} , a mesma pertencerá à classe em consideração (PONTES, 2009).

Dois tipos de erros podem ser apresentados em uma classificação SIMCA: Tipo I: a amostra não é classificada em sua classe alvo; Tipo II: a amostra é classificada em outra(s) classe(s) (HANSEN, 2019; PONTES, 2009; KOWALSKI et al., 1971).

O modelo SIMCA se baseia na ausência de sobreposição entre duas classes diferentes, bem como a maior distância entre as classes possíveis. Basicamente, o modelo modelagem independente flexível por analogia de classes – SIMCA, “independente” pelo fato de construir um modelo para cada classe presente nas amostras (HANSEN, 2019; SABIN et al., 2004; WOLD et al., 1987; KOWALSKI et al., 1971).

DD-SIMCA

A modelagem independente e flexível por analogia de classe direcionado pelos dados – DD-SIMCA é uma modificação do método SIMCA, que apenas diferencia no cálculo do limite de aceitação que limitam a classe autêntica. O método DD-SIMCA é utilizado para problemas de autenticação de classe única (*One class*) utilizado como uma ferramenta RP, sendo apenas a classe alvo modelada, qualquer outra amostra distinta ou não inclusa a essa classe é considerada como uma amostra extrema. A classe alvo é construída de acordo com as propriedades da amostra, essa propriedade podem ser sinais analíticos gerados por equipamentos, como espectroscopia (RODIONOVA et al., 2016).

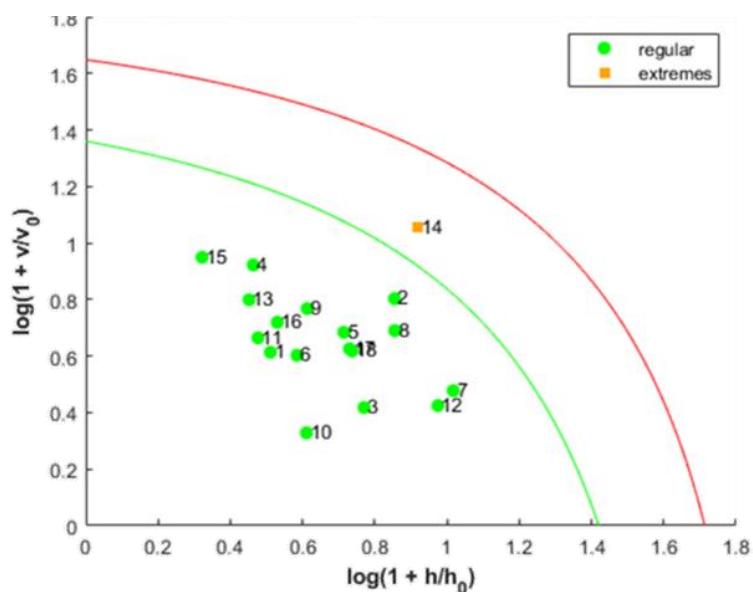
O método DD-SIMCA é baseado em etapas, a primeira aplica a PCA nos dados de treinamento ocorrendo a decomposição da matriz. Quanto maior o número de componentes principais (PC's) maior será a variância explicada. Ao mesmo tempo a inclusão de PC's sem necessidade pode acarretar na interferência da capacidade preditiva do modelo, por ter inserido ruídos irrelevantes. Por tanto a complexidade do modelo influenciará diretamente na qualidade de classificação (RODIONOVA et al., 2014; POMERANTSEV et al., 2013).

Na segunda etapa, é estimada a distância de escores (SD) representa uma amostra dentro do espaço definido pelas PC's e as distâncias ortogonais (OD) é determinada como a soma dos quadrados dos resíduos associados ao modelo PCA para cada amostra pertencente ao conjunto de treinamento. Essas distâncias estabelecem um determinado nível de confiança utilizado como espaço limite a classe alvo. A combinação dessas distâncias SD e OD são descritas como distância global, assumindo valores que seguem uma distribuição qui-quadrado (RODIONOVA et al., 2020; POMERANTSEV et al., 2013).

Na terceira etapa é definida a área de limite para a classe autêntica de acordo com o valor do tipo I (α), que é a taxa de falsa rejeição, quando as amostras não fazem parte da classe alvo. O modelo construído é representado graficamente por uma área de aceitação OD vs distância dos escores, determinado pelo valor α . Para cada amostra do conjunto de treinamento é projetada sua posição nas novas coordenadas do gráfico de aceitação, sendo classificada para a classe alvo como pertencente ou não. Na figura 3 é ilustrado a área de aceitação para um modelo, onde cada amostra pode ser classificada como uma amostra “regular”, ou seja, uma amostra atribuída à classe alvo, ou amostra “extrema”, que está localizada fora da área de aceitação, e é estabelecida como amostra “outliers” (não membro) (RODIONOVA et al., 2020; ZONTOV et al., 2017).

Na figura 3, podemos observar os parâmetros v que é a distância ortogonal (OD), h é a distância de pontuação (SD), v_0 e h_0 são os fatores de escala. A linha verde limita as amostras regulares na área de aceitação e fora dessa área entre as duas linhas verde e vermelha amostra extrema.

Figura 3 - Gráfico da área de aceitação com amostras regulares (verde) e amostra extrema (laranja).



Na última etapa de classificação do conjunto de teste, novas amostras que podem pertencer ou não a classe alvo são empregadas ao modelo construído, o resultado dessa classificação fornece o gráfico de aceitação correspondente. Com o emprego de uma classe alternativa, existe a possibilidade do cálculo do erro tipo II (b), que é a taxa de falsa aceitação quando as amostras não pertencentes a classe alvo são classificadas como positivas. A Sensibilidade e especificidade são os termos frequentemente utilizados para apresentar os resultados da classificação, são conhecidos por figuras de mérito (NEVES et al., 2020; ZONTOV et al., 2017).

As vantagens do DD-SIMCA são: método duplo de estimativa para dados regulares, clássico e para dados contaminados, robusto; abordagem orientada por dados para avaliação de distribuições de distâncias e uma associação entre os níveis de significância extremos e discrepantes (POMERANTSEV et al., 2013).

Validação dos modelos de classificação

Figuras de mérito

O desempenho dos modelos de classificação podem ser avaliados por alguns parâmetros de desempenho analítico, sendo os mais comuns: sensibilidade (SENS) (equação 5), especificidade (ESPEC) (equação 6) e acurácia (AC) (equação 7). Esses parâmetros avaliam a capacidade do modelo identificar as amostras pertencentes a classe alvo ou não (RODIONOVA et al., 2020; ZONTOV et al., 2017).

Os parâmetros são baseados em respostas binárias: verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN) e falso negativo (FN). FP, quando o modelo classifica como positivo uma amostra negativa e FN, quando a amostra positiva é classificada como negativa. Neste trabalho o resultado é considerado como positivo quando a amostra era de paciente saudável e negativo quando o paciente tinha câncer de tireoide (RODIONOVA et al., 2020; ZONTOV et al., 2017).

A SENS, representa o nível de confiança do conjunto da classe modelada (positivo = pacientes saudáveis).

$$Sens (\%) = \frac{VP}{VP+FN} \times 100 \quad (5)$$

Em outras palavras a SENS é a quantidade total de amostras da classe alvo subtraindo a quantidade de amostras extremas. Dividindo tudo pela quantidade total de amostras alvo (ZONTOV et al., 2017).

A ESPEC, representa a habilidade do modelo ao identificar as amostras não pertencentes à classe modelada.

$$Espec (\%) = \frac{VN}{VN+FP} \times 100 \quad (6)$$

Ou seja, ESPEC é calculado a partir do número total de amostras extremas pertencentes a classe não alvo dividida pelo total de amostras correspondente as amostras não alvo do conjunto de teste. É importante ressaltar que o cálculo da ESPEC no modelo DD-SIMCA só poderá ser calculado a partir de amostras pertencentes a classe não alvo, sendo assim o conjunto de treinamento não possuem valores de SPEC (ZONTOV et al., 2017)

Outro parâmetro utilizado para validação é acurácia, que é calculado como a média geométrica entre a sensibilidade e especificidade, em resumo representa a habilidade de classificação total dos modelos qualitativos.

$$AC (\%) = \left[\frac{VP+VN}{VP+FP+VN+FN} \right] \times 100 \quad (7)$$

O cálculo se dá a partir do número de amostras classificadas corretamente (independente da classe), dividido pelo total de amostras (RODIONOVA et al., 2020).

CAPÍTULO 4

METODOLOGIA

METODOLOGIA

- Aquisição das amostras PAAF

Todas as amostras foram coletadas por meio de punção aspirativa por agulha fina (PAAF). Essa técnica permite o corte do tecido tireoidiano pela agulha, obtendo uma quantidade de material celular suficiente para colocar em uma lâmina. Posteriormente passam por um processo de coloração por hematoxilina & eosina (H&E), permitindo corar diferentes células e estruturas tornando-as visíveis ao microscópio. A coloração é empregada para realização de diagnóstico pelo patologista. Quanto maior a quantidade de mitose na amostra, mais corada será. Este procedimento de coleta de amostra e coloração foi realizado pelos funcionários do setor de patologia no Hospital Universitário Lauro Wanderley - HULW. Importante ressaltar que a metodologia proposta neste trabalho não requer necessariamente que as lâminas estejam coradas, já que a assinatura espectral das diferentes células estará presente nos espectros obtidos sem a utilização do corante. Entretanto, como foram utilizadas lâminas do diagnóstico padrão, por esse motivo se verifica a coloração delas. As amostras empregadas neste estudo, foram todas de um único banco de dados, do departamento de patologia do HULW.

Este estudo teve aprovação do Comitê de Ética em Pesquisa – CEP/HULW, parecer nº 3.785.821/2019.

- Equipamento e materiais utilizados

Neste trabalho, utilizaram-se das lâminas de vidro (dimensão: 26 x 76mm e espessura: 1,00mm) contendo a amostra coletada dos pacientes diagnosticados com câncer de tireoide e pacientes controle. Também foi empregada lâmina limpa, ou seja, sem nenhuma amostra, que foi utilizada como o branco nas análises.

Os experimentos foram realizados no equipamento Espectrofotômetro Shimadzu modelo IR Prestige-21 (Kyoto, Japão), com acessório de reflectância total atenuada acoplado, com as seguintes condições de análise: Região 4000 - 600 cm^{-1} ; Resolução: 4

cm⁻¹; Modo: transmitância. 32 varreduras, utilizando a média dessas varreduras. Os espectros de infravermelho das 49 amostras de tecido, obtidos através do ATR-FTIR (4000 – 600 cm⁻¹), foram obtidos em modo de transmitância e convertidos em absorbância ($A = -\log(T)$).

- Tratamento de dados: Organização dos dados e algoritmo de tratamento

O estudo consistiu na participação de 49 amostras de tecido (n = 49), dentre eles 27 pacientes diagnosticados como benigno (n = 27), 11 pacientes diagnosticados com câncer de tireoide (n = 11) e 11 pacientes diagnosticados como indefinidos. Na tabela 1, encontra-se característica dos pacientes como idade, gênero, ano da amostra coletada. A coleta realizada em cada paciente deu origem a seis lâminas. Foram obtidos os espectros de cada lâmina e utilizada a média dos seis espectros na modelagem.

Tabela 2 - Característica clínica dos pacientes.

Amostras	Diagnóstico histopatológico	Ano	Gênero	Idade
1	Benigna	2017	F	65
2	Benigna	2017	F	41
3	Benigna	2017	F	57
4	Benigna	2017	F	45
5	Benigna	2017	F	68
6	Benigna	2017	F	56
7	Benigna	2018	F	69
8	Benigna	2018	F	74
9	Benigna	2018	F	47
10	Benigna	2018	F	69
11	Benigna	2018	F	37
12	Benigna	2018	F	62
13	Benigna	2019	F	50
14	Benigna	2019	F	55
15	Benigna	2019	F	61
16	Benigna	2019	F	47
17	Benigna	2019	F	44
18	Benigna	2019	F	74
19	Benigna	2019	F	46
20	Benigna	2019	F	78

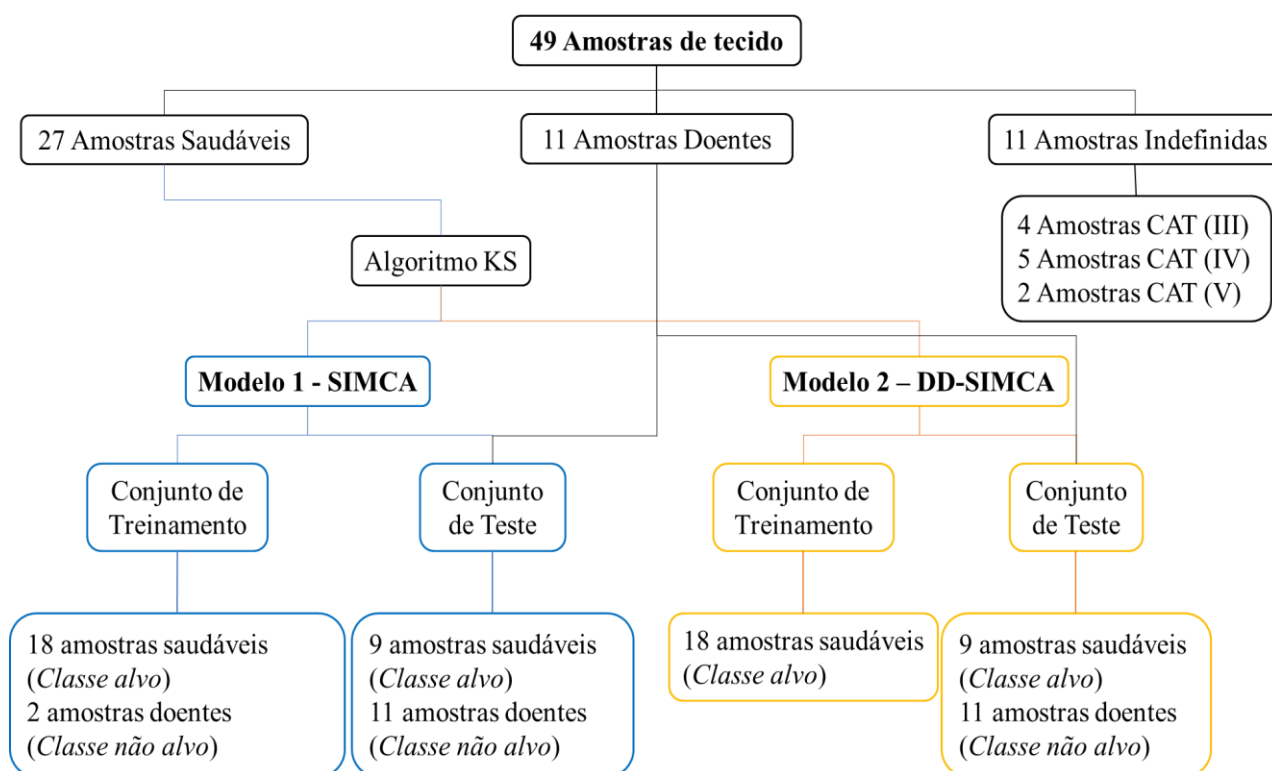
21	Benigna	2020	F	50
22	Benigna	2020	F	48
23	Benigna	2018	F	38
24	Maligna	2017	F	72
25	Maligna	2017	F	38
26	Maligna	2018	F	36
27	Maligna	2019	M	50
28	Maligna	2019	F	X
29	Maligna	2020	F	X
30	Benigna	2017	F	50
31	Maligna	2018	M	50
32	Maligna	2020	M	28
33	Benigna	2017	F	49
34	Benigna	2017	F	42
35	Benigna	2018	M	33
36	Maligna	2018	X	X
37	Maligna	2018	X	X
38	Maligna	2015	X	X

X= não informado; F = Feminino; M = Masculino.

Como podemos observar, o número de amostras de pacientes saudáveis é maior que as amostras de pacientes doentes, sendo assim optou-se por um modelo de classificação de classe única empregando os algoritmos SIMCA e DD-SIMCA. A classe alvo para desenvolvimento do método são as amostras de pacientes saudáveis, sobre as quais a modelagem foi realizada. Para validação do método as amostras de pacientes doentes foram alocadas no conjunto de teste.

Os dados espectrais obtidos foram organizados em matrizes (amostra x variáveis). Antes de cada método aplicado, as amostras espectrais benignas foram separadas em conjunto de treinamento 70%, validação 30%, através da aplicação clássica do Kennard – Stone (KS), que consiste no algoritmo uniforme de amostragem (figura 4). Para o conjunto de treinamento foram separadas 18 amostras saudáveis e para o conjunto de teste foram separadas 9 amostras saudáveis. Ao conjunto de teste adicionaram-se as 11 amostras de pacientes doentes, totalizando 20 amostras no conjunto de teste.

Figura 4 - Conjunto de treinamento, teste e indefinidas para os modelos SIMCA e DD-SIMCA.



Para o cálculo SIMCA o toolbox do Matlab (Classification Toolbox 6.0) requer um número mínimo de 2 amostras de outra classe (no caso deste trabalho, duas amostras de pacientes diagnosticados com câncer) no conjunto de treinamento para a execução do cálculo, para o cálculo das figuras de mérito. Essas duas amostras inseridas no conjunto de treinamento, também foram inseridas no conjunto de teste. Assim, para o modelo SIMCA a modelagem do conjunto de treinamento empregou 18 amostras de tecidos saudáveis e 2 amostras de tecidos doentes, totalizando 20 amostras, e para o conjunto de teste manteve-se as 9 amostras normais e 11 amostras de tecidos doentes, totalizando 20 amostras de tecidos. A modelagem DD-SIMCA não requer duas classes na etapa de treinamento e o conjunto de treinamento consistiu de 18 amostras de tecido saudáveis, e o conjunto de teste continha 9 amostras de tecido saudáveis e 11 amostras de tecido doentes, totalizando 20 amostras de tecido.

Após desenvolvimento e escolha do modelo de classificação para as amostras definidas, foi utilizado um terceiro conjunto contendo amostras classificadas como indefinidas segundo a escala de Bethesda. Foram 11 amostras ($n = 11$) classificadas pelo

patologista entre as categorias (CAT) (III) (benigna), CAT (IV) normalmente essa classificação é a mais indefinida, considerou-se neste trabalho como maligna e CAT (V) (malignas), como apresentado na tabela 2.

Tabela 3 - Característica clínica dos pacientes que apresentam diagnóstico indefinido.

Amostras	Diagnóstico histopatológico (n)	Ano	Gênero	Idade
1	CAT V	2019	M	25
2	CAT V	2019	F	43
3	CAT IV	2018	F	36
4	CAT IV	2019	F	55
5	CAT IV	2019	M	50
6	CAT IV	2019	M	67
7	CAT IV	2019	F	25
8	CAT III	2017	F	53
9	CAT III	2017	F	25
10	CAT III	2018	F	48
11	CAT III	2018	F	34

n = número de lâminas; F = Feminino; M = Masculino; CAT III = Lesão folicular de significado indeterminado; CAT IV = Neoplasia folicular; CAT V = Suspeita de Malignidade.

- Pré-processamentos

Os métodos de pré-processamento aplicados nas amostras do presente estudo foi a filtragem Savitzky-Golay para suavização dos espectros para diminuir os ruídos aleatórios existentes, o qual empregou um polinômio do primeiro grau e janela de 3 pontos. Na segunda etapa aplicou-se o algoritmo da correção de espalhamento multiplicativo (*Multiplicative Scatter Correction*, MSC), sendo obtida por regressão linear para estimar o coeficiente angular e linear entre o espectro médio e o espectro a ser corrigido resultando em uma diminuição no efeito de dispersão da luz decorrente de fenômenos físicos (MORAIS et al., 2020; FERREIRA, 2016).

Posteriormente, optou-se por selecionar a região entre 650 cm^{-1} e 1400 cm^{-1} , por apresentaram picos de absorção intensa. Neste estudo também foram utilizados os dados brutos para cada região selecionada, bem como os dados pré-processados.

O software The Unscrambler® foi utilizado para a aplicação das duas etapas de pré-processamento na matriz de dados. O software OriginPro 8.5 foi utilizado para produção

de alguns gráficos deste trabalho. Os modelos quimiométricos foram desenvolvidos no ambiente Matlab R2020a. Foi usada a caixa de ferramenta para seleção de amostra disponível para aplicação KS, a caixa de ferramenta Classification Toolbox 6.0 disponível em <https://michem.unimib.it/download/matlab-toolboxes/> para utilização do modelo SIMCA, e a caixa de ferramenta DD-SIMCA Master, disponível em <https://github.com/yzontov/dd-simca>, para aplicação da técnica DD - SIMCA. Os dados foram importados para o MATLAB® R2020a, no qual foi realizada separação das amostras os conjuntos de treinamento e teste com o algoritmo KS e os modelos de classificação aplicados.

Para construção do modelo SIMCA foi utilizada a validação cruzada com a exclusão de uma amostra por vez (*leave-one-out cross-validation*) para auxiliar a escolha do número de componentes do modelo. O número de componentes foi escolhido considerando-se a menor taxa de erro de classificação para o conjunto de validação cruzada. Avaliou-se a centralização dos dados e o autoescalamento como técnicas de pré-processamento de variáveis.

Para construção do modelo DD-SIMCA foi adotado o nível de significância de outliers de 0,01 para definição da área de aceitação usando distribuição qui-quadrado no modo clássico. A escolha do número de componentes foi realizada após análise dos valores de sensibilidade para o conjunto de treinamento, e menor número de amostras extremas (falso negativo) ou *outliers* (amostras discrepantes) em relação ao conjunto alvo.

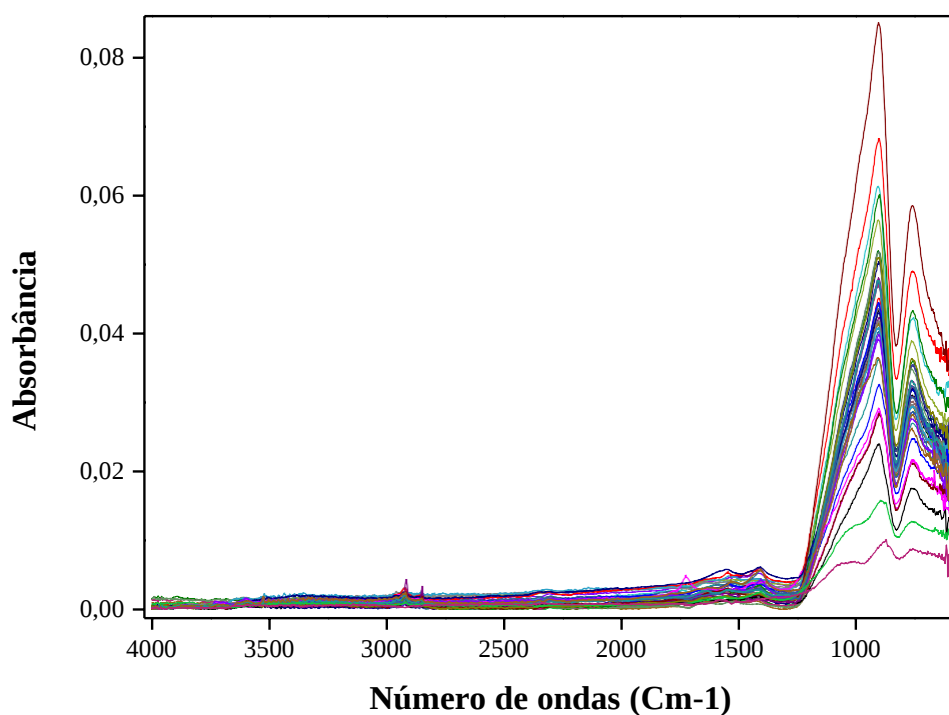
Capítulo 5

RESULTADOS E DISCUSSÃO

RESULTADOS E DISCUSSÃO

A figura 5 apresenta os espectros médios ATR-FTIR dos 38 pacientes obtidos neste trabalho. Nota-se que há uma intensa absorção de radiação na região entre 1500 cm^{-1} e 600 cm^{-1} . Por outro lado, entre 4000 cm^{-1} e 1700 cm^{-1} os espectros são muito ruidosos e não apresentam informações analíticas úteis.

Figura 5 - Espectros ATR-FTIR lâminas de pacientes com câncer de tireoide e pacientes saudáveis, na região de 4000 a 600 cm^{-1} sem pré-processamento.

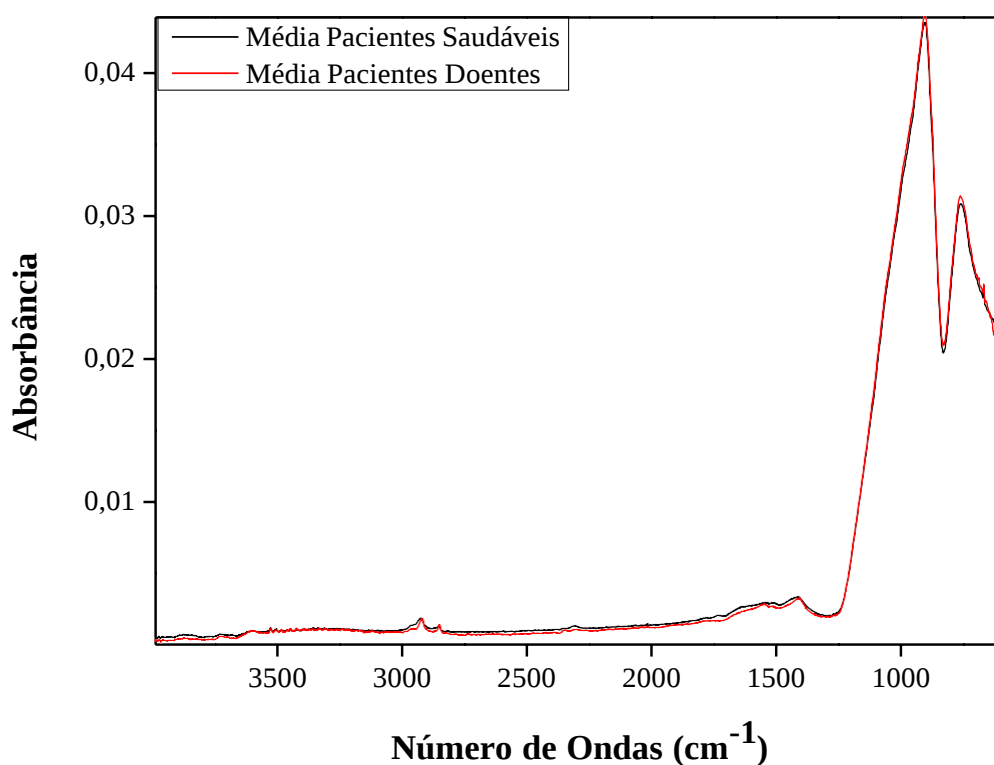


Movasaghi e colaboradores forneceram um banco de dados preciso sobre as frequências de pico característicos de FTIR. São de grande valia para análises de tecidos benignos e cancerosos por espectroscopia FTIR. As frequências vibracionais no pico de $\sim 940\text{ cm}^{-1}$ são referentes a carotenóides (KUMAR, J. K., et al., 2011); $\sim 1400\text{ cm}^{-1}$ $\delta\text{C-H}$, $\delta\text{C-O-H}$, são relacionadas a lipídios; $\sim 1460\text{ cm}^{-1}$ $\delta\text{C-H}$; $\sim 1550\text{ cm}^{-1}$ amida II, proteína

(ZHANG, X., et al., 2010); $\sim 1600\text{ cm}^{-1}$ amida I, proteína (VILLELA, L. F. A., 2017; ZHANG, X., et al., 2010).

Comparam-se as médias dos espectros dos pacientes saudáveis e dos pacientes doentes, figura 6. Verifica-se que o espectro de cor vermelha (média dos pacientes doentes) não apresenta diferença visível de intensidade na absorção em toda região espectral, em comparação com o espectro de cor preta (média dos pacientes saudáveis).

Figura 6 - Média dos Espectros ATR-FTIR para pacientes saudáveis (—) sobreposta a média dos espectros de pacientes doentes (—).

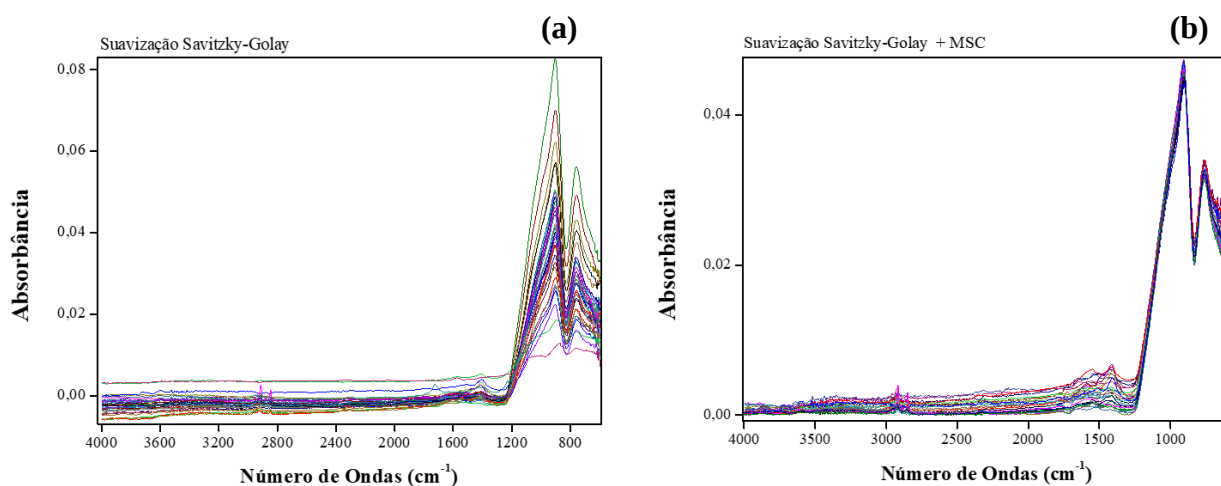


Foi realizado o pré-processamento dos dados espectrais, na figura 7 são mostradas duas etapas de pré-processamento. Na primeira etapa foi aplicada a suavização Savitzky-Golay em uma janela de 3 pontos para um polinômio de grau 1, no intuito de diminuir os ruídos aleatórios em todo espectro (figura 7a). Observa-se que em comparação com os

espectros brutos, não há uma diferença visível significativa após aplicação da suavização Savitzky-Golay.

A segunda etapa foi a aplicação do MSC (figura 7b) com objetivo de diminuir a dispersão de luz existente nos espectros. O espectro corrigido é calculado pela subtração do coeficiente linear para cada número de onda original (para correção dos efeitos aditivos) e, em seguida, divide-se este pelo coeficiente angular (para correção de efeitos multiplicativos).

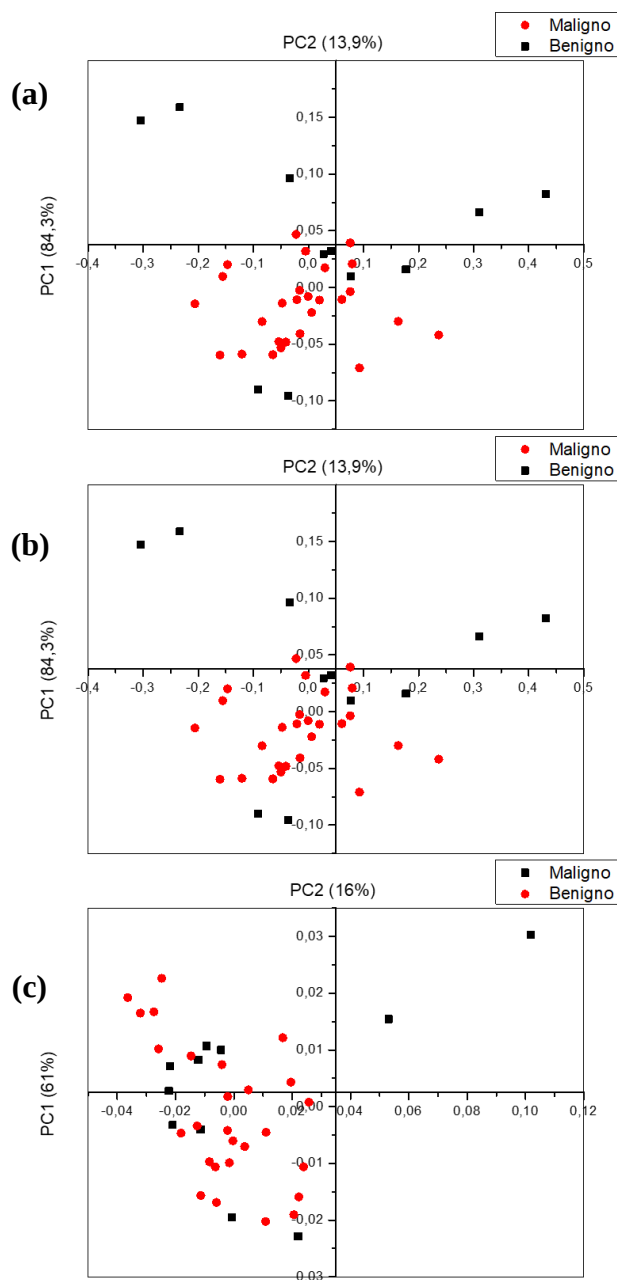
Figura 7 – Espectros com pré-processamentos. (a) Suavização Savitzky-Golay, (b) Suavização + MSC.



A análise exploratória PCA foi realizada nos espectros completos (4000 – 600 cm⁻¹) com intuito de observar comportamentos, para os espectros brutos e com pré-processamento. Na figura 8 são ilustrados os gráficos de escores das duas primeiras componentes principais (PC) para os espectros sem pré-processamento (8a), com suavização Savitzky-Golay (8b) e com suavização Savitzky-Golay + MSC (8c).

Os espectros sem pré-processamento e com suavização Savitzky-Golay requereu 4 PC's com variância explicada de aproximadamente 98%. Esses valores eram esperados por não haver diferenças significativas dos espectros brutos e os espectros com suavização Savitzky-Golay. Para os dados com suavização Savitzky-Golay + MSC para 4 PC's nos forneceu uma variância total dos dados de aproximadamente 96%.

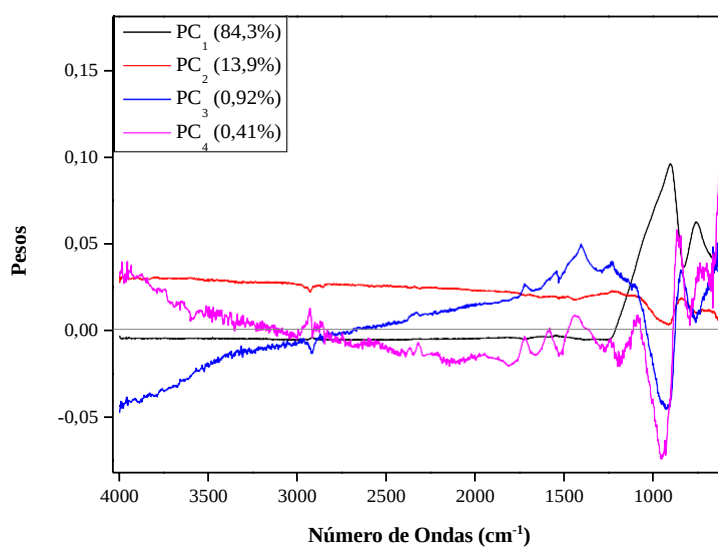
Figura 8 - Gráficos dos escores referentes aos 38 pacientes da PC1 x PC2. Espectros 4000 – 600 cm^{-1} (a) sem pré-processamento; (b) Suavização (Savitzky-Golay) (c) Suavização (Savitzky-Golay) + MSC. Amostras de pacientes saudáveis (●) e amostras de pacientes doentes (■).



É possível notar que na figura 8 as amostras malignas (preto) não apresentam um padrão. Pode-se explicar isso pela maior variabilidade das amostras benignas e por termos um pequeno número de amostras de tecidos doentes, diferente das amostras de tecidos saudáveis (vermelho), que apresentam uma tendência de agrupamento visto na figura 8a, por se esperar que sejam amostras com menor variabilidade.

O gráfico de pesos referente às PC₁, PC₂, PC₃ e PC₄ dos espectros sem pré-processamento são apresentados na figura 9. A partir do gráfico de pesos pode-se encontrar quais as partes do espectro que são responsáveis pela separação entre as amostras, observada no gráfico de escores na figura 8a. As amostras de tecido saudáveis são agrupadas na parte negativa de PC₁ e PC₂ e parte positiva PC₁, as amostras de tecidos doentes são distribuídas na parte negativa PC₁ e PC₂ e na parte positiva PC₁ e PC₂. Trazendo essas informações para o gráfico de pesos, verifica-se que PC₁ (preto) consiste em partes negativas (4000 – 1250 cm⁻¹) e parte positiva (1250 – 600 cm⁻¹) que evidencia duas bandas ~716 cm⁻¹ e ~912 cm⁻¹. Por outro lado, o gráfico de pesos para PC₂ (vermelho) apresenta parte positiva (4000 – 6000 cm⁻¹).

Figura 9 - Gráfico de pesos para PC₁, PC₂, PC₃ e PC₄, obtendo uma variância explicada acima de 90%, para os espectros sem pré-processamento.



Os gráficos de escores PC₁ x PC₃ e PC₁ x PC₄ não são mostrados, no entanto apresentam comportamento similares. Sendo assim, na figura 9 é mostrado o gráfico de pesos PC₃ que apresenta parte negativa (4000 – 2687 cm⁻¹ e 1044 – 879 cm⁻¹) com um pico evidente em ~937 cm⁻¹ e parte positiva (2687 – 1044 cm⁻¹ e 879 – 600 cm⁻¹) com picos centrados em ~1408 cm⁻¹ e 845 cm⁻¹. O gráfico de pesos para PC₄ apresenta parte negativa com pico evidente em 950 cm⁻¹ e parte positiva ~868 cm⁻¹.

Os resultados do modelo de classificação de classe única serão apresentados a seguir. Para os modelos SIMCA e DD-SIMCA foram aplicados nos espectros sem pré-processamento e com pré-processamento. A utilização dos algoritmos quimiométricos teve como classe alvo as amostras benignas, por ter uma quantidade maior de amostras e também por esperar que tenha uma variabilidade menor do que as amostras de tecidos com câncer de tireoide, como mencionado anteriormente. O conjunto de teste consiste em amostras de tecidos saudáveis e doentes.

Na tabela 3 são apresentados os resultados da matriz de confusão obtidos para os modelos de classificação SIMCA para os espectros completos (4000 - 600cm⁻¹) em termos de sensibilidade, especificidade e acurácia para os dados brutos e com os processamentos nos espectros ART-FTIR.

Tabela 4 - Resultados dos valores da matriz de confusão para o modelo SIMCA nos conjuntos de treinamento (20 amostras) e teste (20 amostras) na região entre 4000 e 600 cm⁻¹.

Espectro	PC	Validação	Real/predição	classe 1 (alvo)	classe 2 (não alvo)	SENS (%)	ESPEC (%)	AC (%)
Bruto	4	Treinamento	classe 1 (alvo)	17	1	94	-	-
			classe 2 (não alvo)	1	1			
		Validação cruzada	classe 1 (alvo)	14	4	78	-	-
			classe 2 (não alvo)	1	1			
		Teste	classe 1 (alvo)	7	2	78	64	70
			classe 2 (não alvo)	4	7			
Suavização (Savitzky-Golay)	4	Treinamento	classe 1 (alvo)	17	1	94	-	-
			classe 2 (não alvo)	1	1			
		Validação cruzada	classe 1 (alvo)	14	4	78	-	-
			classe 2 (não alvo)	1	1			
		Teste	classe 1 (alvo)	7	2	78	64	70
			classe 2 (não alvo)	4	7			
Suavização (Savitzky-Golay) + MSC	3	Treinamento	classe 1 (alvo)	9	9	50	-	-
			classe 2 (não alvo)	1	1			
		Validação cruzada	classe 1 (alvo)	5	13	28	-	-
			classe 2 (não alvo)	1	1			
		Teste	classe 1 (alvo)	4	5	44	73	60
			classe 2 (não alvo)	3	8			

PC: Componentes principais; MSC: Correção de espalhamento multiplicativo; SENS: Sensibilidade; ESPEC: Especificidade; AC: Acurácia.

A escolha do número de PC's para o modelo SIMCA foi realizada considerando o valor de classificação da validação cruzada, produzindo consequentemente o maior valor de sensibilidade. Observa-se na tabela 3 que o conjunto de treinamento para os dados espectrais bruto e com suavização Savitzky-Golay apresentaram melhores valores de SENS, ESPEC e AC, para 4 PC's. Para os conjuntos de teste dos espectros brutos e com suavização Savitzky-Golay, há 6 erros de predição, sendo 2 para amostras benignas classificadas como amostra não alvo e 4 amostras malignas que classificou como amostra alvo, sendo obtido valor de SENS 78% e ESPEC 64%. Para os espectros com suavização Savitzky-Golay + MSC, previu erroneamente 8 amostras, sendo 5 de amostras benignas e 3 amostras malignas.

Assim, os modelos foram utilizados na predição das amostras indefinidas (CAT III, IV e V) para observarmos a similaridade do diagnóstico do patologista com o modelo. Na tabela 4 são mostrados valores de classificação para as amostras indefinidas, consideramos os modelos para os espectros brutos e com suavização Savitzky-Golay.

Tabela 5 - Valores de predição para amostras indefinidas, utilizando os modelos SIMCA para os espectros bruto e suavizado (Savitzky-Golay).

Amostras	Classificação	Diagnóstico Patológico	Predição SIMCA
			Espectros Bruto e Suavizado (Savitzky-Golay)
1	CAT (V)	Maligna	Maligna
2	CAT (V)	Maligna	Maligna
3	CAT (IV)	Maligna	Benigna
4	CAT (IV)	Maligna	Maligna
5	CAT (IV)	Maligna	Benigna
6	CAT (IV)	Maligna	Maligna
7	CAT (IV)	Maligna	Benigna
8	CAT (III)	Benigna	Benigna
9	CAT (III)	Benigna	Maligno
10	CAT (III)	Benigna	Maligno
11	CAT (III)	Benigna	Benigna

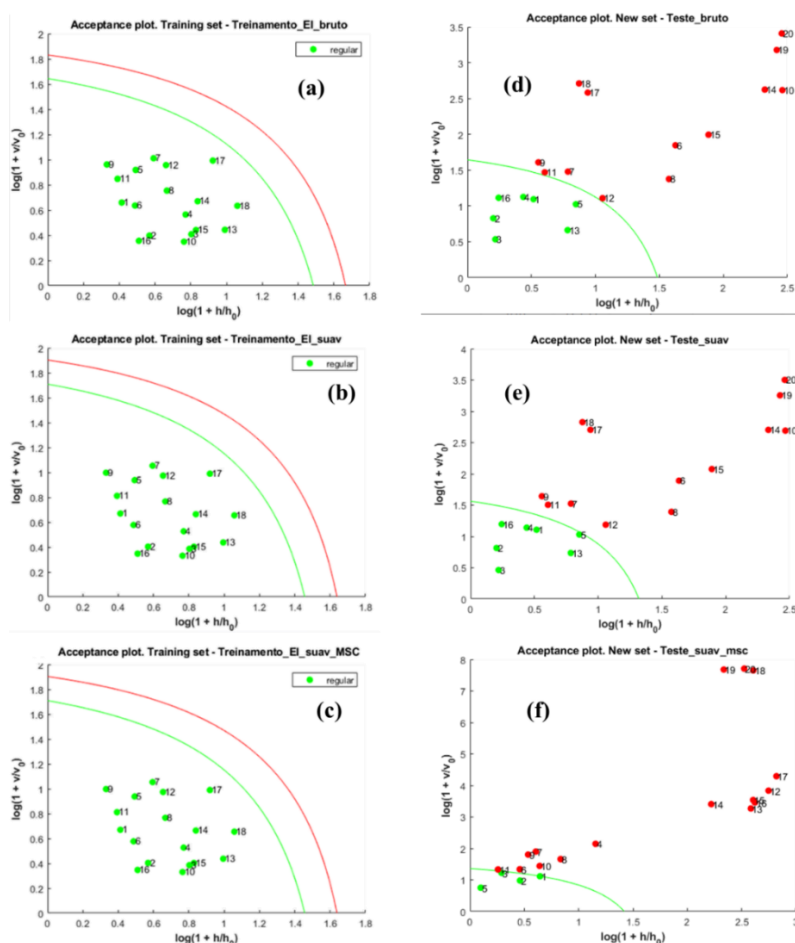
De acordo com o diagnóstico patológico, as amostras de CAT (III), são consideradas como potencialmente benignas, a CAT (IV) apresenta maior indefinição na classificação, consideramos nesse estudo a classificação de maligna e a CAT (V) é

considerada como potencialmente malignas. Desse modo, teremos 4 amostras indefinidas (benignas) e 7 amostras indefinidas (malignas).

Os resultados de classificação mostrados na tabela 4 para os espectros brutos e com suavização Savitzky-Golay apresentam cinco erros de classificação, sendo três para amostras CAT (IV) e duas para amostras CAT (III). De modo que apresenta uma SENS global de 69% e ESPEC global de 61%, considerando o conjunto de teste e amostras indefinidas.

A seguir poderemos observar os gráficos da área de aceitação para os espectros brutos e com pré-processamento na figura 10, obtidos através do DD-SIMCA.

Figura 10 - Gráfico de área de aceitação para o conjunto de treinamento (a, b, c) e teste (d, e, f) para o algoritmo de classificação DD-SIMCA desenvolvidos no ATR-FTIR 4000-600cm⁻¹ (espectros inteiros). (a, d) espectro bruto; (b, e) espectro com suavização; (c, f) espectro com suavização mais MSC. (●) amostra regular e amostra externa (●). As linha verde representa a área de aceitação e as linha vermelha é a fronteira da área de aceitação.



Na figura 10, encontra-se os resultados de classificação para os espectros sem pré-processamento e com pré-processamento. Para os conjuntos de treinamento (figuras 10a, 10b e 10c) nota-se que todas as amostras de tecido saudáveis se encontram dentro da região limitada com a linha verde, sendo considerada como regular. Nas figuras 10d, 10e e 10f, encontram-se a classificação DD-SIMCA para os conjuntos de teste, que consistem em amostras saudáveis e doentes. Os resultados para os modelos de classificação DD-SIMCA podem ser observados na tabela 5.

Tabela 6 - Resultados dos parâmetros de desempenho sensibilidade, especificidade e acurácia para o modelo DD-SIMCA nos conjuntos de treinamento e teste na região entre 4000 e 600 cm⁻¹.

Espectro	PC	Validação	Real/predição	classe 1 (alvo)	classe 2 (não alvo)	SENS (%)	ESPEC (%)	AC (%)
Bruto	3	Treinamento	classe 1 (alvo)	18	0	100	-	-
			classe 2 (não alvo)	0	0			
		Teste	classe 1 (alvo)	7	2	78	64	70
			classe 2 (não alvo)	4	7			
Suavização (Savitzky-Golay)	2	Treinamento	classe 1 (alvo)	18	0	100	-	-
			classe 2 (não alvo)	0	0			
		Teste	classe 1 (alvo)	7	2	78	64	70
			classe 2 (não alvo)	4	7			
Suavização (Savitzky-Golay) + MSC	4	Treinamento	classe 1 (alvo)	18	0	100	-	-
			classe 2 (não alvo)	0	0			
		Teste	classe 1 (alvo)	7	2	78	64	70
			classe 2 (não alvo)	4	7			

Na tabela 5 observa-se que o algoritmo de classificação DD-SIMCA requereu 3 PC's para os espectros brutos, 2 PC's para os espectros com suavizados Savitzky-Golay e 4 PC's para os espectros suavizados Savitzky-Golay + MSC.

Observa-se também na tabela 5 que os valores de SENS para os conjuntos de treinamento é 100% e para os conjuntos de teste são obtidos valores de SENS 78% e

ESPEC 64%. Os modelos DD-SIMCA errou na classificação de 6 amostras sendo, duas amostras benignas e 4 amostras malignas.

Após obtenção dos modelos DD-SIMCA, foi realizado a classificação das amostras indefinidas, de CAT (III), (IV) e (V) , esse resultado pode ser visto na tabela 6.

Tabela 7 - Valores de predição para amostras indefinidas, utilizando os modelos DD-SIMCA para os espectros bruto e suavizado (Savitzky-Golay)

Amostras	Classificação	Diagnóstico Patológico	Predição DD-SIMCA	
			Espectro Bruto	Espectro Suavizado (Savitzky-Golay)
1	CAT (V)	Maligna	Maligna	Maligna
2	CAT (V)	Maligna	Maligna	Maligna
3	CAT (IV)	Maligna	Benigna	Benigna
4	CAT (IV)	Maligna	Maligna	Maligna
5	CAT (IV)	Maligna	Benigna	Benigna
6	CAT (IV)	Maligna	Maligna	Maligna
7	CAT (IV)	Maligna	Benigna	Benigna
8	CAT (III)	Benigna	Benigna	Benigna
9	CAT (III)	Benigna	Benigna	Benigna
10	CAT (III)	Benigna	Benigna	Benigna
11	CAT (III)	Benigna	Benigna	Benigna

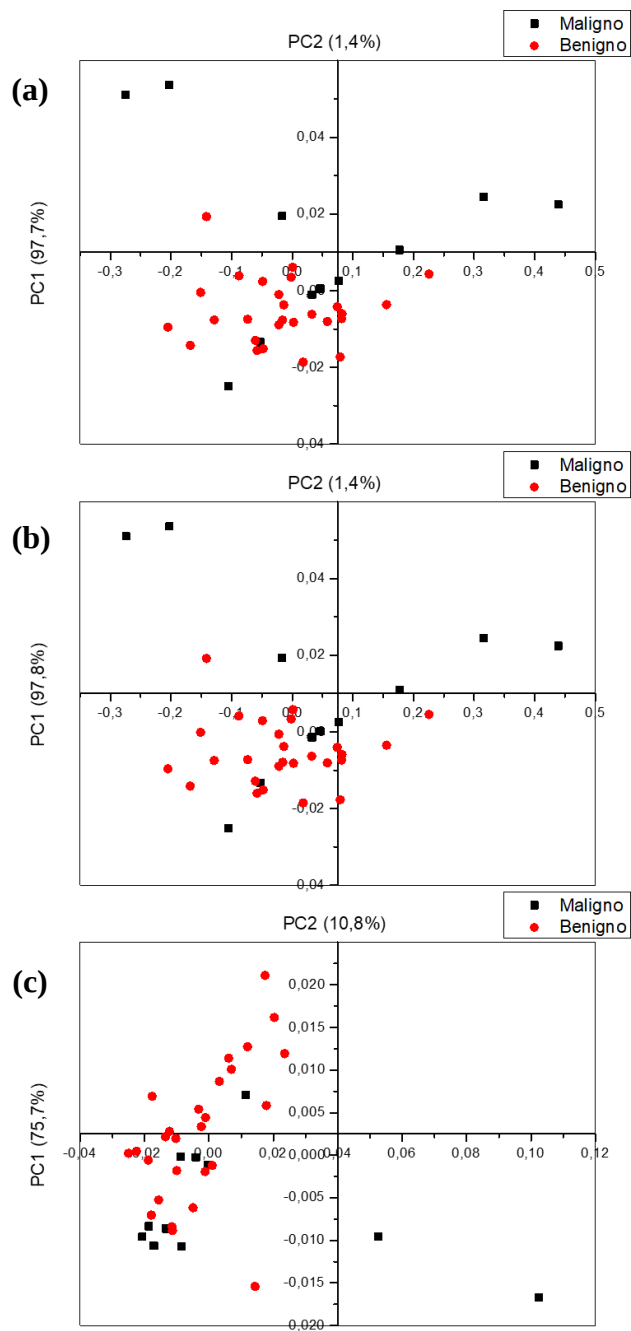
Observa-se que na tabela 6, os espectros brutos e suavizados Savitzky-Golay apresentaram três erros de classificação para as amostras CAT (IV). Para os espectros suavizados Savitzky-Golay + MSC a classificação foi incorreta para todas as amostras de CAT (III). De modo que apresentaram uma SENS global de 84% e ESPEC global de 61%, considerando o conjunto de teste e amostras indefinidas.

De acordo com a tabela 6, os modelos desenvolvidos na predição das amostras de CAT (IV) e CAT (V) não são satisfatório. Ou seja, na prática estes modelos estariam considerando pacientes com câncer de tireoide como pacientes saudáveis, que não é o verdadeiro diagnóstico de acordo com o médico patologista.

Tendo em vista esse panorama e com intuito de melhorar as classificações das amostras indefinidas, optou-se por selecionar a região espectral entre 650 cm^{-1} e 1400 cm^{-1} por ser a região de maior absorção. Do mesmo modo que foi realizado com o

espectro completo, aplicou-se a PCA nos espectros com a região ($1400 - 650 \text{ cm}^{-1}$), podemos observar na figura 11.

Figura 11 - Gráficos dos escores referentes aos espectros ($1400 - 650 \text{ cm}^{-1}$) dos 38 pacientes, PC1 x PC2. (a) sem pré-processamento (b) com suavização Savitzky-Golay; (c) com suavização Savitzky-Golay + MSC. Amostras de tecido saudáveis (●) e amostras de tecido com câncer de tireoide (■)



Após a selecionar a região de espectro (1400 – 650 cm⁻¹) foram utilizados algoritmos de classificação para os espectros sem pré-processamento, com suavização Savitzky-Golay e suavização Savitzky-Golay + MSC. Pode-se observar os resultados obtidos da matriz de confusão na tabela 7, como mostrado na primeira parte dos modelos classificação.

Tabela 8 - Resultados da matriz de confusão para os modelos SIMCA e DD-SIMCA na região entre 1400 e 650 cm⁻¹, para os espectros bruto, com suavização Savitzky-Golay e com suavização Savitzky-Golay + MSC.

SIMCA								
Espectros	PC	Validação	Real/predição	Classe 1 (alvo)	Classe 2 (não alvo)	SENS (%)	ESPEC (%)	AC (%)
Bruto	4	Treinamento	Classe 1 (alvo)	18	0	100	-	-
			Classe 2 (não alvo)	0	2			
		Validação cruzada	Classe 1 (alvo)	18	0	100	-	-
			Classe 2 (não alvo)	1	1			
		Teste	Classe 1 (alvo)	7	2	78	82	78
			Classe 2 (não alvo)	2	9			
Suavização (Savitzky- Golay)	3	Treinamento	Classe 1 (alvo)	18	0	100	-	-
			Classe 2 (não alvo)	0	2			
		Validação cruzada	Classe 1 (alvo)	18	0	100	-	-
			Classe 2 (não alvo)	1	1			
		Teste	Classe 1 (alvo)	6	3	67	82	75
			Classe 2 (não alvo)	2	9			
Suavização (Savitzky- Golay) + MSC	3	Treinamento	Classe 1 (alvo)	9	9	50	-	-
			Classe 2 (não alvo)	1	1			
		Validação cruzada	Classe 1 (alvo)	6	12	33	-	-
			Classe 2 (não alvo)	1	1			
		Teste	Classe 1 (alvo)	4	5	44	73	60
			Classe 2 (não alvo)	3	8			

<i>Cont.</i>		DD-SIMCA						
Espectros	PC	Validação	Real/predição	Classe 1 (alvo)	Classe 2 (não alvo)	SENS (%)	ESPEC (%)	AC (%)
Bruto	3	Treinamento	classe 1 (alvo)	17	1	94	-	-
			classe 2 (não alvo)	0	0			
		Teste	classe 1 (alvo)	6	3	67	91	80
			classe 2 (não alvo)	1	10			
Suavização (Savitzky- Golay)	3	Treinamento	classe 1 (alvo)	18	0	100	-	-
			classe 2 (não alvo)	0	0			
		Teste	classe 1 (alvo)	5	4	67	82	70
			classe 2 (não alvo)	2	9			
Suavização (Savitzky- Golay) + MSC	3	Treinamento	classe 1 (alvo)	18	0	100	-	-
			classe 2 (não alvo)	0	0			
		Teste	classe 1 (alvo)	4	5	44	91	70
			classe 2 (não alvo)	1	10			

De acordo com a tabela 7, os resultados para o modelo SIMCA obtido a partir dos espectros sem pré-processamento requereu 4 PC's, para os espectros com suavização Savitzky-Golay 3 PC's e para os espectros com suavização Savitzky-Golay + MSC 3 PC's. Os valores obtendo de SENS 78% e ESPEC 82% para os espectros sem pré-processamento, apresentando quatro erros de predição para duas amostras malignas e duas benignas. Para os espectros com suavização Savitzky-Golay SENS 67% e ESPEC 82%, apresentando cinco erros de predição para três amostras benignas e duas malignas e para os espectros com suavização Savitzky-Golay + MSC SENS 44% e ESPEC 73%, apresentando oito erros de predição para cinco amostras benignas e três malignas.

O resultado para o modelo DD-SIMCA com espectros brutos requereu de 3 PC's, de modo que o conjunto de treinamento apresentou uma amostra extrema. Devido a esta amostra a SENS resultou em 94%. Esse modelo apresentou 4 erros de predição, sendo três amostras benignas e uma maligna, resultando em SENS e ESPEC de 67% e 91%, respectivamente. Para o modelo obtido com os espectros com suavização Savitzky-Golay foram necessárias 3 PC's. Este modelo apresentou seis erros de predição, sendo quatro

amostras benignas e duas amostras malignas, resultando nos valores de SENS e ESPEC de 67% e 82%, respectivamente. Nos espectros com suavização Savitzky-Golay + MSC requereu 3 PC's, o qual resultou em seis erros de predição, para cinco amostras benignas e uma amostra maligna, obtivemos SENS de 44% e ESPEC de 91%.

Todos os modelos SIMCA e DD-SIMCA obtidos para os espectros (1400 – 650 cm^{-1}), foram empregados no conjunto de amostras indefinidas, assim como anteriormente mostrado, podemos observar na tabela 8. Serão mostrados os resultados que apresentaram menores erros de classificação.

Tabela 9 - Resultado de predição das amostras indefinidas, para o modelo DD-SIMCA para os dados brutos e os espectros suavizados Savitzky-Golay.

Amostras	Classificação	Espectros Suavizados Savitzky-Golay		Amostras	classe	Predição DD-SIMCA
		Diagnóstico Patológico	Predição SIMCA			
1	CAT (V)	Maligna	Maligna	1	Maligna	Maligna
2	CAT (V)	Maligna	Maligna	2	Maligna	Maligna
3	CAT (IV)	Maligna	Maligna	3	Maligna	Maligna
4	CAT (IV)	Maligna	Maligna	4	Maligna	Maligna
5	CAT (IV)	Maligna	Maligna	5	Maligna	Maligna
6	CAT (IV)	Maligna	Maligna	6	Maligna	Maligna
7	CAT (IV)	Maligna	Benigna	7	Maligna	Maligna
8	CAT (III)	Benigna	Benigna	8	Benigna	Benigna
9	CAT (III)	Benigna	Benigna	9	Benigna	Benigna
10	CAT (III)	Benigna	Maligna	10	Benigna	Maligna
11	CAT (III)	Benigna	Benigna	11	Benigna	Benigna

O resultado de classificação para o modelo SIMCA como visto na tabela 8 para os espectros com suavização Savitzky-Golay apresentaram dois erros de classificação, sendo para uma amostra CAT (IV) e uma amostra CAT (III). De modo que para o modelo com os espectros com suavização Savitzky-Golay apresenta uma SENS global de 69% e ESPEC global de 83%. É importante ressaltar que os valores são bons e comparáveis a literatura (DEPCIUCH et al., 2019).

Os resultados de classificação do modelo DD-SIMCA com os espectros com suavização Savitzky-Golay há apenas um erro de classificação para uma amostra CAT

(III). De modo que apresenta uma SENS global de 61% e ESPEC global de 89%, considerando o conjunto de teste e amostras indefinidas. Observe que para os modelos DD-SIMCA não há erro de predição das amostras malignas de categoria (IV) e (V). Podendo ser considerado até o momento o melhor modelo de predição para o conjunto de amostras indefinidas.

Os resultados obtidos com o método proposto são comparáveis aos encontrados na literatura empregando espectroscopia FTIR em conjunto com PCA-LDA, resultando em uma SENS 65% e ESPEC 75% (DEPCIUCH et al., 2019).

Capítulo 6

CONCLUSÕES E PESPECTIVAS

CONCLUSÕES

Com base nos resultados, nota-se que a técnica ATR-FTIR associada com os modelos de classificação SIMCA e DD-SIMCA são capazes de diferenciar os tecidos tireoidianos benignos em comparação aos tecidos patologicamente alterados da glândula tireoide.

Neste estudo foi desenvolvido modelos de classificação de classe única com os algoritmos SIMCA e DD-SIMCA utilizando a técnica de espectroscopia ATR-FTIR. Para o modelo SIMCA com espectros na região $1400 - 650 \text{ cm}^{-1}$, bruto e resultando em valores de SENS global de 77% e ESPEC global de 83%. Para o modelo para os espectros com suavização Savitzky-Golay apresentam uma SENS global de 69% e ESPEC global de 83%.

O resultado para modelo DD-SIMCA para os dados espectrais ($1400 - 650 \text{ cm}^{-1}$) com espectros ($1400 - 650 \text{ cm}^{-1}$), bruto resultando em valores de SENS global de 69% e ESPEC global de 94%. Para os espectros com suavização Savitzky-Golay apresentaram valores de SENS global de 61% e ESPEC global de 89%. Este modelo utilizado em amostras indefinidas que são classificadas pelo patologista de acordo com as categorias III, IV e V. Apresentando menor erro na predição das amostras indefinidas. Levando em consideração que é uma classificação duvidosa para o patologista, podemos considerar que é uma metodologia que poderá auxiliá-lo.

Conclui-se que a metodologia desenvolvida neste estudo é promissora no âmbito de diagnósticos médico para o câncer de tireoide, de modo que o profissional de saúde, poderá executar esta metodologia.

PESPECTIVAS

Embora o método proposto seja promissor para auxiliar no âmbito de diagnóstico de câncer de tireoide, como proposta de continuidade deste trabalho, sugerimos as seguintes propostas para darmos continuidade no desenvolvimento deste trabalho:

- Avaliar a eficiência da técnica de infravermelho próximo (NIR) associada às técnicas de classificação Modelagem Independente Flexível por Analogia de Classes (SIMCA) e Modelagem Independente Flexível por Analogia de Classes Direcionado pelos Dados (Data Driven SIMCA, DD-SIMCA) no desenvolvimento dos modelos;
- Avaliar condições melhores no pré-processamento dos espectros, como suavização Savitzky-Golay em polinômios maiores e maior janela de pontos;
- Avaliar o modelo desenvolvido na classificação de amostras de categorias III, IV e V.

REFERÊNCIAS

BAKER, M. J.; TREVISAN, J.; BASSAN, P.; BHARGAVA, R.; BUTLER, H. J.; DORLING, K. M.; FIELDEN, P. R.; FOGARTY, S. W.; FULLWOOD, N. J.; HEYS, K. A.; HUGHES, C.; LASCH, P.; HIRSCH, P. L. M.; OBINAJU, B.; SOCKALINGUM, G. D.; SUSO, J. S.; STRONG, R. J.; WALSH, M. J.; WOOD, GARDNER, P.; MARTIN, F. L. Using Fourier transform IR spectroscopy to analyze biological materials. **Nature Protocols**, v. 9, p. 1771-1791, 2014.

BRASIL. Conselho Regional de Medicina do Estado de São Paulo – CREMESP. Disponível em: https://www.cremesp.org.br/?siteAcao=Imprensa&aacao=crm_midia&id=706#:~:text=S egundo%20pesquisa%20do%20Conselho%20Federal,mais%20numerosa%2C%20com %2030.112%20titulados>. Acessado em 6 de abril de 2022.

BRASIL. Instituto Nacional de Câncer - INCA. Disponível em: <<https://www.inca.gov.br/tipos-de-cancer/cancer-de-tireoide>>. Acesso em 12 de maio de 2021.

BRASIL. BLOG DA SAÚDE. Disponível em: <<http://www.blog.saude.gov.br/promocao-da-saude/51108-saiba-sobre-a-grande-importancia-da-pequena-glandula-tireoidehtml.html>>. Acesso em 12 de maio de 2021.

BORGES, A. K. M.; FERREIRA, J. D.; KOIFMAN, S.; KOIFMAN, R. J. Thyroid câncer in Brazil: a descriptive study of case held on hospital-based câncer registries, 2000-2016, **Epidemiol. Serv. Saúde**, v. 29(4), p.1-11, 2020.

BURY, D.; MORAIS, C. L. M.; PARASKEVAIDI, M.; ASHTON, K. M.; DAWSON, T. P.; MARTIN, F. L. Spectral classification for diagnosis involving numerous pathologies in a complex clinical setting: A neuro-oncology example. **Spectrochim Acta A Mol Biomol Spectrosc**, v. 206, p. 89-86, 2019.

BRITO, J. P.; MORRIS, J. C.; MONTORI, V. M. Thyroid câncer: zealous imaging has increased detection and treatment of low risk tumours. **BMJ**, v. 347, p. 1-6, 2013.

CAIXETA, D. .G. **Diagnóstico salivar de diabetes mellitus utilizando espectroscopia de reflexão total atenuada no Infravermelho com transformada de fourier (ATR-FTIR)**. 2018. Dissertação (Mestrado em Ciência da Saúde). Universidade Federal de Uberlândia, Uberlândia, 2018.

DAVIES, L.; WELCH, H. G. Increasing Incidence of thyroid câncer in the United State, 1973-2002. **Journal of the American Medical Association**, v. 295, p. 2164-2167, 2006.

DEPCIUCH, J.; STANEK-WIDERA, A.; SKRZYPIEC, D.; LANGE, D.; BISKUP-FRUZYNSKA, M.; KIPER, K.; STAMEK-TARKOWSKA, J.; KULA, M.; CEBULSKI, J. Spectroscopic identification of benign (follicular adenoma) and cancerous lesions (follicular thyroid carcinoma) in thyroid tissues. **Journal of Pharmaceutical and Biomedical Analysis**, v. 170, p. 321-326, 2019.

ENEWOLD, L.; ZHU, K.; RON, ELAINE. Rising thyroid câncer incidence in the United State by demographic and tumor characteristics, 1980-2005. **Cancer Epidemiology, Biomarkers & Prevention**, v. 18, p. 784-791, 2009.

FERREIRA, M. M. C. **Quimiometria: Conceitos, métodos e aplicações**. Edição 1ª. Campinas, SP. Editora da Unicamp, 2015.

GHAZY, E.; KUMAR, A.; BARANI, M.; KAUR, I.; RAHDAR, A. Scrutinizing the therapeutic and diagnostic potential of nanotechnology in thyroid cancer: Edifying drug targeting by nano-oncotherapeutics. **Journal of Drug Delivery Science and Technology**, v. 61, p. 1-14, 2020.

HANSEN, L. **Análise de Dados Físico-Químicos de amostras de leite cru do Sul do Brasil utilizando métodos multivariados exploratórios e classificatórios**. 2019. Dissertação (em Química) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

HARRIS, A. T.; GARG, M.; YANG, X. B.; FISHER, S. E.; KIRHAM, J.; SMITH, D. A.; MARTIN-HIRSCH, D. P.; HIGH, A. Raman spectroscopy and advanced mathematical modelling in the discrimination of human thyroid cell lines. **BioMed Central**, v. 38, p. 1-6, 2009.

KELLY, J. G.; TREVISAN, J.; SCOTT, A. D.; CARMICHAEL, P. L.; POLLOCK, H. M.; HIRSCH, P. L. M.; MARTIN, F. L. Biospectroscopy to metabolically profile biomolecular structure: a multistage approach linking computational analysis with biomarkers. **Journal of proteome research**, v. 10, p. 1437-1438, 2011.

KHAN YS, FARHANA A. **Histology, Thyroid Gland**. 2021 Dec 8. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan–. PMID: 31869123.

KILFOY, B. A.; WARD, M. H.; SABRA, M. M.; DEVESA, S. S. Thyroid Cancer Incidence Patterns in the United States by Histologic Type, 1992–2006. **Thyroid**, v. 21(2), p. 125-134, 2011.

KOWALSKI, B. R.; BENDER, C. F. Pattern Recognition. A Powerful Approach to Interpreting Chemical Data. **Journal of the American Chemical Society**, v. 94, p. 5632-5638, 1971.

KOWALSKI, B. R. Chemometrics: Views and Propositions. **Journal of Chemical Information and Computer Sciences**, v. 15, p. 201-203, 1975.

LA VECCHIA, C.; MALVEZZI, M.; BOSETTI, C.; GARAVELLO, W.; BERTUCCIO, P.; LEVI, F.; NEGRI, EVA. Thyroid câncer mortality and incidence: A global overview. **International Journal of Cancer**, v. 136 (9), p. 2187-2195, 2014.

LAYFIELD, L. J.; CIBAS, E. S.; BALOCH, Z. Thyroid fine needle aspiration cytology: a review of the National Cancer Institute state of the science symposium. **Cytopathology**, v. 21, p. 75-85, 2010.

LAYFIELD, L. J.; ABRAMS, J.; PRIOLLET, B. C.; EVANS, D.; GHARIB, H.; GREENSPAN, F.; HENRY, M.; LIVOLSI, V.; MERINO, M. MICHAEL, C. W.; WANG, H.; WELLS, S. Treatment Options: A Synopsis of the National Cancer Institute Thyroid Fine Needle Aspiration State of the Science Conference. **Diagnostic Cytopathology**, v. 36, p. 442-448, 2008.

LEENHARDT, L.; BERNIER, M. O.; BOIN-PINEAU, M. H.; DEVOLX, B. C.; MARÉCHAUD, R.; NICCOLI-SIRE, P.; NOCAUDIE, M.; ORGIAZZI, J.; SCHLUMBERGER, M.; WÉMEAU, M.; CRÉRIÉ-CHALLINE, L.; DE VATHAIRE, F., Advances in diagnostic practices affect thyroid câncer incidence in france. **European Journal of Endocrinology**, v. 150, p. 133-139, 2004.

LI, L.; WU, J.; YANG, L.; WANG, H.; XU, Y. SHEN, K. Fourier Transform Infrared Spectroscopy: Na Innovative Method for the Diagnosis of Ovarian Cancer. **Cancer Management and Research**, v. 13, p. 2389-2399, 2021.

LIU, Y.; XU, Y.; ZHANG, Y.; WANG, D.; XIU, D.; XU, Z.; ZHOU, X.; WU, J.; LING, X. Detection of cervical metastatic lymph nodes in papillary thyroid carcinoma by Fourier transform infrared spectroscopy. **British Journal of Surgery**, v. 98, p. 380-384, 2011.

MALBURG, C. E. L.; SUPLICI, L. M. E. A. **Correlação entre punção – Biópsia aspirativa por agulha fina e biópsia de congelação em nódulos tireoidianos**. 1991. Monografia (Graduação de Medicina) – Universidade Federal de Santa Catarina, Florianópolis, 1991.

MEDEIROS, A. R. B. **Uso de ATR/FTIR e FTNIR associado a técnica quimiométricas para quantificação de aditivos em gasolina automotiva**. 2009. Dissertação (em Química) – Universidade de Brasília, Brasília, 2009.

MOVASAGHI, Z.; REHMAN, S.; REHMAN, I. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. **Applied Spectroscopy Reviews**, v. 43, p. 134-179, 2008.

NA, D. G.; BAEK, J. H.; JUNG, S. L.; KIM, J. H.; SUNG, J. Y.; KIM, K. S.; LEE, J. H.; SHIN, J. H.; CHOI, Y. J.; HA, E. J.; LIM, H. K.; KIM, S. J.; HAHN, S. Y.; LEE, K. H.; CHOI, Y. J.; YOUN, I.; KIM, Y. J.; AHN, H. S.; RYU, J. H.; BAEK, S. M.; SIM, J. S.; JUNG, C. K.; LEE, J. H. Core Needle Biopsy of the Thyroid: 2016 Consensus Statement and Recommendations from Korean Society of Thyroid Radiology. **Korean Journal of Radiology**, v. 18, p. 217-237, 2017.

NAGAR, S.; ASCHEBROOK-KILFOY, B.; KAPLAN, E. L.; ANGELOS, P.; GROGAN, R. H. Age of diagnosing physician impacts the incidence of thyroid câncer in a population. **Cancer Causer Control**, v.25, p. 1627-1634, 2014.

National Cancer Institute – NIH. Disponível em: <<https://www.cancer.gov/publications/dictionaries/cancer-terms/def/lymph-node>>. Acesso em 29 de março de 2022.

NETO, B. B.; SCARMINIO, I. S.; BRUNS, R. E. 25 anos de Quimiometria no Brasil. **Química Nova**, v. 29, p. 1401-1406, 2006.

NÓBREGA, R. O. **Classificação de cafés solúveis usando espectroscopia NIR e quimiometria**. 2021. Dissertação (em Química) – Universidade Federal da Paraíba, João Pessoa, 2021.

OLIVIERI, P.; DOWNEY, G. Multivariate class modeling for the verification of food-authenticity claims. **Trends in Analytical Chemistry**, v. 35, p. 74-86, 2012.

PARASKEVAIDI, M.; MORAIS, C. L. M.; LIMA, K. M. G.; MANN, D. M. A.; ALLSOP, D.; MARTIN-HIRSCHD, P. L.; MARTIN, F. L. Vibrational spectroscopy of blood plasma for the diagnosis of alzheimer's disease and differentiation from dementia with lewy bodies. **Alzheimer's Association International Conference: P1: Poster Presentations**, 2017.

PARASKEVAIDI, M.; MORAIS, C. L. M.; LIMA, K. M. G.; ASHTON, K. M.; STRINGFELLOW, H. F.; MARTIN-HIRSCHD, P. L.; MARTIN, F. L. Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer. **Royal Society of Chemistry**, v. 143 (13), p. 3156-3163, 2018.

PASSOS, J. O. S.; ALVES, M. V. S.; MORAIS, C. L. M.; MARTIN, F. L.; CAVALCANTE, A. F.; LEMOS, T. M. A. M.; MOURA, S.; FREITAS, D. L. D.; MARIZ, J. V. M.; CARVALHO, J. L.; LIMA, K. M. G.; PEGADO, R. Spectrochemical analysis in blood plasma combines with subsequent chemometrics for fibromyalgia detection. **SCIENTIFIC REPORTS**, v. 10, p. 1-8, 2020.

PECCIN, S.; FURLANETTO, T. W.; DE CASTRO, J. A. S.; FURTADO, A. P. A.; BRASIL, A. A.; CZEPIELEWIKI, M. A. Nódulo de tireoide: Valor da ultra-sonografia e da biópsia por punção aspirativa no diagnóstico de câncer. **Revista da associação médica brasileira**, v. 49 (2), p. 145 – 149, 2003.

PENNA, G. C.; VAISMAN, F. VAISMAN M.; SIMÕES, M. S.; SOARES, P. Molecular Markers Involved in Tumorigenesis of Thyroid Carcinoma: Focus on Aggressive Histotypes. **Cytogenetic and Genome Research**, v. 150, p. 194-207, 2017.

POMERANTSEV, A. L.; RODIONOVA, O. Y. Concept and role of extreme objects in PCA/SIMCA. **Journal of chemometrics**, v. 28, p. 429-238, 2013.

PONTES, M. J. C. **Algoritmo das projeções sucessivas para seleção de variáveis espectrais em problemas de classificação**. 2009. Tese (Doutorado em Química) – Universidade Federal da Paraíba, João Pessoa, 2009.

RAHIB, L.; SHITH, B. D.; AIZENBERG, R.; ROSENZWEIG, A. B.; FLESHMAN, J. M. MATRISIAN, L. M. Projecting Cancer Incidence and Deaths to 2030: The Unexpected Burden of Thyroid, Liver, and Pancreas Cancers in the United States. **American Association for Cancer Research**, v. 74, p. 2913-2923, 2014.

RAU, J. V.; GRAZIANI, V.; FOSCA, M.; TAFFON, C.; ROCCHIA, M.; CRUCITTI, P.; POZZILLI, P.; MUDA, A. O.; CARICATO, M.; CRESCENZI, A. RAMAN spectroscopy imaging improves the diagnosis of papillary thyroid carcinoma. **NATURE - SCIENTIFIC REPORTS**, v. 6, p. 1-10, 2016.

RIBEIRO, F. A. L.; MANTOVANI, G. A.; POPPI, R. J. PCA: UMA FERRAMENTA PARA IDENTIFICAÇÃO DE TRAÇADORES QUÍMICOS PARA ÁGUA DE FORMAÇÃO E ÁGUA DE INJEÇÃO ASSOCIADAS À PRODUÇÃO DE PETRÓLEO. **Química Nova**, v. 36, p. 1281-1287, 2013.

RODIONOVA, O. Y.; BALKLOVA, K. S.; TIVONA, A. V.; POMERANTSEV, A. L. Quantitative risk assessment in classification of drugs with identical API content. **Journal of Pharmaceutical and Biomedical Analysis**, v. 98, p. 186-192, 2014.

RODIONOVA, O. Y.; TIVONA, A. V.; POMERANTSEV, A. L. Discriminant analysis is an inappropriate method of authentication. **TrAC Trends in Analytical Chemistry**, v. 78, p. 17-22, 2016.

RODIONOVA, O. Y.; POMERANTSEV, A. L. Chemometric tools for food fraud detection: The role of target class in non-targeted analysis. **Food Chemistry**, v. 317, p. 126448, 2020.

SABIN, J. G.; FERRÃO, M. F.; FURTADO, J. C. Análise multivariada aplicada na identificação de fármacos antidepressivos. Parte II: Análise por componentes principais (PCA) e o método de classificação SIMCA. **Revista Brasileira de Ciências Farmacêuticas**, v. 40, p. 387-396, 2004.

SANTOS, M. C. D.; NASCIMENTO, Y. M. MONTEIRO, J. D.; ALVES, B. E. B.; MELO, M. F.; PAIVA, A. A. P.; PEREIRA, H. W. B.; FERNANDES, J. V.; ARAÚJO,

J. M. G.; LIMA, K. M. G. ATR-FTIR spectroscopy with chemometric algorithms of multivariate classification in the discrimination between healthy vs. Dengue vs. Chikungunya vs. Zika clinical samples. **Royal Society of Chemistry**, v. 10, p. 1280-1285, 2018.

SCHEFFER, M.; CASSENOTE, A.; GUERRA, A.; GUILLOUX, A. G. A.; BRANDÃO, A. P. D.; MIOTTO, B. A.; ALMEIDA, C. J.; GOMES, J. O.; MIOTTO, R. A. **Demografia Médica no Brasil 2020**. São Paulo, SP: FMUSP, CFM, 2020. 312 p. ISBN:978-65-00-12370-8.

SIQUEIRA, L. F. S.; JÚNIOR, R. F. A.; DE ARAÚJO, A. A.; MORAIS, C. L. M.; LIMA, K. M. G. LDA vs QDA for FT-MIR prostate câncer tissue classification. **Chemometrics and Intelligent Laboratory Systems**, v. 162, p. 123-129, 2017.

SOUZA, A. M.; POPPI, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, Parte I. **Química Nova**, v. 35, p. 223-229, 2012.

SILVA, S. F.. **Análise de triagem para confirmação do estado de conservação de insulina humana NPH injetável**. 2020. Dissertação (Mestrado em Química) – Universidade Federal da Paraíba, João Pessoa, 2020.

SKOOG, D. A.; HOLLER, J. F.; NIEMAN, AT. T. **Princípios de análise instrumental**. 5º edição. Porto Alegre: Bookman, 2002.

ULLAH, R.; KHAN, S.; FARMAN, F.; BILAL, M.; KRAFFT, C.; SHAHZAD, S. Demonstrating the application of Raman spectroscopy together with chemometric

technique for screening of asthma disease. **Biomedical Optics Express**, v. 10 (2), p.600-609, 2019.

VACCARELLA, S.; MASO, L. D.; LAVERSANNE, M.; BRAY, F.; PLUMMER, M.; FRANCESCHI, S. The impact of diagnostic changes on the rise in thyroid cancer incidence: A population-based study in selected high-resource countries. **THYROID**, v. 25 (10), p. 1127-1136, 2015.

VALDERRAMA, L.; PAIVA, V. B.; MARÇO, P. H.; VALDERRAMA, P. PROPOSTA EXPERIMENTAL DIDÁTICA PARA O ENSINO DE ANÁLISE DE COMPONENTES PRINCIPAIS. **Química Nova**, v. 39, p. 245-249, 2016.

WALSH, M. J.; SINGH, M. N.; STRINGFELLOW, H. F.; POLLOCK, H. M.; HAMMICHE, A.; GRUDE, O.; FULLWOOD, N. J.; PITT, M. A.; HIRSCH, P. L. M.; MARTIN, F. L. FTIR Microspectroscopy Coupled with Two-Class Discrimination Segregates Markers Responsible for Inter- and Intra-Category Variance in Exfoliative Cervical Cytology. **Biomarker Insights**, v. 3, p. 179-189, 2008.

WANG, H.; CHEN, C.; TONG, D.; CHEN, C.; GAO, R.; HAN, H.; LV, X. Serum Raman spectroscopy combines with multiple algorithms for diagnosing thyroid dysfunction and chronic renal failure. **Photodiagnosis and Photodynamic Therapy**, v. 34, p. 1-4, 2021.

WARD, L. S.; GRAF, H., Câncer da tireoide: Aumento na Ocorrência da doença ou simplesmente na sua detecção?. Curitiba, Paraná. **Arq. Bras. Endocrinol Metab**, 2008, p. 1515-1516.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal Component Analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 2, p. 37-52, 1987.

XAVIER, M. A. C.; TAVARES, N. N. P.; MATOS, R. A.; GUEDES, V. R. REVISÃO BIBLIOGRÁFICA – CARCINOMA INDIFERENCIADO DA TIREÓIDE. **Revista de Patologia do Tocantins**, v. 1, p. 7-13, 2014.

YOKOZAWA, T. Câncer da tireóide detectado pela punção aspirativa por agulha fina guiada pelo Ultra-som. **Arq. Bras. Endocrinol Metab**, v. 42, p. 296-298, 1998.

ZHANG, X.; XU, Y.; ZHANG, Y.; WANG, L.; HOU, C.; ZHOU, X.; LING, X.; XU, Z. Intraoperative Detection of Thyroid Carcinoma by Fourier Transform Infrared Spectrometry. **Journal of Surgical Research**, v. 171, p. 650-656, 2011.

ZONTOV, Y. V.; RODIONOVA, O. Y.; KUCHERYAVSKIY, S. V.; POMERANTSEV, A. L. DD-SIMCA — a MATLAB GUI tool for data driven SIMCA approach. **Chemometrics and Intelligent laboratory systems**, v. 17, p. 1-14, 2017.